THEMIS SIGNAL ANALYSIS STATISTICS RESEARCH PROGRAM

PARTIAL PRIOR INFORMATION: SOME EMPIRICAL BAYES AND G-MINIMAX DECISION FUNCTIONS

by

Stephen L. George

Technical Report No. 48
Department of Statistics THEMIS Contract

Reproduction in whole or in part is permitted
for any purpose of the United States Government.

This document has been approved for public release
and sale; its distribution is unlimited.

THEMIS SIGNAL ANALYSIS STATISTICS RESEARCH PROGRAM

PARTIAL PRIOR INFORMATION: SOME EMPIRICAL BAYES AND G-MINIMAX DECISION FUNCTIONS

by

Stephen L. George

Technical Report No. 48
Department of Statistics THEMIS Contract

October 30, 1969

Research sponsored by the Office of Naval Research
Contract N00014-68-A-0515
Project NR 042-260

Reproduction in whole or in part is permitted
for any purpose of the United States Government.

This document has been approved for public release
and sale; its distribution is unlimited.

DEPARTMENT OF STATISTICS
Southern Methodist University

PARTIAL PRIOR INFORMATION: SOME EMPIRICAL BAYES AND $\mathcal{G}$-MINIMAX DECISION FUNCTIONS

Approved by:

PARTIAL PRIOR INFORMATION: SOME EMPIRICAL BAYES AND G-MINIMAX DECISION FUNCTIONS

A Thesis Presented to the Faculty of the Graduate School

of

Southern Methodist University

in

Partial Fulfillment of the Requirements

for the degree of

Doctor of Philosophy

with a

Major in Statistics

by

Stephen L. George
(B.A., Texas Technological College, 1965)
(M.E.S., North Carolina State University, 1967)

October 30, 1969

George, Stephen L.

B.A., Texas Technological College, 1965
M.E.S., North Carolina State University, 1967

<u>Partial Prior Information: Some Empirical Bayes and $\mathcal{G}$-Minimax Decision Functions</u>

Adviser: Associate Professor Richard P. Bland

Doctor of Philosophy degree conferred December 20, 1969

Dissertation completed October 30, 1969

If the prior probability distribution function G of the parameters in a statistical decision theory model is not completely specified then a Bayes decision function cannot be obtained. However, in many cases there may be some partial (i.e., incomplete) prior information concerning G.

The types of partial prior information considered in this paper is of two kinds: (i) N past observations on the compound distribution $f_G(t)$ or (ii) knowledge of a restrictive class $\mathcal{G}$ to which the unknown G is assumed to belong.

When past observations are available, empirical Bayes decision functions are found which are: (i) asymptotically optimal (ii) easy to apply and (iii) better than some optimal non-Bayes decision function for reasonably small N. When only knowledge of the class $\mathcal{G}$ is assumed, $\mathcal{G}$-minimax decision functions are found.

In all cases, estimators for the parameters of the normal and Bernoulli processes are found using a quadratic loss function. These

estimators are then compared to each other and to other well-known estimators by means of their Bayes risks and, in the empirical Bayes case, by means of their global risks for small N.

ACKNOWLEDGEMENTS

I wish to thank my major adviser, Dr. Richard P. Bland for his advice and helpful comments given during the writing of this paper. In addition I would like to thank Dr. John T. Webster and Dr. Donald B. Owen for their constructive criticism of certain aspects of the paper.

This paper was supported in part by National Institute of Health grant number 2 TOL GM00951-08

TABLE OF CONTENTS

		Page
ABSTRACT		iv
ACKNOWLEDGMENTS		vi
LIST OF TABLES		ix
LISTS OF ILLUSTRATIONS		x
Chapter		
I.	Introduction	1
	1.1 Bayesian Approach	1
	1.2 G-Minimax Decision Functions	3
	1.3 Empirical Bayes Decision Functions	5
	1.4 Problems Considered	8
II.	PRELIMINARY RESULTS	11
	2.1 General Structural Elements	11
	2.2 Bayes Decision Functions	14
	2.3 Empirical Bayes Decision Functions	18
	2.4 G-Minimax Decision Functions	24
III.	NORMAL PROCESS I: EMPIRICAL BAYES DECISION FUNCTIONS	34
	3.1 Definitions and Initial Results	34
	3.2 Estimation of λ_{N+1}	37
	3.3 Numerical Results	51
IV.	NORMAL PROCESS II: G-MINIMAX DECISION FUNCTIONS	62
	4.1 Introduction	62
	4.2 Estimation of λ_{N+1}: G Conjugate	63
	4.3 Estimation of λ_{N+1}: Moments Known	72
	4.4 Asymptotic G-Minimax Decision Functions	76

Page

V. BERNOULLI PROCESS I: EB DECISION FUNCTIONS 80

5.1 Definitions and Initial results. 80
5.2 Estimation of λ_{N+1} 87
5.3 Numerical Results. 109

VI. BERNOULLI PROCESS II: $\mathcal{G}$-MINIMAX DECISION FUNCTIONS 128

6.1 Introduction 128
6.2 Estimation of λ_{N+1}: G conjugate 129
6.3 Estimation of λ_{N+1}: Moments known 136
6.4 Asymptotic $\mathcal{G}$-Minimax Decision Functions. . . 140

APPENDIX 144

LIST OF REFERENCES 161

LIST OF TABLES

Table Page

1. Global Risks of Various Decision Functions Using a N(0,1) Prior. 58

2. Global Risks of Various Decision Functions Using a N(0,.1) Prior. 59

3. Global Risks (N'=20, R'=10) 115

4. Global Risks (N'=6, R'=3) 116

5. Global Risks (N'=2, R'=1) 117

6. Global Risks (N'=24, R'=6). 118

7. Global Risks (N'=12, R'=3). 119

8. Global Risks (N'=4, R'=1) 120

9. Values of N_0. 121

LIST OF ILLUSTRATIONS

Figure | Page

1. $R(\delta)$ as a Function of N (M'=0, N'=1, M=5). 60
2. $R(\delta)$ as a Function of N (M'=0, N'=10, M=5) 61
3. Bayes Risk of $\delta_{G_1^+}$ and δ_T as a Function of μ_2 68
4. Upper and Lower Limits of μ and μ_2 86
5. $R(\delta)$ as a Function of N (N'=20, R'=10, M=5). 122
6. $R(\delta)$ as a Function of N (N'=6, R'=3, M=5). 123
7. $R(\delta)$ as a Function of N (N'=2, R'=1, M=5). 124
8. $R(\delta)$ as a Function of N (N'=24, R'=1, M=5) 125
9. $R(\delta)$ as a Function of N (N'=12, R'=3, M=5) 126
10. $R(\delta)$ as a Function of N (N'=4, R'=1, M=5). 127
11. Preference Regions for δ_0 and δ_T 132
12. Preference Regions for δ_1 and δ_T 137

INTRODUCTION

1.1 The Bayesian Approach

In the usual formulation of problems in statistical decision theory the probability distribution function of an observable random variable X defined on a sample space $\mathcal{X}$ is assumed to be a member of some specified class $\mathcal{F}$ of probability distribution functions. The elements $F_\lambda \in \mathcal{F}$ are indexed by the so-called "parameter space" Λ with generic element λ. We assume that corresponding to each F_λ in $\mathcal{F}$ there exists a probability density function (p. d. f.) $f_\lambda = f(x|\lambda)$ with respect to some σ-finite measure μ defined on the space $\mathcal{X}$.

In the non-Bayesian formulation no prior information regarding the unknown parameter λ is assumed to exist. Optimality criteria can then be found which make no use of prior information and suitable decision functions derived therefrom.

However, in some situations it is reasonable to assume that λ is itself a random variable so that there exists a prior distribution function G of λ on Λ. In the remainder of this paper we use $\tilde{\lambda}$ to denote a random variable taking values in Λ and λ for a particular outcome of the random variable $\tilde{\lambda}$.

Furthermore we assume that for any G there exists a p.d.f. $g = g(\lambda)$ with respect to some σ-finite measure μ^* defined on the space Λ. In this case, the probability distribution function $F(x|\lambda)$ in effect during a particular experiment is thought of as having been randomly selected from $\mathcal{F}$ according to the distribution function G on Λ. Conceptually, the process leading to a particular observation of the random variable X may be thought of as first randomly selecting an element λ_0 from Λ according to the probability distribution function $G(\lambda)$ and then generating an observation (or observations) on $\mathcal{X}$ according to the p.d.f. $f(\cdot|\lambda_0)$. If the particular G in effect in a given decision situation is completely specified then the problem is reduced to the standard Bayes decision problem and a Bayes decision function can be found. In general, this involves defining a non-negative real-valued loss function $L(\delta(X),\lambda)$ on $A \times \Lambda$ where A is the action space and $\delta(\cdot)$ is a decision function $\delta\colon \mathcal{X} \to A$, then finding the "risk":

$$R(\delta,\lambda) = \int_{\mathcal{X}} L(\delta(x),\lambda) f(x|\lambda)\, d\mu(x) \qquad (1.1.1)$$

for any decision function (d.f.) δ for a particular λ and finally minimizing the "average" or "Bayes" risk

$$R(\delta,G) = \int_{\Lambda} R(\delta,\lambda)\, dG(\lambda). \qquad (1.1.2)$$

That is, we try to find a d.f. δ_G such that

$$R(\delta_G,G) \le R(\delta,G) \tag{1.1.3}$$

for any other d.f. δ. Then

$$R(\delta_G,G) = R(G) = \inf_{\delta} R(\delta,G) \tag{1.1.4}$$

is the minimum Bayes risk attainable and $R(G)$ is often called the Bayes envelope function.

1.2 $\mathcal{G}$-Minimax Decision Functions

Suppose that the Bayesian formulation of the problem is accepted and hence the Bayes d.f. δ_G is the best attainable but G is not completely specified. In this case, it is assumed only that G is a member of some class $\mathcal{G}$ of distribution functions on Λ where $\mathcal{G}$ may be unrestrictive; that is, $\mathcal{G}$ may be the class of <u>all</u> probability distribution functions G on Λ. Of course the minimum Bayes risk cannot be achieved in this situation but by using the information on the class $\mathcal{G}$ it may be possible to obtain a decision function δ_1 which is at least as good as some optimum decision function δ_0 which ignores such information. Hopefully, it will be possible in many cases to have $R(\delta_1,G) \le R(\delta_0,G)$ for all G in $\mathcal{G}$.

To this end, the following concept of a $\mathcal{G}$-minimax (or "minimax in $\mathcal{G}$") decision function will be useful: δ^* is a $\mathcal{G}$-minimax decision function if $\sup_{G \varepsilon \mathcal{G}} R(\delta^*,G) \leq \sup_{G \varepsilon \mathcal{G}} R(\delta,G)$ for any other d.f. δ. If such a d.f. δ^* exists for the class $\mathcal{G}$ it will be used as the "best" obtainable d.f. when no additional prior information is available. Of course the ease of finding such decision functions depends primarily on the class $\mathcal{G}$.

Cote and Skibinsky [7] solve such a problem when $\mathcal{G}$ is assumed to be the family of all distributions on Λ such that $P(\lambda_1 < \tilde{\lambda} < \lambda_2) > 1-\alpha$. Then in the binomial and normal case they find an estimator better than the usual estimator in each case. For the binomial case, Weiler [48] assumes $\mathcal{G}$ to be the Beta family and also assumes one of three additional bits of information are known: (i) prior observations $x_1, \ldots, x_N$ are present (ii) $\int_\Lambda \lambda dG(\lambda) = \mu$ is known (mean of prior) with the probability of exceeding 2μ also known, and (iii) $P(\lambda_1 < \tilde{\lambda} < \lambda_2)$ is known. In each case he obtains estimators of the parameters of the prior distribution without reference to $\mathcal{G}$-minimax optimality. Blum and Rosenblatt [3] specify $\mathcal{G}$ as $\mathcal{G}_{F,p} = \{G: G=pF+(1-p)H\}$ where F and H are distributions on Λ, $o<p<1$, F known, H arbitrary.

The risks for various $\mathcal{G}$-minimax decision functions derived in this paper are compared with the usual decision function in order to demonstrate the reduction in average risk possible for various amounts of prior knowledge concerning the class $\mathcal{G}$. These cases are described more fully below.

1.3 Empirical Bayes Decision Functions

In addition to knowledge of the class $\mathcal{G}$ of distribution functions on Λ we may have N independent "past" observations $x_1,\ldots,x_N$ on the random variable X such that each x_i is a random outcome from the p.d.f. $f(X|\lambda_i)$ with the λ_i's themselves chosen randomly from a fixed G in $\mathcal{G}$. That is, we have an independent sequence of bivariate random variables $(X_1,\tilde{\lambda}_1),\ldots,(X_N,\tilde{\lambda}_N)$, $(X_{N+1},\tilde{\lambda}_{N+1})$ where the X_i's are observable random variables and the $\tilde{\lambda}_i$'s are unobservable random variables. The observations $x_1,\ldots,x_N$ are called "past" observations, x_{N+1} is the "current" observation and the problem consists of making some inference concerning the unknown λ_{N+1}. Note that each x_i has the "compound" or "unconditional" distribution

$$f_G(x) = \int_\Lambda f(x|\lambda)\,dG(\lambda). \tag{1.3.1}$$

Robbins[33] was the first to explore the possibility of using the past observations to approximate to certain Bayes procedures and in so doing is credited by Neyman [26] for achieving a "breakthrough" in the theory of statistical decision making. Robbins coined the phrase "empirical Bayes" (EB) to refer to any procedure which used the past observations in a systematic manner to approximate to the optimum Bayes procedures. Some writers insist that the decision function possess the property of asymptotic optimality (defined below) in order to be called an empirical Bayes d.f. but the term is used in this paper in its looser sense of a sequence $\{\delta_N\}$ of d.f.'s of the form

$$\delta_N(\cdot) = \delta_N(x_1,\ldots,x_N;\cdot) \tag{1.3.2}$$

with values in A. That is, δ_N is a function of x_{N+1} whose form depends on $x_1,\ldots,x_N$. Thus it is possible to speak of asymptotically optimal (a.o.) EB procedures and non-a.o. EB procedures. It is shown later that some non-a.o. EB d.f.'s at least have some desirable properties for small N.

Formally, we say that an EB procedure $\{\delta_N\}$ is asymptotically optimal iff

$$\lim_{N\to\infty} E[R_N(\delta_N,G)] = R(\delta_G,G) \tag{1.3.3}$$

for any $G\varepsilon\mathcal{G}$ where δ_G is the Bayes d.f., E denotes expectation with respect to the N past observations $\{x_1,\ldots,x_N\}$ and $R_N(\delta_N,G)$ is the Bayes risk for the given (fixed) set of past observations. The quantity $E[R_N(\delta_N,G)]$ is called the "global risk" of the d.f. δ_N for a given G and will be denoted by $R(\delta_N,G)$ to distinguish it from the risk function $R_N(\delta_N,G)$.

The empirical Bayes decision situation has received considerable attention from various sources. Prominent contributors to the EB literature and the closely related compound decision theory are Robbins [32], Samuel [39], and Johns [18]. More recently, articles by Deely and Zimmer [12], Maritz [22] and Krutchkoff [20] have appeared.

Many of the early EB d.f.'s are "simple" in the sense that they merely attempt to estimate the Bayes d.f. δ_G and although a.o., often converge so slowly that for small N some optimum non-Bayes procedure which completely ignores the previous samples often has a smaller Bayes risk. In fact, Maritz [24] shows that some of these early proposals require an inordinately large number of past observations (e.g. N = 1000 in one well-known case) to achieve a global risk as small as the Bayes risk for certain non-Bayes procedures.

Also, a substantial amount of attention is given to proving the existence of a.o. EB d.f.'s without actually giving a constructive procedure for explicitly finding the function. Even when constructive procedures are proposed (Deely and Kruse [11] and Maritz [23]), they are often difficult to apply since they involve linear programming and minimization of integrals, problems which have to be submitted to a computer to obtain even the simplest of results. In general, it can be fairly stated that very little attention has been given to "small sample" properties of the decision functions (i.e., small numbers of previous samples) although some recent work appears aimed in that direction.

A major purpose of this paper is to exhibit empirical Bayes procedures in a number of problems that often occur in theory and practice which are:

(i) asymptotically optimal (if possible)

(ii) easy to calculate and

(iii) better than the "usual" d.f. for reasonably small values of N (in some cases, N=1 suffices).

It is shown below that if one is willing to assume a reasonable amount of restriction on $\mathcal{G}$, then one can obtain a substantial reduction in global risks even for small N.

Also, a comparison is made of various proposed EB d.f.'s for small N. In some cases, it is difficult to determine analytically the average risk for small N and either numerical integration or Monte Carlo simulation is used to study the small sample properties of these decision functions.

1.4 Problems Considered

The two types of prior information discussed above, i.e., knowledge of $\mathcal{G}$ and past observations on X, can be thought of loosely as "partial" or "incomplete" prior information. The terms partial and incomplete indicate that we implicitly assume $\mathcal{G}$ to contain more than one element and that the past observations do not allow us to specify with certainty the "correct" G in $\mathcal{G}$. The following specifications of the class $\mathcal{G}$ will be considered both without prior observations (in which case $\mathcal{G}$-minimax d.f.'s are obtained) and with prior observations (in which case empirical Bayes decision functions are obtained):

(i) $\mathcal{G}$ is the conjugate class of priors. That is, $g(\lambda) \propto k(x|\lambda)$ where $k(x|\lambda)$ is the kernel of the likelihood of x given λ and $g(\lambda)$ is the prior density corresponding to G.

(ii) $\int_\Lambda \lambda dG(\lambda) = \mu$ is known.

(iii) $\int_\Lambda (\lambda-\mu)^2 dG(\lambda) = \mu_2$ is known.

(iv) Some combination of the above cases.

The binomial and normal processes will be of primary concern although the general procedure outlined here can be easily applied to other distributions. The loss function to be considered here is $L(\delta(\mathbf{x}),\lambda) = \omega(\lambda)(\delta(\mathbf{x})-\lambda)^2$ where $\omega(\lambda) > 0 \quad \forall\lambda\varepsilon\Lambda$ corresponding to estimation of the parameter λ (see section 2.2 for greater detail).

In addition, the problem of finding a Bayesian confidence interval for the parameter λ_{N+1} is considered. It is shown that prior information can be used to find shorter confidence intervals (in most cases) than the ordinary non-Bayesian confidence interval.

One particular case of this type is considered in Deely and Zimmer [12]. The problem here simply involves finding an interval $I_\beta(x_{N+1})$ of λ depending on x_{N+1} and β such that $\int_{I_\beta(x_{N+1})} g(\lambda|x_{N+1})d\lambda = \beta$ where $0 \leq \beta \leq 1$ and $g(\lambda|x_{N+1})$ is the posterior distribution of $\tilde{\lambda}$ having observed x_{N+1}. $I_\beta(x_{N+1})$ is then called a $100\beta\%$ Bayesian confidence interval for λ (given $\mathbf{x}$). If g is not completely specified but the types of partial prior information concerning G discussed above are available then the methods of this paper can be used to obtain a confidence interval that closely approximates a Bayes confidence interval. In the empirical Bayes case, we find "asymptotically optimal" confidence intervals, that is, confidence intervals whose end points converge to the Bayes confidence interval end points.

CHAPTER II

PRELIMINARY RESULTS, CONCEPTS, AND STRUCTURAL ELEMENTS

2.1 General Structural Elements

The structural elements of the general Bayesian statistical decision problem are used throughout this paper in the following notation:

(i) Λ: a nonempty parameter space. Λ represents the possible "states of nature" or "states of the world" and some $\lambda_0 \varepsilon \Lambda$ is the parameter value in effect at the time of experimentation.

(ii) A: a nonempty "action" or "decision" space. Each $a \varepsilon A$ represents an action or decision available to the statistician which may be chosen as the result of some experiment.

(iii) L: $A \times \Lambda \to R^+$; a non-negative loss function. $L(a,\lambda)$ represents the loss incurred in taking action a when the true state of nature is λ.

In passing, it should be noted that the triplet (Λ, A, L) is the usual definition of a mathematical game. Nature chooses a $\lambda \varepsilon \Lambda$, the statistician chooses an $a \varepsilon A$ and thereby incurs a loss $L(a,\lambda)$.

(iv) X: an observable random variable (or vector) associated with some experiment. X takes values in a sample space $\mathcal{X}$ on which a σ-finite measure μ is defined. When λ_0 is the true parameter value then X has the probability distribution function $F(x|\lambda_0)$ and has p.d.f. $f(x|\lambda_0)$ with respect to the measure μ.

(v) δ: $\mathcal{X} \to A$: a non-randomized decision function. $\delta(x)$ represents the action taken after observing a particular outcome x of the random variable X. The space of all d.f.'s of this type is denoted by D.

Often, it is necessary to consider randomized d.f.'s, with randomization taken either over the space D before experimentation or over the action space A after experimentation (the so-called behavioral decision functions) but in the Bayesian decision problem, as shown below, it is only necessary to consider the space D.

(vi) G: a "prior" probability distribution function of $\tilde{\lambda}$ on Λ .

G may be specified explicitly or only partially.

Of course, the assumption of (vi) is what distinguishes the Bayesian formulation from the non-Bayesian formulation of the problem. Throughout this paper, assumption (vi) is assumed to hold.

The problem is to choose a good (in some sense) decision function δ such that when x is observed action $\delta(x)$ is taken with a consequent loss of $L(\delta(x),\lambda)$. In order to find good decision functions criteria must be established to induce an ordering on the space D. The following notation and terminology for the usual criteria are used throughout this paper:

(i) The <u>risk</u> or <u>average loss</u> in using a d.f.δ when the true parameter value is λ is:

$$R(\delta,\lambda) = E\left[L(\delta(X),\lambda)\right] = \int_{\mathcal{X}} L(\delta(x),\lambda) f(x|\lambda)\, d\mu(x) \tag{2.1.1}$$

In this paper, the problems considered are such that $R(\delta,\lambda) < \infty$ $\forall\ \delta\varepsilon D,\ \lambda\varepsilon\Lambda$

(ii) The <u>Bayes</u> or <u>average risk</u> in using a d.f.δ when G is the prior distribution of λ is:

$$R(\delta,G) = \int_{\Lambda} R(\delta,\lambda)\, dG(\lambda) \tag{2.1.2}$$

$$= \int_{\Lambda}\int_{\mathcal{X}} L(\delta(x),\lambda) f(x|\lambda)\, d\mu(x)\, dG(\lambda)$$

(iii) the <u>Bayes envelope function</u> is

$$R(G) = \inf_{\delta} R(\delta,G) \tag{2.1.3}$$

2.2 Bayes Decision Functions

The desired object in the decision problem outlined above is to find a d.f.δ that has Bayes risk $R(\delta,G)$ as small as possible. For a given G, a d.f.δ_1 is <u>preferable</u> to δ_2 iff $R(\delta_1,G) < R(\delta_2,G)$ and δ_1 is <u>equivalent</u> to δ_2 iff $R(\delta_1,G) = R(\delta_2,G)$. Recall that the ordinary definition of <u>risk equivalence</u> states that δ_1 and δ_2 are risk equivalent iff $R(\delta_1,\lambda) = R(\delta_2,\lambda)\ \forall\ \lambda\varepsilon\Lambda$. Hence risk equivalence implies equivalence but the converse is not true.

In the problems considered in this paper there always exists a <u>best</u> d.f.δ_G; i.e., there exists a δ_G such that

$$R(\delta_G,G) = R(G) = \inf_{\delta} R(\delta,G) \qquad (2.2.1)$$

Such a d.f.δ_G is called a Bayes decision function. Since it is not true in general that δ_G always exists it is sometimes necessary to consider a δ^* such that

$$R(\delta^*,G) \leq \inf_{\delta\varepsilon D} R(\delta,G) + \varepsilon \qquad \text{for} \quad \varepsilon>0. \quad (2.2.2)$$

Such a d.f. is called <u>ε-Bayes</u> and if δ^* is ε-Bayes for all $\varepsilon>0$ then

δ^* is called an extended Bayes decision function.

An important implication of the existence of a Bayes (or ε-Bayes) d.f. with respect to G is the well-known fact (see Ferguson [13], p. 43) that there always exists a non-randomized Bayes d.f. so we may restrict our attention to D.

As described earlier, the process of obtaining an observation x consists of first choosing a λ' by the distribution $G(\lambda)$ followed by an x from $F(x|\lambda')$. On the other hand, the joint distribution of $\tilde{\lambda}$ and X can be determined by first choosing x from

$$F_G(x) = \int_\Lambda F(x|\lambda)dG(\lambda) \tag{2.2.3}$$

and then choosing a λ from $G(\lambda|x)$, the conditional or posterior distribution of $\tilde{\lambda}$ given X=x. Hence, the Bayes risk may be written (assuming the validity of the operation of interchanging integrals):

$$R(\delta,G) = \int_{\mathcal{X}} \int_\Lambda L(\delta(x),\lambda)dG(\lambda|x)dF_G(x) \tag{2.2.4}$$

The Bayes decision function δ_G may then be found by minimizing

$$\int_\Lambda L(\delta(x),\lambda)dG(\lambda|x) \tag{2.2.5}$$

for each x.

We can write (2.2.5) as:

$$\int_\Lambda L(\delta(x),\lambda)\ g(\lambda|x)d\lambda = \int_\Lambda L(\delta(x),\lambda)\ \frac{f(x|\lambda)g(\lambda)d\lambda}{f_G(x)} \tag{2.2.6}$$

$$= \frac{\int_\Lambda L(\delta(x),\lambda)f(x|\lambda)g(\lambda)d\lambda}{\int_\Lambda f(x|\lambda)g(\lambda)d\lambda}$$

The following theorem (see Ferguson [13], p. 60) is useful in later chapters:

<u>Theorem 1</u>

If the Bayes d.f.δ^* with respect to G^* is unique (up to risk equivalence) then δ^* is admissible. (i.e., there does not exist a δ such that $R(\delta,\lambda) \le R(\delta^*,\lambda)$, $\forall\lambda\varepsilon\Lambda$, and $R(\delta,\lambda_0) < R(\delta^*,\lambda_0)$ for some $\lambda_0\varepsilon\Lambda$).

The loss function used for the estimation problems considered in this paper is the following:

$$L(\delta(x),\lambda) = \omega(\lambda)\ (\delta(x) - \lambda)^2 \tag{2.2.7}$$

where $\omega(\lambda)>0$, $\forall\lambda\varepsilon\Lambda$ and Λ is some subset of the real line. In this case, it is desired to find a decision function δ that minimizes

$$R(\delta,G) = \int_{\Lambda}\int_{X}\omega(\lambda)\ (\delta(x)-\lambda)^2\ f(x|\lambda)g(\lambda)dxd\lambda \tag{2.2.8}$$

$$= \int_{X}\int_{\Lambda}\ \omega(\lambda)\ (\delta(x)-\lambda)^2 f(x|\lambda)g(\lambda)d\lambda dx$$

Equivalently, we can minimize (for each x):

$$\int_{\Lambda}\omega(\lambda)\ (\delta(x)-\lambda)^2 f(x|\lambda)g(\lambda)d\lambda \tag{2.2.9}$$

or, dividing by $f_G(x)$, by minimizing

$$\int_{\Lambda}\omega(\lambda)\ (\delta(x)-\lambda)^2 g(\lambda|x)d\lambda \tag{2.2.10}$$

It is easily seen that (2.2.10) is minimized by that $\delta_G(x)$ such that

$$\delta_G(x) = \left\{\int_{\Lambda}\lambda\omega(\lambda)g(\lambda|x)d\lambda\right\}\ \left\{\int_{\Lambda}\omega(\lambda)g(\lambda|x)d\lambda\right\}^{-1} \tag{2.2.11}$$

In particular, if $\omega(\lambda) = C > 0$, $\forall\lambda\varepsilon\Lambda$, then

$$\delta_G(x) = \int_\Lambda \lambda g(\lambda|x)d\lambda = E_{\lambda|x}(\lambda) \tag{2.2.12}$$

That is, when $\omega(\lambda)$ is a positive constant the Bayes d.f. is just the mean of the posterior distribution of $\tilde{\lambda}$ given x. The loss function (2.2.7) is the loss function used most commonly in problems referred to as point estimation problems and a decision function obtained using this loss function is called an "estimator" (of λ).

2.3 Empirical Bayes Decision Functions

A Bayes decision function δ_G of the type described in the preceding section can only be found if G is completely specified. Hence the minimum possible Bayes risk $R(\delta_G,G)$ can only be attained if G is completely specified. Suppose, however, that G is not completely specified but that there exists a sequence of independent "prior" or "past" observations $\underline{x}_1,\ldots,\underline{x}_N$, each $\underline{x}_i$ itself consisting of M_i independent observations $\left\{x_{ij}\right\}_{j=1}^{M_i}$ chosen randomly from $f(x|\lambda_i)$. In this situation some type of empirical Bayes procedure may be applicable.

In many cases (in all cases considered in this paper), there exists a sufficient statistic $t(\underline{x}_i)$ of fixed dimensionality s for λ; that is, by the Neyman factorization theorem we can write

$$f(\underline{x}_i|\lambda) = K(t(\underline{x}_i,\lambda)p(\underline{x}_i) \tag{2.3.1}$$

where K is a function of $t(\underline{x}_i)$ and λ_1, $p(\underline{x}_i)$ is independent of λ, the dimensionality s of t does not depend on $M(M \geq s)$, and

$$t(\underline{x}_i) = t(x_{i1},\ldots,x_{iM}) = (t_1,\ldots,t_s). \tag{2.3.2}$$

Hence all pertinent past (prior) information is contained in the independent observations $t_1,\ldots,t_N$ where each t_i is from the conditional distribution of T given $\tilde{\lambda} = \lambda_i$. For convenience, we continue to use $f(t|\lambda_i)$ as a functional notation for the conditional density function.

Note that $\left\{t_i\right\}_{i=1}^{N}$ may be considered a random sample of size N from the distribution with p.d.f.

$$f_G(t) = \int_{\Lambda} f(t|\lambda)dG(\lambda) \tag{2.3.3}$$

Hence there exist N pairs of past realizations $\left\{(t_i,\lambda_i)\right\}_{i=1}^{N}$ of the random variables $\left\{(T_i,\tilde{\lambda}_i)\right\}$ in addition to the "current" or "present" realization (t_{N+1}, λ_{N+1}) where the T_i's are observable and the λ_i's unobservable. The unknown λ_{N+1} is the parameter of interest. An empirical Bayes d.f. is of the form $\delta_N\ (t_{N+1};\ t_1,\ldots,t_N)$ and the Bayes risk of δ_N for a <u>given</u> set of past observations is just

$$R_N\ (\delta_N(\ \cdot\ ;\ t_1,\ldots,t_N),\ G) \tag{2.3.4}$$

which is a random variable since it is a function of $t_1,\ldots,t_N$.

The <u>global</u> risk $R(\delta_N,G)$ of δ_N is defined by

$$R(\delta_N, G) = E_N \left\{ R_N(\delta_N, G) \right\} \tag{2.3.5}$$

where E_N represents expectation with respect to the N random variables $T_1, \ldots, T_N$. Note that for all N,

$$R(\delta_N, G) \geq R(\delta_G, G) \tag{2.3.6}$$

since δ_G is the Bayes decision function. However, it is hoped to find a sequence $\{\delta_N\}$ such that

$$\lim_{N\to\infty} R(\delta_N, G) = R(\delta_G, G) \quad \forall\ G \varepsilon \mathcal{G} \tag{2.3.7}$$

i.e., an asymptotically optimal EB d.f. This is not always possible. For example, if $M = 1$ and $P(T = 0|\lambda) = 1-\lambda$ and $P(T = 1|\lambda) = \lambda$ then

$$f_G(t) = \int_0^1 \lambda^t (1-\lambda)^{1-t} dG(\lambda) \qquad t = 0, 1 \tag{2.3.8}$$

i.e.,

$$\begin{cases} f_G(0) = 1-\mu \\ f_G(1) = \mu. \end{cases} \tag{2.3.9}$$

But

$$R(\delta,G) = \int_0^1 R(\delta,\lambda)dG(\lambda) \qquad (2.3.10)$$

and for squared error loss

$$R(\delta,G) = \int_0^1 \left\{ (1-\lambda)(\lambda-\delta(0))^2 + \lambda\left(\lambda-\delta(1)\right)^2 \right\} dG(\lambda) \qquad (2.3.11)$$

$$= \delta^2(0) + \mu[\delta^2(1) - 2\delta(0) - \delta^2(0)] + \mu_2 [1-2\delta(1) + 2\delta(0)]$$

$$= \frac{(\mu-\mu_2')(\mu_2'-\mu^2)}{\mu(1-\mu)} + (1-\mu)\left[\delta(0) - \frac{\mu-\mu_2'}{1-\mu}\right]^2 + \mu\left[\delta(1) - \frac{\mu_2'}{\mu}\right]^2$$

Then, for any $G \in \mathcal{G}$,

$$\delta_G(t) = \frac{t(\mu_2'-\mu^2) + \mu(\mu-\mu_2')}{\mu(1-\mu)} \qquad (2.3.12)$$

and

$$R(\delta_G,G) = \frac{(\mu-\mu_2')(\mu_2'-\mu^2)}{\mu(1-\mu)} . \qquad (2.3.13)$$

Now, any EB procedure depends on $f_G(t)$ for information concerning G (or any function of G). Hence, as pointed out by Robbins [32], since $f_G(t)$ depends only on μ and both $\delta_G(t)$ and $R(\delta_G,G)$ depend on μ and μ_2', then, in general, no asymptotically optimal EB decision function exists in this case.

A common computation in working with empirical Bayes decision functions when using a squared error loss function is the following (recall that $\delta_G(t) = E_{\lambda|t}(t)$)

$$\begin{aligned} R(\delta_N,G) &= E_N \left\{ R_N\ (\delta_N,\ G) \right\} && (2.3.14) \\ &= E_N \left\{ E_{\lambda,t}\ (\delta_N(t) - \lambda)^2 \right\} \\ &= E_N \left\{ E_t\ E_{\lambda|t}(\delta_N(t) - \delta_G(t) + \delta_G(t) - \lambda\ )^2 \right\} \\ &= E_N \left\{ E_t(\delta_N(t) - \delta_G(t))^2 + E_t\ E_{\lambda|t}(\delta_G(t) - \lambda)^2 \right. \\ &\qquad \left. + 2E_t\ E_{\lambda|t}\ (\delta_N(t) - \delta_G(t))(\delta_G(t) - \lambda) \right\} \\ &= E_{N+1}\ (\delta_N(t) - \delta_G(t)\ )^2 + R(\delta_G,G) \end{aligned}$$

where E_{N+1} represents expectation with respect to the N+1 independent random variables $T_1,\ldots,T_{N+1}$ each with p.d.f. $f_G(t)$, and $E_{\lambda,t}$, $E_{\lambda|t}$, and E_t represent expectation with respect to (respectively) the joint distribution of $\tilde{\lambda}$ and T, the conditional distribution of $\tilde{\lambda}$ given T = t and the unconditional distribution of T. Often, when no confusion arises, the symbol "E" (without subscripts) is used for expectations with respect to all random variables involved. Also, as in the above, the subscripts on the random variables T_i (and on the t_i's) are often omitted when it is clear which variables are being considered.

It will be seen below that the structure of $\delta_N(t)$ often makes evaluation of $R(\delta_N,G)$ difficult. In these cases it will be necessary to use numerical integration or Monte Carol studies to approximate the value $R(\delta_N,G)$.

Of particular interest is the case in which G is assumed specified except for some unknown parameters. This case is of interest because it is precisely this assumption which is made in many Bayesian-type analyses. However, the usual procedure involved in determining the unknown parameters is of a subjective nature. A typical method of this type is to question the "assessor" in order to determine his prior beliefs (cf. Winkler [49]) with no direct reference to prior observations, if any. Hence a method of directly estimating the parameters of G from past data, assuming reasonable small sample accuracy, is of more than academic interest.

In general, if the form of G is specified except for a finite number of unknown parameters $\underline{\alpha}$ then it should be possible to devise estimators of $\underline{\alpha}$ from the fact that $t_1,\ldots,t_{N+1}$ may be considered a random sample of size N+1 from $f_G(t)$. This could be done, for example, by the well-known method of maximum likelihood or perhaps by the method of moments. In this paper, when the form of G is known, the empirical Bayes procedures proposed can be shown to be essentially equivalent to the method of moments. These procedures are used, even though the method of moments may not have the best asymptotic properties, because they are much simpler than other procedures (such as maximum liklihood) and appear accurate enough for our purposes.

In particular, the decision functions based on this method are shown to be asymptotically optimal while possessing good small (previous) sample properties as well.

2.4 $\mathcal{G}$-Minimax Decision Functions

Assume now that G is unknown and no past observations from $F_G(x)$ are available but the class $\mathcal{G}$ of which G is a member is specified. In this case, it is reasonable (see Robbins [32]) to prefer a d.f.δ_1 to d.f.δ_2 if

$$\sup_{G\varepsilon\mathcal{G}} R(\delta_1,G) \leq \sup_{G\varepsilon\mathcal{G}} R(\delta_2,G). \tag{2.4.1}$$

Hence it would be desirable to find a d.f. δ^* ∋

$$\sup_{G\varepsilon\mathcal{G}} R(\delta^*, G) \leq \sup_{G\varepsilon\mathcal{G}} R(\delta, G) \tag{2.4.2}$$

for all $\delta\varepsilon D$. Such a d.f.δ^* is called a <u>$\mathcal{G}$-minimax</u> (or "minimax in $\mathcal{G}$ ") decision function by Blum and Rosenblatt [3]. In particular, if $\mathcal{G}$ contains only one element G then the $\mathcal{G}$-minimax decision function is identical to the Bayes d.f. The definition may be equivalently stated as: δ^* is $\mathcal{G}$-minimax if

$$\sup_{G\varepsilon\mathcal{G}} R(\delta^*, G) = \inf_{\delta} \sup_{G} R(\delta,G). \tag{2.4.3}$$

We also define the <u>$\mathcal{G}$-minimax value</u> $\overline{V}(\mathcal{G})$ by writing

$$\overline{V}(\mathcal{G}) = \inf_{\delta} \sup_{G} R(\delta,G). \tag{2.4.4}$$

This definition is analagous to the usual definition of the minimax value $\overline{V}$ since

$$\overline{V} = \inf_{\delta} \sup_{\lambda} R(\delta,\lambda) \tag{2.4.5}$$

In fact, if $\mathcal{G}$ contains all degenerate distributions on Λ then $\overline{V} = \overline{V}(\mathcal{G})$.

It is useful later to use the fact that a d.f.δ^* is $\mathcal{G}$-minimax iff

$$R(\delta^*,G') \leq \sup_{G\varepsilon\mathcal{G}} R(\delta,G) \tag{2.4.6}$$

for all $\delta\varepsilon D$ and all $G'\varepsilon\,\mathcal{G}$. Recall that a d.f.$\delta_0$ is called <u>minimax</u> if

$$\sup_{\lambda\varepsilon\Lambda} R(\delta_0,\lambda) \leq \sup_{\lambda\varepsilon\Lambda} R(\delta,\lambda) \tag{2.4.7}$$

or equivalently, if for all $\delta\varepsilon D$,

$$\sup_{\lambda\varepsilon\Lambda} R(\delta_0,\lambda) = \inf_{\delta} \sup_{\lambda\varepsilon\Lambda} R(\delta,\lambda). \tag{2.4.8}$$

Now, for any $\delta \varepsilon D$ we have

$$R(\delta,G) = \int_{\Lambda} R(\delta,\lambda)\, dG(\lambda) \leq \int_{\Lambda} \sup_{\lambda} R(\delta,\lambda) dG(\lambda) \tag{2.4.9}$$

$$= \sup_{\lambda} R(\delta,\lambda)$$

for all $G \varepsilon \mathcal{G}$. Hence,

$$\sup_{G \varepsilon \mathcal{G}} R(\delta,G) \leq \sup_{\lambda \varepsilon \Lambda} R(\delta,\lambda). \tag{2.4.10}$$

and, of course, if $\mathcal{G}$ contains all degenerate distributions on Λ then

$$\sup_{G \varepsilon \mathcal{G}} R(\delta,G) = \sup_{\lambda \varepsilon \Lambda} R(\delta,\lambda) \tag{2.4.11}$$

so that in this case a d.f.δ^* is $\mathcal{G}$-minimax iff it is minimax.

The concept of $\mathcal{G}$-admissibility is defined analagously to ordinary admissibility (see section 2.2).

Definition

A d.f.δ^* is <u>$\mathcal{G}$-admissible</u> if there does not exist δ such that

$$R(\delta,G) \leq R(\delta^*,G) \text{ for all } G \varepsilon \mathcal{G} \tag{2.4.12}$$

and

$$R(\delta,G_0) < R(\delta^*,G_0) \text{ for some } G_0 \varepsilon \mathcal{G} \quad . \tag{2.4.13}$$

Obviously, admissibility implies $\mathcal{G}$-admissibility for any class $\mathcal{G}$. Also, if $\mathcal{G}$ contains all degenerate distributions on Λ then $\mathcal{G}$-admissibility implies admissibility. In fact, all that is really needed for $\mathcal{G}$-admissibility to imply admissibility is that $\forall \lambda \varepsilon \Lambda$ there exists $G_\lambda \; \varepsilon \; \mathcal{G}$ such that $G_\lambda(\lambda) - G_\lambda(\lambda^-) > 0$.

To see this, assume δ^* is $\mathcal{G}$-admissible but not admissible. Then there exists $\delta_0 \; \varepsilon \; D$ and $\lambda_0 \; \varepsilon \; \Lambda$ such that

$$R(\delta_0,\lambda_0) < R(\delta^*,\lambda_0) \tag{2.4.14}$$

and

$$R(\delta_0,\lambda) \leq R(\delta^*,\lambda) \quad \forall \; \lambda \varepsilon \Lambda. \tag{2.4.15}$$

Then there exists $G_{\lambda_0} \; \varepsilon \; \mathcal{G}$ such that

$$R(\delta_0,G_{\lambda_0}) < R(\delta^*,G_{\lambda_0}) \tag{2.4.16}$$

and by (2.4.15),

$$R(\delta_0,G) \leq R(\delta^*,G) \quad \forall G \varepsilon \mathcal{G} \tag{2.4.17}$$

a contradiction to the assumption of the $\mathcal{G}$-admissibility of δ^*. Hence $\mathcal{G}$-admissibility implies admissibility in this case.

As a corollary to theorem 1, we add: if the Bayes d.f. δ^* with respect to G is unique (up to equivalence) then δ^* is $\mathcal{G}$-admissible.

In game theory terminology, a $\mathcal{G}$-minimax strategy for nature is called a <u>least favorable distribution</u>.

<u>Definition</u>

G^* is least favorable in $\mathcal{G}$ if

$$\inf_{\delta} R(\delta,G^*) \geq \inf_{\delta} R(\delta,G) \tag{2.4.18}$$

for all $G \varepsilon \mathcal{G}$ or, equivalently, if

$$\inf_{\delta} R(\delta,G^*) = \sup_{G\varepsilon\mathcal{G}} \inf_{\delta} R(\delta,G) \tag{2.4.19}$$

or, equivalently, if

$$R(G^*) \geq R(G). \tag{2.4.20}$$

Expression (2.4.20) can be expressed in words as: "G* is least favorable in $\mathcal{G}$ if the Bayes envelope function attains its supremum over $\mathcal{G}$ at G*".

The $\mathcal{G}$ -maximin value $\underline{V}$ ($\mathcal{G}$) is defined by:

$$\underline{V}(\mathcal{G}) = \sup_{G} \inf_{\delta} R(\delta,G). \tag{2.4.21}$$

Again, if $\mathcal{G}$ is the class of all distributions on Λ then $\underline{V}(\mathcal{G}) = \underline{V}$, the ordinary maximin value.

The following definition will also be needed:

<u>Definition</u>

A d.f.δ_1 is said to be <u>$\mathcal{G}$ -equivalent</u> to δ_2 if

$$\sup_{G\varepsilon\mathcal{G}} R(\delta_1,G) = \sup_{G\varepsilon\mathcal{G}} R(\delta_2,G). \tag{2.4.22}$$

This definition is identical to G equivalence if the class consists of a single member G.

In view of the similarities of the definitions of $\mathcal{G}$-minimax concepts and ordinary minimax concepts it is not surprising that many well-known results on minimax decision functions carry over with little modification to $\mathcal{G}$-minimax decision functions. In particular, the following $\mathcal{G}$-minimax analogues to certain minimax results are easy to prove and are useful in the later sections:

<u>Theorem 2</u>

If δ^* is a unique (up to equivalence) $\mathcal{G}$-minimax decision function then δ^* is $\mathcal{G}$-admissible.

<u>Proof:</u>

Suppose δ^* is <u>not</u> $\mathcal{G}$-admissible. Then there exists δ_0 such that

$$R(\delta_0,G) \leq R(\delta^*,G) \quad \forall\, G\varepsilon\mathcal{G}$$

$$R(\delta_0,G_0) < R(\delta^*,G_0) \text{ for some } G_0\varepsilon\mathcal{G}$$

then

$$\sup_G R(\delta_0,G) \leq \sup_G R(\delta^*,G). \tag{2.4.23}$$

That is, δ_0 is also $\mathcal{G}$-minimax. But since δ_0 is not equivalent to δ^* this contradicts the uniqueness of δ^*.

q.e.d.

The following theorems (3, 4 and 5) are exact $\mathcal{G}$-minimax analogues of minimax theorems 1, 2 and 3 found in Ferguson [13], pages 90-91, and are invaluable in finding $\mathcal{G}$-minimax decision functions in practice.

<u>Theorem 3</u>

If δ_0 is Bayes with respect to $G_0\varepsilon\mathcal{G}$ and for all $G\varepsilon\mathcal{G}$

$$R(\delta_0,G) \leq R(\delta_0,G_0) \tag{2.4.24}$$

then

$$\overline{V}(\mathcal{G}) = \underline{V}(\mathcal{G}), \tag{2.4.25}$$

δ_0 is $\mathcal{G}$ -minimax and G_0 is least favorable.

<u>Proof</u>

Obviously $\underline{V}(\mathcal{G}) \leq \overline{V}(\mathcal{G})$. Also,

$$\begin{aligned} \overline{V}(\mathcal{G}) &= \inf_{\delta} \sup_{G} R(\delta,G) \\ &\leq \sup_{G} R(\delta_0,G) \\ &\leq R(\delta_0,G_0) \\ &= \inf_{\delta} R(\delta,G_0) \\ &\leq \sup_{G} \inf_{\delta} R(\delta,G) \\ &= \underline{V}(\mathcal{G}). \end{aligned} \tag{2.4.26}$$

Hence, $\overline{V}(\mathcal{G}) = \underline{V}(\mathcal{G})$. Next, from (2.4.24),

$$\begin{aligned} \sup_{G} R(\delta_0,G) &\leq \sup_{G} \inf_{\delta} R(\delta,G) \\ &\leq \sup_{G} R(\delta,G) \quad \forall \delta \varepsilon D \end{aligned} \tag{2.4.27}$$

$$\Rightarrow \delta_0 \text{ is } \mathcal{G} \text{ - minimax}$$

Finally, also from (2.4.24),

$$\begin{aligned} \inf_{\delta} R(\delta,G_0) &\geq \inf_{\delta} \sup_{G} R(\delta,G) \\ &\geq \inf_{\delta} R(\delta,G) \quad \forall G \varepsilon \mathcal{G} \end{aligned} \tag{2.4.28}$$

$\Rightarrow G_0$ is least favorable.

q.e.d.

Theorem 4

If δ_N is Bayes with respect to $G_N \in \mathcal{G}$ and

$$\lim_{N \to \infty} R(\delta_N, G_N) = C \tag{2.4.29}$$

and there exists δ_0 such that

$$R(\delta_0, G) \le C \quad \forall G \in \mathcal{G} \tag{2.4.30}$$

then

$$\overline{V}(\mathcal{G}) = \underline{V}(\mathcal{G})$$

and δ_0 is $\mathcal{G}$-minimax.

Proof $\forall \delta$,

$$\sup_{G \in \mathcal{G}} R(\delta, G) \ge R(\delta, G_N) \ge R(\delta_N, G_N) \quad \forall N \tag{2.4.31}$$

$$\Rightarrow \sup_{G \in \mathcal{G}} R(\delta, G) \ge C \ge R(\delta_0, G) \quad \forall G$$

$$\Rightarrow \sup_{G} R(\delta, G) \ge R(\delta_0, G) \quad \forall G$$

$$\Rightarrow \sup_{G} R(\delta, G) \ge \sup_{G} R(\delta_0, G)$$

therefore,

δ_0 is $\mathcal{G}$-minimax.

q.e.d.

Theorem 5

If δ^* is Bayes (or extended Bayes) with respect to $G^* \in \mathcal{G}$ and

$$R(\delta^*, G) = K \quad \forall\, G \varepsilon \mathcal{G} \tag{2.4.32}$$

where K is some constant then δ^* is $\mathcal{G}$-minimax and G^* is least favorable in $\mathcal{G}$.

Proof:

$$R(\delta^*, G) = K = R(\delta^*, G^*) \quad \forall\, G \varepsilon \mathcal{G}$$

By theorem 3, δ^* is $\mathcal{G}$ -minimax and G^* least favorable.

q.e.d.

The above theorems provide the two principal methods used to find $\mathcal{G}$ -minimax decision functions in this paper. Namely:

(i) Using theorem 3: Find a supposed least favorable distribution G_0, obtain the Bayes d.f. δ_0 for G_0 and check to see if $R(\delta_0, G_0) \geq R(\delta_0, G) \quad \forall\, G \varepsilon \mathcal{G}$.

(ii) Using theorem 5: Find a d.f. δ_0 such that $R(\delta_0, G) = K$ for all $G \varepsilon \mathcal{G}$ and check to see if δ_0 is Bayes (or extended Bayes) for some $G_0 \varepsilon \mathcal{G}$.

If N prior observations on $F_G(t)$ exist but an a.o. EB d.f. δ_N as defined in section 2.3 does not exist or an a.o. EB δ_N exists but $R(\delta_N, G)$ converges too slowly to $R(\delta_G G)$ then it may be desirable (at least for small N) to find a decision function $\delta_N(t_{N+1}; t_1, \dots, t_N)$ with the property:

$$\lim_{N \to \infty} R(\delta_N, G) = R(\delta^*, G) \tag{2.4.33}$$

where δ^* is a $\mathcal{G}$ -minimax decision function. Robbins [31] calls such an estimator <u>asymptotically subminimax</u> but to be consistent with the present terminology a decision function with property (2.4.33) is called asymptotically $\mathcal{G}$-minimax or asymptotically minimax in $\mathcal{G}$ in this paper.

CHAPTER III

NORMAL PROCESS I: EMPIRICAL BAYES DECISION FUNCTIONS

3.1 Introduction

The independent normal process as used in this paper is defined as a generator of independent real-valued variables $X_1, X_2, \ldots$, each with p.d.f.

$$f(x|\lambda,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\lambda)^2}{2\sigma^2}} \tag{3.1.1}$$

where $-\infty < \lambda < \infty$ and $\sigma^2 > 0$. A random sample of size M is drawn from the distribution with p.d.f. (3.1.1) and it is desired to make some inference concerning the unknown parameter value or values in effect at the time of drawing the sample. The inference is always to be made with the aid of the prior information at hand - whether it be past observations on $f_G(x)$ or some knowledge concerning the class $\mathcal{G}$ of which the prior distribution G is a member (as will be discussed in Chapter 4).

The current observations are denoted by $\left\{x_{N+1,j}\right\}_{j=1}^{M}$ and the past observations by $\left\{x_{1,j}\right\}_{j=1}^{M}, \ldots, \left\{x_{N,j}\right\}_{j=1}^{M}$ where and the following notation for the first few moments of G is used:

$$\begin{cases} \int_{\Lambda} \lambda \, dG(\lambda) = \mu \\ \int_{\Lambda} \lambda^2 \, dG(\lambda) = \mu_2' \\ \mu_2 = \mu_2' - \mu^2 \end{cases} \tag{3.1.7}$$

The conjugate prior density in this case is:

$$g(\lambda | m', n') = \frac{\sqrt{n'}}{\sqrt{2\pi\sigma^2}} \, e^{\frac{-n'}{2\sigma^2}(\lambda - m')^2} \tag{3.1.8}$$

that is,

$$\tilde{\lambda} \sim N\left(m', \frac{\sigma^2}{n'}\right) \tag{3.1.9}$$

and

$$\begin{cases} \mu = m' \\ \mu_2 = \dfrac{\sigma^2}{n'} \\ \mu_2' = \dfrac{\sigma^2}{n'} + (m')^2 \end{cases} \tag{3.1.10}$$

When the prior G is conjugate, the compound density (3.1.6) becomes

$$f_G(t) = \frac{1}{\sqrt{2\pi\sigma^2\left(\frac{1}{M} + \frac{1}{n'}\right)}} \, e^{-\frac{(t-\mu)^2}{2\sigma^2\left(\frac{1}{M} + \frac{1}{n'}\right)}} \tag{3.1.11}$$

that is,

$$T \sim N\left(\mu,\ \sigma^2\left[\frac{1}{M} + \frac{1}{n'}\right]\right) . \tag{3.1.12}$$

Also, the posterior density of $\tilde{\lambda}$ given t_{N+1} becomes:

$$g(\lambda \mid t_{N+1}) = \frac{\sqrt{n''}}{\sqrt{2\pi\sigma^2}}\ e^{-\frac{n''(\lambda - m'')^2}{\alpha\sigma^2}} \tag{3.1.13}$$

where $n'' = n' + M$

$$m'' = \frac{1}{n''}\ (n'\mu + Mt_{N+1}) = \frac{\sigma^2\mu + M\mu_2 t_{N+1}}{\sigma^2 + M\mu_2}$$

that is,

$$\tilde{\lambda} \mid t_{N+1} \sim N\left(m'',\ \frac{\sigma^2}{n''}\right) . \tag{3.1.14}$$

3.2 Estimation Of λ_{N+1}

The loss function to be used when estimating λ_{N+1} is:

$$L(\delta(t),\lambda) = (\delta(t) - \lambda)^2 . \tag{3.2.1}$$

Let G^* denote the class of all probability distribution functions on Λ. Now the usual (ML, UMVU) estimator of λ_{N+1} is simply t_{N+1} and is denoted here by:

$$\delta_T(t_{N+1}) = t_{N+1} \tag{3.2.2}$$

The Bayes risk of δ_T is:

$$R(\delta_T,G) = \int_{\Lambda}\int_{-\infty}^{\infty}(t-\lambda)^2 f(t|\lambda)\,dt\,dG(\lambda) \tag{3.2.3}$$

$$= \int_{\Lambda}\mathrm{Var}_{T|\lambda}(T)\,dG(\lambda)$$

$$= \frac{\sigma^2}{M} \qquad \forall G\varepsilon \mathcal{G}^*$$

i.e., δ_T has a constant risk for all $G\varepsilon\mathcal{G}^*$.

Now, if G is in the class of conjugate priors we note that as $n' \to 0$ we have increasingly vague prior information (see (3.1.9)). In fact, we can consider the limiting distribution ($n' = 0$) to be a "uniform" distribution in the sense of Lindley [21]. That is, if F is any set of λ's of finite measure then the distribution (conditional on $\lambda\varepsilon F$) of $\tilde{\lambda}$ has density:

$$g(\lambda|F) = [\mu_L(F)]^{-1} \tag{3.2.4}$$

where $\mu_L(F)$ is the Lebesque measure of F. Hence,

$$\int_F g(\lambda|F)\,d\lambda = [\mu_L(F)]^{-1}\int_F d\lambda = 1. \tag{3.2.5}$$

When we speak of the class of conjugate priors we will include this limiting distribution as a member.

Now the posterior distribution when $n' = 0$ is just a $N\left(t_{N+1}, \frac{\sigma^2}{M}\right)$ distribution (see (3.1.13)) and since the Bayes d.f. for the loss function (3.2.1) is the mean of the posterior, then $\delta_T(t_{N+1})$ is Bayes with respect to this uniform prior and $R(\delta_T, G) = \sigma^2/M$, a constant for all $G \varepsilon \mathcal{G}^*$. Hence, by theorem 5, the density (3.2.4) is least favorable in $\mathcal{G}^*$. Hence, we could restrict our attention to the class of normal priors with the knowledge that the least favorable distribution in $\mathcal{G}^*$ is still under consideration. Also, in many practical cases the conditions of theorem 1, pg. 13-14 of Lindley [21], are satisfied and hence the posterior distribution is approximately normal. Thus the conjugate prior family not only satisfies the usual requirements of richness and tractability but is also a conservative choice for the prior family since it certainly contains the "worst possible" G. Because of the above reasons, for the remainder of this chapter we assume that $\mathcal{G}$ is the class of all conjugate priors, namely all normal distributions on Λ. For squared error loss, the Bayes d.f. δ_G for any $G \varepsilon \mathcal{G}$ is the mean of the posterior distribution of $\tilde{\lambda}$ given t_{N+1} or:

$$\begin{aligned}
\delta_G(t_{N+1}) &= m'' \qquad\qquad (3.2.6)\\
&= \frac{n'\mu + Mt_{N+1}}{n' + M}\\
&= \frac{\sigma^2\mu + M\mu_2 t_{N+1}}{\sigma^2 + M\mu_2}\\
&= \frac{M\mu_2}{\sigma^2 + M\mu_2}\, t_{N+1} + \frac{\sigma^2}{\sigma^2 + M\mu_2}\, \mu \,.
\end{aligned}$$

That is, δ_G is the weighted average of the prior mean μ and the sample mean t_{N+1} with weights proportional to the inverses of their respective variances.

The minimum Bayes risk in this case is:

$$
\begin{aligned}
R(\delta_G, G) &= \int_\Lambda E_{t|\lambda}(\delta_G(t)-\lambda)^2 dG(\lambda) \\
&= \int_{-\infty}^{\infty}\int_\Lambda (\delta_G(t)-\lambda)^2 g(\lambda|t) f_G(t)\,d\lambda\,dt \qquad (3.2.7)\\
&= \int_{-\infty}^{\infty} \frac{\sigma^2}{M+n'} f_G(t)\,dt \\
&= \frac{\sigma^2}{M+n'} \quad .
\end{aligned}
$$

The ratio of $R(\delta_G,G)$ to $R(\delta_T,G)$ is:

$$\frac{R(\delta_G,G)}{R(\delta_T,G)} = \frac{M}{M+n'} \leq 1. \qquad (3.2.8)$$

The following facts are useful in devising an EB d.f.:

$$\hat{\mu}^* = \sum_{i=1}^{N} t_i /N \sim N\left(\mu, \frac{\sigma^2+M\mu_2}{MN}\right) \sim N\left(\mu, \frac{\sigma^2}{N}\left[\frac{1}{M}+\frac{1}{n'}\right]\right) \qquad (3.2.9)$$

since $\left\{t_i\right\}_{i=1}^{N}$ is a random sample from the distribution with p.d.f. (3.1.11). And

$$\hat{\mu} = \sum_{i=1}^{N+1} t_i/(N+1) \sim N\left(\mu, \frac{\sigma^2+M\mu_2}{M(N+1)}\right) \sim N\left(\mu, \frac{\sigma^2}{N+1}\left[\frac{1}{M}+\frac{1}{n'}\right]\right) \qquad (3.2.10)$$

Define:

$$s_t^2 = \sum_{i=1}^{N+1} (t_i - \hat{\mu})^2/N \ . \tag{3.2.11}$$

Now,

$$t_i - \hat{\mu} = t_i - \sum_{j=1}^{N+1} t_j/(N+1) \tag{3.2.12.}$$

$$= \left\{ Nt_i - \sum_{\substack{j=1 \\ j \neq i}}^{N+1} t_j \right\} /(N+1)$$

and since the t_i are all independently distributed $N(\mu, \frac{\sigma^2}{M} + \mu_2)$,

$$t_i - \hat{\mu} \sim N\left(0, \frac{N}{N+1} \left[\frac{\sigma^2}{M} + \mu_2\right]\right) \tag{3.2.13}$$

and hence

$$E(s_t^2) = \frac{\sigma^2}{M} + \mu_2. \tag{3.2.14}$$

Therefore we have shown that

$$E\left(s_t^2 - \frac{\sigma^2}{M}\right) = \mu_2 \tag{3.2.15}$$

and by the strong law of large numbers, as $N \to \infty$:

$$\begin{aligned} \hat{\mu} &\xrightarrow{a.s.} \mu \\ s_t^2 - \frac{\sigma^2}{M} &\xrightarrow{a.s.} \mu_2 \ . \end{aligned} \tag{3.2.16}$$

The general approach used here and in the following sections is to obtain an explicit estimator G_N of the unknown G where we require G_N to be a member of $\mathcal{G}$. We show that the d.f. δ_N, the Bayes d.f. with prior G_N, is: (i) asymptotically optimal (ii) easy to apply and (iii) good for small N. In particular it is desirable to have $R(\delta_N, G) \leq R(\delta_T, G)$ for reasonably small N. If δ_N is a.o. then this last property is obviously true for large N.

As our estimator G_N of G we take a normal distribution with mean = $\hat{\mu}$ and variance = $\hat{\mu}_2$ where

$$\hat{\mu}_2 = \max\left\{0,\ s_t^2 - \sigma^2/n\right\}. \tag{3.2.17}$$

As pointed out in section 2.3, the estimators $\hat{\mu}$ and $\hat{\mu}_2$ are essentially identical to the moment estimators obtained by noting

$$\begin{aligned} E\ (t_G) &= \mu \\ \mathrm{Var}\ (t_G) &= \sigma^2\left(\frac{1}{M} + \frac{1}{N'}\right) = \frac{\sigma^2}{M} + \mu_2 \end{aligned} \tag{3.2.18}$$

and setting

$$\begin{aligned} \hat{\mu} &= \bar{t} \\ \frac{\sigma^2}{M} + \hat{\mu}_2 &= s_t^2 \end{aligned} \tag{3.2.19}$$

i.e.,

$$\begin{aligned} \hat{\mu} &= \bar{t} \\ \hat{\mu}_2 &= s_t^2 - \frac{\sigma^2}{M} \end{aligned} \tag{3.2.20}$$

Note that from (3.2.17) we are estimating μ_2 to be 0 if $s_t^2 \leq \frac{\sigma^2}{M}$. That is, G_N is a degenerate distribution at $\hat{\mu}$ in this case and our empirical Bayes d.f. is just

$$\delta_N(t_{N+1}) = \hat{\mu} = \sum_{i=1}^{N+1} t_i/(N+1). \tag{3.2.21}$$

But since

$$s_t^2 - \frac{\sigma^2}{n} \xrightarrow{a.s.} \mu_2 \geq 0 \tag{3.2.22}$$

then for large N we have with probability one

$$s_t^2 \geq \frac{\sigma^2}{M} . \tag{3.2.23}$$

In any case, our empirical Bayes estimator is:

$$\delta_N(t_{N+1}) = \frac{\sigma^2\hat{\mu} + M\hat{\mu}_2 t_{N+1}}{\sigma^2 + M\hat{\mu}_2} \tag{3.2.24}$$

and, from (2.3.14), we know that the global risk is:

$$R(\delta_N,G) = R(\delta_G,G) + E_{N+1}\left\{\delta_N(t_{N+}) - \delta_G(t_{N+1})\right\}^2 . \tag{3.2.25}$$

Now, by modifying the theorem and its proof given in section 20.6, Cramér [9], for the case of convergence with probability one we obtain the following theorem:

<u>Theorem 6</u>

If the random variables $\xi_N, \ldots, \psi_N$ converge with probability one to the constants x, y, ..., r respectively then any rational function $R(\xi_N, \ldots, \psi_N)$ converges with probability one provided the latter is finite.

Now, from (3.2.16) and theorem 6 we have that

$$\left(\frac{\sigma^2\hat{\mu} + M\hat{\mu}_2 t_{N+1}}{\sigma^2 + M\hat{\mu}_2} - \frac{\sigma^2\mu + M\mu_2 t_{N+1}}{\sigma^2 + M\mu_2}\right) \xrightarrow{a.s.} 0 \tag{3.2.26}$$

i.e.,

$$\left(\delta_N(t_{N+1}) - \delta_G(t_{N+1})\right) \xrightarrow{a.s.} 0. \tag{3.2.27}$$

Defining

$$\alpha_1 = \frac{\sigma^2}{\sigma^2+M\mu_2} \tag{3.2.28}$$

$$\left\{\begin{aligned} \alpha_2 &= \frac{M\mu_2}{\sigma^2+M\mu_2} \\ \alpha_3 &= \frac{\sigma^2 N}{(N+1)(\sigma^2+M\hat{\mu}_2)} \end{aligned}\right.$$

$$\alpha_4 = \frac{M\hat{\mu}_2+\sigma^2/(N+1)}{M\hat{\mu}_2+\sigma^2}$$

we see that

$$\left\{\begin{array}{ll} 0 \le \alpha_i \le 1 & i=1, \ldots, 4 \\ \alpha_1+\alpha_2 = 1 & \\ \alpha_3+\alpha_4 = 1 & \end{array}\right. \tag{3.2.29}$$

and

$$\delta_G(t_{N+1}) = \alpha_1\mu+\alpha_2 t_{N+1} \tag{3.2.30}$$

$$\delta_N(t_{N+1}) = \alpha_3\hat{\mu}^* + \alpha_4 t_{N+1} \tag{3.2.31}$$

where $\hat{\mu}^* = \sum_{i=1}^{N} t_i/N$. Also, defining

$$\begin{cases} \beta_1 = \alpha_3 \\ \beta_2 = -\alpha_1 \\ \beta_3 = \alpha_4 - \alpha_2 \end{cases} \tag{3.2.32}$$

we see that

$$\delta_N(t_{N+1}) - \delta_G(t_{N+1}) = \beta_1\mu^* + \beta_2\mu + \beta_3 t_{N+1} \tag{3.2.33}$$

where $0 \leq \beta_1 \leq 1$ and $\sum_{i=1}^{3}\beta_i = 0$ so that

$-1 \leq \beta_2 \leq 0$

$-1 \leq \beta_3 \leq 1$

$$\left\{\delta_N(t) - \delta_G(t)\right\}^2 \leq (\hat{\mu}^*)^2 + \mu^2 + t^2 + |\mu^*\mu| + |\mu t| + |\hat{\mu}^* t| \ . \tag{3.2.34}$$

Hence,

$$E\left[\delta_N(t) - \delta_G(t)\right]^2 < \infty \ . \tag{3.2.35}$$

Then, (3.2.27) and (3.2.35) imply

$$\lim_{N \to \infty} E\left[\delta_N(t) - \delta_G(t)\right]^2 = 0 \tag{3.2.36}$$

i.e., δ_N is asymptotically optimal in $\mathcal{G}$.

In section 3.3, $R(\delta_N, G)$ is evaluated for small N and for various $G \varepsilon \mathcal{G}$ and a comparison is made with the global risks for other estimators.

Suppose next that additional information concerning $\mathcal{G}$ is known. In particular, suppose that

$$\mu_2 = \int_\Lambda (\lambda-\mu)^2 dG(\lambda) \tag{3.2.37}$$

is known and denote the class of all conjugate priors with variance $= \mu_2$ by G'. In this case, our estimator G_N of G will be a $N(\hat{\mu},\mu_2)$ distribution and our EB d.f. will be

$$\delta_N'(t_{N+1}) = \frac{\sigma^2\hat{\mu}+M\mu_2 t_{N+1}}{\sigma^2+M\mu_2} = \frac{n'\hat{\mu}+Mt_{N+1}}{n'+M}. \tag{3.2.38}$$

The global risk of (3.2.38) is (by (2.3.14))

$$R(\delta_N', G) = E_N R_N(\delta_N, G) \tag{3.2.39}$$

$$= R(\delta_G, G)+E_{N+1}\left\{\frac{n'\mu+Mt_{N+1}}{n'+M} - \frac{n'\hat{\mu}+Mt_{N+1}}{n'+M}\right\}^2$$

$$= R(\delta_G, G)+\left(\frac{n'}{n'+M}\right)^2 E_{N+1}(\mu-\hat{\mu})^2$$

and since $\hat{\mu} \sim N\left(\mu, \frac{\sigma^2}{N+1}\left[\frac{1}{M}+\frac{1}{n'}\right]\right)$ by (3.2.10) then

$$R(\delta_N', G) = R(\delta_G,G)+\left(\frac{n'}{n'+M}\right)^2 \frac{\sigma^2}{N+1}\left(\frac{1}{M}+\frac{1}{n'}\right) \tag{3.2.40}$$

$$= R(\delta_G, G)+\frac{n'\sigma^2}{(N+1)(n'+M)M}$$

$$= \frac{\sigma^2}{M+n'}\left[1+\frac{n'}{M(N+1)}\right].$$

Hence, obviously

$$\lim_{N\to\infty} R(\delta_N', G) = R(\delta_G, G). \tag{3.2.41}$$

That is, δ_N' is asymptotically optimal in $\mathcal{G}'$.

An important characteristic of δ_N' is deduced from the following:

$$\frac{R(\delta_N', G)}{R(\delta_T, G)} = \frac{M}{n'+M}\left[1 + \frac{n'}{M(N+1)}\right] \tag{3.2.42}$$

$$= \frac{M(N+1) + n'}{(n'+M)(N+1)}$$

$$= \frac{M(N+1) + n'}{M(N+1) + n'(N+1)}$$

$$\leq 1 .$$

That is, for <u>any</u> $N \geq 1$, the empirical Bayes estimator δ_N' is better than the usual non-Bayes estimator δ_T. In other words, knowledge of the prior variance enables us to improve on the estimator δ_T regardless of how large that variance may be. Of course, the fact that

$$\lim_{N \to \infty} \frac{R(\delta_N', G)}{R\delta_T, G)} = \frac{M}{M+n'} = \frac{R(\delta_G, G)}{R(\delta_T, G)} \tag{3.2.43}$$

can be deduced directly from (3.2.42) or simply from the knowledge that δ_N' is asymptotically optimal. As a particular example of the possible improvement when μ_2 is known, suppose $M = 3$, $\mu_2 = \sigma^2/9$. In this case,

$$\frac{R(\delta_N', G)}{R(\delta_T, G)} = \frac{N+4}{4N+4} \tag{3.2.44}$$

so that even when N=1 a substantial reduction in the global risk is obtained. See section 3.3 for more complete results.

In addition to computing the expectation of the random variable $\left\{\delta_N'(t)-\lambda\right\}^2$ it is useful to know the exact distribution of $\delta_N'(t)-\lambda$. We can write:

$$\delta_N'(t_{N+1})-\lambda = \left(\frac{n'}{n'+M}\right)\hat{\mu} + \left(\frac{M}{n'+M}\right)t_{N+1} - \lambda \tag{3.2.45}$$

$$= (\alpha_1,\alpha_2,\alpha_3)\begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix} = \underline{\alpha}'\underline{z}$$

where

$$\alpha_1 = \frac{n'}{n'+M} \qquad z_1 = \hat{\mu} \tag{3.2.46}$$

$$\alpha_2 = \frac{M}{n'+M} \qquad z_2 = t_{N+1}$$

$$\alpha_3 = -1 \qquad z_3 = \lambda \quad .$$

We also have:

$$\underline{z} \sim N(\mu_{\underline{z}}, \textstyle\sum_{\underline{z}}) \tag{3.2.47}$$

and hence

$$\underline{\alpha}'\underline{z} \sim N(\underline{\alpha}'\mu_{\underline{z}} \quad \underline{\alpha}'\textstyle\sum_{\underline{z}}\underline{\alpha}) \tag{3.2.48}$$

where

$$\underline{\mu}'_{\underline{z}} = (\mu, \mu, \mu) \tag{3.2.49}$$

$$\Sigma_{\underline{z}} = \begin{pmatrix} \sigma^2_{\hat{\mu}} & \sigma_{\hat{\mu}, t_{N+1}} & \sigma_{\hat{\mu},\lambda} \\ \sigma_{\hat{\mu}, t_{N+1}} & \sigma^2_{t_{N+1}} & \sigma_{t_{N+1},\lambda} \\ \sigma_{\hat{\mu},\lambda} & \sigma_{t_{N+1},\lambda} & \sigma^2_{\lambda} \end{pmatrix}$$

and the elements of $\Sigma_{\underline{z}}$ are simply

$$\sigma^2_{\hat{\mu}} = \frac{\sigma^2}{N+1}\left(\frac{1}{M} + \frac{1}{n'}\right) = \frac{\sigma^2 (M+n')}{(N+1)Mn'} \tag{3.2.50}$$

$$\sigma^2_{t_{N+1}} = \sigma^2\left(\frac{1}{M} + \frac{1}{n'}\right) = \frac{\sigma^2 (M+n')}{Mn'}$$

$$\sigma^2_{\lambda} = \mu_2 = \frac{\sigma^2}{n'}$$

$$\sigma_{\hat{\mu},\lambda} = \sigma_{t,\lambda} = \mu_2 = \frac{\sigma^2}{n'}$$

$$\sigma_{\hat{\mu},t_{N+1}} = \left(\frac{1}{N+1}\right)\sigma^2_{t_{N+1}} = \frac{\sigma^2}{N+1}\;\frac{M+n'}{Mn'}$$

obviously,

$$\underline{\alpha}'\underline{\mu}_{\underline{z}} = 0 \tag{3.2.51}$$

and after some straight - forward manipulation it is easy to see that

$$\underline{\alpha}'\Sigma_{\underline{z}}\,\underline{\alpha} = \sigma^2 \frac{M(N+1) + n'}{M(N+1)(M+n')} \qquad (3.2.52)$$

which is, of course, identical with (3.2.40). That is, we have shown that

$$\delta_N'(t)-\lambda \sim N\left(0,\ \sigma^2\left[\frac{M(N+1) + n'}{M(N+1)(M+n')}\right]\right). \qquad (3.2.53)$$

also, the random variable $\delta_N'(t) - \lambda$ converges in distribution to a $N\left(0,\ \frac{\sigma^2}{n'+m}\right)$ which is the distribution of $\delta_G(t)-\lambda$. This should be expected of course, since $\delta_N'(t)$ is asymptotically optimal.

In a similar fashion we can easily see that

$$\delta_T(t)-\lambda = t-\lambda \sim N\left(0,\ \frac{\sigma^2}{M}\right) \qquad (3.2.54)$$

so that

$$\frac{\delta_T(t)-\lambda}{\sigma/\sqrt{M}} \sim N(0,1)\ . \qquad (3.2.55)$$

The usual non-Bayesian $1-\alpha$ level confidence interval on the unknown λ_{N+1} is simply:

$$\delta_T(t) \pm z_{1-\alpha/2}\ \frac{\sigma}{\sqrt{M}} \qquad (3.2.56)$$

which is obviously longer than the usual Bayesian $(1-\alpha)$ 100 % confidence interval on λ_{N+1}, namely:

$$\delta_T(t) \pm z_{1-\alpha/2} \frac{\sigma}{\sqrt{M+n'}} \tag{3.2.57}$$

Now, from (3.2.53) we have

$$\frac{\delta_N'(t) - \lambda}{\sqrt{\sigma^2 \frac{M(N=1)+n'}{M(N+1)(M+n')}}} \sim N(0,1) \tag{3.2.58}$$

and hence an empirical Bayes confidence interval can be formed using (3.2.58). The result is:

$$\delta_N'(t) \pm z_{1-\alpha/2} \sqrt{\frac{\sigma^2}{(M+n')} \left[1+ \frac{n'}{M(N+1)}\right]} \quad . \tag{3.2.59}$$

The confidence interval represented by (3.2.59) is shorter than the usual confidence interval (3.2.56) and also, as $N \to \infty$, it converges to the Bayes confidence interval (3.2.57). The confidence interval (3.2.59) was also found by Deely and Zimmer [12] using a different approach.

3.3 Numerical Results

In this section the results of several Monte Carlo type simulations relating to the decision functions described in the previous section are given.

Since any asymptotically optimal empirical Bayes decision function δ_N is nearly as good as the Bayes d.f.δ_G for large N, the primary concern here is the case where N is small. (e.g., N=1, 10, 20). Although asymptotically optimality is a desirable property for any decision function δ_N depending on past observations to possess, it is a minimal requirement in that it implies nothing about the behavior of δ_N for small N.

If however, we could find a d.f. δ_N^* such that

$$R(\delta_N^*, G) \leq R(\delta_N, G) \quad \forall N \text{ and } \forall G \varepsilon \mathcal{G} \tag{3.3.1}$$

where δ_N is any other a.o. EB decision function, then one would feel justified in calling δ_N^* the "best" EB d.f. available. However, we are unlikely to be able to find such a δ_N^* except in very restrictive cases.

In practice, one may be willing to use δ_N^* if it can be shown that

$$R(\delta_N^*, G) \leq R(\delta_T, G) \quad \forall N \geq N_0 \text{ and } \forall G \varepsilon \mathcal{G} \tag{3.3.2}$$

where, as usual, δ_T is some optimal non-Bayes decision function. In any case, if a d.f.δ_N^* satisfies (3.3.2) for reasonably small N_0 then δ_N^* is at least worthy of further consideration.

The main purpose in this section is to closely approximate the global risk $R(\delta_N,G)$ for each proposed decision function given in section 3.2 where an exact determination is difficult or impossible by analytical means. Then these decision functions are compared (by means of $R(\delta_N, G)$ to each other and to the Bayes decision function δ_G).

The method used is first to generate a sample $t_1, \ldots, t_N$ of size N, each T_i having the p.d.f. $f_G(t)$, then to generate an observation λ_{N+1} from the distribution with distribution function $G(\lambda)$ and finally to generate a t_{N+1} from the distribution with p.d.f. $f(t|\lambda_{N+1})$. Equivalently, we could have generated a random sample of size N+1 from the distribution with distribution function $G(\lambda)$, say $\left\{ \lambda_i \right\}_{i=1}^{N+1}$ and then generate an observation t_i from the distribution with p.d.f. $f(t|\lambda_i)$ for $i=1, 2, \ldots, N+1$. The decision function value $\delta_N(t_{N+1})$ resulting from this process is then compared with the actual λ_{N+1} (unknown in practice) by means of the squared error loss function:

$$L(\delta_N(t),\lambda) = (\delta_N(t)-\lambda)^2 \tag{3.3.3}$$

and the entire process is then repeated. In this section, all runs involve 100 replications unless otherwise stated.

The actual calculations were performed on an IBM 360 model 44. The programs used and more details can be found in the appendix.

In addition to the decision functions considered in section 3.2, one other decision function is studied here since it exhibited an explicit method for construction of a sequence of distributions G_N which estimate the unknown G where $\mathcal{G}$ is assumed only to be the class of all continuous distribution functions on Λ. This procedure is described by Deely and Kruse [11] and is outlined briefly below.

Define $\mathcal{G}_N$ to be the class of discrete distributions on Λ with weights at $\lambda_N, \ldots, \lambda_{NN}$. We want to find a $G_N^* \varepsilon \mathcal{G}_N$ which minimizes

$$||F_H - F_N|| = \sup_x |F_H(x) - F_N(x)| \quad \text{for } H \varepsilon\, \mathcal{G}_N \tag{3.3.4}$$

where $F_N(x)$ is the usual empirical distribution function. Then under certain fairly general conditions, Deely and Kruse show:

$$P\left\{\lim_{N \to \infty} G_N^*(\lambda) = G(\lambda);\ \lambda \text{ any continuity point of } G\right\} = 1 \ . \tag{3.3.5}$$

The actual mehtod of construction is to denote by $H(\lambda)$ the distribution in $\mathcal{G}_N$ with weights $h_1, \ldots, h_N$ at $\lambda_{1N}, \ldots, \lambda_{NN}$. Then,

$$||F_H - F_N|| = \max_{1 \leq i \leq 2N} \left|\sum_j a_{ij} h_j\right| \tag{3.3.6}$$

where

$$a_{ij} = \begin{cases} F(t_i|\lambda_{jN}) - \dfrac{(i-1)}{N}, & 1 \leq i \leq N \\ F(t_{i-N}|\lambda_{jN}) - \dfrac{(i-N)}{N}, & N+1 \leq i \leq 2N \ . \end{cases} \tag{3.3.7}$$

Finding $G_N^*(\lambda)$ then can be shown to be equivalent to finding an optimal strategy for the second player in a game with payoff matrix

$$\begin{pmatrix} (a_{ij})_{2NXN} \\ (-a_{ij})_{2NXN} \end{pmatrix}_{4NXN} \ . \tag{3.3.8}$$

This optimal strategy may be found by linear programming techniques. The computer program used for this purpose is found in the appendix.

For convenience we let $\sigma^2 = 1$ and consider the following two prior distributions:

(i) $\mu = M' = 0$; $\mu_2 = \frac{\sigma^2}{N'} = \frac{1}{N'} = 1$ $(N'=1)$.

That is, the prior variance and the variance of T are equal.

(ii) $\mu = M' = 0$; $\mu_2 = \frac{\sigma^2}{N'} = .10$ $(N'=10)$.

i.e., the prior variance is 1/10 the variance of T.

The estimators considered are:

(i) δ_G; $R(\delta_G, G) = \frac{\sigma^2}{M+N'}$.

The Bayes estimator.

(ii) $\delta_T = t_{N+}$; $R(\delta_T, G) = \frac{\sigma^2}{M}$.

The "usual" estimator

(iii) δ_N; $R(\delta_N, G)$ determined in Monte Carlo study.

Empirical Bayes estimator, G assumed conjugate.

(iv) δ''_N ; $R(\delta''_N, G)$ determined in Monte Carlo study.

Empirical Bayes estimator, μ known.

(v) δ'_N; $R(\delta'_N, G) = \frac{\sigma^2}{N'+M} \quad \frac{M(N+1)+N'}{M(N+1)}$

empirical Bayes estimator, μ_2 known.

(vi) δ^D_N; $R(\delta^D_N, G)$ determined in Monte Carlo study.

Deely and Kruse's method used, yielding this EB estimator. δ^D_N is just δ_G with G replaced by G_N as described earlier.

Although the number of λ's involved (N1) in finding the G_N used in the d.f. δ^D_N is an additional parameter, it was found that reasonably good approximations are obtained even when N1 is very small (and much less than N). In fact, in every case considered here there was very little difference observed between using N1=3 and using N1=N. Hence, for simplification and ease of presentation we set N1=3 for all computations considered here. This simplication also resulted in much shorter computer run times.

The following tables (1 and 2) present the main results on risks in a tabular form for N=1, 10, 20 and M = 1, 5, 10. The numbers in parentheses are standard errors. If absent, then the results are exact. Besides the value $R(\delta,G)$ for each d.f.δ, the tables give the ratio

$$R(\delta) = R(\delta_G, G)/R(\delta,G). \qquad (3.3.9)$$

That is, the ratio of the minimum possible Bayes risk to that obtained by using the d.f.δ. Note that

$$0 \leq R(\delta) \leq 1 \qquad (3.3.10)$$

and obviously the larger $R(\delta)$, the better the decision function δ.

However, it must be borne in mind that the estimators δ_N' and δ_N'' represent decision situations in which one is prepared to assume more than the usual assumptions concerning G and hence cannot fairly be compared with estimators that do not assume this knowledge.

The figures following the tables (figures 1 and 2) show $R(\delta)$ for each estimator as a function of N. The curves representing $R(\delta_N')$ and $R(\delta_N'')$ are not given since they obviously lie above $R(\delta_N)$ and, as explained above, have an unfair advantage over the other estimators given. These figures not only allow an easy comparison among the various estimators for small N but enable one to check visually the rapidity of convergence of $R(\delta_N)$ to 1.

Note that in the first case (table 1 and figure 1) the relatively large prior variance (N'=1) results in a $R(\delta_T)$ value of .833 so there is not much room for improvement in this case. However, even for N'=1 it is seen that $R(\delta_N) > R(\delta_T)$ for all N. Also, for the small N indicated we see that $R(\delta_N^D) < R(\delta_T)$ even though we know that

$$\lim_{N \to \infty} R(\delta_N^D) = 1.$$

In the second case (N'=10) we have $R(\delta_T) = .333$ so there is considerable room for improvement over δ_T when the prior variance is small. Here we also have $R(\delta_N) > R(\delta_T)$ and, in addition, $R(\delta_N^D) > R(\delta_T)$ for the indicated small N (even for N=1).

TABLE 1

GLOBAL RISKS OF VARIOUS ESTIMATORS USING
A N(0,1) PRIOR

Estimator	N=1			N=10			N=20		
	M=1	M=5	M=10	M=1	M=5	M=10	M=1	M=5	M=10
1) Bayes	.500	.167	.091	.500	.167	.091	.500	.167	.091
2) Usual (mean)	1.000	.200	.100	1.000	.200	.100	1.000	.200	.100
3) G conjugate	.983 (.106)	.195 (.047)	.106 (.011)	.741 (.103)	.183 (.029)	.095 (.018)	.661 (.053)	.178 (.011)	.094 (.009)
4) G conjugate, mean known	.885 (.103)	.183 (.034)	.095 (.013)	.618 (.107)	.172 (.082)	.093 (.010)	.505 (.044)	.166 (.015)	.092 (.007)
5) G conjuate, variance known	.750	.183	.095	.545	.170	.092	.524	.168	.091
6) Deely and Kruse's estimator	1.549 (.170)	.304 (.038)	.196 (.027)	1.249 (.164)	.256 (.026)	.185 (.018)	1.106 (.107)	.245 (.020)	.106 (.016)

* Numbers in parentheses are standard errors. If absent, results are exact.

TABLE 2

GLOBAL RISKS OF VARIOUS ESTIMATORS USING
A N(0,.1) PRIOR

Estimator	N=1			N=10			N=20		
	M=1	M=5	M=10	M=1	M=5	M=10	M=1	M=5	M=10
1) Bayes	.091	.067	.050	.091	.067	.050	.091	.067	.050
2) Usual (mean)	1.000	.200	.100	1.000	.200	.100	1.000	.200	.100
3) G conjugate	.647 (.055)	.152 (.018)	.097 (.008)	.198 (.011)	.103 (.009)	.088 (.007)	.155 (.015)	.074 (.008)	.058 (.006)
4) G conjugate, mean known	.538 (.052)	.148 (.010)	.078 (.007)	.175 (.015)	.091 (.021)	.069 (.006)	.148 (.019)	.007 (.008)	.055 (.006)
5) G conjugate, variance known	.545	.133	.075	.174	.079	.055	.134	.073	.052
6) Deely & Kruse's estimator	.772 (.069)	.176 (.012)	.092 (.006)	.377 (.034)	.151 (.015)	.087 (.004)	.293 (.024)	.124 (.017)	.079 (.004

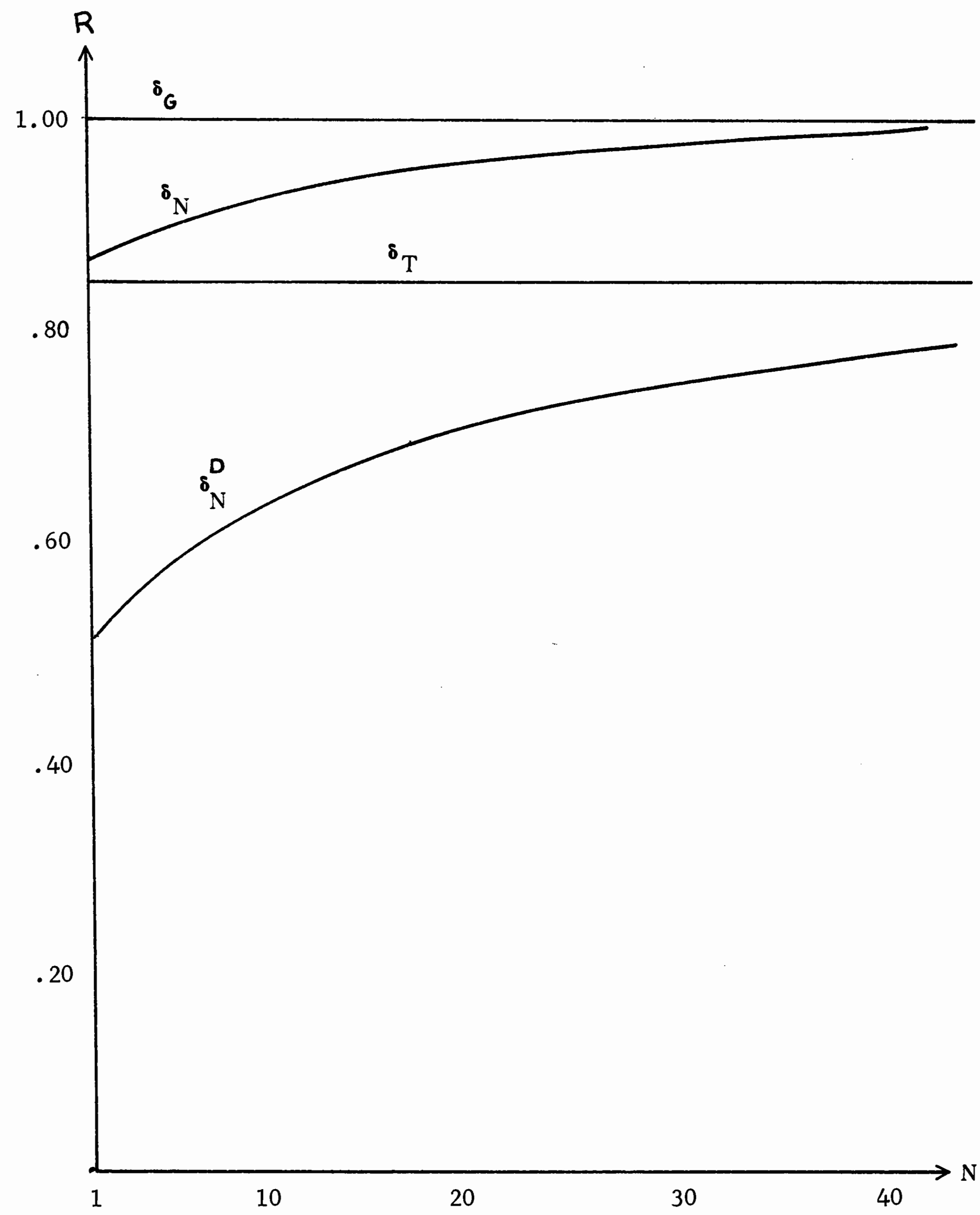

Figure 1. $R(\delta)$ as a function of $N(M'=0, N'=1, M=5)$.

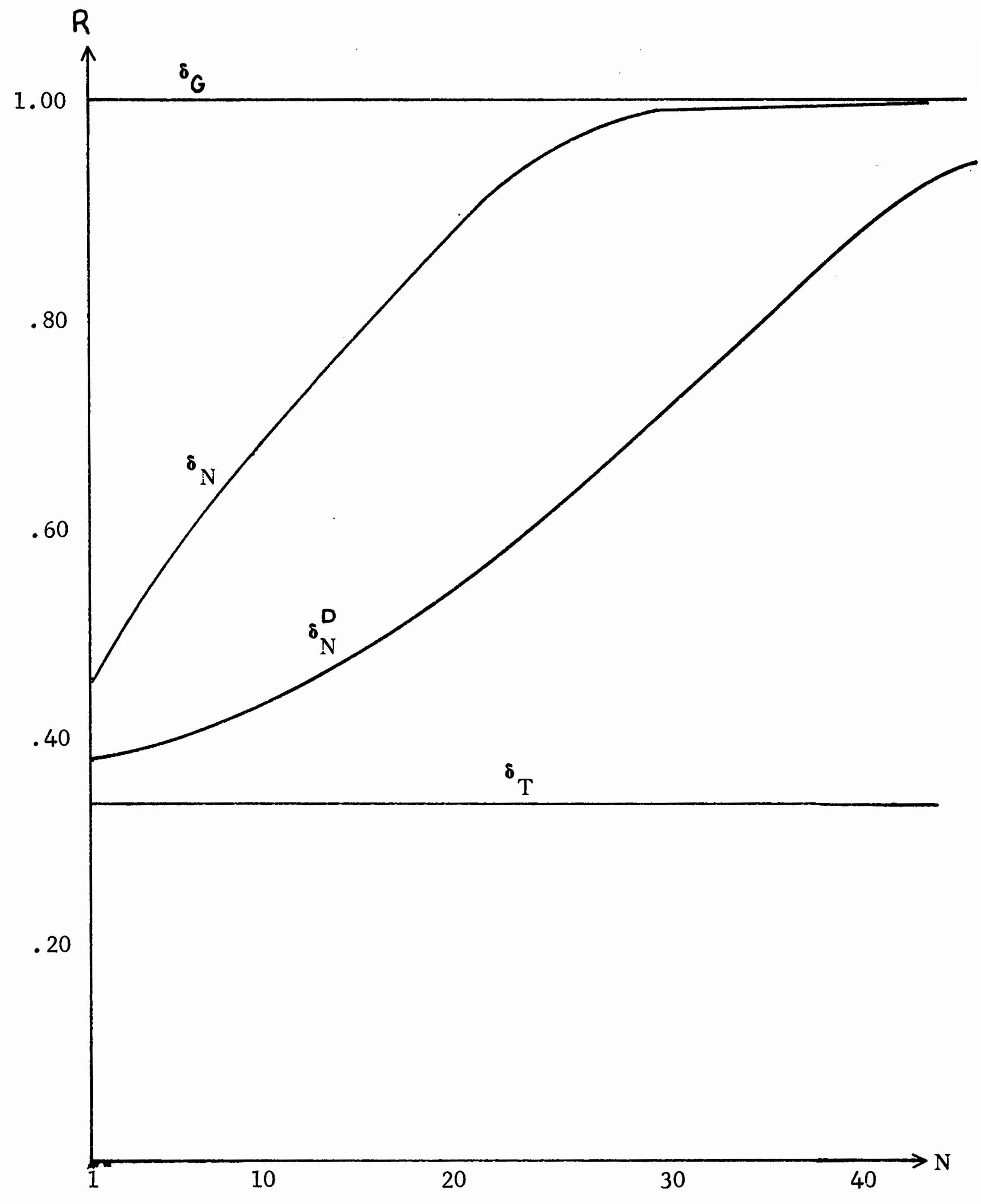

Figure 2. R(δ) as a function of N(M'=0, N'=10, M=5).

CHAPTER IV

NORMAL PROCESS II: $\mathcal{G}$-MINIMAX DECISION FUNCTIONS

4.1 Introduction

Suppose now that instead of N past observations on $f_G(t)$ it is known only that the distribution function G is a member of some class $\mathcal{G}$. The best we can do in this situation, in the spirit of Chapter II, is to find a $\mathcal{G}$-minimax decision function. The different specifications of the class $\mathcal{G}$ to be considered here are:

(i) $\mathcal{G}$: the conjugate class of priors.

(ii) $\mathcal{G}_1$: conjugate class, μ known.

(iii) $\mathcal{G}_1^+$: conjugate class, μ known, $\mu_2 \leq \mu_2^+$.

(iv) $\mathcal{G}_2$: conjugate class, μ_2 known.

(v) $\mathcal{G}_2^+$: conjugate class, μ_2 known, $\mu \varepsilon [\mu_L, \mu_u]$.

(vi) $\mathcal{G}_3$: μ known, $\mu_2 \leq \mu_2^+$.

(vii) $\mathcal{G}_4$: $\mu_2 < \infty$.

It is shown below that cases (i), (ii), and (iv) represent situations in which the $\mathcal{G}$-minimax d.f. is the same in each case. A similar result is true for cases (iii) and (vi). Hence, the above seven catagories yield only four distinct decision functions in this $\mathcal{G}$-minimax context.

In each case, the $\mathcal{G}$-minimax d.f. is obtained, its Bayes risk computed and compared with other decision functions. In section 4.4, asymptotic $\mathcal{G}$-minimax d.f.'s are obtained for the same cases.

The general results of Chapter III, relating to the independent normal process and related topics, will not be repeated here but it will be necessary to refer to them throughout this chapter.

4.2 Estimation Of λ_{N+1}: $\mathcal{G}$ Conjugate

Since the conjugate class $\mathcal{G}$ is the class of all normal distributions on Λ and any degenerate distribution on Λ may be thought of as a normal distribution with zero variance, then the $\mathcal{G}$-minimax estimator is just the ordinary minimax estimator. That is,

$$\delta_{\mathcal{G}}\ (t_{N+1}) = \delta_T(t_{N+1}) = t_{N+1} \quad . \tag{4.2.1}$$

The Bayes risk for $\delta_{\mathcal{G}}$ has already been shown to be:

$$R(\delta_{\mathcal{G}},\ G) = \frac{\sigma^2}{M} \quad . \tag{4.2.2}$$

Next, suppose that the mean μ of G is assumed known and denote the class of all normal distributions with mean μ by $\mathcal{G}_1$. Now, of course, $\mathcal{G}_1$ does not contain all degenerate distributions on Λ.

To find a $\mathcal{G}_1$-minimax estimator we will attempt to find an estimator δ_0 such that $R(\delta_0,G)=k \geq 0$ for all $G\varepsilon\mathcal{G}$, and such that δ_0 is Bayes with respect to some $G\varepsilon\mathcal{G}_1$. Since the Bayes estimator with respect to any $G\varepsilon\mathcal{G}_1$ is a weighted average of μ and t_{N+1} (with weights totalling one) we require δ_0 to be such that:

$$\delta_0(t_{N+1}) = \alpha_0\mu+\beta_0\ t_{N+1} \tag{4.2.3}$$

where

$$\alpha_0,\ \beta_0 \geq 0 \text{ and } \alpha_0+\beta_0=1\ . \tag{4.2.4}$$

The Bayes risk for such a δ_0 is

$$R(\delta_0,G) = E_\lambda E_{t/\lambda}\left\{\delta_0(t)-\lambda\right\}^2 \tag{4.2.5}$$

$$= E_\lambda E_{t/\lambda}\left\{\alpha_0^2\mu^2+\beta_0^2t^2+\lambda^2+2\alpha_0\beta_0\mu t\ 2\alpha_0\mu\lambda-2\beta_0 t\lambda\right\}$$

$$= E_\lambda\left\{\alpha_0^2\mu^2+\beta_0^2\ \ \frac{\sigma^2}{M}+\lambda^2\ \ +\lambda^2+2\alpha_0\beta_0\mu\lambda\, 2\alpha_0\mu\lambda-2\beta_0\lambda^2\right\}$$

$$= \alpha_0^2\mu^2+\frac{\beta_0^2\sigma^2}{M}\ \ +\beta_0^2(\mu_2+\mu^2)+(\mu_2+\mu^2)\ \ +2\alpha_0\beta_0\mu^2\, 2\alpha_0\mu^2$$

$$-2\beta_0(\mu_2+\mu^2)\quad .$$

Hence, for $R(\delta_0,G)$ to be constant for all $G\varepsilon\mathcal{G}_1$ we must have

$$\beta_0^2-2\beta_0+1 = 0 \tag{4.2.6}$$

i.e., $\beta_0 = 1$ and $\alpha_0 = 0$. Hence,

$$\delta_0(t_{N+1}) = \quad 0\cdot\mu + 1\cdot t_{N+1} = t_{N+1} = \delta_G\ (t_{N+1})\,. \tag{4.2.7}$$

That is, when no prior observations exist, knowledge of the mean of the prior does not yield an improved estimator over the case when G is simply known to be a conjugate prior distribution function.

This strange sounding result occurs because the conjugate class as defined contains, as a limiting distribution, a normal distribution with infinite variance. This distribution is our least favorable distribution and knowledge of the mean of a distribution with infinite variance is of no value in the present context.

However, this problem can be easily avoided by simply restricting μ_2 in such a way that:

$$0 \leq \mu_2 \leq \mu_2^+ < \infty \tag{4.2.8}$$

That is, μ_2^+ is some fixed upper bound on the variance μ_2. In practice, (4.2.8) need not be much of a restriction since μ_2^+ may be extremely large.

This additional knowledge about μ_2 would not enable us to do better than δ_T in the case where G is known only to belong to the conjugate family since this class still contains all degenerate distributions.

But if, in addition, the mean is known to be μ_0, say, then one may suspect that the least favorable distribution in this new class, say G_1^+, is a normal distribution with mean μ_0 and variance μ_2^+.

If the prior distribution is a $N(\mu_0, \mu_2^+)$, denoted by G_0, then the Bayes estimator is:

$$\delta_0(t_{N+1}) = \frac{\sigma^2\mu_0 + M\mu_2^+ t_{N+1}}{\sigma^2 + M\mu_2^+} \tag{4.2.9}$$

and the Bayes risk is:

$$R(\delta_0, G_0) = \frac{\sigma^2\mu_2^+}{\mu_2^+ M + \sigma^2} \tag{4.2.10}$$

and for any other $G \varepsilon G_1^+$, we have:

$$R(\delta_0,G) = E_{t/\lambda}(\delta_0(t)-\lambda)^2$$

$$= \frac{\sigma^2 M(\mu_2^+)^2 + \sigma^4\mu_2}{(\sigma^2+M\mu_2^+)^2} \tag{4.2.11}$$

$$\leq \frac{\sigma^2[M(\mu_2^+)^2 + \sigma^2\mu_2^+]}{(\sigma^2+M\mu_2^+)^2}$$

$$= \frac{\sigma^2\mu_2^+}{\sigma^2+M\mu_2^+}$$

$$= R(\delta_0,G_0) \text{ for all } G\varepsilon G_1^+$$

Hence by theorem 3, G_0 is least favorable and δ_0 is G_1^+-minimax.

For convience and consistency we denote δ_0 by

$$\delta_0(t_{N+1}) = \delta_{G_1^+}(t_{N+1}) \quad . \tag{4.2.12}$$

By using $\delta_{G_1^+}$, the worst we can do, relative to δ_T is:

$$\sup_{G\varepsilon G_1^+} \frac{R(\delta_{G_1^+}, G)}{R(\delta_T,G)} = \frac{R(\delta_{G_1^+} G_0)}{R(\delta_T,G)} \tag{4.2.13}$$

$$= \frac{M}{M+\sigma^2/\mu_2^+} < 1.$$

For <u>any</u> $G\varepsilon G_1^+$ we have

$$\frac{R(\delta_{G_1^+},G)}{R(\delta_T,G)} = \frac{\sigma^2 M(\mu_2^+) + \sigma^4\mu_2}{(\sigma^2+M\mu_2^+)^2} \cdot \frac{M}{\sigma^2} \tag{4.2.14}$$

$$= \frac{(M\mu_2^+) \quad + \sigma^2\mu_2 M}{(\sigma^2+M\mu_2^+)} .$$

That is, regardless of which $G\varepsilon\mathcal{G}_1$ is in effect, we find that $\delta_{\mathcal{G}_1^+}$ is preferred to δ_T.

As a particular example, suppose $M=2$, $\sigma^2=4$ and $\mu_2^+ = 10.00$. Then figure 3 gives $R(\delta_{\mathcal{G}_1^+},G)$ and $R(\delta_T,G)$ as a function of μ_2.

Suppose next that μ_2, but not μ, is known. We denote the class of conjugate priors with variance μ_2 by $\mathcal{G}_2$.

Now, a Bayes estimator δ_0 for any $G\varepsilon\mathcal{G}_2$ is

$$\delta_0(t_{N+1}) = \frac{\sigma^2\alpha+M\mu_2 t_{N+1}}{\sigma^2+M\mu_2} \tag{4.2.15}$$

where α is the (unknown) mean of G. If $\alpha = t_{N+1}$ then

$$\delta_0(t_{N+1}) = t_{N+1} = \delta_T(t_{N+1}) \tag{4.2.16}$$

and

$$R(\delta_0,G) = \frac{\sigma^2}{M} \text{ for all } G\varepsilon\mathcal{G}_2. \tag{4.2.17}$$

Since t_{N+1} is Bayes with respect to a $N(t_{N+1},\mu_2)$ distribution then, by theorem 5, δ_T is $\mathcal{G}_2$-minimax.

Another way to see this is to assume $\alpha \neq t_{N+1}$. Then,

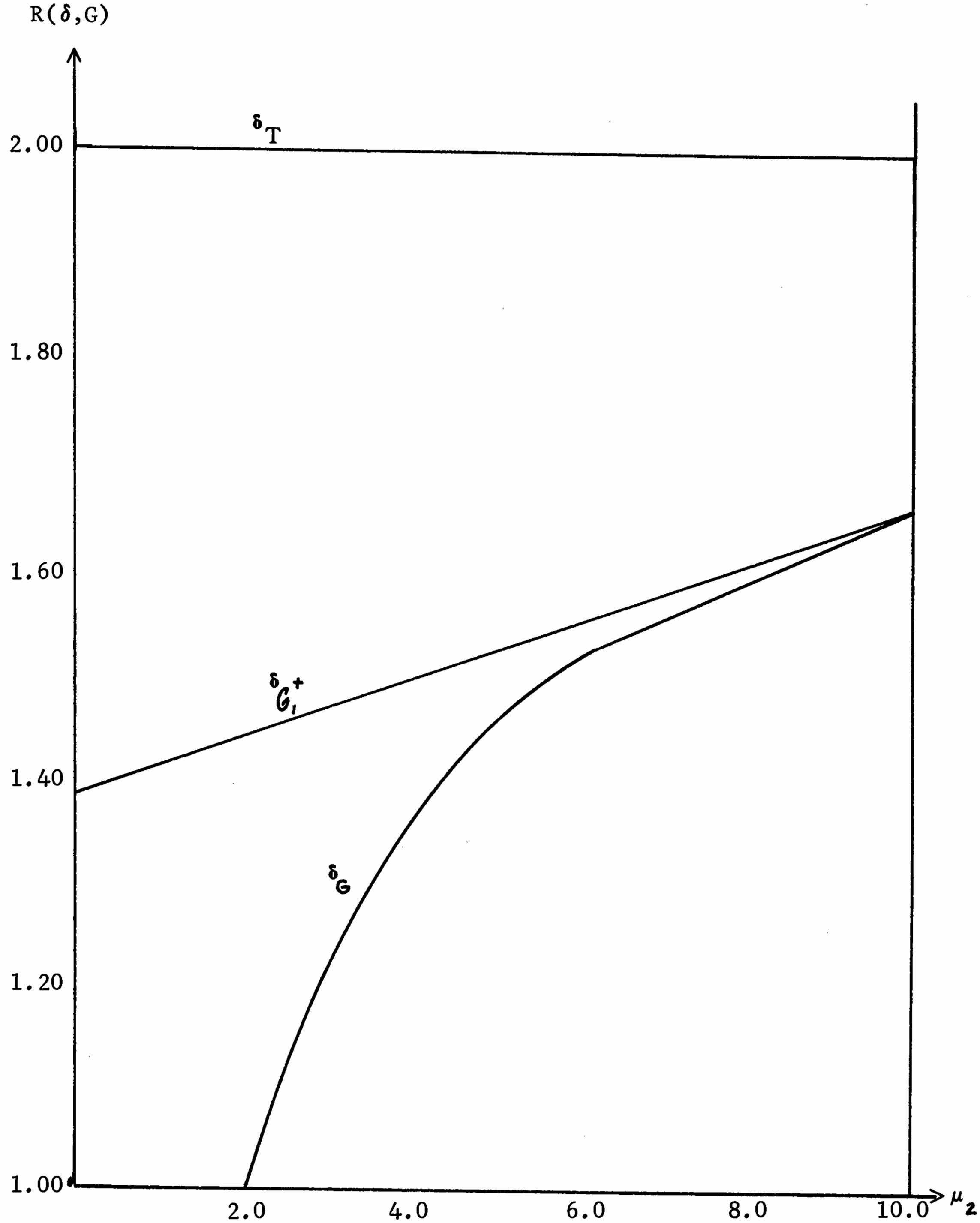

Figure 3. Bayes risk of $\delta_{G_1^+}$ and δ_T as a function of μ_2

$$R(\delta_0,G) = E_{t,\lambda}\left\{\frac{\sigma^2\alpha+M\mu_2 t_{N+1}}{\sigma^2+M\mu_2} - \lambda\right\}^2 \tag{4.2.18}$$

$$= E_\lambda\left\{\left(\frac{M\mu_2}{\sigma^2+M\mu_2}\right)^2 \cdot \frac{\sigma^2}{M} + \frac{\sigma^4(\alpha-\lambda)^2}{(\sigma^2+M\mu_2)^2}\right\}$$

$$= \frac{\sigma^2 M(\mu_2)^2+\sigma^4[\mu_2+(\mu-\alpha)^2]}{(\sigma^2+M\mu_2)^2}$$

and for any $G\varepsilon\mathcal{G}_2$ such that

$$(\mu-\alpha)^2 \geq \frac{\sigma^2}{M} + \mu_2 \tag{4.2.19}$$

we have

$$R(\delta_0,G) \geq \frac{\sigma^2}{M} = R(\delta_T, G) \quad . \tag{4.2.20}$$

Hence, there obviously always exists a $G\varepsilon\mathcal{G}_2$ such that (4.2.20) is true so that

$$\sup_{G\varepsilon\mathcal{G}_2} R(\delta_0,G) \geq \sup_{G\varepsilon\mathcal{G}_2} R(\delta_T,G). \tag{4.2.21}$$

That is, if we choose $\alpha = t_{N+1}$ then we can never be sure that δ_0 is better than δ_T and, in fact, δ_0 may be much worse.

Just as in the previous case (μ known) we can improve upon δ_0 defined by (4.2.16) if we are willing to assume

$$-\infty < \mu_L \leq \mu \leq \mu_u < \infty. \tag{4.2.22}$$

That is, if we are willing to place upper and lower limits within which the unknown μ is assumed to lie. Again, this need not be a strong restriction in practice since $|\mu_L|$ and $|\mu_u|$ can be quite large.

We denote the class of conjugate priors with variance μ_2 and with mean in the interval $[\mu_L, \mu_u]$ by G_2^+.

Under the restriction (4.2.22), the only estimators we need consider are of the form (4.2.15) where $\mu_L \leq \alpha \leq \mu_u$. From (4.2.18) it is easy to see that

$$\sup_{G\varepsilon\, G_2^+} R(\delta_0, G) = \frac{M(\mu_2)^2\sigma^2 + \sigma^4[\mu_2 + \sup_G(\mu-\alpha)^2]}{(\sigma^2 + M\mu_2)^2} \tag{4.2.23}$$

where

$$\sup_{G\varepsilon\, G_2^+} (\mu-\alpha)^2 = \max\left\{(\mu_L-\alpha)^2,\ (\mu_u-\alpha)^2\right\} \leq (\mu_u-\mu_L)^2. \tag{4.2.24}$$

If we let

$$\alpha^* = \mu_u - \frac{|\mu_u-\mu_L|}{2} = \mu_u + \mu_L \tag{4.2.25}$$

then we see that

$$\sup_{G\varepsilon G_2^+} (\mu-\alpha^*)^2 = (\mu_u-\mu_L)^2 / 4. \tag{4.2.26}$$

Defining δ_0^* by:

$$\delta_0^*(t_{N+1}) = \frac{\sigma^2\alpha^* + M\mu_2 t_{N+1}}{\sigma^2+M\mu_2} \tag{4.2.27}$$

then we have:

$$\sup_{G\varepsilon\, G_2^+} R(\delta_0^*, G) = \inf_{\delta} \sup_{G\varepsilon\, G_2^+} R(\delta,G) \tag{4.2.28}$$

that is, δ_0^* is G_2^+-minimax. As a comparison with the usual estimator we note:

$$\sup_{G\varepsilon\, G_2^+} \frac{R(\delta_0^*,G)}{R(\delta_T,G)} = \frac{(M\mu_2)^2+\sigma^2 M\mu_2+\sigma^2 M\left[\frac{(\mu_L-\mu_u)^2}{4}\right]}{(\sigma^2+M\mu_2)^2} \tag{4.2.29}$$

and the above "worst possible" or maximum ratio is ≤ 1 iff:

$$(M\mu_2)^2 + \sigma^2 M\mu_2 + \sigma^2 M\left[\frac{(\mu_L-\mu_u)^2}{4}\right] \leq (\sigma^2+M\mu_2)^2 \tag{4.2.30}$$

$$<=> \frac{M}{4}(\mu_L-\mu_u)^2 \leq M\mu_2 + \sigma^2$$

$$<=> (\mu_L-\mu_u)^2 \leq 4\left(\frac{\sigma^2}{M} + \mu_2\right) = 4 \text{ Var } (T)$$

i.e., if $(\mu_L-\mu_u)^2$ is less than or equal to four times the (unconditional) variance of T then the G_2^+-minimax estimator δ_0^* is to be preferred to δ_T for all $G\varepsilon G_2^+$. Note, however, that even if the inequality (4.2.30) is not satisfied δ_0^* may still be preferred to δ_T. This depends on just how close α^* is to the true μ.

As a particular example, suppose $\mu_2=4$, $\sigma^2=1$ and $M=2$. Then, $(\mu_L-\mu_u)^2$ must be less than or equal to 14. If it is true that $\mu_L=-3$, $\mu_u=0$ then $(\mu_L-\mu_u)^2=9<16$ and regardless of the true μ,

$$\delta_0^*(t_{N+1}) = \left[\sigma^2(-3/2) + M\mu_2 t_{N+1}\right] / (\sigma^2+M\mu_2)^2 \tag{4.2.31}$$

is a better estimator than δ_T. In fact,

$$R(\delta_T,G) = \frac{1}{2} = .50 \;\forall\; G\varepsilon\mathcal{G}_2 \tag{4.2.32}$$

and

$$\sup_{G\varepsilon\mathcal{G}_2} R(\delta_0^*,G) = \frac{51}{108} = .47 \tag{4.2.33}$$

while the Bayes envelope function is: (by (3.2.7):

$$R(G) = \frac{\sigma^2}{M+n'} = \frac{1}{2+\frac{1}{4}} = \frac{4}{9} = .44 \tag{4.2.34}$$

regardless of the particular $G\varepsilon\mathcal{G}_2^+$.

4.3 Estimation Of λ_{N+1}: Moments Known

In this section we do not assume that G is conjugate but only that certain prior moments are known (or at least known to lie in some interval of possible values).

For example, if μ is known then we can do no better than t_{N+1} mainly because we have no way to estimate other moments, e.g., the variance μ_2. However, as in the last section, suppose it is known that

$$0 \leq \mu_2 \leq \mu_2^+ < \infty \quad . \tag{4.3.1}$$

We denote the class of all priors with mean μ and variance bounded by μ_2^+ by G_3.

Note that although G_3 does not contain all degenerate distributions on Λ, it is true that G_3-admissibility implies admissibility since for every $\lambda_0 \in \Lambda$ there exists a $G_{\lambda_0} \in G_3$ such that $G_{\lambda_0}(\lambda_0) - G_{\lambda_0}(\lambda_0^-) > 0$.

To see this, we let $\lambda_0 = \mu-\alpha$ and choose G_{λ_0} as that distribution function that puts weight ε at λ_0 and weight $1-\varepsilon$ at $\lambda_0^* = \mu+\beta$ where $\beta = \left(\frac{\varepsilon}{1-\varepsilon}\right)\alpha$. Then,

$$\int_\Lambda \lambda dG\lambda_0(\lambda) = \lambda_0 \ \varepsilon + (\lambda_0^*)(1-\varepsilon) \tag{4.3.2}$$

$$= (\mu-\alpha)\varepsilon + \left(\mu - \frac{\varepsilon}{1-\varepsilon} \cdot \alpha\right)(1-\varepsilon) \quad = \mu$$

as required of any distribution in G_3. Also, we can always choose ε so that the variance is less than μ_2^+ as required by (4.3.1). This is true since

$$\int (\lambda-\mu)^2 dG_{\lambda_0}(\lambda) = (\lambda_0-\mu)^2\varepsilon + (\lambda_0^*-\mu)^2(1-\varepsilon) \tag{4.3.3}$$

$$= \alpha \left(\alpha\varepsilon + \frac{\varepsilon^2}{1-\varepsilon} \right)$$

and the above variance approaches zero as ε approaches zero.

Hence, regardless of the size of $|\lambda-\lambda_0|$ we can always choose ε so that the variance is less than μ_2^+. That is, $G_{\lambda_0} \varepsilon \mathcal{G}_3$ for any $\lambda_0 \varepsilon \Lambda$.

A good guess at a least favorable distribution in $\mathcal{G}_3$ is any distribution with mean μ and variance μ_2^+. For example, if we consider a $N(\mu,\mu_2^+)$ distribution, say G_0, then δ_0 (given by (4.2.9)) is the Bayes estimator with respect to G_0 and the Bayes risk for δ_0 is given by (4.2.10).

Also, just as for the class $\mathcal{G}_1^+$ in the previous section, it is obvious (by (4.2.11)) that for all $G \varepsilon \mathcal{G}_3$:

$$R(\delta_0,G) \leq R(\delta_0,G_0) \qquad (4.3.4)$$

so that G_0 is least favorable in $\mathcal{G}_3$ (as well as $\mathcal{G}_1^+$) and δ_0 is $\mathcal{G}_3$-minimax (as well as $\mathcal{G}_1^+$-minimax). Hence, although $\mathcal{G}_1^+ \subset \mathcal{G}_3$, the restriction $G\varepsilon\mathcal{G}_1^+$ does not enable us to lower the Bayes risk from the case $G\varepsilon\mathcal{G}_3$. That is, knowledge that $G\varepsilon\mathcal{G}_1^+$ does not enable us to find a better estimator (in the current context) then that attainable when $G\varepsilon\mathcal{G}_3$. This situation arises because restriction of G to the conjugate class still entails consideration of the least favorable distribution in $\mathcal{G}_3$. Hence we have one more example of a situation in which restriction of consideration to the conjugate class is really not a stringent restraint.

We note that if the true G in effect is a $N(\mu,\mu_2^*)$, say G^*, then we have:

$$\frac{R(\delta_o,G^*)}{R(\delta_{G^*},G^*)} = \frac{\sigma^2 M(\mu_2^+)^2+\sigma^4\mu_2^*}{(\sigma^2+M\mu_2^+)} \cdot \frac{\sigma^2+M\mu_2^*}{\sigma^2\mu_2^*} \tag{4.3.5}$$

$$= \frac{\sigma^2+M\mu_2^*}{(\sigma^2+M\mu_2^+)^2} \cdot \frac{M(\mu_2^+)^2+\sigma^2\mu_2^*}{\mu_2^*}$$

$$= 1 + \frac{(\alpha-1)^2}{\alpha} \cdot \frac{M\sigma^2\mu_2^+}{(\sigma^2+M\mu_2^+)^2}$$

where $\alpha=\mu_2^*/\mu_2^+$. The above ratio is always ≥ 1 but is small if M or α is large. In fact, as would be expected, $R(\delta_0,G^*)/R(\delta_{G^*},G^*)$ approaches one as $M \to \infty$ or $\alpha \to 1$.

We next assume that μ and μ_2 are both known. The class of all distributions on Λ with mean = μ and variance = μ_2 will be denoted by $\mathcal{G}_4$.

Now, the estimator

$$\delta_0^*(t_{N+1}) = \frac{\sigma^2\mu+M\mu_2 t_{N+1}}{\sigma^2+M\mu_2} \tag{4.3.6}$$

has a Bayes risk of

$$R(\delta_0^*,G) = \frac{\sigma^2\mu_2}{\sigma^2+M\mu_2} \quad \forall\ G\epsilon\mathcal{G}_4 \tag{4.3.7}$$

i.e., the Bayes risk for δ_0^* is constant over the class $\mathcal{G}_4$. Also, δ_0^* is Bayes with respect to a $N(\mu,\mu_2)$ and hence by theorem 5, δ_0^* is $\mathcal{G}_4$-minimax and a $N(\mu,\mu_2)$ distribution is least favorable. Here, again, we note that:

$$R(\delta_0^*,G) \leq R(\delta_T,G) \quad \forall G \varepsilon \mathcal{G}_4 \tag{4.3.8}$$

so that regardless of the true G in effect, δ_0^* is preferred to δ_T.

4.4 Asymptotic $\mathcal{G}$-Minimax Decision Functions

In this section we derive various asymptotic $\mathcal{G}$-minimax estimators. Here, we have N past observations on $f_G(t)$ and we wish to use these past observations to form an estimator δ_N, in such a way that

$$\lim_{N \to \infty} R(\delta_N,G) \to R(\delta_{\mathcal{G}},G) \quad \forall G \varepsilon \mathcal{G} \tag{4.4.1}$$

where $\delta_{\mathcal{G}}$ is the $\mathcal{G}$-minimax estimator for the class $\mathcal{G}$.

Assume, in the first case, that the unknown μ_2 is such that

$$0 \leq \mu_2 \leq \mu_2^+ < \infty \tag{4.4.2}$$

where μ_0^+ is some known finite upper bound for μ_2. Next we define

$$\hat{\mu} = \sum_{i=1}^{N+1} t_i \,/(N+1) \tag{4.4.3}$$

as in previous sections and consider the estimator:

$$\delta_N(t_{N+1}) = \frac{\sigma^2\hat{\mu}+M\mu_2^+ t_{N+1}}{\sigma^2+M\mu_2^+} . \tag{4.4.4}$$

We have seen that

$$\hat{\mu} \xrightarrow{\text{a.s.}} \mu \tag{4.4.5}$$

and hence that

$$\delta_N(t_{N+1}) \xrightarrow{\text{a.s.}} \delta_0(t_{N+1}) \tag{4.4.6}$$

where $\delta_0\ (t_{N+1})$ is defined by (4.2.9). Now, we can rewrite (4.4.4) as:

$$\delta_N(t_{N+1}) = \frac{\sigma^2\left(\frac{\sum_{i=1}^{N} t_i + t_{N+1}}{N+1}\right) + M\mu_2^+(t_{N+1})}{\sigma^2+M\mu_2^+} \tag{4.4.7}$$

$$= \frac{\sigma^2 \sum_{i=1}^{N} t_i}{N+1} + t_{N+1}\left(M\mu_2^+ + \frac{\sigma^2}{N+1}\right) .$$

The global risk of δ_N is computed by first finding:

$$R_N(\delta_N, G) = E_{t,\lambda}\left\{\delta_N(t_{N+1}) - \lambda\right\}^2 \tag{4.4.8}$$

$$= E_\lambda\left\{\frac{\left(M\mu_2^+ + \frac{\sigma^2}{N+1}\right)^2 \frac{\sigma^2}{M}}{(\sigma^2+M\mu_2^+)} + \left[\frac{\frac{\sigma^2\Sigma t_i}{N+1} + \lambda\left(M\mu_2^+ + \frac{\sigma^2}{N+1}\right)}{(\sigma^2+M\mu_2^+)} - \lambda\right]^2\right\}$$

$$= \frac{\left(M\mu_2^+ + \frac{\sigma^2}{N+1}\right)^2 \frac{\sigma^2}{M} \quad + \quad \frac{N^2\sigma^4 E_\lambda\{\Sigma t_i - N_\lambda\}^2}{(N+1)^2(\sigma^2+M\mu_2^+)^2}}{(\sigma^2+M\mu_2^+)^2}$$

$$= \frac{\left[M\mu_2^+(N+1)+\sigma^2\right]^2\sigma^2 + M\sigma^4N^2\left\{\mu_2 + \left[\mu - \frac{\Sigma t_i}{N}\right]^2\right\}}{M(N+1)^2(\sigma^2+M\mu_2^+)^2}$$

and $\frac{\Sigma t_i}{N} \sim N\left(\mu, \frac{\sigma^2+M\mu_2}{MN}\right)$. Hence,

$$R(\delta_N,G) = E[R_N(\delta_N,G)] \tag{4.4.9}$$

$$= \frac{(M(N+1)\mu_2^+ + \sigma^2)^2\sigma^2 + M\sigma^4N^2\left\{\mu_2 + \frac{\sigma^2+M\mu_2}{MN}\right\}}{M(N+1)^2(\sigma^2+M\mu_2^+)^2}$$

and therefore we have:

$$\lim_{N\to\infty} R(\delta_N,G) = \frac{M^2(\mu_2^+)^2\sigma^2+M\sigma^4\mu_2}{M(\sigma^2+M\mu_2^+)^2} \tag{4.4.10}$$

$$= \frac{M(\mu_2^+)\sigma^2+\sigma^4\mu_2}{(\sigma^2+M\mu_2^+)^2}$$

$$= R(\delta_0,G), \quad \forall G\varepsilon \mathcal{G}_3$$

i.e., δ_N is asymptotically $\mathcal{G}_3$-minimax. As a matter of fact, δ_N is also asymptotically G_1^+-minimax (see (4.2.11)).

Next, consider the case where we are unwilling to assume known bounds for μ or μ_2. Denote the class of all distribution functions on Λ with μ_2 finite (but unknown) by G_4. The estimator used in this case is:

$$\delta_N(t_{N+1}) = \frac{\sigma^2\hat{\mu}+M\hat{\mu}t_{N+1}}{\sigma^2+M\hat{\mu}_2} \tag{4.4.11}$$

where $\hat{\mu}$ and $\hat{\mu}_2$ are defined by:

$$\hat{\mu} = \sum_{i=1}^{N+1} t_i/(N+1) \tag{4.4.12}$$

$$\hat{\mu}_2 = \max\left\{0,\ s_t - \frac{\sigma^2}{M}\right\},$$

as in (3.2.20).

Now, since the steps following (3.2.25) establishing the asymptotic optimality of the estimator δ_N for the class of conjugate priors are independent of the normality assumtion, we have that

$$\lim_{N\to\infty} R(\delta_N^*,G) = R(\delta_0^*,G) \quad \text{for all } G\varepsilon G_4 \tag{4.4.13}$$

where δ_0^* is defined by

$$\delta_0^*(t_{N+1}) = \frac{\mu\sigma^2+M\mu_2 t_{N+1}}{\sigma^2+M\mu_2}. \tag{4.4.14}$$

That is, δ_N^* is an asymptotic G_4- minimax estimator.

CHAPTER V

BERNOULLI PROCESS I: EMPIRICAL BAYES DECISION FUNCTIONS

5.1 Definitions And Preliminary Results

The Bernoulli process as used in this chapter is defined as a generator of X's with probability mass function:

$$\lambda^{x}(1-\lambda)^{1-x} \qquad x=0,1 \qquad 0<\lambda<1 \quad . \tag{5.1.1}$$

The Bernoulli process could also be defined as a generator of $\tilde{N}$'s with probability mass function (p.m.f.)

$$\lambda(1-\lambda)^{N-1} \qquad N=1,2,\ldots \qquad 0<\lambda<1 \quad . \tag{5.1.2}$$

Here, for example, N may represent the number of trials required to obtain the first success with a probability of success λ at each trial. If we let t' equal the number of N's observed and $M=\Sigma N_i$ then the likelihood of the sample is proportional to:

$$\lambda^{t'}(1-\lambda)^{M-t'} \tag{5.1.3}$$

which in turn is proportional to (5.1.1). Since the prior densities will be the same in either definition of the Bernoulli process then by the likelihood principle (Lindley [21], p. 59) any statement made about the parameter λ using t' should be the same as that made using Σx_i. Hence, for convenience, we consider only the first definition of the Bernoulli process.

A current random sample $\left\{x_{N+1,j}\right\}_{j=1}^{M}$ of size M is chosen, each $x_{N+1,j}$ having the p.m.f. given by (5.1.1). The past observations are denoted, as in previous chapters, by

$$\left\{x_{ij}\right\}_{j=1}^{M}, \ldots, \left\{x_{N,j}\right\}_{j=1}^{M} \tag{5.1.4}$$

where in each sample, for $i \in \left\{1, \ldots, N+1\right\}$,

$$t_i = \sum_{j=1}^{M} x_{ij} \tag{5.1.5}$$

is a sufficient statistic for λ_i. The problem, as usual, is to make some inference concerning the unknown λ_{N+1}.

It is well-known that

$$T_i/\lambda_i \sim B(M,\lambda_i) \ . \tag{5.1.6}$$

That is, each T_i has a binomial distribution with (known) parameter M and unknown parameter λ_i. The p.m.f. for each T_i is simply

$$f(t|\lambda) = \binom{M}{t}\lambda^t(1-\lambda)^{M-t} \tag{5.1.7}$$

$$t \in \{0,1,\ldots,M\}$$

$$0 < \lambda < 1$$

with $E(T) = M\lambda$ and $E(T-M\lambda)^2 = M\lambda(1-\lambda)$. Also, the set $\left\{t_i\right\}_{i=1}^{N+1}$ can be considered a random sample from the compound or unconditional distribution with p.m.f.:

$$f_G(t) = \int_\Lambda \binom{M}{t}\lambda^t(1-\lambda)^{M-t}dG(\lambda) \tag{5.1.8}$$

$$\text{where } \Lambda = [0,1]$$

and $G(\lambda)$ is the prior probability distribution function of the parameter λ. The conjugate prior distribution in this case is just a beta distribution with p.d.f.

$$g(\lambda|R',N') = \frac{\Gamma(N')}{\Gamma(R')\Gamma(N'-R')} \lambda^{R'-1}(1-\lambda)^{N'-R'-1} \tag{5.1.9}$$

$$0 < \lambda < 1, \qquad N' > R' > 0$$

and is denoted by $Be(R',N'-R')$. The first few moments of the beta distribution are:

$$\mu = \frac{R'}{N'} \tag{5.1.10}$$

$$\mu_2' = \frac{R'(R'+1)}{N'(N'+1)} = \frac{\mu(N'\mu+1)}{N'+1}$$

$$\mu_2 = \frac{R'(N'-R')}{(N')^2(N'+1)} = \frac{\mu(1-\mu)}{N'+1}$$

and the parameters R' and N' in terms of the first two moments are:

$$R' = \frac{\mu(\mu-\mu_2')}{\mu_2} \qquad (5.1.11)$$

$$N' = \frac{\mu-\mu_2'}{\mu_2} \quad .$$

Some useful facts concerning distributions on [0,1] in general and beta distributions in particular are derived below.

For any distribution on [0,1] it is obvious that

$$\int_\Lambda \lambda^j dG(\lambda) > \int_\Lambda \lambda^{j+1} dG(\lambda) \qquad (5.1.12)$$

$$j=0,1,2,\ldots$$

since for $0 < \lambda < 1$,

$$\lambda^j > \lambda^{j+1} \qquad j=0,1,2,\ldots \quad . \qquad (5.1.13)$$

Also, since the range of $\tilde{\lambda}$ is bounded it is true that

$$\mu_j' < \infty \qquad j=1,2,\ldots \quad . \qquad (5.1.14)$$

In fact, an equivalent way to write (5.1.12) is:

$$\mu_j' > \mu_{j+1}' \qquad j=0,1,2,\ldots \qquad (5.1.15)$$

and Skibinsky [40] shows that the maximum range for μ_j' is 2^{-2j+2}.

From (5.1.15) we have that

$$\mu_2' \leq \mu \qquad \text{i.e.,} \mu - \mu_2' \geq 0 \qquad (5.1.16)$$

although, of course,

$$\mu_2 = \mu_2' - \mu^2 \geq 0 \qquad (5.1.17)$$

as is true for any distribution.

The following inequalities are also useful in the later sections. For any distribution on $\Lambda = [0,1]$:

$$\mu_2 = \mu_2' - \mu^2 \leq \mu - \mu^2 \leq \frac{1}{4} \qquad (5.1.18)$$

$$\mu - \mu_2' \leq \mu - \mu^2 \leq \frac{1}{4}$$

$$\mu^2 \leq \mu_2' \leq \mu$$

For a Be $(R', N'' - R'')$ distribution (see(5.1.10)):

$$\mu_2 = \mu_2' - \mu^2 = \frac{R'(N'-R')}{(N')^2(N'+1} \qquad (5.1.19)$$

$$= \frac{\mu(N'\mu+1)}{N'+1} - \mu^2$$

$$= \frac{\mu-\mu^2}{N'+1}$$

$$\mu-\mu_2' = \frac{R'(N'-R')}{N'(N'+1)} = \mu - \frac{\mu(N'\mu+1)}{N'+1}$$

$$= \frac{N'}{N'+1} (\mu-\mu^2)$$

$$= N'\mu_2$$

and if $N' \geq 1$

$$\mu_2 \leq \mu - \mu_2' \leq \mu \qquad (5.1.20)$$

Note that for any distribution on [0,1] we have (by (5.1.18)

$\mu_2 \leq \mu - \mu^2$; i.e., $\frac{1-\sqrt{1-4\mu_2}}{2} \leq \mu \leq \frac{1+\sqrt{1-4\mu_2}}{2}$ and $\mu^2 \leq \mu_2' \leq \mu$.

That is, for a given μ_2, μ can take values only in a strictly defined interval and for a given μ, μ_2' (or μ_2) can take values only in a fixed interval. See figure 4. Hence, for any beta distribution

$$\mu_2 \leq \frac{1}{4(N'+1)} \tag{5.1.21}$$

$$\mu - \mu_2' \leq \left(\frac{N'}{N'+1}\right) \frac{1}{4} \quad .$$

When $G(\lambda)$ is a Be(R', N'-R') distribution the unconditional density of t becomes

$$f_G(t) = \frac{\Gamma(t+R')\Gamma(M+N'-t-R')\Gamma(M+1)\Gamma(N')}{\Gamma(t+1)\Gamma(M-t+1)\Gamma(N'-R')\Gamma(M+N')} \quad . \tag{5.1.22}$$

That is, T has a beta - binomial distribution with parameters R', N' and M. This distribution has the following first few moments:

$$E(T) = \frac{MR'}{N'} = M\mu \tag{5.1.23}$$

$$E(T^2) = M(M-1)\mu_2' + M\mu$$

$$\text{Var }(T) = M(M-1)\mu_2 + M\mu(1-\mu)$$

$$= \frac{M(M+N')R'(N'-R')}{(N')^2(N'+1)} \quad .$$

The posterior density of $\tilde{\lambda}$ given t_{N+1} when the prior density is a Be(R',N'-R') is:

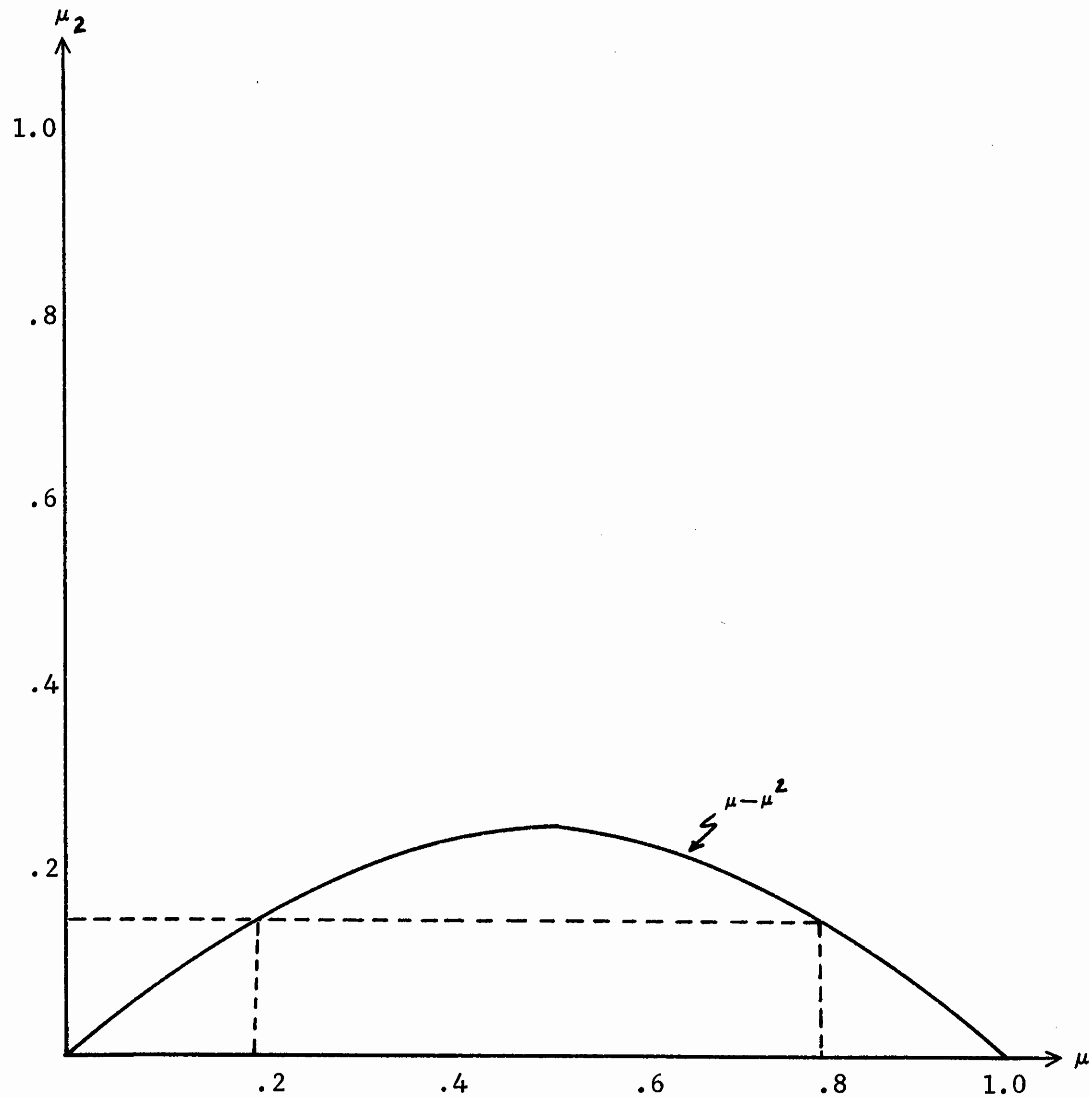

Figure 4. Upper and lower limits for μ and μ_2.

$$g(\lambda \mid t_{N+1}) = \frac{\Gamma(N'')}{\Gamma(R'')\Gamma(N''-R'')} \lambda^{R''-1} (1-\lambda)^{N''-R''-1} \tag{5.1.24}$$

that is, $\tilde{\lambda} \mid t_{N+1} \sim Be(R'', N''-R'')$ where

$$R'' = R' + t_{N+} \tag{5.1.25}$$

$$N'' = N' + M$$

and this posterior distribution has a mean and variance of:

$$E(\tilde{\lambda}) = \frac{R''}{N''} = \frac{R'+t_{N+}}{N' + M} \tag{5.1.26}$$

$$E\left\{\tilde{\lambda}-\frac{R''}{N''}\right\}^2 = \frac{R''(N''-R'')}{(N'')^2(N''+1)}$$

$$= \frac{(R'+t_{N+1})(N'+M-R'-t_{N+1})}{(N'+M)^2(N'+M+1)} .$$

Notice that for $M_1 < M_2$

$$Var_{M_1}(\tilde{\lambda} \mid t_{N+1}) > Var_{M_2}(\tilde{\lambda} \mid t_{N+1}) \tag{5.1.27}$$

where $Var_{M_i}(\cdot \mid t)$ represents the posterior variance of $\tilde{\lambda}$ for a sample size M_i.

5.2 Estimation Of λ_{N+1}

When attempting to derive empirical Bayes decision functions in the binomial case, it is important to note that

$$f_G(t) = \int_\Lambda \binom{M}{t} \lambda^t (1-\lambda)^{M-t} dG(\lambda) \tag{5.2.1}$$

$$= \binom{M}{t} \int_\Lambda \lambda^t \left[\sum_{j=0}^{M-t} \binom{M-t}{j} (-1)^j \lambda^j \right] dG(\lambda)$$

$$= \binom{M}{t} \sum_{j=0}^{M-t} (-1)^j \binom{M-t}{j} \int_\Lambda \lambda^{t+j} dG(\lambda)$$

$$= \binom{M}{t} \sum_{j=0}^{M-t} (-1)^j \binom{M-t}{j} \mu'_{t+j}$$

and hence if any two distribution functions G_1 and G_2 are such that the first M moments of G_1 are identical with the first M moments of G_2 then $f_{G_1}(t) \equiv f_{G_2}(t)$ so that $f_G(t)$ is not "identifiable" in the class of all distirbution functions on Λ.

Definition

For a given class $\mathcal{G}$ of distribution functions on Λ, if $f_{G_1}(t) = f_{G_2}(t)$, $G_1, G_2 \in \mathcal{G} \Rightarrow G_1 \equiv G_2$ then the mixture $f_G(t)$ is said to be identifiable in $\mathcal{G}$. Now, in general since $f_{G_1}(t) \equiv f_{G_2}(t)$ does not imply that $\delta_{G_1}(t) \equiv \delta_{G_2}(t)$ then if $\mathcal{G}$ is unrestricted it is impossible to find an asymptotically optimal EB decision function for this case since the EB procedure depends on $f_G(t)$. However, it may be possible to find an a.o. EB d.f. for some restrictive class $\mathcal{G}$.

The loss function to be used in estimating λ_{N+1} is:

$$L(\delta(t),\lambda) = (\delta(t)-\lambda)^2 \tag{5.2.2}$$

The usual (minimum variance unbiased) estimator of λ_{N+1} is denoted by δ_T:

$$\delta_T(t_{N+1}) = t_{N+1}/ M \tag{5.2.3}$$

and has a Bayes risk of

$$\begin{aligned} R(\delta_T,G) &= E_\lambda E_{t/\lambda} (t/M-\lambda)^2 && (5.2.4)\\ &= E_\lambda \text{ Var } (T/M) \\ &= E_\lambda \frac{\lambda(1-\lambda)}{M} \\ &= \frac{\mu-\mu_2'}{M} \end{aligned}$$

for any prior G with mean μ and second moment μ_2'. (we are always assuming $\int_\Lambda dG(\lambda) = 1$). From (5.1.17) and (5.1.18) we note that for any distribution G it is true that

$$0 \leq R(\delta_T,G) \leq \frac{\mu-\mu^2}{M} \leq \frac{1}{4M} \tag{5.2.5}$$

and for a Beta (R', N'-R') prior

$$0 \leq R(\delta_T,G) = \frac{1}{M} \cdot \frac{N'}{N'+1} \quad (\mu-\mu^2) \leq \frac{1}{4M} \cdot \frac{N'}{N'+1} \quad . \tag{5.2.6}$$

The Bayes estimator δ_G with respect to a Be(R', N'-R') prior is:

$$\delta_G(t_{N+1}) = \frac{t_{N+1} + R'}{M + N'} \tag{5.2.7}$$

$$= \frac{t_{N+1} + \frac{\mu(\mu-\mu_2')}{\mu_2}}{M + \frac{(\mu-\mu_2')}{\mu_2}}$$

$$= \frac{t_{N+1}\,\mu_2 + \mu(\mu-\mu_2')}{M\mu_2 + (\mu-\mu_2')}$$

with a risk function of

$$R(\delta_G,\lambda) = E_{t/\lambda}(\delta_G(t) - \lambda)^2 \tag{5.2.8}$$

$$= \frac{M\lambda(1-\lambda) + (R'-\lambda N')^2}{(M+N')^2}$$

and a Bayes risk of

$$R(\delta_G,G) = E_\lambda\, R(\delta_G,\lambda) \tag{5.2.9}$$

$$= \frac{\mu_2'[(N')^2-M]+\mu[M-2R'N']+(R')^2}{(M+N')^2}$$

$$= \frac{(\mu-\mu_2')(\mu_2'-\mu^2)}{(M-1)\mu_2' + (\mu-\mu^2 M)}$$

$$= \frac{\mu-\mu_2'}{M+\frac{\mu-\mu_2'}{\mu_2}} = \frac{N'\mu_2}{M+N'} = \frac{R'(N'-R')}{(M+N')N'(N'+1)} .$$

the various forms of $R(\delta_G,G)$ given above are useful in the following text. From (5.1.18) we note that for a given μ,

$$0 \leq \mu_2 < \mu-\mu^2 \tag{5.2.10}$$

and $\mu_2 \to 0$ as $N' \to \infty$, $\mu_2 \to \mu-\mu^2$ as $N' \to 0$. Hence as $N' \to 0$ we have increasingly vague prior information. Of course, $N'=0$ is not allowed in a proper beta distribution as given by (5.1.9). However, just as in the normal case a "conditional" density corresponding to $N' = R' = 0$ can be defined. By (5.1.9) this density is proportional to

$$\lambda^{-1}(1-\lambda)^{-1} \qquad 0 < \lambda < 1 \quad . \tag{5.2.11}$$

Now, for any λ_0 and λ_1 such that

$$0 < \lambda_0 < \lambda_1 < 1 \tag{5.2.12}$$

we can define, for $\lambda \varepsilon [\lambda_0, \lambda_1]$,

$$\begin{aligned} g(\lambda|\lambda_0,\lambda_1) &= \left\{\lambda(1-\lambda)\right\}^{-1} \Big/ \int_{\lambda_0}^{\lambda_1} \frac{d\lambda}{\lambda(1-\lambda)} \\ &= \left\{\lambda(1-\lambda)\right\}^{-1} \Big/ \ln\left(\frac{1-\lambda_0}{\lambda_0} \cdot \frac{\lambda_1}{1-\lambda_1}\right) \end{aligned} \tag{5.2.13}$$

and hence

$$\int_{\lambda_0}^{\lambda_1} g(\lambda|\lambda_0,\lambda_1)\, d\lambda = 1 \quad . \tag{5.2.14}$$

In the above sense then, a beta prior with $N'=R'=0$ is defined and this prior corresponds to the most vague information in the beta class of priors. The posterior distribution here is a $Be(t_{N+1}, M-t_{N+1})$ and is itself an improper density unless $M > t_{N+1} > 0$; i.e., unless both a "success" and a "failure" have been observed. It should be noted in passing that a commonly used distribution representing vague prior information, namely a uniform or $Be(1,1)$ distribution, results in a posterior distribution of $Be(t_{N+1},+1,M-t_{N+1}+1)$ so that the difference in posterior distributions resulting from using a $Be(0,0)$ and a $Be(1,1)$ prior is slight (see Lindley [21], p. 145). Also, since uniform priors are intuitively appealing it should be noted that when $N'=R'=0$ the random variable $\Phi=\ln\tilde{\lambda}-\ln(1-\tilde{\lambda})$ has a ("conditional") uniform distribution on the real line and that writing the density for T as a member of the exponential family, we find

$$\begin{aligned} f(t|\lambda) &\propto \binom{M}{t} e^{t[\ln\lambda - \ln(1-\lambda)]} \\ &\propto \binom{M}{t} e^{t\Phi} \end{aligned} \tag{5.2.15}$$

so that Φ, rather than the usual λ, could easily be thought of as the "natural" parameter for the binomial distribution.

Next, consider the estimator

$$\delta_0(t_{N+1}) = \frac{t_{N+1} + \sqrt{M}/2}{M + \sqrt{M}} \quad . \tag{5.2.16}$$

This is the usual minimax estimator as given by Ferguson [13] (pages 93-94) and in Lehmann's "Notes on Estimation". For any distribution function G on Λ (not necessarily beta) the Bayes risk is

$$R(\delta_0,G) = E_{t,\lambda}\left[\delta_0(t)-\lambda\right]^2 \tag{5.2.17}$$

$$= E_\lambda\left\{\frac{M[\lambda(1-\lambda)] + M(\frac{1}{2}-\lambda)^2}{(M+\sqrt{M})^2}\right\}$$

$$= \frac{1}{4(\sqrt{M}+1)^2} \quad .$$

It should be noted that in some cases δ_0 is preferred to δ_T and especially for small M one should be aware of this fact.

To see this, consider

$$\frac{R(\delta_0,G)}{R(\delta_T,G)} = \frac{M}{4(\sqrt{M}+1)^2(\mu-\mu_2')} \quad . \tag{5.2.18}$$

Now the above is ≤ 1 (i.e., δ_0 preferred to δ_T) iff

$$\frac{M}{4(\sqrt{M}+1)^2} \leq \mu-\mu_2' \quad . \tag{5.2.19}$$

For example, if M=1, δ_0 is preferred if $\mu-\mu_2' \geq \frac{1}{16}$ and if M=25, δ_0 is preferred if $\mu-\mu_2' \geq \frac{25}{144}$.

Of course, for large M it is unlikely that (5.2.19) holds (or, at least, unlikely to be known) since

$$\lim_{N\to\infty} \frac{M}{4(\sqrt{M}+1)^2} = \frac{1}{4} \tag{5.2.20}$$

and we know that $\mu-\mu_2' \leq \frac{1}{4}$. However, for large M even δ_T has a Bayes risk close to that of δ_G (actually $R(\delta_T,G)$ is only $\frac{M+N'}{M}$ times as large as $R(\delta_G,G)$ in any case) so that our main concern is with small M. (see section 6.2 for a more complete discussion of δ_0.)

Since δ_0 has a constant Bayes risk for all G and since δ_0 is Bayes with respect to a $Be(\sqrt{M}/2, \sqrt{M}/2)$ then, by theorem 5, the least favorable distribution in the class of all distribution functions is a $Be(\sqrt{M}/2, \sqrt{M}/2)$.

Therefore, just as in the normal case, the least favorable distribution among all distribution functions is a member of the conjugate class. That is, the conjugate class is not only rich (many different possible distributions) and tractable (e.g., the posterior distribution is also a member of the conjugate class) but is also conservative - in the sense that is we restrict attention to the conjugate class we must have the "worst possible" (i.e., least favorable) distribution still under consideration.

Also, by restricting our attention to the class $\mathcal{G}$ of beta priors then any two distribution functions with the same two first moments are identical. Hence, if M > 1 there is no problem of identifiability in the class $\mathcal{G}$. (i.e., $f_{G_1} \equiv f_{G_2} \Rightarrow G_1 \equiv G_2$). Throughout the remainder of this chapter then, we will restrict our attention to the class $\mathcal{G}$ of beta priors.

The following inequality on the Bayes risk of a Bayes d.f. (assuming a beta prior) is useful:

$$R(\delta_G, G) = \frac{\mu - \mu_2'}{M + \dfrac{\mu - \mu_2'}{\mu_2}} \tag{5.2.21}$$

$$= \frac{N'\mu_2}{M + N'} \quad \text{by (5.2.9)}$$

$$= \frac{N'}{M+N'} \cdot \frac{\mu-\mu^2}{N'+1} \quad \text{by (5.1.10)}$$

$$\leq \frac{1}{4} \cdot \frac{N'}{N'+1} \cdot \frac{1}{M+N'}$$

where $N' > 0$, $M \geq 1$.

Now, setting

$$\frac{d}{dN'} \left\{ \frac{1}{4} \cdot \frac{N'}{N'+1} \cdot \frac{1}{M+N'} \right\} = 0 \tag{5.2.22}$$

implies $N' = \sqrt{M}$ and since the upper bound in (5.2.21) is always positive and approaches zero as $N' \to 0$ and as $N' \to \infty$ then for all N' we have

$$R(\delta_G, G) \leq \frac{1}{4} \cdot \frac{\sqrt{M}}{\sqrt{M}+1} \cdot \frac{1}{M+\sqrt{M}} \tag{5.2.23}$$

$$= \frac{1}{4} \left[\frac{1}{\sqrt{M}+1} \right]^2$$

$$= R(\delta_0, G) .$$

Robbins [32] showed that for $M=1$, $R(\delta_G, G) \leq \frac{1}{16}$ which agrees with (5.2.23).

To form an empirical Bayes estimator we will first obtain an estimator G_N of the unknown G from the observations $t_1, \ldots, t_{N+1}$. Then, since the Bayes estimator is, in general, some function $h(G)$ of the prior G our empirical Bayes estimator will be $h(G_N)$. For example, in the estimation situation as considered here, $h(G)$ is the mean of the posterior distribution with prior G and hence $h(G_N)$ is the mean of the posterior with prior G_N. Since we are restricting $\mathcal{G}$ to be the beta class, we also require $G_N \in \mathcal{G}$. To find a consistent estimator G_N we need the following expectations (with respect to the joint distributions of T and $\tilde{\lambda}$):

$$E\left(\frac{T_i}{M}\right) = E_\lambda\left(\frac{M\lambda_i}{M}\right) = \mu \text{ for } i \in 1,2,\ldots,N+1 \qquad (5.2.24)$$

$$E\left(\frac{T_i(T_i-1)}{M(M-1)}\right) = E_\lambda E_{t/\lambda}\left(\frac{t^2-t}{M(M-1)}\right) \qquad (5.2.25)$$

$$= E_\lambda\left\{\frac{M\ (1-\lambda)+M^2\lambda^2-\lambda M}{M(M-1)}\right\}$$

$$= E_\lambda(\lambda^2)$$

$$= \mu_2' \qquad \text{for } i \in \{1,2,\ldots,N+1\} .$$

Then, by the strong law of large numbers:

$$\frac{1}{N+1} \sum_{i=1}^{N+1} \frac{T_i}{M} \xrightarrow{\text{a.s.}} \mu \tag{5.2.26}$$

$$\frac{1}{N+1} \sum_{i=1}^{N+1} \frac{T_i(T_i-1)}{M(M-1)} \xrightarrow{\text{a.s.}} \mu_2' \; .$$

Hence, as our estimate G_N of G we use a beta $(\hat{R}', \hat{N}' - \hat{R}')$ where

$$\hat{R}' = \frac{\hat{\mu}(\hat{\mu}-\hat{\mu}_2')}{\hat{\mu}_2' - \hat{\mu}^2} \tag{5.2.27}$$

$$\hat{N}' = \frac{\hat{\mu} - \hat{\mu}_2'}{\hat{\mu}_2' - \hat{\mu}^2}$$

with

$$\hat{\mu} = \frac{\Sigma t_i}{(N+1)M}$$

$$\hat{\mu}_2' = \begin{cases} \hat{\mu}^2 & \text{if } \frac{1}{N+1} \cdot \frac{\Sigma t_i(t_i-1)}{M(M-1)} \leq \hat{\mu}^2 \\ \hat{\mu} & \text{if } \frac{1}{N+1} \cdot \frac{\Sigma t_i(t_i-1)}{M(M-1)} \geq \hat{\mu} \\ \frac{1}{N+1} \cdot \frac{\Sigma t_i(t_i-1)}{M(M-1)} & \text{otherwise} \end{cases} \tag{5.2.28}$$

where M > 1 and Σ refers to summation over all N+1 variables. A definition for $\hat{\mu}_2'$ that is equivalent to (5.2.28) is:

$$\mu_2' = \max\left\{ \hat{\mu}^2, \frac{1}{N+1} \frac{\Sigma t_i(t_i-1)}{M(M-1)} \right\} \tag{5.2.29}$$

since obviously (forM > 1)

$$\frac{\Sigma t_i^2}{(N+1)M(M-1)} \leq \frac{\Sigma t_i (M)}{(N+1)(M)(M-1)} = \frac{\Sigma t_i}{M(N+1)} + \frac{\Sigma t_i}{(M-1)M(N+1)} \tag{5.2.30}$$

and hence

$$\frac{\Sigma t_i^2}{(N+1)M(M-1)} - \frac{\Sigma t_i}{(M-1)M(N+1)} \leq \frac{\Sigma t_i}{M(N+1)} = \hat{\mu} \tag{5.2.31}$$

and the last restriction in (5.2.28) is superfulous.

If it turns out that $\hat{\mu}^2 \geq \frac{1}{N+1} \cdot \frac{\Sigma t_i(t_i-1)}{M(M-1)}$ then our estimate G_N of G will be a degenerate beta at $\hat{\mu}$. However, this will usually happen only if N is small since

$$\frac{\Sigma t_i(t_i-1)}{(N+1)M(M-1)} \geq \hat{\mu}^2 = \left[\frac{\Sigma t_i}{(N+1)M}\right]^2 \tag{5.2.32}$$

$$\begin{aligned} &\iff (N+1)M\,\Sigma t_i^2 - (N+1)M\Sigma t_i \\ &\geq (M-1)\Sigma t_i^2 + (M-1)\sum_{i\neq j} t_i t_j \\ &\iff 1 - \frac{M-1}{(N+1)M}\,\Sigma t_i^2 - \left(\frac{M-1}{(N+1)M}\right)\sum_{i\neq j} t_i t_j \geq \Sigma t_i \end{aligned}$$

and, as $N \to \infty$ the above inequality becomes

$$\Sigma t_i^2 \geq \Sigma t_i \tag{5.2.33}$$

which is always true since $T_i \in \{0,1,2,\ldots\}$, $\forall i$. That is, for large N we can be nearly certain that $\hat{\mu}_2' = \frac{1}{N+1} \sum_{i=1}^{N+1} \frac{t_i\,(t_i-1)}{M(M-1)}$. In any event, the empirical Bayes estimator to be used is:

$$\delta_N(t_{N+1}) = \frac{t_{N+1} + \hat{R}'}{M + \hat{N}'} \tag{5.2.34}$$

$$= \frac{t_{N+1}(\hat{\mu}_2) + \hat{\mu}\ (\hat{\mu}-\hat{\mu}_2')}{M\hat{\mu}_2 + (\hat{\mu}-\hat{\mu}_2')}$$

and the global risk is:

$$R(\delta_N, G) = E_N E_{\lambda,t}\left\{\delta_N(t)-\lambda\right\}^2 \tag{5.2.35}$$

$$= R(\delta_G, G) + E_N E_{\lambda,t}\left\{\delta_N(t)-\delta_G(t)\right\}^2$$

$$= R(\delta_G, G) + W(\delta_N, G)$$

with the second term above $W(\delta_N,G)$ being the addition to the risk or the "penalty" for not knowing G exactly.

As might be expected, the term $W(\delta_N,G)$ in (5.2.35) is extremely difficult to evaluate analytically for a given finite N (although $\lim_{N \to \infty} W(\delta_N,G)=0$ as is seen below). To evaluate $W(\delta_N,G)$ for small N, a Monte Carlo study was undertaken, the results of which are given in section 5.3. There it appears that δ_N is not only a.o. but usually better than the ordinary estimator δ_T even for small N (see section 5.3 for further details).

Now, by theorem 6 and (5.2.26), it follows that

$$\hat{R}' \xrightarrow{a.s.} \hat{R}' \qquad (5.2.36)$$
$$\hat{N}' \xrightarrow{a.s.} \hat{N}'$$

and hence

$$\left[\frac{t_{N+1} + \hat{R}'}{M + \hat{N}'} - \frac{t_{N+1} + R'}{M + \hat{N}'} \right] \xrightarrow{a.s.} 0 \qquad (5.2.37)$$

or, equivalently,

$$\delta_N(t_{N+1}) - \delta_G(t_{N+1}) \xrightarrow{a.s.} 0 \; . \qquad (5.2.38)$$

Also, (5.2.38) implies (see Tucker [46], p. 105) that

$$\delta_N(t) - \delta_G(t) \xrightarrow{d} 0 \; . \qquad (5.2.39)$$

Finally, since

$$\{\delta_N(t) - \delta_G(t)\}^2 \leq 1 \qquad (5.2.40)$$

for all $t \varepsilon \{o, 1, \ldots, M\}$ then by theorem 3A, Parzen [27], p. 424, it is true that

$$E\{\delta_N(t) - \delta_G(t)\}^2 \longrightarrow 0 \qquad (5.2.41)$$

as $N \longrightarrow \infty$. That is, δ_N is asymptotically optimal.

It may not be immediately obvious that, as claimed in section 2.3, the estimators for μ and μ_2' given earlier are essentially the same as those obtained by the method of moments. However, by recalling that

$$\begin{cases} E(T) = M\mu \\ E(T-M\mu)^2 = M(M-1)\mu_2' - M^2\mu^2 + M\mu \end{cases} \tag{5.2.42}$$

and defining

$$\bar{t} = \sum_{i=1}^{N+1} t_i \,/\, (N+1) \tag{5.2.43}$$

$$s_t^2 = \sum_{i=1}^{N+1} (t_i - \bar{t})^2 / (N+1)$$

and finally setting

$$\begin{cases} M\hat{\hat{\mu}} = \bar{t} \\ M(M-1)\hat{\hat{\mu}}_2' - M^2\hat{\hat{\mu}}^2 + M\hat{\hat{\mu}} = s_t^2 \end{cases} \tag{5.2.44}$$

we obtain

$$\begin{cases} \hat{\hat{\mu}}_2 = \bar{t}/M \\ \hat{\hat{\mu}}_2' = \dfrac{1}{M-1} \left[\dfrac{s_t^2}{M} + \dfrac{\bar{t}^2}{M} - \dfrac{\bar{t}}{M} \right] \end{cases} \tag{5.2.45}$$

or, equivalently,

$$\hat{\hat{\mu}} = \sum t_i / M(N+1) \tag{5.2.46}$$

$$\hat{\hat{\mu}}_2' = \sum t_i (t_i - 1) \,/\, (N+1)(M)(M-1)$$

as estimators of μ and μ_2', respectively. Hence the two methods are essentially the same.

After some manipulation we could also show the estimators for N' and R' to be:

$$\hat{\hat{N}}' = \frac{M(N+1)(M\Sigma t_i - \Sigma t_i^2)}{M(N+1)\left[\Sigma t_i(t_i-1)-(M-1)(\Sigma t_i)^2\right]} \tag{5.2.47}$$

$$\hat{\hat{R}}' = \frac{(\Sigma t_i)(\hat{\hat{N}}')}{M(N+1)}$$

where we note that $\Sigma t_i/M(N+1) \leq 1$ so that $\hat{\hat{R}}' \leq \hat{\hat{N}}'$. However, $\hat{\hat{N}}'$ as defined above must be modified so that $\hat{\hat{N}}' \geq 0$ just as we require $\hat{\mu}_2'$ to be such that $\hat{\mu}^2 \leq \hat{\mu}_2' \leq \hat{\mu}$.

It is interesting to note that when the Bayesian assumptions hold even a naive use of the past observations can result in a global risk smaller than that of δ_T. To see this, suppose we decide to use the estimator

$$\delta_N'(t_{N+1}) = \sum_{i=1}^{N+1} t_i/(N+1)(M) \tag{5.2.48}$$

as if the N+1 samples had all been chosen from the same distribution. That is, we just pool the N+1 samples into one sample of size M(N+1) and apply the usual estimator δ_T. The global risk for δ_N' is computed by first finding

$$\begin{aligned} R_N(\delta_N', G) &= E_{t,\lambda}(\delta_N'(t)-\lambda)^2 \\ &= E_\lambda\left\{\mathrm{Var}_{t|\lambda}(\delta_N'(t))+[E_{t|\lambda}\delta_N'(t)-\lambda]^2\right\} \\ &= \frac{\mu-\mu_2'}{M(N+1)^2} + \left(\frac{N}{N+1}\right)^2\left(\mu_2+\mu-\left[\frac{\overset{N}{\Sigma} t_i}{MN}\right]^2\right). \end{aligned} \tag{5.2.49}$$

Hence the global risk is:

$$R(\delta_N',G) = E_N R_N(\delta_N',G) \tag{5.2.50}$$

$$= \frac{\mu-\mu_2'}{M(N+1)^2} + \left[\frac{N}{N+1}\right]^2 \mu_2 + \left(\frac{N}{N+1}\right)^2 \frac{(M+N')\mu_2}{(MN)}$$

$$= \frac{R'(N'-R')}{M(N+1)^2(N')^2(N'+1)} [N'+MN^2+NM+NN']$$

$$= R(\delta_T,G) \left\{ \frac{N'+MN^2+NM+NN'}{[N+1]^2 N'} \right\}$$

$$= R(\delta_G,G) \left\{ \frac{(M+N')(N'+MN^2+NM+NN')}{N'M(N+1)^2} \right\} .$$

The estimator δ_N' is <u>not</u> asymptotically optimal since

$$\lim_{N\to\infty} R(\delta_N',G) = \frac{M+N'}{N'} R(\delta_G,G) > R(\delta_G,G) . \tag{5.2.51}$$

It should also be noted that δ_N' is not, in general, unbiased for λ_{N+1} since

$$E_N(\delta_N'(t_{N+1})) = \mu \tag{5.2.52}$$

although this is no real drawback since also

$$E(\delta_G(t_{N+1})) = \frac{\lambda_{N+1} \cdot M+R'}{M+N'} \neq \lambda_{N+1} .$$

However, even this simple-minded approach often yields a better estimator than δ_T. In fact we have $R(\delta_N',G) \leq R(\delta_T,G)$, that is, δ_N is better than δ_T if and only if

$$\frac{N'+MN^2+NM+NN'}{(N+1)^2N'} < 1 \qquad (5.2.54)$$

$$\iff N'+MN^2+NM+NN' < N^2N'+2NN'+N'$$

$$\iff M(N^2+N) < N'(N^2+N) \iff M < N' .$$

Hence, if the prior $Be(R',N'-R')$ is such that $N' > M$ then δ_N' is better than δ_T for all N regardless of how small N may be. See section 5.3 for some numerical comparison of the Bayes risks.

Next, suppose we have the additional information that the prior variance is μ_2. The class of all beta priors with variance μ_2 will be denoted by G'. In this case, one would be tempted to use a beta with mean $\Sigma t_i/(N+1)M$ and variance μ_2 as our estimate G_N of G. However, we must be careful since from (5.1.18) we have

$$\mu_2-\mu+\mu^2 \leq 0 \qquad (5.2.55)$$

or, equivalently,

$$\frac{1}{2} - \frac{\sqrt{1-4\mu_2}}{2} \leq \mu \leq \frac{1}{2} + \frac{\sqrt{1-4\mu_2}}{2} \quad . \qquad (5.2.56)$$

Now, since $\mu \in \left[\frac{1-\sqrt{1-4\mu_2}}{2}, \frac{1+\sqrt{1-4\mu_2}}{2}\right]$ then it is reasonable to require $\hat{\mu}$ to be in the same interval. Hence, our estimator $\hat{\mu}$ of the prior mean μ is defined by:

$$\hat{\mu} = \begin{cases} \dfrac{\Sigma t_i}{(N+1)(M)} & \text{if } \dfrac{\Sigma t_i}{(N+1)M} \in \left[\dfrac{1-\sqrt{1-4\mu_2}}{2}, \dfrac{1+\sqrt{1-4\mu_2}}{2}\right] \\[2ex] \dfrac{1+\sqrt{1-4\mu_2}}{2} & \text{if } \dfrac{\Sigma t_i}{(N+1)M} \geq \dfrac{1+\sqrt{1-4\mu_2}}{2} \\[2ex] \dfrac{1-\sqrt{1-4\mu_2}}{2} & \text{if } \dfrac{\Sigma t_i}{(N+1)M} \leq \dfrac{1-\sqrt{1-4\mu_2}}{2} \end{cases} \qquad (5.2.57)$$

Note that when using the above $\hat{\mu}$ the maximum possible difference between $\hat{\mu}$ and the true μ is:

$$\left| \frac{1}{2} + \frac{\sqrt{1-4\mu_2}}{2} - \left(\frac{1}{2} - \frac{\sqrt{1-4\mu_2}}{2}\right)\right| = \sqrt{1-4\mu_2} \quad . \qquad (5.2.58)$$

Knowledge that μ_2 is small usually gives one more confidence in the estimator $\hat{\mu}$ of μ. However, from (5.2.58) and figure 4 see that the interval $\left[\frac{1-\sqrt{1-4\mu_2}}{2}, \frac{1+\sqrt{1-4\mu_2}}{2}\right]$, from which $\hat{\mu}$ is chosen, has a short length for large μ_2. This is so since as $\mu_2 \to 1/4$, the maximum possible variance, we know that $G(\lambda)$ is approaching the distribution $G_0(\lambda)$ where

$$G_0(\lambda) = \begin{cases} 1/2 & \text{if } 0 \leq \lambda < 1 \\ 1 & \text{if } \lambda = 1 \end{cases} \qquad (5.2.59)$$

i.e., the distribution that assigns all probability to the points $\lambda=0$ and $\lambda=1$ with equal weights at 0 and 1. And in the case that μ_2 is known to be = 1/4 then of course we also know $\mu = 1/2$ and the Bayes risk is attainable.

Using the estimator G_N, i.e., a beta with mean defined by (5.2.57) and variance μ_2 we find an empirical Bayes estimator of

$$\delta''_N(t_{N+1}) = \frac{t_{N+1}\mu_2 + \hat{\mu}(\hat{\mu}-\hat{\mu}'_2)}{M\mu_2 + \hat{\mu} - \hat{\mu}'_2} \tag{5.2.60}$$

where $\hat{\mu}'_2 = \mu_2 + \hat{\mu}^2$

and the global risk is

$$R(\delta''_N, G) = R(\delta_G, G) + W(\delta''_N, G) \tag{5.2.61}$$

where

$$W(\delta''_N, G) = E\left\{\delta''_N(t) - \delta_G(t)\right\}^2 \tag{5.2.62}$$

where E represents expectation with respect to the random variables $T_1, \ldots, T_{N+1}$ each with p.m.f. $f_G(t)$.

A Monte Carlo evaluation of $W(\delta''_N, G)$ is given in section 5.3 for various values of N and different G. The fact that δ''_N is a.o. in $\mathcal{G}'$ can be shown in exactly the same manner used to show δ_N is a.o. in $\mathcal{G}$. See (5.2.36) to (5.2.41) for details.

Next assume μ is known but μ_2 is unknown. Denote this class of priors by $\mathcal{G}''$. In this case, our G_N is chosen to be a beta with mean μ and second moment: (assuming $M > 1$)

$$\hat{\mu}_2' = \begin{cases} \mu^2 & \text{if } \dfrac{1}{(N+1)} \dfrac{\Sigma t_i(t_i-1)}{M(M-1)} \le \mu^2 \\[2ex] \dfrac{\Sigma t_i(t_i-1)}{(N+1)m(M-1)} & \text{if } \mu^2 \le \dfrac{\Sigma t_i(t_i-1)}{(N+1)M(M-1)} \le \mu \\[2ex] \mu & \text{if } \dfrac{1}{N+1} \cdot \dfrac{\Sigma t_i(t_i-1)}{M(M-1)} \ge \mu \end{cases} \tag{5.2.63}$$

where $\hat{\mu}_2'$ is required to lie in the interval $[\mu^2,\mu]$ since $\mu_2' \varepsilon [\mu^2,\mu]$.

The empirical Bayes estimator in this case is then just

$$\delta_N''' = \begin{cases} \dfrac{t_{N+1} + \dfrac{\mu(\mu-\hat{\mu}_2')}{\hat{\mu}' - \mu^2}}{M + \dfrac{\mu-\hat{\mu}_2'}{\hat{\mu}_2'-\mu^2}} & \text{if } M > 1 \\[2ex] \mu & \text{if } M = 1 \end{cases} \tag{5.2.64}$$

If M=1 then we find that $R(\delta_N''',G) = \mu_2$ so that $\dfrac{R(\delta_G,G)}{R(\delta_N''',G)} = \dfrac{N'}{1+N'}$ and $\dfrac{R(\delta_T,G)}{R(\delta_N''',G)} = N'$. Note that by defining $\hat{\mu}_2'$ by (5.2.63) the maximum possible difference between $\hat{\mu}_2'$ and the true μ_2' is:

$$|\hat{\mu}_2'-\mu_2'| = \mu-\mu^2. \tag{5.2.65}$$

Hence if μ is either very small or very large then the estimator δ_N''' should have a value very close to that of the Bayes estimator δ_G.

Also, if $\frac{1}{N+1} \cdot \frac{\Sigma t_i(t_i-1)}{M(M-1)} \leq \mu^2$ then $\delta_N'''(t_{N+1}) =$ and if $\frac{1}{N+1} \cdot$

$\frac{\Sigma t_i(t_i-1)}{M(M-1)} \geq \mu$ then $\delta'''(t_{N+1}) = \frac{t_{N+}}{M} = \delta_T(t_{N+1})$.

The global risk of δ_N''' is:

$$R(\delta_N''',G) = R(\delta_G,G)+E\{\delta_N'''(t)-\delta_G(t)\}^2 \tag{5.2.66}$$

$$= R(\delta_G,G) + W(\delta_N''',G).$$

Again, section 5.3 contains Monte Carlo evaluations of $R(\delta_N''',G)$ for various small values of N and different prior G. Also, $\delta_N'''(t)$ can be shown to be asymptotically optimal by following the steps outlined in equations (5.2.36) to 5.2.41).

Just as in the normal case it is possible to obtain empirical Bayes confidence intervals on the parameter λ_{N+1}. By corollary 1, Lindley [21], p. 141, the posterior distribution of

$$\tilde{F} = \frac{N'-R'+M-t}{R'+t} \left(\frac{\tilde{\lambda}}{1-\tilde{\lambda}}\right) \tag{5.2.67}$$

is an F-distribution with parameters $2(R'+t)$ and $2(N'-R'+M-t)$. That is $\tilde{F} \sim F[2(R'+t), 2(N'-R'+M-t)]$. Hence an exact β Bayesian confidence interval for λ can be seen to be:

$$\frac{\underline{F}\ (R' + t)}{(N'-R'+M-t)+\underline{F}(R'+t)} < \lambda < \frac{\overline{F}\ (R' + t)}{(N'-R'+M-t)+\overline{F}(R'+t)} \tag{5.2.68}$$

where $\alpha=1-\beta$ and, using Lindley's notation,

$$\underline{F} = \underline{F}_{\alpha/2} \quad [2(R'+t),\ 2(N'-R'+M-t)]$$

$$\bar{F} = \bar{F}_{\alpha/2} \quad [2(R'+t),\ 2(N'-R'+M-t)]$$

or, since $\underline{F}_{\alpha}(x_1, x_2) = \left\{ \bar{F}_{\alpha}(x_2, x_1) \right\}^{-1}$, (5.2.68) can be written:

$$\frac{1}{\dfrac{\bar{F}'(N'-R'+M-t)}{(R'+t)} + 1} < \lambda < \frac{1}{\dfrac{N'-R'+M-t}{(R'+t)\bar{F}} + 1} \qquad (5.2.69)$$

where

$$\bar{F}' = \left\{ \underline{F}\ [2(R'+t),\ 2(N'-R'+M-t)] \right\}^{-1}$$

$$= \bar{F}\ [2(N'-R'+M-t),\ 2(R'+t)]$$

and an empirical Bayes confidence interval for λ can be formed by replacing R' and N' in (5.2.69) with their estimates $\hat{R}'$ and $\hat{N}'$. The confidence level of the EB confidence interval will not, in general, be β. However, the end points of the EB confidence interval will converge to the end points of (5.2.59) and hence the confidence interval is asymptotically optimal.

5.3 Numerical Results

In this section the results of several Monte Carlo experiments relating to the estimators described in the previous section are given. The main purposes of this study are:

(i) to closely approximate the global risk for each proposed estimator when an exact determination is difficult

(ii) to study the behavior of the estimators for small N

(iii) to compare the various proposed estimators for different priors and different values of M and N.

The general comments concerning methods used and other items given in section 3.3 are applicable here and will not be repeated.

In addition to the estimators considered in section 5.2, two additional ones are studied here. The first one, denoted by δ_N^D, is found using Deely and Kruse's method of estimating G by G_N as described in section 3.3.

The second one, denoted by δ_N^S, is called the "simple" empirical Bayes estimator because it makes no attempt to form an explicit estimate G_N of G. Rather, δ_N^S bypasses this point and attempts to estimate $\delta_G(t)$ directly. This estimator was first given by Robbins [33] and is defined by:

$$\delta_N^S(t_{N+1}) = \frac{t'_{N+1}+1}{M} \cdot \frac{P_{N+1,M}(t'_{N+1}+1)}{P_{N+1,M-1}(t'_{N+1})} \tag{5.3.1}$$

where

$$P_{N+1,M}(t) = \frac{\text{no. of terms of } t_1,\ldots,t_{N+1} = t}{N+1} \tag{5.3.2}$$

$$P_{N+1,M-1}(t) = \frac{\text{no. of terms of } t'_1,\ldots,t'_{N+1} = t}{N+1}$$

and t_i' is the number of successes in the first M-1 out of M trials in the i^{th} sample $(i=1,\ldots,N+1)$.

Robbins does not point out that his proposed estimator δ_N^S can be greater than 1. For example suppose, N=2, M=5 and

$$\begin{cases} t_1' = 3, \quad t_2' = 3, \quad t_3' = 2 \\ t_1 = 3, \quad t_2 = 3, \quad t_3 = 2 \end{cases} \tag{5.3.3}$$

then,

$$\begin{cases} P_{3,5}(t_3'+1) = P_{3,5}(3) = 2 \\ P_{3,4}(t_3') = P_{3,4}(2) = 1 \; . \end{cases} \tag{5.3.4}$$

Hence,

$$\delta_N^S(t_{N+1}) = \delta_N^S(2) = \frac{3}{5} \cdot \frac{2}{1} = \frac{6}{5} > 1 \quad . \tag{5.3.5}$$

That is, we are estimating the binomial parameter λ to be greater than 1. The obvious modification of δ_N^S to correct this defect (and the one used in this study) is to simply restrict the values of δ_N^S to the interval [0, 1].

The prior distributions considered are all beta distributions with parameters:

(i) $N' = 20$, $R' = 10$

(ii) $N' = 6$, $R' = 3$

(iii) $N' = 2$, $R' = 1$

(iv) $N' = 24$, $R' = 6$

(v) $N' = 12,\ R' = 3$

(vi) $N' = 4,\ R' = 1$

The estimators under consideration are:

(i) δ_G - the Bayes estimator with $R(\delta_G, G) = \dfrac{R'(N'-R')}{(M+N')N'(N'+1)}$

(ii) δ_T - the ordinary mean with $R(\delta_T, G) = \dfrac{R'(N'-R')}{MN'(N'+1)}$

(iii) δ_0 - the minimax estimator; $R(\delta_0, G) = \dfrac{1}{4(\sqrt{M}+1)^2}$

(iv) δ_N - empirical Bayes estimator, G assumed conjugate

(v) δ_N''' - empirical Bayes estimator, mean known

(vi) δ_N'' - empirical Bayes estimator, variance known

(vii) δ_N^S - Robbins simple EB estimator

(viii) δ_N' - a non-a.o. EB estimator (pooled mean)

$$R(\delta_N', G) = \frac{R'(N'-R)(N'+MN^2+NM+NN')}{M(N+1)^2(N')^2(N'+1)}$$

(ix) δ_D^D - empirical Bayes estimator obtained by using Deely and Kruse's method (again, N1=3 in this study).

The following tables (tables 5-10) present the main results on risks for the different prior distributions and for N=1, 10, 20 and M=1, 5, 10. Again, the numbers in parentheses are standard errors and if these numbers are absent then the risk given is exact. Also note the numbers given are actually 10 times the risk for ease of comparison. The tables also give $R(\delta) = R(\delta_G, G)/R(\delta, G)$ as defined in section 3.3.

The figures (3-8) following the tables show $R(\delta)$ as a function of N and allow us simultaneously to compare two estimators for a small N and to see the rate of convergence of $R(\delta_N,G)$ to $R(\delta_G,G)$, (assuming the estimator is asymptotically optimal). Just as in chapter III, however, the figures will not contain the curves of $R(\delta_N'')$ and $R(\delta_N''')$ since, as is obvious from the tables, they lie above $R(\delta_N)$ and they are not applicable unless we assume the prior mean or variance to be known.

Some points to keep in mind when reading the tables and figures are:

(i) $\delta_N = \delta_N' = \delta_N^S$ for $M = 1$.

This implies that δ_N is not asymptotically optimal in this case but this is not surprising since it was shown earlier that no asymptotically optimal empirical Bayes estimator exists in this situation ($M=1$).

(ii) δ_0 is Bayes with respect to a Be $(\sqrt{M}/2, \sqrt{M}/2)$. Hence, for certain $M, R(\delta_0)$ may be very close to 1 although it should be remembered that it is also possible for $R(\delta_0)$ to be quite small.

(iii) $\delta_N''' = \mu$ for $M = 1$. This fact is implicit in the definition of δ_N''' but is pointed out here since $\delta_N''' = \mu$ for all t implies $R(\delta_N''', G) = \mu_2$.

(iv) $\lim_{N \to \infty} R(\delta_N') = \dfrac{N'}{M+N'}$

(v) δ'_N is preferred over δ_T if $M \leq N'$.

(vi) δ_0 is preferred over δ_T if $\mu-\mu_2' \leq \frac{M}{4(\sqrt{M}+1)^2}$.

An obvious, but nonetheless important fact concerning these studies is that $R(\delta_N,G) \leq R(\delta_T,G)$ for all N and for every prior distribution considered. As a matter of fact, by (v) above, even δ'_N is better than δ_T for small M although as M increases this advantage disappears.

For any given empirical Bayes estimator δ and a given prior G it is interesting to know the value N_0 where N_0 is defined to be the smallest integer N such that

$$R(\delta,G) \leq R(\delta_T,G) \tag{5.3.6}$$

or equivalently, the smallest N such that

$$R(\delta) \geq R(\delta_T). \tag{5.3.7}$$

The number N_0 is then the number of prior observations necessary for us to "break-even" using δ instead of the usual δ_T. If N_0 is very large for a given decision function δ, the fact that δ is asymptotically optimal may mean very little in practice.

Approximations to the value of N_0 were obtained from this study by simply reading the point of intersection of the curves $R(\delta)$ and $R(\delta_T)$ from the following figures. The results are given in table 9.

TABLE 3

GLOBAL RISKS* (N'=20, R'=10)

Estimator	N=1			N=10			N=20		
	M=1	M=5	M=10	M=1	M=5	M=10	M=1	M=5	M=10
1) Bayes	.113	.095	.079	.113	.095	.079	.113	.095	.079
2) Usual (mean)	2.381	.476	.238	2.381	.476	.238	2.381	.476	.238
3) Minimax	.625	.239	.144	.625	.239	.144	.625	.239	.144
4) G conjugate	1.250	.297	.169	.325	.150	.112	.227	.129	.105
5) G conjugate, mean known	.119	.236 (.048)	.151 (.040)	.119	.138 (.029)	.108 (.028)	.119	.118 (.031)	.093 (.029)
6) G conjugate, variance known	.674 (.087)	.268 (.045)	.158 (.037)	.312 (.034)	.146 (.033)	.106 (.027)	.205 (.032)	.121 (.029)	.097 (.028)
7) Pooled mean	1.250	.297	.178	.324	.151	.129	.226	.136	.124
8) Robbins' simple EB	1.250	1.338 (.101)	.954 (.098)	.324	.653 (.072)	.277 (.039)	.226	.423 (.044)	.239 (.032)
9) Deely and Kruse	1.307 (.111)	.588 (.049)	.218 (.027)	.629 (.052)	.296 (.038)	.193 (.028)	.413 (.038)	.183 (.035)	.177 (.021)

* Entries are actually 10 times risk for ease of comparison

TABLE 4

GLOBAL RISKS (N' = 6, R' = 3)

Estimator	N=1			N=10			N=20		
	M=1	M=5	M=10	M=1	M=5	M=10	M=1	M=5	M=10
1) Bayes	.306	.199	.133	.306	.194	.133	.306	.194	.133
2) Usual (mean)	2.142	.428	.214	2.142	.428	.214	2.142	.428	.214
3) Minimax	.625	.238	.144	.625	.238	.144	.625	.238	.144
4) G conjugate	1.250	.389 (.031)	.200 (.031)	.519	.314 (.043)	.189 (.042)	.442	.281 (.037)	.172 (.036)
5) G conjugate, mean known	.357	.329 (.029)	.186 (.030)	.357	.306 (.031)	.174 (.028)	.357	.274 (.036)	.161 (.035)
6) G conjugate, variance known	.675 (.089)	.278 (.024)	.179 (.020)	.487 (.046)	.279 (.035)	.177 (.031)	.429 (.037)	.260 (.034)	.160 (.036)
7) Pooled mean	1.250	.392	.285	.519	.363	.344	.442	.360	.350
8) Robbins' simple EB	1.250	1.215 (.100)	.906 (.085)	.519	.762 (.095)	.343 (.034)	.442	.558 (.076)	.276 (.029)
9) Deely and Kruse	1.329 (.101)	.590 (.051)	.247 (.022)	.657 (.051)	.322 (.042)	.207 (.025)	.459 (.041)	.296 (.043)	.193 (.017)

TABLE 5

GLOBAL RISKS (N' = 2, R' = 1)

Estimator	N+1			N=10			N=20		
	M=1	M=5	M=10	M=1	M=5	M=10	M=1	M=5	M=10
1) Bayes	.555	.238	.138	.555	.238	.138	.555	.238	.138
2) Usual (mean)	1.666	.333	.166	1.666	.333	.166	1.66	.333	.166
3) Minimax	.625	.238	.144	.625	.238	.144	.625	.238	.144
4) G conjugate	1.250	.323 (.042)	.165 (.032)	.909	.290 (.039)	.158 (.035)	.873	.259 (.038)	.150 (.032)
5) G conjugate, mean known	.833	.308 (.050)	.164 (.032)	.833	.287 (.038)	.155 (.029)	.833	.273 (.041)	.145 (.029)
6) G conjugate, variance known	1.123 (.108)	.254 (.033)	.158 (.028)	.846 (.073)	.243 (.032)	.153 (.027)	.810 (.058)	.244 (.038)	.143 (.021)
7) Pooled mean	1.250	.583	.500	.909	.787	.772	.873	.809	.801
8) Robbins' simple EB	1.250	1.041 (.133)	.926 (.079)	.909	.761 (.113)	.356 (.042)	.873	.659 (.100)	.300 (.033)
9) Deely and Kruse	1.357 (.112)	.609 (.052)	.273 (.021)	.713 (.068)	.282 (.052)	.218 (.011)	.484 (.030)	.336 (.047)	.205 (.011)

TABLE 6

GLOBAL RISKS (N' = 24, R' = 6)

Estimator	N=1			N=10			N=20		
	M=1	M=5	M=10	M=1	M=5	M=10	M=1	M=5	M=10
1) Bayes	.072	.062	.052	.072	.062	.052	.072	.062	.052
2) Usual (mean)	1.800	.360	.180	1.800	.360	.180	1.800	.360	.180
3) Minimax	.625	.238	.144	.625	.238	.144	.625	.238	.144
4) G conjugate	.937	.210 (.027)	.105 (.018)	.231	.100 (.013)	.064 (.007)	.157	.088 (.011)	.062 (.015)
5) G conjugate, mean known	.075	.138 (.024)	.102 (.016)	.075	.094 (.081)	.063 (.011)	.075	.084 (.071)	.060 (.014)
6) G conjugate, variance known	.508 (.070)	.204 (.026)	.103 (.017)	.201 (.024)	.079 (.009)	.056 (.007)	.151 (.025)	.063 (.008)	.054 (.012)
7) Pooled mean	.937	.217	.127	.231	.100	.084	.157	.088	.080
8) Robbins' simple EB	.937	.586 (.077)	.516 (.052)	.231	.333 (.073)	.332 (.043)	.157	.205 (.034)	.307 (.031)
9) Deely and Kruse	.958 (.070)	.281 (.025)	.142 (018)	.303 (.023)	.187 (.032)	.138 (.013)	.219 (.026)	.126 (.018)	.120 (.013)

TABLE 7

GLOBAL RISKS (N' = 12, R' = 3)

Estimator	N=1			N=10			N=20		
	M=1	M=5	M=10	M=1	M=5	M=10	M=1	M=5	M=10
1) Bayes	.133	.101	.078	.133	.101	.078	.133	.101	.078
2) Ususal (mean)	1.730	.346	.173	1.730	.346	.173	1.730	.346	.173
3) Minimax	.625	.238	.144	.625	.238	.144	.625	.238	.144
4) G conjugate	.937	.209 (.028)	.154 (.026)	.288	.151 (.025)	.109 (.017)	.219	.139 (.728)	.104 (.016)
5) G conjugate, mean known	.144	.198 (.026)	.144 (.022)	.144	.147 (.019)	.101 (.018)	.144	.136 (.023)	.092 (.012)
6) G conjuguate, variance known	.683 (.063)	.193 (.025)	.143 (.022)	.294 (.038)	.148 (.023)	.093 (.019)	.202 (.021)	.129 (.024)	.089 (.081)
7) Pooled mean	.937	.245	.158	.288	.162	.146	.219	.153	.145
8) Robbins' simple EB	.937	.903 (.085)	.847 (.083)	.288	.604 (.056)	.251 (.028)	.219	.404 (.030)	.214 (.016)
9) Deely and Kruse	1.011 (.088)	.321 (.025)	.162 (.017)	.351 (.023)	.305 (.031)	.159 (.013)	.281 (.027)	.200 (.021)	.133 (.011)

TABLE 8

GLOBAL RISKS (N' = 4, R' = 1)

	N=1			N=10			N=20		
	M=1	M=5	M=10	M=1	M=5	M=10	M=1	M=5	M=10
1) Bayes	.300	.166	.107	.300	.166	.107	.300	.166	.107
2) Usual (mean)	1.500	.300	.150	1.500	.300	.150	1.500	.300	.150
3) Minimax	.625	.238	.144	.625	.238	.144	.625	.238	.144
4) G conjugate	.937	.257 (.024)	.148 (.024)	.477	.214 (.020)	.139 (.021)	.428	.178 (.029)	.120 (.020)
5) G conjugate, mean known	.652	.237 (.023)	.141 (.021)	.456	.204 (.025)	.131 (.019)	.412	.168 (.023)	.118 (.013)
6) G conjugate, variance known	.652 (.080)	.237 (.019)	.141 (.022)	.456 (.045)	.204 (.021)	.131 (.021)	.412 (.038)	.168 (.019)	.118 (.010)
7) Pooled mean	.937	.337	.262	.477	.368	.354	.428	.371	.364
8) Robbins' simple EB	.937	.909 (.080)	.856 (.077)	.477	.721 (.062)	.310 (.053)	.428	.502 (.048)	.249 (.020)
9) Deely and Kruse	1.084 (.076)	.349 (.031)	.171 (.019)	.377 (.027)	.278 (.037)	.167 (.012)	.310 (.021)	.237 (.030)	.152 (.012)

TABLE 9

VALUES OF N (M = 5)

Prior	Estimator			
	Empirical Bayes	Pooled mean*	Robbins	Deely & Kruse
1) N'+20, R'=10	1	1	16	3
2) N'=6, R'=3	1	1	35	5
3) N'=2, R'=1	1		250	7
4) N'=24, R'=6	1	1	9	1
5) N'=12, R'=3	1	1	28	1
6) N'=4, R'=1	1		165	8

* Recall that the pooled mean is better than the usual estimator if and only if M N'.

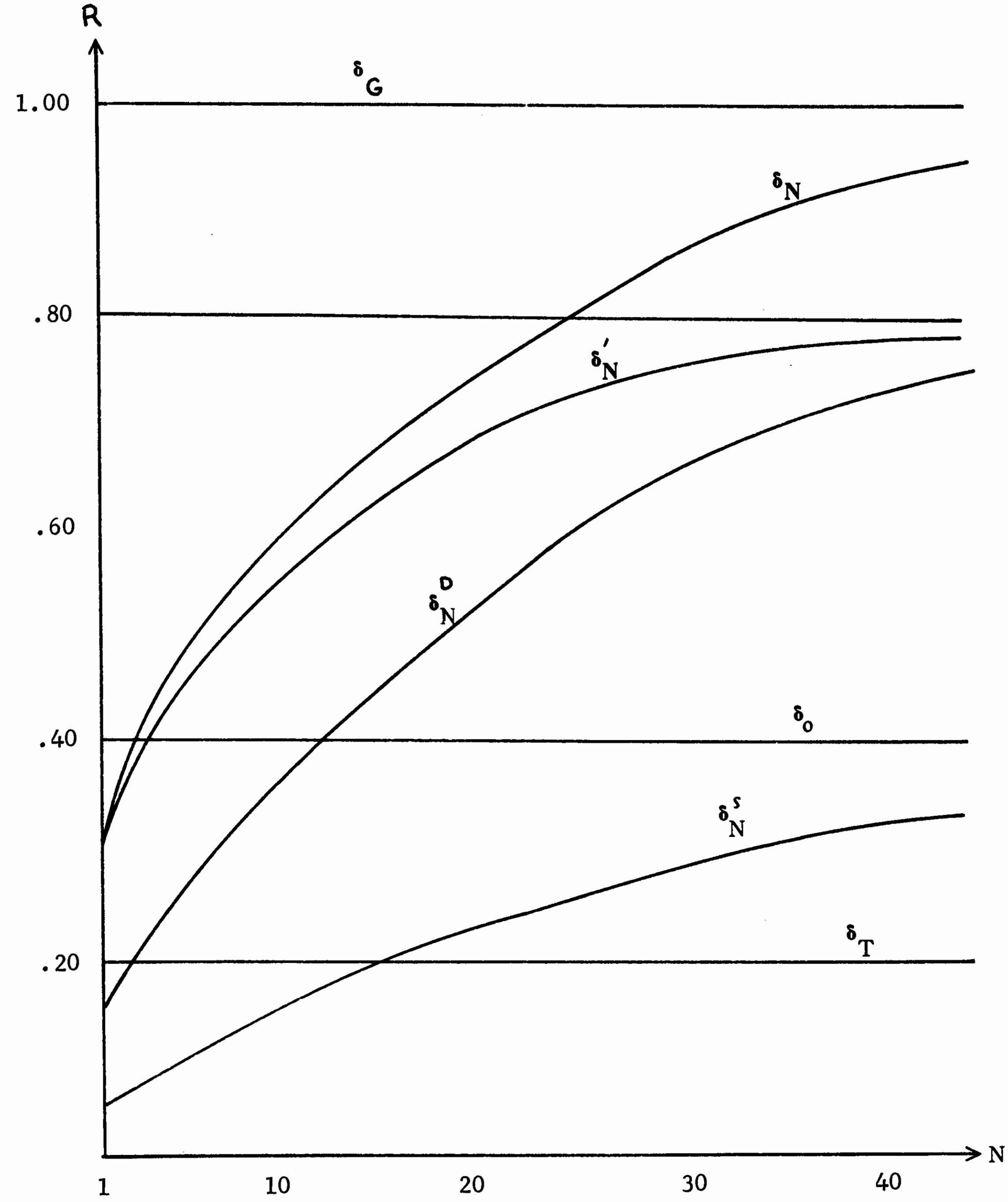

Figure 5. R(δ) as a function of N(N'=20, R'=10, M=5).

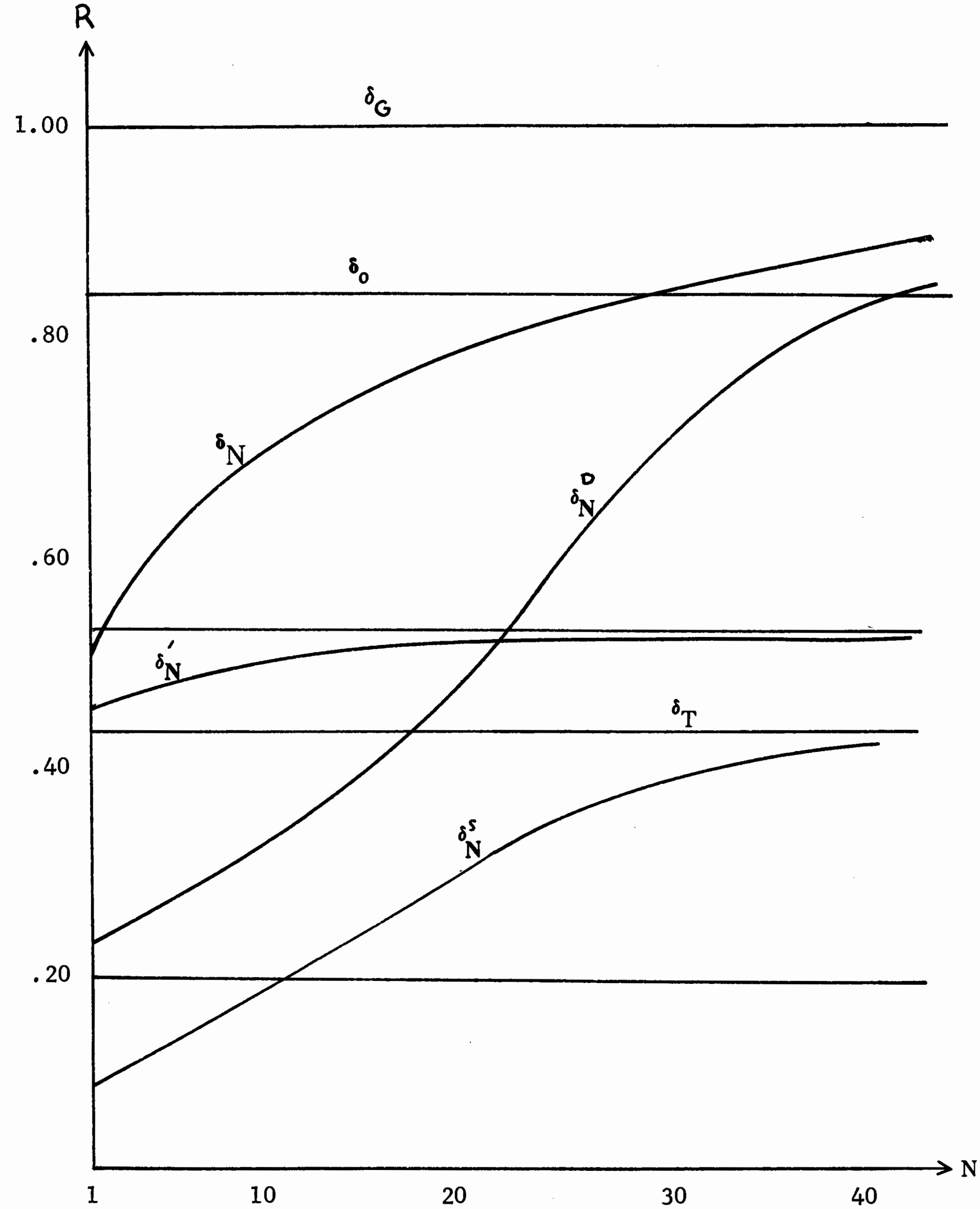

Figure 6. R(δ) as a function of N (N'=6, R'=3, M=5).

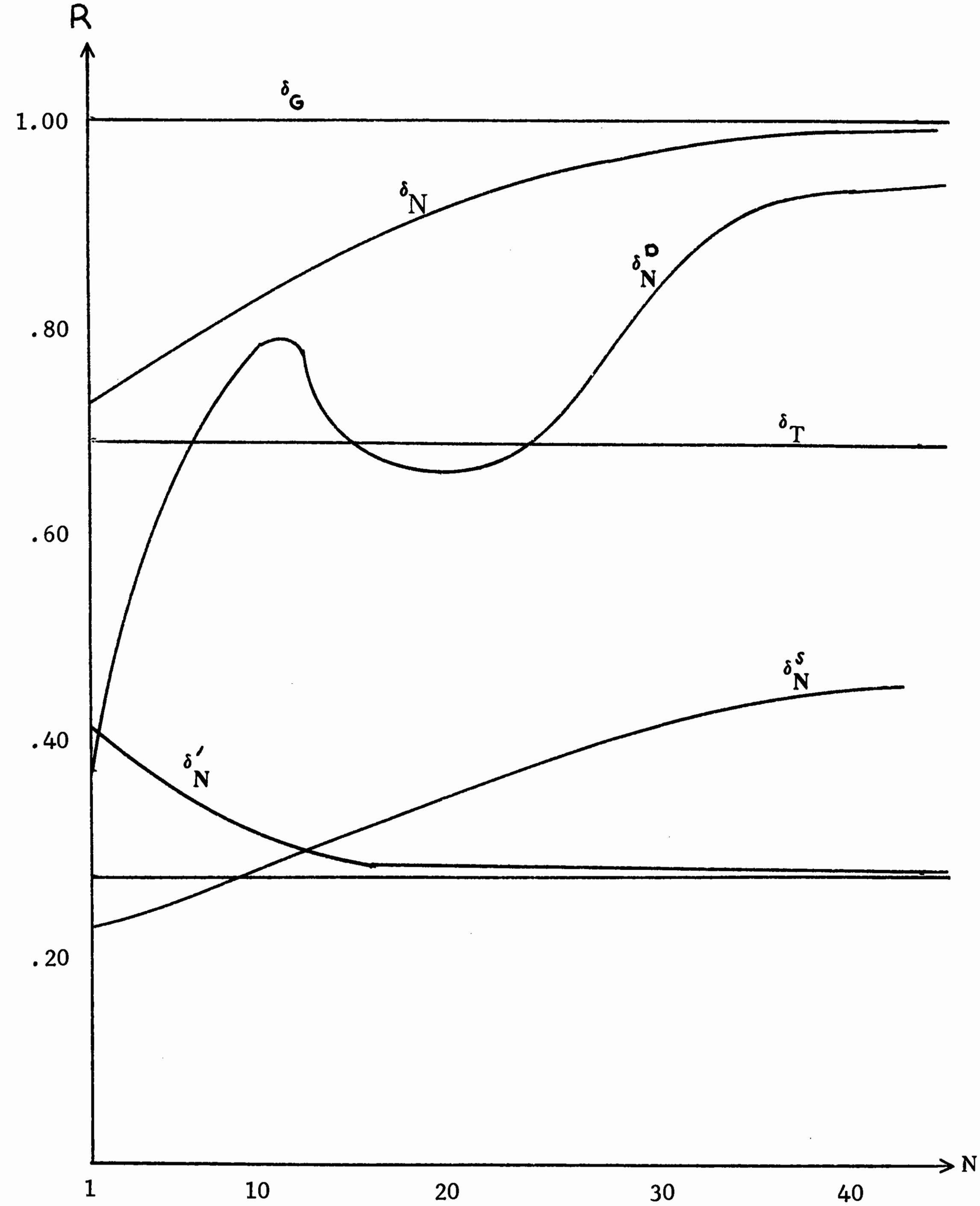

Figure 7. R(δ) as a function of N(N'=2, R'=1, M=5).

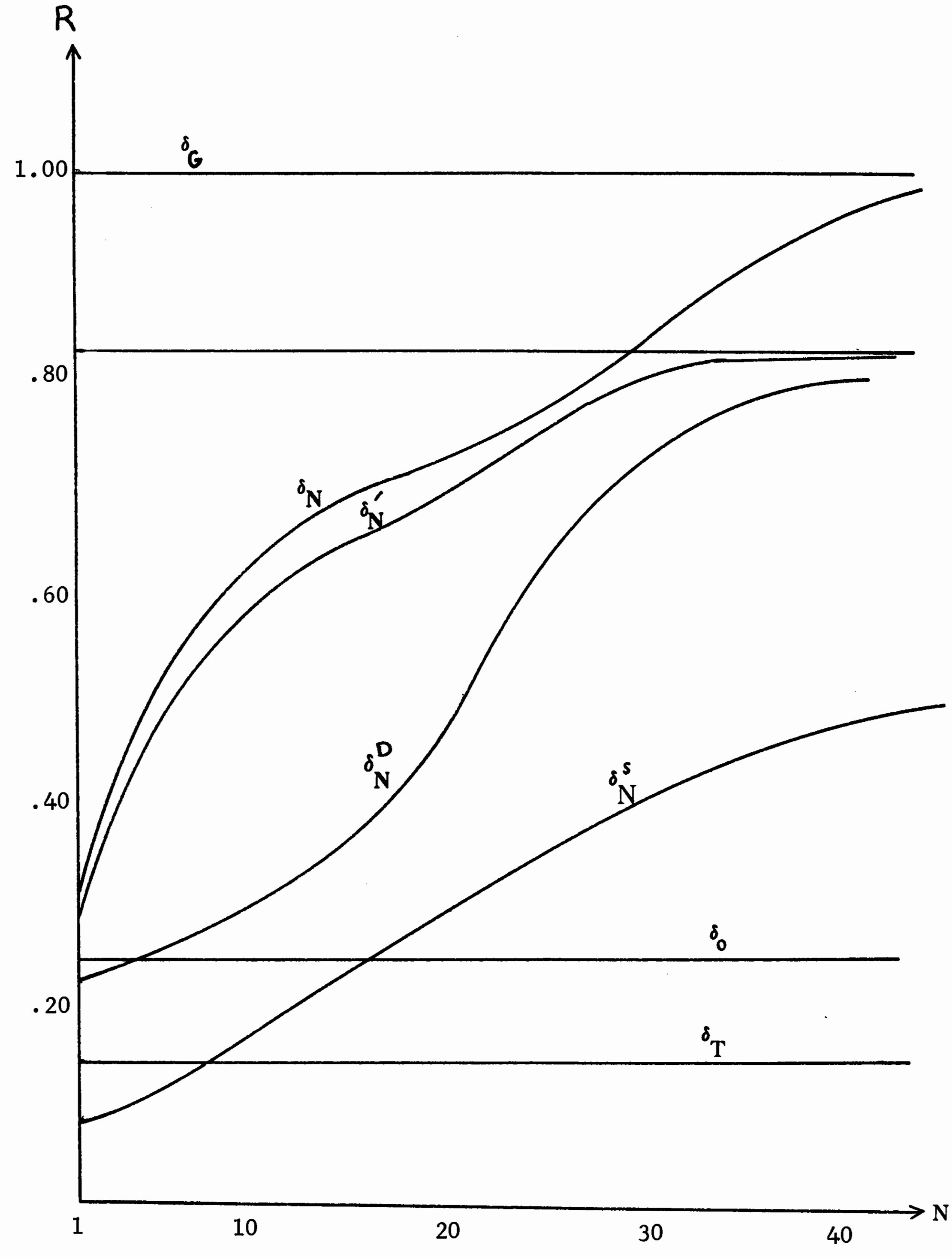

Figure 8. R(δ) as a function of N(N'=24, R'=6, M=5).

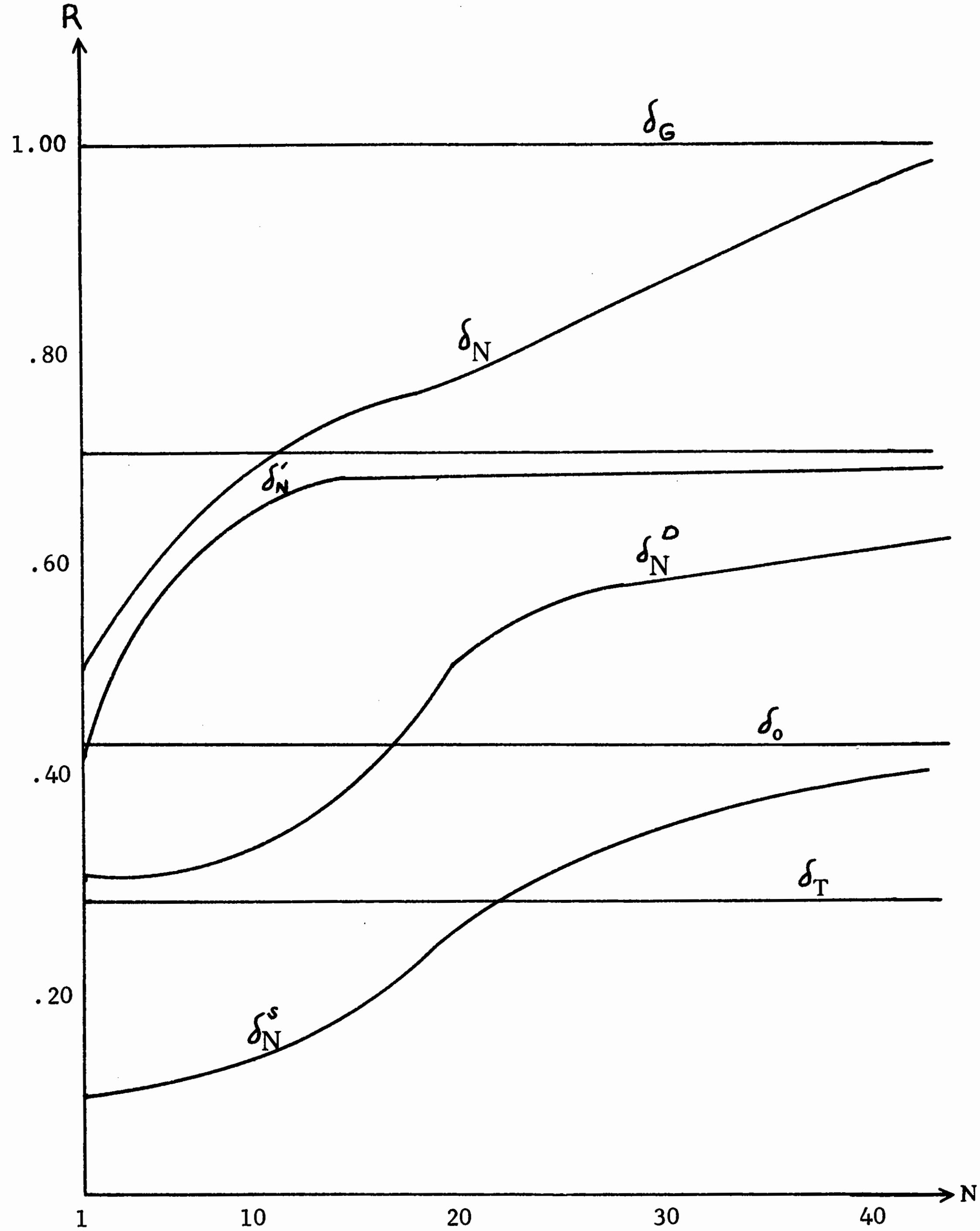

Figure 9. R (δ) as a function of N(N'=12, R'=3, M=5).

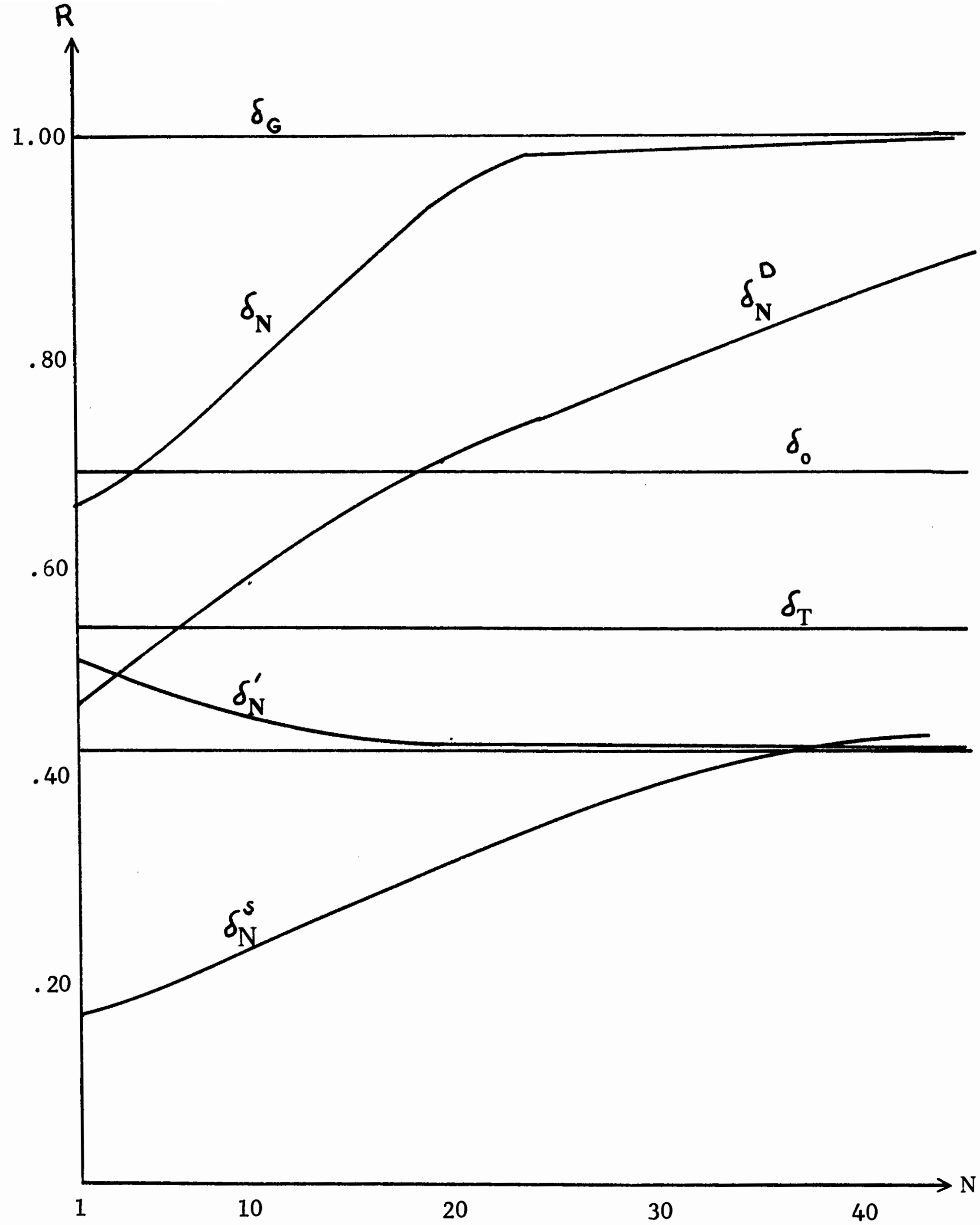

Figure 10. R (δ) as a function of N(N'=4, R'=1, M=5).

CHAPTER VI

BERNOULLI PROCESS II: $\mathcal{G}$-MINIMAX DECISION FUNCTIONS

6.1 Introduction

As in Chapter IV we consider here the case that G is known only to be a member of some class $\mathcal{G}$ of distribution functions on Λ and no prior observations on $f_G(x)$ exist. The applicable theory is developed in section 2.4. The classes $\mathcal{G}$ to be considered are:

(i) The class of all probability distributions functions on Λ.

(ii) $\mathcal{G}$: the conjugate class.

(iii) $\mathcal{G}'$: the conjugate class, μ known.

(iv) $\mathcal{G}''$: μ known.

(v) $\mathcal{G}'''$: μ and μ_2 known.

In each case $\mathcal{G}$-minimax estimators of λ are found and all are shown to be admissible. In addition, it is shown below that cases (i) and (ii) and cases (iii) and (iv) yield identical estimators so that only three distinct estimators are found in the five cases.

Asymptotic $\mathcal{G}$-minimax estimators are given in section 6.4.

The first case to be considered, $\mathcal{G}$ unrestrictive, is given in this introductory section because of its simplicity. When $\mathcal{G}$ is the class of all probability distribution functions on Λ then obviously

$\mathcal{G}$ contains all degenerate distributions and hence the $\mathcal{G}$-minimax estimator is just the ordinary minimax estimator

$$\frac{t_{N+1}+\sqrt{M}/2}{M + \sqrt{M}} \tag{6.1.1}$$

This particular estimator is discussed in greater detail in the next section.

6.2 Estimation of λ_{N+1}: $\mathcal{G}$ Conjugate

Suppose it is known only that G is a member of the conjugate family of distributions on $\Lambda = [0,1]$. That is, G is a distribution function in the beta class with parameters R' and N'-R'. Denote the class of all beta distributions by $\mathcal{G}$.

For any given $\lambda_0 \varepsilon [0,1]$, a $Be(R_0', N_0' - R_0')$ distribution with $R_0' = \lambda_0 N_0'$ will have a mean of λ_0 and variance (by (5.1.10)) of $\frac{\lambda_0(1-\lambda_0)}{N_0'+1}$. And if we require $R' = \lambda_0 N'$ as $N \to \infty$ then, in the limit, this is the degenerate distribution at λ_0. Hence, since $\mathcal{G}$ contains all degenerate distributions, the $\mathcal{G}$-minimax estimator in this case is just the ordinary minimax estimator

$$\delta_0(t_{N+1}) = \frac{t_{N+1}+\sqrt{M}/2}{M + \sqrt{M}} = \frac{t_{N+1}}{\sqrt{M}\ (1+\sqrt{M})} + \frac{1}{2(1+\sqrt{M})} \tag{6.2.1}$$

as described earlier in (5.2.16) and (6.1.1).

Since the risk (see (5.2.17)) of δ_0 is constant for all $G \varepsilon \mathcal{G}$ the conclusion that δ_0 is $\mathcal{G}$-minimax also follows from theorem 5. And, since δ_0 is Bayes with respect to a $Be(\sqrt{M}/2, \sqrt{M}/2)$, denoted by G_0, and

$$R(\delta_0, G_0) = \frac{1}{4(\sqrt{M}+1)^2} = R(\delta_0, G) \quad \forall G \varepsilon \mathcal{G} \tag{6.2.2}$$

then it also follows from theorem 3 that δ_0 is $\mathcal{G}$-minimax.

Further, since the estimator obtained when G is conjugate is the same as that obtained with no restrictions on G, then this estimator is obviously both $\mathcal{G}$-admissible and admissible.

It should be noted that

$$\sup_{G \varepsilon \mathcal{G}} R(\delta, G) \geq \frac{1}{4(\sqrt{M}+1)^2} \tag{6.2.3}$$

since δ_0 is $\mathcal{G}$-minimax and from (5.2.23)

$$R(\delta_G, G) \leq R(\delta_0, G) \quad \forall G \varepsilon \mathcal{G} . \tag{6.2.4}$$

That is, the Bayes risk of δ_0, namely (6.2.2), is the maximum risk for a Bayes d.f. with respect to any $G \varepsilon \mathcal{G}$. Recall (see 5.2.19) that δ_0 is better than δ_T for prior distributions with large variances and for samples with a small number (M) of observations. A rewriting of (5.2.19) allows us to state that δ_0 is better than δ_T if

$$\frac{1}{\left(1 + \frac{1}{\sqrt{M}}\right)^2} \leq 4(\mu - \mu_2') \tag{6.2.5}$$

For example, if $\mu-\mu_2' = \frac{1}{4}$ then δ_0 is better than δ_T iff

$$\frac{1}{\left(1 + \frac{1}{\sqrt{M}}\right)^2} \leq 1 \tag{6.2.6}$$

which is true for all M. If $\mu-\mu_2' = \frac{1}{8}$ then δ_0 is better than δ_T if and only if:

$$\frac{1}{\left(1 + \frac{1}{\sqrt{M}}\right)^2} \leq \frac{1}{2} \quad \text{i.e.,} \quad M \leq 5. \tag{6.2.7}$$

In general, if

$$\mu-\mu_2' = \alpha \tag{6.2.8}$$

$$\text{where} \quad 0 \leq \alpha \leq \frac{1}{4}$$

then δ_0 is preferred to δ_T iff:

$$\frac{1}{\left(1 + \frac{1}{\sqrt{M}}\right)^2} \leq 4\alpha \tag{6.2.9}$$

or, equivalently, iff

$$M \leq \left(\frac{8\alpha}{1-4\alpha}\right) \sqrt{M} + \frac{4\alpha}{1-4\alpha} \quad . \tag{6.2.10}$$

In terms of α, δ_0 is preferred to δ_T if and only if:

$$\alpha \geq \frac{M}{4M + 8\sqrt{M} + 4} = \frac{M}{4\left[1 + \frac{2}{\sqrt{M}} + \frac{1}{M}\right]} \tag{6.2.11}$$

Figure 11 graphically illustrates the preference regions.

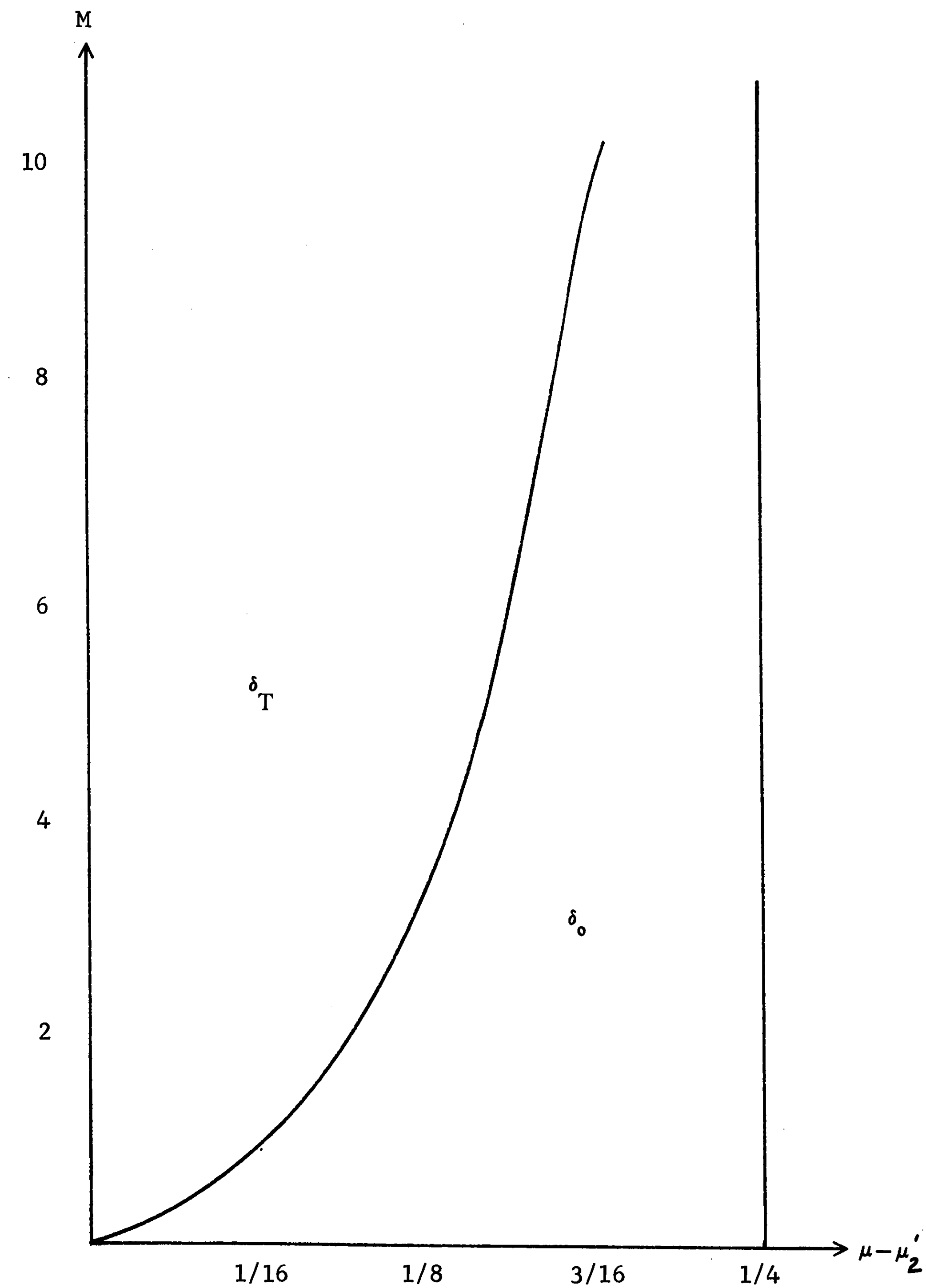

Figure 11. Preference regions for δ_o and δ_T

Of course, M can take only integral values but for purposes of illustration we have considered (6.2.11) for all real $M \geq 0$.

Next, suppose that the mean, say μ_0, is also known and let G' denote the class of all conjugate prior distributions with mean μ_0. That is, G' is the class of all $Be(\mu_0 N', (1-\mu_0)N')$ distributions.

Although G' neither contains all degenerate distributions nor satisfies the sufficient condition for admissibility given after (2.4.13), it is shown below that G'-admissibility implies admissibility.

The Bayes estimator for any $G \varepsilon G'$ is of the form (see(5.2.7))

$$\delta_G(t_{N+1}) = \frac{t_{N+1} + \mu_0 \alpha}{M+\alpha} \tag{6.2.12}$$

where $\alpha = N'$. For a given α, and any $G' \varepsilon G'$, not necessarily G, we have

$$R(\delta_G, G') = E_\lambda\left\{Var(\delta_G(t))\right\} + E_\lambda\left\{\frac{\lambda M + \mu_0\alpha}{M+\alpha} - \lambda\right\}^2 \tag{6.2.13}$$

$$= \frac{M\mu_0 - M\mu_2 - M\mu_0^2 + \alpha^2\mu_2}{(M+\alpha)^2}$$

and in order that the above Bayes risk be independent of the particular G' in effect we must have

$$\alpha^2 - M = 0 \qquad \text{i.e., } \alpha = \sqrt{M}. \tag{6.2.14}$$

The estimator

$$\delta_1(t_{N+1}) = \frac{t_{N+1} + \sqrt{M}\mu_0}{M+\sqrt{M}} \tag{6.2.15}$$

is Bayes with respect to $Be(\mu_0\sqrt{M},\sqrt{M}(1-\mu_0))$, say G_1, and has constant Bayes risk

$$R(\delta_1,G) = \frac{M\mu_0 - M\mu_0^2}{(M+\sqrt{M})^2} \tag{6.2.16}$$

$$= \frac{\mu_0-\mu_0^2}{(\sqrt{M}+1)^2}$$

over the class $\mathcal{G}'$. Hence, by theorem 5, δ_1 is $\mathcal{G}'$-minimax and G_1 is least favorable in $\mathcal{G}'$. It is obvious that

$$\inf_{\delta} R(\delta,G_1) = R(\delta_1,G_1) = \sup_{G\varepsilon\mathcal{G}'} \inf_{\delta} R(\delta,G) < \sup_{G\varepsilon\mathcal{G}'} R(\delta,G) \tag{6.2.17}$$

i.e., that

$$\sup_{G\varepsilon\mathcal{G}'} R(\delta,G) > \frac{\mu_0 - \mu_0^2}{(\sqrt{M}+1)^2} \tag{6.2.18}$$

$$\text{for every } \delta \neq \delta_1 \quad .$$

As a comparison between δ_1 and δ_T, consider the ratio

$$\frac{R(\delta_1,G)}{R(\delta_T,G)} = \frac{\mu_0 - \mu_0^2}{(\sqrt{M}+1)^2} \cdot \frac{M}{\mu_0 - \mu_2'} \quad . \tag{6.2.19}$$

This ratio is ≤ 1(i.e., δ_1 is better than δ_T) iff

$$\frac{\mu_0-\mu_0^2}{(\sqrt{M}+1)^2} \cdot \frac{M}{\mu_0-\mu_2'} \leq 1 \tag{6.2.20}$$

$$<=> \mu_2' \leq \mu_0-\mu_0\left[\frac{M}{\sqrt{M}+1}\right]^2 + \mu_0^2\left[\frac{\sqrt{M}}{\sqrt{M}+1}\right]^2$$

$$<=> \mu_2 \leq \mu_0-\mu_0\left[\frac{\sqrt{M}}{\sqrt{M}+1}\right]^2 + \mu_0^2\left[\left\{\frac{\sqrt{M}}{\sqrt{M}+1}\right\}^2 -1\right]$$

$$<=> \mu_2 \leq \left\{\frac{2\sqrt{M}+1}{(\sqrt{M}+1)}\right\}\mu_0\ (1-\mu_0) .$$

But from (5.1.20) we have

$$\mu_2 = \frac{\mu_0(1-\mu_0)}{N'+1} \qquad \text{for all } G\varepsilon\mathcal{G}' \tag{6.2.21}$$

so that the inequality (6.2.20) becomes

$$\frac{1}{N'+1} \leq \frac{2\sqrt{M}+1}{(\sqrt{M}+1)^2} \tag{6.2.22}$$

that is,

$$N' \geq \frac{(\sqrt{M}+1)^2}{2\sqrt{M}+1} - 1 \tag{6.2.23}$$

$$= \frac{M}{2\sqrt{M}+1}$$

Hence, even for reasonably large M the required N' (required to make δ_1

better than δ_T) is small. For example, even if M=100, δ_1 is better than δ_T if N' is such that $N' \geq \frac{100}{2(10)+1} = 4.76$. Figure 12 shows the preference regions for δ_1 and δ_T for any given pair (N',M). Again, even though M is of necessity an integer it is treated here as any non-negative real number.

6.3 Estimation Of λ_{N+1}: Moments Known

Suppose that it is known only that μ_0 is the mean of the unknown G. Let $\mathcal{G}''$ represent the class of all prior distributions on [0,1] with mean μ_0.

It is interesting to note that the estimator δ_1 defined by (6.2.15) not only has constant risk over the class $\mathcal{G}'$ but also over the class $\mathcal{G}''$. To see this, we write, for all $G \in \mathcal{G}''$,

$$R(\delta_1,G) = E_\lambda E_{t/\lambda}\left\{\frac{t_{N+1}+\sqrt{M}\mu_0}{M+\sqrt{M}} - \lambda\right\}^2 \tag{6.3.1}$$

$$= E_\lambda\left\{\frac{\lambda M(1-\lambda)}{(M+\sqrt{M})^2}\right\} + E_\lambda\left\{\frac{\sqrt{M}\mu_0-\lambda\sqrt{M}}{(M+\sqrt{M})}\right\}^2$$

$$= \frac{M(\mu_0-\mu_2') + M\mu_2}{(M+\sqrt{M})^2}$$

$$= \frac{\mu_0-\mu_0^2}{(\sqrt{M}+1)^2} .$$

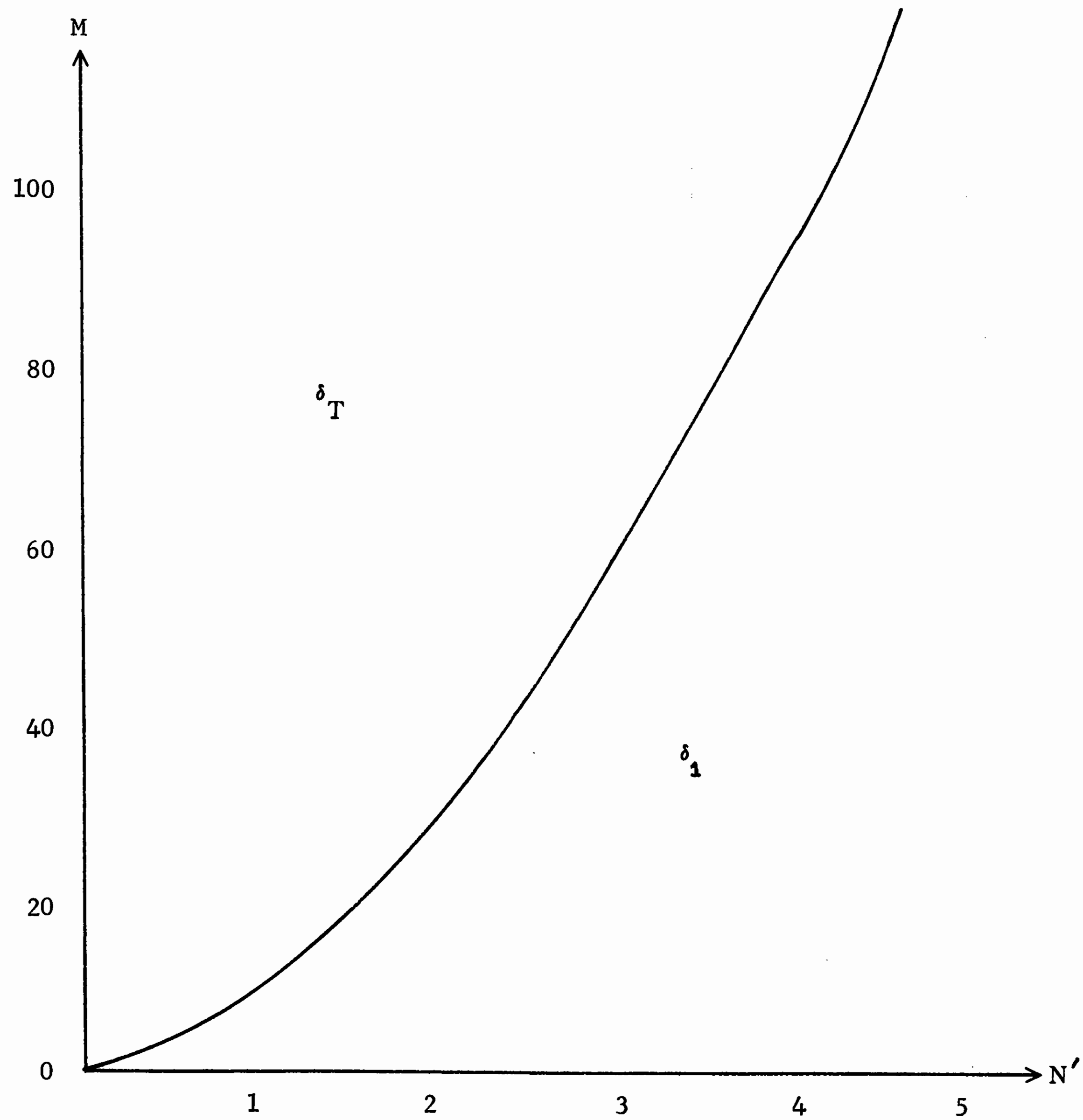

Figure 12. Preference regions for δ_1 and δ_T

And, since δ_1 is Bayes with respect to a $Be(\sqrt{M}\mu_0, \sqrt{M}(1-\mu_0))$ distribution, say G_0, then by theorem 5, δ_1 is $\mathcal{G}''$-minimax and G_0 is least favorable in $\mathcal{G}''$.

Just as in the normal theory case, the above result further justifies the use of the conjugate class of distributions as the class of prior distributions. Although $\mathcal{G}' \subset \mathcal{G}''$, the least favorable distribution in the class $\mathcal{G}''$ is also in the class $\mathcal{G}'$ so the assumption $G\varepsilon\mathcal{G}'$ does not enable us to do any better than the assumption $G\varepsilon\mathcal{G}''$.

To see that $\mathcal{G}''$-admissibility implies admissibility let λ_0 be some point in Λ. Without loss of generality we can assume that $\lambda_0 < \mu_0$ and write

$$\mu_0 - \lambda_0 = \alpha > 0. \tag{6.3.2}$$

Also, define λ_1 by

$$\lambda_1 = \mu_0 + \frac{\alpha\varepsilon}{1-\varepsilon} \tag{6.3.3}$$

where $0 < \varepsilon < 1$.

Then, for any probability distribution function G on Λ such that

$$G(\lambda_0) - G(\lambda_0^-) = \varepsilon \tag{6.3.4}$$

$$G(\lambda_1) - G(\lambda_1^-) = 1-\varepsilon$$

we have

$$\int_\Lambda \lambda dG(\lambda) = \lambda_0\varepsilon + \lambda_1(1-\varepsilon) = \mu_0. \tag{6.3.5}$$

And if we choose ε so that

$$\int_{\Lambda} (\lambda-\mu_0)\,dG(\lambda) = (\lambda_0-\mu_0)^2\varepsilon+(\lambda_1-\mu_0)^2(1-\varepsilon) \tag{6.3.6}$$

$$= \frac{\alpha^2\varepsilon}{1-\varepsilon}$$

$$\leq \mu-\mu^2$$

as required by (5.1.18) then we see that $G\varepsilon\mathcal{G}''$. Finally, since the above argument holds for an arbitrary $\lambda_0\varepsilon\Lambda$, by the comments following (2.4.13) we know that $\mathcal{G}''$-admissibility implies admissibility. By Ferguson [13], p. 62, theorem 3, we know that δ_1 is admissible.

Next, suppose that μ_0 and μ_2 are both known and denote the class of all probability distribution functions with mean μ_0 and variance μ_2 by $\mathcal{G}'''$. The estimator

$$\delta_2(t_{N+1}) = \frac{t_{N+1} + \dfrac{\mu(\mu-\mu_2')}{\mu_2}}{M + \dfrac{(\mu-\mu_2')}{\mu_2}} \tag{6.3.7}$$

has a constant Bayes risk of

$$R(\delta_2,G) = \frac{(\mu-\mu_2')\left\{M + \dfrac{\mu-\mu_2'}{\mu_2}\right\}}{\left[M + \dfrac{\mu-\mu_2'}{\mu_2}\right]^2} \tag{6.3.8}$$

for all $G\varepsilon\mathcal{G}'''$ and is Bayes with respect to a prior Be(R',N'-R'), say G_2. Hence, by theorem 5, δ_2 is $\mathcal{G}'''$-minimax and G_2 is least favorable in $\mathcal{G}'''$.

In addition, if we refer back to expressions (6.3.2) to (6.3.5) and alter (6.3.6) to read

$$\frac{\alpha^2\varepsilon}{1-\varepsilon} = \mu_2 \tag{6.3.9}$$

that is,

$$\varepsilon = \frac{\mu_2}{\alpha^2+\mu_2} \tag{6.3.10}$$

then we have the fact that G'''-admissibility implies admissibility. consequently, we also have the fact that δ_2 is admissible.

If it is also known that G is a member of the conjugate class then G is completely specified since a beta distribution is completely determined by its first two moments. That is, we would be in the simple Bayesian estimation problem with G completely known and could attain the minimum possible Bayes risk.

Also note that

$$R(\delta_2,G) = \frac{\mu-\mu_2'}{M+\frac{\mu-\mu_2'}{\mu_2}} < \frac{\mu-\mu_2'}{M} = R(\delta_T,G). \tag{6.3.11}$$

That is, δ_2 is better than the usual estimator δ_T for all M and all N. Hence knowledge of the mean and variance of the prior is quite valuable, considerably more so than knowledge of the mean alone.

6.4 Asymptotic G-Minimax Decision Functions

As pointed out at the end of section 2.4, even in the presence of prior observations $t_1,\ldots,t_N$ on $f_G(t)$ it may be desirable (at least for small N) to construct asymptotically $\mathcal{G}$-minimax decision functions instead of finding asymptotically optimal EB decision functions.

Suppose that it is known that $\mathcal{G}$ is a class of priors and there exist N past observations $\{t_i\}$ on $f_G(t)$. Then, analogous to the δ_1 estimator defined by (6.2.15) we can write:

$$\delta_0(t_{N+1}) = \frac{t_{N+1}+\sqrt{M}\,\hat{\mu}_0}{M+\sqrt{M}} \tag{6.4.1}$$

where $\hat{\mu}_0$ is defined by:

$$\hat{\mu}_0 = \frac{1}{N+1}\sum_{i=1}^{N+1}\frac{t_i}{M} \; . \tag{6.4.2}$$

It was shown earlier that

$$\hat{\mu}_0 \xrightarrow{\text{a.s.}} \mu_0 \tag{6.4.3}$$

and hence, by theorem 6,

$$\delta_0(t_{N+1}) \ \xrightarrow{\text{a.s.}} \ \delta_0(t_{N+1}) = \frac{t_{N+1}+\sqrt{M}\,\mu_2}{M+\sqrt{M}} \; . \tag{6.4.4}$$

Now, $\delta_0(t_{N+1})$ can be rewritten

$$\delta_0(t_{N+1}) = \frac{t_{N+1}\left(1+\dfrac{1}{(N+1)\sqrt{M}}\right) + \hat{\mu}'\,M}{M+\sqrt{M}} \tag{6.4.5}$$

where

$$\hat{\mu}' = \sum_{i=1}^{N} t_i/M(N+1) \tag{6.4.6}$$

and the global risk involved is found to be:

$$R(\delta_0,G) = R(\delta_1,G) + \left\{\frac{\sqrt{M}}{M+\sqrt{M}}\right\}^2 E_{N+1} (\hat{\mu}_0-\mu_0)^2 \tag{6.4.7}$$

where

$$E_{N+1}(\hat{\mu}_0-\mu_0)^2 = \mathrm{Var}_T(\hat{\mu}_0) \tag{6.4.8}$$

$$= \frac{(M+N')R'(N'-R')}{(N+1)(N')^2(N'+1)M} .$$

Therefore,

$$\lim_{N \to \infty} R(\delta_0,G) = R(\delta_1,G) \qquad \forall G \varepsilon \mathcal{G}'' . \tag{6.4.9}$$

That is, δ_0 is asymptotically $\mathcal{G}''$-minimax.

For a given prior $G\varepsilon\mathcal{G}''$ and a fixed M and N we have

$$R(\delta_0,G) - R(\delta_1,G) = \frac{(M+N')\mu_2}{(M+\sqrt{M})^2(N+1)} \tag{6.4.10}$$

$$\leq \frac{1}{4(\sqrt{M}+1)^2(N+1)}$$

so that the maximum possible difference between the two risk functions decreases rapidly as either M or N increases. And, in the extreme case M=N=1 we find

$$R(\delta_0,G)-R(\delta_1,G) = \frac{\mu-\mu^2}{8} \leq \frac{1}{32} \tag{6.4.11}$$

so that even in this case the difference is slight.

Next, analogous to the $\mathcal{G}'''$-minimax estimator δ_2 given by (6.3.7) we can write (assuming M > 1)

$$\delta_2'(t_{N+1}) = \frac{t_{N+1}+\hat{R}'}{M+\hat{N}'} \tag{6.4.12}$$

where

$$\begin{cases} \hat{R}' = \dfrac{\hat{\mu}(\hat{\mu}-\hat{\mu}_2')}{\hat{\mu}_2} \\ \hat{N}' = \dfrac{\hat{\mu}-\hat{\mu}_2'}{\hat{\mu}_2} \end{cases} \tag{6.4.13}$$

and

$$\hat{\mu} = \frac{\sum_{i=1}^{N+1} T_i}{(N+1)M} \tag{6.4.14}$$

$$\hat{\mu}_2' = \max\left\{\hat{\mu}^2, \left(\frac{1}{N+1}\right) \frac{\Sigma t_i(t_i-1)}{M(M-1)}\right\} .$$

Now, $\delta_2'(t_{N+1}) \xrightarrow{a.s.} \delta_2(t_{N+1})$ and therefore

$$\left\{\delta_2'(t_{N+1}) - \lambda\right\}^2 \xrightarrow{a.s.} \left\{\delta_2(t_{N+1}) - \lambda\right\}^2 \tag{6.4.15}$$

then, since

$$|\delta_2'(t_{N+1})| \leq 1 \quad \forall\ N \text{ and all } t_{N+1} \tag{6.4.16}$$

it follows that

$$\lim_{N\to\infty} R(\delta_2',G) = R(\delta_2,G) \quad \forall G \in \mathcal{G}''' . \tag{6.4.17}$$

That is, δ_2' is an asymptotically $\mathcal{G}'''$-minimax decision function.

APPENDIX

COMPUTER PROGRAMS

All of the computer work in this paper was performed on an IBM 360 model 44 with the programs written in Fortran IV language. The programs used are listed below with brief explanations.

Program 1. RANDU subroutine

This well-known program is designed to generate a random observation from a $U(0,1)$ distribution. The method used is the power residue method described in IBM manual C20-8011, "Random number generation and testing".

Program 2. GAUSS subroutine

This program generates a normally distributed random variable with a given mean μ and standard deviation σ (i.e, a $N(\mu,\sigma^2)$) and requires the RANDU subroutine listed above. It should be noted that this program is not very efficient since it requires 12 uniform random numbers to compute 1 normal random number but it was found to be adequate for the purposes of this paper.

Program 3. The main program

The purposes of this program are:

i) to generate a random sample $\lambda_1,\ldots,\lambda_{N+1}$ from $G(\lambda)$. In the beta example given here this is accomplished by use of the well-known fact that the Kth order statistic in a sample of N from a U(0,1) distribution has a Be(K, N-K+1) distribution.

ii) to generate observations $t_1,\ldots,t_{N+1}$ from $f(t|\lambda_1),\ldots,f(t|\lambda_{N+1})$ respectively.

iii) to compute each decision function under consideration at t_{N+1}.

iv) to estimate (or calculate directly) the global risk for each decision function.

The beta example given here has the following parameters:

Input

NP, RP - prior parameters; Be(RP,NP-RP).

M - number of observations in each t_i (i.e., $t_i \sim B_i(M,\lambda)$).

N - the total number of past and current observations (the same as N+1 in the text of the paper).

IREP - number of replications.

IX - entry value for RANDU subroutine.

Output

R(I) - uniform random numbers I=1, ..., N.

T1(I,J) - the Jth observation in the Ith sample. I=1,...,N; J=1,...,M T1=0 or 1.

T(I) - the Ith binomial obersvation. I=1,...,N.

B(I) - beta random numbers. (B(I) $\equiv \lambda_i$).

EST1,...,EST8 / RISK1,...,RISK8 } the various estimators and their respective risks (either exact or estimated) in the following order:

1. δ_T 2. δ_G 3. δ_N^D 4. δ_N^S 5. δ_0 6. δ_N 7. δ_N''' 8. δ_N''

Program 4. Linkage program

This program provides the link between programs 3 and 5 by computing the entry values for program 5.

The K λ_i's needed to compute G_K were found by evaluating δ_T at t[1] and t[N+1], the smallest and largest order statistics in the sample, and setting

$$\lambda_j = \delta_T(t[1]) + \frac{(j-1)\left|\delta_T(t[N+1])-\delta_T(t[1])\right|}{K-1} \tag{A.1}$$

$$j=1,\ldots,K \leq N \quad .$$

In the binomial case, for example, $\delta_T(t)=t$ and hence

$$\lambda_j = \frac{(K-j)t[1]+(j-1)t[N+1]}{(K-1)}$$

$$\lambda_1 = t[1]$$

$$\lambda_2 = t[1] + \frac{t[N+1]-t[1]}{K-1}$$

$$\vdots$$

$$\lambda_{K-1} = t[1]+ \frac{K-2}{K-1} (t[N+1]-t[1])$$

$$\lambda_K = t[N+1]$$

Parameters computed

XLAMDA - these are the λ_j described above.

A(I,J) - elements of the payoff matrix.

AS(I,J) - linear programming matrix computed directly from the payoff matrix.

BS(I) - the $\underline{b}$ vector in the linear constrain equation $A\underline{x} = \underline{b}$ (A is the AS matrix).

C(I) - the $\underline{c}$ vector where $\underline{c}'\underline{x}$ is to be minimized.

Program 5. SIMPLE subroutine

In general, this program can supply a simplex solution to a linear programming problem. In the present context, however, it is used to compute the step function approximation G_K to the unknown G.

Before calling this subroutine, JH(J) and X(J) must be zeroed out.

Input parameters:

AS,BS,C - described in program 4 explanation.

INFLAG = 0 if no basis provided

1 if basis provided

MX - number of rows of linear programming matrix (4·N).

NN - number of columns of linear programming matrix (4N+K); the total number of variables.

Output parameters

F - Player B's strategies. The first K of these values are the desired weights $h_1,\ldots,h_K$ at the points $\lambda_1,\ldots,\lambda_K$.

PS - Player A's strategies.

Listings

Program 1 RANDU

```
      SUBROUTINE RANDU(IX,IY,YFL)
      IY=IX*65539
      IF(IY)5,6,6
    5 IY=IY+2147483647+1
    6 YFL=IY
      YFL=YFL*.4656613E-9
      RETURN
      END
```

Program 2 GAUSS

```
      SUBROUTINE GAUSS(IX,S,AM,V)
      A=0.0
      DØ50I=1,12
      CALL RANDU(IX,IY,Y)
      IX=IY
   50 A=A+Y
      V=(A-6.0)*S+AM
      RETURN
      END
```

```
C Program 3  The main program
C
      DIMENSION R(100),T(100),TL1(100)
      DIMENSION EST1(150),EST2(150),EST3(150),EST4(150),EST5(150)
      DIMENSION EST6(150),EST7(150),EST8(150)
      DIMENSION XLOSS4(150),XLOSS6(150),XLOSS7(150),XLOSS8(150)
      DIMENSION B(100),T1(100,10)
      NP = 12
      RP = 3
      M = 5
      N = 2
C     N IS THE TOTAL NO. OF CURRENT AND PAST OBSERVATIONS
      ZNP = NP
      ZM = M
      ZN = N
      IRP = RP
      NPL1 = NP - 1
      V1 = RP/ZNP
      V2 = RP*(RP + 1.0)/(ZNP*(ZNP+1.0))
      VAR = V2 - V1**2
      WRITE(6,19)
   19 FORMAT(//7X,'BETA PRIOR DIST.   =   B(R,N-R) ')
      WRITE(6,62) RP,ZNP
   62 FORMAT(//7X,'R = ',F10.4,10X,' N = ' ,F10.4)
      WRITE(6,20) V1,VAR
   20 FORMAT(//7X,'MEAN = ',F15.6,10X,'VARIANCE =',F15.6)
      WRITE(6,900)
  900 FORMAT(/3X,'TRUE PARAMETER',4X,'USUAL',8X,'BAYES',9X,'SEB',
     1PLE',5X,'MEAN ONLY',4X,'CONJUGATE',4X,'CONJ.,MEAN',4X,'CON.
      IREP = 100
      IX = 423
C
      DO 1000 JJ = 1,IREP
      DO 53 K = 1,N
      DO 51 I = 1,NPL1
      CALL RANDU(IX,IY,YFL)
      IX = IY
   51 R(I) = YFL
      DO 52 I = 1,NPL1
      IPL = I + 1
      DO 52 J = 1,NPL1
      IF(J.LT.IPL) GO TO 52
      IF(R(I).LE.R(J)) GO TO 52
      TEMP = R(I)
      R(I) = R(J)
      R(J) = TEMP
   52 CONTINUE
   53 B(K) = R(IRP)
      DO 54 I=1,N
```

```
      DO 54 J=1,M
      CALL RANDU(IX,IY,YFL)
      IX = IY
      IF(YFL.LE.B(I)) T1(I,J) = 1.0
      IF(YFL.GT.B(I)) T1(I,J) = 0.0
   54 CONTINUE
      DO 55 I=1,N
      T(I) = 0.0
      DO 55 J = 1,M
      T(I) = T(I) +T1(I,J)
   55 CONTINUE
C
      DO 42 I = 1,N
      EST1(I) = T(I)/ZM
      EST2(I) = (RP + T(I))/(ZNP + ZM)
      EST5 (I) =(SQRT(ZM)*V1    +T(I))/(ZM+SQRT(ZM))
   42 CONTINUE
C
C  COMPUTATION OF ROBBINS SIMPLE EB ESTIMATOR
      DO 400 I = 1,N
  400 TL1(I) = T(I) - T1(I,M)
      SUM1 = 0.0
      SUM2 = 0.0
      TEMP = TL1(N) + 1.0
      DO 402 J=1,N
      IF(T(J).NE.TEMP) GO TO 401
      SUM1 = SUM1 + 1.0
  401 IF(TL1(J).NE.TL1(N)) GO TO 402
      SUM2 = SUM2 + 1.0
  402 CONTINUE
      EST4(JJ) = ((TL1(N)+1.0)/ZM)*(SUM1/SUM2)
      IF(EST4(JJ).GT.1.0) EST4(JJ) = 1.0
      XLOSS4(JJ) =(EST4(JJ)-B(N))**2
      TEMP1 = 0.0
      TEMP2 = 0.0
      DO 420 I = 1,N
      TEMP1 = TEMP1 + T(I)
  420 TEMP2 = TEMP2 + T(I)*(T(I)-1.0)
      V11 = TEMP1/(ZM*ZN)
      IF(ZM.EQ.1.0) GO TO 430
      V22 = TEMP2/(ZN*ZM*(ZM-1.0))
      V112 = V11**2
      IF(V22.LE.V112)  GO TO 430
      ZNUM = V11*(V11-V22)+T(N)*(V22-V11**2)
      DENOM = ZM*(V22-V11**2)+(V11-V22)
      IF(DENOM.LT.1.0E-10) GO TO 430
      EST6 (JJ) = ZNUM/DENOM
      XLOSS6(JJ) = (EST6(JJ)-B(N))**2
      GO TO 431
```

```
  430 EST6(JJ) = V11
      XLOSS6(JJ)= (V11 - B(N))**2
  431 CONTINUE
      V12 = V1**2
      IF(ZM.EQ.1.0) GO TO 432
      IF(V22.LE.V12) GO TO 432
      IF(V22.GT.V1) GO TO 433
      ZNUM = V1*(V1-V22)+T(N)*(V22-V1**2)
      DENOM = ZM*(V22-V1**2)+(V1-V22)
      EST7(JJ) = ZNUM/DENOM
      GO TO 434
  432 EST7(JJ) = V1
      GO TO 434
  433 EST7(JJ) = T(N)/ZM
  434 XLOSS7(JJ) = (EST7(JJ) - B(N))**2
      VV = V11
      UP = .5 + .5*SQRT(1.0-4.0*VAR)
      DOWN=.5 - .5*SQRT(1.0-4.0*VAR)
      IF(V11.GT.UP)                    VV =.5+.5*SQRT(1.0-4.0*VAR)
      IF(V11.LT.DOWN)                  VV =.5-.5*SQRT(1.0-4.0*VAR)
      ZNUM = VAR*(T(N)-VV )+(VV **2)*(1.0-VV )
      DENOM = VAR*(ZM-1.0)+VV *(1.0-VV )
      EST8(JJ) = ZNUM/DENOM
      XLOSS8(JJ) = (EST8(JJ) - B(N))**2
 1000 CONTINUE
C
      WT = RP*(ZNP-RP)/((ZM*ZNP)*(ZNP + 1.0))
      WG = (RP*(ZM/ZNP-RP+(RP+1.0)/(ZNP+1.0)*((ZNP)**2-ZM)/ZNP))/(ZNP
     1          ZM)**2
      WRITE (6,18) WT,WG
   18 FORMAT('1',////1X,'AVERAGE RISK OF USUAL ESTIMATOR =',E15.6,///
     1'AVERAGE RISK OF BAYES ESTIMATOR =',E15.6)
      ZZ = IREP
      RISK5 = (V1*(1.0-V1))/(SQRT(ZM)+1.0)**2
      WRITE(6,423) RISK5
  423 FORMAT(//1X,'AVG.RISK FOR ONLY MEAN KNOWN EST.=  ',F15.6)
      SUM4 = 0.0
      SUM6 = 0.0
      SUM7 = 0.0
      SUM8 = 0.0
      DO 408 I = 1,IREP
      SUM4 = SUM4 + XLOSS4(I)
      SUM6 = SUM6 +XLOSS6(I)
      SUM7 = SUM7 + XLOSS7(I)
  408 SUM8 = SUM8 + XLOSS8(I)
      RISK4 = SUM4/ZZ
      RISK6 = SUM6/ZZ
      RISK7 = SUM7/ZZ
      RISK8 = SUM8/ZZ
```

```
      TEMP4 = 0.0
      TEMP6 = 0.0
      TEMP7 = 0.0
      TEMP8 = 0.0
      DO 416 I=1,IREP
      TEMP4 = TEMP4 + (XLOSS4(I)- RISK4)**2
      TEMP6 = TEMP6 +(XLOSS6(I) - RISK6)**2
      TEMP7 = TEMP7 + (XLOSS7(I)-RISK7)**2
  416 TEMP8 = TEMP8 + (XLOSS8(I) - RISK8)**2
      VAR4 = TEMP4/(ZZ-1.0)
      VAR6 = TEMP6/(ZZ-1.0)
      VAR7 = TEMP7/(ZZ-1.0)
      VAR8 = TEMP8/(ZZ-1.0)
      SE4 = SQRT(VAR4/ZZ)
      SE6 = SQRT(VAR6/ZZ)
      SE7 = SQRT(VAR7/ZZ)
      SE8 = SQRT(VAR8/ZZ)
      WRITE (6,409) IREP
  409 FORMAT(//2X,'MONTE CARLO ESTIMATES OF BAYES RISKS',10X,'NO. OF
     1LICATIONS  = ',I4,//21X,'AVG. RISK',9X,'VAR',13X,'SE')
      WRITE(6,417) RISK4,VAR4,SE4
  417 FORMAT(/1X,'SIMPLE EB',6X,3F15.6)
      WRITE(6,418) RISK6,VAR6,SE6
  418 FORMAT(/1X,'CONJUGATE PRIOR',3F15.6)
      WRITE(6,419) RISK7,VAR7,SE7
  419 FORMAT(/1X, 'CONJUGATE,MEAN',1X,3F15.6)
      WRITE(6,421) RISK8,VAR8,SE8
  421 FORMAT(/1X,'CONJUGATE,VAR',2X,3F15.6)
      DO 200 J = 5,25,5
      ZJ = J
      DO 200 I = 1,20
      ZI = I
      ZNUM = RP*(ZNP-RP)*(ZNP+ZJ*(ZI**2)+ZI*ZJ+ZI*ZNP)
      DENOM = ZJ*((ZI+1.0)*ZNP)**2*(ZNP+1.0)
      RISK = ZNUM/DENOM
      WRITE(6,100) RISK
  100 FORMAT(1X,F16.7)
  200 CONTINUE
      STOP
      END
C
C Program 4  Linkage program
      TMIN = 1.0E75
      TMAX = -1.0E75
      DO 6 I = 1,N
      IF(EST1(I)-TMIN) 3,4,4
    3 TMIN = EST1(I)
    4 IF (EST1(I)-TMAX) 6,6,5
    5 TMAX = EST1(I)
```

```
    6 CONTINUE
      DO 38 J=1,N1
      XLAMDA(J) = TMIN +((J-1)*ABS(TMAX-TMIN))/(N1-1)
   38 CONTINUE
C
C
C  ORDERING OF THE N SAMPLES
      DO 106 I = 1,N
  106 T2(I) = T(I)
      DO 103 I = 1,N
      IPL = I + 1
      DO 103 J = 1,N
      IF(J.LT.IPL) GO TO 103
      IF (T2(I).LE.T2(J)) GO TO 103
      TEMP = T2(I)
      T2(I) = T2(J)
      T2(J) = TEMP
  103 CONTINUE
C
      DO 101 J = 1,N1
      DO 101 I = 1,N2
      ZI = I
      IF(I.LE.N) A(I,J) = SFB(T2(I),XM,XLAMDA(J)) -(ZI-1.0)/ZN
      IF(I.GT.N) A(I,J) = SFB(T2(I-N),XM,XLAMDA(J)) - (ZI-ZN)/ZN
      A(I+N2,J) = -A(I,J)
  101 CONTINUE
C  THE FOLLOWING LOCATES LARGEST ELEMENT IN FIRST COL, OF THE A MATRI
C        PLACES THE CORRESPONDING ROW IN THE LAST POSITION
      DO 120 I = 1,N4
      DO 120 J = 1,N4
      IF(J.LE.I) GO TO 120
      IF(A(I,1).LE.A(J,1)) GO TO 120
      DO 121 K = 1,N1
      ATEMP(K) = A(I,K)
      A(I,K) = A(J,K)
  121 A(J,K) = ATEMP(K)
  120 CONTINUE
C
C
      DO 140 I= 1,N4
      IF(I.GT.1) GO TO 132
      DO 133 J = 1,N5
      IF(J.LE.N1) GO TO 134
      AS(I,J) = 0.0
      GO TO 133
  134 AS(I,J) = 1.0
  133 CONTINUE
      GO TO 140
  132 CONTINUE
```

```
      DØ 140 J=1,N5
      IF(J.LE.N1) GØ TØ 136
      NTEMP = N1 + I - 1
      IF (J.EQ.NTEMP) GØ TØ 137
      IF (J.EQ.N5) GØ TØ 138
      AS(I,J) = 0.0
      GØ TØ 140
  136 AS(I,J) = A(N4,J) - A(I-1,J)
      GØ TØ 140
  137 AS(I,J) = -1.0
      GØ TØ 140
  138 AS(I,J) = 1.0
  140 CØNTINUE
      DØ 130 I=1,N4
      IF(I-1) 127,127,126
  126 BS(I) = 0.0
      GØ TØ 130
  127 BS(I) = 1.0
  130 CØNTINUE
      DØ 150 I=1,N5
      IF(I-N1) 145,145,146
  145 C(I) = A(N4,I)
      GØ TØ 150
  146 IF(I.EQ.N5) GØ TØ 147
      C(I) = 0.0
      GØ TØ 150
  147 C(I) = 1.0
  150 CØNTINUE
C
C
      DØ 200 J#1,N5
      JH(J) = 0.0
      X(J) = 0.0
      Z(J) = 0.0
  200 CØNTINUE
      CALL SIMPLE(0,N4,N5,AS,BS,C,KØ,KB,PS,JH,X,Y,PE,E)
      DØ 210 J = 1,N5
      IF(JH(J).EQ.0)  GØ TØ 210
      I = JH(J)
      Z(I) = X(J)
  210 CØNTINUE
C
C
C
C THE FØLLØWING CØMPUTES THE SMØØTH EB ESTIMATØR
      SUM1 = 0.0
      SUM2 = 0.0
      DØ 300 J = 1,N1
      IF (XLAMDA(J).EQ.0.0.AND.T(N).EQ.0.0)  GØ TØ 302
```

```
      IF (XLAMDA(J).EQ.1.0.AND.T(N).EQ.XM)    GO TO 303
      ADD1    =       XLAMDA(J)*FB(T(N),XM,XLAMDA(J))*Z(J)
      ADD2    =       FB(T(N),XM,XLAMDA(J))*Z(J)
      GO TO 304
  302 ADD1 = 0.0
      ADD2 = Z(J)
      GO TO 304
  303 ADD1 = Z(J)
      ADD2 = Z(J)
  304 SUM1 = SUM1 + ADD1
  300 SUM2 = SUM2 + ADD2
      IF(SUM2.LT.1.0E-10)  GO TO 309
      EST3(JJ) = SUM1/SUM2
      GO TO 306
  309 EST3(JJ) = EST1(N)
  306 WSEB(JJ) = (EST3(JJ)-B(N))**2
C
C Program 5  SIMPLE subroutine
C
      SUBROUTINE SIMPLE(INFLAG,MX,NN,AS,BS,C,KO,KB,PS,JH,X,Y,PE,E)
C  AUTOMATIC SIMPLEX       REDUNDANT EQUATIONS CAUSE INFEASIBILITY
      REAL BS(1),C(1),PS(1),X(1),Y(1),PE(1),E(1)
      INTEGER INFLAG,MX,NN,KO(6),KB(1),JH(1)
      EQUIVALENCE (XX,LL)
C THE FOLLOWING DIMENSION SHOULD BE THE SAME HERE AS IT IS IN CALLER.
      REAL AS(80,90)
      REAL AA,AIJT,BB,COST,DT,RCOST,TEXP,TPIV,TY,XOLD,XX,XY,YI,YMAX
      INTEGER I,IA,INVC,IR,ITER,J,JT,K,KBJ,L,LL,M,M2,MM,N
      INTEGER NCUT,NUMPV,NUMVR,NVER
      LOGICAL FEAS,VER,NEG,TRIG,KQ,ABSC
C
C          SET INITIAL VALUES, SET CONSTANT VALUES
      ITER = 0
      NUMVR = 0
      NUMPV = 0
      M = MX
      N = NN
      TEXP  =  .5**16
      NCUT  =  4*M + 10
      NVER  =  M/2  +  5
      M2 = M**2
      FEAS = .FALSE.
      IF  (INFLAG.NE.0)  GO TO 1400
      DO 1402  J = 1,N
       KB(J) = 0
       KQ = .FALSE.
       DO 1403  I = 1,M
      IF(AS(I,J).EQ.0.0)  GO TO 1403
      IF(KQ.OR.AS(I,J).LT.0.0)  GO TO 1402
```

```
         KQ = .TRUE.
 1403   CONTINUE
        KB(J) = 1
 1402 CONTINUE
 1400 DO 1401  I = 1,M
        JH (I) = -1
 1401 CONTINUE
C* 'VER'    CREATE INVERSE FROM 'KB' AND 'JH'      (STEP 7)
 1320 VER = .TRUE.
      INVC = 0
      NUMVR   =  NUMVR  +1
      TRIG = .FALSE.
      DO 1101  I = 1,M2
        E(I) = 0.0
 1101 CONTINUE
      MM=1
      DO 1113  I = 1,M
        E(MM) = 1.0
        PE(I)  =  0.0
        X(I) = BS(I)
        IF (JH(I) .NE.0) JH(I) = -1
        MM = MM + M + 1
 1113 CONTINUE
C                 FORM INVERSE
      DO 1102  JT = 1,N
        IF (KB(JT).EQ.0)  GO TO 1102
        GO TO 600
C 600   CALL JMY
C                            CHOOSE PIVOT
 1114   TY = 0.0
        KQ = .FALSE.
        DO 1104 I = 1,M
        IF (JH(I).NE.(-1).OR.ABS(Y(I)).LE.TPIV) GO TO 1104
          IF (KQ) GO TO 1116
          IF (X(I).EQ.0.)  GO TO 1115
          IF (ABS(Y(I)/X(I)).LE.TY)  GO TO 1104
          TY = ABS(Y(I)/X(I))
          GO TO 1118
 1115     KQ = .TRUE.
          GO TO 1117
 1116     IF (X(I).NE.0..OR.ABS(Y(I)).LE.TY)  GO TO 1104
 1117     TY = ABS(Y(I))
 1118     IR = I
 1104   CONTINUE
        KB(JT) = 0
C                            TEST PIVOT
        IF (TY.LE.0.)   GO TO 1102
C                            PIVOT
        GO TO 900
```

```
C 900   CALL PIV
 1102 CONTINUE
C                    RESET ARTIFICIALS
      DO 1109  I = 1,M
        IF (JH(I).EQ.(-1))  JH(I) = 0
        IF (JH(I).EQ.0)  FEAS = .FALSE.
 1109 CONTINUE
 1200 VER = .FALSE.
C                ***         PERFORM ONE ITERATION                  ***
C* !XCK!    DETERMINE FEASIBILITY                              (STEP 1)
      NEG = .FALSE.
      IF (FEAS)  GO TO 500
      FEAS= .TRUE.
      DO 1201  I = 1,M
        IF (X(I).LT.0.0)  GO TO 1250
        IF (JH(I).EQ.0)  FEAS = .FALSE.
 1201 CONTINUE
C* !GET!    GET APPLICABLE PRICES                              (STEP 2)
      IF (.NOT.FEAS)  GO TO 501
  500 DO 503 I = 1,M
       PS(I) = PE(I)
        IF (X(I).LT.0.)  X(I) = 0.
  503 CONTINUE
      ABSC = .FALSE.
      GO TO 599
 1250 FEAS = .FALSE.
      NEG  = .TRUE.
  501 DO 504  J = 1, M
       PS(J) = 0.
  504 CONTINUE
      ABSC = .TRUE.
      DO 505  I = 1,M
        MM = I
        IF (X(I).GE.0.0)  GO TO 507
        ABSC = .FALSE.
        DO 508 J = 1,M
         PS(J) =PS(J) + E(MM)
          MM = MM + M
  508   CONTINUE
        GO TO 505
  507   IF (JH(I).NE.0) GO TO 505
        IF (X(I).NE.0.) ABSC = .FALSE.
        DO 510  J = 1,M
         PS(J) =PS(J) - E(MM)
          MM = MM + M
  510   CONTINUE
  505 CONTINUE
C* !MIN!    FIND MINIMUM REDUCED COST                          (STEP 3)
  599 JT = 0
```

```
      BB   =  0.0
      DO 701  J =1,N
        IF (KB(J).NE.0)   GO TO 701
        DT = 0.0
        DO 303 I = 1,M
          DT = DT +PS(I) *AS(I,J)
  303   CONTINUE
        IF (FEAS)  DT = DT + C(J)
        IF  (ABSC)  DT = - ABS(DT)
        IF (DT.GE.BB)  GO TO 701
        BB = DT
        JT = J
  701 CONTINUE
C  TEST FOR NO PIVOT COLUMN
      IF (JT.LE.0)  GO TO 203
C  TEST FOR ITERATION LIMIT EXCEEDED
      IF (ITER.GE.NCUT)  GO TO 160
      ITER = ITER +1
C* 'JMY'    MULTIPLY INVERSE TIMES A(.,JT)                (STEP 4)
  600 DO 610  I= 1,M
        Y(I) = 0.0
  610 CONTINUE
      LL = 0
      COST = C(JT)
      DO 605   I= 1,M
        AIJT =AS(I,JT)
        IF (AIJT.EQ.0.) GO TO 602
        COST = COST + AIJT * PE(I)
        DO  606  J = 1,M
          LL = LL + 1
          Y(J) = Y(J) + AIJT * E(LL)
  606   CONTINUE
        GO TO 605
  602   LL = LL + M
  605 CONTINUE
C        COMPUTE PIVOT TOLERANCE
      YMAX = 0.0
      DO 620  I  =  1,M
        YMAX = AMAX1( ABS(Y(I)),YMAX )
  620 CONTINUE
      TPIV  =  YMAX * TEXP
C            RETURN TO INVERSION ROUTINE, IF INVERTING
      IF (VER)  GO TO 1114
C       COST TOLERANCE CONTROL
      RCOST = YMAX/BB
      IF (TRIG.AND.BB.GE.(-TPIV)) GO TO 203
      TRIG = .FALSE.
      IF (BB.GE.(-TPIV))  TRIG = .TRUE.
C* 'ROW'    SELECT PIVOT ROW                              (STEP 5)
```

```
C AMONG EQS. WITH X=0, FIND MAXIMUM Y  AMONG ARTIFICIALS, OR, IF NONE
C  GET MAX POSITIVE Y(I) AMONG REALS.
      IR = 0
      AA = 0.0
      KQ = .FALSE.
      DO  1050  I =1,M
        IF (X(I).NE.0.0.OR.Y(I).LE.TPIV)  GO TO 1050
        IF (JH(I).EQ.0)  GO TO 1044
        IF (KQ) GO TO 1050
 1045   IF (Y(I).LE.AA)  GO TO 1050
        GO TO 1047
 1044   IF (KQ) GO TO 1045
        KQ = .TRUE.
 1047   AA = Y(I)
        IR = I
 1050 CONTINUE
      IF (IR.NE.0)  GO TO 1099
      AA = 1.0E+20
C               FIND MIN. PIVOT AMONG POSITIVE EQUATIONS
      DO 1010  I  = 1,M
        IF (Y(I).LE.TPIV.OR.X(I).LE.0.0.OR.Y(I)*AA.LE.X(I) ) GO TO 10
        AA = X(I)/Y(I)
        IR = I
 1010 CONTINUE
      IF (.NOT.NEG)  GO TO 1099
C  FIND PIVOT AMONG NEGATIVE EQUATIONS, IN WHICH X/Y IS LESS THAN THE
C MINIMUM X/Y IN THE POSITIVE EQUATIONS, THAT HAS THE LARGEST ABSF(Y)
      BB = - TPIV
      DO 1030  I = 1,M
        IF (X(I).GE.0..OR.Y(I).GE.BB.OR.Y(I)*AA.GT.X(I) )  GO TO 1030
        BB = Y(I)
        IR = I
 1030 CONTINUE
C  TEST FOR NO PIVOT ROW
 1099 IF  (IR.LE.0)  GO TO 207
C* 'PIV'     PIVOT ON (IR,JT)                             (STEP 6)
      IA = JH(IR)
      IF (IA.GT.0)  KB(IA) = 0
  900 NUMPV  =  NUMPV  +  1
      JH(IR) = JT
      KB(JT) = IR
      YI = -Y(IR)
      Y(IR) = -1.0
      LL = 0
C                                 TRANSFORM INVERSE
      DO  904  J = 1,M
        L = LL + IR
        IF (E(L).NE.0.0)  GO TO 905
        LL = LL + M
```

```
      GØ TØ 904
 905  XY = E(L) / YI
      PE(J) = PE(J) + CØST * XY
      E(L) = 0.0
      DØ 906  I = 1,M
        LL   =  LL + 1
        E(LL) = E(LL)  + XY * Y(I)
 906  CØNTINUE
 904 CØNTINUE
C                              TRANSFØRM X
     XY   =  X(IR) / YI
     DØ 908  I  = 1, M
       XØLD = X(I)
       X(I) = XØLD + XY * Y(I)
       IF (.NØT.VER.AND.X(I).LT.0..AND.XØLD.GE.0.)  X(I) = 0.
 908 CØNTINUE
     Y(IR) = -YI
     X(IR) = -XY
     IF (VER) GØ TØ 1102
     IF (NUMPV.LE.M) GØ TØ 1200
C  TEST FØR INVERSIØN ØN THIS ITERATIØN
     INVC    = INVC   +1
     IF  (INVC.EQ.NVER)   GØ TØ 1320
     GØ  TØ  1200
C*  END ØF ALGØRITHM, SET EXIT VALUES                 ***
 207 IF (.NØT.FEAS.ØR.RCØST.LE.(-1000.)) GØ TØ 203
C            INFINITE SØLUTIØN
     K  =  2
     GØ TØ 250
C          PRØBLEM IS CYCLING
 160 K  =  4
     GØ TØ 250
C          FEASIBLE ØR INFEASIBLE SØLUTIØN
 203 K  =  0
 250 IF (.NØT.FEAS)  K = K + 1
     DØ 1399  J = 1,N
       XX  =  0.0
       KBJ  =  KB(J)
       IF (KBJ.NE.0)  XX = X(KBJ)
       KB(J) = LL
1399 CØNTINUE
     KØ(1) = K
     KØ(2) = ITER
     KØ(3) = INVC
     KØ(4) = NUMVR
     KØ(5) = NUMPV
     KØ(6) = JT
     RETURN
      END
STØP
```

LIST OF REFERENCES

[1] Blischke, W.R. "Estimating the Parameters of Mixtures of Binomial Distributions." JASA, 59 (1964), 510-528.

[2] Blischke, W.R. "Moment Estimators for the Parameters of a Mixture of Two Binomial Distributions." AMS, 33(1962), 444-454.

[3] Blum, J.R. and Judah Rosenblatt. "On Partial a Prior Information in Statistical Inference." AMS, 38 (1967), 1671-1678.

[4] Boes, D.C. "Minimax Unbiased Estimator of Mixing Distribution for Finite Mixtures." Sankhya, A 29 (1967), 417-420.

[5] Boes, D.C. "On the Estimation of Mixing Distributions." AMS, 37 (1966), 177-188.

[6] Clemmer, B.A. and R.G. Krutchkoff. "The Use of Empirical Bayes Estimators in a Linear Regression Model." Biometrika, 55 (1968), 525-535.

[7] Cote, L. and M. Skibinsky. "On the inadmissibility of Some Standard Estimates in the Presence of Prior Information." AMS, 34 (1963), 539-548.

[8] Cozui, C.H. and G.R. Cooper. "The Use of Prior Estimates in Successive Point Estimations of a Random Signal." J. SIAM, 14 (1966), 112-130.

[9] Cramer, H. Mathematical Methods of Statistics. Princeton: Princeton University Press, 1946.

[10] Deely, John J. "Non-Parametric Empirical Bayes Procedures for Selecting the Best of K Populations." AMS, 37 (1966), 540-541.

[11] Deely, J.J. and R.L. Kruse. "Construction of Sequences Estimating the Mixing Distribution." AMS, 39 (1968), 286-288.

[12] Deely, J.J. and W.J. Zimmer. "Shorter Confidence Intervals Using Prior Observations." JASA, 64 (1969), 378-386.

[13] Ferguson, T.S. Mathematical Statistics: A Decision Theoretic Approach. New York: Academic Press, 1967.

[14] Hald, A. "The Mixed Binomial Distribution and the Posterior Distribution of ρ for a Continuous Prior Distribution." JRSS (B), 30 (1968), 359-367.

[15] Hoadley, B. "The Compound Multinomial Distribution and Bayesian Analysis of Categorical Data From Finite Populations." JASA, 64 (1969), 216-229.

[16] Hodges, J.L. and E.L. Lehmann. "The Use of Previous Experience in Reaching Statistical Decisions." AMS, 23 (1952), 396-407.

[17] Johns, M.V. "An Empirical Bayes Approach to Non-Parametric Two-Way Classification" in Studies in Stem Analysis and Prediction (ed. by H. Solomon). Stanford U. Press, 1961, 221-232.

[18] Johns, M.V. "Non-Parametric Empirical Bayes Procedures." AMS, 28 (1957), 649-669.

[19] Kagan, A.M. "An Empirical Bayes Approach to the Estimation Problem." Doklady of the Academy of Sciences of the USSR, 147: 5 (1962), 1020-1021.

[20] Krutchkoff, R.G. "A Supplementary Sample Non-Parametric Empirical Bayes Approach to Some Statistical Decision Problems." Biometrika, 54 (1967), 451-458.

[21] Lindley, D.V. Introduction to Probability and Statistics from a Bayesian Viewpoint: Volumes I and II. Cambridge: University Press, 1965.

[22] Maritz, J.S. "On the Smooth Empirical Bayes Approach to Testing of Hypotheses and the Compound Decision Problem." Biometrika, 55 (1968), 83-100.

[23] Maritz, J.S. "Smooth Empirical Bayes Estimation for Continuous Distributions." Biometrika, 54 (1967), 435-450.

[24] Maritz, J.S. "Smooth Empirical Bayes Estimation for One-Parameter Discrete Distributions." Biometrika, 53 (1966), 417-429.

[25] Miyasawa, K. "An Empirical Bayes Estimator of the Mean of a Normal Population." Bull. Inst. Int. Stat., 38 (1961), 181-188.

[26] Neyman, J. "Two Breakthroughs in the Theory of Statistical Decision Making." Rev. Int. Statist. Inst., 30 (1962), 11-27.

[27] Parzen, E. Modern Probability Theory and Its Applications. New York: John Wiley and Sons, 1960.

[28] Raiffa, Howard and R. Schlaifer. Applied Statistical Decision Theory. Boston: Harvard University Press, 1961.

[29] Rider, Paul R. "Estimating the Parameters of Mixed Poisson, Binomial and Weibull Distributions by the Method of Moments." BISI, 39 (1962), 225-232.

[30] Rider, Paul R. "The Method of Moments Applied to a Mixture of 2 Exponential Distributions." AMS, 32 (1961), 143-147.

[31] Robbins, H. "Asymptotically Subminimax Solutions of Compound Statistical Decision Problems " in Proc. of 2nd Berkeley Symp. on Math. Stat. and Prob., Berkeley, 1951, 131-148.

[32] Robbins, H. "The Empirical Bayes Approach to Statistical Decision Problems." AMS, 35 (1964) 1-20.

[33] Robbins, H. "An Empirical Bayes Approach to Statistics" in Proc. 3rd Berkeley Symp. on Math. Stat. and Prob., 1956, 157-164.

[34] Robbins, H. "The Empirical Bayes Approach to Testing Statistical Hypothesis." Rev. Int. Stat. Inst., 31 (1963), 195-208.

[35] Rolph, J.E. "Bayesian Estimation of Mixing Distributions." AMS, 39 (1968), 1289-1302.

[36] Rutherford, J.R. "Monte Carlo Analysis of the Rate of Convergence of Some Empirical Bayes Point Estimators." AMS, 37 (1966), 766-767.

[37] Rutherford, J.R. and R.G. Krutchkoff. "ϵ-Asymptotic Optimality of Empirical Bayes Estimators." Biometrika, 56 (1969), 220-223.

[38] Rutherford, J.R. and R.G. Krutchkoff. "Some Empirical Bayes Techniques in Point Estimation." Biometrika, 56 (1969), 133-137.

[39] Samuel, E. "An Empirical Bayes Approach to the Testing of Certain Parametric Hypotheses." AMS, 34 (1963), 1370-1385.

[40] Skibinsky, M. "Minimax Estimation of a Random Probability." SIAM Journal Appl. Math, 16 (1968), 134-145.

[41] Suzuki, G. "On Testing Certain Hypotheses." Ann. I. Stat., 20 (1968), 71-80.

[42] Teicher, H. "Indentifiability of Finite Mixtures." AMS, 34 (1963), 1265-1269.

[43] Teicher, H. "Indentifiability of Mixtures." AMS, 32 (1961), 244-248.

[44] Teicher, H. "On the Mixtures of Distribution." AMS, 31 (1960), 55-73.

[45] Tucker, H.G. "An Estimate of the Compounding Distribution of a Compound Poisson Process." Theor. Prob. Appl., 8 (1963), 195-200.

[46] Tucker, H.G. A Graduate Course in Probability. New York: Academic Press, 1967.

[47] Ula, N. and M. Kim. "An Empirical Bayes Approach to Adaptive Control." Journal of the Franklin Institute, 280 (1965), 189-204.

[48] Weiler, H. "The Use of Incomplete Beta Functions for Prior Distributions in Binomial Sampling." Technometrics, 7 (1965), 335-347.

[49] Winkler, R.L. "The Assessment of Prior Distributions in Bayesian Analysis." JASA, 62 (1967), 776-800.

UNCLASSIFIED
Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)
SOUTHERN METHODIST UNIVERSITY

2a. REPORT SECURITY CLASSIFICATION
UNCLASSIFIED

2b. GROUP
UNCLASSIFIED

3. REPORT TITLE
Partial Prior Information: Some Empirical Bayes and G-Minimax Decision Functions

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)
Technical Report

5. AUTHOR(S) (First name, middle initial, last name)
Stephen L. George

6. REPORT DATE
October 30, 1969

7a. TOTAL NO. OF PAGES
165

7b. NO. OF REFS
49

8a. CONTRACT OR GRANT NO.
N00014-68-A-0515

b. PROJECT NO.
NR 042-260

c.

d.

9a. ORIGINATOR'S REPORT NUMBER(S)
48

9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)

10. DISTRIBUTION STATEMENT
This document has been approved for public release and sale; its distribution is unlimited.

11. SUPPLEMENTARY NOTES

12. SPONSORING MILITARY ACTIVITY
Office of Naval Research

13. ABSTRACT

If the prior probability distribution function G of the parameters in a statistical decision theory model is not completely specified then a Bayes decision function cannot be obtained. However, in many cases there may be some partial (i.e., imcomplete) prior information concerning G.

The types of partial prior information considered in this paper is of two kinds: (i) N past observations on the compound distribution $f_G(t)$ or (ii) knowledge of a restrictive class G to which the unknown G is assumed to belong.

When past observations are available, empirical Bayes decision functions are found which are: (i) asymptotically optimal (ii) easy to apply and (iii) better than some optimal non-Bayes decision function for reasonably small N. When only knowledge of the class G is assumed, G-minimax decision functions are found.

In all cases, estimators for the parameters of the normal and Bernoulli processes are found using a quadratic loss function. These estimators are then compared to each other and to other well-known estimators by means of their Bayes risks and, in the empirical Bayes case, by means of their global risks for small N.

DD FORM 1 NOV 65 1473

UNCLASSIFIED
Security Classification