A Bayesian Analysis of Some Nonparametric Problems

Thomas S. Ferguson

The Annals of Statistics, Vol. 1, No. 2 (Mar., 1973), 209-230.

Stable URL:
http://links.jstor.org/sici?sici=0090-5364%28197303%291%3A2%3C209%3AABAOSN%3E2.0.CO%3B2-U

The Annals of Statistics
1973, Vol. 1, No. 2, 209-230

A BAYESIAN ANALYSIS OF SOME NONPARAMETRIC PROBLEMS[1]

BY THOMAS S. FERGUSON

University of California, Los Angeles

1. Introduction and summary. The Bayesian approach to statistical problems, though fruitful in many ways, has been rather unsuccessful in treating nonparametric problems. This is due primarily to the difficulty in finding workable prior distributions on the parameter space, which in nonparametric ploblems is taken to be a set of probability distributions on a given sample space. There are two desirable properties of a prior distribution for nonparametric problems.

(I) The support of the prior distribution should be large—with respect to some suitable topology on the space of probability distributions on the sample space.

(II) Posterior distributions given a sample of observations from the true probability distribution should be manageable analytically.

These properties are antagonistic in the sense that one may be obtained at the expense of the other. This paper presents a class of prior distributions, called Dirichlet process priors, broad in the sense of (I), for which (II) is realized, and for which treatment of many nonparametric statistical problems may be carried out, yielding results that are comparable to the classical theory.

In Section 2, we review the properties of the Dirichlet distribution needed for the description of the Dirichlet process given in Section 3. Briefly, this process may be described as follows. Let $\mathscr{X}$ be a space and $\mathscr{A}$ a σ-field of subsets, and let α be a finite non-null measure on $(\mathscr{X}, \mathscr{A})$. Then a stochastic process P indexed by elements A of $\mathscr{A}$, is said to be a Dirichlet process on $(\mathscr{X}, \mathscr{A})$ with parameter α if for any measurable partition $(A_1, \cdots, A_k)$ of $\mathscr{X}$, the random vector $(P(A_1), \cdots, P(A_k))$ has a Dirichlet distribution with parameter $(\alpha(A_1), \cdots, \alpha(A_k))$. P may be considered a random probability measure on $(\mathscr{X}, \mathscr{A})$, The main theorem states that if P is a Dirichlet process on $(\mathscr{X}, \mathscr{A})$ with parameter α, and if $X_1, \cdots, X_n$ is a sample from P, then the posterior distribution of P given $X_1, \cdots, X_n$ is also a Dirichlet process on $(\mathscr{X}, \mathscr{A})$ with parameter $\alpha + \sum_1^n \delta_{X_i}$, where δ_x denotes the measure giving mass one to the point x.

In Section 4, an alternative definition of the Dirichlet process is given. This definition exhibits a version of the Dirichlet process that gives probability one to the set of discrete probability measures on $(\mathscr{X}, \mathscr{A})$. This is in contrast to Dubins and Freedman [2], whose methods for choosing a distribution function on the interval [0, 1] lead with probability one to singular continuous distributions. Methods of choosing a distribution function on [0, 1] that with probability one is absolutely continuous have been described by Kraft [7]. The

Received September 17, 1969; revised April 3, 1972.

[1] The preparation of this paper was supported in part by NSF Grant No. GP-8049.

general method of choosing a distribution function on [0, 1], described in Section 2 of Kraft and van Eeden [10], can of course be used to define the Dirichlet process on [0, 1].

Special mention must be made of the papers of Freedman and Fabius. Freedman [5] defines a notion of tailfree for a distribution on the set of all probability measures on a countable space $\mathscr{X}$. For a tailfree prior, posterior distribution given a sample from the true probability measure may be fairly easily computed. Fabius [3] extends the notion of tailfree to the case where $\mathscr{X}$ is the unit interval [0, 1], but it is clear his extension may be made to cover quite general $\mathscr{X}$. With such an extension, the Dirichlet process would be a special case of a tailfree distribution for which the posterior distribution has a particularly simple form.

There are disadvantages to the fact that P chosen by a Dirichlet process is discrete with probability one. These appear mainly because in sampling from a P chosen by a Dirichlet process, we expect eventually to see one observation exactly equal to another. For example, consider the goodness-of-fit problem of testing the hypothesis H_0 that a distribution on the interval [0, 1] is uniform. If on the alternative hypothesis we place a Dirichlet process prior with parameter α itself a uniform measure on [0, 1], and if we are given a sample of size $n \geqq 2$, the only nontrivial nonrandomized Bayes rule is to reject H_0 if and only if two or more of the observations are exactly equal. This is really a test of the hypothesis that a distribution is continuous against the hypothesis that it is discrete. Thus, there is still a need for a prior that chooses a continuous distribution with probability one and yet satisfies properties (I) and (II).

Some applications in which the possible doubling up of the values of the observations plays no essential role are presented in Section 5. These include the estimation of a distribution function, of a mean, of quantiles, of a variance and of a covariance. A two-sample problem is considered in which the Mann–Whitney statistic, equivalent to the rank-sum statistic, appears naturally. A decision theoretic upper tolerance limit for a quantile is also treated. Finally, a hypothesis testing problem concerning a quantile is shown to yield the sign test.

In each of these problems, useful ways of combining prior information with the statistical observations appear.

Other applications exist. In his Ph. D. dissertation [1], Charles Antoniak finds a need to consider mixtures of Dirichlet processes. He treats several problems, including the estimation of a mixing distribution, bio-assay, empirical Bayes problems, and discrimination problems.

2. The Dirichlet distribution. The discussion of this section is well known but our definition of the Dirichlet distribution is slightly more general than the usual one. The Dirichlet distribution makes its appearance in problems involving order statistics. A discussion of these applications and of the main properties of the Dirichlet distribution may be found in the book of S. S. Wilks [11]. The

Dirichlet distribution is known to Bayesians as the conjugate prior for the parameters of a multinomial distribution. See, for example, the book by I. J. Good [7].

We denote by $\mathscr{G}(\alpha, \beta)$ the gamma distribution with shape prarameter $\alpha \geqq 0$, and scale parameter $\beta > 0$. For $\alpha = 0$, this distribution is degenerate at zero; for $\alpha > 0$, this distribution has density with respect to Lebesgue measure on the real line

$$f(z \mid \alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} e^{-z/\beta} z^{\alpha-1} I_{(0,\infty)}(z) \tag{1}$$

where $I_S(z)$ represents the indicator function of the set S.

We define the Dirichlet distribution slightly more generally than in Wilks [11], by allowing some of the variables to be degenerate at zero.

Let $Z_1, Z_2, \cdots, Z_k$ be independent random variables with $Z_j \in \mathscr{G}(\alpha_j, 1)$, where $\alpha_j \geqq 0$ for all j, and $\alpha_j > 0$ for some j, $j = 1, 2, \cdots, k$.

The Dirichlet distribution with parameter $(\alpha_1, \cdots, \alpha_k)$, *denoted by* $\mathscr{D}(\alpha_1, \cdots, \alpha_k)$, *is defined as the distribution of* $(Y_1, \cdots, Y_k)$, *where*

$$Y_j = Z_j / \textstyle\sum_{i=1}^{k} Z_i \qquad \textit{for } j = 1, 2, \cdots, k\,. \tag{2}$$

Use of the notation $\mathscr{D}(\alpha_1, \cdots, \alpha_k)$ is taken to imply that $\alpha_j \geqq 0$ for all j, and $\alpha_j > 0$ for some j. This distribution is always singular with respect to Lebesgue measure in k-dimensional space since $Y_1 + \cdots + Y_k = 1$. In addition, if any $\alpha_j = 0$, the corresponding Y_j is degenerate at zero. However, if $\alpha_j > 0$ for all j, the $(k-1)$-dimensional distribution of $(Y_1, \cdots, Y_{k-1})$ is absolutely continuous with density

$$\begin{aligned} f(y_1, \cdots, y_{k-1} \mid \alpha_1, \cdots, \alpha_k) \\ = \frac{\Gamma(\alpha_1 + \cdots + \alpha_k)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_k)} (\textstyle\prod_{j=1}^{k-1} y_j^{\alpha_j - 1})(1 - \sum_{j=1}^{k-1} y_j)^{\alpha_k - 1} I_{\mathbb{S}}(y_1, \cdots, y_{k-1}) \end{aligned} \tag{3}$$

where $\mathbb{S}$ is the simplex

$$\mathbb{S} = \{(y_1, \cdots, y_{k-1}) : y_j \geqq 0, \textstyle\sum_{j=1}^{k-1} y_j \leqq 1\}\,.$$

For $k = 2$, (3) reduces to the Beta distribution, denoted by $\mathscr{B}e(\alpha_1, \alpha_2)$.

The main property of the Dirichlet distribution as used below is

i°. *If* $(Y_1, \cdots, Y_k) \in \mathscr{D}(\alpha_1, \cdots, \alpha_k)$ *and* $r_1, \cdots, r_l$ *are integers such that* $0 < r_1 < \cdots < r_l = k$, *then*

$$(\textstyle\sum_1^{r_1} Y_i, \sum_{r_1+1}^{r_2} Y_i, \cdots, \sum_{r_{l-1}+1}^{r_l} Y_i) \in \mathscr{D}(\sum_1^{r_1} \alpha_i, \sum_{r_1+1}^{r_2} \alpha_i, \cdots, \sum_{r_{l-1}+1}^{r_l} \alpha_i)\,.$$

This follows directly from the definition of the Dirichlet distribution and the additive property of the gamma distribution: If $Z_1 \in \mathscr{G}(\alpha_1, 1)$, if $Z_2 \in \mathscr{G}(\alpha_2, 1)$, and if Z_1 and Z_2 are independent, then $Z_1 + Z_2 \in \mathscr{G}(\alpha_1 + \alpha_2, 1)$.

In particular, the marginal distribution of each Y_j is Beta: $Y_j \in \mathscr{B}e(\alpha_j, (\sum_1^k \alpha_i) - \alpha_j)$.

We record for future use the first two moments of the Dirichlet distribution.

ii°. *If* $(Y_1, \cdots, Y_k) \in \mathscr{D}(\alpha_1, \cdots, \alpha_k)$, *then*

$$\mathscr{E}Y_i = \alpha_i/\alpha$$
$$\mathscr{E}Y_i^2 = \alpha_i(\alpha_i + 1)/(\alpha(\alpha + 1)) \qquad \textit{and}$$
$$\mathscr{E}Y_iY_j = \alpha_i\alpha_j/(\alpha(\alpha + 1)) \qquad \textit{for } i \neq j$$

where $\alpha = \sum_1^k \alpha_i$.

The following Bayes property of the Dirichlet distribution is well known.

iii°. *If the prior distribution of* $(Y_1, \cdots, Y_k)$ *is* $\mathscr{D}(\alpha_1, \cdots, \alpha_k)$ *and if*

$$\mathscr{P}\{X = j \mid Y_1, \cdots, Y_k\} = Y_j \quad \text{a.s.} \qquad \textit{for } j = 1, \cdots, k,$$

then the posterior distribution of $(Y_1, \cdots, Y_k)$ *given* $X = j$ *is* $\mathscr{D}(\alpha_1^{(j)}, \cdots, \alpha_k^{(j)})$, *where*

$$\alpha_i^{(j)} = \alpha_i \qquad \textit{if } i \neq j$$
$$= \alpha_j + 1 \qquad \textit{if } i = j.$$

Contained in this property is a formula that will prove useful. Let us use $D(y_1, \cdots, y_k \mid \alpha_1, \cdots, \alpha_k)$ to denote the distribution function of the Dirichlet distribution, $\mathscr{D}(\alpha_1, \cdots, \alpha_k)$. Then, the equality

$$\mathscr{P}\{X = j, Y_1 \leqq z_1, \cdots, Y_k \leqq z_k\}$$
$$= \mathscr{P}\{X = j\}\mathscr{P}\{Y_1 \leqq z_1, \cdots, Y_k \leqq z_k \mid X = j\}$$

may be expressed in terms of $D(y_1, \cdots, y_k \mid \alpha_1, \cdots, \alpha_k)$, using ii° and iii°, as

$$\int_0^{z_1} \cdots \int_0^{z_k} y_j \, dD(y_1, \cdots, y_k \mid \alpha_1, \cdots, \alpha_k) \tag{4}$$
$$= \frac{\alpha_j}{\alpha} D(z_1, \cdots, z_k \mid \alpha_1^{(j)}, \cdots, \alpha_k^{(j)}).$$

It should be noted that this formula is true even if $\alpha_j = 0$.

3. The Dirichlet process. Let $\mathscr{X}$ be a set and let $\mathscr{A}$ be a σ-field of subsets of $\mathscr{X}$. We define below a random probability, P, on $(\mathscr{X}, \mathscr{A})$ by defining the joint distribution of the random variables $(P(A_1), \cdots, P(A_m))$ for every m and every sequence $A_1, \cdots, A_m$ of measurable sets ($A_i \in \mathscr{A}$ for all i). We then verify the Kolmogorov consistency conditions to show there exists a probability, $\mathscr{P}$, on $([0, 1]^{\mathscr{A}}, B\mathscr{T}^{\mathscr{A}})$ yielding these distributions. Here, $[0, 1]^{\mathscr{A}}$ represents the space of all functions from $\mathscr{A}$ into $[0, 1]$, and $B\mathscr{T}^{\mathscr{A}}$ represents the σ-field generated by the field of cylinder sets (Kolmogorov [8]).

For our purposes, it is more convenient to define the random probability, P, by defining the joint distribution of $(P(B_1), \cdots, P(B_k))$ for all k and all measurable partitions $(B_1, \cdots, B_k)$ of $\mathscr{X}$. (We say $(B_1, \cdots, B_k)$ is a measurable partition of $\mathscr{X}$ if $B_i \in \mathscr{A}$ for all i, $B_i \cap B_j = \varnothing$ for $i \neq j$, and $\bigcup_{j=1}^k B_j = \mathscr{X}$.) From these distributions, the joint distribution of $(P(A_1), \cdots, P(A_m))$ for arbitrary measurable sets $A_1, \cdots, A_m$ may be defined using the hoped for finite additivity of P as follows.

Given arbitrary measurable sets $A_1, \cdots, A_m$, form the $k = 2^m$ sets obtained by taking intersections of the A_i and their complements; that is, define $B_{\nu_1,\cdots,\nu_m}$ for each $\nu_j = 0$ or 1 as

$$B_{\nu_1,\cdots,\nu_m} = \bigcap_{j=1}^{m} A_j^{\nu_j} \tag{1}$$

where A_j^1 is interpreted as A_j, and A_j^0 is interpreted as A_j^c, the complement of A_j. Thus the $\{B_{\nu_1,\cdots,\nu_m}\}$ form a measurable partition of $\mathscr{X}$. If we are given the joint distribution of

$$\{P(B_{\nu_1,\cdots,\nu_m});\ \nu_j = 0 \text{ or } 1, j = 1, \cdots, m\}, \tag{2}$$

then we may define the joint distribution of $(P(A_1), \cdots, P(A_m))$ to be that obtainable from (2) upon defining for $i = 1, \cdots, m$

$$P(A_i) = \sum_{(\nu_1,\cdots,\nu_m)\ni \nu_i=1} P(B_{\nu_1,\cdots,\nu_m}). \tag{3}$$

We note that if $(A_1, \cdots, A_m)$ was a measurable partition to start with, then this does not lead to contradictory definitions of the distribution of $(P(A_1), \cdots, P(A_m))$ provided

$$P(\varnothing) \quad \text{is degenerate at} \quad 0. \tag{4}$$

(We are assuming that the distributions of the random variables are defined free of their order, so that Kolmogorov's condition (2) ([8] page 29) is automatic.) Under condition (4), the distribution of $(P(A_1), \cdots, P(A_m))$ for arbitrary measurable $A_1, \cdots, A_m$ is defined uniquely, once the distributions of $(P(B_1), \cdots, P(B_k))$ are given for arbitrary measurable partitions $(B_1, \cdots, B_k)$.

If we are given a system of distributions of $(P(B_1), \cdots, P(B_k))$ for all k and all measurable partitions $(B_1, \cdots, B_k)$, there is one consistency criterion we would certainly like to have satisfied; namely,

CONDITION C. If $(B_1', \cdots, B_{k'}')$ and $(B_1, \cdots, B_k)$ are measurable partitions, and if $(B_1', \cdots, B_{k'}')$ is a refinement of $(B_1, \cdots, B_k)$ with $B_1 = \bigcup_1^{r_1} B_i'$, $B_2 = \bigcup_{r_1+1}^{r_2} B_i'$, $\cdots$, $B_k = \bigcup_{r_{k-1}+1}^{k'} B_i'$, then the distribution of

$$(\textstyle\sum_1^{r_1} P(B_i'), \sum_{r_1+1}^{r_2} P(B_i'), \cdots, \sum_{r_{k-1}+1}^{k'} P(B_i'))$$

as determined from the joint distribution of $(P(B_1'), \cdots, P(B_{k'}'))$, is identical to the distribution of $(P(B_1), \cdots, P(B_k))$.

As the following lemma shows, this condition is sufficient for the validity of the Kolmogorov consistency conditions for the distributions of $(P(A_1), \cdots, P(A_m))$ defined as in (2) and (3). In fact the lemma is valid also as a description of a random finitely additive set function, with finite values (by letting $P(A_i)$ take values in the real line, $\mathbb{R}$). However our interest in the present paper is with random probability measures. We will say that P is a random probability measure on $(\mathscr{X}, \mathscr{A})$, if C is satisfied, if $P(A)$ takes values only in $[0, 1]$, and if $P(\mathscr{X})$ is degenerate at 1.

LEMMA 1. *If a system of joint distributions of* $(P(B_1), \cdots, P(B_k))$ *for all* k *and*

measurable partitions $(B_1, \cdots, B_k)$ *is defined satisfying Condition* C, *and if for arbitrary measurable sets* $A_1, \cdots, A_m$, *the distribution of* $(P(A_1), \cdots, P(A_m))$ *is defined as in* (1), (2), *and* (3), *then there exists a probability* $\mathscr{P}$ *on* $([0, 1]^{\mathscr{A}}, B.\mathscr{F}^{\mathscr{A}})$ *yielding these distributions.*

PROOF. Since $\mathscr{C} \cup \emptyset = \mathscr{C}$, it follows from Condition C that $P(\emptyset)$ is degenerate at zero, and thus that the distribution of $(P(A_1), \cdots, P(A_m))$ is well-defined by (2) and (3). To check the Kolmogorov consistency conditions, we must show that, for arbitrary m and measurable sets $A_1, \cdots, A_m$, the marginal distribution of $(P(A_1), \cdots, P(A_{m-1}))$ derived from the distribution of $(P(A_1), \cdots, P(A_m))$ is identical to the defined distribution of $(P(A_1), \cdots, P(A_{m-1}))$.

The marginal distribution of $(P(A_1), \cdots, P(A_{m-1}))$ derived from the distribution of $(P(A_1), \cdots, P(A_m))$ is identical to the distribution of

$$(5) \qquad \left(\sum_{(\nu_1,\cdots,\nu_m),\nu_1=1} P(B_{\nu_1,\cdots,\nu_m}), \cdots, \sum_{(\nu_1,\cdots,\nu_m),\nu_{m-1}=1} P(B_{\nu_1,\cdots,\nu_m})\right)$$

derived from the distribution of (2). The distribution of $(P(A_1), \cdots, P(A_{m-1}))$ is defined as the distribution of

$$(6) \qquad \left(\sum_{(\nu_1,\cdots,\nu_{m-1}),\nu_1=1} P(B_{\nu_1,\cdots,\nu_{m-1}}), \cdots, \sum_{(\nu_1,\cdots,\nu_{m-1}),\nu_{m-1}=1} P(B_{\nu_1,\cdots,\nu_{m-1}})\right)$$

derived from the distribution of $\{P(B_{\nu_1,\cdots,\nu_{m-1}}); \nu_i = 0 \text{ or } 1, i = 1, \cdots, m-1\}$ where

$$B_{\nu_1,\cdots,\nu_{m-1}} = \bigcup_{j=1}^{m-1} A_j^{\nu_j} .$$

Since $B_{\nu_1,\cdots,\nu_{m-1}} = B_{\nu_1,\cdots,\nu_{m-1},1} \cup B_{\nu_1,\cdots,\nu_{m-1},0}$, Condition C implies that the distribution of $\{P(B_{\nu_1,\cdots,\nu_{m-1}}); \nu_j = 0 \text{ or } 1, j = 1, \cdots, m-1\}$ is identical to the distribution of $\{P(B_{\nu_1,\cdots,\nu_{m-1},1}) + P(B_{\nu_1,\cdots,\nu_{m-1},0}); \nu_j = 0 \text{ or } 1, j = 1, \cdots, m-1\}$ as determined from the distribution of (2). Thus, the distribution of (6) can also be found from the distribution of (2) upon replacing $P(B_{\nu_1,\cdots,\nu_{m-1}})$ by $P(B_{\nu_1,\cdots,\nu_{m-1},1}) + P(B_{\nu_1,\cdots,\nu_{m-1},0})$. With this replacement, (6) becomes formally identical to (5), proving that their distributions are identical.

DEFINITION 1. Let α be a non-null finite measure (nonnegative and finitely additive) on $(\mathscr{X}, \mathscr{A})$. We say P is a Dirichlet process on $(\mathscr{X}, \mathscr{A})$ with parameter α if for every $k = 1, 2, \cdots$, and measurable partition $(B_1, \cdots, B_k)$ of $\mathscr{X}$, the distribution of $(P(B_1), \cdots, P(B_k))$ is Dirichlet, $\mathscr{D}(\alpha(B_1), \cdots, \alpha(B_k))$.

The consistency Condition C for the Dirichlet process is exactly property i° of the Dirichlet distribution. It follows from Lemma 1 that the Kolmogorov consistency conditions are satisfied so that this actually defines a random process. In addition, since $P(\mathscr{X})$ is degenerate at 1, we call P a random probability measure.

The following three propositions show a close relationship between properties of the random probability measure, P, and properties of the parameter of the process, α.

PROPOSITION 1. *Let* P *be a Dirichlet process on* $(\mathscr{X}, \mathscr{A})$ *with parameter* α,

and let $A \in \mathscr{A}$. *If* $\alpha(A) = 0$, *then* $P(A) = 0$ *with probability one. If* $\alpha(A) > 0$, *then* $P(A) > 0$ *with probability one. Furthermore,* $\mathscr{E}P(A) = \alpha(A)/\alpha(\mathscr{X})$.

PROOF. By considering the partition (A, A^c), it is seen that $P(A)$ has a Beta distribution, $\mathscr{B}e(\alpha(A), \alpha(A^c))$. The proof follows immediately.

This proposition would seem to say that α and P have the same null sets, in other words, that α and P are mutually absolutely continuous. This interpretation is false; in fact, it is shown in the next section that P is essentially a discrete distribution. Thus, α and P may be mutually singular. The point is that the null set outside of which the conclusion of Proposition 1 holds may depend upon A.

PROPOSITION 2. *Let* P *be a Dirichlet process on* $(\mathscr{X}, \mathscr{A})$ *with parameter* α. *If* α *is* σ*-additive, then so is* P *in the sense that for a fixed decreasing sequence of measurable sets* $A_n \searrow \emptyset$, *we have* $P(A_n) \to 0$ *with probability one.*

PROOF. Since $A_n \searrow \emptyset$ and α is additive, $\alpha(A_n) \to 0$. Hence there exists a subsequence $\{n_j\}$ such that $\sum_1^\infty \alpha(A_{n_j}) < \infty$. For fixed $\varepsilon > 0$,

$$\sum_1^\infty \mathscr{P}\{P(A_{n_j}) > \varepsilon\} \leqq \sum_1^\infty \varepsilon^{-1}\mathscr{E}P(A_{n_j}) = \varepsilon^{-1} \sum_1^\infty \alpha(A_{n_j})/\alpha(\mathscr{X}) < \infty .$$

Hence, from the Borel–Cantelli lemma, $\mathscr{P}\{P(A_{n_j}) > \varepsilon \text{ i.o.}\} = 0$. This proves that $P(A_{n_j}) \to 0$ with probability one. The proof is completed by noting that $P(A_n) > P(A_{n+1})$ with probability one for all n, and hence, $P(A_1) > P(A_2) > \cdots$ with probability one. The converse is also true: if α is not σ-additive, then with probability one P is not σ-additive.

PROPOSITION 3. *Let* P *be a Dirichlet process on* $(\mathscr{X}, \mathscr{A})$ *with parameter* α, *and let* Q *be a fixed probability measure on* $(\mathscr{X}, \mathscr{A})$ *with* $Q \ll \alpha$. *Then, for any positive integer* m *and measurable sets* $A_1, \cdots, A_m$, *and* $\varepsilon > 0$,

$$\mathscr{P}\{|P(A_i) - Q(A_i)| < \varepsilon \text{ for } i = 1, \cdots, m\} > 0 .$$

PROOF. Form $B_{\nu_1,\cdots,\nu_m}$ as in (1) and note that

$$\mathscr{P}\{|P(A_i) - Q(A_i)| < \varepsilon \text{ for } i = 1, \cdots, m\}$$
$$\geqq \mathscr{P}\{\textstyle\sum_{(\nu_1,\cdots,\nu_m)\ni\nu_i=1} |P(B_{\nu_1,\cdots,\nu_m}) - Q(B_{\nu_1,\cdots,\nu_m})| < \varepsilon \text{ for } i = 1, \cdots, m\} .$$

Therefore, it is sufficient to show

$$\mathscr{P}\{|P(B_{\nu_1,\cdots,\nu_m}) - Q(B_{\nu_1,\cdots,\nu_m})| < 2^{-m}\varepsilon \text{ for all } (\nu_1, \cdots, \nu_m)\} > 0 .$$

If $\alpha(B_{\nu_1,\cdots,\nu_m}) = 0$, then $Q(B_{\nu_1,\cdots,\nu_m}) = 0$ and $P(B_{\nu_1,\cdots,\nu_m}) = 0$ with probability one, so that $|P(B_{\nu_1,\cdots,\nu_m}) - Q(B_{\nu_1,\cdots,\nu_m})| = 0$ with probability one. For those $(\nu_1, \cdots, \nu_m)$ for which $\alpha(B_{\nu_1,\cdots,\nu_m}) > 0$, the distribution of the corresponding $P(B_{\nu_1,\cdots,\nu_m})$ gives positive weight to all open sets in the set

$$\textstyle\sum_{(\nu_1,\cdots,\nu_m)\ni\alpha(B_{\nu_1,\cdots,\nu_m})>0} P(B_{\nu_1,\cdots,\nu_m}) = 1$$

completing the proof.

This proposition is a version of the desirable property (I) mentioned in the introduction. To discuss the support of a random probability measure, the topology on the space of probability measures on $(\mathscr{X}, \mathscr{A})$ must be specified. If the topology is chosen to be that of pointwise convergence ($Q_n \to Q$, if for every $A \in \mathscr{A}$, $Q_n(A) \to Q(A)$), Proposition 3 states that the support of the Dirichlet process on $(\mathscr{X}, \mathscr{A})$ with parameter α contains the set of all probability measures absolutely continuous with respect to α. It is easy to see conversely, that any measure Q not absolutely continuous with respect to α is not in the support of P.

If $(\mathscr{X}, \mathscr{A})$ is the real line with the Borel sets, we may consider the topology of convergence in distribution on the set of σ-additive probability measures on $(\mathscr{X}, \mathscr{A})$. With this topology, it may be shown that if α is σ-additive, the support of P is the set of all σ-additive probability measures whose support is contained in the support of α.

DEFINITION 2. Let P be a random probability measure on $(\mathscr{X}, \mathscr{A})$. We say that $X_1, \cdots, X_n$ is a sample of size n from P if for any $m = 1, 2, \cdots$ and measurable sets $A_1, \cdots, A_m, C_1, \cdots, C_n$,

$$\mathscr{P}\{X_1 \in C_1, \cdots, X_n \in C_n \mid P(A_1), \cdots, P(A_m), P(C_1), \cdots, P(C_n)\} = \prod_{j=1}^{n} P(C_j) \quad \text{a.s.} \tag{7}$$

Roughly, $X_1, \cdots, X_n$ is a sample of size n from P, if, given $P(C_1), \cdots, P(C_n)$, the events $\{X_1 \in C_1\}, \cdots, \{X_n \in C_n\}$ are independent of the rest of the process, and are independent among themselves, with $\mathscr{P}\{X_j \in C_j \mid P(C_1), \cdots, P(C_n)\} = P(C_j)$ a.s. for $j = 1, \cdots, n$. This definition determines the joint distribution of $X_1, \cdots, X_n, P(A_1), \cdots, P(A_m)$, once the distribution of the process is given, since

$$\mathscr{P}\{X_1 \in C_1, \cdots, X_n \in C_n, P(A_1) \leqq y_1, \cdots, P(A_m) \leqq y_m\} \tag{8}$$

may be found by integrating (7) with respect to the joint distribution of $P(A_1), \cdots, P(A_m), P(C_1), \cdots, P(C_n)$ over the set $[0, y_1] \times \cdots \times [0, y_m] \times [0, 1] \times \cdots \times [0, 1]$. The Kolmogorov consistency conditions may easily be checked to show that (8) determines a probability $\mathscr{P}$ over $(\mathscr{X}^n \times [0, 1]^{\mathscr{A}}, \mathscr{A}^n \times B\mathscr{T}^{\mathscr{A}})$.

PROPOSITION 4. *Let P be a Dirichlet process on $(\mathscr{X}, \mathscr{A})$ with parameter α and let X be a sample of size 1 from P. Then for $A \in \mathscr{A}$,*

$$\mathscr{P}(X \in A) = \alpha(A)/\alpha(\mathscr{X}) .$$

PROOF. Since $\mathscr{P}(X \in A \mid P(A)) = P(A)$ a.s.,

$$\begin{aligned} \mathscr{P}(X \in A) &= \mathscr{E}\mathscr{P}(X \in A \mid P(A)) \\ &= \mathscr{E}P(A) \\ &= \alpha(A)/\alpha(\mathscr{X}) , \end{aligned}$$

completing the proof.

PROPOSITION 5. *Let P be a Dirichlet process on $(\mathscr{X}, \mathscr{A})$ with parameter α, and let X be a sample of size 1 from P. Let $(B_1, \cdots, B_k)$ be a measurable partition of*

$\mathscr{X}$, and let $A \in \mathscr{A}$. Then,

$$(9) \qquad \mathscr{P}\{X \in A, P(B_1) \leqq y_1, \cdots, P(B_k) \leqq y_k\} = \sum_{j=1}^{k} \frac{\alpha(B_j \cap A)}{\alpha(\mathscr{X})} D(y_1, \cdots, y_k \mid \alpha_1^{(j)}, \cdots, \alpha_k^{(j)})$$

where $D(y_1, \cdots, y_k \mid \alpha_1, \cdots, \alpha_k)$ is the distribution function of the Dirichlet distribution, $\mathscr{D}(\alpha_1, \cdots, \alpha_k)$, and where

$$\alpha_i^{(j)} = \alpha(B_i) \quad \text{if} \quad i \neq j$$
$$= \alpha(B_j) + 1 \quad \text{if} \quad i = j.$$

PROOF. Define $B_{j,1} = B_j \cap A$, and $B_{j,0} = B_j \cap A^c$ for $j = 1, \cdots, k$. Let $Y_{j,\nu} = P(B_{j,\nu})$ for $j = 1, \cdots, k$ and $\nu = 0, 1$. Then, from (7)

$$(10) \qquad \mathscr{P}\{X \in A \mid Y_{j,\nu}, j = 1, \cdots, k, \text{ and } \nu = 0, 1\} = \sum_{j=1}^{k} Y_{j,1} \quad \text{a.s.}$$

Hence for arbitrary $y_{j,\nu} \in [0, 1]$, for $j = 1, \cdots, k$ and $\nu = 0, 1$,

$$\mathscr{P}\{X \in A, Y_{j,\nu} \leqq y_{j,\nu} \text{ for } j = 1, \cdots, k, \text{ and } \nu = 0, 1\}$$

can be found by integrating (10) with respect to the distribution of the $Y_{j,\nu}$ over the set $\{Y_{j,\nu} \leqq y_{j,\nu}, j = 1, \cdots, k \text{ and } \nu = 0, 1\}$. This integration turns out to be (see (4) of Section 2)

$$\sum_{j=1}^{k} \frac{\alpha(B_{j,1})}{\alpha(\mathscr{X})} D(\mathbf{y} \mid \boldsymbol{\alpha}^{(j)})$$

where $\mathbf{y} = (y_{1,0}, \cdots, y_{k,0}, y_{1,1}, \cdots, y_{k,1})$ and $\boldsymbol{\alpha}^{(j)} = (\alpha_{1,0}^{(j)}, \cdots, \alpha_{k,0}^{(j)}, \alpha_{1,1}^{(j)}, \cdots, \alpha_{k,1}^{(j)})$, and where

$$\alpha_{i,\nu}^{(j)} = \alpha(B_{i,\nu}) \quad \text{if} \quad i \neq j$$
$$= \alpha(B_{j,\nu}) + 1 \quad \text{if} \quad i = j.$$

The conclusion of the proposition follows from this using property i° of the Dirichlet distribution, since $P(B_j) = Y_{j,0} + Y_{j,1}$ a.s., and since the process of finding marginal distributions of random variables is linear.

We are now prepared to find the conditional distribution of a Dirichlet process P, given a sample $X_1, \cdots, X_n$ from P. It turns out that this conditional distribution is also a Dirichlet process.

For $x \in \mathscr{X}$, let δ_x denote the measure on $(\mathscr{X}, \mathscr{A})$ giving mass one to the point x:

$$\delta_x(A) = 1 \quad \text{if} \quad x \in A$$
$$= 0 \quad \text{if} \quad x \notin A.$$

THEOREM 1. *Let P be a Dirichlet process on $(\mathscr{X}, \mathscr{A})$ with parameter α, and let $X_1, \cdots, X_n$ be a sample of size n from P. Then the conditional distribution of P given $X_1, \cdots, X_n$, is as a Dirichlet process with parameter $\alpha + \sum_1^n \delta_{X_i}$.*

PROOF. It is sufficient to prove the theorem for $n = 1$, since the theorem would then follow by induction upon repeated application of the case $n = 1$. Let $(B_1, \cdots, B_k)$ be a measurable partition of $\mathscr{X}$ and let $A \in \mathscr{A}$. It is easy to

check that the marginal distributions of a conditional distribution of a process are identical to the conditional distributions of the marginals. Hence, we must show that the conditional distribution of $P(B_1), \cdots, P(B_k)$ given X, a sample of size one from P, has distribution function

$$D(y_1, \cdots, y_k \mid \alpha(B_1) + \delta_X(B_1), \cdots, \alpha(B_k) + \delta_X(B_k)) . \tag{11}$$

This may be done by showing that the integral of (11) with respect to the marginal distribution of X over A is equal to the probability (9). Using the marginal distribution of X as found in Proposition 4, we compute

$$\begin{aligned}
&\int_A D(y_1, \cdots, y_k \mid \alpha(B_1) + \delta_x(B_1), \cdots, \alpha(B_k) + \delta_x(B_k))\, d\alpha(x)/\alpha(\mathscr{X}) \\
&\quad = \sum_{j=1}^{k} \int_{B_j \cap A} D(y_1, \cdots, y_k \mid \alpha_1^{(j)}, \cdots, \alpha_k^{(j)})\, d\alpha(x)/\alpha(\mathscr{X}) \\
&\quad = \sum_{j=1}^{k} \frac{\alpha(B_j \cap A)}{\alpha(\mathscr{X})} D(y_1, \cdots, y_k \mid \alpha_1^{(j)}, \cdots, \alpha_k^{(j)}) ,
\end{aligned}$$

completing the proof.

4. An alternative definition of the Dirichlet process. In this section, we define a random probability measure which is a Dirichlet process on $(\mathscr{X}, \mathscr{A})$ with parameter α and which with probability one is a discrete probability measure on $(\mathscr{X}, \mathscr{A})$.

The basic idea is that since the Dirichlet distribution is definable, as in (2) of Section 2, as the joint distribution of a set of independent gamma variables divided by the sum, so also should the Dirichlet process be definable as a gamma process with independent "increments" divided by the sum. Using a representation of a process with independent increments as a sum of a countable number of jumps of random height at a countable number of random points, as found in [4], we may divide by the total heights of the jumps and obtain a discrete probability measure, which should be distributed as a Dirichlet process.

The gamma distribution, $\mathscr{G}(\alpha, 1)$, $\alpha > 0$, has characteristic function (see Gnedenko and Kolmogorov ([6] pages 86–87)),

$$\begin{aligned}
\varphi(a) &= (1 - it)^{-\alpha} \\
&= \exp \int_0^\infty (e^{iux} - 1)\, dN(x)
\end{aligned} \tag{1}$$

where

$$N(x) = -\alpha \int_x^\infty e^{-y} y^{-1}\, dy \qquad \text{for } 0 < x < \infty . \tag{2}$$

We define the distribution of random variables $J_1, J_2, \cdots$ as follows.

$$\mathscr{P}(J_1 \leqq x_1) = e^{N(x_1)} \qquad \text{for } x_1 > 0 \tag{3}$$

and for $j = 2, 3, \cdots$

$$\mathscr{P}(J_j \leqq x_j \mid J_{j-1} = x_{j-1}, \cdots, J_1 = x_1) = \exp[N(x_j) - N(x_{j-1})] \qquad \text{for } 0 < x_j < x_{j-1} . \tag{4}$$

In other words, the distribution function of J_1 is $\exp N(x_1)$ and for $j = 2, 3, \cdots$,

the distribution of J_j given $J_{j-1}, \cdots, J_1$, is the same as the distribution of J_1 truncated above at J_{j-1}. The following theorem is taken from the main theorem of [4].

THEOREM 1. *Let $G(t)$ be a distribution function on* $[0, 1]$. *Let*

$$Z_t = \sum_{j=1}^{\infty} J_j I_{[0, G(t))}(U_j) , \tag{5}$$

where (i) *the distribution of $J_1, J_2, \cdots$ is given in* (3) *and* (4), *and*

(ii) *$U_1, U_2 \cdots$ are independent identically distributed variables, uniformly distributed on* $[0, 1]$, *and independent of $J_1, J_2, \cdots$. Then, with probability one, Z_t converges for all $t \in [0, 1]$ and is a gamma process with independent increments, with $Z_t \in \mathcal{G}(\alpha G(t), 1)$.*

In particular, $Z_1 = \sum_1^{\infty} J_j$ converges with probability one and $Z_1 \in \mathcal{G}(\alpha, 1)$. If we define

$$P_j = J_j / Z_1 , \tag{6}$$

then $P_j \geqq 0$ and $\sum_1^{\infty} P_j = 1$ with probability one.

We now define the Dirichlet process. As before, let $(\mathcal{X}, \mathcal{A})$ be a measurable space, and let $\alpha(\cdot)$ be a finite non-null measure on $\mathcal{A}$. Let $V_1, V_2, \cdots$ be a sequence of independent identically distributed random variables with values in $\mathcal{X}$, and with probability measure Q, where $Q(A) = \alpha(A)/\alpha(\mathcal{X})$. (More specifically, let $(\mathcal{X}_j, \mathcal{A}_j, Q_j)$ be identical copies of $(\mathcal{X}, \mathcal{A}, Q)$, and let V_j be the identity map from $\mathcal{X}_j$ to $\mathcal{X}$. Then the V_j are extended to be defined on the infinite product space $(\prod \mathcal{X}_j, \prod \mathcal{A}_j, \prod Q_j)$ in the usual manner.)

We identify the α in formulas (1) and (2) with $\alpha(\mathcal{X})$, and define the random probability measure, P, on $(\mathcal{X}, \mathcal{A})$, as

$$P(A) = \sum_{j=1}^{\infty} P_j \delta_{V_j}(A) . \tag{7}$$

THEOREM 2. *The random probability measure defined by* (7) *is a Dirichlet process on $(\mathcal{X}, \mathcal{A})$ with parameter α.*

PROOF. Let $(B_1, \cdots, B_k)$ be a measurable partition of $\mathcal{X}$. Then

$$(P(B_1), \cdots, P(B_k)) = \frac{1}{Z_1} \sum_{j=1}^{\infty} J_j(\delta_{V_j}(B_1), \cdots, \delta_{V_j}(B_k)) .$$

From the assumption on the distribution of $V_1, V_2, \cdots$,

$$\mathbf{M}_j = (\delta_{V_j}(B_1), \cdots, \delta_{V_j}(B_k))$$

are independent identically distributed random vectors having a multinomial distribution with probability vector $(Q(B_1), \cdots, Q(B_k))$. Hence, the distribution of $\sum_1^{\infty} J_j \mathbf{M}_j$ must be the same as the distribution of

$$(Z_{1/k}, Z_{2/k} - Z_{1/k}, \cdots, Z_1 - Z_{(k-1)/k})$$

where Z_t is the gamma process defined by (5) with $G(t)$ chosen so that

$$G\left(\frac{j}{k}\right) - G\left(\frac{j-1}{k}\right) = Q(B_j) \qquad j = 1, \cdots, k .$$

Hence, $\sum_{j=1}^{\infty} J_j \delta_{V_j}(B_i)$ are, for $i = 1, \cdots, k$, independent random variables, with $\sum_{j=1}^{\infty} J_j \delta_{V_j}(B_i) \in \mathcal{G}(\alpha(B_i), 1)$. Since Z_1 is the sum of these independent gamma variables, $(P(B_1), \cdots, P(B_k)) \in \mathcal{D}(\alpha(B_1), \cdots, \alpha(B_k))$ from the definition of the Dirichlet distribution. Thus, P satisfies the definition of the Dirichlet process.

THEOREM 3. *Let P be the Dirichlet process defined by* (7), *and let Z be a measurable real valued function defined on $(\mathcal{X}, \mathcal{A})$. If $\int |Z|\, d\alpha < \infty$, then $\int |Z|\, dP < \infty$ with probability one, and*

$$\mathcal{E} \int Z\, dP = \int Z\, d\mathcal{E}P = \alpha(\mathcal{X})^{-1} \int Z\, d\alpha .$$

PROOF. From (7)

$$\int |Z|\, dP = \sum_{j=1}^{\infty} |Z(V_j)| P_j \tag{8}$$

so that the monotone convergence theorem gives

$$\begin{aligned} \mathcal{E} \int |Z|\, dP &= \sum_{j=1}^{\infty} \mathcal{E}|Z(V_j)| \mathcal{E}P_j \\ &= \alpha(\mathcal{X})^{-1} \int |Z|\, d\alpha \sum_{j=1}^{\infty} \mathcal{E}P_j \\ &= \alpha(\mathcal{X})^{-1} \int |Z|\, d\alpha \end{aligned}$$

where we have used the independence of the V_j and the P_j. Therefore, $\int |Z|\, dP$ is finite with probability one, and hence

$$\int Z\, dP = \sum_{j=1}^{\infty} Z(V_j) P_j$$

is absolutely convergent with probability one. Since this series is bounded by (8), which is integrable, the bounded convergence theorem implies

$$\begin{aligned} \mathcal{E} \int Z\, dP &= \sum_{j=1}^{\infty} \mathcal{E}Z(V_j) \mathcal{E}P_j \\ &= \alpha(\mathcal{X})^{-1} \int Z\, d\alpha \end{aligned}$$

completing the proof.

This theorem emphasises the close relationship between α and the random probability measure P. It implies, in particular, that if $(\mathcal{X}, \mathcal{A})$ were the real line and the Borel sets, and if α has a finite kth moment, then with probability one P has a finite kth moment.

THEOREM 4. *Let P be the Dirichlet process defined by* (7), *and let Z_1 and Z_2 be measurable real valued functions defined on $(\mathcal{X}, \mathcal{A})$. If $\int |Z_1|\, d\alpha < \infty$, $\int |Z_2|\, d\alpha < \infty$ and $\int |Z_1 Z_2|\, d\alpha < \infty$, then*

$$\mathcal{E} \int Z_1\, dP \int Z_2\, dP = \frac{\sigma_{12}}{\alpha(\mathcal{X}) + 1} + \mu_1 \mu_2$$

where

$$\mu_i = \alpha(\mathcal{X})^{-1} \int Z_i\, d\alpha \qquad i = 1, 2 \text{ and}$$

$$\sigma_{12} = \alpha(\mathcal{X})^{-1} \int Z_1 Z_2\, d\alpha - \mu_1 \mu_2 .$$

PROOF. As in Theorem 3

$$\begin{aligned} \int Z_1\, dP \int Z_2\, dP &= \sum_i Z_1(V_i) P_i \sum_j Z_2(V_j) P_j \\ &= \sum_i \sum_j Z_1(V_i) Z_2(V_j) P_i P_j \end{aligned} \tag{11}$$

since both series are absolutely convergent with probability one. This is bounded in absolute value by

$$\sum_i \sum_j |Z_1(V_i)Z_2(V_j)|P_i P_j \,. \tag{12}$$

If this is an integrable random variable, we may take an expectation of (11) inside the summation sign and obtain

$$\begin{aligned}\mathscr{E} \int Z_1\,dP \int Z_2\,dP &= \sum_i \sum_j \mathscr{E}(Z_1(V_i)Z_2(V_j))\mathscr{E}(P_i P_j) \\ &= \sum_{i\neq j}\sum \mathscr{E}Z_1(V_i)\mathscr{E}Z_2(V_j)\mathscr{E}(P_i P_j) \\ &\quad + \sum_i \mathscr{E}(Z_1(V_i)Z_2(V_i))\mathscr{E}P_i^2\end{aligned}$$

using the independence of the P_i and the V_i, and the independence of the V_i among themselves. The equation continues

$$\begin{aligned}&= \mu_1\mu_2 \sum_{i\neq j}\sum \mathscr{E}(P_i P_j) + (\sigma_{12} + \mu_1\mu_2)\sum_i \mathscr{E}P_i^2 \\ &= \mu_1\mu_2 + \sigma_{12}\mathscr{E}\sum_i P_i^2 \,.\end{aligned}$$

An analogous equation shows that (12) is integrable. The proof will be complete when we show

$$\mathscr{E}\sum_{j=1}^{\infty} P_i^2 = \frac{1}{\alpha(\mathscr{X}) + 1} \,.$$

This seems difficult to show directly from the definition of the P_i, so we proceed as follows. The distribution of the P_i depends on α only through the value of $\alpha(\mathscr{X})$. So choose $\mathscr{X}$ to be the real line, α to give mass $\alpha(\mathscr{X})/2$ to -1 and mass $\alpha(\mathscr{X})/2$ to $+1$, and $Z_1(x) = Z_2(x)$ to be identically x. Then $\mu_1 = \mu_2 = 0$ and $\sigma_{12} = 1$. Hence

$$\begin{aligned}\mathscr{E}\sum_1^{\infty} P_i^2 = \mathscr{E}(\int x\,dP(x))^2 &= \mathscr{E}(2P(\{1\}) - 1)^2 \\ &= \frac{1}{\alpha(\mathscr{X}) + 1}\end{aligned}$$

since $P(\{1\}) \in \mathscr{B}e(\alpha(\mathscr{X})/2, \alpha(\mathscr{X})/2)$, completing the proof.

This theorem states that the covariance of the random variables $\int Z_1\,dP$ and $\int Z_2\,dP$ is equal to $(\alpha(\mathscr{X}) + 1)^{-1}$ times the covariance of Z_1 and Z_2 as random variables on $(\mathscr{X}, \mathscr{A}, Q)$ where $Q = \mathscr{E}P = \alpha/\alpha(\mathscr{X})$. In particular, the correlation coefficient of $\int Z_1\,dP$ and $\int Z_2\,dP$ is equal to the correlation coefficient of Z_1 and Z_2 as random variables on $(\mathscr{X}, \mathscr{A}, Q)$ (assuming the finiteness of $\int Z_1^2\,d\alpha$ and $\int Z_2^2\,d\alpha$).

5. Applications. Throughout this section, α is taken to denote a σ-additive non-null finite measure on $(\mathscr{X}, \mathscr{A})$. We write $P \in \mathscr{D}(\alpha)$ as a notation for the phrase "P is a Dirichlet process on $(\mathscr{X}, \mathscr{A})$ with parameter α." We let $\mathbb{R}$ denote the real line and $\mathscr{B}$ the σ-field of Borel sets. In most of the applications we take $(\mathscr{X}, \mathscr{A}) = (\mathbb{R}, \mathscr{B})$.

The nonparametric statistical decision problems we consider are typically

described as follows. The parameter space is the set of all probability measures P on $(\mathscr{X}, \mathscr{A})$. The statistician is to choose an action a in some space, thereby incurring a loss, $L(P, a)$. There is a sample $X_1, \cdots, X_n$ from P available to the statistician, upon which he may base his choice of action. He seeks a Bayes rule with respect to the prior distribution, $P \in \mathscr{D}(\alpha)$.

With such a prior distribution, the posterior distribution of P given the observations is $\mathscr{D}(\alpha + \sum_1^n \delta_{X_i})$, where δ_x denotes the measure giving mass one to the point x. Thus, if we can find a Bayes rule for the no-sample problem (with $n = 0$), a Bayes rule for the general problem may be found by replacing α with $\alpha + \sum_1^n \delta_{X_i}$. In the problems considered below, we first find the Bayes rule for the no-sample problem, and then state the Bayes rule for the general problem.

(a) *Estimation of a distribution function.* Let $(\mathscr{X}, \mathscr{A}) = (\mathbb{R}, \mathscr{B})$, and let the space of actions of the statistician be the space of all distribution functions on $\mathbb{R}$. Let the loss function be

$$L(P, \hat{F}) = \int (F(t) - \hat{F}(t))^2 \, dW(t)$$

where W is a given finite measure on $(\mathbb{R}, \mathscr{B})$ (a weight function), and where

$$F(t) = P((-\infty, t]) \,. \tag{1}$$

If $P \in \mathscr{D}(\alpha)$, then $F(t) \in \mathscr{B}e(\alpha((-\infty, t]), \alpha((t, \infty)))$ for each t. The Bayes risk for the no-sample problem,

$$\mathscr{E}L(P, \hat{F}) = \int \mathscr{E}(F(t) - \hat{F}(t))^2 \, dW(t) \,,$$

is minimized by choosing $\hat{F}(t)$ for each t to minimize $\mathscr{E}(F(t) - \hat{F}(t))^2$. This is achieved by choosing $\hat{F}(t)$ to be $\mathscr{E}F(t)$. Thus, the Bayes rule for the no-sample problem is

$$\hat{F}(t) = \mathscr{E}F(t) = F_0(t)$$

where

$$F_0(t) = \alpha((-\infty, t])/\alpha(\mathbb{R}) \tag{2}$$

represents our prior guess at the shape of the unknown $F(t)$.

For a sample of size n, the Bayes rule is therefore

$$\begin{aligned} \hat{F}_n(t \mid X_1, \cdots, X_n) &= \frac{\alpha((-\infty, t]) + \sum_1^n \delta_{X_i}((-\infty, t])}{\alpha(\mathbb{R}) + n} \\ &= p_n F_0(t) + (1 - p_n) F_n(t \mid X_1, \cdots, X_n) \end{aligned} \tag{3}$$

where

$$p_n = \alpha(\mathbb{R})/(\alpha(\mathbb{R}) + n) \tag{4}$$

and

$$F_n(t \mid X_1, \cdots, X_n) = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}((-\infty, t])$$

is the empirical distribution function of the sample.

The Bayes rule (3) is a mixture of our prior guess at F and of the empirical distribution function, with respective weights p_n and $(1 - p_n)$. If $\alpha(\mathbb{R})$ is large compared to n, little weight is given to the observations. If $\alpha(\mathbb{R})$ is small compared to n, little weight is given to the prior guess at F. One might interpret $\alpha(\mathbb{R})$ as a measure of faith in the prior guess at F measured in units of numbers of observations. As $\alpha(\mathbb{R})$ tends to zero (the "noninformative" Dirichlet prior), the Bayes estimate converges to the empirical distribution function.

It is interesting to note that whatever be the true distribution function, the Bayes estimate (3) converges to it uniformly almost surely. This follows from the Glivenko–Cantelli theorem and the observation that $p_n \to 0$ as $n \to \infty$.

The results for estimating a k-dimensional distribution function are completely analogous.

(b) *Estimation of the mean.* Again let $(\mathscr{X}, \mathscr{A}) = (\mathbb{R}, \mathscr{B})$, and suppose the statistician is to estimate the mean with squared error loss

$$L(P, \hat{\mu}) = (\mu - \hat{\mu})^2$$

where

$$\mu = \int x \, dP(x) . \tag{5}$$

We assume $P \in \mathscr{D}(\alpha)$, where α has finite first moment. The mean of the corresponding probability measure $\alpha(\cdot)/\alpha(\mathbb{R})$ is denoted by μ_0:

$$\mu_0 = \int x \, d\alpha(x)/\alpha(\mathbb{R}) . \tag{6}$$

By Theorem 3, the random variable μ defined by (5) exists. The Bayes rule for the no-sample problem is the mean of μ, which, again by Theorem 3, is $\hat{\mu} = \mu_0$.

For a sample of size n, the Bayes rule is therefore

$$\begin{aligned} \hat{\mu}_n(X_1, \cdots, X_n) &= (\alpha(\mathbb{R}) + n)^{-1} \int x \, d(\alpha(x) + \textstyle\sum_1^n \delta_{X_i}(x)) \\ &= p_n \mu_0 + (1 - p_n)\bar{X}_n \end{aligned} \tag{7}$$

where p_n is given by (4) and $\bar{X}_n$ is the sample mean,

$$\bar{X}_n = \frac{1}{n} \textstyle\sum_{i=1}^n X_i .$$

The Bayes estimate is thus between the prior guess at μ, namely μ_0, and the sample mean. As $\alpha(\mathbb{R}) \to 0$, $\hat{\mu}_n$ converges to $\bar{X}_n$. Also, as $n \to \infty$, $p_n \to 0$ so that, in particular, the Bayes estimate (7) is strongly consistent within the class of distributions with finite first moment.

More generally, for arbitrary $(\mathscr{X}, \mathscr{A})$, if Z is real-valued measurable defined on $(\mathscr{X}, \mathscr{A})$, and if we are to estimate

$$\theta = \int Z \, dP$$

with squared error loss and prior $P \in \mathscr{D}(\alpha)$, where α is such that

$$\theta_0 = \int Z \, d\alpha/\alpha(\mathscr{X}) < \infty ,$$

then the estimate $\hat{\theta} = \theta_0$ is Bayes for the no-sample problem. For a sample of size n,

$$\hat{\theta}_n(X_1, \cdots, X_n) = p_n \theta_0 + (1 - p_n) \frac{1}{n} \sum_1^n Z(X_i)$$

is Bayes, where $p_n = \alpha(\mathscr{X})/(\alpha(\mathscr{X}) + n)$. Results for estimating a mean vector in k-dimensions are completely analogous.

(c) *Estimation of the median.* Again let $(\mathscr{X}, \mathscr{A}) = (\mathbb{R}, \mathscr{B})$ and suppose the statistician is to estimate the median

$$m = \text{med } P \tag{8}$$

of an unknown probability measure P on $(\mathbb{R}, \mathscr{B})$. If $P \in \mathscr{D}(\alpha)$, the median of P is unique with probability one. To see this, note that $F(t)$, defined by (1), increases whenever α increases, with probability one. Any multiple medians of P must occur on an interval of measure zero of α. There are only a countable number of such intervals, and the probability that any such interval is an interval of medians is just $\mathscr{P}\{F(t) = \frac{1}{2}\}$ for t any interior point of the interval. This is zero since $F(t)$ has a Beta distribution. Thus, for $P \in \mathscr{D}(\alpha)$, m defined by (8) is a random variable.

The Bayes estimate of m for the no-sample problem with Dirichlet process prior and squared error loss is the expectation of m. Unfortunately, this expectation is difficult to compute, and may, in fact, not even exist. Instead, we seek the Bayes estimate of m under absolute error loss,

$$L(P, \hat{m}) = |m - \hat{m}| .$$

As is well known, any median of the distribution of m is a Bayes estimate of m. For the Dirichlet process prior, any median of the distribution of m is a median of the expectation of P, and conversely:

$$\text{med (dist. med } P) = \text{med } \mathscr{E}P . \tag{9}$$

This may be seen as follows. A number t is a median of the distribution of m if and only if

$$\mathscr{P}\{m < t\} \leqq \tfrac{1}{2} \leqq \mathscr{P}\{m \leqq t\} . \tag{10}$$

Since $\mathscr{P}\{m \leqq t\} = \mathscr{P}\{F(t) > \frac{1}{2}\}$ by the definition of m, and since $F(t)$ has a $\mathscr{B}e(\alpha((-\infty, t]), \alpha((t, \infty)))$ distribution, whose median is a non-decreasing function of t with value one-half if and only if the two parameters are equal, we see that t satisfies (10) if and only if

$$\frac{\alpha((-\infty, t))}{\alpha(\mathbb{R})} \leqq \tfrac{1}{2} \leqq \frac{\alpha((-\infty, t])}{\alpha(\mathbb{R})} . \tag{11}$$

Such t are exactly the medians of $\mathscr{E}P$, proving (9).

Thus, any number t satisfying (11) is a Bayes estimate of m for prior $\mathscr{D}(\alpha)$ and absolute error loss. For F_0 defined by (2),

$$\hat{m} = \text{median of } \quad F_0 .$$

For a sample of size n, the Bayes estimate is therefore

$$\hat{m}_n(X_1, \cdots, X_n) = \text{median of} \quad \hat{F}_n$$

where $\hat{F}_n$ is the Bayes estimate of F given by (3).

(d) *Estimation of quantiles.* We extend the analysis of part (c) to the estimation of the qth quantile of P, denoted by t_q:

$$P((-\infty, t_q)) \leqq q \leqq P((-\infty, t_q]) .$$

As in the case of the median, it is easy to see that for $0 < q < 1$, the qth quantile of $P \in \mathscr{D}(\alpha)$ is unique with probability one, so that t_q is a well-defined random variable.

We consider the problem of estimating t_q with loss for some p, $0 < p < 1$,

$$\begin{aligned} L(P, \hat{t}_q) &= p(t_q - \hat{t}_q) && \text{if} \quad t_q \geqq \hat{t}_q \\ &= (1 - p)(\hat{t}_q - t_q) && \text{if} \quad t_q < \hat{t}_q . \end{aligned}$$

As is well known, any pth quantile of the distribution of t_q is a Bayes estimate of t_q under this loss. The distribution of t_q may be found from the formula

$$\begin{aligned} \mathscr{P}\{t_q \leqq t\} &= \mathscr{P}\{F(t) > q\} \\ &= \int_q^1 \frac{\Gamma(M)}{\Gamma(uM)\Gamma((1-u)M)} z^{uM-1}(1-z)^{(1-u)M-1}\, dz \end{aligned} \tag{12}$$

where

$$\begin{aligned} M &= \alpha(\mathbb{R}) \\ u &= \alpha((-\infty, t])/\alpha(\mathbb{R}) = F_0(t) . \end{aligned}$$

To find the pth quantile of t_q, we set (12) equal to p and solve for t.

$$\int_q^1 \frac{\Gamma(M)}{\Gamma(uM)\Gamma((1-u)M)} z^{uM-1}(1-z)^{(1-u)M-1}\, dz = p . \tag{13}$$

For fixed p, q, and M, we let (13) define a function $u(p, q, M)$. The Bayes estimate of t_q for the no-sample problem is the uth quantile of F_0,

$$\hat{t}_q = u(p, q, \alpha(\mathbb{R}))\text{th quantile of} \quad F_0 .$$

For a sample of size n, the Bayes estimate of t_q is therefore

$$\hat{t}_q(X_1, \cdots, X_n) = u(p, q, \alpha(\mathbb{R}) + n)\text{th quantile of} \quad \hat{F}_n \tag{14}$$

where $\hat{F}_n$ is the Bayes estimate of F given by (3). If p and q are both $\frac{1}{2}$, this reduces to the estimate of (c), since $u(\frac{1}{2}, \frac{1}{2}, M) = \frac{1}{2}$ for all M, as is seen from (13).

If tables of the function $u(p, q, M)$ were available, it would be an easy matter to find the estimate (14). Unfortunately, it is difficult to obtain values of $u(p, q, M)$ from existing tables of the incomplete Beta function. The author has tables of u for $p = .05(.05).95$, $q = .05(.05).95$, and $M = 1(1)10$.

(e) *Estimation of a variance and a covariance.* Consider the problem of

estimating the variance of an unknown probability distribution P with squared error loss

$$L(P, \hat{\sigma}^2) = (\text{Var } P - \hat{\sigma}^2)^2 .$$

If $P \in \mathscr{D}(\alpha)$, and if α has a finite second moment, then

$$\text{Var } P = \int x^2 \, dP(x) - (\int x \, dP(x))^2$$

is a random variable whose expectation (a Bayes estimate for the no-sample problem) may be computed from Theorems 3 and 4 as follows

$$\begin{aligned} \mathscr{E} \text{ Var } P &= \mathscr{E} \int x^2 P(x) - \mathscr{E}(\int x \, dP(x))^2 \\ &= (\sigma_0^2 + \mu_0^2) - \left(\frac{\sigma_0^2}{\alpha(\mathbb{R}) + 1} + \mu_0^2 \right) \\ &= \frac{\alpha(\mathbb{R})}{\alpha(\mathbb{R}) + 1} \sigma_0^2 \end{aligned}$$

where μ_0 is given by (6) and where

$$\sigma_0^2 = \alpha(\mathbb{R})^{-1} \int x^2 \, d\alpha(x) - \mu_0^2$$

is the variance of F_0, the prior guess at F.

For a sample of size n, the Bayes rule is therefore

$$\hat{\sigma}_n^2(X_1, \cdots, X_n) = \frac{\alpha(\mathbb{R}) + n}{\alpha(\mathbb{R}) + n + 1} \text{ Var } \hat{F}_n$$

where $\hat{F}_n$ is given by (3). Using an easy formula for the variance of a mixture, we find

$$\begin{aligned} \hat{\sigma}_n^2(X_1, \cdots, X_n) &= \frac{\alpha(\mathbb{R}) + n}{\alpha(\mathbb{R}) + n + 1} \text{ Var } (p_n F_0 + (1 - p_n) F_n) \\ &= \frac{\alpha(\mathbb{R}) + n}{\alpha(\mathbb{R}) + n + 1} (p_n \sigma_0^2 + (1 - p_n) s_n^2 \\ &\qquad + p_n(1 - p_n)(\mu_0 - \bar{X}_n)^2) \end{aligned} \tag{15}$$

where μ_0, σ_0^2, p_n, and $\bar{X}_n$ are as before, and where s_n^2 is the sample variance

$$s_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 .$$

An alternative form of the estimate (15) expresses $\hat{\sigma}_n^2$ as a mixture of three different estimates of the variance:

$$\begin{aligned} &\hat{\sigma}_n^2(X_1, \cdots, X_n) \\ &\quad = \frac{\alpha(\mathbb{R}) + n}{\alpha(\mathbb{R}) + n + 1} \left(p_n \sigma_0^2 + (1 - p_n) \left(p_n \frac{1}{n} \sum_1^n (X_i - \mu_0)^2 + (1 - p_n) s_n^2 \right) \right) . \end{aligned}$$

If we let our prior sample size $\alpha(\mathbb{R})$ tend to zero, keeping F_0 fixed, we find that $\hat{\sigma}_n^2$ converges to the estimate

$$\frac{1}{n + 1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 .$$

This estimate is well known as the best invariant or minimax estimate of the variance of a normal distribution under relative squared error loss (squared error divided by (Var P)^2). Its appearance in this problem is rather surprising.

To estimate the covariance of a distribution P in the plane,

$$\text{Cov } P = \int xy\, dP - \int x\, dP \int y\, dP\,,$$

a similar analysis may be carried out. Here, $\mathscr{X}$ represents the Euclidean plane $\mathbb{R}^2$, $\mathscr{A}$ represents the Borel subsets, and α represents a finite measure thereon. With squared error loss, the Bayes estimate of Cov P with respect to the prior $P \in \mathscr{D}(\alpha)$, is for the no-sample problem

$$\mathscr{E}\,\text{Cov } P = \frac{\alpha(\mathbb{R}^2)}{\alpha(\mathbb{R}^2) + 1}\,\sigma_{12}$$

where, as in Theorem 4, σ_{12} is the covariance of the distribution $\mathscr{E}P$,

$$\begin{aligned}
\sigma_{12} &= \alpha(\mathbb{R}^2)^{-1} \int xy\, d\alpha(x, y) - \mu_1 \mu_2 \\
\mu_1 &= \alpha(\mathbb{R}^2)^{-1} \int x\, d\alpha(x, y) \\
\mu_2 &= \alpha(\mathbb{R}^2)^{-1} \int y\, d\alpha(x, y)\,.
\end{aligned}$$

For a sample of size n, the Bayes estimate is

$$\begin{aligned}
\hat{\sigma}_{12}(X_1, Y_1, \cdots, X_n, Y_n) = \frac{\alpha(\mathbb{R}^2) + n}{\alpha(\mathbb{R}^2) + n + 1} &(p_n \sigma_{12} + (1 - p_n)s_{12} \\
&+ p_n(1 - p_n)(\mu_1 - \bar{X}_n)(\mu_2 - \bar{Y}_n))
\end{aligned}$$

where s_{12} is the sample covariance

$$s_{12} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)\,.$$

(f) *Estimation of* $\int F\, dG$ *for a two-sample problem.* Let F and G be two distribution functions on the real line, and let $X_1, \cdots, X_m$ be a sample from F and $Y_1, \cdots, Y_n$ a sample from G. Consider the problem of estimating probability that $X_1 \leqq Y_1$, denoted by Δ,

$$\Delta = \int F\, dG$$

with squared error loss. As a prior distribution for (F, G), we assume that F is the distribution function of $P_1 \in \mathscr{D}(\alpha_1)$, that G is the distribution function of $P_2 \in \mathscr{D}(\alpha_2)$ and that P_1 and P_2 are independent. A computation similar to that found in Theorems 3 and 4 shows that the Bayes rule for the no-sample problem is

$$\Delta_0 = \mathscr{E}\Delta = \int F_0\, dG_0$$

where $F_0 = \mathscr{E}F$ and $G_0 = \mathscr{E}G$.

Given the two samples, the Bayes rule is

$$\hat{\Delta}(X_1, \cdots, X_m, Y_1, \cdots, Y_n) = \int \hat{F}_m\, d\hat{G}_n$$

where $\hat{F}_m$ and $\hat{G}_n$ are the respective Bayes estimates of F and G, as in (3). This may be written

$$\hat{\Delta}(X_1, \cdots, X_m, Y_1, \cdots, Y_n) = p_{1,m}p_{2,n}\Delta_0 + p_{1,m}(1 - p_{2,n})\frac{1}{n}\sum_1^n F_0(Y_j)$$
$$+ (1 - p_{1,m})p_{2,n}\frac{1}{m}\sum_1^m (1 - G_0(X_i^-))$$
$$+ (1 - p_{1,m})(1 - p_{2,n})\frac{1}{mn}U$$

where $$p_{1,m} = \frac{\alpha_1(\mathbb{R})}{\alpha_1(\mathbb{R}) + m}, \qquad p_{2,n} = \frac{\alpha_2(\mathbb{R})}{\alpha_2(\mathbb{R}) + n}$$

and where U, the number of pairs (X_i, Y_j) for which $X_i \leqq Y_j$,

$$U = \sum_i \sum_j I_{(-\infty, Y_j]}(X_i)$$

is the Mann–Whitney statistic, a linear function of the rank-sum statistic. It is interesting to note that the estimate $\hat{\Delta}$ is a simple mixture of four separate estimates of Δ. As both $\alpha_1(\mathbb{R})$ and $\alpha_2(\mathbb{R})$ tend to zero, the estimate $\hat{\Delta}$ converges to $(mn)^{-1}U$, the usual nonparametric estimate.

(g) *"Tolerance" regions.* The notion of tolerance regions that we treat is not the usual one, but rather the decision theoretic analogue. Consider the problem of estimating the qth quantile t_q of an unknown distribution P on the real line by an upper "tolerance" point $\underline{a}$ with loss function

$$L(P, a) = pP((-\infty, a]) + qI_{(a,\infty)}(t_q)$$

where p is a constant $0 < p < 1$. If t_q is known exactly, L is minimized by choosing $\underline{a} = t_q$. But if t_q is only vaguely known, it is best to "overestimate" t_q to keep the expectation of the second term small, provided the expectation of the first term is not too enlarged.

If $P \in \mathscr{Q}(\alpha)$, the Bayes risk for the no-sample problem is

$$\begin{aligned} \mathscr{E}L(P, \underline{a}) &= p\mathscr{E}P((-\infty, \underline{a}]) + q\mathscr{P}\{t_q > \underline{a}\} \\ &= pF_0(\underline{a}) + q\mathscr{P}\{F(\underline{a}) \leqq q\} \\ &= pu + q\int_0^q \frac{\Gamma(M)}{\Gamma(uM)\Gamma((1-u)M)} z^{uM-1}(1 - z)^{(1-u)M-1}\,dz \end{aligned} \tag{16}$$

where u represents $F_0(a)$ and $M = \alpha(\mathbb{R})$. Since this Bayes risk depends on $\underline{a}$ only through $F_0(a)$, we seek that value of u in $[0, 1]$ that minimizes (16). It is not difficult to show uniqueness of the point at which the minimum occurs, but assuming uniqueness, and letting $u = f(p, q, M)$ denote the point at which the minimum occurs, the Bayes rule for the no-sample problem is

$$a = f(p, q, \alpha(\mathbb{R}))\text{th quantile} \quad F_0 .$$

For a sample of size n, the Bayes rule is

$$\hat{a}_n(X_1, \cdots, X_n) = f(p, q, \alpha(\mathbb{R}) + n)\text{th quantile of} \quad \hat{F}_n .$$

(h) *Tests of hypotheses involving quantiles.* Consider the problem of testing the hypothesis that the qth quantile t_q of an unknown distribution P on the real line does not exceed a given constant, taken without loss of generality to be zero. There are two actions available to the statistician, a_0 and a_1, with loss functions

$$L(P, a_0) = w_0 I_{(0,\infty)}(t_q)$$
$$L(P, a_1) = w_1 I_{(-\infty,0]}(t_q)$$

where w_0 and w_1 are positive constants. Suppose $P \in \mathscr{D}(\alpha)$. The Bayes rule for the no-sample problem is to select the action with smaller expected risk. This rule is: take action a_0 if

$$\mathscr{P}\{F(0) > q\} > \frac{w_0}{w_1 + w_0}$$

and take action a_1 otherwise. In terms of the function $u(p, q, M)$ defined by (13), we would take action a_0 if

$$u\left(\frac{w_0}{w_1 + w_0}, q, \alpha(\mathbb{R})\right) < F_0(0) .$$

For a sample of size n, the Bayes rule is: take action a_0 if

$$u\left(\frac{w_0}{w_0 + w_1}, q, \alpha(\mathbb{R}) + n\right) < p_n F_0(0) + (1 - p_n)\frac{1}{n} W_n$$

where p_n is an in (4), and where W_n is the number of X_i less than or equal to zero. This is essentially the sign test.

Since $u(\frac{1}{2}, \frac{1}{2}, M) = \frac{1}{2}$ for all M, the above formula simplifies when testing for the median with $w_0 = w_1$. This test becomes: accept the hypothesis that the median does not exceed zero if

$$W_n > \tfrac{1}{2}n + \alpha(\mathbb{R})(\tfrac{1}{2} - F_0(0)) .$$

Acknowledgments. The Dirichlet process was uncovered in the course of several long conversations with James B. MacQueen, who has an interesting way of viewing the Dirichlet process as a limit of Pólya urn schemes. The problem was suggested to me by David Blackwell, who is a great source of stimulating problems. Discussions with Charles Antoniak have helped clarify the ideas. To all three, my gratitude.

REFERENCES

[1] ANTONIAK, CHARLES (1969). Mixtures of Dirichlet processes with application to some nonparametric problems. Ph. D. Dissertation, Univ. of California, Los Angeles.

[2] DUBINS, LESTER E. and FREEDMAN, DAVID A. (1966). Random distribution functions. *Proc. Fifth Berkeley Symp. Math. Statist. Prob.* **2** 183–214. Univ. of California Press.

[3] FABIUS, J. (1964). Asymptotic behavior of Bayes' estimates. *Ann. Math. Statist.* **35** 846–856.

[4] FERGUSON, THOMAS S. and KLASS, MICHAEL J. (1972). A representation of independent increment processes without Gaussian components. *Ann. Math. Statist.* **43** 1634–1643.

[5] FREEDMAN, DAVID A. (1963). On the asymptotic behavior of Bayes' estimates in the discrete case. *Ann. Math. Statist.* **34** 1194–1216.

[6] GNEDENKO. B. V. and KOLMOGOROV, A. N. (1949). *Limit Distributions for Sums of Independent Random Variables*, trans. K. L. Chung (1954). Addison-Wesley, Reading.

[7] GOOD, I. J. (1965). *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*. Research Monograph No. 30, M.I.T. Press.

[8] KOLMOGOROV, A. N. (1933). *Foundations of the Theory of Probability*, 2nd ed., trans. Nathan Morrison (1956). Chelsea, New York.

[9] KRAFT, CHARLES H. (1964). A class of distribution function processes which have derivatives. *J. Appl. Probability* **1** 385–388.

[10] KRAFT, CHARLES H. and VAN EEDEN, CONSTANCE (1964). Bayesian bio-assay. *Ann. Math. Statist.* **35** 886–890.

[11] WILKS, SAMUEL S. (1962). *Mathematical Statistics*. Wiley, New York.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF CALIFORNIA
LOS ANGELES, CALIFORNIA 90024