AN ABSTRACT OF THE THESIS OF

CHARLES JAMES PHILLIPS for the DOCTOR OF PHILOSOPHY
(Name) (Degree)

in STATISTICS presented on April 9, 1969
(Major) (Date)

Title: CONSISTENT EMPIRICAL APPROXIMATION OF A-PRIORI DISTRIBUTIONS

Abstract approved: Redacted for Privacy
Donald Guthrie, Jr.

Suppose we have a repeated decision problem based on the random variable X whose distribution has density $f(x|\lambda) \in \mathcal{F}$, where the parameter λ has an unknown cumulative distribution function $G(\lambda) \in \mathcal{G}$. The unconditional density of X is a G-mixture over $\mathcal{F}$, $f_G(x) = \int_\Lambda f(x|\lambda)dG(\lambda)$. The estimation of $G(\lambda)$ using successive observations of the random variable X is developed in this thesis.

To get approximate estimates of G for the identifiable subfamily of the exponential family of distributions, a certain function of λ is approximated by a set of functions derived from the parameterization of the exponential family and a sequence of consistent estimators of the expectation of this function is developed.

In the case where G is a discrete distribution over a finite number of points, we derive consistent estimators based on a sequence of samples whose sizes depend on previously observed

samples. Application to acceptance sampling is indicated. In this latter case the family $\mathcal{F}$ is enlarged to include all densities which are square integrable and bounded almost everywhere with respect to a suitable measure.

Consistent Empirical Approximation of A-priori Distributions

by

Charles James Phillips

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

June 1969

APPROVED:

Redacted for Privacy

Associate Professor of Mathematics and Statistics

in charge of major

Redacted for Privacy

Chairman of Department of Statistics

Redacted for Privacy

Dean of Graduate School

Date thesis is presented April 9, 1969

Typed by Clover Redfern for Charles James Phillips

ACKNOWLEDGMENT

The author wishes to express his thanks to his wife, Jeanette, for her patience and forbearance during the long travail, to his major professor, Dr. Donald Guthrie, Jr., without whose confidence and generous help this thesis would not have been possible, and to the National Science Foundation for support of this research under grants GP-6142 and GP-7491.

TABLE OF CONTENTS

CONSISTENT EMPIRICAL APPROXIMATION OF A-PRIORI DISTRIBUTIONS

I. INTRODUCTION

In the traditional formulation of statistical decision problems the probability distribution of the observed random variable is assumed to be a member of some class $\{F(x|\lambda),\ \lambda \in \Lambda\}$ of distribution functions. No prior information is usually assumed to be known concerning which element of the index set is the true one in any given situation. It is possible, of course, to introduce a priori probability measures defined on Λ as a technical device for generating complete classes of decision functions, minimax decision rules, etc., and in some cases as subjective measures of prior information. It would seem quite reasonable, however, that we may assume in certain experimental situations that such an a priori probability measure actually exists on Λ in the sense that the distributions of the observed random variables occurring in different experiments are selected according to some probability distribution $G(\lambda)$ defined on the index set Λ.

To make such an assumption seems particularly proper in the case where one wishes to make a statistical decision about some characteristic or characteristics of a population at each step of a sequence of independent observations of populations, the decision at

each step depending on some population parameter of the particular population occurring at that step. If the probability distribution of the parameter is known to the investigator then an optimum Bayes decision procedure can be determined. Usually, though, such information will not be available to the investigator, but previous observations on the populations may be and under the proper circumstances these observations may be used to approximate the optimum Bayes procedure. The possibility of using previous observations in such approximations was first conjectured by H. Robbins (1956).

In this paper then we will discuss decision problems arising with the following structure:

a) An observable random variable x from an σ-finite measure space (X, χ, μ).

b) The random variable x has a probability density function $f_\lambda(x)$ depending on a parameter λ.

c) The parameter λ is an element of a space Λ on which is defined a known or unknown a-priori distribution $G(\lambda)$. The specific value of λ in any situation is unknown to us.

d) An action space A with generic element a.

e) A loss function $L(a, \lambda) \geq 0$ which represents the loss incurred in taking action a when the parameter value is λ.

f) A decision function space T such that for $t \in T$; $t : X \to A$.

The problem will be to determine a decision function $t \in T$

such that upon observing x we take action $t(x)$ and incur loss $L(t(x), \lambda)$. Thus for any t the expected loss for a given λ is

$$R(t, \lambda) = \int_X L(t(x), \lambda) f_\lambda(x) d\mu(x) \tag{1.1}$$

which leads to the Bayes risk

$$R(t, G) = \int_\Lambda R(t, \lambda) dG(\lambda). \tag{1.2}$$

If we set

$$\phi_G(a, x) = \int_\Lambda L(a, \lambda) f_\lambda(x) dG(\lambda) \tag{1.3}$$

we may then write the Bayes risk as

$$R(t, G) = \int_X \phi_G(t(x), x) d\mu(x). \tag{1.4}$$

For simplicity we assume the existence of the Bayes decision function t_G such that a. e. $\mu(x)$

$$\phi_G(t_G(x), x) = \min_A \phi_G(a, x). \tag{1.5}$$

Then for any t

$$(1.6) \qquad R(t_G, G) = \int_X \min_A \phi_G(a, x) d\mu(x) \leq R(t, G)$$

and by defining

$$(1.7) \qquad R(G) = R(t_G, G) = \int_X \phi_G(t_G(x), x) d\mu(x)$$

we obtain

$$(1.8) \qquad R(G) = \min_T R(t, G).$$

As was pointed out above under the proper circumstances we may assume the existence of a distribution $G(\lambda)$ on Λ where we are faced repeatedly and independently with the above type decision problem, the value of λ occurring according to $G(\lambda)$. This problem was first considered by Robbins (1956) in a more restricted form and in the more general form again by Robbins (1963). The nonparametric case was considered by Johns (1956) and the parametric case in a paper slightly less general than Robbins' by Samuel (1963).

In this more general case we have a sequence of pairs of random variables $(\lambda_1, x_1), (\lambda_2, x_2), \ldots,$ each pair independent of all others, the λ_n having the common a-priori distribution G, and the conditional density of x_n given λ_n being $f_{\lambda_n}(x) = f(x|\lambda = \lambda_n)$. Given the observation $x_1, x_2, \ldots, x_n$ (the values λ_i remaining unknown) we wish to make a decision about the value λ_{n+1} when

observation x_{n+1} is obtained. Thus we use the decision function

$$t_n(\cdot) = t_n(x_1, x_2, \ldots, x_n; \cdot), \tag{1.9}$$

i.e., we take action $t_n(x_{n+1})$ and incur the loss $L(t_n(x_{n+1}), \lambda_{n+1})$.

Due to this Robbins has defined an "empirical decision procedure" to be a sequence $T^* = \{t_n\}$ of functions of the form (1.9) with values in A. We then havw the expected loss

$$\int_X \phi_G(t_n(x), x)d\mu(x) \tag{1.10}$$

and over all expected loss

$$R_n(T^*, G) = \int_X E\phi_G(t_n(x), x)d\mu(x) \tag{1.11}$$

with expectation taken w.r.t. $x_1, \ldots, x_n$ where

$$f_G(x) = \int_\Lambda f_\lambda(x)dG(\lambda) \tag{1.12}$$

and clearly

$$R_n(T^*, G) \geq R(G). \tag{1.13}$$

Under this formulation then there is a sequence of risk functions $\{R_n\}$ so the following definition is made: If $\lim_{n\to\infty} R_n(T^*, G) = R(G)$

then we say T^* is asymptotically optimal relative to G. Let

$$\Delta_G(a, x) = \int_\Lambda [L(a, \lambda) - L(a_0, \lambda)] f_\lambda(x) dG(\lambda) \tag{1.14}$$

for a fixed arbitrary $a_0 \in A$ and

$$L_0(x) = \int_\Lambda L(a_0, \lambda) f_\lambda(x) dG(\lambda). \tag{1.15}$$

Then if

$$\int_\Lambda L(\lambda) dG(\lambda) < \infty, \quad L(\lambda) = \sup_A L(a, \lambda)$$

$$\phi_G(a, x) = L_0(x) + \Delta_G(a, x), \quad \text{a. e.} \quad \mu(x). \tag{1.16}$$

If we have a sequence of functions

$$\Delta_n(a, x) = \Delta_n(x_1, \ldots, x_n; a, x) \tag{1.17}$$

such that a. e. $\mu(x)$

$$\operatorname*{p\,lim}_{n \to \infty} \sup_A |\Delta_n(a, x) - \Delta_G(a, x)| = 0 \tag{1.18}$$

and a sequence

$$t_n(x) = t_n(x_1, \ldots, x_n; x) = \text{any element } \bar{a} \text{ of } A \text{ such that} \tag{1.19}$$

$$\Delta_n(\bar{a}, n) \leq \inf_A \Delta_n(a, x) + \varepsilon_n \quad \text{where} \quad \{\varepsilon_n\} \to 0$$

then by Theorem 1 of Robbins paper (1963) $T^* = \{t_n\}$ is said to be asymptotically optimal relative to G.

If we restrict A to the case $A = \{a_0, a_1\}$, we may proceed as follows. Find a sequence $G_n(\lambda) = G_n(x_1, x_2, \ldots, x_n; \lambda)$ of random distribution functions in λ such that

$$(1.20) \quad P\left[\lim_{n\to\infty} G_n(\lambda) = G(\lambda) \text{ at every continuity point of } G\right] = 1.$$

One immediately notices that we are then proceeding with a sequence of estimators $G_n(\lambda)$ which is a consistent estimator of $G(\lambda)$, i.e., pointwise consistency at continuity points of $G(\lambda)$. Set

$$(1.21) \quad \Delta_n(x) = \int_{-\infty}^{\infty} [L(a_1, \lambda) - L(a_0, \lambda)] f_\lambda(x) dG_n(\lambda)$$

and if a.e. $\mu(x)$, fixed x,

$$(1.22) \quad [L(a_1, \lambda) - L(a_0, \lambda)] f_\lambda(x)$$

is continuous and bounded in λ then

$$(1.23) \quad p\lim_{n\to\infty} \Delta_n(x) = \Delta_G(x) = \int_{-\infty}^{\infty} [L(a_1, \lambda) - L(a_0, \lambda)] f_\lambda(x) dG(\lambda)$$

which by (1.19) gives us asymptotic optimality of T^* relative to G.

Thus we see that a pointwise consistent (at points of continuity of $G(\lambda)$ sequence of pointwise estimators of $G(\lambda)$) produces asymptotic optimality of T^* relative to G.

In the construction of a sequence of consistent estimators we will need the concept of identifiability of mixtures. There are two particularly good papers on this topic by Teicher (1960, 1963). Let $\mathcal{F} = \{f(x|\lambda), \lambda \in \Lambda\}$ where $f(x|\lambda)$ is a density function in the random variable x for each $\lambda \in \Lambda$ and measurable on the product space $X \times \Lambda$. Then for any non-degenerate (c.d.f. G such that

$$f_G(x) = \int_\Lambda f(x|\lambda)dG(\lambda) \tag{1.24}$$

is a density function of the random variable x, $f_G(x)$ is called a G-mixture of $\mathcal{F}$ or, simply, a mixture.

Let $\mathcal{G}$ denote a class of such c.d. f's $\{G\}$, $\mathcal{H}$ the induced class of mixtures $\{f\}$ and $\mathcal{Q}$ the class of degenerate distributions in Λ. Then $\mathcal{H}$ will be called identifiable in $\mathcal{G}$ (with respect to $\mathcal{F}$) if (1.24) effects a one-to-one correspondence between $\mathcal{H} \cup \mathcal{F}$ and $\mathcal{G} \cup \mathcal{Q}$; equivalently, if

$$f(x) = \int_\Lambda f(x|\lambda)dG_1(\lambda) = \int_\Lambda f(x|\lambda)dG_2(\lambda)$$

implies $G_1 = G_2$ for all $G_1, G_2 \in \mathcal{G} \cup \mathcal{Q}$. If $\mathcal{H}$ is identifiable in the class of all $G \notin \mathcal{Q}$, it is simply called identifiable.

In particular, Teicher shows that the class of mixtures generated by the normal distribution with either the mean or variance fixed is identifiable and the class of gamma mixtures with one of the parameters fixed also identifiable.

Another pair of concepts to be used are those of completeness and closure of sequences in a normed linear space. Given a normed linear space X we say a sequence $\{x_n\}$ of elements of X is closed if every element of X can be approximated arbitrarily closely by finite linear linear combinations of the elements of the sequence, where the degree of approximation is measured in terms of the norm of the space. We also say that a sequence $\{x_n\}$ is complete if $L(x_n) = 0$, $n = 0, 1, \ldots, L \in X^*$, implies $L = 0$. Here of course, L is a linear functional in the conjugate space of X. The fundamental relationship here is the equivalence of these two concepts in a normed linear space. For an excellent discussion of these concepts and their relationships and applications to approximation theory see Davis (1963).

We will address ourselves to two problems in this paper, the first being the construction of a sequence of consistent estimates of $G(\lambda)$, considered in Chapter II. Some work has been done on this for finite mixtures by Robbins (1964), and by Rolph (1968) for

mixtures with discrete domain. We will initially assume the family $\{f_G\}$ of G mixtures.

$$f_G(x) = \int_\Lambda f(x|\lambda)dG(\lambda) \tag{1.25}$$

is identifiable. We then construct a sequence of pointwise estimators $\{G_n(\lambda)\}$ of $G(\lambda)$, where $f(x|\lambda)$ is a member of the exponential family, by a method of approximation in a normed linear space of certain functions of the parameter. We use the strong law of large numbers to establish consistency.

In Chapter III we discuss a generalization of Robbins's (1964) work on discrete distributions to the case of variable dimensional random vectors with applications to the theory of acceptance sampling.

II. ESTIMATION OF A-PRIORI DISTRIBUTIONS OVER COMPACT SUBSETS OF THE PARAMETER SPACE Λ

Suppose that we have repeated observations of the random variable distributed with density

$$f_G(x) = \int_\Lambda f(x|\lambda)dG(\lambda), \tag{2.1}$$

where

$$f(x|\lambda) \in \mathcal{F} \tag{2.2}$$

and the parameter λ has the unknown cumulative distribution function

$$G(\lambda) \in \mathcal{G}. \tag{2.3}$$

The problem is, of course, to estimate $G(\lambda)$ using observations of the random variable X. We will assume a measure space (X, χ, μ) and cumulative distribution function

$$F(x|\lambda) = P_\lambda(X \leq x) \tag{2.4}$$

which is absolutely continuous with respect to μ and

$$\int_X f^2(x|\lambda)d\mu(x) < \infty, \quad \lambda \in \Lambda. \tag{2.5}$$

We further assume that the family $\mathcal{F}$ is identifiable in $\mathcal{G}$, i.e.,

$$\int_{\Lambda} f(x|\lambda)dG_1(\lambda) = \int_{\Lambda} f(x|\lambda)dG_2(\lambda) \tag{2.6}$$

implies

$$G_1(\lambda) = G_2(\lambda), \quad f \in \mathcal{F}, \quad G_i \in \mathcal{G},$$

Suppose further we assume there exists a compact subset L of Λ such that $P(\lambda \in \tilde{L}) = 0$. For our explicit purpose Λ will be the real line and L a closed interval $[\lambda_\ell, \lambda_r]$, $-\infty < \lambda_\ell < \lambda_r < \infty$.

We estimate $G(\lambda)$ and other expressions involving λ using linear combinations of $f(x|\lambda_i)$. Let

$$\Phi_n(x|\lambda^*) = \sum_{i=1}^{k_n} a_{i\lambda^*}^{(n)} f(x|\lambda_i^{(n)}) \tag{2.7}$$

where

$$\lambda_i^{(n)} \in \Lambda, \quad i = 1, \ldots, k_n, \quad n \in I^+,$$

and

$$\bar{G}_m^n(\lambda^*) = \frac{1}{m} \sum_{\nu=1}^{m} \Phi_n(x_\nu|\lambda^*). \tag{2.8}$$

Consider

$$E_X \Phi_n(x|\lambda^*) = \int_X \Phi_n(x|\lambda^*) \int_{\Lambda} f(x|\lambda)dG(\lambda)d\mu(x). \tag{2.9}$$

In (2.1) we assumed

$$(2.10) \qquad 1 = \int_X f_G(x)d\mu(x) = \int_X \int_\Lambda f(x|\lambda)dG(\lambda)d\mu(x)$$

$$= \int_Y f(x|\lambda)d\pi(Y)$$

where (Y, y, π) is the measure space generated by $X \times \Lambda$.

Clearly $f(x|\lambda_i^{(n)})f(x|\lambda)$ is π integrable. Thus by Fubini's theorem

$$(2.11) \qquad E_X\Phi_n(x|\lambda^*) = \int_\Lambda \int_X \Phi_n(x|\lambda^*)f(x|\lambda)d\mu(x)dG(\lambda)$$

$$= \int_\Lambda H_n(\lambda|\lambda^*)dG(\lambda)$$

where we let

$$(2.12) \qquad \int_X \Phi_n(X|\lambda^*)f(x|\lambda)d\mu(x) = H_n(\lambda|\lambda^*) .$$

If

$$H_n(\lambda|\lambda^*) = \begin{cases} 1, & \lambda \leq \lambda^* \\ 0, & \lambda > \lambda^* \end{cases}$$

then

$$E_X\Phi_n(x|\lambda^*) = G(\lambda^*).$$

Thus, we wish to choose the $a_{i\lambda^*}^{(n)}$ such that

$$H_n(\lambda|\lambda^*) \doteq H(\lambda|\lambda^*) \tag{2.13}$$

where

$$H(\lambda|\lambda^*) = \begin{cases} 1, & \lambda \leq \lambda^* \\ 0, & \lambda > \lambda^* \end{cases} . \tag{2.14}$$

(Clearly, in general we would not be able to choose the $a_{i\lambda^*}^{(n)}$ so that $H_n(\lambda|\lambda^*) = H(\lambda|\lambda^*)$ for all values of λ.) Now

$$H_n(\lambda|\lambda^*) = \int_X \sum_{i=1}^{k_n} a_{i\lambda^*}^{(n)} f(x|\lambda_i^{(n)}) f(x|\lambda) d\mu(x) \tag{2.15}$$

$$= \sum_{i=1}^{k_n} a_{i\lambda^*}^{(n)} \int_X f(x|\lambda_i^{(n)}) f(x|\lambda) d\mu(x) .$$

Let

$$\int_X f(x|\lambda_i^{(n)}) f(x|\lambda) d\mu(x) = h_{\lambda_i^{(n)}}(\lambda) \tag{2.16}$$

so that

$$H_n(\lambda|\lambda^*) = \sum_{i=1}^{k_n} a_{i\lambda^*}^{(n)} h_{\lambda_i^{(n)}}(\lambda) . \tag{2.17}$$

We have the following

Lemma 1. $\{f(x|\lambda_i^{(n)}), \ldots, f(x|\lambda_{k_n}^{(n)})\}$

is a linearly independent set of functions for $\lambda_i^{(n)} \neq \lambda_j^{(n)}$, $i \neq j$.

Proof. Suppose $C_1 f_1^{(n)} + C_2 f_2^{(n)} + \ldots + C_{k_n} f_{k_n}^{(n)} = 0$ for

$$\sum_{i=1}^{k_n} C_i^2 \neq 0 .$$

Then we may rearrange and renumber so that

$$C_1 f_1^{(n)} + \ldots + C_\ell f^{(n)} = C_{\ell+1} f_{\ell+1}^{(n)} + \ldots + C_q f_q^{(n)} ,$$

where $C_i > 0 \; \forall_i \leq q$. Then integrating with respect to $\mu(x)$ we have

$$C_1 + C_2 + \ldots + C_\ell = C_{\ell+1} + \ldots + C_q = C > 0$$

and

$$\frac{C_1}{C} + \ldots + \frac{C_\ell}{C} = \frac{C_{\ell+1}}{C} + \ldots + \frac{C_q}{C} = 1 .$$

Let $G_1(\lambda)$ assign probability $\frac{C_i}{C}$ to $f(x|\lambda)$ if $\lambda = \lambda_i^{(n)}$, $i = 1, 2, \ldots, \ell$ and 0 to $f(x|\lambda)$ otherwise; $G_2(\lambda)$ assigns probability $\frac{C_i}{C}$ to $f(x|\lambda)$ if $\lambda = \lambda_i^{(n)}$, $i = \ell+1, \ldots, q$ and 0 otherwise. Then

$$f_{G_1}(x) = \sum_{i=1}^{\ell} f(x|\lambda_i^{(n)}) \frac{C_i}{C} = \sum_{i=\ell+1}^{q} f(x|\lambda_i^{(n)}) \frac{C_i}{C} = f_{G_2}(x)$$

which by (2.6) implies $G_1(\lambda) = G_2(\lambda)$, and thus we have a contradiction. Then $C_i = 0$, $\forall_i$ and the lemma is proved.

We immediately have

Lemma 2. The set of functions $\{h^{(n)}_{\lambda_1}(\lambda), \ldots, h^{(n)}_{\lambda k_n}(\lambda)\}$ is a linearly independent set of functions.

Proof. Suppose not. Then

$$\sum_{i=1}^{k_n} b_i^{(n)} h^{(n)}_{\lambda i}(\lambda) = 0,$$

where

$$\sum_{i=1}^{k_n} (b_i^{(n)})^2 \neq 0$$

thus

$$\int_X \sum_{i=1}^{k_n} b_i^{(n)} f(x|\lambda_i^{(n)}) f(x|\lambda) \, d\mu(x) = 0$$

This says

$$\sum_{i=1}^{k_n} b_i^{(n)} f(x|\lambda_i^{(n)}) = 0$$

which contradicts the preceding lemma. Thus the lemma is proved.

Let us now restrict ourselves to the subset of the exponential

family which is identifiable. We shall use the representation where the sigma-finite measure $\mu(x)$ determines the density function of a member of the family in the form

$$f(x|\lambda) = \beta(\lambda)e^{\lambda x} \tag{2.18}$$

where

$$\beta(\lambda) = \left(\int_X e^{\lambda x} d\mu(x)\right)^{-1} \tag{2.19}$$

Lehmann (1959) has shown that $\beta(\lambda)$ is a continuous and differentiable function in λ, and thus that

$$h_{\lambda_i^{(n)}}(\lambda) = \frac{\beta(\lambda_i^{(n)})\beta(\lambda)}{\beta(\lambda_i^{(n)}+\lambda)} \tag{2.20}$$

is a continuous function on any finite closed interval $[\lambda_i^{(n)}, \lambda_{k_n}^{(n)}]$. Thus

$$\int_{\lambda_1^{(n)}}^{\lambda_{k_n}^{(n)}} [h_{\lambda_i^{(n)}}(\lambda)]^2 d\lambda < \infty, \quad \text{(Lebesgue measure)}, \tag{2.21}$$

and the set of $h_{\lambda_i^{(n)}}(\lambda)$ generates a finite dimensional subspace of the Hilbert space H of square integrable functions on $[\lambda_1^{(n)}, \lambda_{k_n}^{(n)}]$. Clearly $H(\lambda|\lambda^*) \in H$.

Let us consider the case of

$$f(x|\lambda) = \frac{1}{\sqrt{2\pi}\, e^{\lambda^2/2}} e^{\lambda x}, \quad -\infty < \lambda < \infty, \tag{2.22}$$

where

$$\frac{d\mu}{dm}(x) = e^{-x^2/2}, \tag{2.23}$$

where x is real and m denotes Lebesgue measure.

It is shown by Waterman (1966) that this represents the normal distribution with mean λ and variance 1. Immediately we observe that

$$\beta(\lambda) = \frac{1}{\sqrt{2\pi}\, e^{\lambda^2/2}} \tag{2.24}$$

so that

$$\frac{\beta(\lambda_i^{(n)})\beta(\lambda)}{\beta(\lambda+\lambda_i^{(n)})} = \frac{1}{\sqrt{2\pi}} e^{\lambda_i^{(n)}\lambda} . \tag{2.25}$$

If one wishes to consider the case where the variance is $\sigma^2 \neq 1$, but fixed, then merely define $\mu(x)$ by

$$\frac{d\mu}{dm}(x) = e^{-x^2/2\sigma^2},$$

so that

$$\beta(\lambda) = \frac{1}{\sigma\sqrt{2\pi}\, e^{\lambda^2\sigma^2/2}}$$

and

$$\frac{\beta(\lambda_i^{(n)})\beta(\lambda)}{\beta(\lambda+\lambda_i^{(n)})} = \frac{1}{\sigma\sqrt{2\pi}} e^{\sigma^2\lambda_i^{(n)}\lambda} .$$

This change does not effect the subsequent analysis so it will not be pursued further. Thus let us look at

$$\frac{1}{\sqrt{2\pi}} e^{\lambda_i^{(n)}\lambda} , \tag{2.26}$$

where

$$\lambda_\ell < \lambda_r, \quad \lambda_\ell \leq \lambda \leq \lambda_r, \quad \lambda_\ell \leq \lambda_i^{(n)} \leq \lambda_r, \tag{2.27}$$

and

$$|\lambda_i^{(n)}| > 0 .$$

Let

$$\lambda = (\lambda_r - \lambda_\ell)\omega + \lambda_\ell . \tag{2.28}$$

Then

$$0 \leq \omega \leq 1 \quad \text{for} \quad \lambda_\ell \leq \lambda \leq \lambda$$

and

$$\frac{1}{\sqrt{2\pi}} e^{\lambda_i^{(n)}\lambda} = \frac{1}{\sqrt{2\pi}} e^{\lambda_i^{(n)}\lambda(\lambda_r-\lambda_\ell)\omega+\lambda_i^{(n)}\lambda_\ell} \tag{2.29}$$

$$= \frac{1}{\sqrt{2\pi}} e^{\lambda_i^{(n)}\lambda_\ell} e^{\lambda_i^{(n)}(\lambda_r-\lambda_\ell)\omega} .$$

In any linear combination of the above terms the constants

$e^{\lambda_i^{(n)}\lambda}$ may be absorbed into the coefficients of $e^{\lambda_i^{(n)}(\lambda_r-\lambda_\ell)\omega}$ so that without loss of generality we consider the independent set $\{e^{\lambda_i^{(n)*}\omega}\}$ where

(2.30) $$\lambda_i^{(n)*} = \lambda_i^{(n)}(\lambda_r-\lambda_\ell) .$$

The following theorem by Szàsz is proven in Davis (1963):

Let

$$F(Z) = \sum_{k=0}^{\infty} C_k Z^k, \quad C_k \text{ real},$$

be a fixed power series and let $r > 0$ be its radius of convergence. Assume that

$$\sum_{n=1}^{\infty} \frac{1}{k_n} = \infty$$

where $k_1 < k_2 < \ldots$ is the sequence of all those integers ≥ 1 for which $C_k \neq 0$. If $\{t_n\}$ is a sequence of distinct real numbers satisfying $0 < |t_n| \leq r_1 < r$, then the sequence of functions $\{f_n(x)\} = \{F(t_n x)\}$, $n = 1, 2, \ldots,$ is complete in $L^2[0, 1]$. If $C_0 \neq 0$, it is also complete in $C[0, 1]$.

Now

$$e^z = 1 + \sum_{k=1}^{\infty} \frac{1}{k!} z^k$$

with

$$r = \infty, \quad \text{i.e.,} \quad |z| < \infty, \quad C_0 = 1, \quad \text{and} \quad C_k = \frac{1}{k!} .$$

Thus

$$k_n = n \quad \text{and} \quad \sum_{n=1}^{\infty} \frac{1}{k_n} = \sum_{n=1}^{\infty} \frac{1}{n} = \infty.$$

Now

$$0 < |\lambda_\ell(\lambda_r - \lambda_\ell)| \leq |\lambda_i^{(n)*}| < |\lambda_r(\lambda_r - \lambda_\ell)| < \infty. \tag{2.31}$$

So let $i = n$, $k_n = n$,

$$\lambda_n^{(n)*} = \lambda_\ell(\lambda_r - \lambda_\ell) + \frac{n-1}{n}(\lambda_r - \lambda_\ell)^2, \quad n = 1, 2, \ldots . \tag{2.32}$$

Then by Szàsz' theorem $\{e^{\lambda_n^{(n)*}\omega}\}$, $n = 1, 2, \ldots,$ is complete in $L^2[0, 1]$ and also in $C[0, 1]$. Thus with

$$\lambda_n^{(n)} = \lambda_\ell + \frac{n-1}{n}(\lambda_r - \lambda_\ell) \tag{2.33}$$

we see that $\{e^{\lambda_n^{(n)}\omega}\}$ is complete in $L^2[\lambda_\ell, \lambda_r]$ and $C[\lambda_\ell, \lambda_r]$.

The following theorem due to Banach is also proven in Davis (1963):

"Let X be a normed linear space (real or complex). A sequence of elements $\{x_k\}$ is closed if and only if it is complete."

Thus from the above theorem the sequence $\{e^{\lambda^{(n)}\lambda}\}$ is closed in

$L^2[\lambda_\ell, \lambda_r]$ and $C[\lambda_\ell, \lambda_r]$.

Finally, Davis proves the following theorem:

If X is a normed linear space, y an arbitrary element of X, and

$$E_n(y) = \min_{a_i} \| y - \sum_{i=1}^{m} a_i x_i \|$$

$\lim_{n \to \infty} E_n(y) = 0$ for all $y \in X$ if and only if the independent sequence $\{x_i\}$ is closed in X.

Thus $\{e^{\lambda_n^{(n)} \lambda}\}$ is dense in $L^2[\lambda_\ell, \lambda_r]$ and in $C[\lambda_\ell, \ell_r]$ and hence we see that this method of estimation is tractable for the normal distribution. It is not to be understood that the particular sequence $\{\lambda_n^{(n)}\}$ here used is the only, or in some sense best, sequence for this purpose. For example, if $\lambda_\ell \neq -\lambda_r$, then the sequence

$$\{\lambda_n^{(n)}\} = \lambda_\ell, \lambda_r, \frac{\lambda_\ell + \lambda_r}{2}, \frac{\lambda_\ell + \lambda_r}{4}, \frac{3(\lambda_\ell + \lambda_r)}{4}, \ldots,$$

would also serve.

Now let us consider the Gamma family:

$$(2.34) \qquad f(x|\lambda) = \frac{|\lambda|^r}{\Gamma(r)} x^{r-1} e^{-|\lambda| x}, \quad r > 0, \ x \in [0, \infty), \ \lambda \in (-\infty, 0).$$

For this family $\beta(\lambda) = \frac{|\lambda|^r}{\Gamma(r)}$ and thus

$$\frac{\beta(\lambda_i)\beta(\lambda)}{\beta(\lambda_i^{(n)}+\lambda)} = \frac{|\lambda_i^{(n)}|^r |\lambda|^r}{\Gamma(r)(|\lambda_i^{(n)}+\lambda|)^r} . \tag{2.35}$$

Now consider

$$\frac{1}{|\lambda_i^{(n)}+\lambda|^r} .$$

Since $\lambda_i^{(n)}$ and λ are always less than 0, let $\lambda' = -\lambda$, i.e., $\lambda' = |\lambda|$ and then $|\lambda_i^{(n)}+\lambda| = \lambda_i^{(n)'} + \lambda'$. Now dropping the primes let

$$\lambda = (\lambda_r - \lambda_\ell)\omega + \lambda_\ell, \quad \lambda_\ell < \lambda_r, \tag{2.36}$$

so that we have to consider

$$\frac{1}{(\lambda_i^{(n)}+\lambda_\ell)^r \left[1+\frac{\lambda_r-\lambda_\ell}{\lambda_i^{(n)}+\lambda_\ell}\omega\right]^r} \tag{2.37}$$

or in orther words, without loss of generality

$$\frac{1}{\left[1+\frac{\lambda_r-\lambda_\ell}{\lambda_i^{(n)}+\lambda_\ell}\omega\right]^r}, \quad (0 \le \omega \le 1). \tag{2.38}$$

Now $F(z) = (1+z)^s$, $s \neq 1, 2, \ldots,$ has all Maclaurin's coefficients non-zero and the radius of convergence $r = 1$. Thus by Szàsz' theorem we have $\{(1+t_n x)^s\}$ complete in $L^2[0, 1]$ and $C[0, 1]$ for

$0 < |t_n| \leq 1 - \epsilon, \quad \epsilon > 0.$

Since $\lambda_r > \lambda_\ell$ and $\lambda_\ell > 0$

$$\left| \frac{\lambda_r - \lambda_\ell}{\lambda_i^{(n)} + \lambda_\ell} \right| = \frac{\lambda_r - \lambda_\ell}{\lambda_i^{(n)} + \lambda_\ell} \tag{2.39}$$

and

$$\frac{\lambda_r - \lambda_\ell}{\lambda_i^{(n)} + \lambda_\ell} < 1 \tag{2.40}$$

for

$$\lambda_i^{(n)} > \lambda_r - 2\lambda_\ell. \tag{2.41}$$

Thus let

$$t_n = \frac{\lambda_r - \lambda_\ell}{\lambda_n^{(n)} + \lambda_\ell}$$

where $\lambda_n^{(n)} = n\lambda_r > \lambda_r - 2\lambda_\ell$, $n = 1, 2, \ldots,$ and then $\dfrac{1}{(1+t_n\omega)^r}$ is complete in $L^2[0, 1)$ and $C[0, 1)$, i.e.,

$$\left\{ \frac{1}{\left[1 + \dfrac{\lambda_r - \lambda_\ell}{\lambda_n^{(n)} + \lambda_\ell} \omega \right]^r} \right\} = \left\{ \frac{(\lambda_n^{(n)} + \lambda_\ell)^r}{((\lambda_n^{(n)} + \lambda_\ell) + (\lambda_r - \lambda_\ell)\omega)^r} \right\}. \tag{2.42}$$

is complete in $L^2[0, 1]$ and $C[0, 1]$.

Then so is

$$\left\{\frac{1}{((\lambda_n^{(n)}+\lambda_\ell+(\lambda_r-\lambda_\ell)\omega)^r}\right\}$$

and thus

$$\left\{\frac{\lambda_n^{(n)r}}{\Gamma(r)(\lambda_n^{(n)}+\lambda)^r}\right\} \tag{2.43}$$

is complete in $L^2[\lambda_\ell, \lambda_r]$.

Lemma. Let $w(x) \in C[0, 1]$ and $\frac{1}{w(x)} \geq \varepsilon > 0$ there. Then $\{w(x)f_n(x)\}$ is closed (complete) in $L^2[0, 1]$ if and only if $\{f_n(x)\}$ is.

Proof. (Although the proofs of both parts of the lemma are easy we will give only the "if" part here.) Let $g(x)$ be any element of $L^2[0, 1]$. Then

$$\int_0^1 \frac{g^2(x)}{w^2(x)}\, d\mu(x) \leq \frac{1}{\varepsilon^2}\int_0^1 g^2(x)d\mu(x) < \infty.$$

Therefore $\frac{g(x)}{w(x)} \in L^2[0, 1]$. Thus $\frac{g(x)}{w(x)}$ can be approximated arbitrarily closely by finite linear combinations of elements of $\{f_n(x)\}$. Using these linear combinations which approximate $\frac{g(x)}{w(x)}$ and multiplying them by $w(x)$ we approximate $g(x)$ by $\{w(x)f_n(x)\}$.

Clearly this lemma generalizes to $L^2[\lambda_\ell, \lambda_r]$. Let $w(\lambda) = \lambda^r$.

Then since

$$f_n(\lambda) = \frac{\lambda_n^r}{\Gamma(r)(\lambda_n^{(n)}+\lambda)^r}$$

we have

$$\left\{w(\lambda)f_n(\lambda)\right\} = \left\{\frac{\lambda_n^{(n)}\lambda^r}{\Gamma(r)(\lambda_n^{(n)}+\lambda)^r}\right\} \tag{2.44}$$

is closed (complete) in $L^2[\lambda_\ell, \lambda_r]$. Note that since $\frac{1}{\lambda^r}$ is continuous on $[\lambda_\ell, \lambda_r]$, $(\lambda_\ell > 0)$ a slight modification of this proof for $[\lambda_\ell, \lambda_r]$ yields closure in $C[\lambda_\ell, \lambda_r]$. Remembering that the λ's above represent $|\lambda|$ and $\lambda^{(n)} + \lambda$ represents $|\lambda_n^{(n)}+\lambda|$ for $\lambda < 0$ we have proven the following theorem:

> The $\beta(\cdot)$ functions of the Gamma family are closed in $L^2[\lambda_\ell, \ell_r]$ and $C[\lambda_\ell, \lambda_r]$.

In accordance with the above we make the following definition:

> Any member of the exponential family which is identifiable and is such that there exists a sequence
>
> $$\left\{\frac{\beta(\lambda_n)\beta(\lambda)}{\beta(\lambda_n+\lambda)}\right\}$$
>
> which is closed in $L^2[\lambda_\ell, \lambda_r]$ and $C[\lambda_\ell, \lambda_r]$ will be called

approximable.

Let us assume now that $G(\lambda)$ is known to be continuous.

Theorem. Let $f(x|\lambda)$ be approximable and $G(\lambda)$ continuous. Then given $\varepsilon > 0$ and $\lambda^* \in [\lambda_\ell, \lambda_r]$ there exists an N^* such that $|E_X(\Phi_{N^*}(x|\lambda^*) - G(\lambda^*)| < \varepsilon$.

Proof. Since $f(x|\lambda)$ is approximable there exists a sequence

$$\left\{\frac{\beta(\lambda_n)\beta(\lambda)}{\beta(\lambda_n+\lambda)}\right\}$$

which converges in L^2. Thus we may form a sequence $\{H_n(\lambda|\lambda^*)\}$ such that

$$H_{n_k}(\lambda|\lambda^*) \xrightarrow{L^2} H(\lambda|\lambda^*). \tag{2.45}$$

Therefore there exists a subsequence $\{H_{n_k}(\lambda|\lambda^*)\}$ such that

$$H_{n_k}(\lambda|\lambda^*) \xrightarrow{\text{a.e.}} H(\lambda|\lambda^*). \tag{2.46}$$

Thus we may choose a positive integer K such that

$$|H_{n_K}(\lambda|\lambda*) - H(\lambda|\lambda*)| < \varepsilon \quad \text{for all} \quad \lambda \in [\lambda_\ell, \lambda_r] \setminus A$$

where $A \subset [\lambda_\ell, \lambda_r]$ and $m(A) = 0$. Now by (2.11) and (2.14) we

have

$$(2.47)\quad \left|E_X(\Phi_N(x|\lambda^*)) - G(\lambda^*)\right| = \left|\int_{\lambda_\ell}^{\lambda_g} H_N(\lambda|\lambda^*)dG(\lambda) - \int_{\lambda_\ell}^{\lambda_r} H(\lambda|\lambda^*)dG(\lambda)\right|$$

where $N = n_K$ So

$$(2.48)\quad |E_X(\Phi_N(x|\lambda^*)) - G(\lambda^*)| \le \int_{\lambda_\ell}^{\lambda_r} |H_N(\lambda|\lambda^*) - H(\lambda|\lambda^*)|dG(\lambda)\,.$$

Now

$$(2.49)\quad \int_{\lambda_\ell}^{\lambda_r} |H_N(\lambda|\lambda^*) - H(\lambda|\lambda^*)|dG(\lambda)$$

$$= \int_{[\lambda_\ell, \lambda_r]\setminus A} |H_N(\lambda|\lambda^*) - H(\lambda|\lambda^*)|dG(\lambda) + \int_A |H_N(\lambda|\lambda^*) - H(\lambda|\lambda^*)|dG(\lambda)\,.$$

But both $H_N(\lambda|\lambda^*)$ and $H(\lambda|\lambda^*)$ are bounded on $[\lambda_\ell, \lambda_r]$. Therefore

$$(2.50)\quad \int_A |H_N(\lambda|\lambda^*) - H(\lambda|\lambda^*)|dG(\lambda) \le M \int_A dG(\lambda)$$

where

$$|H_N(\lambda|\lambda^*) - H(\lambda|\lambda^*)| \le M < \infty$$

for

$$\lambda \in [\lambda_\ell, \lambda_r].$$

Since $G(\lambda)$ is continuous A is a set of G measure 0, and thus

$$M \int_A dG(\lambda) = M \cdot 0 = 0. \tag{2.51}$$

Also we have

$$\int_{[\lambda_\ell, \lambda_r] \setminus A} |H_N(\lambda|\lambda^*) - H(\lambda|\lambda^*)| dG(\lambda) \leq \varepsilon \int_{[\lambda_\ell, \ell_r] \setminus A} dG(\lambda) \leq \varepsilon . \tag{2.52}$$

Therefore from (2.48), (2.49), (2.51), and (2.52) we have

$$|E_X(\Phi_N(\lambda|\lambda^*)) - G(\lambda^*)| < \varepsilon . \tag{2.53}$$

As a generalization of the above theorem suppose the points of discontinuity of $G(\lambda)$ are known. (Actually the theorem to be proven will hold even if the discontinuity points are unknown.)

Theorem. Let $f(x|\lambda)$ be approximable and λ^* be a point of continuity of $G(\lambda)$. Then given $\varepsilon > 0$ there exists an N such that $|E_X(\Phi_N(x|\lambda^*)) - G(\lambda^*)| < \varepsilon$.

Proof. Suppose λ^* is a continuity point of $G(\lambda)$. Then given $\varepsilon > 0$ there exists a closed interval $[\lambda^*, \lambda^* + \delta]$ such that $G(\lambda)$ is continuous on $[\lambda^*, \lambda^* + \delta]$ and $G(\lambda^* + \delta) - G(\lambda^*) < \frac{\varepsilon}{2}$. Consider the function

$$(2.54)\qquad H^*(\lambda) = \begin{cases} 1, & \lambda_\ell \le \lambda \le \lambda^* \\ 1 - \frac{1}{\delta}(\lambda - \lambda^*), & \lambda^* < \lambda \le \lambda^* + \delta \\ 0, & \lambda^* + \delta < \lambda \le \lambda_r \end{cases}.$$

Choose N so that

$$(2.55)\qquad |H_N(\lambda|\lambda^*) - H^*(\lambda)| < \frac{\varepsilon}{2}, \quad \lambda \in [\lambda_\ell, \lambda_r].$$

Then

$$(2.56)\qquad |H_N(\lambda|\lambda^*) - H(\lambda|\lambda^*)| < \frac{\varepsilon}{2}, \quad \lambda \in [\lambda_\ell, \lambda^*], \lambda \in (\lambda^* + \delta, \lambda_r]$$

and

$$(2.57)\qquad |H_N(\lambda|\lambda^*) - H(\lambda|\lambda^*)| < 1 + \frac{\varepsilon}{2}, \quad \lambda \in (\lambda^*, \lambda^* + \delta].$$

Now from (2.48)

$$|E_X(\Phi_N(x|\lambda^*)) - G(\lambda^*)| \le \int_{\lambda_\ell}^{\lambda_r} |H_N(\lambda|\lambda^*) - H(\lambda|\lambda^*)| dG(\lambda).$$

Further by (2.56) and (2.57) we have

$$(2.58)\qquad \int_{\lambda_\ell}^{\lambda_r} |H_N(\lambda|\lambda^*) - H(\lambda|\lambda^*)| dG(\lambda)$$

$$= \int_{\lambda_\ell}^{\lambda^*} |H_N(\lambda|\lambda^*) - H(\lambda|\lambda^*)| dG(\lambda) + \int_{\lambda_\ell}^{\lambda^*+\delta} |H_N(\lambda|\lambda^*) - H(\lambda|\lambda^*)| dG(\lambda) +$$

$$+ \int_{\lambda^*+\delta}^{\lambda_r} |H_N(\lambda|\lambda^*) - H(\lambda|\lambda^*)| dG(\lambda)$$

$$\leq \frac{\varepsilon}{2} \int_{\lambda_\ell}^{\lambda^*} dG(\lambda) + \frac{\varepsilon}{2} \int_{\lambda^*+\delta}^{\lambda_r} dG(\lambda) + (1+\frac{\varepsilon}{2}) \int_{\lambda^*}^{\lambda^*+\delta} dG(\lambda)$$

$$= \frac{\varepsilon}{2} \int_{\lambda_\ell}^{\lambda_r} dG(\lambda) + \int_{\lambda^*}^{\lambda^*+\delta} dG(\lambda) \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon .$$

Thus by (2.48) and (2.58)

$$|E_X(\Phi_N(x|\lambda^*)) - G(\lambda^*)| \leq \varepsilon . \tag{2.59}$$

Since λ^r, $r \in I^+$, is continuous on $[\lambda_\ell, \lambda_r]$ we may approximate any r^{th} moment of λ.

We have the

Theorem. Let $f(x|\lambda)$ be approximable. Then given $\varepsilon > 0$ there exists an N such that $|E_X(\Phi_N(x|\lambda_r)) - E_\Lambda(\lambda^r)| < \varepsilon$.

Proof. In (2.14) replace λ^* by λ_r and let

$$H(\lambda|\lambda_r) = \lambda^r. \tag{2.60}$$

Thus in (2.17) we have

$$(2.61) \qquad H_n(\lambda|\lambda_r) = \sum_{i=1}^{k_n} a_{i\lambda_r}^{(n)} h_{\lambda_i^{(n)}}(\lambda) .$$

Now since $f(x|\lambda)$ is approximable there exists an N such that

$$(2.62) \qquad |H_N(\lambda|\lambda_r) - H(\lambda|\lambda_r)| < \varepsilon$$

we have

$$(2.63) \quad |E_X(\Phi_N(x|\lambda_r)) - E_\Lambda(\lambda^r)| = |\int_\Lambda H_N(\lambda|\lambda_r)dG(\lambda) - \int_\Lambda \lambda^r dG(\lambda)|$$

$$\leq \int_\Lambda |H_N(\lambda|\lambda_r) - H(\lambda|\lambda_r)|dG(\lambda)$$

$$\leq \varepsilon \int_\Lambda dG(\lambda) = \varepsilon .$$

From what has been done above it is clear that we may replace $H(\lambda|\lambda^*)$ by any function of λ, $\psi(\lambda)$, which is continuous on $[\lambda_\ell, \lambda_r]$ and thus estimate $E_\Lambda[\psi(\lambda)]$.

At (2.8) we defined

$$\bar{G}_m^n(\lambda^*) = \frac{1}{m} \sum_{\nu=1}^{m} \Phi_n(x_\nu|\lambda^*) .$$

This function is not necessarily a distribution function since it may not be monotonic increasing, nor is it necessarily non-negative

nor bounded by one. Thus it is necessary to define a function $G_m^n(\lambda^*)$ which will be a distribution function.

Let C_G = set of continuity points of $G(\lambda)$ in $[\lambda_\ell, \lambda_r]$. We have seen by a previous theorem that given $\varepsilon > 0$, $\lambda^* \in C_G$ there exists an $n(\varepsilon, \lambda^*)$ such that $|E_X(\Phi_{n(\varepsilon, \lambda^*)}(x|\lambda^*)) - G(\lambda^*)| < \varepsilon$.

Lemma. Given $\varepsilon > 0$ let

$$N_\varepsilon = \sup_{\lambda^* \in C_G} \{n(\varepsilon, \lambda^*)\}. \tag{2.64}$$

Then $N_\varepsilon < \infty$.

Proof. Suppose not. Then there exists $\lambda^* \in C_G$ such that

$$|E_X(\Phi_n(x|\lambda^*)) - G(\lambda^*)| \geq \varepsilon$$

for all $n \in I^+$. But this contradicts the completeness of

$$\left\{\frac{\beta(\lambda_n)\beta(\lambda)}{\beta(\lambda+\lambda_n)}\right\}$$

in $C[\lambda_\ell, \lambda_r]$. Thus $N_\varepsilon < \infty$. We may write

$$\overline{G}_m^{N_\varepsilon}(\lambda^*) = \frac{1}{m}\sum_{\nu=1}^{m} \Phi_{N_\varepsilon}(x_\nu|\lambda^*), \tag{2.65}$$

and

$$\left| E_X(\Phi_{N_\varepsilon}(x|\lambda^*)) - G(\lambda^*) \right| < \varepsilon$$

for all $\lambda^* \in C_G$. Now let

$$E_X(\Phi_{N_\varepsilon}(x|\lambda^*)) = F^{N_\varepsilon}(\lambda^*) . \tag{2.66}$$

Then we have the

Theorem.

$$\bar{G}_m^{N_\varepsilon}(\lambda^*) \xrightarrow{\text{a. s.}} F^{N_\varepsilon}(\lambda^*)$$

where

$$G(\lambda^*) - \varepsilon \leq F^{N_\varepsilon}(\lambda^*) \leq G(\lambda^*) + \varepsilon$$

for al $\lambda^* \in C_G$.

Proof. By the Kolmogorov strong law of large numbers

$$\bar{G}_m^{N_\varepsilon}(\lambda^*) \xrightarrow{\text{a. s.}} F^{N_\varepsilon}(\lambda^*)$$

and by the preceding theorem

$$G(\lambda^*) - \varepsilon \leq F^{N_\varepsilon}(\lambda^*) \leq G(\lambda^*) + \varepsilon.$$

Let

$$(2.67)\qquad [\bar{G}_m^{N_\varepsilon}(\lambda^*)] = \begin{cases} 0, & \bar{G}_m^{N_\varepsilon}(\lambda^*) \le 0 \\ \bar{G}_m^{N_\varepsilon}(\lambda^*), & 0 < \bar{G}_m^{N_\varepsilon}(\lambda^*) < 1 \\ 1, & \bar{G}_m^{N_\varepsilon}(\lambda^*) \ge 1 \end{cases}$$

and

$$G_m^{N_\varepsilon}(\lambda^*) = \begin{cases} 0, & \lambda^* < \lambda_\ell \\ \max\limits_{\lambda_\ell \le \theta \le \lambda^*} [\bar{G}_m^{N_\varepsilon}(\theta)], & \lambda_\ell \le \lambda^* \le \lambda_r \\ 1, & \lambda^* > \lambda_r \end{cases}$$

Clearly $G_m^{N_\varepsilon}(\lambda^*)$ a.s. $\to$ $F^{N_\varepsilon}(\lambda^*)$ and further $G_m^{N_\varepsilon}(\lambda^*)$ is a distribution function. Thus we have proved the following

Theorem. $G_m^{N_\varepsilon}(\lambda^*)$ is a distribution function such that

(2.69)

$$P\left[\lim_{m\to\infty} G_m^{N_\varepsilon}(\lambda^*) = F^{N_\varepsilon}(\lambda^*) \text{ at every continuity point of } G(\lambda^*)\right] = 1,$$

where

$$G(\lambda^*) - \varepsilon \le F^{N_\varepsilon}(\lambda^*) \le G(\lambda^*) + \varepsilon .$$

III. ESTIMATION OF DISCRETE PRIOR DISTRIBUTIONS WITH VARYING DIMENSIONAL RANDOM VECTORS

Consider the measure space (X, χ, μ) with μ σ-finite and the random variable $x \in X$ known to have one of a finite number of probability distributions

$$P_i, \quad 1 \leq i \leq r, \tag{3.1}$$

such that all P_i are absolutely continuous with respect to μ and their densities $f_i = \frac{dP_i}{d\mu}$ square integrable and bounded a.e. $\mu(x)$. Which P_i the random variable x has depends on a random parameter $\lambda \in \Lambda = \{\lambda_1, \ldots, \lambda_r\}$ which has an unknown a-priori probability vector

$$G = \{g_1, g_2, \ldots, g_r\}, \tag{3.2}$$

$$g_i \geq 0, \quad \sum_{i=1}^{r} g_i = 1$$

such that

$$P(\lambda = \lambda_i) = g_i. \tag{3.3}$$

We observe a sequence of random vectors $x_1^{(\nu_1)}, x_2^{(\nu_2)}, \ldots$ as follows: the values of λ_j are drawn independently from $G(\lambda)$ and the random sample sizes ν_j and determined by the sequence $\{x_k^{(\nu_k)}\}_{k=1}^{j-1}$. Given λ_j and ν_j, the ν_j components of $x_j^{(\nu_j)}$

are conditionally independent with common density function $f(x|\lambda_j)$, with

$$(3.4) \qquad x_j^{(\nu_j)} \in (X^{(\nu_j)}, \chi^{(\nu_j)}, \mu^{(\nu_j)})$$

the ν_j fold product space generated by ν_j fold Cartesian products of (X, χ, μ), $1 \leq \nu_j \leq N$. That is, at the jth instant we have a choice of observing any one of N random vectors

$$(3.5) \qquad x_j^{(\nu_j=1)} \in X, \quad x_j^{(\nu_j=2)} \in X \times X, \ldots,$$

$$x_j^{(\nu_j)} \in \underbrace{X \times X \times \ldots \times X}_{\nu_j \text{ fold}}, \ldots, x_j^{(\nu_j=N)} \in \underbrace{X \times \ldots \times X}_{N \text{ fold}}$$

with the dimension ν_j of the vector observed at the jth instant being determined by a rule depending on the preceding observations. The jth of these random vectors has the conditional distribution, given ν_j,

$$(3.6) \qquad P_G^{(\nu_j)}(x_j^{(\nu_j)} \in B^{(\nu_j)}) = \sum_{i=1}^{r} g_i P_i^{(\nu_j)}(B^{(\nu_j)}),$$

where $B^{(\nu_j)} \in \chi^{(\nu_j)}$ and $P_i^{(\nu_j)}(\cdot)$ is the probability distribution generated on $(X^{(\nu_j)}, \chi^{(\nu_j)}, \mu^{(\nu_j)})$ by the $P_i(\cdot)$ on (X, χ, μ).

Our problem is to construct functions

$$(3.7) \qquad g_{i,n} = g_{i,n}(x_1^{(\nu_1)}, x_2^{(\nu_2)}, \ldots, x_n^{(\nu_n)})$$

such that

$$g_{i,n} \geq 0, \quad \sum_{i=1}^{r} g_{i,n} = 1,$$

and whatever be G,

$$(3.8) \qquad P\left[\lim_{n\to\infty} g_{i,n} = g_i\right] = 1, \quad (i = 1, 2, \ldots, r).$$

Since we are dealing with random samples the components of the random vectors are independent random variables and the conditional (given ν_j) density function of $x_i^{(\nu_j)}$ is then

$$(3.9) \qquad f_i^{(\nu_j)}(x_j^{(\nu_j)}) = \prod_{k=1}^{\nu_j} f_i(x_k).$$

Now due to (3.9) and the fact that

$$\int_X f_i^2(x)d\mu(x) < \infty$$

we have

$$(3.10) \qquad \int_{X^{(\nu_j)}} [f_i^{(\nu_j)}(x_j^{(\nu_j)})]^2 d\mu^{(\nu_j)}(x^{(\nu_j)}) < \infty, \quad (i = 1, 2, \ldots, r)$$

Thus the functions $f_i^{(\nu_j)}$ are elements of the Hilbert space $H^{(\nu_j)}$ of square integrable functions over the measure space $(X^{(\nu_j)}, \chi^{(\nu_j)}, \mu^{(\nu_j)})$.

We make at this point the further assumption of identifiability, i.e., if $G = \{g_1, \ldots, g_r\}$ and $\bar{G} = \{\bar{g}_1, \ldots, \bar{g}_r\}$ are any two probability vectors such that for every $B_j^{(\nu_j)}$,

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

then $G = \bar{G}$. The class of measures P which are identifiable under such finite mixtures is a subclass of those which are identifiable under general mixtures.

Lemma 1 of Chapter II allows us to denote by $H_k^{(\nu_j)}$ the linear manifold spanned by the $r - 1$ functions $f_1^{(\nu_j)}, \ldots, f_{k-1}^{(\nu_j)}, f_{k+1}^{(\nu_j)}, \ldots, f_r^{(\nu_j)}$. We can then write uniquely

$$f_k^{(\nu_j)} = f_k^{(\nu_j)'} + f_k^{(\nu_j)''}, \quad (k = 1, 2, \ldots, r) \tag{3.11}$$

where

$$f_k^{(\nu_j)'} \in H_k^{(\nu_j)}, \quad f_k^{(\nu_j)''} \in H_k^{(\nu_j)\perp}, \quad \text{and} \quad f_k^{(\nu_j)''} \neq 0. \tag{3.12}$$

Let

$$\phi_{kj}(x^{(\nu_j)}) = \frac{f_k^{(\nu_j)''}(x^{(\nu_j)})}{\int_{X^{(\nu_j)}} [f_k^{(\nu_j)''}(x^{(\nu_j)})]^2 d\mu^{(\nu_j)}(x^{(\nu_j)})}, \tag{3.13}$$

then

$$\int_{X^{(\nu_j)}} \phi_{kj}(t) f_\ell^{(\nu_j)}(t) d\mu^{(\nu_j)} = \begin{cases} 1, & \text{if } k = \ell, \\ 0, & \text{if } k \neq \ell \end{cases} . \tag{3.14}$$

Now define

$$\bar{g}_{i,n} = \frac{1}{n} \sum_{j=1}^{n} \phi_{ij}(x_j^{(\nu_j)}) , \tag{3.15}$$

$$g_{i,n} = \frac{[\bar{g}_{i,n}]^+}{\sum_{j=1}^{r} [\bar{g}_{i,n}]^+} \cdot \quad ([a]^+ = \max(a, 0))$$

Note that any convergence properties of $\bar{g}_{i,n}$ will be shared by $g_{i,n}$. Now

$$E_{X^{(\nu_j)}} \phi_{ij}(x_j^{(\nu_j)}) = \int_{X^{(\nu_j)}} \phi_{ij}(x^{(\nu_j)}) \sum_{k=1}^{r} g_k f_k^{(\nu_j)}(x^{(\nu_j)}) d\mu^{(\nu_j)}(x^{(\nu_j)}) \tag{3.16}$$

$$= \sum_{k=1}^{r} g_k \int_{X^{(\nu_j)}} \phi_{ij}(x^{(\nu_j)}) f_k(x^{(\nu_j)}) d\mu^{(\nu_j)}(x^{(\nu_j)})$$

$$= g_i .$$

Note that whatever value of the random variable ν_j may occur the conditional expectation of $\phi_{ij}(x_j^{(\nu_j)})$ is g_i. Thus the unconditional expectation of $\phi_{ij}(x_j^{(\nu_j)})$ is g_i. Also

$$(3.17)\quad E_{X^{(\nu_j)}}\,\phi_{ij}^{2}(x_j^{(\nu_j)}) = \int_{X^{(\nu_j)}} \phi_{ij}^{2}(x^{(\nu_j)}) \sum_{k=1}^{r} g_k f_k^{(\nu_j)}(x^{(\nu_j)})\, d\mu_j^{(\nu_j)}(x^{(\nu_j)})$$

$$= \sum_{k=1}^{r} g_k \int_{X^{(\nu_j)}} \phi_{ij}^{2}(x^{(\nu_j)}) f_k^{(\nu_j)}(x^{(\nu_j)})\, d\mu_j^{(\nu_j)}(x^{(\nu_j)}).$$

By the initially assumed conditions there exists $M_k < \infty$ such that

$$(3.18)\quad 0 \leq f_k^{(\nu_j)}(x^{(\nu_j)}) \leq M_k < \infty, \quad \text{a.e.} \quad \mu_j^{(\nu_j)}(x^{(\nu_j)}),$$

therefore

$$(3.19)\quad E_{X^{(\nu_j)}}\,\phi_{ij}^{2}(x_j^{(\nu_j)}) \leq \sum_{k=1}^{r} g_k M_k \int_{X^{(\nu_j)}} \phi_{ij}^{2}\, d\mu_j^{(\nu_j)}(x^{(\nu_j)}).$$

But

$$(3.20)\quad \int_{X^{(\nu_j)}} \phi_{ij}^{2}\, d\mu_j^{(\nu_j)}(x^{(\nu_j)})$$

$$= \frac{\int_{X^{(\nu_j)}} \left[f_i^{(\nu_j)''}(x^{(\nu_j)})\right]^2 d\mu_j^{(\nu_j)}(x^{(\nu_j)})}{\int_{X^{(\nu_j)}} \left[f_i^{(\nu_j)''}(x^{(\nu_j)})\right]^2 d\mu_j^{(\nu_j)}(x^{(\nu_j)})^2}$$

$$= \frac{1}{\int_{X^{(\nu_j)}} \left[f_i^{(\nu_j)''}(x^{(\nu_j)})\right]^2 d\mu_j^{(\nu_j)}(X^{(\nu_j)})},$$

and from (3.12), we have

$$\int_{X^{(\nu_j)}} \left[f_i^{(\nu_j)''}(x^{(\nu_j)}) \right]^2 d\mu^{(\nu_j)}(x^{(\nu_j)}) > 0, \tag{3.21}$$

so that

$$\frac{1}{\int_{X^{(\nu_j)}} \left[f_i^{(\nu_j)''}(X^{(\nu_j)}) \right]^2 d\mu^{(\nu_j)}(x^{(\nu_j)})} = M_{ij}^* < \infty. \tag{3.22}$$

This produces

$$E_{X^{(\nu_j)}} \phi_{ij}^2(x^{(\nu_j)}) \leq \sum_{k=1}^{r} g_k M_k M_{ij}^* \tag{3.23}$$

and letting $M' = \max_k \{M_k\}$ and $M^* = \max_{i,j} \{M_{ij}^*\}$ we have

$$E_{X^{(\nu_j)}} \phi_{ij}^2(x^{(\nu_j)}) \leq M'M^* < \infty. \tag{3.24}$$

Now

$$\operatorname{Var} \phi_{ij}(x_j^{(\nu_j)}) \leq E_{X^{(\nu_j)}} \phi_{ij}^2(x_j^{(\nu_j)}) < \infty.$$

So

$$\sum_{j=1}^{\infty} \frac{\operatorname{Var} \phi_{ij}(x_j^{(\nu_j)})}{j^2} \leq \sum_{j=1}^{\infty} E\phi_{ij}^2(x_j^{(\nu_j)}) \leq M'M^* \sum_{j=1}^{\infty} \frac{1}{j^2} < \infty. \tag{3.25}$$

Feller (1966, p. 238) gives the theorem: Let $\{x_k\}$ be a sequence of random variables with $E(x_k | x_1, \ldots, x_{k-1}) = 0$, and let

$$S_n = \sum_{k=1}^{n} x_k .$$

If $b_1 < b_2 < \ldots \to \infty$ and

$$\sum_{k=1}^{\infty} E(x_k^2)/b_k^2 < \infty, \tag{3.26}$$

Then

$$S_n/n \xrightarrow{a.s.} 0 .$$

If we let x_k represent $\phi_{ij}(x_j^{(\nu_j)}) - g_i$ and $b_n = n$, then (3.25) implies that (3.26) is satisfied. Furthermore (3.16) implies that $E(x_k | x_1, \ldots, x_{k-1}) = 0$, therefore we have

$$\frac{1}{n} \sum_{j=1}^{n} \phi_{ij}(x_j^{(\nu_j)}) - g_i \xrightarrow{a.s.} 0 \tag{3.27}$$

and

$$\frac{1}{n} \sum_{j=1}^{n} \phi_{ij}(x_j^{(\nu_j)}) \xrightarrow{a.s.} g_i . \tag{3.28}$$

We have therefore completed the construction of consistent estimates

of the g_i.

Guthrie and Johns (1959) have studied the problem of choice of sample size in a Bayesian approach to lot-by-lot acceptance sampling. They derive asymptotic expressions for the optimal sample sizes and decision procedures as the lot size becomes large, assuming throughout that the a-priori distribution of expected lot quality is known. They study two cases: when the a-priori distribution is continuously differentiable with respect to Lebesgue measure; and when the a-priori distribution belongs to a family of atomic distributions whose support contains no point of accumulation at a certain critical point. Included in the latter class is the class of distributions over a finite number of points.

A critical parameter in the development of the asymptotic sample size is

$$A_G = (s_1 - r_1)E(\Lambda) + s_2 - r_2 + (r_1 - a_1)\int_0^C (\lambda - C)dG(\lambda)$$

where r_1, r_2, s_1, s_2, and C are given known constants.

We may use the previous development to give empirical estimates for A_G in the case where many lots are inspected, one at a time, with each sample size based on the previously estimated value of A_G. Let

$$E_n(\lambda) = \sum_{i=1}^{r} \lambda_i g_{i,n} . \tag{3.29}$$

Then clearly

$$E_n(\lambda) \xrightarrow{\text{a. s.}} E(\lambda). \tag{3.30}$$

Also

$$\int_0^C (\lambda - C)dG(\lambda) = \sum_{i=1}^{[k]} (\lambda_i - C)(g_i - g_{i-1}) \tag{3.31}$$

where $[k]$ = the greatest integer such that $\lambda_i \leq C$. So

$$\sum_{i=1}^{[k]} (\lambda_i - C)(g_{i,n} - g_{i-1,n}) \xrightarrow{\text{a. s.}} \sum_{i=1}^{[k]} (\lambda_i - C)(g_i - g_{i-1}) = \int_0^C (\lambda - C)dG(\lambda). \tag{3.32}$$

Thus we define

$$A_{G_n} = (s_1 - r_1)E_n(\lambda) + s_2 - r_2 + (r_1 - a_1) \sum_{i=1}^{[k]} (\lambda_i - C)(g_{i,n} - g_{i-1,n})$$

and we have

$$A_{G_n} \xrightarrow{\text{a. s.}} A_G. \tag{3.33}$$

If we will now let

$$\nu_n = \begin{cases} N, & A_{G_n} \leq 0 \\ K_n \ell n\, N - \dfrac{K_n}{2} \ell n\, \ell n\, N, & A_{G_n} > 0 \end{cases} . \tag{3.34}$$

(For definition of $K_{\cdot}$ see definition of K on p. 922 of the Guthrie-Johns paper), we immediately see that $\nu_n \xrightarrow{a.s.} \nu = n^*(N)$, the approximate Bayes sample size for a-priori G, omitting the remainder terms in the Guthrie-Johns expansion.

It is readily verified that their loss structure satisfies

$$0 \leq L(a, \lambda_i) \leq L < \infty, \quad \forall a \in A, \quad i = 1, 2, \ldots, r . \tag{3.35}$$

We therefore have

$$\Delta_G(a, x^{(\nu)}) = \sum_{i=1}^{r} [L(a, \lambda_i) - L(a_0, \lambda_i)] \cdot f_i^{(\nu)}(x^{(\nu)}) g_i . \tag{3.36}$$

Set

$$\Delta_n(a, x^{(\nu_n)}) = \sum_{i=1}^{r} [L(a, \lambda_i) - L(a_0, \lambda_i)] \cdot f_i^{(\nu_n)}(x^{(\nu_n)}) g_{i,n} , \tag{3.37}$$

so that

$$\sup_A |\Delta_n(a, x^{(\nu_n)}) - \Delta_G(a, x^{(\nu)})| \leq L \sum_{i=1}^{r} |f_i^{(\nu_n)}(x^{(\nu_n)}) g_{i,n} - f_i^{(\nu)}(x^{(\nu)}) g_i| . \tag{3.38}$$

Since $\nu_n \to \nu$, $f_i^{(\nu_n)}(x^{(\nu_n)})$ and $f_i^{(\nu)}(x^{(\nu)})$ are $< \infty$ a.e., we see that

$$p \lim_{n \to \infty} \sup_A |\Delta_n(a, x^{(\nu_n)}) - \Delta_G(a, x^{(\nu)})| = 0 \tag{3.39}$$

Thus our sequence of decision functions is asymptotically optimal in the sense of Robbins, hence the empirical acceptance sampling procedure so generated is asymptotically equivalent to that for known a-priori distributions, omitting remainder terms.

BIBLIOGRAPHY

Cheney, E. W. 1966. Introduction to approximation theory. New York, McGraw-Hill. 259 p.

Davis, P. J. 1963. Interpolation and approximation. New York, Blaisdell. 385 p.

Feller, William. 1966. An introduction to probability theory and its applications. Vol. 2. New York, Wiley. 626 p.

Guthrie, D., Jr. and M. V. Johns, Jr. 1959. Bayes acceptance sampling procedures for large lots. Annals of Mathematical Statistics 30:896-925.

Johns, M. V., Jr. 1957. Non-parametric empirical Bayes procedures. Annals of Mathematical Statistics 28:649-669.

Lehmann, E. L. 1959. Testing statistical hypotheses. New York, Wiley. 369 p.

Loeve, M. 1960. Probability theory. 2d ed. Princeton, Van Nostrand. 685 p.

Robbins, Herbert. 1956. An empirical Bayes approch to statistics. In: Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, 1955. Vol. 1. Berkeley, University of California. p. 157-163.

_____ 1964. The empirical Bayes approach to statistical decision problems. Annals of Mathematical Statistics 35:1-20.

Rolph, John E. 1968. Bayesian estimation of mixing distributions. Annals of Mathematical Statistics 39:1289-1302.

Samuel, Esther. 1963. An empirical Bayes approach to the testing of certain parametric hypotheses. Annals of Mathematical Statistics 34:1370-1385.

Teicher, Henry. 1960. On the mixtures of distributions. Annals of Mathematical Statistics 31:55-73.

_____ 1961. Identifiability of mixtures. Annals of Mathematical Statistics 32:244-248.

_____ 1963. Identifiability of finite mixtures. Annals of Mathematical Statistics 34:1265-1269.

Tucker, Howard G. 1967. A graduate course in probability. New York, Academic. 273 p.

Waterman, Michael Spencer. 1966. The exponential family of probability distributions generated by σ-infinite measures. Master's thesis. Corvallis, Oregon State University. 50 numb. leaves.