The Annals of Statistics
1975, Vol. 3, No. 2, 504-511

UPPER AND LOWER POSTERIOR PROBABILITIES FOR DISCRETE DISTRIBUTIONS

BY ROBERT KLEYLE

University of Massachusetts

Dempster (1966) defines two sampling models which he labels structures of the first and second types. In this paper we consider the application of the type one model to the problem of finding upper and lower posterior probabilities when sampling from discrete univariate distributions whose support consists of the set of nonnegative integers. Specific results are obtained for the Poisson and geometric distributions.

1. Introduction. The basic components of the inference system proposed by Dempster (1966), (1967) consist of a population space U, an observation space X, and a class $\mathscr{M}$ of measurable mappings of U onto X. As is usually the case with type I structures, the population space can be taken to be the unit interval $U = (0, 1)$ (cf. Dempster (1966)) and in this application the observation space X consists of all nonnegative integers. The random sampling process is governed by the uniform probability measure μ over the Borel sets of U. The distributions over space X induced by the measure μ over U are assumed to be in one-to-one correspondence with the class $\mathscr{M}$.

Let Θ denote a space which is in one-to-one correspondence with $\mathscr{M}$, so that $\theta \in \Theta$ if and only if $m_\theta \in \mathscr{M}$. Θ is the parameter space indexing the class $\mathscr{M}$ and also the family of probability distributions induced on X by μ and $\mathscr{M}$. Now for each $\theta \in 0$ let $\{I(x, \theta), x = 0, 1, 2, \cdots\}$ denote a countable partition of U, where $I(x, \theta) = (a(x - 1, \theta), a(x, \theta)]$, and

$$0 = a(-1, \theta) < a(0, \theta) < \cdots < a(k, \theta) < \cdots \leqq 1 \qquad \text{for all} \quad \theta .$$

The functions $a(x, \theta)$ are picked so that, given $x \in X$,

$$m_\theta : u \to x \Leftrightarrow u \in I(x, \theta) \qquad \text{for all} \quad \theta ,$$

where m_θ denotes the mapping corresponding to θ. Thus, if $p(x, \theta)$ denotes the probability function of the distribution over X induced by μ and m_θ,

$$p(x, \theta) = \mu(I(x, \theta)) = a(x, \theta) - a(x - 1, \theta) \qquad \text{for all} \quad \theta \in \Theta ,$$

and

$$a(x, \theta) = F(x, \theta), \quad x = 0, 1, 2, \cdots ,$$

where $F(x, \theta)$ denotes the distribution function of the induced probability distribution.

Let us now assume that $a(x, \theta)$ is continuous and strictly monotonic in θ for

Received October 1972; revised April 1974.

AMS 1970 *subject classifications.* Primary 62; Secondary A99.

Key words and phrases. Population space, observation space, type one structures, consistent with the data, upper and lower preimages, upper and lower posterior probabilities.

all $x \in X$. Without loss of generality we can assume that $a(x, \theta)$ is strictly decreasing in θ since, if the opposite were the case, we could reparametrize in terms $\phi = 1/\theta$. Since $a(x, \theta)$ is strictly decreasing in θ for each $x \in X$, there exists an inverse function $b(x, \theta)$ which is itself continuous and strictly decreasing in θ for each $x \in X$ and has the property that for any $t \in (0, 1)$,

$$a(x, \theta) < t \Leftrightarrow \theta > b(x, t) , \tag{1.1}$$
$$a(x, \theta) > t \Leftrightarrow \theta < b(x, t) .$$

Now suppose that a random sample of size n is drawn. That is, a sample point $(u_1, \cdots, u_n)$ is drawn from U^n in accordance with probability measure μ^n, and the corresponding data point $(x_1, \cdots, x_n)$ is observed. Given the observed data $(x_1, \cdots, x_n)$ and an arbitrary sample point $(u_1, \cdots, u_n)$, we wish to determine the values of θ for which the sample point is *consistent with the data.* (For a definition of consistency see Dempster (1966).) Recall that for given θ,

$$m_\theta : u \to x \Leftrightarrow a(x - 1, \theta) < u < a(x, \theta) .$$

Referring to (1.1), we see that the above line can be rewritten

$$m_\theta : u \to x \Leftrightarrow b(x - 1, u) < \theta < b(x, u) ,$$

and for any given $(u_1, \cdots, u_n)$ and $(x_1, \cdots, x_n)$ the set of consistent θ is given by

$$\Gamma(u_1, \cdots, u_n) = \{\theta : \max_i b(x_i - 1, u_i) < \theta < \min_j b(x_j, u_j)\} .$$

It is possible that $\max_i b(x_i - 1, u_i) > \min_j b(x_j, u_j)$. In this case $\Gamma = \varnothing$. For notational convenience we define

$$Y = \max_i b(x_i - 1, u_i) , \qquad Z = \min_j b(x_j, u_j) . \tag{1.2}$$

Thus, for given $(x_1, \cdots, x_n)$ and $(u_1, \cdots, u_n)$,

$$\Gamma = \{\theta : Y < \theta < Z\} , \qquad \text{for} \quad Y < Z , \tag{1.3}$$
$$= \varnothing , \qquad \text{for} \quad Y > Z .$$

Furthermore, if T denotes the set of all sample points which are consistent with the data for some θ,

$$T = \{(u_1, \cdots, u_n) : 0 < Y < Z < \infty\} .$$

All upper and lower probabilities will be conditional on T.

Since any given sample point may be consistent with the data for more than one θ, the concept of consistency sets up a one-many mapping of U^n into Θ. Consequently, a set $S \subset \Theta$ will not have a unique preimage in U^n. However, upper and lower preimage can be defined (cf. Dempster (1967)). Given any set $S \subset \Theta$, let

$$S^* = \{(u_1, \cdots, u_n) : \Gamma(u_1, \cdots, u_n) \cap S \neq \varnothing\} , \tag{1.4}$$
$$S_* = \{(u_1, \cdots, u_n) : \varnothing \subset \Gamma(u_1, \cdots, u_n) \subseteqq S\} .$$

Let C denote the class of sets $S \subset \theta$ such that both S^* and S_* are Borel subsets

of U^n. Then for any set $S \in C$,

$$P^*(S) = P(S^*)/P(T)\,, \qquad P_*(S) = P(S_*)/P(T)\,,$$

where $P = \mu^n$.

A system of upper and lower posterior probabilities, rather than a unique posterior probability for each event on Θ, may seem cumbersome at first, but it offers greater flexibility in representing the *degree of certainty* in a particular situation. If the experimenter were in a state of total ignorance, the set of consistent θ would be Θ itself, resulting in an upper probability of one and a lower probability of zero. On the other extreme, perfect knowledge of the nature of a given situation implies that there can be only one consistent mapping for each sample point. This results in the $U^n \to \Theta$ mapping being one-to-one, and the system of upper and lower probabilities collapses into a single posterior probability. For a fuller discussion of these points the reader is referred to Dempster (1968).

2. Upper and lower posterior distributions. Let us assume that $\theta \subseteqq (0, \infty)$ and consider events of the types $\{\theta < \lambda\}$ and $\{\lambda_1 < \theta < \lambda_2\}$. It is straightforward to check that these events are members of class C. Now if $S = (\lambda_1, \lambda_2)$,

$$S^* = T - A(\lambda_2, \infty) - A(0, \lambda_1)\,, \qquad S_* = A(\lambda_1, \lambda_2)\,,$$

where

$$A(\lambda_1, \lambda_2) = \{(u_1, \cdots, u_n)\colon \lambda_1 < U < Z < \lambda_2\}\,. \tag{2.1}$$

The calculation of upper and lower posterior probabilities for both types of events is seen, therefore, to be essentially the problem of calculating $P(A(\lambda_1, \lambda_2))$ since $T = A(0, \infty)$, and

$$\{\theta < \lambda\} = \lim_{\lambda_1 \to 0} \{\lambda_1 < \theta < \lambda\}\,.$$

However, as will be shown later in this section, it is virtually impossible to calculate this probability directly when $0 < \lambda_1 < \lambda_2 < \infty$.

Let us first consider one-sided events. Define

$$A_1(\lambda) = \{(u_1, \cdots, u_n)\colon 0 < Y < Z < \lambda\}\,, \tag{2.2}$$
$$A_2(\lambda) = \{(u_1, \cdots, u_n)\colon 0 < Y < \lambda < Z\}\,,$$

and note that if $S = (0, \lambda)$, (1.3), (1.4), and (1.5) imply that

$$S^* = A_1(\lambda) \cup A_2(\lambda)\,, \qquad S_* = A_1(\lambda)\,.$$

Thus,

$$P^*(\theta < \lambda) = [P(A_1(\lambda)) + P(A_2(\lambda))]/P(T)\,, \qquad P_*(\theta < \lambda) = P(A_1(\lambda))/P(T)\,. \tag{2.3}$$

Now suppose that $S = (\lambda_1, \lambda_2)$. In this instance,

$$S^* = A_1(\lambda_2) \cup A_2(\lambda_2) - A_1(\lambda_1)\,,$$

and from (2.3),

$$P^*(\lambda_1 < \theta < \lambda_2) = P^*(\theta < \lambda_2) - P_*(\theta < \lambda_1)\,. \tag{2.4}$$

We find the lower probability of $S(\lambda_1, \lambda_2)$ by finding the upper probability of its complement $\bar{S}$ since it is easily demonstrated that

$$P_*(S) = 1 - P^*(\bar{S}) \qquad \text{for all} \quad S \in C . \tag{2.5}$$

Definition (1.4) implies that

$$\bar{S}^* = A_1(\lambda_1) \cup A_2(\lambda_1) \cup A_3(\lambda_2) \cup [A_2(\lambda_2) - A_4(\lambda_1, \lambda_2)] ,$$

where

$$\begin{aligned} A_3(\lambda) &= \{(u_1, \cdots, u_n) : \lambda < Y < Z\} , \\ A_4(\lambda_1, \lambda_2) &= \{(u_1, \cdots, u_n) : 0 < Y < \lambda_1 < \lambda_2 < Z\} . \end{aligned} \tag{2.6}$$

Note that the sets in the above union are mutually disjoint and that $A_4(\lambda_1, \lambda_2) \subset A_2(\lambda_2)$. Thus,

$$P(\bar{S}^*) = P(A_1(\lambda_1)) + P(A_2(\lambda_1)) + P(A_2(\lambda_2)) + P(A_3(\lambda_2)) - P(A_4(\lambda_1, \lambda_2)) . \tag{2.7}$$

The next step is to determine the P measure of sets $A_i(\lambda)$, $i = 1, 2, 3$, $A_4(\lambda_1, \lambda_2)$, and T. Recall from (1.2) and (2.2) that

$$A_2(\lambda) = \{(u_1, \cdots, u_n) : b(x_i - 1, u_i) < \lambda < b(x_i, u_i) \quad \text{for all} \quad i\} .$$

It therefore follows from (1.1) that

$$A_2(\lambda) = \{(u_1, \cdots, u_n) : a(x_i - 1, \lambda) < u_i < a(x_i, \lambda) \quad \text{for all} \quad i\} ,$$

so that

$$P(A_2(\lambda)) = \prod_{i=1}^{n} [a(x_i, \lambda) - a(x_i - 1, \lambda)] = \prod_{i=1}^{n} p(x_i, \lambda) , \tag{2.8}$$

where $p(x, \lambda)$ denotes the probability function of the distribution over the observation space.

By a similar argument, we derive

$$\begin{aligned} P(A_4(\lambda_1, \lambda_2)) &= \prod_{i=1}^{n} [a(x_i, \lambda_2) - a(x_i - 1, \lambda_1)] , \\ &\qquad \text{for} \quad a(x_i, \lambda_2) < a(x_i - 1, \lambda_1) \qquad \text{for all } i \\ &= 0 , \qquad \text{otherwise.} \end{aligned} \tag{2.9}$$

Furthermore, since

$$A_3(\lambda) = T - A_1(\lambda) - A_2(\lambda) ,$$

and

$$A_1(\lambda) \cap A_2(\lambda) = \varnothing ,$$

$$P(A_3(\lambda)) = P(T) - P(A_1(\lambda)) - P(A_2(\lambda)) . \tag{2.10}$$

From (2.10), we may rewrite (2.7) as

$$P(\bar{S}^*) = P(T) + P(A_1(\lambda_1)) + P(A_2(\lambda_1)) - P(A_1(\lambda_2)) - P(A_4(\lambda_1, \lambda_2)) , \tag{2.11}$$

and using (2.3), (2.5) and (2.11), we have

$$P_*(\lambda_1 < \theta < \lambda_2) = P_*(\theta < \lambda_2) - P^*(\theta < \lambda_1) + P(A_4(\lambda_1, \lambda_2))/P(T) . \tag{2.12}$$

It remains only to find $P(A_1(\lambda))$, but this task is much more difficult than that

of finding the probabilities of $A_2(\lambda)$ and $A_4(\lambda_1, \lambda_2)$. The reason for this difficulty is that unlike $A_2(\lambda)$ and $A_4(\lambda_1, \lambda_2)$, there is no λ separating Y and Z in the definition of $A_1(\lambda)$.

The first step in finding the P-measure of $A_1(\lambda)$ is to derive the joint density of (Y, Z). We see from (1.1) and (1.2) that

$$
\begin{aligned}
&P(Y \leqq y, Z > z) \\
&\quad = P(b(x_i - 1, u_i) \leqq y \text{ for all } i,\ b(x_j, u_j) > z \text{ for all } j) \\
&\quad = P(a(x_i - 1, y) \leqq u_i < a(x_i, z) \text{ for all } i) \qquad (2.13) \\
&\quad \begin{cases} = \prod_{i=1}^{n} [a(x_i, z) - a(x_i - 1, y)], & \text{for } (y, z) \in R \\ = 0, & \text{otherwise,} \end{cases}
\end{aligned}
$$

where

$$R = \{(y, z) : a(x_i - 1, y) < a(x_i, z) \text{ for all } i = 1, \cdots, n\}.$$

The joint distribution function of (Y, Z) is

$$
\begin{aligned}
G(y, z) &= P(Y \leqq y, Z \leqq z) \\
&= \prod_{i=1}^{n} [1 - a(x_i - 1, y)] - \prod_{i=1}^{n} [a(x_i, z) - a(x_i - 1, y)], && (y, z) \in R \\
&= \prod_{i=1}^{n} [1 - a(x_i - 1, y)], && (y, z) \notin R,
\end{aligned}
$$

and the joint density is

$$
\begin{aligned}
g(y, z) &= -\partial^2/\partial y\, \partial z \left[\prod_{i=1}^{n} [a(x_i, z) - a(x_i - 1, y)]\right], && (y, z) \in R \\
&= 0, && (y, z) \notin R.
\end{aligned}
$$

Now

$$P(A_1(\lambda)) = \int_0^{\lambda} \int_0^{z} g(y, z)\, dy\, dz = \int_0^{\lambda} \left.\frac{\partial G}{\partial z}\right|_{y=z} dz,$$

and using the product rule along with (2.13), we obtain

$$
\begin{aligned}
\partial G(y, z)/\partial z &= -\sum_{i=1}^{n} a'(x_i, z) \prod_{1, j \neq i}^{n} [a(x_j, z) - a(x_j - 1, z)], && (y, z) \in R \\
&= 0, && (y, z) \notin R,
\end{aligned}
$$

where $a'(x, z) = \partial a(x, z)/\partial z$. Thus, since $a(x_j - 1, z) < a(x_j, z)$ for all j,

$$P(A_1(\lambda)) = -\int_0^{\lambda} \sum_{i=1}^{n} \prod_{1, j \neq i}^{n} p(x_j, z) a'(x_i, z)\, dz, \qquad (2.14)$$

where $p(x, z)$ denotes the discrete probability function defined in Section 1.

To obtain a more explicit expression for $P(A_1(\lambda))$, one must know the exact form of $a(x, z)$. In the next section we consider examples in which $a(x, z)$ denotes the distribution functions of the Poisson and geometric distributions.

Before proceeding to the next section, we note that if the above approach were applied to the problem of finding the P-measure of $A(\lambda_1, \lambda_2)$ given by (2.1),

$$P(A(\lambda_1, \lambda_2)) = \int_{\lambda_2}^{\lambda_1} \left.\frac{\partial G}{\partial z}\right|_{y=z} dz - \int_{\lambda_2}^{\lambda_1} \left.\frac{\partial G}{\partial z}\right|_{y=\lambda_1} dz.$$

The second term on the rhs of the above line is extremely difficult to calculate, even if the exact form of $a(x, z)$ is known.

3. Examples.

The Poisson distribution. Consider the situation when the probability distribution on the observation space is the familiar Poisson. In this case,

$$a(x, \theta) = e^{-\theta} \sum_{j=0}^{x} \theta^j/j! = (1/x!) \int_{\theta}^{\infty} u^x e^{-u}\, du\,, \tag{3.1}$$
$$a'(x, \theta) = -e^{-\theta}\theta^x/x!\,.$$

It follows from (2.14) that

$$P(A_1(\lambda)) = (n/\prod_1^n x_j!) \int_0^{\lambda} z^t e^{-nz}\, dz \tag{3.2}$$
$$= k_n(x_1, \cdots, x_n)\gamma_{n\lambda}(t + 1)\,,$$

where $t = \sum_1^n x_j$, $k_n(x_1, \cdots, x_n) = t!/n^t \prod_{j=1}^n x_j!$, and $\gamma_a(b)$ denotes the incomplete gamma function

$$\gamma_a(b) = (1/\Gamma(b)) \int_0^a u^{b-1}e^{-u}\, du\,.$$

Recalling that $T = A_1(\infty)$, we see from (3.2) that

$$P(T) = k_n(x_1, \cdots, x_n)\,.$$

Furthermore, applying (2.8) we see that

$$P(A_2(\lambda)) = \prod_{j=1}^n e^{-\lambda}\lambda^{x_j}/x_j! = k_n e^{-n\lambda}(n\lambda)^t/t!\,,$$
$$P(A_1(\lambda)) + P(A_2(\lambda)) = k_n[\gamma_{n\lambda}(t + 1) + e^{-n\lambda}(n\lambda)^t/t!] = k_n\gamma_{n\lambda}(t)\,,$$

and from (2.3),

$$P^*(\theta < \lambda) = \gamma_{n\lambda}(t)\,, \qquad P_*(\theta < \lambda) = \gamma_{n\lambda}(t + 1)\,.$$

Finally, (2.4) and (2.14) imply that

$$P^*(\lambda_1 < \theta < \lambda_2) = \gamma_{n\lambda_2}(t) - \gamma_{n\lambda_1}(t + 1)\,,$$

while

$$P_*(\lambda_1 < \theta < \lambda_2) = \gamma_{n\lambda_2}(t + 1) - \gamma_{n\lambda_1}(t) + \phi(\lambda_1, \lambda_2 : x_1, \cdots, x_n)\,,$$

where

$$\phi(\lambda_1, \lambda_2 : x_1, \cdots, x_n) = \prod_{j=1}^n [\gamma_{\lambda_1}(x_j) - \gamma_{\lambda_2}(x_j + 1)]/k_n(x_1, \cdots, x_n)\,,$$
$$\text{for} \quad \gamma_{\lambda_1}(x_j) < \gamma_{\lambda_2}(x_j + 1) \quad \text{for all} \quad j = 1, \cdots, n\,,$$
$$= 0\,, \qquad \text{otherwise.}$$

From the above results, it is obvious that

$$\lim_{n\to\infty} [P^*(\lambda_1 < \theta < \lambda_2) - P_*(\lambda_1 < \theta < \lambda_2)] = 0 \qquad \text{for all} \quad (\lambda_1, \lambda_2)\,.$$

In fact, it is apparent from lines (2.4) through (2.9) that this result holds even when $a(x, \theta)$ is not a Poisson distribution function.

It is interesting to note that aside from $\phi(\lambda_1, \lambda_2 : x_1, \cdots, x_n)$, which is zero for many data points $(x_1, \cdots, x_n)$ and λ values and is asymptotically zero for all λ values, all posterior probabilities depend on the data only through the statistic $t = \sum_1^n x_i$ which is sufficient for θ.

The geometric distribution. For the geometric distribution,

$$p(x, \theta) = (1 - \theta)\theta^x, \quad x = 0, 1, 2, \cdots, \qquad 0 < \theta < 1,$$

and

$$a(x, \theta) = 1 - \theta^{x+1}.$$

Thus, for $\lambda \in (0, 1)$, (2.14) implies that

$$\begin{aligned} P(A_1(\lambda)) &= \textstyle\sum_{i=1}^{n} (x_i + 1) \int_0^\lambda z^t(1 - z)^{n-1}\, dz \\ &= I_\lambda(t + 1, n)/\binom{t+n-1}{t}, \end{aligned}$$

where $t = \sum_1^n x_j$, and I_λ denotes the incomplete beta function

$$I_\lambda(a, b) = (1/B(a, b)) \int_0^\lambda u^{a-1}(1 - u)^{b-1}\, du.$$

Furthermore,

$$P(T) = P(A_1(1)) = 1/\binom{t+n-1}{t},$$

and from (2.3),

$$\begin{aligned} P^*(\theta < \lambda) &= I_\lambda(t + 1, n) + \binom{t+n-1}{t}\lambda^t(1 - \lambda)^n \\ &= \textstyle\sum_{j=t}^{t+n} \binom{t+n}{j}\lambda^j(1 - \lambda)^{t+n-j} - t(t + n)^{-1}\binom{t+n}{t}\lambda^t(1 - \lambda)^n, \end{aligned}$$

while

$$P_*(\theta < \lambda) = I_\lambda(t + 1, n) = \textstyle\sum_{j=t+1}^{t+n} \binom{t+n}{j}\lambda^j(1 - \lambda)^{t+n-j}.$$

The second equality in both of the above equations follows from the well-known fact that

$$I_\lambda(k, n - k + 1) = \textstyle\sum_{j=k}^{n} \binom{n}{j}\lambda^j(1 - \lambda)^{n-j}.$$

Applying the above results to (2.4) and (2.12), we get

$$P^*(\lambda_1 < \theta < \lambda_2) = I_{\lambda_2}(t + 1, n) - I_{\lambda_1}(t + 1, n) + \binom{t+n-1}{t}\lambda_2{}^t(1 - \lambda_2)^n,$$

and

$$\begin{aligned} P_*(\lambda_1 < \theta < \lambda_2) = I_{\lambda_2}(t + 1, n) - I_{\lambda_1}(t + 1, n) - \binom{t+n-1}{t}\lambda_1{}^t(1 - \lambda_1)^n \\ + \psi(\lambda_1, \lambda_2 : x_1, \cdots, x_n), \end{aligned}$$

where

$$\begin{aligned} \psi(\lambda_1, \lambda_2 : x_1, \cdots, x_n) &= \textstyle\prod_{j=1}^{n} (\lambda_1{}^{x_j} - \lambda_2{}^{x_j+1}), \quad \text{for } \lambda_2{}^{x_j+1} < \lambda_1{}^{x_j} \text{ for all } j, \\ &= 0, \quad \text{otherwise.} \end{aligned}$$

As in the Poisson example, the difference between the upper and lower probabilities is asymptotically zero, and these probabilities depend almost entirely on $t = \sum_1^n x_j$ which is again sufficient.

Acknowledgment. The author wishes to thank the referee for suggesting an approach which greatly simplified the derivation of $P(A_1(\lambda))$.

REFERENCES

[1] DEMPSTER, A. P. (1966). New approaches for reasoning toward posterior distributions based on sample data. *Ann. Math. Statist.* **37** 355–374.

[2] DEMPSTER, A. P. (1967). Upper and lower probabilities induced by a multivalued mapping. *Ann. Math. Statist.* **38** 325–339.

[3] DEMPSTER, A. P. (1968). A generalization of Bayesian inference (with discussion). *J. Roy. Statist. Soc. Ser. B* **30** 205–247.

DEPT. OF MATHEMATICAL SCIENCES
INDIANA UNIVERSITY/PURDUE UNIVERSITY
1201 EAST 38TH STREET
INDIANAPOLIS, INDIANA 46205