The Annals of Mathematical Statistics
1968, Vol. 39, No. 4, 1289–1302

BAYESIAN ESTIMATION OF MIXING DISTRIBUTIONS[1]

By John E. Rolph[2]

University of California, Berkeley and University College, London

1. Introduction and summary. Let $\mathcal{Q} = \{Q(t): 0 \leqq t \leqq 1\}$ be a family of probability distributions on the positive integers parameterized on [0, 1], that is

$$P_t(X = x) = q_x(t); \qquad x = 1, 2, \cdots. \tag{1}$$

If G is a distribution on [0, 1] the distribution of X is a G-mixture over $\mathcal{Q}$ if

$$P_G(X = x) = \int_0^1 q_x(t)\, dG(t) = q_x(G). \tag{2}$$

G is called the mixing distribution. It is assumed at the outset that the family $\mathcal{Q}$ is known to be identifiable, that is if $q_x(G_1) = q_x(G_2)$ for $x = 1, 2, \cdots$, then $G_1 = G_2$. See [12] and [13] for conditions insuring identifiability. Thus it makes sense to attempt to estimate G when one has independent observations on X. Some work on estimating G has been done when the mixture is finite [4], [2], [9] and for special $\mathcal{Q}$'s [6], [14]. The problem is of interest not only in an estimation context, but also in the construction of empirical Bayes decision procedures [9].

Our approach is to define a prior distribution on possible values of G and then construct consistent Bayes estimates of G from the posterior distribution. Section 2 gives the needed background on moment spaces, sets up the prior distribution and derives the posterior distribution. In Section 3, the Bayes estimates are defined while Section 4 proves the consistency of the posterior distribution and thus of the estimates. Here, Theorem 1 is not directly applicable to our problem, but is included because of its possible independent interest. Sections 5 and 6 generalize the earlier results.

2. The prior distribution on moment space. We begin by assuming that the frequency function given the parameter t is a polynomial in t. So $q_x(t) = \sum_{i=0}^{k_x} a_{xi} t^i$. More general $q_x(t)$ are considered in Section 5. Identifiability clearly places restrictions on k_x and the a_{xi}. The negative binomial distribution with parameters t and r which arises when waiting for the rth success in Bernoulli trials is a common identifiable situation. Here

$$q_x(t) = \binom{x-1}{r-1} t^r (1-t)^{x-r} = \sum_{i=r}^{x} \binom{x-1}{r-1}\binom{x-r}{i-r}(-1)^{i-r} t^i$$

for $x \geqq r$. So $a_{xi} = 0$ for $i < r$ and $a_{xi} = \binom{x-1}{r-1}\binom{x-r}{i-r}(-1)^{i-r}$ for $r \leqq i \leqq x$, thus $k_x = x$.

Received 20 February 1967; revised 18 December 1967.

[1] This paper is a revised version of the author's doctoral dissertation at the University of California, Berkeley, and was written with the partial support of the National Science Foundation (Grants GP-2593 and GP-5059).

[2] Now at Columbia University, New York.

If G is the mixing distribution on [0, 1],

$$P_G(X = x) = \int_0^1 \left(\sum_{i=0}^{k_x} a_{xi} t^i\right) dG(t) = \sum_{i=0}^{k_x} a_{xi} m_i(G) \tag{3}$$

where $m_i(G) = \int_0^1 t^i \, dG(t)$ is the ith moment of G.

The parameter set $\mathcal{G}$ is the set of distribution functions on [0, 1]. To calculate Bayes estimates, we must put a prior probability distribution on $\mathcal{G}$ and then compute the posterior distribution on $\mathcal{G}$ given the observations. The polynomial form of $q_x(t)$ suggests putting the prior distribution on sequences of possible moments of a distribution on [0, 1]. It is well known that the moments of a distribution on [0, 1] determine it uniquely, so a prior distribution on possible moment sequences induces a distribution on $\mathcal{G}$.

We need some preliminaries on moment sequences to define such a prior. Let D be the subset of the infinite dimensional unit cube $[0, 1]^\infty$ whose elements are possible moment sequences and let $m_1, m_2, \cdots$ be a sequence of real numbers. Define the Hankel determinants for $k = 1, 2, \cdots$ by

$$\underline{\Delta}_{2k} = \begin{vmatrix} 1 & m_1 & \cdots & m_k \\ m_1 & m_2 & \cdots & m_{k+1} \\ \vdots & \vdots & & \vdots \\ m_k & m_{k+1} & \cdots & m_{2k} \end{vmatrix}, \tag{4}$$

$$\underline{\Delta}_{2k+1} = \begin{vmatrix} m_1 & m_2 & \cdots & m_{k+1} \\ m_2 & m_3 & \cdots & m_{k+2} \\ \vdots & \vdots & & \vdots \\ m_{k+1} & m_{k+2} & \cdots & m_{2k+1} \end{vmatrix},$$

$$\bar{\Delta}_{2k} = \begin{vmatrix} m_1 - m_2 & m_2 - m_3 & \cdots & m_k - m_{k+1} \\ m_2 - m_3 & m_3 - m_4 & \cdots & m_{k+1} - m_{k+2} \\ \vdots & \vdots & & \vdots \\ m_k - m_{k+1} & m_{k+1} - m_{k+2} & \cdots & m_{2k-1} - m_{2k} \end{vmatrix},$$

$$\bar{\Delta}_{2k+1} = \begin{vmatrix} 1 - m_1 & m_1 - m_2 & \cdots & m_k - m_{k+1} \\ m_1 - m_2 & m_2 - m_3 & \cdots & m_{k+1} - m_{k+2} \\ \vdots & \vdots & & \vdots \\ m_k - m_{k+1} & m_{k+1} - m_{k+2} & \cdots & m_{2k} - m_{2k+1} \end{vmatrix},$$

where k is a non-negative integer and set $\underline{\Delta}_{-1} = \bar{\Delta}_{-1} = \underline{\Delta}_0 = \bar{\Delta}_0 = 1$. Let D^N be the projection of D onto its first N coordinates so if $m = (m_1, m_2, \cdots) \varepsilon D$ then $m^N = (m_1, \cdots, m_N) \varepsilon D^N$. Necessary and sufficient conditions that an N-vector m^N be in D^N are that $\underline{\Delta}_i \geqq 0$ and $\bar{\Delta}_i \geqq 0$ for all i, $1 \leqq i \leqq N$ (see [7] or 11]). This implies that if $(m_1, \cdots, m_n)$ are the first n moments of a distribution function, then m_{n+1} is a possible n + 1st moment if

$$\underline{m}_{n+1}(m_1, \cdots, m_n) \leqq m_{n+1} \leqq \bar{m}_{n+1}(m_1, \cdots, m_n) \tag{5}$$

where $\underline{m}_{n+1} = m_{n+1} - (\underline{\Delta}_{n+1}/\underline{\Delta}_{n-1})$ and $\bar{m}_{n+1} = m_{n+1} + (\bar{\Delta}_{n+1}/\underline{\Delta}_{n-1})$.
It is easy to check that the m_{n+1} in $\underline{m}_{n+1}$ and $\bar{m}_{n+1}$ cancels so that these bounds are actually only a function of $(m_1, \cdots, m_n)$. The first few bounds are

$$0 \leqq m_1 \leqq 1, \quad m_1^2 \leqq m_2 \leqq m_1, \quad m_2^2/m_1 \leqq m_3 \leqq m_2 - (m_1 - m_2)^2/(1 - m_1).$$

Thus for n a non-negative integer,

$$(6) \qquad \bar{m}_{n+1} - \underline{m}_{n+1} = \bar{\Delta}_{n+1}/\bar{\Delta}_{n-1} + \underline{\Delta}_{n+1}/\underline{\Delta}_{n-1} = d_{n+1}(m_1, \cdots, m_n).$$

To characterize distributions having given moments, we define the polynomials $\underline{\Delta}_n(t)$ and $\bar{\Delta}_n(t)$ analogously to $\underline{\Delta}_n$ and $\bar{\Delta}_n$ by replacing the last column by the vector $(1, t, \cdots, t^m)$ with appropriate m. That is

$$(7) \qquad \underline{\Delta}_{2k}(t) = \begin{vmatrix} 1 & m_1 & \cdots & m_{k-1} & 1 \\ m_1 & m_2 & \cdots & m_k & t \\ \vdots & \vdots & & \vdots & \vdots \\ m_k & m_{k+1} & \cdots & m_{2k-1} & t^k \end{vmatrix},$$

$$\underline{\Delta}_{2k+1}(t) = \begin{vmatrix} m_1 & m_2 & \cdots & m_k & 1 \\ m_2 & m_3 & \cdots & m_{k+1} & t \\ \vdots & \vdots & & \vdots & \vdots \\ m_{k+1} & m_{k+2} & \cdots & m_{2k} & t^k \end{vmatrix},$$

$$\bar{\Delta}_{2k}(t) = \begin{vmatrix} m_1 - m_2 & m_2 - m_3 & \cdots & m_{k-1} - m_k & 1 \\ m_2 - m_3 & m_3 - m_4 & \cdots & m_k - m_{k+1} & t \\ \vdots & \vdots & & \vdots & \vdots \\ m_k - m_{k+1} & m_{k+1} - m_{k+2} & \cdots & m_{2k} - m_{2k-1} & t^{k-1} \end{vmatrix},$$

$$\bar{\Delta}_{2k}(t) = \begin{vmatrix} 1 - m_1 & m_1 - m_2 & \cdots & m_{k-1} - m_k & 1 \\ m_1 - m_2 & m_2 - m_3 & \cdots & m_k - m_{k+1} & t \\ \vdots & \vdots & & \vdots & \vdots \\ m_k - m_{k+1} & m_{k+1} - m_{k+2} & \cdots & m_{2k-1} - m_{2k} & t^k \end{vmatrix}.$$

Define the polynomials $\underline{P}_n(t)$ and $\bar{P}_n(t)$ by

$$(8) \qquad \begin{aligned} \underline{P}_n(t) &= \underline{\Delta}_n(t)^2, && \text{if } n \text{ even}, \\ &= t\underline{\Delta}_n(t)^2, && \text{if } n \text{ odd}; \\ \bar{P}_n(t) &= t(1-t)\bar{\Delta}_n(t)^2, && \text{if } n \text{ even}, \\ &= (1-t)\bar{\Delta}_n(t)^2, && \text{if } n \text{ odd}. \end{aligned}$$

A point $m^N \varepsilon D^N$ is called a boundary point if $m_i = \bar{m}_i$ or $m_i = \underline{m}_i$ for some i, $1 \leqq i \leqq N$. Define $\bar{m}^{N+1} = (m_1, \cdots, m_N, \bar{m}_{N+1})$ and $\underline{m}^{N+1} = (m_1, \cdots, m_n, \underline{m}_{N+1})$. Let $\bar{G}_{N+1}$ and $\underline{G}_{N+1}$ be distribution functions having $\bar{m}^{N+1}$ and $\underline{m}^{N+1}$ as moments. The following facts are found in [7]. Points in D^N which correspond to unique distributions in $\mathcal{G}$ are precisely the boundary points. If m^N is not a boundary point in D^N then the $\bar{G}_{N+1}$ and $\underline{G}_{N+1}$ derived from m^N are unique finite valued distribution functions whose steps occur at the roots of the polynomials $\bar{P}_{N+1}(t)$ and $\underline{P}_{N+1}(t)$ respectively. Clearly any point m^N in D^N can be represented as a convex combination of $\bar{m}^N$ and $\underline{m}^N$ so that a distribution function which has m^N as it first N moments is

$$(9) \qquad G_N = [(\bar{\Delta}_N/\bar{\Delta}_{N-2})\underline{G}_N + (\underline{\Delta}_N/\underline{\Delta}_{N-2})\bar{G}_N][(\bar{\Delta}_N/\bar{\Delta}_{N-2}) + (\underline{\Delta}_N/\underline{\Delta}_{N-2})]^{-1}$$

where the Δ's are defined in (4). The set of distribution functions with first N

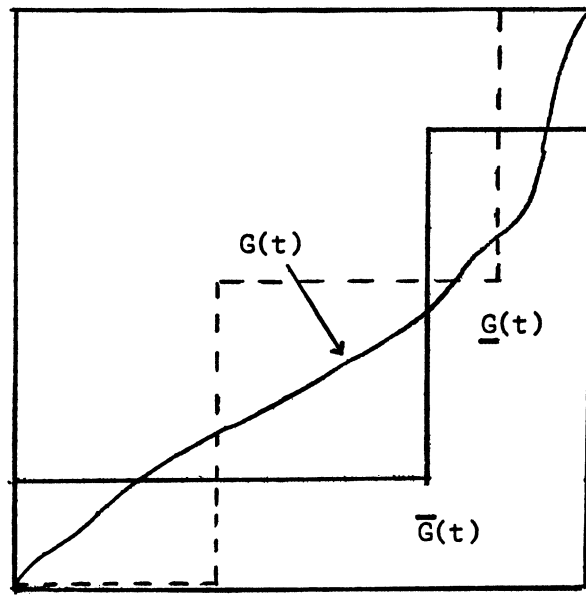

FIG. 1

moments $m^N = (m_1, \cdots, m_N)$ can be characterized through $\bar{G}_{N+1}$ and $\underline{G}_{N+1}$. If $\underline{G}_{N+1}$ and $\bar{G}_{N+1}$ are distinct their spectra form strictly interlocking sets in [0, 1] and the sets of values of $\underline{G}_{N+1}$ and $\bar{G}_{N+1}$ are strictly interlocking. Thus $\underline{G}_{N+1}$ and $\bar{G}_{N+1}$ control the shape of any G having m^N as moments by the requirement that $G(x)$ must cross every step of $\underline{G}_{N+1}(x)$ and $\bar{G}_{N+1}(x)$.

We now put a prior distribution on the moment space D. A uniform type of prior is used here to make formulae simpler, but the consistency results will hold for more general priors (see Section 6). The distribution ν is defined on D via the coordinates. The density ν_1 on the first coordinate with respect to Lebesgue measure is

$$\nu_1(m) = 1, \qquad 0 \leqq m \leqq 1, \tag{10}$$
$$= 0, \qquad \text{elsewhere,}$$

and on the kth given $(m_1, \cdots, m_{k-1})$ is

$$\nu_k(m \mid m_1, \cdots, m_{k-1}) = (\bar{m}_k - \underline{m}_k)^{-1}, \qquad \underline{m}_k \leqq m \leqq \bar{m}_k,$$
$$= 0, \qquad \text{elsewhere.}$$

The first N moments map $\mathcal{G}$ onto D^N. Let $\mathcal{B}_D{}^N$ be the Borel sigma field on D^N in the product topology. We derive the posterior distribution on $(D^N, \mathcal{B}_D{}^N)$ and appeal to the Kolmogorov extension theorem to define it on $(D, \mathcal{B}_D)$ with $\mathcal{B}_D$ being the Borel sigma field on D. In Section 4 it is shown that $(D^N, \mathcal{B}_D{}^N)$ transforms into the appropriate topology on $\mathcal{G}$ itself. If $X_1(\omega), \cdots, X_n(\omega)$

are independent random variables defined on a probability pair $(\Omega, \mathfrak{F})$, let $n_j(\omega)$ be the number of $X_i(\omega) = j$. Our sample can be written $(n_1, n_2, \cdots, n_c, 0, \cdots)$ where c is the largest observation in the sample of size n.

The joint frequency function of $(n_1, n_2, \cdots)$ given n and $m_1, \cdots, m_N$ for $N \geqq K = \max_{\{x: n_x > 0\}} (k_x)$ is

$$(11) \quad f(n_1, \cdots, n_c, \cdots \mid n, m_1, \cdots, m_N) = \binom{n}{n_1 \cdots n_c} \prod_{x=1}^{c} \left(\sum_{i=0}^{k_x} a_{xi} m_i\right)^{n_x}.$$

Letting ν^N be the marginal density of ν on D^N

$$(12) \quad \nu^N(m_1, \cdots, m_N) = \prod_{i=1}^{N} (d_i(m_1, \cdots, m_{i-1}))^{-1}, \quad \text{for } m^N \varepsilon D^N$$
$$= 0, \quad \text{for } m^N \varepsilon D^N$$

where d_i is defined in (6). The posterior density of m^N given $(n_1, \cdots, n_c, \cdots)$ on D^N for $N \geqq K$ is

$$(13) \quad g^N(m_1, \cdots, m_N \mid n_1, \cdots, n_c, \cdots)$$
$$= \prod_{x=1}^{c} \left(\sum_{i=1}^{k_x} a_{xi} m_i\right)^{n_x} \prod_{i=1}^{N} d_i^{-1} / (I(n_1, \cdots, n_c, \cdots))$$

where

$$I(n_1, \cdots, n_c, \cdots)$$
$$= \int_{D^N} \prod_{x=1}^{c} \left(\sum_{i=1}^{k_x} a_{xi} m_i\right)^{n_x} \prod_{i=1}^{N} (d_i(m_1, \cdots, m_{i-1}))^{-1} \, dm_1, \cdots, dm_N.$$

This is a posterior distribution on $(D^N, \mathfrak{B}_D{}^N)$ and hence $(D, \mathfrak{B}_D)$.

3. Bayes estimates. Let $L(G, G_0)$ be the loss incurred by using G as an estimate when G_0 is the true parameter value (mixing distribution). The Bayes estimate $\hat{G}(X)$ of G_0 is usually defined as the parameter value which minimizes the Bayes risk $R(\hat{G})$.

$$(14) \quad R(\hat{G}(X)) = \int_{\mathcal{G}} \left[\sum_X L(\hat{G}(X), G_0) q_X(G_0)\right] d\mu(G_0)$$

where μ is the prior distribution on $\mathcal{G}$. By taking the summation outside it is easily seen that this is equivalent to minimizing the expected loss under the posterior distribution given the sample. Say the loss function is

$$(15) \quad L(\hat{G}, G_0) = \sum_{i=1}^{\infty} \lambda_i [m_i(\hat{G}) - m_i(G_0)]^2$$

where $\lambda_i \geqq 0$ and $\sum \lambda_i$ is finite. Then the Bayes estimate of G_0 is just the expectation of the posterior distribution. The Bayes estimate $\hat{G}$ is determined by taking the distribution function with $(\hat{m}_1, \hat{m}_2, \cdots)$ as its moments where

$$(16) \quad \hat{m}_j = \int_D m_j \prod_{i=1}^{n} q_{X_i}(m) \, d\nu(m) / \int_D \prod_{i=1}^{N} q_{X_i}(m) \, d\nu(m)$$

is the expected jth moment under the posterior distribution. Here we think of $q_x(\cdot)$ as a function of m, the moment sequence of G. If on the other hand the summation in (15) stops at N, then any moments of $\hat{G}$ of higher order than N are immaterial to minimizing the risk. This idea helps motivate our estimates.

Recall that K is the maximal degree of $q_x(t)$ for the x values which occur. To

get any of the posterior expected moments, we must integrate over at least the first K moments. Thus $(\hat{m}_1, \cdots, \hat{m}_k)$ is a convenient quantity to base our estimates on. It is given by (see (13))

$$(17)\quad (\hat{m}_2, \cdots, \hat{m}_k) = \int_{D^K} (m_1, \cdots, m_K) \cdot g^K(m_1, \cdots, m_K \mid n_1, \cdots, n_c, \cdots)\, dm_1, \cdots, dm_K .$$

To get our estimate we calculate the two boundary distributions $\bar{G}_K$ and $\underline{G}_K$ having $(\hat{m}_1, \cdots, \hat{m}_{K-1})$ as their moments (see Figure 1). Then define our estimate $\hat{G}$ as the unique convex combination of $\bar{G}_K$ and $\underline{G}_K$ which has $\hat{m}_K$ for its Kth moment (9). The estimate is the finite valued distribution function

$$(18)\qquad \hat{G}(X_1, \cdots, X_n) = G^K(X_1, \cdots, X_n) = \alpha \underline{G}_K + (1 - \alpha)\bar{G}_K$$

where

$$\alpha = (\bar{\Delta}_K/\bar{\Delta}_{K-2})/[(\bar{\Delta}_K/\bar{\Delta}_{K-2}) + (\underline{\Delta}_K/\underline{\Delta}_{K-2})].$$

Note that $m_i(\hat{G}) = \hat{m}_i$ for $i \leqq K$. This estimate has the advantage of being easy to calculate and since identifiability insures $K \to \infty$ as $n \to \infty$, it becomes progressively smoother as the number of observations increases. One can gain a better moment fit by using $G_N(X_1, \cdots, X_n)$ where $N > K$. If the smoothness of estimate is important it might be worth taking a larger N like $2K$ or $3K$. In Section 4 the consistency of the posterior distribution and consequent consistency of $\hat{G}$ will be precisely stated and proved.

EXAMPLES. Let X have a mixed geometric distribution.

$$P_G(X = x) = \int_0^1 (1 - t)t^{x-1}\, dG(t) = m_{x-1}(G) - m_x(G).$$

Let $I_c(t) = I_{\{t \geqq c\}}(t)$, the function equal to one on the set $\{t\colon t \geqq c\}$ and zero outside it. We calculate the estimate in two simple cases:

(a) $n = 1, X_1 = 1$. Here $\hat{m}_1 = \frac{1}{3}, \underline{\Delta}_1 = m_1, \bar{\Delta}_1 = 1 - m_1, \underline{\Delta}_1(t) = \bar{\Delta}_1(t) = 1$, $\underline{P}_1(t) = t, \bar{P}_1(t) = 1 - t, \underline{G}_1(t) = I_0(t), \bar{G}_1(t) = I_1(t)$; then $\hat{G}_1 = \frac{2}{3}I_0 + \frac{1}{3}I_1$.

(b) $n = 2, X_1 = 1, X_2 = 2$. Here $\hat{m}_1 = (\frac{3}{7}), \hat{m}_2 = (\frac{2}{7}), \underline{\Delta}_2 = m_2 - m_1^2$, $\bar{\Delta}_2 = m_1 - m_2, \underline{\Delta}_2(t) = t - m_1, \bar{\Delta}_2(t) = 1, \underline{P}_2(t) = (t - m_1)^2, \bar{P}_2(t) = t(1 - t)$, $\underline{G}_2(t) = I_{3/7}(t), \bar{G}_2(t) = (\frac{4}{7})I_0(t) + (\frac{3}{7})I_1(t)$; then $\hat{G}_2 = (\frac{5}{21})I_0 + (\frac{7}{12})I_{3/7} + (\frac{5}{28})I_1$.

4. Consistency. A desirable property for estimators to possess is consistency. This means that as the number of observations becomes large, the estimator converges to the true parameter value with probability one for any possible value of the parameter. To show consistency of our estimates, the consistency of the posterior distribution is first proved. We turn to a more general setting for this proof. Since $q_x(G) = \sum_{i=0}^{k_x} a_{xi} m_i(G)$, denote by $Q(G)$ this probability distribution on the positive integers. Viewed in this way a prior distribution has been put on the set of probability distributions Λ living on I, the positive integers. As a first step it will be shown that the posterior distribution on Λ converges to $Q(G)$ with probability 1. If S is the space of functions from I to $[0, 1]$ in the product topology let $L = \{\lambda\colon \lambda \in S, \sum \lambda_i \leqq 1\}$ in the relative topology so

$\Lambda = \{\lambda: \lambda \varepsilon L, \sum \lambda_i = 1\}$. Use μ to denote a probability on $\mathfrak{B}$ the Borel sigma field of L. It is convenient for technical reasons to define μ on L instead of just Λ. Suppose $\{X_n : n = 1, 2, \cdots\}$ is a sequence of I valued random variables on $(\Omega, \mathfrak{F})$ which are independent with common distribution $P_\lambda\{\omega: \omega \varepsilon \Omega, X_n(\omega) = i\} = \lambda_i$, $i \varepsilon I$, then the posterior distribution $\mu_{n,\omega}$ of λ given $X_1(\omega), \cdots, X_n(\omega)$ is

$$\int_A \prod_{j=1}^n \lambda[X_j(\omega)]\mu(d\lambda) / \int_L \prod_{j=1}^n \lambda[X_j(\omega)]\mu(d\lambda) \tag{19}$$

for $A \varepsilon \mathfrak{B}$.

The weak* topology is used on the space of probabilities on $\mathfrak{B}$ so that $\mu_n \to \mu$ means $\int_L f\, d\mu_n \to \int_L f\, d\mu$ for every continuous function f on L. Let δ_λ be a point mass at λ.

DEFINITION. If μ is a probability on $\mathfrak{B}$ and $\lambda \varepsilon \Lambda$, the pair (λ, μ) is consistent if and only if $\mu_{n,\omega} \to \delta_\lambda$ for P_λ-almost all ω. That is, the $\mu_{n,\omega}$ measure of every L-neighborhood of λ converges to 1 for all but a P_λ-null set of ω.

Theorem 1 is the same as Theorem 2 of Freedman (1963) with the hypothesis $-\sum p_i \log p_i < \infty$ dropped. The notation here follows Freedman's as closely as possible.

The relative entropy between two distribution functions F and G is defined as $I(F, G) = \int_{-\infty}^{\infty} \log (dF/dG)\, dF$ with $\log 0 = -\infty$, $0(-\infty) = 0$, so $0 \leqq I(F, G) \leqq \infty$. Here $I(p, \lambda) = \sum_{i=1}^{\infty} p_i \log (p_i/\lambda_i)$.

THEOREM 1. *Let μ be a probability on $\mathfrak{B}$ such that for any $\epsilon > 0$, $\mu\{\lambda: I(p, \lambda) < \epsilon\} > 0$. Then (p, μ) is consistent.*

PROOF. Let $I_+ = \{i: i \varepsilon I \text{ and } p_i > 0\}$. Define S_+, L_+, Λ_+, and $\mathfrak{B}_+$ as before with I replaced by I_+; p^+ and λ^+ are the projections of p and λ onto I_+. Denote by ν and $\nu_{n,\omega}$ the projections of μ and $\mu_{n,\omega}$ on $\mathfrak{B}_+$. Emunerate I so that $I_+ = \{1, 2, \cdots, N\}$ if I_+ is finite and $I_+ = \{1, 2, \cdots\}$ if it is infinite. We now need

LEMMA 1. *Let $P_n = \sum_{i=1}^n p_i$ and $\Lambda_n = \sum_{i=1}^n \lambda_i$. If $\sup_{1 \leqq n \leqq \infty} |P_n - \Lambda_n| \geqq \epsilon$ then $I(p, \lambda) \geqq 2\epsilon^2 + (\frac{4}{9})\epsilon^4 + o(\epsilon^4) = \epsilon'$.*

PROOF. We transform our situation to the uniform distribution on [0, 1] and apply a slight modification of a theorem of Abrahamson (1965). The distribution functions of p and λ are $P(x) = \sum_{i=1}^{[x]} p_i$ and $\Lambda(x) = \sum_{i=1}^{[x]} \lambda_i$ where $[x]$

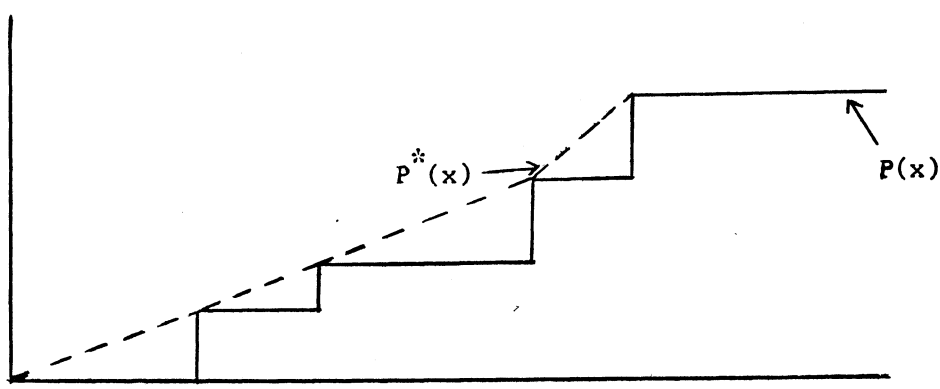

FIG. 2

is the greatest integer $\leqq x$. Define the distribution function $P^*(x)$ by $P^*(x) = ([x] + 1 - x)P([x]) + (x - [x])P([x] + 1)$ and $\Lambda^*(x)$ likewise with $p^*(x)$ and $\lambda^*(x)$ being the densities. Then $I(p, \lambda) = I(P^*, \Lambda^*)$ since

$$\begin{aligned} I(P^*, \Lambda^*) &= \int_0^\infty p^*(x) \log (p^*(x)/\lambda^*(x))\, dx \\ &= \sum_{i=1}^\infty \int_{i-1}^i p^*(x) \log (p^*(x)/\lambda^*(x))\, dx \\ &= \sum_{i=1}^\infty p_i \log (p_i/\lambda_i) \\ &= I(p, \lambda). \end{aligned}$$

Thus $I(p, \lambda) = \int_0^\infty p^*(x) \log (p^*(x)/\lambda^*(x))\, dx$. Making the transformation $y = \Lambda^*(x)$ yields

$$I(p, \lambda) = \int_0^1 p^*(\Lambda^{*-1}(y))[\lambda^*(\Lambda^{*-1}(y))]^{-1} \log p^*(\Lambda^{*-1}(y))[\lambda^*(\Lambda^{*-1}(y))]^{-1}\, dy.$$

But $\int_0^1 p^*(\Lambda^{*-1}(y))[\lambda^*(\Lambda^{*-1}(y))]^{-1}\, dy = 1$ so

$$h(y) = p^*(\Lambda^{*-1}(y))[\lambda^*(\Lambda^{*-1}(y))]^{-1}$$

is a density with distribution function H. Now if U_0 is the uniform distribution on [0, 1] and H is a continuous distribution on [0, 1] with density h then

$$I(H, U_0) = \int_0^1 h(y) \log h(y)\, dy = I(p, \lambda).$$

So we apply

LEMMA 2. (Abrahamson, pp. 28–33).

$$\inf_F \{I(F, U_0) : \sup_{0 \leqq x \leqq 1} |x - F(x)| \geqq \epsilon\} = 2\epsilon^2 + (\tfrac{4}{3})\epsilon^4 + o(\epsilon^4) = \epsilon'.$$

Thus

$$\inf_\lambda \{I(p, \lambda) : \sup_{0 \leqq x < \infty} |P(x) - \Lambda(x)| \geqq \epsilon\} \geqq \inf_H \{I(H, U_0) : \sup_{0 \leqq x \leqq 1} |x - H(x)| \geqq \epsilon\} = \epsilon'.$$

The proof of Lemma 1 is complete

LEMMA 3. *Given* $\epsilon > 0$, $\sup_i |p_i - \lambda_i| \geqq \epsilon$ *implies* $\sup_n |P_n - \Lambda_n| \geqq \epsilon/2$.

PROOF. Assume contrary. $|P_n - \Lambda_n| < \epsilon$ for all n. So $|p_i - \lambda_i| = |P_i - \Lambda_i - (P_{i-1} - \Lambda_{i-1})| \leqq |P_i - \Lambda_i| - |P_{i-1} - \Lambda_{i-1}| < \epsilon/2 + \epsilon/2 = \epsilon$; a contradiction.

Moving to the proof of Theorem 1, fix a small $\epsilon > 0$ and let

$$V = \{\lambda^+ : \sup_{i \varepsilon I_+} |\lambda_i - p_i| < 2\epsilon\}.$$

The aim is to show that $\lim_{n \to \infty} \nu_{n,\omega}(V) = 1$ for a.e. $[P_p]\omega$.

LEMMA 4. *If* $\lambda^+ \varepsilon L_+ - V$, *then for sufficiently large* n,

$$\prod_{i=1}^n \lambda[X_i(\omega)]\{p[X_i(\omega)]\}^{-1} e^{n\epsilon'/2} \leqq 1 \quad \textit{for} \quad \text{a.e.} \quad [P_p]\omega. \tag{20}$$

PROOF. Let $Z_i = \lambda(X_i)/p(X_i)$, $\lambda^+ \varepsilon L_+ - V$. Lemma 3 implies

$$\sup_n |P_n - \Lambda_n| \geqq \epsilon;$$

so by Lemma 1, $E(\log Z) = -I(p, \lambda) \leqq -\epsilon'$. By SLLN for a.e. $[P_p]\omega$.

$$\lim_{n\to\infty} \sup n^{-1} \sum_{i=1}^{n} \log Z_i(\omega) \leqq -\epsilon' \tag{21}$$

which implies (20) holds eventually for a.e. $[P_p]\omega$ proving the lemma.

Here the null set of ω's depends on λ. By Fubini's theorem applied to $(\Omega \times L_+, \mathfrak{F} \times \mathfrak{B}_+)$ there is a P_p-null set $E \in \mathfrak{F}$ such that for $\omega \notin E$, (21) and therefore (20) hold for all but a ν null set of λ's in $L_+ - V$. This set of λ's depend on ω.

Integrating (20) over $L_+ - V$ for $\omega \notin E$ yields

$$\int_{L_+-V} \lim \sup_{n\to\infty} \{\prod_{i=1}^{n} \lambda(X_i)[p(X_i)]^{-1} e^{n\epsilon'/2}\} \nu(d\lambda) \leqq \nu(L_+ - V).$$

By Fatou

$$\lim \sup_{n\to\infty} \int_{L_+-V} \prod_{i=1}^{n} \lambda(X_i)[p(X_i)]^{-1} e^{n\epsilon'/2} \nu(d\lambda) \leqq \nu(L_+ - V)$$

for $\omega \notin E$. Thus for sufficiently large n

$$\int_{L_+-V} \prod_{i=1}^{n} \lambda(X_i)(p(X_i))^{-1} \nu(d\lambda) \leqq e^{-n\epsilon'/3} \nu(L_+ - V). \tag{22}$$

Turning to posterior distributions

$$\begin{aligned} \nu_{n,\omega}(L_+ - V)/\nu_{n,\omega}(V) &= \int_{L_+-V} \prod_{i=1}^{n} \lambda[X_i(\omega)]\nu(d\lambda) [\int_V \prod_{i=1}^{n} \lambda[X_i(\omega)]\nu(d\lambda)]^{-1} \\ &= \int_{L_+-V} \prod_{i=1}^{n} \lambda[X_i(\omega)](p[X_i(\omega)])^{-1} \nu(d\lambda) \\ &\quad \cdot [\int_V \prod_{i=1}^{n} \lambda[X_i(\omega)](p[X_i(\omega)])^{-1} \nu(d\lambda)]^{-1}. \end{aligned}$$

Let $V_0 = \{\lambda^+: \lambda^+ \in V \text{ and } I(p, \lambda) \leqq \epsilon'/8\} = \{\lambda^+: I(p, \lambda) \leqq \epsilon'/8\}$ by Lemmas 1 and 3. By hypothesis $\nu(V_0) > 0$. Using a similar argument to the above,

$$\int_{V_0} \prod_{i=1}^{n} \lambda[X_i(\omega)](p[X_i(\omega)])^{-1} \nu(d\lambda) \geqq e^{-(n\epsilon'/6)} \nu(V_0) \tag{23}$$

for a.e. $[P_p]\omega$ eventually. Combining (22) and (23) and noting $V_0 \subset V$ we have eventually

$$\begin{aligned} &\nu_{n,\omega}(L_+ - V)/\nu_{n,\omega}(V) \\ &\quad \leqq \nu(L_+ - V) e^{-n\epsilon'/3} (\nu(V_0) \epsilon^{-(n\epsilon')/6})^{-1} = \nu(L_+ - V)(\nu(V_0))^{-1} e^{-(n\epsilon')/6} \to 0 \end{aligned} \tag{24}$$

as $n \to \infty$ for a.e. $[P_p]\omega$. Thus $\nu_{n,\omega}(V) \to 1$ as $n \to \infty$ for a.e. $[P_p]\omega$.

Now let $\epsilon = \epsilon_k$ where $\epsilon_k \to 0$ as $k \to \infty$ and apply this argument to get an L_+-neighborhood V_k of p^+ and a P_p-null set E_k such that $\omega \notin E_k$ implies $\lim_{n\to\infty} \nu_{n,\omega}(V_k) = 1$. So for $\omega \notin \bigcup_1^\infty E_k$, $\lim_{n\to\infty} \nu_{n,\omega}(V_k) = 1$ for all k. Since $\{V_k: 1 \leqq k < \infty\}$ is a basis for L-neighborhoods of p^+, $\nu_{n,\omega}$ converges to δ_{p^+}. The proof is complete if $I - I_+$ is empty.

Otherwise let $J \subset I - I_+$ be a finite set. Then if $\epsilon > 0$ and V is an L_+-neighborhood of p^+, $W = \{\lambda: \lambda \in L, \max_{i\in J} \lambda_i < \epsilon, \lambda^+ \in V\}$ is an L-neighborhood of p. Varying J, V and ϵ yields a basis of L-neighborhoods of p. Since $\nu_{n,\omega} \to \delta_{p^+}$, $\mu_{n,\omega}(W) \geqq \nu_{n,\omega}\{\lambda: \lambda^+ \in L_+, \lambda^+ \in V, \sum_{i\in I_+} \lambda_i > 1 - \epsilon/2\} \to 1$ so $\mu_{n,\omega}(W) \to 1$ and $\mu_{n,\omega} \to \delta_p$ thus completing the proof of Theorem 1.

An alternative method of proof of Theorem 1 is to demonstrate the existence of a uniformly consistent sequence of tests of p against $L - V$ for each ϵ, then to apply Theorem 6.1 of Schwartz (1965) to prove the consistency of (p, μ).

It is not clear that our prior measure on L satisfies the hypothesis of Theorem 1, so we need

THEOREM 2. *If μ is a probability on $\mathfrak{B}$ such that $\mu\{\lambda: \lambda \varepsilon L, \sum_{i=0}^{m} p_i \log (p_i/\lambda_i) < \epsilon$ where $p_0 = \sum_{i=m+1}^{\infty} p_i$ and $\lambda_0 = \sum_{i=m+1}^{\infty} \lambda_i\} > 0$ for all m, then (p, μ) is consistent.*

PROOF. Define I_+ etc. as in Theorem 1. Fix m and for the X's defined on I_+ let

$$X_i^{(m)} = X_i, \qquad \text{if } X_i \leqq m - 1,$$
$$= m, \qquad \text{if } X_i \geqq m.$$

Let S^m be the space of functions from $\{1, \cdots, m\}$ to $[0, 1]$ and define L^m and $\mathfrak{B}^m$ accordingly. ν^m is the measure induced by ν on $(L^m, \mathfrak{B}^m)$ by the stopping of X at m. If V^m is a neighborhood of $p^m = (p_1, \cdots, p_m')$ in L^m where $p_m' = p_m + p_0$, then

$$\nu_{n,\omega}^m(V^m) \to 1 \quad \text{for} \quad \text{a.e.} \quad [P_p]_\omega. \tag{25}$$

This follows by copying the proof of Theorem 1 using the stopped X's and corresponding L^m, $\mathfrak{B}^m$ etc. Here note that $\sum_{i=0}^{m} p_i \log (p_i/\lambda_i) = \sum_{i=1}^{m-1} p_i \log (p_i/\lambda_i) + p_m' \log (p_m'/\lambda_m')$.

For each m there is a P_p-null set E_m so if $\omega \varepsilon E_m$, (25) holds. Letting $E = \bigcup_1^\infty E_m$, (25) holds for all m if $\omega \notin E$.

If V is a neighborhood of p^+ in L_+, the problem is to establish

$$\nu_{n,\omega}(V) \to 1 \quad \text{for} \quad \text{a.e.} \quad [P_p]\omega. \tag{26}$$

For any such neighborhood V, there is a sufficiently large k and a corresponding set $V^k \varepsilon L^k$ so that if $V_k = V^k \times L_+$ we have $V_k \subset V$. This is true since the cylinder sets from a basis in $(L_+, \mathfrak{B}_+)$. Thus for $\omega \notin E$, $\nu_{n,\omega}(V) \geqq \nu_{n,\omega}(V_k) = \nu_{n,\omega}^k(V^k) \to 1$. If $I - I_+$ is empty the proof ends. Otherwise by exactly the same argument as in Theorem 1 it is proven that $\mu_{n,\omega}(W) \to 1$ and $\mu_{n,\omega} \to \delta_p$ for a.e. $[P_p]\omega$. Q.E.D.

To verify the hypothesis of Theorem 2 for our problem we observe that the first N moments map $\mathcal{G}$ onto D^N. The inverse moment map say T^N is a set function on $\mathfrak{B}_D{}^N$, the Borel sigma field of D^N. Set $\mathfrak{A}^N = T^N(\mathfrak{B}_D{}^N)$, then $T^N: \mathfrak{B}_D{}^N \to \mathfrak{A}^N$. The $\{\mathfrak{A}^N\}_{N=1}^\infty$ are an increasing sequence of sigma fields and we let $\mathfrak{A}$ be the smallest sigma field containing them. For each N, the prior distribution is a measure on $(\mathcal{G}, \mathfrak{A}^N)$, thus from the Kolmogorov extension theorem it is a measure on $\mathfrak{A}$. For an arbitrary fixed m, we let $U^m = \{G: G \varepsilon \mathcal{G}, I_m'(G_0, G) = \sum_{x=0}^{m} q_x(G_0) \cdot \log [q_x(G_0)/q_x(G)] < \epsilon\}$ with q_0 defined as above. Since $q_x(G) = \sum_{i=0}^{k_x} a_{xi} x_i(G)$ for $x > 0$, $I_m'(G_0, G)$ is a continuous function of the first K^* moments of G where $K^* = \max_{1 \leqq x \leqq m} (k_x)$. Thus $U^m \varepsilon \mathfrak{A}^{K^*}$ and if μ is the original prior probability distribution on $(\mathcal{G}, \mathfrak{A})$ $\mu(U^m) > 0$. Set $V^m = Q(U^m)$ and denote by μQ^{-1} the measure induced through Q on $(L, \mathfrak{B})$ by μ. So $\mu Q^{-1}(V^m) = \mu(U^m) > 0$ for each m, thus satisfying the hypothesis of Theorem 2. The conclusion is

$$(Q(G), \mu Q^{-1}) \quad \text{is consistent.} \tag{27}$$

Now put the weak* topology on $\mathcal{G}$. It is routine to show that the topology induced by moment convergence is the same as the weak* topology. By Helly's theorem $\mathcal{G}$ is sequentially compact. Since this topology is metric then $\mathcal{G}$ is compact. The map Q of $\mathcal{G}$ into L is continuous in this topology since each $q_x(G)$ is a continuous function. Let W = Range Q with the relative topology $\mathcal{B}_W$. Since the problem is identifiable, Q is one to one, so $Q^{-1}: W \to \mathcal{G}$ exists. $\mathcal{B}_W$ is Hausdorff so by a standard theorem (e.g. Kelley, p. 141) Q^{-1} is a continuous function.

THEOREM 3. *If* (a) $q_x(t)$ *is continuous for each* x, (b) Q *is one to one and* (c) $(Q(G_0), \mu Q^{-1})$ *is consistent, then* (G_0, μ) *is consistent.*

NOTE. Consistency on $(\mathcal{G}, \mathcal{A})$ is defined analogously to consistency on $(L, \mathcal{B})$ as at the beginning of this section.

PROOF. Let U be a neighborhood of G_0 in $(\mathcal{G}, \mathcal{A})$. Then by the continuity of Q^{-1} there exists a neighborhood V of $Q(G_0)$ in$(W, \mathcal{B}_W)$ so that $Q^{-1}(V) \subset U$.

$$\mu_{n,\omega}(U) \geqq \mu_{n,\omega}(Q^{-1}(V)) = \mu Q_{n,\omega}^{-1}(V) \to 1$$

as $n \to \infty$ for a.e. $[P_{Q(G_0)}]\omega$. But $P_{Q(G_0)}$ is the distribution on $(\Omega, \mathcal{F})$ corresponding to G_0 so the proof is terminated.

Theorem 3 and (27) yield

(28) (G_0, μ) is consistent for every $G_0 \varepsilon \mathcal{G}$.

REMARK. Note that Theorems 1, 2 and 3 assume only that $q_x(t)$ is continuous and that the family $\mathcal{Q}$ is identifiable. Thus under these assumptions the posterior distribution is still consistent. Section 5 gives the estimates for this case. Having proved consistency of the posterior distribution on $(\mathcal{G}, \mathcal{A})$ we apply it to get consistency of our estimator $\hat{G}$.

THEOREM 4. $\hat{G}(X_1, \cdots, X_n)$ *is a consistent sequence of estimators for all* $G \varepsilon \mathcal{G}$. *In other words, if* U *is a neighborhood of* G *in* $\mathcal{G}$, *there exists a* P_G*-null set* E *in* $(\Omega, \mathcal{F})$ *so if* $\omega \varepsilon E$ *then* $\hat{G}(X_1(\omega), \cdots, X_n(\omega)) \varepsilon U$ *eventually.*

PROOF. Let U be a neighborhood of G in $(\mathcal{G}, \mathcal{A})$. There is a G-neighborhood $U^r \varepsilon \mathcal{A}^r$ for some r with $U^r \subset U$. Since the posterior distribution is weak* consistent and $K \to \infty$ as $n \to \infty$, $(\hat{m}_1, \cdots, \hat{m}_r)$ is consistent in estimating $(m_1(G), \cdots, m_r(G))$. Thus by the definition of $\mathcal{A}^r$, $\hat{G} \varepsilon U^r \subset U$ eventually for a.e. $[P_G]\omega$. Q.E.D.

5. Extension to continuous functions. We begin by examining the case where $q_x(t)$ is a power series in t, that is, $q_x(t) = \sum_{i=0}^{\infty} a_{xi} t^i$. The prior distribution and computation of the posterior distribution are the same as in Section 2 except that K, the maximal degree of the polynomial $q_x(t)$ for the x values which occur, is now $+\infty$. Thus the posterior density on D can be obtained by taking the limit of g^N in (13) as $N \to \infty$ and noting that the contribution of the parameters, $m_{N+1}, m_{N+2}, \cdots$ decreases to zero as $N \to \infty$. As remarked above, the posterior distribution is still consistent. The computation of the estimates changes because K is no longer finite. This can be solved by choosing a non-decreasing sequence of positive integers $\{K_n\}$ with $K_n \to \infty$ as $n \to \infty$ and truncating all the series $q_x(t)$ after K_n terms. The analogues to (13), (17) and (18) are then com-

puted with these $q_x(t)$. For $N > K_n$(13) becomes

$$g_n{}^N(m_1, \cdots, m_N \mid n_1, \cdots, n_c, \cdots)$$
$$= \prod_{x=1}^{c} \left(\sum_{i=1}^{K_n} a_{xi} m_i\right)^{n_x} \prod_{i=1}^{N} d_i^{-1} / I_n(n_1, \cdots, n_c, \cdots))$$

where $I_n(n_1, \cdots, n_c, \cdots) = \int_{D^N} \prod_{x=1}^{c} \left(\sum_{i=1}^{K_n} a_{xi} m_i\right)^{n_x} \prod_{i=1}^{N} d_i^{-1}\, dm_1 \cdots dm_N$. The expected value of the first N moments under the posterior distribution is approximated by

$$(\hat{m}_1, \cdots, \hat{m}_N)$$
$$= \int_{D^N} (m_1, \cdots, m_N) g_n{}^N(m_1, \cdots m_N \mid n_1, \cdots, n_c, \cdots)\, dm_1 \cdots dm_N.$$

The estimate is then

$$\hat{G}(X_1, \cdots, X_n) = G_{K_n}(\hat{m}_1, \cdots, \hat{m}_{K_n}) = \alpha \underline{G}_{K_n} + (1 - \alpha)\bar{G}_{K_n}$$

with $$\alpha = (\bar{\Delta}_{K_n}/\bar{\Delta}_{K_n-2})[(\bar{\Delta}_{K_n}/\bar{\Delta}_{K_n-2}) + (\underline{\Delta}_{K_n} + \underline{\Delta}_{K_n-2})]^{-1}.$$

As can be seen from the argument for arbitrary positive continuous functions given below, these estimates are indeed consistent for any $\{K_n\} \to \infty$. If one cares about more than just asymptotic properties, the sequence $\{K_n\}$ should be chosen to make the calculated posterior distribution and corresponding estimates close to the true ones for small and moderate n.

An example of $q_x(t)$ being a power series is the Poisson distribution with parameter t restricted to [0, 1]. Here

$$q_x(t) = t^x(x!)^{-1}e^{-tx} = t^x(x!)^{-1} \sum_{i=0}^{\infty} [(-1)^i x^i (i!)^{-1}] t^i.$$

This is identifiable so the estimates are consistent. In the Poisson case, the method used by Tucker [14] combined with a prior distribution might well yield a closed form solution, thus avoiding the need for the sequence $\{K_n\}$.

More generally assume only that $q_x(t)$ is continuous and positive on [0, 1]. Consistency holds but the $\hat{m}_j$'s may be difficult to calculate. The approach is to approximate the posterior expected moments by the polynomial case and thus derive approximating estimates.

For any $\epsilon > 0$, Weierstrass's theorem says that there exist polynomials $p_{x\epsilon}(t)$ for each x such that $|p_{x\epsilon}(t)/q_x(t) - 1| < \epsilon$ for all $t \in [0, 1]$. Let

$$(29) \quad \hat{m}_j{}^*(X_1, \cdots, X_n) = \hat{m}_j{}^*(n)$$
$$= \int_{\mathcal{G}} m_j(G_0) \prod_{i=1}^{n} p_{X_i\epsilon}(G_0)\, d\mu(G_0) / \int_{\mathcal{G}} \prod_{i=1}^{n} p_{X_i\epsilon}(G_0)\, d\mu(G_0),$$

then $((1 - \epsilon)/(1 + \epsilon))^n < \hat{m}_j{}^*(n)/\hat{m}_j(n) < ((1 + \epsilon)/(1 - \epsilon))^n$. Let $\delta(n) \to 0$ as $n \to \infty$ be some arbitrary sequence.

It is enough to pick an $\epsilon(n)$ for the approximating polynomials such that $(1 + \epsilon(n)/1 - \epsilon(n))^n \leqq 1 + \delta(n)$ to have $|\hat{m}_j{}^*(n)/\hat{m}_j(n) - 1| < \delta(n)$ for all j.

THEOREM 5. *For each j, $\hat{m}_j{}^*(n)$ is a consistent sequence of estimators of $m_j(G_0)$ for any G_0 in $\mathcal{G}$.*

PROOF. Let

$$U = \{G: |m_j(G) - m_j(G_0)| < \epsilon\}.$$

For a fixed sequence $\delta(n)$, $\delta(n) < \epsilon/2$ eventually. Since $\hat{m}_j$ is consistent, $|\hat{m}_j - m_j(G_0)| < \epsilon/2$ eventually a.s. Thus eventually a.s.

$$|\hat{m}_j{}^*(n) - m_j(G_0)|$$

$$\leqq |\hat{m}_j{}^*(n) - \hat{m}_j(n)| + |\hat{m}_j(n) - m_j(G_0)| < \hat{m}_j(n)\delta(n) + \epsilon/2 \leqq \epsilon/2 + \epsilon/2 = \epsilon.$$

So $\hat{m}_j{}^*(n) \,\varepsilon\, U$ eventually a.s.; proving the theorem.

Note that once the approximation to $q_x(t)$ is made, $\hat{m}_j{}^*$ involves just the same calculation as before. The estimate of G is defined as in (18)

$$\hat{G}^*(X_1, \cdots, X_n) = G_{K'}(\hat{m}_1{}^*(n), \cdots, \hat{m}_K{}^*(n)) \tag{30}$$

where $K' = \max_{\{x:n_x>0\}} (k_x')$ with k_x' = the degree of the polynomial approximating $q_x(t)$. By the same argument as in Theorem 4, $\hat{G}^*$ is a consistent sequence of estimators of G for all $G \,\varepsilon\, \mathcal{G}$.

6. Generalizations. The prior distribution ν on $\mathcal{G}$ introduced in Section 2 is in a certain sense uniform on the moment space D. Consider an alternative prior distribution ξ defined as follows. Let $H_1, H_2, \cdots$ be a sequence of measures on [0, 1] having everywhere positive densities $h_1, h_2, \cdots$ with respect to Lebesgue measure. Define the distribution ξ on D by its conditional densities on the ith moments.

$$\xi_i(m \mid m_1, \cdots, m_{i-1}) = h_i(m)/\textstyle\int_{\underline{m}_i}^{\bar{m}_i} h_i(m)\, dm, \qquad \text{if } \underline{m}_i \leqq m \leqq \bar{m}_i, \tag{31}$$

$$= 0, \qquad \text{elsewhere.}$$

THEOREM 6. *For any such prior distribution ξ as above, (ξ, G) is consistent.*

PROOF. ξ and the original ν are mutually absolutely continuous on D. Since νQ^{-1} satisfies Theorem 2, so does ξQ^{-1} and the consistency proof goes through.

REMARK. If one wishes the prior distribution to reflect his state of knowledge or information about the true G_0, Theorem 6 yields a rich class of priors from which to choose. H_1 is interpreted as prior information on location; it together with H_2 as the prior information on scale. Beyond this one may not care to go, particularly since the computation of the posterior becomes successively more difficult. A suggested procedure is to choose H_1 and H_2 and then return to the uniform ν given in Section 2.

In this paper the problem is assumed to be identifiable; indeed if it is not, there is no hope of estimating G_0 from the observations. If the problem is not identifiable, $Q(G_1) = Q(G_2)$ does not imply $G_1 = G_2$, the best one can hope for is to estimate G up to inverse images of Q; that is up to classes of the form $U_\lambda = \{G: Q(G) = \lambda\}$ for all $\lambda \,\varepsilon\, \Lambda$.

Take the example of binomial n,

$$P_G(X = x) = q_x(G) = \textstyle\int_0^1 \binom{n}{x} p^x(1-p)^x(1-p)^{n-x}\, dG(p).$$

The distribution of X depends only on the first n moments of G and the problem is identifiable up to the first n moments of G. Cogburn [3] has used the notion of stringency to handle this situation. G is estimated through the observations up to the equivalence class U_λ then some other criterion is used to estimate G within the class. Cogburn's criterion is to use minimax estimates with respect to some given loss function. The method presented here can, I think, be extended to get consistent estimates of the classes. In the binomial example, estimating U_λ amounts to estimating only the first n moments which is already part of our procedure.

Turning to spaces other than [0, 1], Tucker [14] has given a consistent non-Bayesian estimate in the Poisson case using moment type estimates. It is of course crucial that moment sequences yield unique distribution functions. Using known facts about moment sequences [11], one may be able to apply our method to other spaces than [0, 1].

7. Acknowledgment. I wish to thank Professor David Blackwell for his guidance throughout this research. Thanks are also due to the referee whose careful reading lead to a number of improvements.

REFERENCES

[1] ABRAHAMSON, I. G. (1965). On the stochastic comparison of tests of hypotheses. Ph.D. Dissertation, Univ. of Chicago.

[2] BLISCHKE, W. R. (1964). Estimating the parameters of mixtures of binomial distributions. *J. Amer. Statist. Assn.* **59** 510–528.

[3] COGBURN, R. (1967). Stringent solutions to statistical decision problems. *Ann. Math. Statist.* **38** 447–463.

[4] DEELY, J. J. and KRUSE, R. L. (1967). Construction of sequences for estimating the mixing distribution. Mimeographed.

[5] FREEDMAN, D. A. (1963). On the asymptotic behaviour of Bayes' estimates in the discrete case. *Ann. Math. Statist.* **34** 1386–1403.

[6] GAFFEY, W. R. (1959). A consistent estimator of a component of a convolution. *Ann. Math. Statist.* **30** 198–205.

[7] KARLIN, S. and SHAPLEY, L. S. (1953). Geometry of a moment spaces. *Memoirs Amer. Math. Soc.* **12** 1–93.

[8] KELLY, J. L. (1955). *General Topology.* Van Nostrand, Princeton.

[9] ROBBINS, H. (1964). The empirical Bayes approach to statistical decision problems. *Ann. Math. Statist.* **35** 1–20.

[10] SCHWARTZ, L. (1965). On Bayes Procedures. *Z. Wahrscheinlichkeitstheorie und Verew Giebete* **4** 10–26.

[11] SHOAHAT, J. A. and TAMARKIN, J. D. (1943). *The Problem of Moments.* Amer. Math. Soc., New York.

[12] TEICHER, H. (1960). On the mixtures of distributions. *Ann. Math. Statist.* **31** 55–73.

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

[14] TUCKER, H. G. (1963). An estimate of the compounding distribution of a compound Poisson distribution. *Theor. Prob. Appl.* **8** 195–200.