A Generalization of Bayesian Inference

By A. P. DEMPSTER

Harvard University and London School of Economics

[Read at a RESEARCH METHODS MEETING of the Society, February 14th, 1968, Professor R. L. PLACKETT in the Chair]

SUMMARY

Procedures of statistical inference are described which generalize Bayesian inference in specific ways. Probability is used in such a way that in general only bounds may be placed on the probabilities of given events, and probability systems of this kind are suggested both for sample information and for prior information. These systems are then combined using a specified rule. Illustrations are given for inferences about trinomial probabilities, and for inferences about a monotone sequence of binomial p_i. Finally, some comments are made on the general class of models which produce upper and lower probabilities, and on the specific models which underlie the suggested inference procedures.

1. INTRODUCTION

REDUCED to its mathematical essentials, Bayesian inference means starting with a global probability distribution for all relevant variables, observing the values of some of these variables, and quoting the conditional distribution of the remaining variables given the observations. In the generalization of this paper, something less than a global probability distribution is required, while the basic device of conditioning on observed data is retained. Actually, the generalization is more specific. The term *Bayesian* commonly implies a global probability law given in two parts, first the marginal distribution of a set of parameters, and second a family of conditional distributions of a set of observable variables given potential sets of parameter values. The first part, or *prior distribution*, summarizes a set of beliefs or state of knowledge in hand before any observations are taken. The second part, or *likelihood function*, characterizes the information carried by the observations. Specific generalizations are suggested in this paper for both parts of the common Bayesian model, and also for the method of combining the two parts. The components of these generalizations are built up gradually in Section 2 where they are illustrated on a model for trinomial sampling.

Inferences will be expressed as *probabilities* of events defined by unknown values, usually unknown parameter values, but sometimes the values of observables not yet observed. It is not possible here to go far into the much-embroiled questions of whether probabilities are or are not objective, are or are not degrees of belief, are or are not frequencies, and so on. But a few remarks may help to set the stage. I feel that the proponents of different specific views of probability generally share more attitudes rooted in the common sense of the subject than they outwardly profess, and that careful analysis renders many of the basic ideas more complementary than contradictory. Definitions in terms of frequencies or equally likely cases do illustrate clearly how reasonably objective probabilities arise in practice, but they fail in

themselves to say what probabilities mean or to explain the pervasiveness of the concept of probability in human affairs. Another class of definitions stresses concepts like degree of confidence or degree of belief or degree of knowledge, sometimes in relation to betting rules and sometimes not. These convey the flavour and motivation of the science of probability, but they tend to hide the realities which make it both possible and important for cognizant people to agree when assigning probabilities to uncertain outcomes. The possibility of agreement arises basically from common perceptions of symmetries, such as symmetries among cases counted to provide frequencies, or symmetries which underlie assumptions of exchangeability or of equally likely cases. The importance of agreement may be illustrated by the statistician who expresses his inferences about an unknown parameter value in terms of a set of betting odds. If this statistician accepts any bet proposed at his stated odds, and if he wagers with colleagues who consistently have more information, perhaps in the form of larger samples, then he is sure to suffer disaster in the long run. The moral is that probabilities can scarcely be "fair" for business deals unless both parties have approximately the same probability assessments, presumably based on similar knowledge or information. Likewise, probability inferences can contribute little to public science unless they are as objective as the web of generally accepted fact on which they are based. While knowledge may certainly be personal, the communication of knowledge is one of the most fundamental of human endeavours. Statistical inference can be viewed as the science whose formulations make it possible to communicate partial knowledge in the form of probabilities.

Generalized Bayesian inference seeks to permit improvement on classical Bayesian inference through a complex trade-off of advantages and disadvantages. On the credit side, the requirement of a global probability law is dropped and it becomes possible to work with only those probability assumptions which are based on readily apparent symmetry conditions and are therefore reasonably objective. For example, in a wide class of sampling models, including the trinomial sampling model analysed in Section 2, no probabilities are assumed except the familiar and non-controversial representation of a sample as n independent and identically distributed random elements from a population. Beyond this, further assumptions like specific parametric forms or prior distributions for parameters need be put in only to the extent that they appear to command a fair degree of assent.

The new inference procedures do not in general yield exact probabilities for desired inferences, but only bounds for such probabilities. While it may count as a debit item that inferences are less precise than one might have hoped, it is a credit item that greater flexibility is allowed in the representation of a state of knowledge. For example, a state of total ignorance about an uncertain event T is naturally represented by an upper probability $P^*(T) = 1$ and a lower probability $P_*(T) = 0$. The new flexibility thus permits a simple resolution of the old controversy about how to represent total ignorance via a probability distribution. In real life, ignorance is rarely so total that (0, 1) bounds are justified, but ignorance is likely to be such that a precise numerical probability is difficult to justify. I believe that experience and familiarity will show that the general range of bounds $0 \leqslant P_*(T) \leqslant P^*(T) \leqslant 1$ provides a useful tool for representing degrees of knowledge.

Upper and lower probabilities apparently originated with Boole (1854) and have reappeared after a largely dormant period in Good (1962) and Smith (1961, 1965). In this paper upper and lower probabilities are generated by a specific mathematical device whereby a well-defined probability measure over one sample space becomes

diffused in its application to directly interesting events. In order to illustrate the idea simply, consider a map showing regions of land and water. Suppose that 0·80 of the area of the map is visible and that the visible area divides in the proportions 0·30 to 0·70 of water area to land area. What is the probability that a point drawn at random from the *whole* map falls in a region of water? Since the visible water area is 0·24 of the total area of the map, while the unobserved 0·20 of the total area could be water or land, it can be asserted only that the desired probability lies between 0·24 and 0·44. The model supposes a well-defined uniform distribution over the whole map. Of the total measure of unity, the fraction 0·24 is associated with water, the fraction 0·56 is associated with land, and the remaining fraction 0·20 is ambiguously associated with water or land. Note the implication of total ignorance of the unobserved area. There would be no objection to introducing other sources of information about the unobserved area. Indeed, if such information were appropriately expressed in terms of an upper and lower probability model, it could be combined with the above information using a rule of combination defined within the mathematical system. A correct analogy can be drawn with prior knowledge of parameter values, which can likewise be formally incorporated into inferences based on sample data, using the same rule of combination. The general mathematical system, as given originally in Dempster (1967a), will be unfolded in Section 2 and will be further commented upon in Section 4.

If the inference procedures suggested in this paper are somewhat speculative in nature, the reason lies, I believe, not in a lack of objectivity in the probability assumptions, nor in the upper and lower probability feature. Rather, the source of the speculative quality is to be found in the logical relationships between population members and their observable characteristics which are postulated in each model set up to represent sampling from a population. These logical relationships are conceptual devices, which are not regarded as empirically checkable even in principle, and they are somewhat arbitrary. Their acceptability will be analysed in Section 5 where it will be argued that the arbitrariness may correspond to something real in the nature of an uncertainty principle.

A degree of arbitrariness does not in itself rule out a method of statistical inference. For example, confidence statements are widely used in practice despite the fact that many confidence procedures are often available within the same model and for the same question, and there is no well-established theory for automatic choice among available confidence procedures. In part, therefore, the usefulness of generalized Bayesian inference procedures will require that practitioners experiment with them and come to feel comfortable with them. Relatively few procedures are as yet analytically tractable, but two examples are included, namely, the trinomial sampling inference procedures of Section 2, and a procedure for distinguishing between monotone upward and monotone downward sequences of binomial p_i as given in Section 3. Another model is worked through in detail in Dempster (1967b).

Finally, an acknowledgement is due to R. A. Fisher who announced with characteristic intellectual boldness, nearly four decades ago, that probability inferences were indeed possible outside of the Bayesian formulation. Fisher compiled a list of examples and guide-lines which seemed to him to lead to acceptable inferences in terms of probabilities which he called *fiducial probabilities*. The mathematical formulation of this paper is broad enough to include the fiducial argument in addition to standard Bayesian methods. But the specific models which Fisher advocated, depending on ingenious but often controversial *pivotal quantities*, are replaced here by models which start further back at the concept of a population explicitly represented

by a mathematical space. Fisher did not consider models which lead to separated upper and lower probabilities, and indeed went to some lengths, using sufficiency and ancillarity, and arranging that the spaces of pivotal quantities and of parameters be of the same dimension, in order to ensure that ambiguity did not appear. This paper is largely an exploration of fiducial-like arguments in a more relaxed mathematical framework. But, since Bayesian methods are more in the main stream of development, and since I do explicitly provide for the incorporation of prior information, I now prefer to describe my methods as extensions of Bayesian methods rather than alternative fiducial methods. I believe that Fisher too regarded fiducial inference as being very close to Bayesian inference in spirit, differing primarily in that fiducial inference did not make use of prior information.

2. Upper and Lower Probability Inferences Illustrated on a Model for Trinomial Sampling

A pair of sample spaces X and S underlie the general form of mathematical model appearing throughout this work. The first space X carries an ordinary probability measure μ, but interest centres on events which are identified with subsets of S. A bridge is provided from X to S by a logical relationship which asserts that, if x is the realized sample point in X, then the realized sample point s in S must belong to a subset Γx of S. Thus a basic component of the model is a mathematical transformation which associates a subset Γx of S with each point x of X. Since the Γx determined by a specific x contains in general many points (or *branches* or *values*), the transformation $x \to \Gamma x$ may be called a *multivalued mapping*. Apart from measurability considerations, which are ignored in this paper, the general model is defined by the elements introduced above and will be labelled (X, S, μ, Γ) for convenient reference. Given (X, S, μ, Γ), upper and lower probabilities $P^*(T)$ and $P_*(T)$ are determined for each subset T of S.

In the cartographical example of Section 1, X is defined by the points of the map, S is defined by two points labelled "water" and "land", μ is the uniform distribution of probability over the map, and Γ is the mapping which associates the single point "water" or "land" in S with the appropriate points of the visible part of X and associates both points of S with the points of the unseen part of X. For set-theoretic consistency, Γx should be regarded as a single point subset of S, rather than a single point itself, over the visible part of X, but the meaning is the same either way.

The general definitions of $P^*(T)$ and $P_*(T)$ as given in Dempster (1967a) are repeated below in more verbal form. For any subset T of S, define T^* to be the set of points x in X for which Γx has a non-empty intersection with T, and define T_* to be the set of points x in X for which Γx is contained in T but is not empty. In particular, the sets S^* and S_* coincide. The complement $X - S^*$ of S^* consists of those x for which Γx is the empty set. Now define the *upper probability* of T to be

$$P^*(T) = \mu(T^*)/\mu(S^*) \tag{1}$$

and the *lower probability* of T to be

$$P_*(T) = \mu(T_*)/\mu(S^*). \tag{2}$$

Note that, since $T_* \subset T^* \subset S^*$, one has

$$0 \leqslant P_*(T) \leqslant P^*(T) \leqslant 1. \tag{3}$$

Also, if $\bar{T}$ is the complement of T in S, then $\bar{T}_*$ and $\bar{T}^*$ are respectively the complements of T^* and T_* in S^*, so that

$$P_*(\bar{T}) = 1 - P^*(T) \quad \text{and} \quad P^*(\bar{T}) = 1 - P_*(T). \tag{4}$$

Other formal consequences of the above definitions are explored in Dempster (1967a).

The heuristic conception which motivates (1) and (2) is the idea of carrying probability elements $d\mu$ from X to S along the branches of the mapping Γx. The ambiguity in the consequent probability measure over S occurs because the probability element $d\mu(x)$ associated with x in X may be carried along any branch of Γx or, more generally, may be distributed over the different branches of Γx for each x. Part of the μ measure, namely the measure of the set $X - S^*$ consisting of points x such that Γx is empty, cannot be moved from X at all. Since there is an implicit assumption that some s in S is actually realized, it is appropriate to condition by S^* when defining relevant probabilities. This explains the divisor $\mu(S^*)$ appearing in (1) and (2). Among all the ways of transferring the relevant probability $\mu(S^*)$ from X to S along branches of Γx, the largest fraction which can possibly follow branches into T is $P^*(T)$, while the smallest possible fraction is $P_*(T)$. Thus conservative probability judgements may be rendered by asserting only that the probability of T lies between the indicated upper and lower bounds.

It may also be illuminating to view Γx as a random *set* in S generated by the random *point* x in X, subject to the condition that Γx is not empty. After conditioning on S^*, $P^*(T)$ is the probability that the random set Γx intersects the fixed set T, while $P_*(T)$ is the probability that the random set Γx is contained in the fixed set T.

A probability model like (X, S, μ, Γ) may be modified into other probability models of the same general type by conditioning on subsets of S. Such conditioning on observed data defines the generalized Bayesian inferences of this paper. Beyond and generalizing the concept of conditioning, there is a natural rule for combining or multiplying several independent models of the type (X, S, μ, Γ) to obtain a single product model of the same type. For example, the models for n independent sample observations may be put together by the product rule to yield a single model for a sample of size n, and the model defining prior information may be combined with the model carrying sample information by the same rule. The rules for conditioning and multiplying will be transcribed below from Dempster (1967a) and will be illustrated on a model for trinomial sampling. First, however, the elements of the trinomial sampling model will be introduced for a sample of size one.

Each member of a large population, shortly to be idealized as an infinite population, is supposed known to belong to one of three identifiable categories c_1, c_2 and c_3, where the integer subscripts do not indicate a natural ordering of the categories. Thus the individuals of the population could be balls in an urn, identical in appearance apart from their colours which are red (c_1) or white (c_2) or blue (c_3). A model will be defined which will ultimately lead to procedures for drawing inferences about unknown population proportions of c_1, c_2 and c_3, given the categories of a random sample of size n from the population. Following Dempster (1966), the individuals of the population will be explicitly represented by the points of a space U, and the randomness associated with a sample individual drawn from U will be characterized by a probability measure over U. Thus, a finite population of size N could be represented by any finite space U with N elements, with random sampling represented by the uniform distribution of probability over the N elements of U. Such a finite population

model is analysed in detail in Dempster (1967b). Here, however, the population is treated as infinite, and, for reasons tied up with the trinomial observable, the space U is identified with a triangle. Convenient barycentric coordinates for a general point of U are

$$\mathbf{u} = (u_1, u_2, u_3), \tag{5}$$

where $0 \leqslant u_1$, $0 \leqslant u_2$, $0 \leqslant u_3$ and $u_1 + u_2 + u_3 = 1$. See Fig. 1. It is further supposed that a random sample of size one means an individual $\mathbf{u}$ drawn according to the uniform

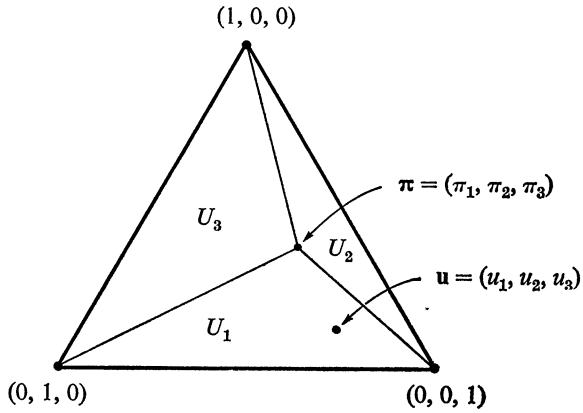

FIG. 1. A triangle representing the space U, showing the barycentric coordinates of the three vertices of U together with a general point $\mathbf{u} = (u_1, u_2, u_3)$. The three closed sub-triangles labelled U_1, U_2 and U_3 with a common vertex at $\boldsymbol{\pi}$ represent the subsets of U consisting of points $\mathbf{u}$ such that $B\mathbf{u}$ contains $(c_1, \boldsymbol{\pi})$, $(c_2, \boldsymbol{\pi})$ and $(c_3, \boldsymbol{\pi})$, respectively.

distribution ρ over the triangle U. In the model (X, S, μ, Γ) representing a random sample of size one from a trinomial population the roles of X and μ will be played by U and ρ.

Two further spaces enter naturally into the model for a single trinomial observation. The first is the three-element space $C = \{c_1, c_2, c_3\}$ whose general member c represents the observable category of the sample individual. The second is the space Π whose general point is

$$\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3), \tag{6}$$

with $0 \leqslant \pi_1$, $0 \leqslant \pi_2$, $0 \leqslant \pi_3$ and $\pi_1 + \pi_2 + \pi_3 = 1$, where π_i is to be interpreted for $i = 1, 2, 3$ as the proportion of the population falling in category c_i. Note that Π is a mathematical copy of U, but its applied meaning is distinct from that of U. The role of S in the general model (X, S, μ, Γ) will be played by the product space $C \times \Pi$ which represents jointly the observation on a single random individual together with the population proportions of c_1, c_2 and c_3. Finally, the role of Γ is played by B where, for any $\mathbf{u}$ in U, the set $B\mathbf{u}$ in $C \times \Pi$ consists of the points $(c_i, \boldsymbol{\pi})$ such that

$$\frac{\pi_i}{u_i} = \max\left(\frac{\pi_1}{u_1}, \frac{\pi_2}{u_2}, \frac{\pi_3}{u_3}\right), \tag{7}$$

for $i = 1, 2, 3$, To understand the definition of B, but not yet the motivation for the definition, it is helpful to visualize $C \times \Pi$ as a stack of three triangles as in Fig. 2 where the three levels correspond to the three points of C. The contributions to $B\mathbf{u}$

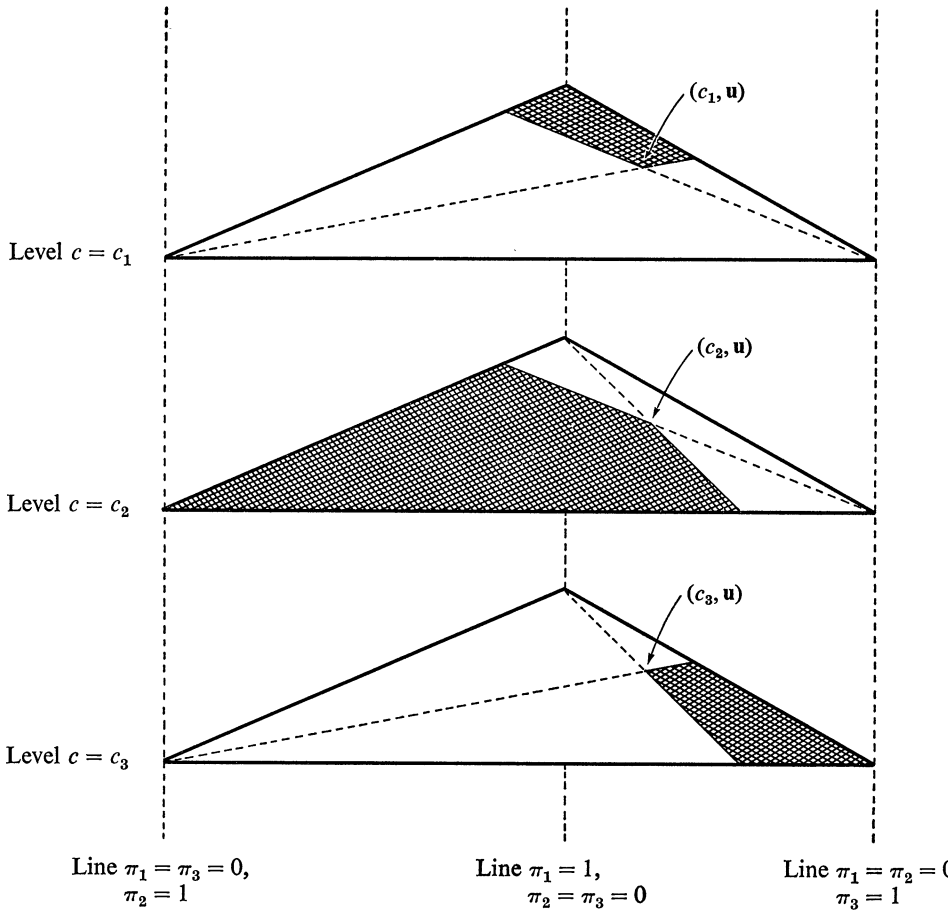

FIG. 2. The space $C \times \Pi$ represented as triangles on three levels. The three closed shaded regions together make up the subset $B\mathbf{u}$ determined from a given $\mathbf{u}$.

from the three levels of $C \times \Pi$ are shown as shaded areas in Fig. 2. It is important also to understand the inverse mapping B^{-1} which carries points of $C \times \Pi$ to subsets of U, where

$$U_i = B^{-1}(c_i, \boldsymbol{\pi}) \tag{8}$$

is defined to be the subset of U consisting of points $\mathbf{u}$ for which $B\mathbf{u}$ contains $(c_i, \boldsymbol{\pi})$. The subsets U_1, U_2, U_3 defined by a given $\boldsymbol{\pi}$ in Π are illustrated in Fig. 1.

It is easily checked with the help of Fig. 1 that

$$\rho(U_i) = \pi_i \quad \text{and} \quad \rho(U_i \cap U_j) = 0 \tag{9}$$

for $i, j = 1, 2, 3$ and $i \neq j$. It will be shown later that the property (9) is a basic requirement for the mapping B defined in (7). Other choices of U and B could be made which would also satisfy (9). Some of these choices amount to little more than adopting different coordinate systems for U, but other possible choices differ in a more fundamental way. Thus an element of arbitrariness enters the model for trinomial sampling at the point of choosing U and B. The present model was introduced in Dempster (1966) under the name *structure of the second kind.* Other possibilities will be mentioned in Section 5.

All of the pieces of the model $(U, C \times \Pi, \rho, B)$ are now in place, so that upper and lower probabilities may be computed for subsets T of $C \times \Pi$. It turns out, however, that $P^*(T) = 1$ and $P_*(T) = 0$ for interesting choices of T, and that interesting illustrations of upper and lower probabilities are apparent only after conditioning. For example, take T to be the event that category c_1 will be observed in a single drawing from the population, i.e. $T = C_1 \times \Pi$, where C_1 is the subset of C consisting of c_1 only. To check that $P^*(T) = 1$ and $P_*(T) = 0$, note (i) that $T^* = U$ because every $\mathbf{u}$ in U lies in U_1 of Fig. 1 for some $(c_1, \boldsymbol{\pi})$ in $C_1 \times \Pi$, and (ii) that T_* is empty because no $\mathbf{u}$ in U lies in U_1 for all $(c_1, \boldsymbol{\pi})$ in $C_1 \times \Pi$. In general, any non-trivial event governed by C alone or by Π alone will have upper probability unity and lower probability zero. Such a result is sensible, for if no information about $\boldsymbol{\pi}$ is put into the system no information about a sample observation should be available, while if no sample observation is in hand there should be no available information about $\boldsymbol{\pi}$. (Recall the interpretation suggested in Section 1 that $P^*(T) = 1$ and $P_*(T) = 0$ should convey a state of complete ignorance about whether or not the real world outcome s will prove to lie in T.)

Turning now to the concept of upper and lower *conditional* probabilities, the definition which fits naturally with the general model (X, S, μ, Γ) arises as follows. If information is received to the effect that sample points in S–T are ruled out of consideration, then the logical assertion "x in X must correspond to s in $\Gamma x \subset S$" is effectively altered to read "x in X must correspond to s in $\Gamma x \cap T \subset S$". Thus the original model (X, S, μ, Γ) is *conditioned on* T by altering (X, S, μ, Γ) to $(X, S, \mu, \tilde{\Gamma})$, where the multivalued mapping $\tilde{\Gamma}$ is defined by

$$\tilde{\Gamma} x = \Gamma x \cap T. \tag{10}$$

Under the conditioned model, an outcome in S–T is regarded as impossible, and indeed the set S–T has upper and lower conditional probabilities both zero. It is sufficient for practical purposes, therefore, to take the conditional model to be $(X, T, \mu, \tilde{\Gamma})$ and to consider upper and lower conditional probabilities only for subsets of T.

Although samples of size one are of little practical interest, the model for a single trinomial observation provides two good illustrations of the definition of a conditioned model. First, it will be shown that conditioning on a fixed value of $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3)$ results in π_i being both the upper and lower conditional probability of an observation c_i, for $i = 1, 2, 3$. This result is equivalent to (9) and explains the importance of (9), since any reasonable model should require that the population proportions be the same as the probabilities of the different possible outcomes in a single random drawing *when the population proportions are known.* Second, it will be shown that non-trivial inferences about $\boldsymbol{\pi}$ may be obtained by conditioning on the observed category c of a single individual randomly drawn from U.

In precise mathematical terms, to condition the trinomial sampling model $(U, C\times\Pi, \rho, B)$ on a fixed $\boldsymbol{\pi}$ is to condition on $T = C\times\tilde{\Pi}$, where $\tilde{\Pi}$ is the subset of Π consisting of the single point $\boldsymbol{\pi}$. T itself consists of the three points $(c_1, \boldsymbol{\pi})$, $(c_2, \boldsymbol{\pi})$ and $(c_3, \boldsymbol{\pi})$ which in turn define single point subsets T_1, T_2 and T_3 of T. The conditioned model may be written $(U, T, \rho, \tilde{B})$ where $\tilde{B}\mathbf{u} = B\mathbf{u}\cap T$ for all $\mathbf{u}$. By referring back to the definition of B as illustrated in Figs. 1 and 2, it is easily checked that the set of $\mathbf{u}$ in U such that $\tilde{B}\mathbf{u}$ intersects T_i is the closed triangle U_i appearing in Fig. 1, while the set of $\mathbf{u}$ in U such that $\tilde{B}\mathbf{u}$ is contained in T_i is the open triangle U_i, for $i = 1, 2, 3$. Whether open or closed, the triangle U_i has measure π_i, and it follows easily from (9) that the upper and lower conditional probabilities of T_i given T are

$$P^*(T_i|T) = P_*(T_i|T) = \pi_i, \tag{11}$$

for $i = 1, 2, 3$, Note that $\tilde{B}\mathbf{u}$ is not empty for any $\mathbf{u}$ in U, so that the denominators in (1) and (2) are both unity in the application (11).

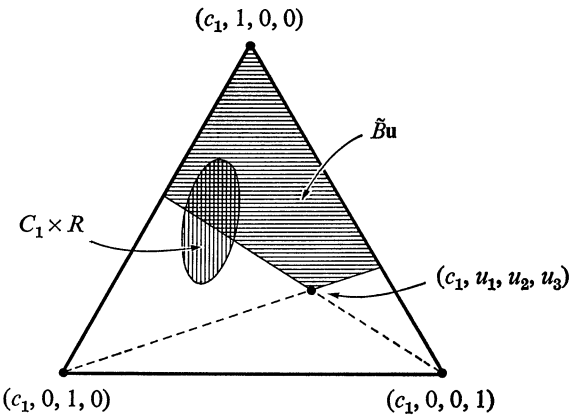

FIG. 3. The triangle $\tilde{T} = C_1\times\Pi$ for the model conditioned on the observation c_1. Horizontal shading covers the region $\tilde{B}\mathbf{u}$, while vertical shading covers a general fixed region $C_1\times R$.

Consider next the details of conditioning the trinomial model on a fixed observation c_1. The cases where a single drawing produces c_2 or c_3 may be handled by permuting indices. Observing c_1 is formally represented by conditioning on $\tilde{T} = C_1\times\Pi$ where C_1 as above is the subset of C consisting of c_1 alone. In the conditional model $(U, \tilde{T}, \rho, \tilde{B})$, the space $\tilde{T}$ is represented by the first level in Fig. 2 while $\tilde{B}\mathbf{u}$ is represented by the closed shaded region in that first level. Since $\tilde{B}\mathbf{u}$ is non-empty for all $\mathbf{u}$ in U, the ρ measure may be used directly without renormalization to compute upper and lower conditional probabilities given $\tilde{T}$. An event R defined as a subset of Π is equivalently represented by the subset $C_1\times R$ of $\tilde{T}$. The upper conditional probability of $C_1\times R$ given $\tilde{T}$ is the probability that the random region $\tilde{B}\mathbf{u}$ intersects $C_1\times R$ where $(c_1, \mathbf{u})$ is uniformly distributed over $C_1\times\Pi$. See Fig. 3. Similarly, the lower conditional probability of $C_1\times R$ given $\tilde{T}$ is the probability that the random region $\tilde{B}\mathbf{u}$ is contained in $C_1\times R$. For example, if R is the lower portion of the triangle where $0\leqslant\pi_1\leqslant\pi_1''$, then

$$P^*(C_1\times R|\tilde{T}) = 1-(1-\pi_1'')^2 = \pi_1''(2-\pi_1'') \quad \text{and} \quad P_*(C_1\times R|\tilde{T}) = 0.$$

Or, in more colloquial notation,

$$P^*(0 \leqslant \pi_1 \leqslant \pi_1'' \mid c = c_1) = \pi_1''(2-\pi_1'') \quad \text{and} \quad P_*(0 \leqslant \pi_1 \leqslant \pi_1'' \mid c = c_1) = 0.$$

More generally, it can easily be checked that

$$P^*(\pi_1' \leqslant \pi_1 \leqslant \pi_1'' \mid c = c_1) = \pi_1''(2-\pi_1''), \tag{12}$$

while

$$\left.\begin{aligned} P_*(\pi_1' \leqslant \pi_1 \leqslant \pi_1'' \mid c = c_1) &= 0 \qquad \text{if } \pi_1'' < 1 \\ &= (1-\pi_1')^2 \quad \text{if } \pi_1'' = 1, \end{aligned}\right\} \tag{13}$$

for any fixed π_1' and π_1'' satisfying $0 \leqslant \pi_1' \leqslant \pi_1'' \leqslant 1$. Likewise,

$$P^*(\pi_2' \leqslant \pi_2 \leqslant \pi_2'' \mid c = c_1) = 1-\pi_2', \tag{14}$$

while

$$\left.\begin{aligned} P_*(\pi_2' \leqslant \pi_2 \leqslant \pi_2'' \mid c = c_1) &= 0 \qquad \text{if } \pi_2' > 0 \\ &= \pi_2'' \quad \text{if } \pi_2' = 0, \end{aligned}\right\} \tag{15}$$

for any fixed π_2' and π_2'' satisfying $0 < \pi_2' < \pi_2'' < 1$. Relations (14) and (15) also hold when subscripts 2 and 3 are interchanged. Formulae (12) to (15) are the first instances of generalized Bayesian inferences reached in this paper, where, as will shortly be explained, prior knowledge of $\boldsymbol{\pi}$ is tacitly assumed to have the null form such that all upper probabilities are unity and all lower probabilities are zero. For example, the model asserts that, if a single random individual is observed to belong in category c_1, and no prior knowledge of $\boldsymbol{\pi}$ is assumed, it may be inferred that at least half the population belongs in c_1 with probability between $\frac{1}{4}$ and 1.

A collection of n models $(X^{(i)}, S, \mu^{(i)}, \Gamma^{(i)})$ for $i = 1, 2, \ldots, n$ may be *combined* or *multiplied* to obtain a *product model* (X, S, μ, Γ). The formal definition of (X, S, μ, Γ) is given by

$$\left.\begin{aligned} X &= X^{(1)} \times X^{(2)} \times \ldots \times X^{(n)}, \\ \text{and} \qquad \mu &= \mu^{(1)} \times \mu^{(2)} \times \ldots \times \mu^{(n)} \\ \Gamma\mathbf{x} &= \Gamma^{(1)} x^{(1)} \cap \Gamma^{(2)} x^{(2)} \cap \ldots \cap \Gamma^{(n)} x^{(n)}, \end{aligned}\right\} \tag{16}$$

where $\mathbf{x} = (x^{(1)}, x^{(2)}, \ldots, x^{(n)})$ denotes a general point of the product space X. The product model is appropriate where the realized values $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$ are regarded as independently random according to the probability measures $\mu^{(1)}, \mu^{(2)}, \ldots, \mu^{(n)}$, while the logical relationships implied by $\Gamma^{(1)}, \Gamma^{(2)}, \ldots, \Gamma^{(n)}$ are postulated to apply simultaneously to a common realized outcome s in S. It may be helpful to view the models $(X^{(i)}, S, \mu^{(i)}, \Gamma^{(i)})$ as separate sources of information about the unknown s in S. In such a view, if the n sources are genuinely independent, then the product rule (16) represents the legitimate way to pool their information.

The concept of a product model actually includes the concept of a conditioned model which was introduced earlier. Proceeding formally, the information that T occurs with certainty may be represented by a degenerate model (Y, S, ν, Δ), where Y consists of a single point y, while $\Delta y = T$ and y carries ν measure unity. Multiplying a general model (X, S, μ, Γ) by (Y, S, ν, Δ) produces essentially the same result as conditioning the general model (X, S, μ, Γ) on T. For $X \times Y$ and $\mu \times \nu$ are isomorphic

in an obvious way to X and μ, while $\Gamma x \cap \Delta y = \Gamma x \cap T = \tilde{\Gamma} x$ as in (10). Thus the objective of taking account of information in the special form of an assertion that T must occur may be reached either through the rule of conditioning or through the rule of multiplication, with identical results. In particular, when $T = S$ the degenerate model (Y, S, ν, Δ) conveys no information about the uncertain outcome s in S, both in the heuristic sense that upper and lower probabilities of non-trivial events are unity and zero, and in the formal sense that combining such a (Y, S, ν, Δ) with any information source (X, S, μ, Γ) leaves the latter model essentially unaltered.

Product models are widely used in mathematical statistics to represent random samples of size n from infinite populations, and they apply directly to provide the general sample size extension of the trinomial sampling model $(U, C \times \Pi, \rho, B)$. A random sample of size n from the population U is represented by $\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \ldots, \mathbf{u}^{(n)}$ independently drawn from U according to the same uniform probability measure ρ. More precisely, the sample $(\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \ldots, \mathbf{u}^{(n)})$ is represented by a single random point drawn from the product space

$$U^n = U^{(1)} \times U^{(2)} \times \ldots \times U^{(n)} \tag{17}$$

according to the product measure

$$\rho^n = \rho^{(1)} \times \rho^{(2)} \times \ldots \times \rho^{(n)}, \tag{18}$$

where the pairs $(U^{(1)}, \rho^{(1)}), (U^{(2)}, \rho^{(2)}), \ldots, (U^{(n)}, \rho^{(n)})$ are n identical mathematical copies of the original pair (U, ρ). In a similar way, the observable categories of the n sample individuals are represented by a point in the product space

$$C^n = C^{(1)} \times C^{(2)} \times \ldots \times C^{(n)}, \tag{19}$$

where $C^{(i)}$ is the three-element space from which the observable category $c^{(i)}$ of the sample individual $\mathbf{u}^{(i)}$ is taken. The interesting unknowns before sampling are $c^{(1)}, c^{(2)}, \ldots, c^{(n)}$ and $\boldsymbol{\pi}$, which define a point in the space $C^n \times \Pi$. Accordingly, the model which represents a random sample of size n from a trinomial population is of the form $(U^n, C^n \times \Pi, \rho^n, B^n)$, where it remains only to define B^n. In words, B^n is the logical relationship which requires that (7) shall hold for each $\mathbf{u}^{(i)}$. In symbols,

$$B^n(\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \ldots, \mathbf{u}^{(n)}) = B^{(1)}\mathbf{u}^{(1)} \cap B^{(2)}\mathbf{u}^{(2)} \cap \ldots \cap B^{(n)}\mathbf{u}^{(n)}, \tag{20}$$

where $B^{(i)}\mathbf{u}^{(i)}$ consists of those points $(c^{(1)}, c^{(2)}, \ldots, c^{(n)}, \boldsymbol{\pi})$ in $C^n \times \Pi$ such that

$$\pi_k / u_k^{(i)} = \max\{(\pi_1 / u_1^{(i)}), (\pi_2 / u_2^{(i)}), (\pi_3 / u_3^{(i)})\} \tag{21}$$

for $k = 1, 2, 3$.

The model $(U^n, C^n \times \Pi, \rho^n, B^n)$ now completely defined provides in itself an illustration of the product rule. For (17), (18) and (20) are instances of the three lines of (16), and hence show that $(U^n, C^n \times \Pi, \rho^n, B^n)$ is the product of the n models $(U^{(i)}, C^n \times \Pi, \rho^{(i)}, B^{(i)})$ for $i = 1, 2, \ldots, n$, each representing an individual sample member.

As in the special case $n = 1$, the model $(U^n, C^n \times \Pi, \rho^n, B^n)$ does not in itself provide interesting upper and lower probabilities. Again, conditioning may be illustrated either by fixing $\boldsymbol{\pi}$ and asking for probability judgments about $c^{(1)}, c^{(2)}, \ldots, c^{(n)}$ or conversely by fixing $c^{(1)}, c^{(2)}, \ldots, c^{(n)}$ and asking for probability judgments (i.e.

generalized Bayesian inferences) about π. Conditioning on fixed π leads easily to the expected generalization of (11). Specifically, if T is the event that π has a specified value, while $\tilde{T}$ is the event that $c^{(1)}, c^{(2)}, \ldots, c^{(n)}$ are fixed, with n_i observations in category c_i for $i = 1, 2, 3$, then

$$P^*(\tilde{T} \mid T) = P_*(\tilde{T} \mid T) = \pi_1^{n_1} \pi_2^{n_2} \pi_3^{n_3}. \tag{22}$$

The converse approach of conditioning on $\tilde{T}$ leads to more difficult mathematics.

Before $c^{(1)}, c^{(2)}, \ldots, c^{(n)}$ are observed, the relevant sample space $C^n \times \Pi$ consists of 3^n triangles, each a copy of Π. Conditioning on a set of recorded observations $c^{(1)}, c^{(2)}, \ldots, c^{(n)}$ reduces the relevant sample space to the single triangle associated with those observations. Although this triangle is actually a subset of $C^n \times \Pi$, it is essentially the same as Π and will be formally identified with Π for the remainder of this discussion. Conditioning the model $(U^n, C^n \times \Pi, \mu^n, B^n)$ on $c^{(1)}, c^{(2)}, \ldots, c^{(n)}$ leads therefore to the model $(U^n, \Pi, \mu^n, \tilde{B}^n)$ where $\tilde{B}^n$ is defined by restricting B^n to the appropriate copy of Π. The important random subset $\tilde{B}^n(\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \ldots, \mathbf{u}^{(n)})$ of Π defined by the random sample $\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \ldots, \mathbf{u}^{(n)}$ will be denoted by V for short. V determines the desired inferences, that is, the upper and lower probabilities of a fixed subset R of Π are respectively the probability that V intersects R and the probability that V is contained in R, both conditional on V being non-empty.

V is the intersection of the n random regions $B^{(i)}\mathbf{u}^{(i)}$ for $i = 1, 2, \ldots, n$ where each $B^{(i)}\mathbf{u}^{(i)}$ is one of the three types illustrated on the three levels of Fig. 2, the type and level depending on whether the observation $c^{(i)}$ is c_1, c_2 or c_3. Fig. 4 illustrates one

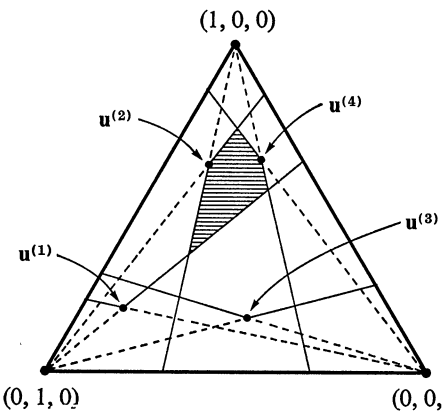

FIG. 4. The triangle Π representing the sample space of unknowns after $n = 4$ observations $c^{(1)} = c_1$, $c^{(2)} = c_3$, $c^{(3)} = c_1$, $c^{(4)} = c_2$ have been taken. The shaded region is the realization of V determined by the illustrated realization of $\mathbf{u}^{(1)}$, $\mathbf{u}^{(2)}$, $\mathbf{u}^{(3)}$ and $\mathbf{u}^{(4)}$.

such region for $n = 4$. It is easily discovered by experimenting with pictures like Fig. 4 that the shaded region V may have 3, 4, 5 or 6 sides, but most often is empty. It is shown in Appendix A that V is non-empty with probability $n_1!\, n_2!\, n_3!/n!$ under independent uniformly distributed $\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \ldots, \mathbf{u}^{(n)}$. Moreover, conditional on non-empty V, six random vertices of V are shown in Appendix A to have Dirichlet distributions. Specifically, define $\mathbf{W}^{(i)}$ for $i = 1, 2, 3$ to be the point π in V with maximum coordinate π_i and define $\mathbf{Z}^{(i)}$ for $i = 1, 2, 3$ to be the point π in V with

minimum coordinate π_i. These six vertices of V need not be distinct, but are distinct with positive probability and so have different distributions. Their distributions are

$$\left.\begin{aligned} \mathbf{W}^{(1)}&: \ D(n_1+1, n_2, n_3), \\ \mathbf{W}^{(2)}&: \ D(n_1, n_2+1, n_3), \\ \mathbf{W}^{(3)}&: \ D(n_1, n_2, n_3+1), \\ \mathbf{Z}^{(1)}&: \ D(n_1, n_2+1, n_3+1), \\ \mathbf{Z}^{(2)}&: \ D(n_1+1, n_2, n_3+1), \\ \mathbf{Z}^{(3)}&: \ D(n_1+1, n_2+1, n_3), \end{aligned}\right\} \tag{23}$$

where $D(r_1, r_2, r_3)$ denotes the Dirichlet distribution over the triangle Π whose probability density function is proportional to

$$\pi_1^{r_1-1}\pi_2^{r_2-1}\pi_3^{r_3-1}.$$

The Dirichlet distribution is defined as a continuous distribution over Π if $r_i > 0$ for $i = 1, 2, 3$. Various conventions, not listed here, are required to cover the distributions of the six vertices when some of the n_i are zero.

Many interesting upper and lower probabilities follow from the distributions (23). For example, the upper probability that π_1 exceeds π_1' is the probability that V intersects the region where $\pi_1 \geqslant \pi_1'$ which is, in turn, the probability that the first coordinate of $\mathbf{W}^{(1)}$ exceeds π_1'. In symbols,

$$\begin{aligned} P^*(\pi_1 \geqslant \pi_1' | n_1, n_2, n_3) &= \int_{\pi_1'}^1 \int_0^1 \frac{n!}{n_1!\,(n_2-1)!\,(n_3-1)!} \pi_1^{n_1}\pi_2^{n_2-1}\pi_3^{n_3-1}\, d\pi_1\, d\pi_2 \\ &= \int_{\pi_1'}^1 \frac{n!}{n_1!\,(n_2+n_3-1)!} \pi_1^{n_1}(1-\pi_1)^{n_2+n_3-1}\, d\pi_1 \end{aligned} \tag{24}$$

if $n_2 > 0$ and $n_3 > 0$. Similarly, $P_*(\pi_1 \geqslant \pi_1' | n_1, n_2, n_3)$ is the probability that the first coordinate of $\mathbf{Z}^{(1)}$ exceeds π_1', that is,

$$P_*(\pi_1 \geqslant \pi_1' | n_1, n_2, n_3) = \int_{\pi_1'}^1 \frac{(n+1)!}{(n_1-1)!\,(n_2+n_3+1)!} \pi_1^{n_1-1}(1-\pi_1)^{n_2+n_3+1}\, d\pi_1, \tag{25}$$

again assuming no prior information about $\boldsymbol{\pi}$. Two further analogues of the pair (24) and (25) may be obtained by permuting the indices so that the role of 1 is played successively by 2 and 3. In a hypothetical numerical example with $n_1 = 2$, $n_2 = 1$, $n_3 = 1$ as used in Fig. 4, it is inferred that the probability of at least half the population belonging in c_1 lies between $\frac{3}{16}$ and $\frac{11}{16}$. In passing, note that the upper and lower probabilities (24) and (25) are formally identical with Bayes posterior probabilities corresponding to the pseudo-prior distributions $D(1, 0, 0)$ and $D(0, 1, 1)$, respectively. This appears to be a mathematical accident with a limited range of applicability, much like the relations between fiducial and Bayesian results pointed out by Lindley (1958). In the present situation, it could be shown that the relations no longer hold for events of the form $(\pi_1' \leqslant \pi_1 \leqslant \pi_1'')$.

The model $(U^n, C^n \times \Pi, \rho^n, B^n)$ has the illuminating feature of remaining a product model *after* conditioning on the sample observations. Recall that the original model $(U^n, C^n \times \Pi, \rho^n, B^n)$ is expressible as the product of the n models

$(U^{(i)}, C^n \times \Pi, \rho^{(i)}, B^{(i)})$ for $i = 1, 2, \ldots, n$. Conditioning the original model on the observations yields $(U^n, \tilde{T}, \rho^n, \tilde{B}^n)$ where, as above, $\tilde{T}$ is the subset of $C^n \times \Pi$ with $c^{(1)}, c^{(2)}, \ldots, c^{(n)}$ fixed at their observed values and

$$\tilde{B}^n(\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \ldots, \mathbf{u}^{(n)}) = B^n(\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \ldots, \mathbf{u}^{(n)}) \cap \tilde{T}. \tag{26}$$

Conditioning the ith component model on the ith sample observation yields $(U^{(i)}, \tilde{T}^{(i)}, \rho^{(i)}, \tilde{B}^{(i)})$, where $\tilde{T}^{(i)}$ is the subset of $C^n \times \Pi$ with $c^{(i)}$ fixed at its observed value, and

$$\tilde{B}^{(i)}\,\mathbf{u}^{(i)} = B^{(i)}\,\mathbf{u}^{(i)} \cap \tilde{T}^{(i)}, \tag{27}$$

for $i = 1, 2, \ldots, n$. It is clear that

$$\tilde{T} = \tilde{T}^{(1)} \cap \tilde{T}^{(2)} \cap \ldots \cap \tilde{T}^{(n)}, \tag{28}$$

and from (20), (26), (27) and (28) it follows that

$$\tilde{B}^n(\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \ldots, \mathbf{u}^{(n)}) = \tilde{B}^{(1)}\,\mathbf{u}^{(1)} \cap \tilde{B}^{(2)}\,\mathbf{u}^{(2)} \cap \ldots \cap \tilde{B}^{(n)}\,\mathbf{u}^{(n)}. \tag{29}$$

From (28) and (29) it is immediate that the model $(U^n, \tilde{T}, \rho^n, \tilde{B}^n)$ is the product of the n models $(U^{(i)}, \tilde{T}^{(i)}, \rho^{(i)}, \tilde{B}^{(i)})$ for $i = 1, 2, \ldots, n$. The meaning of this result is that inferences about $\boldsymbol{\pi}$ may be calculated by traversing two equivalent routes. First, as above, one may multiply the original n models and condition the product on $\tilde{T}$. Alternatively, one may condition the original n models on their associated $\tilde{T}^{(i)}$ and then multiply the conditioned models. The availability of the second route is conceptually interesting, because it shows that the information from the ith sample observation $c^{(i)}$ may be isolated and stored in the form $(U^{(i)}, \tilde{T}^{(i)}, \rho^{(i)}, \tilde{B}^{(i)})$, and when the time comes to assemble all the information one need only pick up the pieces and multiply them. This basic result clearly holds for a wide class of choices of U and B, not just the particular trinomial sampling model illustrated here.

The separability of sample information suggests that prior information about $\boldsymbol{\pi}$ should also be stored as a model of the general type (X, Π, μ, Γ) and should be combined with sample information according to the product rule. Such prior information could be regarded as the distillation of previous empirical data. This proposal brings out the full dimensions of the generalized Bayesian inference scheme. Not only does the product rule show how to combine individual pieces of sample information: it handles the incorporation of prior information as well. Moreover, the sample information and the prior information are handled symmetrically by the product rule, thus banishing the asymmetric appearance of standard Bayesian inference. At the same time, if the prior information is given in the standard form of an ordinary probability distribution, the methods of generalized Bayesian inference reproduce exactly the standard Bayesian inferences.

A proof of the last assertion will now be sketched in the context of trinomial sampling. An ordinary prior distribution for an unknown $\boldsymbol{\pi}$ is represented by a model of the form (X, Π, μ, Γ) where Γ is single-valued and hence no ambiguity is allowed in the computed probabilities. Without loss of generality, the model (X, Π, μ, Γ) may be specialized to (Π, Π, μ, I), where I is the identity mapping and μ is the ordinary prior distribution over Π. For simplicity, assume that μ is a discrete distribution with probabilities $p_1, p_2, \ldots, p_d$ assigned to points $\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots, \boldsymbol{\pi}_d$ in Π. From (16) it follows that the mapping associated with a product of models is single-valued if the mapping associated with any component model is single-valued. If a component model not only has a single-valued mapping, but has a discrete measure μ as well,

then the product model is easily seen to reduce to another discrete distribution over the same carriers $\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots, \boldsymbol{\pi}_d$. Indeed the second line of (16) shows that the product model assigns probabilities $P(\boldsymbol{\pi}_i)$ to $\boldsymbol{\pi}_i$ which are proportional to $p_i l_i$, where l_i is the probability that the random region V includes the point $\boldsymbol{\pi}_i$. Setting $\boldsymbol{\pi}_i = (\pi_{i1}, \pi_{i2}, \pi_{i3})$, it follows from the properties of the random region V that

$$l_i = \pi_{i1}^{n_1} \pi_{i2}^{n_2} \pi_{i3}^{n_3}, \tag{30}$$

which is just the probability that all of the independent random regions whose intersection is V include $\boldsymbol{\pi}_i$. Normalizing the product model as indicated in (1) or (2) leads finally to

$$P(\boldsymbol{\pi}_i) = \frac{p_i l_i}{p_1 l_1 + p_2 l_2 + \ldots + p_d l_d} \tag{31}$$

for $i = 1, 2, \ldots, d$, which is the standard form of Bayes's theorem. This result holds for any choices of U and B satisfying (9). Note that l_i is identical with the likelihood of $\boldsymbol{\pi}_i$.

Generalized Bayesian inference permits the use of sample information alone, which is mathematically equivalent to adopting the informationless prior model in which all upper probabilities are unity and all lower probabilities are zero. At another extreme, it permits the incorporation of a familiar Bayesian prior distribution (if it is a genuine distribution) and then yields the familiar Bayesian inferences. Between these extremes a wide range of flexibility exists. For example, a prior distribution could be introduced for the coordinate π_1 alone, while making no prior judgment about the ratio π_2/π_3. Alternatively, one could specify prior information to be the same as that contained in a sample of size m which produced m_i observations in category c_i for $i = 1, 2, 3$. In the analysis of quite small samples, it would be reasonable to attempt to find some characterization of prior information which could reflect tolerably well public notions about $\boldsymbol{\pi}$. In large samples, the inferences clearly resemble Bayesian inferences and are insensitive to prior information over a wide range.

3. A Second Illustration

Consider a sequence of independent Bernoulli trials represented by z_i with

$$P(z_i = 1 | p_i) = p_i \quad \text{and} \quad P(z_i = 0 | p_i) = 1 - p_i, \quad \text{for } i = 1, 2, \ldots, n, \tag{32}$$

where it is suspected that the sequence p_i is subject to a monotone upward drift. In this situation, the common approach to a sequence of observations z_i is to apply a test of the null hypothesis $\{p_1 = p_2 = \ldots = p_n\}$ designed to be sensitive against the alternative hypothesis $\{p_1 \leqslant p_2 \leqslant \ldots \leqslant p_n\}$. The unorthodox approach suggested here is to compute upper and lower probability inferences for the pair of symmetric hypotheses $\{p_1 \geqslant p_2 \geqslant \ldots \geqslant p_n\}$ and $\{p_1 \leqslant p_2 \leqslant \ldots \leqslant p_n\}$ under the overall prior assumption that the sequence p_i is monotone, either increasing or decreasing, with probability one. A small upper probability for either of these hypotheses would be evidence for drift in the direction contrary to that indicated by the hypothesis. Upper and lower probabilities may also be computed for the null hypothesis $\{p_1 = p_2 = \ldots = p_n\}$, but the upper probability will usually be vanishingly small in sample sequences of moderate length however little trend is apparent, while the lower probability is always zero.

The model described could apply in simple bioassays or learning situations. A wider range of applications could be achieved in several ways, for example by allowing several observations at each p_i or postulating Markov-type dependence in the z_i sequence. But the aim here is to focus attention as simply as possible on one feature of the new methods, namely their ability to handle the problem of many nuisance parameters which plagues the more traditional forms of statistical inference. Plausible inferences may be obtained despite the presence of as many continuous parameters as there are dichotomous observables.

Under the binomial analogue of the trinomial model treated in Section 2, a single binomial observable z is represented before observation by the model $(U, Z \times P, \rho, B)$ where

$$U = \{u: 0 \leqslant u \leqslant 1\}, \tag{33}$$

$$Z = \{z: z = 0 \text{ or } z = 1\}, \tag{34}$$

$$P = \{p: 0 \leqslant p \leqslant 1\}, \tag{35}$$

ρ is the uniform distribution over U, and

$$Bu = \{(z,p): z = 0 \text{ and } u \leqslant p \leqslant 1, \text{ or } z = 1 \text{ and } 0 \leqslant p \leqslant u\}. \tag{36}$$

After conditioning on z, this model becomes effectively (U, P, ρ, B_z), where

$$B_z u = \{p: u \leqslant p \leqslant 1\} \quad \text{if } z = 0,$$
$$= \{p: 0 \leqslant p \leqslant u\} \quad \text{if } z = 1. \tag{37}$$

A conditioned model of this kind may be constructed for each of n independent observations z_i and associated parameters p_i. Combining these n sources of information about $p_1, p_2, \ldots, p_n$ yields a single model $(U^n, P^n, \rho^n, B_{(z_1,z_2,\ldots,z_n)})$, where

$$U^n = \{(u_1, u_2, \ldots, u_n): 0 \leqslant u_i \leqslant 1 \text{ for } i = 1,2,\ldots,n\}, \tag{38}$$

$$P^n = \{(p_1, p_2, \ldots, p_n): 0 \leqslant p_i \leqslant 1 \text{ for } i = 1,2,\ldots,n\}, \tag{39}$$

ρ^n is the uniform distribution over the cube U, and

$$B_{(z_1,z_2,\ldots,z_n)}(u_1, u_2, \ldots, u_n) = \{(p_1, p_2, \ldots, p_n): p_i \in B_{z_i} u_i \text{ for } i = 1,2,\ldots,n\}. \tag{40}$$

The combined model would be appropriate for unrestricted inferences about an unknown $(p_1, p_2, \ldots, p_n)$ based on observations $(z_1, z_2, \ldots, z_n)$. However, when consideration is restricted to the subset S of P^n in which $p_1, p_2, \ldots, p_n$ is a monotone sequence, the sharpness of the inferences is much improved.

Define T_1 and T_2 to be the subsets of S for which $p_1 \leqslant p_2 \leqslant \ldots \leqslant p_n$ and $p_1 \geqslant p_2 \geqslant \ldots \geqslant p_n$, respectively. Define $T_{12} = T_1 \cap T_2$ to be the subset of S for which $p_1 = p_2 = \ldots = p_n$. An immediate objective is to characterize T_1^*, T_2^* and T_{12}^*, from whose ρ^n measure the desired inferences will follow. For example, T_1^* consists of all points $(u_1, u_2, \ldots, u_n)$ for which there exists some $(p_1, p_2, \ldots, p_n)$ satisfying $p_1 \leqslant p_2 \leqslant \ldots \leqslant p_n$ and such that p_i lies in $B_{z_i} u_i$, for $i = 1, 2, \ldots, n$. With the help of Fig. 5 it is easily checked that

$$T_1^* = \{(u_1, u_2, \ldots, u_n): u_i \leqslant u_j, \text{ whenever } z_i = 1,\ z_j = 0,\ i < j\}. \tag{41}$$

By symmetry,

$$T_2^* = \{(u_1, u_2, \ldots, u_n)\colon u_i \geqslant u_j, \text{ whenever } z_i = 0,\ z_j = 1,\ i<j\}. \tag{42}$$

Finally,

$$T_{12}^* = \{(u_1, u_2, \ldots, u_n)\colon u_i \leqslant u_j, \text{ whenever } z_i = 1,\ z_j = 0\}. \tag{43}$$

It is clear that $T_{12}^* = T_1^* \cap T_2^*$ and that T_{12}^*, $T_1^* - T_{12}^*$ and $T_2^* - T_{12}^*$ are disjoint sets whose union is S^*.

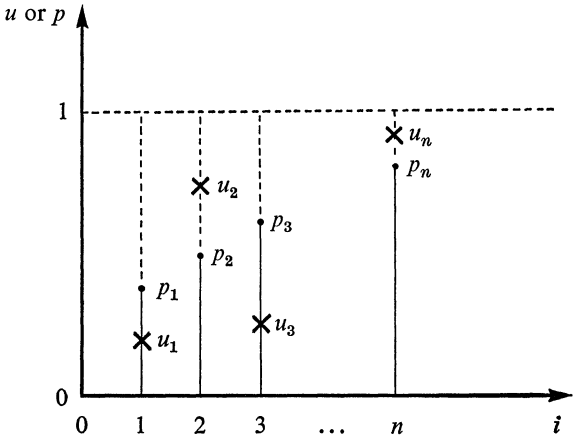

FIG. 5. The plotted values $p_1, p_2, \ldots, p_n$ determine a point P^n for which

$$p_1 \leqslant p_2 \leqslant \ldots \leqslant p_n.$$

The plotted values $u_1, u_2, \ldots, u_n$ determine a point of U^n for which p_1 lies in $B_1\, z_1$, p_2 lies in $B_0\, z_2$, p_3 lies in $B_1\, z_3$, ..., p_n lies in $B_0\, z_n$. The interpretation is that $(u_1, u_2, \ldots, u_n)$ lies in the region T_1^* determined by the observation $z_1 = 1$, $z_2 = 0$, $z_3 = 1$, ..., $z_n = 0$.

U^n may be decomposed into $n!$ geometrically similar simplexes, each characterized by a particular ordering of the values of the coordinates $(u_1, u_2, \ldots, u_n)$. These simplexes are in one-to-one correspondence with the permutations

$$(1, 2, \ldots, n) \to (1^*, 2^*, \ldots, n^*),$$

where for every $(u_1, u_2, \ldots, u_n)$ in a given simplex the corresponding permutation obeys $u_{1^*} \leqslant u_{2^*} \leqslant \ldots \leqslant u_{n^*}$. Since the characterizations (41), (42) and (43) involve only order relations among coordinates u_i, each of the simplexes is either included or excluded as a unit from T_1^* or T_2^* or T_{12}^*. And since each of the $n!$ simplexes has ρ^n measure $1/n!$, the ρ^n measures of T_1^* or T_2^* or T_{12}^* may be found by counting the appropriate number of simplexes and dividing by $n!$. Or, instead of counting simplexes, one may count the permutations to which they correspond. The permutation

$$(1, 2, \ldots, n) \to (1^*, 2^*, \ldots, n^*)$$

carries the observed sequence $(z_1, z_2, \ldots, z_n)$ of zeros and ones into another sequence $(z_{1^*}, z_{2^*}, \ldots, z_{n^*})$ of zeros and ones. According to the definition of T_1^*, a simplex is contained in T_1^* if and only if its corresponding permutation has the property that $i^* < j^*$ for all $i<j$ such that $z_i = 1$ and $z_j = 0$, i.e. any pair ordered $(1, 0)$ extracted from $(z_1, z_2, \ldots, z_n)$ must retain the same order in the permuted sequence

$(z_{1*}, z_{2*}, \ldots, z_{n*})$. Similarly, to satisfy T_2^* any pair ordered $(0,1)$ extracted from $(z_1, z_2, \ldots, z_n)$ must have its order reversed in the permuted sequence, while to satisfy $T_{12}^* = T_1^* \cap T_2^*$ the sequence $(z_{1*}, z_{2*}, \ldots, z_{n*})$ must consist of all ones followed by all zeros.

If $(z_1, z_2, \ldots, z_n)$ contains n_1 ones and n_2 zeros, then a simple counting of permutations yields

$$\rho(T_{12}^*) = \frac{n_1!\, n_2!}{n!}. \tag{44}$$

A simple iterative procedure for computing $\rho^n(T_1^*)$ or $\rho^n(T_2^*)$ is derived in Appendix B by Herbert Weisberg. The result is quoted below and illustrated on a numerical example.

For a given sequence of observations $z_1, z_2, \ldots$ of indefinite length define $N(n)$ to be the number of permutations of the restricted type counted in T_1^*. $N(n)$ may be decomposed into

$$N(n) = \sum_{k=0}^{r} N(k,n), \tag{45}$$

where $N(k,n)$ counts the subset of permutations such that $(z_{1*}, z_{2*}, \ldots, z_{n*})$ has k zeros preceding the rightmost one. Since no zero which follows the rightmost one in the original sequence $(z_1, z_2, \ldots, z_n)$ can be permuted to the left of any one under any allowable permutation, the upper limit r in (45) may be taken as the number of zeros preceding the rightmost one in the original sequence $(z_1, z_2, \ldots, z_n)$. In the special case of a sequence consisting entirely of zeros, all of the zeros will be assumed to follow the rightmost one so that $N(k,n) = 0$ for $k > 0$ and indeed $N(n) = N(0,n) = n!$. Weisberg's iterative formula is

$$\begin{aligned} N(k,n+1) &= \sum_{j=0}^{k-1} N(j,n) + (n_1 + 1 + k)\, N(k,n) \quad \text{if } z_{n+1} = 1 \\ &= (n_2 + 1 - k)\, N(k,n) \quad \text{if } z_{n+1} = 0, \end{aligned} \tag{46}$$

where n_1 and n_2 denote as above the numbers of ones and zeros, respectively, in $(z_1, z_2, \ldots, z_n)$.

Formula (46) has the pleasant feature that the counts for the sequences $(z_1), (z_1, z_2), (z_1, z_2, z_3), \ldots$ may be built up successively, and further observations may be easily incorporated as they arrive. Consider, for example, the hypothetical observations

$$(z_1, z_2, \ldots, z_7) = (0,0,1,1,0,1,1).$$

Table 1 shows

$$z_n, N(0,n), \ldots, N(r,n)$$

on line n, for $n = 1, 2, \ldots, 7$, from which $N(7) = 1680$. The number of permutations consistent with T_2^* is found by applying the same iterative process to the sequence $(1,1,0,0,1,0,0)$ with zeros and ones interchanged. This yields Table 2 from which $N(7) = 176$. The number of permutations common to T_1^* and T_2^* is $3!\,4! = 144$. Thus $\rho^n(T_1^*) = 1680/7!$, $\rho^n(T_2^*) = 176/7!$, $\rho^n(T_{12}^*) = 144/7!$, and $\rho^n(S^*) = (1680 + 176 - 144)/7! = 1712/7!$. Consequently, the upper and lower

TABLE 1

n	z_n	$N(0,n)$	$N(1,n)$	$N(2,n)$	$N(3,n)$
1	0	1			
2	0	2			
3	1	2,	2,	2	
4	1	4,	8,	12	
5	0	12,	16,	12	
6	1	36,	76,	84,	40
7	1	144,	416,	640,	480

TABLE 2

n	z_n	$N(0,n)$	$N(1,n)$	$N(2,n)$
1	1	1		
2	1	2		
3	0	2		
4	0	4		
5	1	12,	4,	4
6	0	36,	8,	4
7	0	144,	24,	8

probabilities of T_1, T_2 and T_{12} conditional on S and $(z_1, z_2, \ldots, z_7) = (0,0,1,1,0,1,1)$ are

$$P^*(T_1) = \frac{1680}{1712}, \quad P_*(T_1) = \frac{1536}{1712}, \quad P^*(T_2) = \frac{176}{1712}, \quad P_*(T_2) = \frac{32}{1712},$$

$$P^*(T_{12}) = \frac{144}{1712}, \quad P_*(T_{12}) = 0.$$

Since more than 10 per cent of the measure could apply to a monotone non-increasing sequence, the evidence for an increasing sequence is not compelling.

TABLE 3

n	$P_*(T_2)$	$P^*(T_2)$
1	0	1
2	0	1
3	0	0·333
4	0	0·167
5	0·167	0·417
6	0·048	0·190
7	0·019	0·103
8	0·188	0·319
9	0·065	0·148
10	0·028	0·080
11	0·014	0·047

For the extended sequence of observations $0,0,1,1,0,1,1,0,1,1,1,\ldots$, the lower and upper probabilities of a monotone downward sequence after n observations are exhibited in Table 3.

4. Comments on the Method of Generating Upper and Lower Probabilities

Although often notationally convenient, it is unnecessary to use models (X, S, μ, Γ) outside of the subclass where the inverse of Γ is single-valued. For the model $(X, \tilde{S}, \mu, \tilde{\Gamma})$ with

$$\tilde{S} = X \times S \tag{47}$$

and

$$\tilde{\Gamma} x = \{x\} \times \Gamma x \tag{48}$$

does belong to the stated subclass, and yields

$$(P^*(T), P_*(T)) = (\tilde{P}^*(X \times T), \tilde{P}_*(X \times T)) \tag{49}$$

for any $T \subset S$, where the left side of (49) refers to any original model (X, S, μ, Γ) and the right side refers to the corresponding model $(X, \tilde{S}, \mu, \tilde{\Gamma})$. Moreover, the model $(X, \tilde{S}, \mu, \tilde{\Gamma})$ provides upper and lower probabilities for all subsets of $X \times S$, not just those of the form $X \times T$. On the other hand, it was assumed in applying the original form (X, S, μ, Γ) that the outcome x in X is conceptually unobservable, so that no operational loss is incurred by the restriction to subsets of the form $X \times T \subset \tilde{S}$.

Underlying the formalism of (X, S, μ, Γ) or its equivalent $(X, \tilde{S}, \mu, \tilde{\Gamma})$ is the idea of a probability model which assigns a distribution only over a partition of a complete sample space, specifically the distribution μ over the partition of $\tilde{S} = X \times S$ defined by X. Thus the global probability law of an ordinary probability measure space is replaced by a marginal distribution or what might be called a *partial* probability law. The aim therefore is to establish a useful probability calculus on marginal or partial assumptions.

I believe that the most serious challenges to applications of the new calculus will come not from criticism of the logic but from the strong form of ignorance which is necessarily built into less-than-global probability laws. To illustrate, consider a simple example where w_1 denotes a measured weight, w_2 denotes a true weight, and $x = w_1 - w_2$ denotes a measurement error. Assume that ample relevant experience is available to justify assigning a specific error distribution μ over the space X of possible values of x. The situation may be represented by the model (X, W, μ, Γ) with X and μ as defined, with $W = \{(w_1, w_2);\ w_1 \geqslant 0,\ w_2 \geqslant 0\}$, and Γ defined by the relation $x = w_1 - w_2$. Conditioning the model on an observed w_1 leaves one with the same measure μ applied to $w_1 - w_2$, except for renormalization which restricts the measure to $w_1 \geqslant 0$. The result is very much in the spirit of the fiducial argument (although there is some doubt about Fisher's attitude to renormalization). I am unable to fault the logic of this fiducial-like argument. Rather, some discomfort is produced by distrust of the initial model, in particular by its implication that every uncertain event governed by the true weight w_2 has initial upper and lower probabilities one and zero. It would be hard to escape a feeling in most real situations that a good bit of information about a parameter is available, even if difficult to formalize objectively, and that such information should clearly alter the fiducial-like inference if it could be incorporated. One way to treat this weakness is openly to eschew the use of prior information, while not necessarily denying its existence, that is, to assert that the statistician should summarize only that information which relies on the observation w_2 and the objectively based error distribution μ. Because of the conservatism

implicit in the definition of upper and lower probabilities, the approach of rejecting soft information seems likely to provide conservative inferences on an average, but I have not proved theorems to this effect. The difficulty is that the rejection of all soft information, including even information about parametric forms, may lead to unrealistically weak inferences. The alternative approach is to promote vague information into as precise a model as one dares and combine it in the usual way with sample information.

Some comments on the mathematics of upper and lower probabilities are appropriate. A very general scheme for assigning upper and lower probabilities to the subsets of a sample space S is to define a family $\mathscr{C}$ of measures P over S and to set

$$P^*(T) = \sup_{\mathscr{C}} P(T), \quad P_*(T) = \inf_{\mathscr{C}} P(T). \tag{50}$$

Within the class of systems of upper and lower probabilities achieved in this way for different $\mathscr{C}$, there is a hierarchical scheme of shrinking subclasses ending with the class of systems defined by models like (X, S, μ, Γ). (see Dempster, 1967a). The family $\mathscr{C}$ corresponding to a given (X, S, μ, Γ) consists of all measures P which for each x distribute the probability element $d\mu(x)$ in some way over Γx. Some readers may feel that all systems should be allowed, not just the subclass of this paper. In doing so, however, one loses the conception of a source of information as being a single probability measure. For, in the unrestricted formulation of (50), the class $\mathscr{C}$ consists of conceptually distinct measures such as might be adopted by a corresponding class of personalist statisticians, and the conservatism in the bounds of (50) amounts to an attempt to please both extremes in the class of personalist statisticians. I believe that the symmetry arguments underlying probability assignments do not often suggest hypothetical families $\mathscr{C}$ demanding simultaneous satisfaction. Also, the rules of conditioning and, more generally, of combination of independent sources of information do not extend to the unrestricted system (50), and without these rules the spirit of the present approach is lost.

The aim of this short section has been to suggest that upper and lower probabilities generated by multivalued mappings provide a flexible means of characterizing limited amounts of information. They do not solve the difficult problems of what information should be used, and of what model appropriately represents that information. They do not provide the only way to discuss meaningful upper and lower probabilities. But they do provide an approach with a well-rounded logical structure which applies naturally in the statistical context of drawing inferences from samples to populations.

5. Comments on the Models used for Inference

The models used here for the representation of sampling from a population take as their point of departure a space whose elements correspond to the members of the population. In addition to the complex of observable characteristics usually postulated in mathematical statistics, each population member is given an individual identity. In conventional mathematical statistics the term *hypothesis* is often used for an unknown population distribution of observable characteristics, but the presence of the population space in the model leads directly to the more fundamental question of how each hypothesized population distribution applies to the elements of the population space, that is, under a given hypothesis what are the observable characteristics of each population member? In the trinomial sampling model of Section 2, the question

is answered by the multivalued mapping B defined in (7). As illustrated in Fig. 1, B asserts that for each hypothesis $\boldsymbol{\pi}$ the population space U partitions into three regions U_1, U_2, U_3 corresponding to the observable characteristics c_1, c_2, c_3. More generally, the observable characteristics may be multinomial with k categories $c_1, c_2, \ldots, c_k$ and the population space U may be any space with an associated random sampling measure ρ. For a given hypothesis $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_k)$ the question is answered by determining subsets $U_1, U_2, \ldots, U_k$ of U which specify that a population member in U_i is permitted to have characteristic c_i under $\boldsymbol{\pi}$, for $i = 1, 2, \ldots, k$. Having reached this point in building the model, it seems reasonable to pose the restriction which generalizes (9), namely,

$$\rho(U_i) = \pi_i \quad \text{and} \quad \rho(U_i \cap U_j) = 0 \tag{51}$$

for $i, j = 1, 2, \ldots, k$ and $i \neq j$. The reason for (51) as with (9) is simply to have π_i represent both upper and lower probabilities of c_i for a single drawing with a given $\boldsymbol{\pi}$.

Now it is evident that the above information by no means uniquely determines a model for multinomial sampling. Indeed, one may start from any continuous space U with measure ρ, and for each $\boldsymbol{\pi}$ specify a partition $U_1, U_2, \ldots, U_k$ satisfying (51) but otherwise completely arbitrary. In other words, there is a huge bundle of available models. In Dempster (1966), two choices were offered which I called *models of the first kind* and *models of the second kind.* The former assumes that the multinomial categories $c_1, c_2, \ldots, c_k$ have a meaningful order, and is uniquely determined by the assumption that the population members have an order consistent with the order of their observable characteristics under any hypothesis $\boldsymbol{\pi}$. (See Dempster, 1967b). The restriction to ordered categories implies essentially a univariate characteristic, and because that restriction is so severe the following discussion is mostly aimed at a general multinomial situation with no mathematical structure assumed on the space of k categories. The general model of the second kind is defined by extending (5), (6) and (7) in the obvious way from $k = 3$ to general k. This model treats the k categories with complete symmetry, but it is not the only model to do so, for one can define B^{-1} arbitrarily for $\boldsymbol{\pi}$ such that $\pi_1 \leqslant \pi_2 \leqslant \ldots \leqslant \pi_k$, and define B^{-1} for other $\boldsymbol{\pi}$ by symmetry. But the general model of the second kind is strikingly simple, and I recommend it because I can find no competitor with comparable aesthetic appeal.

The status of generalized Bayesian inference resembles that of Bayesian inference in the time of Bayes, by which I mean that Bayes must have adopted a uniform prior distribution because no aesthetically acceptable competitor came to mind. The analogy should be carried further, for even the principles by which competitors should be judged were not formulated by Bayes, nor have the required principles been well formulated for the models discussed here. I believe that the principles required by the two situations are not at all analogous, for the nature and meaning of a prior distribution has become quite clear over the last two centuries and the concept may be carried more or less whole over to generalized Bayesian inference. The choice of a model satisfying (51), on the other hand, has no obvious connection with prior information as the term is commonly applied relative to information about postulated unknowns. In the case of generalized Bayesian inference, I believe the principles for choosing a model to be closely involved with an *uncertainty principle* which can be stated loosely as: *The more information which one extracts from each sample individual in the form of observable characteristics, the less information about any given aspect of the population distribution may be obtained from a random sample of fixed size.*

For example, a random sample of size $n = 1000$ from a binomial population yields quite precise and nearly objective inferences about the single binomial parameter p involved. On the other hand, if a questionnaire given to a sample of $n = 1000$ has been sufficient to identify each individual with one of 1,000,000 categories, then it may be foolhardy to put much stock in the sample information about a binomial p chosen arbitrarily from among the $2^{1,000,000} - 2$ non-trivial available possibilities. Conceptually, at least, most real binomial situations are of the latter kind, for a single binomial categorization can be achieved only at the expense of suppressing a large amount of observable information about each sample individual. The uncertainty principle is therefore a specific instance of the general scientific truism that an investigator must carefully delimit and specify his area of investigation if he is to learn anything precise.

Generalized Bayesian inference makes possible precise formulations of the uncertainty principle. For example, the model of the second kind with $k = 2$ and $n = 1000$ yields inferences which most statisticians would find nearly acceptable for binomial sampling. On the other hand, it is a plausible conjecture that the model of the second kind with $k = 1{,}000{,}000$ and $n = 1000$ would yield widely separated upper and lower probabilities for most events. The high degree of uncertainty in each inference compensates for the presence of a large number of nuisance parameters, and protects the user against selection effects which would produce many spurious inferences. Use of the model of the first kind with $k = 1{,}000{,}000$ and $n = 1000$ would very likely lead to closer bounds than the model of the second kind for binomial inferences relating to population splits in accord with the given order of population members. And it is heuristically clear that models could be constructed which for each $\boldsymbol{\pi}$ would place each point of U in each of $U_1, U_2, \ldots, U_k$ as $\boldsymbol{\pi}^*$ varies over an arbitrarily small neighbourhood about $\boldsymbol{\pi}$. Such a model would present an extreme of uncertainty, for all upper and lower probability inferences would turn out to be one and zero, respectively. It is suggested here that the choice of a model can only be made with some understanding of the specific reflections of the uncertainty principle which it provides. For the time being, I judge that the important task is to learn more about the inferences yielded by the aesthetically pleasing models of the second kind. Eventually, intuition and experience may suggest a broader range of plausible models.

Models of the second kind were introduced above for sampling from a general multinomial population with k categories and unknown $1 \times k$ parameter vector $\boldsymbol{\pi}$. But the range of application of these models is much wider. First, one may restrict $\boldsymbol{\pi}$ to parametric hypotheses of the general form $\boldsymbol{\pi} = \boldsymbol{\pi}(\theta, \phi, \ldots)$. More important, the multinomial may be allowed to have an infinite number of categories, as explained in Dempster (1966), so that general spaces of discrete and continuous observable characteristics are permissible. It is possible therefore to handle the standard parametric hypotheses of mathematical statistics. Very few of these have as yet proved analytically tractable.

At present, mainly qualitative insights are available into the overview of statistical inference which the sampling models of generalized Bayesian inference make possible. Some of these insights have been mentioned above, such as the symmetric handling of prior and sample information, and the uncertainty principle by which upper and lower probabilities reflect the degree of confusion produced by small samples from complex situations. It is interesting to note also that parametric hypotheses and prior distributions, which are viewed as quite different in conventional statistical theory, play indistinguishable roles in the logical machinery of generalized Bayesian inference. For a parametric hypothesis such as $\boldsymbol{\pi} = \boldsymbol{\pi}(\theta, \phi, \ldots)$ may be represented by a model of

the general type (X, S, μ, Γ), which assigns all of its probability ambiguously over the subset of π allowed by $\pi(\theta, \phi, \ldots)$ as $\theta, \phi, \ldots$ range over their permitted values, and this model combines naturally with sample information using the rule of combination defined in Section 2 and suggested there to be appropriate for the introduction of prior information.

Concepts which appear in standard theories of inference may reappear with altered roles in generalized Bayesian inference. *Likelihood* is a prime example. The ordinary likelihood function $L(\pi)$ based on a sample from a general multinomial population is proportional to the upper probability of the hypothesis π. This may be verified in the trinomial example of Section 2 by checking that the random region illustrated in Fig. 4 covers the point π with probability $\pi_1^{n_1}\pi_2^{n_2}\pi_3^{n_3}$. The general result is hardly more difficult to prove. Now the upper probability of π for all π does not contain all the sample information under generalized Bayesian inference. Thus the likelihood principle fails in general, and the usual sets of sufficient statistics under exponential families of parametric hypotheses no longer contain all of the sample information. The exception occurs in the special case of ordinary Bayesian inference with an ordinary prior distribution, as illustrated in (31). Thus the failure of the likelihood principle is associated with the uncertainty which enters when upper and lower probabilities differ. In passing, note that marginal likelihoods are defined in the general system, that is, the upper probabilities of specific values of θ from a set of parameters $\theta, \phi, \ldots$ are well defined and yield a function $L(\theta)$ which may be called the marginal likelihood of θ alone. If the prior information consists of an ordinary prior distribution of θ alone, with no prior information about the nuisance parameters, then $L(\theta)$ contains all of the sample information about θ.

Unlike frequency methods, which relate to sequences of trials rather than to specific questions, the generalized Bayesian inference framework permits direct answers to specific questions in the form of probability inferences. I find that significance tests are inherently awkward and unsatisfying for questions like that posed in the example of Section 4, and the main reason that Bayesian inference has not replaced most frequency procedures has been the stringent requirement of a precise prior distribution. I hope that I have helped to reduce the stringency of that requirement.

REFERENCES

BOOLE, G. (1854). *An Investigation of the Laws of Thought.* New York: Reprinted by Dover (1958).

DEMPSTER, A. P. (1966). New methods for reasoning towards posterior distributions based on sample data. *Ann. Math. Statist.*, **37**, 355–74.

—— (1967a). Upper and lower probabilities induced by a multivalued mapping. *Ann. Math. Statist.*, **38**, 325–39.

—— (1967b). Upper and lower probability inferences based on a sample from a finite univariate population. *Biometrika*, **54**, 515–528.

GOOD, I. J. (1962). The measure of a non-measurable set. *Logic, Methodology and Philosophy of Science* (edited by Ernest Nagel, Patrick Suppes and Alfred Tarski), pp. 319–329. Stanford University Press.

LINDLEY, D. V. (1958). Fiducial distributions and Bayes' theorem. *J. R. Statist. Soc.* B, **20**, 102–107.

SMITH, C. A. B. (1961). Consistency in statistical inference and decision (with discussion). *J. R. Statist. Soc.* B, **23**, 1–25.

—— (1965). Personal probability and statistical analysis (with discussion). *J. R. Statist. Soc.* A, **128**, 469–499.

APPENDIX A

A derivation is sketched here for the distributions (23) relating to specific vertices of the random region R defined by (20). R is the intersection of n regions $B^{(i)}\mathbf{u}^{(i)}$, for $i = 1, 2, \ldots, n$, as illustrated in Fig. 4. The region $B^{(i)}\mathbf{u}^{(i)}$ corresponding to $\mathbf{u}^{(i)}$, which gives rise to an observation c_1, consists of points $\mathbf{u}$ such that $u_3/u_1 \leqslant u_3^{(i)}/u_1^{(i)}$ and $u_2/u_1 \leqslant u_2^{(i)}/u_1^{(i)}$. The intersection of the n_1 regions corresponding to the n_1 observations c_1 is a region R_1 consisting of points $\mathbf{u}$ such that

$$u_3/u_1 \leqslant c_{13} \quad \text{and} \quad u_2/u_1 \leqslant c_{12}, \tag{A.1}$$

where $c_{13} = \min(u_3^{(i)}/u_1^{(i)})$ and $c_{12} = \min(u_2^{(i)}/u_1^{(i)})$, the minimization being over the subset of i corresponding to observations c_1. Note that R_1 together with the n_1 regions which define it are all of the type pictured on level 1 of Fig. 2. By permuting subscripts, define the analogous regions R_2 with coordinates c_{23}, c_{21} and R_3 with coordinates c_{31}, c_{32}, where R_2 and R_3 are of the types pictured on levels 2 and 3 of Fig. 2, respectively. One is led thus to the representation

$$R = R_1 \cap R_2 \cap R_3. \tag{A.2}$$

Any particular instance of the region R which contains at least one point is a closed polygon whose sides are characterized by fixed ratios of pairs of coordinates u_i, u_j. Thus R may be described by a set of six coordinates

$$b_{ij} = \max_{\mathbf{u} \in R}(u_j/u_i) \tag{A.3}$$

for $i \neq j$. From (A.1), (A.2), and (A.3) it follows that

$$b_{ij} \leqslant c_{ij} \tag{A.4}$$

for $i \neq j$. Moreover, equality holds if the corresponding side of R_i is also a side of R, while the inequality is strict if the side of R_i misses R entirely. The reader may wish to satisfy himself that R may have 3, 4, 5 or 6 sides in which case the strict inequality in (A.4) holds for 3, 4, 5 or 6 pairs i, j (with probability one).

If R is considered a random region, while R^0 is a fixed region of the same type with coordinates b_{ij}^0, then

$$\begin{aligned} \mathrm{P}(R \supset R^0) &= \mathrm{P}(b_{ij} \geqslant b_{ij}^0) \quad \text{for all } i \neq j \\ &= (1 + b_{12}^0 + b_{13}^0)^{-n_1}(1 + b_{21}^0 + b_{23}^0)^{-n_2}(1 + b_{31}^0 + b_{32}^0)^{-n_3}. \end{aligned} \tag{A.5}$$

To prove (A.5) note first that the three events

$$\{b_{12} \geqslant b_{12}^0, b_{13} \geqslant b_{13}^0\}, \quad \{b_{21} \geqslant b_{21}^0, b_{23} \geqslant b_{23}^0\}, \quad \{b_{31} \geqslant b_{31}^0, b_{32} \geqslant b_{32}^0\}$$

are equivalent respectively to the three events

$$\{c_{12} \geqslant b_{12}^0, c_{13} \geqslant b_{13}^0\}, \quad \{c_{21} \geqslant b_{21}^0, c_{23} \geqslant b_{23}^0\}, \quad \{c_{31} \geqslant b_{31}^0, c_{32} \geqslant b_{32}^0\}.$$

In the latter form, the three events are clearly independent, for they depend on disjoint sets of independent $\mathbf{u}^{(i)}$, and their three probabilities are the three factors in (A.5). For example, the first event says that the n_1 points $\mathbf{u}^{(i)}$ corresponding to observations c_1 fall in the subtriangle $u_2/u_1 \geqslant b_{12}^0$ and $u_3/u_1 \geqslant b_{13}^0$ whose area is the fraction $(1 + b_{12}^0 + b_{13}^0)^{-1}$ of the area of the whole triangle U.

It will be convenient to denote the right side of (A.5) by $F(b^0_{12}, b^0_{13}, b^0_{21}, b^0_{23}, b^0_{31}, b^0_{32})$ which defines, as the b^0_{ij} vary, a form of the joint cumulative distribution function of the b_{ij}. This c.d.f. should be handled with care. First, it is defined only over the subset of the positive orthant in six dimensions such that the b^0_{ij} define a non-empty R^0. Many points in the orthant are ruled out by relations like $b^0_{12} \leqslant b^0_{13} b^0_{32}$ which are implicit in (A.3). Second, the distribution of the b_{ij} is not absolutely continuous over its six-dimensional domain, but assigns finite probability to various boundary curved surfaces of dimensions 5, 4 and 3, corresponding to random R with 5, 4 and 3 sides. Nevertheless it is not difficult to deduce (23) from (A.5).

Suppose that u^* denotes the vertex of R with maximum first coordinate. This vertex lies, with probability one, at the intersection of two of the six sides of R_1, R_2 and R_3. By looking at the vertices defined by all possible pairs of sides it is easily checked that exactly three possibilities exist for $\mathbf{u}^*$, namely,

$$\left.\begin{array}{lll} \text{(i)} & u^*_1/u^*_2 = c_{21} \text{ and } u^*_1/u^*_3 = c_{31}, & \\ \text{(ii)} & u^*_3/u^*_2 = c_{23} \text{ and } u^*_1/u^*_3 = c_{31}, & \text{or} \\ \text{(iii)} & u^*_1/u^*_2 = c_{21} \text{ and } u^*_2/u^*_3 = c_{32}. & \end{array}\right\} \tag{A.6}$$

The probability density function of u^* may be formed by summing the contributions from the three possibilities (i), (ii), (iii). The contribution from case (i) will be expressed first in terms of c_{21}, c_{31} and then transformed to u^*_1, u^*_2. Consider the event E that *both* $\{b^0_{21} < c_{21} < b^0_{21} + \delta, b^0_{31} < c_{31} < b^0_{31} + \varepsilon\}$ *and* that the lines c_{21} and c_{31} intersect in a point which maximizes the first coordinate. The latter condition may be written

$$\{c_{12} \geqslant v_2/v_1, c_{13} \geqslant v_3/v_1, c_{23} \geqslant v_3/v_2, c_{32} \geqslant v_2/v_3\}, \tag{A.7}$$

where $\mathbf{v} = (v_1, v_2, v_3)$ is the point at which the lines c_{21} and c_{31} intersect, or

$$\{c_{12} \geqslant c_{21}^{-1}, c_{13} \geqslant c_{31}^{-1}, c_{23} \geqslant c_{21} c_{31}^{-1}, c_{32} \geqslant c_{21}^{-1} c_{31}\}. \tag{A.8}$$

Thus, apart from terms of second order and higher in δ and ε,

$$\begin{aligned} \Pr(E) = {} & F\{(b^0_{21}+\varepsilon)^{-1}, (b^0_{31}+\delta)^{-1}, b^0_{21}+\varepsilon, (b^0_{21}+\varepsilon)(b^0_{31}+\delta)^{-1}, \\ & b^0_{31}+\delta, (b^0_{21}+\varepsilon)^{-1}(b^0_{31}+\delta)\} \\ & - F\{(b^0_{21}+\varepsilon)^{-1}, (b^0_{31})^{-1}, b^0_{21}+\varepsilon, (b^0_{21}+\varepsilon)(b^0_{31})^{-1}, \\ & b^0_{31}, (b^0_{21}+\varepsilon)^{-1} b^0_{31}\} \\ & - F\{(b^0_{21})^{-1}, (b^0_{31}+\delta)^{-1}, b^0_{21}, b^0_{21}(b^0_{31}+\delta), b^0_{31}+\delta, \\ & (b^0_{21})^{-1}(b^0_{31}+\delta)\} \\ & + F\{(b^0_{21})^{-1}, (b^0_{31})^{-1}, b^0_{21}, b^0_{21}(b^0_{31})^{-1}, b^0_{31}, (b^0_{21})^{-1} b^0_{31}\}. \end{aligned} \tag{A.9}$$

That is, the required case (i) contribution is found in terms of c_{21}, c_{31} represented by b^0_{21}, b^0_{31} by differentiating F with respect to its third and fifth arguments and then substituting $(b^0_{21})^{-1}, (b^0_{31})^{-1}, b^0_{21}(b^0_{31})^{-1}, (b^0_{21})^{-1} b^0_{31}$ in order for the other four arguments.

Expressing the result in terms of the coordinates $\mathbf{u} = (u_1, u_2, u_3)$ at which the lines b^0_{21} and b^0_{31} intersect, one finds

$$n_2 n_3 u_1^{n_1} u_2^{n_2+1} u_3^{n_3+1}$$

which, after multiplying by

$$\partial(u_1, u_2)/\partial(b^0_{21}, b^0_{31}) = u_1 u_2^{-2} u_3^{-2}$$

gives the density contribution

$$n_2 n_3 u_1^{n_1+1} u_2^{n_2-1} u_3^{n_3-1} \tag{A.10}$$

expressed in terms of u_1, u_2 and of course $u_3 = 1-u_1-u_2$. The contributions from cases (ii) and (iii) may be found similarly to be

$$n_2 n_3 u_1^{n_1} u_2^{n_2-1} u_3^{n_3} \quad \text{and} \quad n_2 n_3 u_1^{n_1} u_2^{n_2} u_3^{n_3-1}. \tag{A.11}$$

Since

$$u_1+u_2+u_3 = 1,$$

the sum of the three parts is

$$n_2 n_3 u_1^{n_1} u_2^{n_2-1} u_3^{n_3-1},$$

or

$$\frac{n_1!\, n_2!\, n_3!}{n!} \left\{ \frac{n!}{n_1!\,(n_2-1)!\,(n_3-1)!} u_1^{n_1} u_2^{n_2-1} u_3^{n_3-1} \right\}, \tag{A.12}$$

where the first term is the probability that $\mathbf{u}^*$ is anywhere, i.e. that R is not empty, while the second is the Dirichlet density given in (23).

The density of the point with minimum first coordinate may be found by a similar argument. The analogue of (A.6) is

$$\left.\begin{array}{lll} \text{(i)} & u_2^*/u_1^* = c_{12} \quad \text{and} \quad u_3^*/u_1^* = c_{13}, & \\ \text{(ii)} & u_2^*/u_3^* = c_{32} \quad \text{and} \quad u_3^*/u_1^* = c_{13}, & \text{or} \\ \text{(iii)} & u_2^*/u_1^* = c_{12} \quad \text{and} \quad u_3^*/u_2^* = c_{23}, & \end{array}\right\} \tag{A.13}$$

and the corresponding three components of density turn out to be

$$n_1(n_1+1) u_1^{n_1-1} u_2^{n_2} u_3^{n_3}, \quad n_1 n_3 u_1^{n_1-1} u_2^{n_2} u_3^{n_3}, \quad \text{and} \quad n_1 n_2 u_1^{n_1-1} u_2^{n_2} u_3^{n_3} \tag{A.14}$$

which sum to

$$\frac{n_1!\, n_2!\, n_3!}{n!} \left\{ \frac{(n+1)!}{(n_1-1)!\, n_2!\, n_3!} u_1^{n_1-1} u_2^{n_2} u_3^{n_3} \right\}, \tag{A.15}$$

which, like (A.12) is the product of the probability that R is not empty and the Dirichlet density specified in (23).

The remaining four lines of (23) follow by symmetry. The probability that R is not empty may be obtained directly by an argument whose gist is that, for any set of n points in U, there is exactly one way to assign them to three cells of sizes n_1, n_2, n_3 corresponding to observations c_1, c_2, c_3 in such a way that R is not empty. This latter assertion will not be proved here.