The Annals of Statistics
1978, Vol. 6, No. 3, 649-657

BAYESIAN CONFIDENCE BANDS FOR A DISTRIBUTION FUNCTION

BY M. BRETH

Victoria, Australia

A set of recurrences is developed for computing the probability content of any prior or posterior confidence bands for a distribution function assuming the parameter of the prior Dirichlet process known. When the prior parameter is unknown and is estimated from the sample, nonparametric estimates of the probability content of the posterior bands are obtained.

1. The Dirichlet process and Bayesian estimation. Let $F(t)$ be a random distribution function on the real line. For $A(\infty) = a + 1 > 0$, let $A(t)$ be defined so that $A(t)/A(\infty)$ is a distribution function. $F(t)$ is a Dirichlet process with parameter $A(t)$ (see Ferguson (1973), for example) if for every $m = 1, 2, \cdots$ and $-\infty = t_0 < t_1 < \cdots < t_m < t_{m+1} = \infty$ the distribution of $(F(t_1), F(t_2), \cdots, F(t_m))$ is, in the notation of Wilks (1962, page 182), of the ordered Dirichlet type with parameter $(a_1, a_2, \cdots, a_m; a_{m+1})$ where for $j = 1, 2, \cdots, m$,

$$a_j = A(t_j) - A(t_{j-1}) . \tag{1.1}$$

Suppose that in a Bayesian estimation situation, a distribution function $F(t)$ has as a prior distribution the Dirichlet process with parameter $A(t)$. Let $F_n(t)$ denote the empirical distribution function corresponding to a random sample of size n from F. Ferguson (1973) has shown that the posterior distribution of F is a Dirichlet process with parameter $A(t) + nF_n(t)$.

To aid in the selection of a suitable prior parameter it is useful to quantify how the choice of a particular parameter affects the prior variability in F at various points in its domain simultaneously.

Similarly, when posterior estimates of F are to used in practice, it is desirable to quantify simultaneous posterior variability in F. In this paper, general probabilistic statements on such variability are derived. We commence by explicitly defining these statements.

2. Bayesian confidence bands. Let m be a fixed positive integer and for $i = 1, 2, \cdots, m$ define u_i and v_i so that $u_i < v_i$ for all i and

$$\begin{aligned} 0 = u_0 \leqq u_1 \leqq u_2 \leqq \cdots \leqq u_m < 1 \\ 0 < v_1 \leqq v_2 \leqq \cdots \leqq v_m \leqq v_{m+1} = 1 . \end{aligned} \tag{2.1}$$

Received January 1976; revised June 1977.

AMS 1970 *subject classifications.* Primary 62C10; Secondary 60G17, 62G15, 65C10.

Key words and phrases. Bayesian confidence bands, probability content, boundary crossing probabilities, nonparametric estimates, simulation.

Further, define

$$I(x) = 1 \quad x \geqq 0 \tag{2.2}$$
$$= 0 \quad x < 0$$

and $J(x)$ to be the function which equals one when $x > 0$ and zero otherwise. Let $t_1 < t_2 < \cdots < t_m$, and finally define

$$S(x) = \sum_{i=1}^{m} (u_i - u_{i-1}) I(x - t_i) \tag{2.3}$$

and

$$B(x) = v_1 + \sum_{i=1}^{m} (v_{i+1} - v_i) J(x - t_i) . \tag{2.4}$$

DEFINITION. Suppose the $F(t)$ has a Dirichlet process as its prior distribution. Let $S(t)$ and $B(t)$ be defined by (2.3) and (2.4) respectively. Then if

$$P\{S(t) \leqq F(t) \leqq B(t) \text{ for all } t\} = L \tag{2.5}$$

and P is a prior (posterior) probability, the functions $S(t)$ and $B(t)$ constitute the boundaries of a fixed region within which the random distribution function lies with prior (posterior) probability L. $S(t)$ and $B(t)$ are defined to be a pair of Bayesian confidence bands for the random distribution function, $F(t)$, with prior (posterior) probability content L.

It is to be emphasised that in the Bayesian context it is the distribution function which is random whilst the Bayesian confidence bands are fixed and known quantities. These bands are to be contrasted with the Neyman–Pearson type random confidence bands for a fixed, but unknown, distribution function which are discussed in Steck (1971).

There are many possible pairs of Bayesian confidence bands for F with probability content L. In practice, a particular pair of these bands must be chosen to express quantitative confidence in F. This situation has its parallel in the Neyman–Pearson theory (see Steck (1971)). In (2.5) the choice of the t_i is open and applied statisticians, when describing simultaneous posterior variability in F, may choose the t_i to correspond to the ith order statistic of the sample. In this case, it is to be borne in mind that repetition of observations may occur due to the discrete nature of the Dirichlet process. A detailed discussion of this case is given in Section 5.

Since $F(t)$ in (2.5) is a random distribution function it is clear that

$$P\{S(t) \leqq F(t) \leqq B(t) \text{ for all } t\} = P\{u_j \leqq F(t_j) \leqq v_j \text{ for all } j\} . \tag{2.6}$$

It then follows from Section 1 that to be able to calculate probabilities of the type (2.6), it suffices to be able to calculate general rectangle probabilities over the ordered Dirichlet distribution. A method for such calculation is presented in the following section.

3. Ordered Dirichlet rectangle probabilities. For m a positive integer, suppose that the joint distribution of Y_i, $i = 1, 2, \cdots, m$, is of the ordered Dirichlet

type with parameter $(a_1, a_2, \cdots, a_m; a_{m+1})$. Let $s_i = a_1 + a_2 + \cdots + a_i$ and for $i = 1, 2, \cdots, m$, let $G_i(x)$ denote the marginal distribution function of Y_i which, in the notation of Wilks (1962, page 174), is beta with parameter $(s_i, s_{m+1} - s_i)$. For integers i and j such that $1 \leqq i < j \leqq m$, let $G_{i,j}(x, y)$ denote the joint marginal distribution function of Y_i and Y_j which is bivariate ordered Dirichlet with parameter $(s_i, s_j - s_i; s_{m+1} - s_j)$. When $a_i = 1$ for all i, Y_j has the same distribution as the jth order statistic of a random sample of size m from the uniform distribution on the unit interval. Define the random step function $H_m(x)$ by

$$mH_m(x) = \sum_{j=1}^{m} I(x - Y_j)$$

where $I(x)$ is given by (2.2). When $a_i = 1$ for all i, $H_m(x)$ is just the empirical distribution function of the uniform random sample.

Let $U(x)$ and $V(x)$ be strictly increasing and continuous functions satisfying $U(0) > 0 > V(0)$ and $U(1) > 1 > V(1)$. The restrictions that $U(x)$ and $V(x)$ be continuous and strictly increasing may be relaxed (see Steck (1971) for example) but this is really tangential to our discussion. Denoting by $[y]$ the integral part of y, let $p = [mU(0)] + 1$ and $q = [mV(1)]$. Define

$$\begin{aligned} u_i &= 0\,, & i &= 1, 2, \cdots, p-1 \\ mU(u_i) &= i\,, & i &= p, p+1, \cdots, m \\ mV(v_j) &= j-1\,, & j &= 1, 2, \cdots, q+1 \\ v_j &= 1\,, & j &= q+2, q+3, \cdots, m \end{aligned}$$

and assume that $U(x)$ and $V(x)$ obey the additional restrictions that $u_i < v_i$ for all i. Let

$$1 - g = P\{U(x) > H_m(x) > V(x) \text{ for all } x \in [0, 1]\}\,. \tag{3.1}$$

Then since, for all x, $mH_m(x) \in \{0, 1, \cdots, m\}$ it is evident that $1 - g = P\{u_i < Y_i < v_i \text{ for all } i\}$. This is a probability of the type (2.6). In the special case when $a_i = 1$ for all i, the probability $1 - g$ has been calculated by Steck (1971), Durbin (1971) and others.

Define for $0 < x < y < 1$, $i \leqq j$ and $i, j \in \{0, 1, \cdots, m\}$ the following probabilities:

$$\begin{aligned} p_i(x) &= P\{mH_m(x) = i\} \\ q_{i,j}(x, y) &= P\{mH_m(x) = i,\, mH_m(y) = j\} \\ p_{i,j}(x, y) &= P\{mH_m(x) = i \mid mH_m(y) = j\} = q_{i,j}(x, y)/p_j(y)\,. \end{aligned} \tag{3.2}$$

Observe that

$$p_i(x) = P\{Y_i \leqq x,\, Y_{i+1} > x\} = G_i(x) - G_{i+1}(x) \tag{3.3}$$

whilst using a similar argument

$$q_{i,j}(x, y) = G_{i,j}(x, y) + G_{i+1,j+1}(x, y) - G_{i,j+1}(x, y) - \Delta_{i,j}(x, y) \tag{3.4}$$

where $\Delta_{i,j}(x, y)$ is the function which equals $G_j(x)$ when $i = j - 1$ and $G_{i+1,j}(x, y)$ otherwise. If $a_i = 1$ for all i, use of integration by parts in conjunction with (3.2) yields the simple expressions:

$$p_i(x) = b(i; m, x) \quad \text{and} \quad p_{i,j}(x, y) = b(i; j, x/y) \tag{3.5}$$

where $b(r; n, p)$ is the probability of r successes in n independent Bernoulli trials when the probability of success is p.

We are now in a position to obtain a set of recurrences for computing g from (3.1). The argument used is not a new one. Illustrations of its use are given in Feller (1948) and Dempster (1959). To obtain the desired recurrences for g, it is to be noted from (3.1) that g is the probability that $H_m(x)$ crosses either, or both, of $U(x)$ and $V(x)$. For $i = p, p + 1, \cdots, m$ let E_i denote the point $(u_i, i/m)$ and for $j = 0, 1, \cdots, q$ let F_j denote the point $(v_{j+1}, j/m)$. These $m + q + 2 - p$ points will be called critical points in the path of $H_m(x)$; for $H_m(x)$ crosses either, or both, of $U(x)$ and $V(x)$ if and only if it passes through at least one of these critical points.

Let e_i denote the probability that conditional on $H_m(x)$ passing through E_i, E_i is the first critical point it passes through. Similarly, let f_j denote the probability that conditional on $H_m(x)$ passing through F_j, F_j is the first critical point it passes through. Then taking probabilities,

$$g = \sum_{i=p}^{m} e_i p_i(u_i) + \sum_{j=0}^{q} f_j p_j(v_{j+1}) . \tag{3.6}$$

Since $p_i(x)$ is given by (3.3) it suffices to calculate recurrences for e_i and f_j. Conditional on $H_m(x)$ passing through E_k, the first critical point it passes through could be any one of E_i, $i = p, p + 1, \cdots, k$, or $F_j, j = 0, 1, \cdots, r(k)$, where $r(k) = \max\{j: v_{j+1} < u_k\}$. On taking probabilities one arrives at, for $k = p, p + 1, \cdots, m$:

$$1 = e_k + \sum_{i=p}^{k-1} e_i p_{i,k}(u_i, u_k) + \sum_{j=0}^{r(k)} f_j p_{j,k}(v_{j+1}, u_k) . \tag{3.7}$$

Similarly, by considering the path of $H_m(x)$ conditional on it passing through F_k, it follows that for $k = 0, 1, \cdots, q$:

$$1 = f_k + \sum_{i=p}^{k} e_i p_{i,k}(u_i, v_{k+1}) + \sum_{j=0}^{k-1} f_j p_{j,k}(v_{j+1}, v_{k+1}) . \tag{3.8}$$

The probability $p_{i,j}(x, y)$ is obtained from (3.2), (3.3) and (3.4). Then (3.6), (3.7) and (3.8) constitute a set of recurrences for calculating g. In the special case when $a_i = 1$ for all i, the probabilities in (3.5) are to be used in the recurrences. In this case, the recurrences are equivalent to those given by Durbin (1971), which he had obtained by considerations based on the Poisson process. The basic probabilistic simplicity of the technique employed in this section shows that, for the finite case at least, there is no basic need to involve the Poisson process.

4. Kolmogorov Bayesian confidence bands. The results of the previous section when used in conjunction with (2.6) enable the construction of any prior or

posterior confidence bands for the distribution function F. In this section a numerical example of the theory is considered.

Suppose that $A(t)$, the parameter of the Dirichlet process is continuous and chosen so as to make $a = A(\infty) - 1$ a positive integer. It is then possible to define t_j by (1.1) so as to make $a_j = 1$ for all $j = 1, 2, \cdots, a$. For such a_j and t_j, statistical tables are available which give the probability (2.6) for the Kolmogorov bands which are defined, for $e > 0$, by

$$u_i = 0, \qquad i = 1, 2, \cdots, a - h - 1 \tag{4.1}$$

$$a(u_i + e) = i, \qquad i = a - h, a - h + 1, \cdots, a$$

and $v_i = 1 - u_{a-i+1}$ for all i, where $h = [a(1 - e)]$. Such tables are given in Miller (1956). For a fixed probability content, α, the e defining these Kolmogorov bands is the solution to

$$P\{\sup_t |H_a(t) - t| \leqq e\} = \alpha \tag{4.2}$$

where $H_a(t)$ is the empirical distribution function of Section 3.

We shall now consider an application of this theory. Suppose that the prior parameter of the Dirichlet process is given by $A(t) = 10\Phi(t)$ where $\Phi(t)$ is the standard normal distribution function. Further suppose that a random sample of size five taken from this prior distribution resulted in the observations $\{-1.281, -0.524, 0.000, 0.253, 0.842\}$. It is desired to construct prior and posterior confidence bands of probability content 0.95 for the random distribution function.

The Kolmogorov prior bands are most easily calculated for this example. From the tables of Miller (1956) it is found that for $a = 9$, $e = 0.430$ is the solution to (4.2) in the case when $\alpha = 0.95$. A substitution of these values in (4.1) yields the required bands. These bands are displayed in Table 1. On inspecting the data, it is seen that jumps in the Kolmogorov prior bands occur at all the data points. It is thus possible to construct the Kolmogorov posterior bands for this example. Proceeding as before it is found that for $a = 14$, $e = 0.349$ is the solution to (4.2) in the case when $\alpha = 0.95$. A substitution of these

TABLE 1

95% prior bands for the Dirichlet process with random distribution function $F(t)$ having parameter $A(t) = 10\Phi(t)$

i	t_i	u_i	v_i
1	−1.281	0.000	0.430
2	−0.842	0.000	0.541
3	−0.524	0.000	0.652
4	−0.253	0.014	0.763
5	0.000	0.126	0.874
6	0.253	0.237	0.986
7	0.524	0.348	1.000
8	0.842	0.459	1.000
9	1.281	0.570	1.000

TABLE 2

95% posterior bands for the Dirichlet process with random distribution function $F(t)$ based on the sample $\{-1.281, -0.524, 0.000, 0.253, 0.842\}$ and prior parameter $A(t) = 10\Phi(t)$

i	t_i	u_i	v_i
1	−1.281—	0.000	0.349
2	−1.281	0.000	0.420
3	−0.842	0.000	0.492
4	−0.524—	0.000	0.563
5	−0.524	0.008	0.635
6	−0.253	0.080	0.706
7	0.000—	0.151	0.778
8	0.000	0.222	0.849
9	0.253—	0.294	0.920
10	0.253	0.365	0.992
11	0.524	0.437	1.000
12	0.842—	0.508	1.000
13	0.842	0.580	1.000
14	1.281	0.651	1.000

values in (4.1) gives the required posterior bands which are displayed in Table 2. An entry in column 2 of Table 2 of the form y- is to be regarded as a number imperceptibly smaller than y.

5. Posterior nonparametric estimates. When applied statisticians construct posterior bands, they may prefer to choose the t_i for the bands (2.3) and (2.4) to correspond to the ith order statistic of the sample. Further, they may not be prepared to stipulate a prior parameter for the Dirichlet process. In this case $A(t)$ must be estimated from the sample.

In a random sample of size n from the Dirichlet process with continuous parameter $A(t)$, suppose that there are m distinct observations and denote them by $x_1 < x_2 < \cdots < x_m$. Let n_i denote the number of members of the sample which assume the value x_i. Korwar and Hollander (1973) give an estimate of $A(t)$ as

$$A_1(t) = \sum_{i=1}^{m} I(t - x_i)/\log n$$

where $I(x)$ is given by (2.2). This estimate necessitates the structure of ties which the Dirichlet process forces. Further, the parameter $A(t)$ must be continuous. Then the joint posterior distribution of $F(x_i)$ for $i = 1, 2, \cdots, m-1$ is estimated to be ordered Dirichlet with parameter $(q_1, q_2, \cdots, q_{m-1}; q_m)$ where for $j = 1, 2, \cdots, m$

$$q_j = n_j + (1/\log n) \tag{5.1}$$

and $F(x_m)$ is estimated to be one. The theory of Sections 2 and 3 may then be applied to this estimated distribution. A major advantage of this technique is that regardless of what the prior parameter is, an estimate of the probability

content of the posterior bands may be obtained. This procedure is thus non-parametric, though only approximate.

There are two problems with the preceding approximations. The first is that $F(x_m)$ is always estimated to equal one; the second is that the recurrences (3.6), (3.7) and (3.8) are difficult to use in practice because the q_j, $j = 1, 2, \cdots, m$, are not all integers. These problems may both be overcome by taking the estimate of the prior parameter, $A_2(t)$, to equal $I(t - x_{m+1})$ if $c(m) = 0$ and to equal

$$I(t - x_{m+1}) + \sum_{k=1}^{c(m)} I(t - x_{j(k)})$$

otherwise, where x_{m+1} is chosen so that $x_{m+1} > x_m$, $c(j) = [j/\log n]$ for $j = 1, 2, \cdots, m$, and $j(k) = \min\{i: c(i) = k\}$ for $k = 1, 2, \cdots, c(m)$. To see how well $A_2(t)$ approximates $A_1(t)$ note that $[A_1(\infty)] = c(m) = A_2(\infty) - 1$ and when $c(m) > 0$, $[A_1(x_{j(k)})] = k = A_2(x_{j(k)})$ for $k = 1, 2, \cdots, c(m)$.

Taking $A_2(t)$ as the estimate of $A(t)$, the joint posterior distribution of $F(x_i)$, $i = 1, 2, \cdots, m$, is estimated to be ordered Dirichlet with parameter $(h_1, h_2, \cdots, h_m; 1)$ where $h_i = n_i + s_i$. If $c(m) = 0$, then $s_i = 0$ for all i whilst if $c(m) > 0$, s_i equals one for $i \in \{j(k): k = 1, 2, \cdots, c(m)\}$ and equals zero otherwise. Let $w_0 = 0$ whilst for $i = 1, 2, \cdots, m$ let $w_i = h_1 + \cdots + h_i$. Define $w_{m+1} = w_m + 1$. As all the h_i are integers, so too are all the w_i. Then from Section 3, $F(x_i)$ is equivalent to the w_ith order statistic from a uniform random sample of size w_m. It the follows from Steck (1971) that, under restrictions (2.1),

$$P\{u_i \leqq F(x_i) \leqq v_i \text{ for all } i\} \approx w_m! \det(k_{ij}) \tag{5.2}$$

where for $i, j = 1, 2, \cdots, w_m$, $(j - i + 1)!\, k_{ij}$ equals $(P_i - L_j)^{j-i+1}$ when $j - i + 1 > 0$ and $P_i > L_j$, whilst it equals zero otherwise. For $j = 0, 1, \cdots, m$, L_i equals u_j when $i = w_j, w_j + 1, \cdots, w_{j+1} - 1$. Similarly, for $j = 0, 1, \cdots, m - 1$, P_i equals v_{j+1} when $i = w_j + 1, w_j + 2, \cdots, w_{j+1}$. Applying (5.2) to (2.6) completes the posterior probability content estimate.

Practitioners may feel uneasy about using the above estimates because the weights of the posterior measures used exceed the sample size. A suitable readjustment would alleviate this source of concern. Some statisticians may also be reluctant to accept that observations which are used to estimate the prior parameter can be employed again to update it. This problem, as well as the previous one, may be overcome by dividing the data into two groups; the first group being used to estimate the prior parameter whilst members of the second group are treated as the observations.

Suppose that such a data allocation results in N observations belonging to the first group. Let $D_i = 1$ if x_i appears in the first group and zero otherwise. Let N_i denote the number of times x_i appears in the second group. Then the posterior parameter is estimated to equal

$$\sum_{i=1}^{m} (D_i + N_i \log N) I(t - x_i)/\log N$$

so that the joint posterior distribution of $F(x_i)$ for $i = 1, 2, \cdots, m - 1$ is estimated to be ordered Dirichlet with parameter $(Q_1, \cdots, Q_{m-1}; Q_m)$ where for

$j = 1, \cdots, m$, $Q_j = N_j + (D_j/\log N)$. This result is the split sample analogue to (5.1). Further analysis along the lines followed for the estimate (5.1) is easy to pursue.

The relative sizes of the groups, as well as the method chosen to allocate data to a group, will affect the eventual nonparametric estimate obtained. However, a study of the advantages and disadvantages of the various estimates proposed in this section is felt to lie beyond the scope of this paper. The purpose of the discussion has been to point out that the construction of such nonparametric estimates is feasible.

6. Simulation: an alternative. A requirement of the above theory is the ability to calculate general ordered Dirichlet rectangle probabilites. Recurrences in Section 3 were developed for this purpose. In situations where electronic computers are available it is perhaps simpler to approach the problem via simulation; that is, the generation of random samples from any ordered Dirichlet distribution.

Let Z_i, $i = 1, 2, \cdots, m + 1$ be independent gamma random variables with parameters a_i respectively. It follows from Wilks (1962, page 179) that the joint distribution of $Y_i = (Z_1 + \cdots + Z_i)/(Z_1 + \cdots + Z_{m+1})$ for $i = 1, 2, \cdots, m$ is ordered Dirichlet with parameter $(a_1, \cdots, a_m; a_{m+1})$. This transformation facilitates the required data generation. To estimate the probability, L, in (2.5) through its form (2.6), one simply computes the proportion, $\hat{L}$, of m-dimensional observations generated that satisfy the event on the right-hand side of (2.6). Let n denote the number of m-dimensional observations generated. Then if n is sufficiently large, $\hat{L}$ is approximately normally distributed with mean L and variance $\sigma^2 = L(1 - L)/n$. This variance may be estimated from the sample by $\hat{\sigma}^2 = \hat{L}(1 - \hat{L})/n$. In practice the simulated sample size, n, will be determined by the accuracy required of L. When this technique is employed in practice, it is advisable to obtain simulation estimates for bands of known probability content. These estimates may then be compared with the actual figures to ensure that the programme is operational. The Kolmogorov bands (4.1) are suitable for this check. In the case when $m = 5$, an actual data generation was carried out for $n = 100$ and the results of the simulation are shown in Table 3.

TABLE 3
Simulated Kolmogorov region probability content ($m = 5$, $n = 100$)

L	$\hat{L}$	σ	$\hat{\sigma}$
0.80	0.79	0.040	0.041
0.90	0.90	0.030	0.030
0.95	0.94	0.022	0.024

Acknowledgments. The author would like to thank an associate editor for suggesting the problem considered in Section 5.

REFERENCES

DEMPSTER, A. P. (1959). Generalized D_n^+ statistics. *Ann. Math. Statist.* **30** 593–597.

DURBIN, J. (1971). Boundary-crossing probabilities for the Brownian motion and Poisson processes and techniques for computing the power of the Kolmogorov–Smirnov test. *J. Appl. Probability* **8** 431–453.

FELLER, W. (1948). On the Kolmogorov–Smirnov limit theorems for empirical distributions. *Ann. Math. Statist.* **19** 177–189.

FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230.

KORWAR, R. M. and HOLLANDER, M. (1973). Contributions to the theory of Dirichlet processes. *Ann. Probability* **1** 705–711.

MILLER, L. H. (1956). Tables of percentage points of Kolmogorov statistics. *J. Amer. Statist. Assoc.* **51** 111–121.

STECK, G. P. (1971). Rectangle probabilities for uniform order statistics and the probability that the empirical distribution function lies between two distribution functions. *Ann. Math. Statist.* **42** 1–11.

WILKS, S. S. (1962). *Mathematical Statistics*, 2nd ed. Wiley, New York.

18, HAMMERDALE AVENUE
BALACLAVA 3183
VICTORIA, AUSTRALIA