NEW METHODS FOR REASONING TOWARDS POSTERIOR DISTRIBUTIONS BASED ON SAMPLE DATA[1]

By A. P. Dempster

Harvard University

0. Summary. This paper redefines the concept of sampling from a population with a given parametric form, and thus leads up to some proposed alternatives to the existing Bayesian and fiducial arguments for deriving posterior distributions. Section 2 spells out the basic assumptions of the suggested class of sampling models, and Section 3 suggests a mode of inference appropriate to the sampling models adopted. A novel property of these inferences is that they generally assign upper and lower probabilities to events concerning unknowns rather than precise probabilities as given by Bayesian or fiducial arguments. Sections 4 and 5 present details of the new arguments for binomial sampling with a continuous parameter p and for general multinomial sampling with a finite number of contemplated hypotheses. Among the concluding remarks, it is pointed out that the methods of Section 5 include as limiting cases situations with discrete or continuous observables and continuously ranging parameters.

1. Introduction. Consider an observable x, a parameter θ, and a specified family of distributious $\mathfrak{F}_\theta$ over x-space. A conventional way of thinking about sample observations $x_1, x_2, \cdots, x_n$ from an unknown member of the family of distributions $\mathfrak{F}_\theta$ is roughly as follows. First, a specific θ is determined by a process which need not be specified. Then, using this θ, the observations $x_1, x_2, \cdots, x_n$ are drawn independently at random each with the distribution $\mathfrak{F}_\theta$. I believe that this attitude is held almost universally, where the schools of Fisher and Neyman usually think rather vaguely about θ as "chosen by Nature," while the Bayesian school specifies a prior distribution governing the random choice of θ. Some Bayesians prefer to think of θ as not fixed at all while $x_1, x_2, \cdots, x_n$ are governed by their joint marginal distribution. I do not see any operational importance in this distinction, since I assume that a parameter value may be fixed and still legitimately be assigned a probability distribution, as long as the fixed value remains unknown.

The inference methods of this paper rest on a weaker definition of sample than that of the conventional model. The revised model gives up the idea that $x_1, x_2, \cdots, x_n$ are independently distributed according to $\mathfrak{F}_\theta$ for fixed θ while retaining the feature that any observed sample $x_1, x_2, \cdots, x_n$ shall appear consistent with a distribution $\mathfrak{F}_\theta$ for some θ regardless of the size n of the sample. Thus, a single observed sample can never be used to distinguish between the more relaxed model and the conventional model.

Received 30 June 1964; revised 8 October 1965.

[1] Research supported in part by the Office of Naval Research through Contract Nonr 1866(37).

A trivial example will serve here to illustrate the new approach, the general theory being defined in Section 2. Suppose that x and θ take values on the real line. Suppose that $\mathfrak{F}_\theta$ is the normal distribution $N(\theta, 1)$ with mean θ and variance unity. In contrast to the conventional approach of fixing θ and drawing $x_1, x_2, \cdots, x_n$ independently from the corresponding fixed $N(\theta, 1)$ distribution, an example of the new model is provided by asserting that $x_1 - \theta, x_2 - \theta, \cdots, x_n - \theta$ are governed by the law of n independent $N(0, 1)$ random variables, *but asserting no further laws whether deterministic or probabilistic about the variables* $x_1, x_2, \cdots, x_n, \theta$. Such an assumption no doubt appears artifical as stated here, but the discussion of Section 2 will provide a general foundation for it. The immediate purpose is to remark that, however one may think of determining θ, whether from a known process or from a black box, and whether dependent on $x_1, x_2, \cdots, x_n$ or not, the observed sample should in no way look unlike repeated drawings from some normal distribution with variance unity. In the absence of further empirical data involving repeated choices of θ, I do not see why the conventional model should be preferred over the new model.

The new model was first introduced in Dempster (1963), but with a further assumption. In the earlier paper it would have been assumed, for example, that $\theta, x_1, x_2, \cdots, x_n$ were jointly distributed random variables, i.e., that there existed a probability law simultaneously governing all of the variables $\theta, x_1, x_2, \cdots, x_n$. This joint distribution would have been specified only to the extent that $x_1 - \theta, x_2 - \theta, \cdots, x_n - \theta$ were asserted to be independently $N(0, 1)$ distributed while the conditional distribution of θ given $x_1 - \theta, x_2 - \theta, \cdots, x_n - \theta$ was not specified in any way. I now find it more satisfying to avoid extraneous complications due to assuming the existence of unknown laws. According to the present approach, it is correct to regard variables, such as parameters or yet-to-be-observed sample variables, as having existing but unknown real-world values. But it is seen as intellectually wasteful and possibly deceptive to assume the existence of probability laws governing such variables, unless these laws may be specified. This change has in turn suggested the more satisfying methods of defining posterior probabilities given in this paper.

An underlying motivation for this work is to be found in the need to break the serious deadlock between those statisticians who prefer Bayesian formulations and those who prefer formulations relying on the repeated sampling aspects of probability laws. These two traditions have a longer history of conflict than is generally realized. Todhunter, writing *circa* 1865, traced what would now be called a confidence or fiducial argument about binomial p to J. Bernoulli *circa* 1700. In correspondence, Leibniz questioned Bernoulli's method. Of more interest here is the fact that Laplace *circa* 1813 used both the Bernoullian and Bayesian approaches to estimate p and presented slightly discrepant normal approximations without comment. Poisson in 1830 also used both methods but achieved normal approximations which were in agreement. De Morgan in 1837 drew attention to the differences in logical processes used and queried Poisson's results. Todhunter himself believed Poisson to have been correct. Unfortunately, the question of the differences in Bernoullian and Bayesian approaches was confounded with the question of accuracy of normal approximations and was destined to remain obscure for around 100 years. See Todhunter (1865) pp. 57, 73, 554–558, for discussion and references.

At present, the Bayesian school is showing renewed vigor and is increasingly in conflict with what I have called above the Bernoullian school. Within the latter school there are disagreements between the many who generally follow Neyman and the few who prefer R. A. Fisher. The following two statements summarize a previously given (Dempster (1964)) attitude to the Neyman-Fisher differences: (i) Neyman's methods while often available and useful are not fully satisfying, and (ii) Fisher, while extraordinarily inventive and mostly on the right track, was unable to give coherence to his system and in particular failed to perfect his fiducial argument.

I believe that the methods of this paper are close to Fisher's viewpoint. The arguments given here resemble the fiducial argument in that they produce posterior probabilities using the sampling hypothesis and parametric hypotheses but no prior distribution. I believe also that the basic reasoning principle described in Section 3 is essentially what Fisher relied on in his fiducial argument.

At the same time the new methods of this paper can be viewed as belonging under a common umbrella with the Bayesian methods. This umbrella is described in a later paper (Dempster (1965)). There the logic underlying upper and lower probability systems is given more generally. Rules are given for combining independent sources of information. The methods of this paper implicitly apply these rules to the combination of information from individual sample observations. If a prior distribution is available, it may be combined with the sample information *according to the same rules*, and the result is the standard Bayesian answer (Dempster (1965)).

2. Construction of the sampling model. Throughout the following discussion measure-theoretic details are not supplied, mostly because they are obvious in the range of examples of present interest.

The basic components of the theory are a pair of spaces $\mathfrak{A}$ and $\mathfrak{X}$. $\mathfrak{A}$ represents the *population* being sampled, and each *population individual* $a \varepsilon \mathfrak{A}$ has a corresponding *observable characteristic* $x \varepsilon \mathfrak{X}$. The mapping $a \rightarrow x$ thus assumed to exist is regarded as unknown but subject to certain restrictions posed below. The statement that a population individual a comes under observation as part of a sample is construed to mean that the x corresponding to a becomes known to the observer. The observer is not allowed, however, to identify a.

A unique probability measure μ over $\mathfrak{A}$ is assumed given. This plays the role of the law governing the random sampling operation. A *finite* population of size N is represented by a set of N elements, and the natural measure μ governing random sampling is the measure assigning probability $1/N$ to each of the N elements. The reader may supply the obvious definitions of a random sample $a_1, a_2, \cdots, a_n$ from $\mathfrak{A}$, sampling either with replacement or without replacement as desired. When an infinite population is postulated, an appropriate choice of

$\mathcal{A}$ and μ is less clear, and, to the extent that various choices may be transformed into one another, the choice is more or less arbitrary. A convenient representation for the infinite population structures used in this paper takes $\mathcal{A}$ to be a simplex and μ to be the uniform distribution over the simplex. A random sample $a_1, a_2, \cdots, a_n$ from an infinite population $\mathcal{A}$ is defined, as one would expect, to be a drawing from the product measure μ^n over the product space $\mathcal{A}^n$.

Besides $\mathcal{A}$, $\mathcal{X}$ and μ, the user of the theory must specify in each instance (i) a class of contemplated mappings $a \rightarrow x$, and (ii) a family of probability measures over $\mathcal{X}$ whose typical member may be denoted by $\mathcal{F}_\theta$ where θ ranges over a space Θ. The family of measures $\mathcal{F}_\theta$ is used in the theory to define two postulates restricting the class of contemplated mappings $a \rightarrow x$, namely

(P1) the probability measure over $\mathcal{X}$ induced by the measure μ over $\mathcal{A}$ under any contemplated mapping $a \rightarrow x$ must be $\mathcal{F}_\theta$ for some $\theta \varepsilon \Theta$, and

(P2) exactly one mapping in the class of contemplated mappings $a \rightarrow x$ leads to the induced measure $\mathcal{F}_\theta$ over $\mathcal{X}$ for each $\theta \varepsilon \Theta$.

(P1) and (P2) together imply a one-one correspondence between the class of contemplated mappings $a \rightarrow x$ and the family of measures $\mathcal{F}_\theta$. The two postulates are kept separate in the exposition because (P1) is easier to swallow than (P2). A discussion of (P2) will be given shortly.

In any application of the theory, a random sample $a_1, a_2, \cdots, a_n$ is drawn from $\mathcal{A}$ as specified above. The observer identifies the corresponding $x_1, x_2, \cdots, x_n$ under the true mapping $a \rightarrow x$. He is then asked to draw inferences concerning which member of the class of contemplated mappings is the true member or, equivalently, concerning which θ in Θ is the true θ. The suggested mode of inference is given in Section 3.

The $N(\theta, 1)$ example of Section 1 may be used as a first illustration of the theory. Take $\mathcal{A}$ to be the whole real line and take μ to be the $N(0, 1)$ distribution over $\mathcal{A}$. Take $\mathcal{X}$ to be the whole real line, and take $\mathcal{F}_\theta$ to be the $N(\theta, 1)$ distribution, where the range space Θ of θ is also the whole real line. Finally, define the class of contemplated mappings $a \rightarrow x$ to be

$$a \rightarrow x = \theta + a, \tag{2.1}$$

where the dual interpretation of θ as a parameter for the class of mappings and as a parameter for the class of distributions $\mathcal{F}_\theta$ defines the one-one correspondence satisfying (P1) and (P2). The essential feature of this illustration is the preservation of the natural orderings on $\mathcal{A}$ and $\mathcal{X}$ under the whole class of mappings from $\mathcal{A}$ to $\mathcal{X}$. The particular representation of $\mathcal{A}$ and μ is not essential, and any monotone one-one transformation, for example carrying μ on $\mathcal{A}$ into a uniform distribution on (0, 1), could be used to obtain an alternative representation. This example will be termed a *structure of the first kind* in the later discussion of this section. Note that, as remarked in Section 1, the only probability law operating is the law of n independent $N(0, 1)$ random variables applied to $a_1, a_2, \cdots, a_n$.

The sampling model proposed above differs from the conventional formulation of mathematical statistics in that the population being sampled is explicitly represented by a mathematical space, namely the space $\mathcal{A}$ of population individuals. The presence of this space makes it possible to ask certain questions within the framework of the model which were only dimly conceivable under the old formulation. Specifically, the old formulation provided a mathematical representation of a population distribution such as $\mathfrak{F}_\theta$ for an observable characteristic, but it did not describe how each population individual contributed to the overall distribution. In real life, however, it is legitimate to ask at least what each individual's x might be under a contemplated hypothesis $\mathfrak{F}_\theta$. In other words, what mapping or mappings $a \rightarrow x$ should be regarded as permissible for a given θ within the limits specified by (P1)?

One answer to this question is to allow *any* set of mappings consistent with (P1). This is tantamount to refusing to be interested in the question. Postulate (P2) goes to the other end of the spectrum and requires that *only one* mapping shall be allowed for each given θ. An underlying motivation for this directive is the general principle that parsimony is a good thing in model-building. Of course, (P2) goes only part way to answering the question, since it does not say which mapping $a \rightarrow x$ shall be the only one allowed for a given θ. Two classes of specific answers, hence specific instances of the theory, will shortly be given. (P2) itself provides a guideline, adopted in a speculative spirit by this investigation in order to examine the statistical methodology which follows naturally from it.

Another consequence of explicitly introducing the population space $\mathcal{A}$ is the insertion of the random sampling hypothesis into the model where it naturally belongs. In the conventional formulation, a distinct law based on independent and identically distributed random variables is assumed to govern $x_1, x_2, \cdots, x_n$ for each distinct θ. In the present formulation, the collection of distinct laws is replaced by a single law μ^n which is overtly meant to describe the operation of sampling from $\mathcal{A}$. Note especially that in the new approach the $\mathfrak{F}_\theta$ are not regarded as probability laws in the ordinary sense, i.e., a random variable x governed by the law $\mathfrak{F}_\theta$ is nowhere postulated. The $\mathfrak{F}_\theta$ play the roles not of sampling distributions but rather of deterministic laws describing the contemplated population distributions of x.

The remainder of this section describes two classes of completely specified sampling models of the proposed kind. These will be called the class of *structures of the first kind* and the class of *structures of the second kind*. The first class, which has been illustrated above, assumes $\mathcal{A}$ and $\mathfrak{X}$ to be ordered. Unfortunately such an ordering of $\mathfrak{X}$ restricts consideration essentially to a univariate characteristic. The second class is designed to remove this restriction so that either multivariate or univariate x may be handled. To keep the discussion simple, $\mathfrak{X}$ will be assumed finite of size k where $k \geqq 2$. In other words the observable characteristic is multinomial, assuming values in one of k *categories* which constitute $\mathfrak{X}$. In this multinomial context the use of a structure of the first kind presupposes that the k categories possess a natural order, while the use of a structure of the second kind poses no such restriction and treats all k categories symmetrically.

Motivation and definition will now be given for the class of structures of the

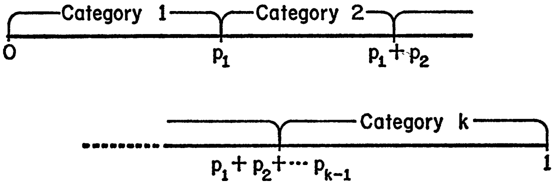

FIG. 1. The interval (0, 1) of population individuals and their corresponding multinomial categories for a given $(p_1, p_2, \cdots, p_k)$ in a structure of the first kind.

first kind. When the observable characteristic is assumed to classify the population individuals into k ordered categories, it is not implausible to suppose that the population individuals possess an ordering consistent with the partial ordering induced by the mapping $a \rightarrow x$, with the same basic ordering of $\mathcal{A}$ holding whatever mapping $a \rightarrow x$ is contemplated. It is then but a short step to suppose that the population individuals are distributed over a real line and a further short step to regard this distribution as being monotonely transformable and thence transformed into a uniform distribution over the interval (0, 1). Such a uniform distribution over (0, 1) induces a given $\mathcal{F}_\theta$ over $\mathcal{X}$ under a mapping $a \rightarrow x$ such that a on the intervals $(0, p_1)$, $(p_1, p_1 + p_2)$, $\cdots$, $(p_1 + p_2 + \cdots + p_{k-1}, 1)$ map respectively into categories $1, 2, \cdots, k$ of $\mathcal{X}$, where p_i defines the probability of category i under $\mathcal{F}_\theta$ for $i = 1, 2, \cdots, k$. This mapping is illustrated in Figure 1. Except for its indeterminacy at a finite set of points of $\mathcal{A}$, this is the only mapping $a \rightarrow x$ which satisfies (P1) for a given $\mathcal{F}_\theta$ and which preserves the ordering on $\mathcal{A}$ and $\mathcal{X}$. Any resolution of the indeterminacy for each $\mathcal{F}_\theta$ yields a class of contemplated mappings in the desired one-one correspondence with the class of all distributions over $\mathcal{X}$. To complete the definition of a structure of the first kind it remains only to specify a family of distributions $\mathcal{F}_\theta$, and this may be done arbitrarily.

Consider now the class of structures of the second kind. Here, the k multinomial categories are to be treated without regard to order. A natural means to this end is to increase the dimension of the proposed $\mathcal{A}$ so it may have the capability to reflect a multivariate observable characteristic. The following simple scheme is proposed: Suppose that $\mathcal{A}$ consists of the points of a $(k - 1)$-dimensional simplex. Using barycentric coordinates, the general point of such a simplex may be represented by a k-tuple of real numbers $(\alpha_1, \alpha_2, \cdots, \alpha_k)$ where

$$(2.2) \qquad \alpha_j \geqq 0 \quad \text{for} \quad j = 1, 2, \cdots, k, \quad \text{and} \quad \textstyle\sum_{j=1}^k \alpha_j = 1.$$

The vertices $I_1, I_2, \cdots, I_k$ of the simplex are represented by the k-tuples $(1, 0, \cdots, 0), (0, 1, \cdots, 0), \cdots, (0, 0, \cdots, 1)$. Suppose that μ is defined to be the uniform probability measure over the simplex $\mathcal{A}$. Specifying a mapping $a \rightarrow x$ is equivalent to specifying a partition of $\mathcal{A}$ into $\pi_1, \pi_2, \cdots, \pi_k$ where $a \varepsilon \pi_i$ maps into category $i \varepsilon \mathcal{X}$, for $i = 1, 2, \cdots, k$. The mapping $a \rightarrow x$ corresponding to a given $\mathcal{F}_\theta$ under (P1) and (P2) must have an associated partition satisfying

$$(2.3) \qquad \mu(\pi_i) = p_i,$$

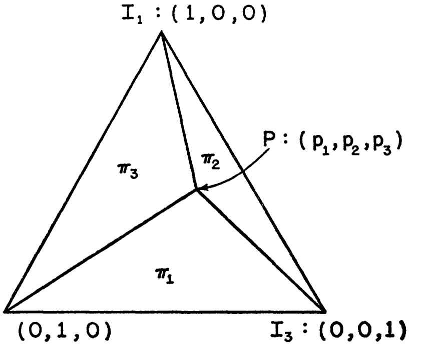

FIG. 2. The triangle of population individuals associated with a structure of the second kind when $k = 3$.

where p_i is the probability of category i under $\mathfrak{F}_\theta$, for $i = 1, 2, \cdots, k$. Such a partition is defined by considering the point P in $\mathfrak{a}$ with coordinates $(p_1, p_2, \cdots, p_k)$ and defining π_i for $i = 1, 2, \cdots, k$ to be the simplex with vertices P and I_j for $1 \leqq j \leqq k, j \neq i$. (Points on the common boundaries of the π_i may be arbitrarily assigned.) A set of mappings of this type, in one-one correspondence with a specified family of distributions $\mathfrak{F}_\theta$, will be said to define a structure of the second kind.

The case $k = 3$ is illustrated in Figure 2.

There are many other structures satisfying postulates (P1) and (P2). The two special classes of structures proposed above were selected because of their mathematical simplicity. I have been unable to find any others with comparably clean properties. The idea behind the class of structures of the first kind, namely the idea of monotonely transforming the distribution of an observable into a uniform distribution on (0, 1), is a familiar one in statistical theory, and some of the resulting inferences resemble those coming from confidence and fiducial arguments. The idea behind the class of structures of the second kind is unfamiliar, but, I think, not drastically different from the idea behind the first class and worth developing so that its potential may be understood.

3. Inference methods for the proposed sampling model. The first task here is to define inferences about an unknown parameter θ, given an observed sample $x_1, x_2, \cdots, x_n$, when a model of the type defined in Section 2 is assumed. Later in the section the discussion will be broadened to include inferences made jointly about θ and a future sample $y_1, y_2, \cdots, y_m$ from the same population.

As conceived here, the aim of inference is to assign a probability distribution to θ. Any probability deduced from such a distribution is intended for interpretation in the usual prospective way as long as θ remains unknown. For example, if the statement $\Pr(\theta > 5.1) = .035$ should be made about a real parameter θ, this statement would be intended to convey the same type of information as the statement that the probability is .035 of drawing a white ball from an urn containing 35 white balls and 965 black balls.

It turns out that the reasoning developed here leads in general not to precise probability statements but to bounded probability statements about any event determined by θ and $y_1, y_2, \cdots, y_m$. For example, in place of a statement such as Pr $(\theta > 5.1) = .035$, a statement such as $.010 \leqq \Pr(\theta > 5.1) \leqq .063$ might be found. The aim of inference and the interpretation of probability remains as before. The difference is simply that the logical apparatus carried by the statistician is able to produce only bounds for the desired posterior probabilities.

The central idea follows. Throughout this section an infinite population is assumed, so that the sample is represented by a point drawn at random from the space $\mathfrak{A}^n$ according to the measure μ^n. That is, *before the sample is drawn*, prospective probability judgments concerning which sample $a_1, a_2, \cdots, a_n$ will appear are governed by the measure μ^n over $\mathfrak{A}^n$. *After the sample is drawn*, this law is generally not appropriate for prospective probability judgments because the observations $x_1, x_2, \cdots, x_n$ typically rule out many of the points of $\mathfrak{A}^n$ as possible samples. It is proposed here to consider the subspace of $\mathfrak{A}^n$ which does represent the range of samples still possible after $x_1, x_2, \cdots, x_n$ become known, to restrict the measure μ^n to this subspace, and to use the restricted measure for prospective probability judgments after $x_1, x_2, \cdots, x_n$ are known.

Accordingly, define R_n to be the subspace of $\mathfrak{A}^n$ consisting of points $a_1, a_2, \cdots, a_n$ such that

$$a_1 \rightarrow x_1, a_2 \rightarrow x_2, \cdots, a_n \rightarrow x_n \tag{3.1}$$

under some mapping $a \rightarrow x$ in the class of contemplated mappings, i.e., R_n consists of the set of samples which could have produced the observed data $x_1, x_2, \cdots, x_n$. Define the measure ν_n over R_n from

$$\nu_n(A) = \mu^n(A)/\mu^n(R_n) \tag{3.2}$$

for $A \subset R_n$. This is just the familiar device of conditioning by R_n. The restricted measure ν_n over R_n is regarded here as appropriate for prospective probability judgments about $a_1, a_2, \cdots, a_n$ after $x_1, x_2, \cdots, x_n$ are known.

It is assumed in (3.2) that $\mu^n(R_n) > 0$. This assumption is essentially met by the structures of the first and second kinds as defined in Section 2 when $\mathfrak{X}$ is finite. For these structures, either $\mu^n(R_n) > 0$ or an observation x_i has fallen in a category of $\mathfrak{X}$ assigned zero measure by all $\mathfrak{F}_\theta$, and the latter possibility means that the data contradict the model with certainty. The extension of the theory to cover continuous observables is touched on in Section 6.

A sample $a_1, a_2, \cdots, a_n$ will be called *consistent with the data* $x_1, x_2, \cdots, x_n$ *and with* θ in Θ if (3.1) holds for the mapping $a \rightarrow x$ corresponding to θ. After the data are fixed, this consistency concept defines a mapping from R_n to Θ. If the mapping should be one-one, then the measure ν_n over R_n induces a measure over Θ which may be used for prospective probability judgments about the unknown θ. In general, however, this mapping from R_n to Θ is one-many, with the consequence that ν_n induces a system of upper and lower probability judgments about θ rather than a single measure.

This system of upper and lower probabilities is defined as follows. Given any event Σ determined by θ, i.e., any subset Σ of Θ belonging to an appropriate class of subsets, define $\bar{R}_n(\Sigma)$ to be the set of points of R_n which are consistent with the data for at least one θ in Θ, and define $\underline{R}_n(\Sigma)$ to be the set of points of R_n which are consistent with the data for no θ not in Σ. Thence define *the upper probability* $\bar{P}(\Sigma)$ *of* Σ and *the lower probability* $\underline{P}(\Sigma)$ *of* Σ to be

$$\bar{P}(\Sigma) = \nu_n(\bar{R}_n(\Sigma)) \quad \text{and} \quad \underline{P}(\Sigma) = \nu_n(\underline{R}_n(\Sigma)). \tag{3.3}$$

The rationale behind the definitions (3.3) is that $\bar{P}(\Sigma)$ includes "as much" of the measure ν_n as can be transferred from R_n to Θ under the various one-one mappings consistent with the one-many consistency mapping from R_n to Θ prescribed above. Similarly, $\underline{P}(\Sigma)$ includes "as little" of the measure as can be transferred under the same circumstances. Thus, prospective probability judgments based on ν_n transfer naturally into a system of upper and lower probability judgments applied to events $\Sigma \subset \Theta$.

The calculus of these upper and lower probability judgments is developed more fully in a later paper (Dempster (1965)), but a few obvious properties are included here.

Since $R_n \supset \bar{R}_n(\Sigma) \supset \underline{R}_n(\Sigma)$ it follows that

$$0 \leqq \underline{P}(\Sigma) \leqq \bar{P}(\Sigma) \leqq 1. \tag{3.4}$$

Also it is easily checked that $\bar{R}_n(\Sigma)$ and $\underline{R}_n(\Theta - \Sigma)$ form a disjoint pair with union R_n so that

$$\bar{P}(\Sigma) = 1 - \underline{P}(\Theta - \Sigma). \tag{3.5}$$

Finally, since $R_n = \bar{R}_n(\Theta) = \underline{R}_n(\Theta)$, it follows that

$$\bar{P}(\Theta) = \underline{P}(\Theta) = 1. \tag{3.6}$$

For any real parameter ϕ determined by θ, upper and lower cumulative distribution functions may be defined as

$$\bar{H}(Z) = \bar{P}(\phi \leqq Z), \quad \text{and} \quad \underline{H}(Z) = \underline{P}(\phi \leqq Z). \tag{3.7}$$

Corresponding upper and lower expectations of ϕ may then be defined as

$$\bar{E}(\phi) = \textstyle\int_{-\infty}^{\infty} Z \, d\underline{H}(Z) \quad \text{and} \quad \underline{E}(\phi) = \int_{-\infty}^{\infty} Z \, d\bar{H}(Z). \tag{3.8}$$

The behavior of these operators under linear transformations is governed by

$$\begin{aligned} \bar{E}(a + b\phi) &= a + b\bar{E}(\phi), \qquad \text{if} \quad b > 0, \\ &= a + b\underline{E}(\phi), \qquad \text{if} \quad b < 0, \end{aligned} \tag{3.9}$$

where a and b are real constants. The expectations (3.8) are suggested as guides for betting or decision procedures whose loss functions are linear in ϕ.

Inferences about further sample observations $y_1, y_2, \cdots, y_m$ may be defined using ideas very similar to those above. The observed sample $a_1, a_2, \cdots, a_n$ and a future sample $b_1, b_2, \cdots, b_m$ are governed prior to any sampling by the law μ^{n+m} over $\mathcal{A}^{n+m}$.

The observations are $x_1, x_2, \cdots, x_n$ as before, but the unknowns are now $\theta, y_1, y_2, \cdots, y_m$ in the space $\Theta \times \mathcal{X}^m$. The space $R_{n,m}$ of samples possible after observation consists of those $a_1, a_2, \cdots, a_n, b_1, b_2, \cdots, b_m$ satisfying

$$(3.10) \qquad a_1 \to x_1, a_2 \to x_2, \cdots, a_n \to x_n,$$
$$b_1 \to y_1, b_2 \to y_2, \cdots, b_m \to y_m$$

for some mapping $a \to x$ in the class of contemplated mappings and for some $\theta, y_1, y_2, \cdots, y_m$. The initial measure μ^{n+m} over $\mathcal{A}^{n+m}$ leads to a measure $\nu_{n,m}$ appropriate for postsample judgments. Given any event Σ^* determined by $\theta, y_1, y_2, \cdots, y_m$ the subsets $\bar{R}_{n,m}(\Sigma^*)$ and $\underline{R}_{n,m}(\Sigma^*)$ of $R_{n,m}$ are defined analogously to $\bar{R}_n(\Sigma)$ and $\underline{R}_n(\Sigma)$ above, i.e., $\bar{R}_{n,m}(\Sigma^*)$ is the set of points in $R_{n,m}$ which could have given rise to $x_1, x_2, \cdots, x_n$ for some $\theta, y_1, y_2, \cdots, y_m$ in Σ^* and $\underline{R}_{n,m}(\Sigma^*)$ is the set of points in $R_{n,m}$ which could have given rise to $x_1, x_2, \cdots, x_n$ for no $\theta, y_1, y_2, \cdots, y_m$ not in Σ^*. Thence

$$(3.11) \qquad \bar{P}(\Sigma^*) = \nu_{n,m}(\bar{R}_{n,m}(\Sigma^*)) \quad \text{and} \quad \underline{P}(\Sigma^*) = \nu_{n,m}(\underline{R}_{n,m}(\Sigma^*)).$$

Any event determined by θ alone has upper and lower probabilities derivable by (3.11) or by (3.3). It is clear, however, that the two sets of inferences concur, as would be desired.

The following two sections are intended to illustrate the foregoing definitions in a pair of non-trivial situations. Section 4 deals with finite $\mathcal{X}$ of size $k = 2$ (binomial sampling), the family $\mathcal{F}_\theta$ consisting of all possible distributions over the two categories of $\mathcal{X}$. A structure of the first kind is assumed, but this is also trivially a structure of the second kind when $k = 2$. The illustration of Section 5 assumes a structure of the second kind with general k but finite Θ.

4. Binomial sampling. Illustrative inferences are worked out here for the structure of the first kind defined by setting $k = 2$ and allowing $\mathcal{F}_\theta$ to range over all distributions on the two categories of $\mathcal{X}$. As is usually done with binomial sampling, the parameter p on $0 \leqq p \leqq 1$ will be used for the distributions over $\mathcal{X}$, where p denotes the probability of category 1 and $1 - p$ denotes the probability of category 2. The population individuals are supposed uniformly distributed on the interval (0, 1) under this structure of the first kind. (The corresponding structure of the second kind would differ only in the nonessential way that the population individuals would be uniformly distributed over the line segment (one-dimensional simplex) joining the points with coordinates (0, 1) and (1, 0).) The mapping $a \to x$ corresponding to a given p is ambiguous at $a = p$. This ambiguity does not affect the resulting inference, but for definiteness $a = p$ will be assumed to map into category 1.

The population individuals $a_1, a_2, \cdots, a_n, b_1, b_2, \cdots b_m$ representing the observed sample of size n and a future sample of size m are supposed drawn at random according to a uniform distribution over $\mathcal{A}^{n+m}$ which is here a unit cube in

$n + m$ dimensions. The sample data $x_1, x_2, \cdots, x_n$ marks each individual of the observed sample as belonging to category 1 or category 2. The observation vector $x_1, x_2, \cdots, x_n$ will be replaced here by the single quantity T defined to be the total number of sample observations in category 1. To assume that only T is observed, rather than the actual configuration $x_1, x_2, \cdots, x_n$, has no effect on the resulting inferences because the spaces R_n and $R_{n,m}$ corresponding to each of the $\binom{n}{T}$ configurations with given T are disjoint and isomorphic. Consequently, the only effect on (3.2) is to multiply both numerator and denominator of the right side by $\binom{n}{T}$. It is a theorem, not proved here, that the inferences based on multinomial data, represented by either a structure of the first kind or a structure of the second kind, are not affected if the individual sample observations $x_1, x_2, \cdots, x_n$ are thrown away and only $T_1, T_2, \cdots, T_k$ retained, where T_i denotes the number of sample observations in category i.

Upper and lower probabilities will be computed for the events

$$\Sigma = \{\alpha \leqq p \leqq \beta\} \tag{4.1}$$

and

$$\Sigma^* = \{r \leqq S \leqq t\}, \tag{4.2}$$

where S is the number of category 1 observations in a future sample of size m. These upper and lower probabilities depend of course on the observed T.

Consider first (4.1). A point $a_1, a_2, \cdots, a_n$ in $\mathfrak{A}^n$ is consistent with the observed T and the parameter value p if and only if

$$a_{(T)} \leqq p < a_{(T+1)} \tag{4.3}$$

where $a_{(1)} \leqq a_{(2)} \leqq \cdots \leqq a_{(n)}$ denote the ordered random variables a_1, $a_2, \cdots, a_n$ and where $a_{(0)} = 0$ and $a_{(n+1)} = 1$. It follows that $\bar{R}_n(\Sigma)$ is the subset of $\mathfrak{A}^n$ such that the intersection of the intervals $[a_{(T)}, a_{(T+1)})$ and $[\alpha, \beta]$ is nonempty. R_n is the special case of $\bar{R}_n(\Sigma)$ when $\alpha = 0$ and $\beta = 1$ so that $R_n = \mathfrak{A}^n$.

Since the measures ν_n and μ^n coincide here, the definition (3.3), reduces to

$$\bar{P}(\Sigma) = \mu^n(\bar{R}_n(\Sigma)). \tag{4.4}$$

To calculate this it is convenient to write

$$\bar{R}_n(\Sigma) = \{\alpha < a_{(T)} \leqq \beta\} \cup \{a_{(T)} \leqq \alpha < a_{(T+1)}\} \tag{4.5}$$

which is a union of disjoint sets. Thus

$$\begin{aligned} \bar{P}(\Sigma) &= T\tbinom{n}{T} \textstyle\int_\alpha^\beta p^{T-1}(1-p)^{n-T}\,dp \\ &\quad + \tbinom{n}{T}\alpha^T(1-\alpha)^{n-T}, && \text{if } 1 \leqq T \leqq n, \\ &= (1-\alpha)^n, && \text{if } T = 0. \end{aligned} \tag{4.6}$$

An alternative to (4.5) is

$$\bar{R}_n(\Sigma) = \{\alpha < a_{(T+1)} \leqq \beta\} \cup \{a_{(T)} \leqq \beta < a_{(T+1)}\} \tag{4.7}$$

which leads to an alternative to (4.6), namely

$$
\begin{aligned}
\bar{P}(\Sigma) &= (n-T)\tbinom{n}{T}\textstyle\int_\alpha^\beta p^T(1-p)^{n-T-1}\,dp \\
&\qquad + \tbinom{n}{T}\beta^T(1-\beta)^{n-T}, && \text{if } 0 \leqq T \leqq n-1, \\
&= \beta^n, && \text{if } T = n.
\end{aligned} \tag{4.8}
$$

Another alternative may be found by replacing the beta integrals in (4.6) or (4.8) by binomial sums, yielding

$$
\bar{P}(\Sigma) = \textstyle\sum_{i=0}^{T}\tbinom{n}{i}\alpha^i(1-\alpha)^{n-i} + \sum_{i=T}^{n}\tbinom{n}{i}\beta^i(1-\beta)^{n-i} - 1. \tag{4.9}
$$

A similar variety of forms is possible for $\underline{P}(\Sigma)$. $\underline{R}_n(\Sigma)$ is the event that the interval $[a_{(T)}, a_{(T+1)})$ is contained in the interval $[\alpha, \beta]$. Writing

$$
\underline{R}_n(\Sigma) = \{\alpha \leqq a_{(T)} \leqq \beta\} - \{\alpha \leqq a_{(T)} \leqq \beta, a_{(T+1)} > \beta\} \tag{4.10}
$$

and

$$
\begin{aligned}
&\{\alpha \leqq a_{(T)} \leqq \beta, a_{(T+1)} > \beta\} \\
&\qquad = \{a_{(T)} \leqq \beta < a_{(T+1)}\} - \{a_{(T)} < \alpha, a_{(T+1)} > \beta\},
\end{aligned} \tag{4.11}
$$

it is seen that

$$
\underline{P}(\Sigma) = T\tbinom{n}{T}\textstyle\int_\alpha^\beta p^{T-1}(1-p)^{n-T}\,dp - \tbinom{n}{T}[\beta^T - \alpha^T](1-\beta)^{n-T}, \tag{4.12}
$$

at least if $1 \leqq T \leqq n-1$. Special interpretations are needed if $T = 0$ and $T = n$, and these may be handled by checking directly that

$$
\begin{aligned}
\underline{P}(\Sigma) &= 0, && \text{if } T = 0, \ \alpha > 0, \\
&= 1-(1-\beta)^n, && \text{if } T = 0, \ \alpha = 0, \\
&= 0, && \text{if } T = n, \ \beta < 1, \\
&= 1-\alpha^n, && \text{if } T = n, \ \beta = 1,
\end{aligned} \tag{4.13}
$$

Just as (4.6) may be replaced by (4.8) and (4.9), (4.12) may be replaced by

$$
\begin{aligned}
\underline{P}(\Sigma) &= (n-T)\tbinom{n}{T}\textstyle\int_\alpha^\beta p^T(1-p)^{n-T-1}\,dp \\
&\qquad - \tbinom{n}{T}\alpha^T[(1-\alpha)^{n-T} - (1-\beta)^{n-T}],
\end{aligned} \tag{4.14}
$$

or, replacing the integrals by sums,

$$
\begin{aligned}
\underline{P}(\Sigma) &= \textstyle\sum_{i=0}^{T-1}\tbinom{n}{i}\alpha^i(1-\alpha)^{n-i} \\
&\qquad + \tbinom{n}{T}\alpha^T(1-\beta)^{n-T} + \textstyle\sum_{i=T+1}^{n}\tbinom{n}{T}\beta^i(1-\beta)^{n-i} - 1.
\end{aligned} \tag{4.15}
$$

Note that

$$
\bar{P}(\Sigma) - \underline{P}(\Sigma) = \tbinom{n}{T}[\alpha^T(1-\alpha)^{n-T} + \beta^T(1-\beta)^{n-T} - \alpha^T(1-\beta)^{n-T}]. \tag{4.16}
$$

Also, in the special case $\alpha = \beta$,

$$
\bar{P}(p = \alpha) = \tbinom{n}{T}\alpha^T(1-\alpha)^{n-T} \quad \text{and} \quad \underline{P}(p = \alpha) = 0. \tag{4.17}
$$

Several general features of the above inferences are worthy of remark. The small upper probability (4.17) assigned to any particular value $p = \alpha$ is proportional to the conventional likelihood at $p = \alpha$. This likelihood is the probability content of the region $\bar{R}_n(p = \alpha)$ in $\mathcal{A}^n$, and such regions sweep out the region R_n as α ranges over $0 \leqq \alpha \leqq 1$. If the regions $\bar{R}_n(p = \alpha)$ had not overlapped for different α, then all upper and lower probabilities would have coincided and both would have been derivable from a posterior density proportional to likelihood. It will next be shown that the overlapping decreases as n increases in the sense that the upper and lower probabilities tend towards agreement with a distribution whose density is proportional to likelihood.

Large sample behavior may be studied by supposing that $n \longrightarrow \infty$ and $T \longrightarrow \infty$ in such a way that $T/n \longrightarrow \rho$. By considering the limiting normal behavior of binomial distributions, it becomes clear that

$$(4.18) \qquad \bar{P}(\Sigma) - \underline{P}(\Sigma) = O(1/n^{\frac{1}{2}})$$

uniformly in α and β, and consequently that either $\bar{P}(\Sigma)$ or $\underline{P}(\Sigma)$ may be approximated by

$$(4.19) \qquad \bar{P}(\Sigma) \sim \underline{P}(\Sigma) \sim \Phi(\beta^*) - \Phi(\alpha^*)$$

where Φ denotes the cdf of the $N(0, 1)$ distribution,

$$(4.20) \qquad \beta^* = (\beta - T/n)/[n(T/n)(1 - T/n)]^{\frac{1}{2}},$$

and

$$(4.21) \qquad \alpha^* = (\alpha - T/n)/[n(T/n)(1 - T/n)]^{\frac{1}{2}}.$$

(The symbols $\sim$ in (4.19) mean that the ratios tend to unity as $n \longrightarrow \infty$.) The normal approximation (4.19) extends to show that, if the arguments α^* and β^* tend to constants as $n \longrightarrow \infty$, the posterior inferences may be computed from a normal density function whose ratio to the likelihood tends to a constant. The limiting posterior inference considered here is also that reached by a Bayesian argument with any well-behaved prior density and the same limiting conditions. That is, in a circumstance where the Bayesian would say that the choice of a prior distribution does not matter, the present theory yields the same answer.

Consider next how to find $\bar{P}(\Sigma^*)$ and $\underline{P}(\Sigma^*)$ where Σ^* was defined in (4.2). This requires consideration of the sample space $\mathcal{A}^{n+m}$ from which the pair of samples is drawn. For Σ^* to hold it is necessary and sufficient that

$$(4.22) \qquad b_{(r)} \leqq p < b_{(t+1)}$$

where $b_{(1)} \leqq b_{(2)} \leqq \cdots \leqq b_{(m)}$ denote the ordered random variables $b_1, b_2, \cdots, b_m$ with the additional conventions that $b_{(0)} = 0$ and $b_{(m+1)} = 1$. On the other hand, (4.3) must hold if p is to be consistent with the observation T. Thus $\bar{R}_{n,m}(\Sigma^*)$ is the subset of $\mathcal{A}^{n+m}$ such that the intervals $[b_{(r)}, b_{(t+1)})$ and $[a_{(T)}, a_{(T+1)})$ are not disjoint. Writing

$$(4.23) \qquad \bar{R}_{n,m}(\Sigma^*) = \{b_{(r)} \leqq a_{(T)} < b_{(t+1)}\} \cup \{a_{(T)} < b_{(r)} < a_{(T+1)}\},$$

it is seen that finding $\bar{P}(\Sigma^*)$ is reducible to a combinatorial problem concerning the $\binom{n+m}{n}$ equally likely relative orderings of the samples $a_1, a_2, \cdots, a_n$ and $b_1, b_2, \cdots, b_m$.

For example, the event $\{a_{(T)} \leqq b_{(r)} < a_{(T+1)}\}$ may be expressed as the event that $b_{(r)}$ has rank $r + T$ in the combined samples. Under this event, the first $r + T - 1$ members of the combined sample consist of T of the a_i and $r - 1$ of the b_j, and the last $m + n - r - T$ members of the combined sample consist of $n - T$ of the a_i and $m - r$ of the b_j. Thus

$$(4.24) \qquad \mu^{n+m}(a_{(T)} \leqq b_{(r)} < a_{(T+1)}) = \binom{r+T-1}{T}\binom{m+n-r-T}{n-T}\Big/\binom{m+n}{n}.$$

This and subsequent formulas apply generally when $0 \leqq r \leqq t \leqq m$ *and* $0 \leqq T \leqq n$ *provided that* $\binom{x}{y}$ *is regarded as zero when* $x < y$.

By reasoning similar to that producing (4.24), one finds from (4.23) that

$$(4.25) \quad \bar{P}(\Sigma^*) = \sum_{i=r}^{t} \binom{i+T-1}{i}\binom{m+n-i-T}{m-i}\Big/\binom{m+n}{n} + \binom{r+T-1}{T}\binom{m+n-r-T}{n-T}\Big/\binom{m+n}{n}.$$

Similarly $\underline{R}_{n,m}(\Sigma^*)$ may be expressed as the event that the interval $[a_{(T)}, a_{(T+1)})$ is contained in the interval $[b_{(r)}, b_{(t)})$, so that $\underline{R}_{n,m}(\Sigma^*)$ may be written

$$(4.26) \qquad \{b_{(r)} \leqq a_{(T)} < b_{(t+1)}\} - \{b_{(r)} \leqq a_{(T)} < b_{(t+1)} < a_{(T+1)}\}$$

while the second event on the right side of (4.26) may be written

$$(4.27) \qquad \{a_{(T)} < b_{(t+1)} < a_{(T+1)}\} - \{a_{(T)} < b_{(r)}, b_{(t+1)} < a_{(T+1)}\}.$$

From (4.26) and (4.27) one has

$$(4.28) \quad \underline{P}(\Sigma^*) = \sum_{i=r}^{t} \binom{i+T-1}{i}\binom{m+n-i-T}{m-i}\Big/\binom{m+n}{n} - \left[\binom{t+T}{T} - \binom{r+T-1}{T}\right]\binom{m+n-T-t-1}{n-T}\Big/\binom{m+n}{n}.$$

From (4.25) and (4.28) it follows that

$$(4.29) \quad \bar{P}(\Sigma^*) - \underline{P}(\Sigma^*) = \left[\binom{r+T-1}{T}\binom{m+n-r-T}{n-T} + \binom{t+T}{T}\binom{m+n-T-t-1}{n-T} - \binom{r+T-1}{T}\binom{m+n-T-t-1}{n-T}\right]\Big/\binom{m+n}{n}.$$

There is an obvious analogy between the set of formulas (4.1), (4.5), (4.6), (4.10), (4.11), (4.12), (4.16) and (4.2), (4.23), (4.25), (4.26), (4.27), (4.28), (4.29), respectively. This analogy has an important statistical consequence. As $m \to \infty$, $r \to \infty$ and $t \to \infty$ in such a way that $r/m \to \alpha$ and $t/m \to \beta$, one might conjecture that $\bar{P}(\Sigma^*) \to \bar{P}(\Sigma)$ and $\underline{P}(\Sigma^*) \to \underline{P}(\Sigma)$, i.e., that inferences about p should be the same as inferences about the proportion of category 1 observations in a subsequent infinite sample. The validity of these limiting properties is evident from the fact that $b_{(r)}$ and $b_{(t+1)}$ converge in probability to α and β together with the fact that the events governing $\bar{P}(\Sigma)$ and $\underline{P}(\Sigma)$ depend on the interval (α, β) in precisely the same way that the events governing $\bar{P}(\Sigma^*)$ and $\underline{P}(\Sigma^*)$ depend on the interval $(b_{(r)}, b_{(t+1)})$. Thus $\bar{P}(\Sigma^*)$ and $\underline{P}(\Sigma^*)$ actually cover $\bar{P}(\Sigma)$ and $\underline{P}(\Sigma)$ as limiting cases.

Finally, to present a simple result, suppose that $\bar{P}_1$ and $\underline{P}_1$ denote upper and lower probabilities that the next sample individual will be observed in category 1 given that T of the first n sample individuals were observed in category 1. From (4.25) and (4.26) with $m = 1$ and $r = t = 1$,

$$\bar{P}_1 = (T + 1)/(n + 1) \quad \text{and} \quad \underline{P}_1 = T/(n + 1). \tag{4.30}$$

5. Structures of the second kind with finite Θ. Let $\mathfrak{X}$ be a set of k observable categories. Let

$$\Theta = \{1, 2, \cdots, q\} \tag{5.1}$$

index a set of q specified distributions over $\mathfrak{X}$, say $\mathfrak{F}_1, \mathfrak{F}_2, \cdots, \mathfrak{F}_q$. Let Σ be any subset of Θ. The aim here is to develop formulas for $\bar{P}(\Sigma)$ and $\underline{P}(\Sigma)$ based on sample observations $x_1, x_2, \cdots, x_n$ where the sampling model is a structure of the second kind as defined in Section 2.

According to these definitions, $\mathfrak{A}$ is represented by a $(k - 1)$-dimensional simplex with vertices $I_1, I_2, \cdots, I_k$ and μ is the uniform probability measure over $\mathfrak{A}$. Each distribution $\mathfrak{F}_i$ determines a point

$$P_i = (p_{i1}, p_{i2}, \cdots, p_{ik}) \tag{5.2}$$

of $\mathfrak{A}$ where, for $i = 1, 2, \cdots, q$ and $j = 1, 2, \cdots, k$, the probability of category j under $\mathfrak{F}_i$ is denoted by p_{ij}. Each P_i determines a partition of $\mathfrak{A}$ into simplexes $\pi_{i1}, \pi_{i2}, \cdots, \pi_{ik}$ where π_{ij} denotes the simplex with the same vertices as $\mathfrak{A}$ except that I_j is replaced by P_i. The mapping $a \rightarrow x$ corresponding to $\theta = i$ is the mapping which sends $a \,\varepsilon\, \pi_{ij}$ into category j (with some rule to make the mapping specific on the boundaries of the π_{ij}). In accordance with the postulate (P1)

$$\mu(\pi_{ij}) = p_{ij}, \tag{5.3}$$

for $i = 1, 2, \cdots, q$ and $j = 1, 2, \cdots, k$ (c.f., (2.3)).

Consider first inferences based on a sample of size $n = 1$ when the sample observation x_1 falls in category j of $\mathfrak{X}$. The regions R_1, $\bar{R}_1(\Sigma)$ and $\underline{R}_1(\Sigma)$ whose measures determine $\bar{P}(\Sigma)$ and $\underline{P}(\Sigma)$ are given by

$$R_1 = \textstyle\bigcup_{i\varepsilon\Theta} \pi_{ij}, \tag{5.4}$$

$$\bar{R}_1(\Sigma) = \textstyle\bigcup_{i\varepsilon\Sigma} \pi_{ij}, \tag{5.5}$$

and

$$\underline{R}_1(\Sigma) = R_1 - \bar{R}_1(\Theta - \Sigma). \tag{5.6}$$

It turns out to be simpler to characterize intersections of the π_{ij} for given j rather than unions. The intersections are also important for understanding the passage from $n = 1$ to general n. The approach therefore will be to express the probabilities of the unions (5.4) and (5.5) in terms of the probabilities of intersections.

A simplex with the same vertices as $\mathfrak{A}$ except that the vertex I_j of $\mathfrak{A}$ is replaced by a general point of $\mathfrak{A}$ will be called for short a *simplex of type j*. The vertex which replaces I_j will be called the *free vertex*. For convenience, the simplex of type j with free vertex P will be denoted by $\pi_j(P)$. For example, π_{ij} above may also be denoted by $\pi_j(P_i)$.

Using obvious vector space operations of addition and multiplication by a scalar, a general point Q of the simplex $\pi_j(P)$ may be characterized as

$$Q = r_j P + \textstyle\sum_{l=1,\, l\neq j}^{k} r_l I_l \tag{5.7}$$

where $r_l \geqq 0$ for $l = 1, 2, \cdots, k$ and $\sum_1^k r_l = 1$. The following two lemmas will be deduced from (5.7).

LEMMA 5.1. *If Q lies in $\pi_j(P)$ then $\pi_j(Q) \subset \pi_j(P)$, and conversely.*

LEMMA 5.2. *Suppose that $P = \sum_1^k p_l I_l$ and $Q = \sum_1^k q_l I_l$ where $p_l \geqq 0$ and $q_l \geqq 0$ for $l = 1, 2, \cdots, k$ and $\sum_1^k p_l = \sum_1^k q_l = 1$. Then Q lies in $\pi_j(P)$ if and only if*

$$q_l/q_j \geqq p_l/p_j \tag{5.8}$$

for $l = 1, 2, \cdots, k$.

The converse part of Lemma 5.1 is immediate and the direct part requires only a simple application of the definitions of $\pi_j(Q)$ and $\pi_j(P)$, and so is omitted.

To prove Lemma 5.2, note that the comparison of (5.7) with $Q = \sum_1^k q_l I_l$ yields

$$q_j = r_j p_j, \quad \text{and} \quad q_l = r_l + r_j p_l \quad \text{for} \quad l \neq j. \tag{5.9}$$

If Q lies in $\pi_j(P)$ then (5.9) holds with $r_l \geqq 0$ and $r_j = q_j/p_j$, so that (5.8) follows. Conversely, starting from (5.8) and defining $r_1, r_2, \cdots, r_k$ from (5.9) it follows that $r_l \geqq 0$ for $l = 1, 2, \cdots, k$ and

$$\begin{aligned} \textstyle\sum_{l=1}^k r_l &= r_j + \textstyle\sum_{l=1, l\neq j}^k (q_l - r_j p_l) \\ &= r_j + \textstyle\sum_{l=1}^k (q_l - r_j p_l) \\ &= r_j + 1 - r_j \cdot 1 \\ &= 1, \end{aligned} \tag{5.10}$$

as required.

The basic result about intersections, which is stated in Theorem 5.1, asserts that the intersection of a finite set of simplexes of type j is again a simplex of type j.

THEOREM 5.1. *Suppose that P_i is defined by (5.2) for i in a subset Σ of the integers $1, 2, \cdots, q$. Suppose that $Q = \sum_1^k q_l I_l$ is defined by*

$$q_l = \max_{i\varepsilon\Sigma} \{p_{il}/p_{ij}\} / \textstyle\sum_{u=1}^k \max_{i\varepsilon\Sigma} \{p_{iu}/p_{ij}\} \tag{5.11}$$

for $l = 1, 2, \cdots, k$. Then

$$\pi_j(Q) = \textstyle\bigcap_{i\varepsilon\Sigma} \pi_j(P_i). \tag{5.12}$$

To prove Theorem 5.1, consider finding a point Q lying in the desired intersection and having maximum coordinate q_j. From (5.8) it follows that

$$q_l/q_j \geqq p_{il}/p_{ij} \tag{5.13}$$

for i in Σ and hence that

$$q_l/q_j \geqq \max_{i\epsilon\Sigma} \{p_{il}/p_{ij}\} \tag{5.14}$$

for $l = 1, 2, \cdots, k$. Summing and using $\sum_1^k q_l = 1$ gives

$$q_j \leqq [\textstyle\sum_{l=1}^k \max_{i\epsilon\Sigma} \{p_{il}/p_{ij}\}]^{-1}. \tag{5.15}$$

Moreover, if (5.15) is changed to an equality, it is easily seen that Q defined by (5.11) is the only point consistent with (5.14) and $\sum_1^k q_l = 1$, i.e., Q is the unique point in the desired intersection with maximum coordinate q_j. This explains where (5.11) came from.

That

$$\pi_j(Q) \subset \textstyle\bigcap_{i\epsilon\Sigma} \pi_j(P_i) \tag{5.16}$$

follows from Lemma 5.1. That

$$\pi_j(Q) \supset \textstyle\bigcap_{i\epsilon\Sigma} \pi_j(P_i) \tag{5.17}$$

follows by applying Lemma 5.2 to a general point in the intersection and showing that it satisfies the requirement of the converse application of Lemma 5.2 to $\pi_j(Q)$. Thus Theorem 5.1 is proved.

Returning now to *inference* from a sample of size $n = 1$ with x_1 in category j, and reverting to the notation π_{ij} in place $\pi_j(P_i)$, the important consequence of Theorem 5.1 is that

$$\mu(\textstyle\bigcap_{i\epsilon\Sigma} \pi_{ij}) = [\sum_{u=1}^k \max_{i\epsilon\Sigma} \{p_{iu}/p_{ij}\}]^{-1}. \tag{5.18}$$

This follows from (2.3), which shows that $\mu(\pi_j(Q)) = q_j$, and from (5.11) with $l = j$. Note that the numerator of (5.11) is unity when $l = j$.

Since $\bar{R}_1(\Sigma)$ is a union of simplexes of type j as in (5.5), $\mu(\bar{R}_1(\Sigma))$ may be expressed in terms of the quantities defined by (5.18) applied to all subsets of Σ. Specifically, suppose that Σ_{1a} for $a = 1, 2, \cdots$ denote the single element subsets of Σ, that Σ_{2b} for $b = 1, 2, \cdots$ denote the two-element subsets of Σ, that Σ_{3c} for $c = 1, 2, \cdots$ denote the three-element subsets of Σ, and so on. Then

$$\begin{aligned} \mu(\bar{R}_1(\Sigma)) = \textstyle\sum_a \mu(\bigcap_{i\epsilon\Sigma_{1a}} \pi_{ij}) - \sum_b \mu(\bigcap_{i\epsilon\Sigma_{2b}} \pi_{ij}) \\ + \textstyle\sum_c \mu(\bigcap_{i\epsilon\Sigma_{3c}} \pi_{ij}) - \cdots. \end{aligned} \tag{5.19}$$

Formula (5.19) may also be applied when Σ is replaced successively by Θ and by $\Theta - \Sigma$ to determine $\mu(R_1)$ and $\mu(\bar{R}_1(\Theta - \Sigma)) = \mu(R_1) - \mu(\underline{R}_1(\Sigma))$. These along with $\mu(\bar{R}_1(\Sigma))$ determine $\bar{P}(\Sigma)$ and $\underline{P}(\Sigma)$.

Consideration of the simplest case $q = 2$ may help to illuminate the foregoing. Here Θ consists of two elements and there are two non-trivial subsets namely Σ_1 consisting of $i = 1$ and Σ_2 consisting of $i = 2$. Thus only three numbers are required to determine all upper and lower probabilities, namely

$$\begin{aligned} \mu(\bar{R}_1(\Sigma_1)) &= p_{1j}\ ; \\ \mu(\bar{R}_1(\Sigma_2)) &= p_{2j}\ ; \\ \mu(\bar{R}(\Sigma_1) \cap \bar{R}(\Sigma_2)) &= [\textstyle\sum_{u=1}^k \max \{p_{1u}/p_{1j}, p_{2u}/p_{2j}\}]^{-1}. \end{aligned} \tag{5.20}$$

Denoting $\mu(\bar{R}(\Sigma_1) \cap \bar{R}(\Sigma_2))$ by p_{12j} for short, it follows that

$$\mu(R_1) = p_{1j} + p_{2j} - p_{12j} \tag{5.21}$$

and hence that

$$\begin{aligned} \bar{P}(\Sigma_1) &= p_{1j}/(p_{1j} + p_{2j} - p_{12j}), \quad \text{and} \\ \underline{P}(\Sigma_1) &= (p_{1j} - p_{12j})/(p_{1j} + p_{2j} - p_{12j}). \end{aligned} \tag{5.22}$$

For samples of general size n, consideration must be directed to regions in the product space $\mathfrak{A}^n$. If the observations $x_1, x_2, \cdots, x_n$ fall in categories $c_1, c_2, \cdots, c_n$, respectively, and if $\Sigma_i \subset \Theta$ is the subset consisting of i only, then

$$\bar{R}_n(\Sigma_i) = \pi_{ic_1} \times \pi_{ic_2} \times \cdots \times \pi_{ic_n}, \tag{5.23}$$

for $i = 1, 2, \cdots, q$. For general Σ, unions of regions like (5.23) are needed. As already mentioned it is easier to first find intersections. In fact, for general Σ,

$$\textstyle\bigcap_{i\epsilon\Sigma} \bar{R}_n(\Sigma_i) = (\bigcap_{i\epsilon\Sigma} \pi_{ic_1}) \times (\bigcap_{i\epsilon\Sigma} \pi_{ic_2}) \times \cdots \times (\bigcap_{i\epsilon\Sigma} \pi_{ic_n}) \tag{5.24}$$

and thence

$$\textstyle\mu^n(\bigcap_{i\epsilon\Sigma} \bar{R}_n(\Sigma_i)) = \prod_{m=1}^{n} \mu(\bigcap_{i\epsilon\Sigma} \pi_{ic_m}). \tag{5.25}$$

Each term in the product on the right side of (5.25) is of the form (5.18) for different j. Formula (5.19) must be generalized by replacing the terms on the right side by products of n terms as in (5.25). Then the computation of upper and lower probabilities proceeds as before.

The task of determining inferences for a sample of size n may therefore be summarized as follows. For the mth sample individual with observation x_m in category c_m, compute the vector of $2^q - 1$ quantities (5.18) with $j = c_m$ and Σ ranging over the $2^q - 1$ non-empty subsets of $\{1, 2, \cdots, q\}$. Having such a vector for each sample individual, combine these n vectors into a single vector by multiplying the corresponding elements as indicated by (5.25). From this sample vector compute $\mu(\bar{R}_n(\Sigma))$ for any Σ as in the generalization of (5.19) and thence determine upper and lower probabilities as required.

Again the case $q = 2$ is especially simple because the vector of $2^q - 1$ quantities required for each individual reduces to three quantities as in (5.20). Thus for each sample individual s there is a triple $(p_{1j(s)}, p_{2j(s)}, p_{12j(s)})$ for $s = 1, 2, \cdots, n$ where $j(s)$ denotes the observational category into which individual s falls. The inferences (5.22) are modified by replacing $(p_{1j}, p_{2j}, p_{12j})$ with

$$\textstyle(\prod_{s=1}^{n} p_{1j(s)}, \prod_{s=1}^{n} p_{2j(s)}, \prod_{s=1}^{n} p_{12j(s)}). \tag{5.26}$$

6. Concluding remarks. The following discussion of qualitative aspects of the proposed inference methods may help the reader to evaluate these methods.

Unlike the fiducial argument which Fisher limited to continuous observables only, the present methods have been developed above in detail only for finite $\mathfrak{X}$. Interestingly enough, the extension to continuous observables poses greater difficulty in the case of structures of the first kind than in the case of structures of the second kind.

Pick up again the $N(\theta, 1)$ example of Section 2 which illustrates a structure of the first kind extended in the obvious way to cover real x. If the procedures of Section 3 are applied to the $N(\theta, 1)$ example, it is found for $n = 1$ that $R_1 = \mathfrak{A}$ and that ν_1 is the $N(0, 1)$ distribution over $\mathfrak{A}$ which, from (2.1), induces a $N(x_1, 1)$ distribution for θ. In other words, Fisher's fiducial argument is reproduced in this simple case. For general sample sizes, R_n becomes the line in n-space consisting of samples $a_1, a_2, \cdots, a_n$ satisfying $x_i = a_i + \theta$ for $i = 1, 2, \cdots, n$. Unfortunately $\mu^n(R_n)$ is now zero and (3.2) cannot be used. This breakdown is not fatal because the observable x may be approximated by a multinomial observable specifying which of a large set of k mutually exclusive and exhaustive intervals contains x. In this way R_n is approximated by a cylinder with $\mu^n(R_n) > 0$. As $k \rightarrow \infty$ in an appropriate way, the cross-section of the cylinder shrinks to the vanishing point and the posterior distribution induced on θ approaches the $N(\bar{x}, 1/n)$ distribution. This answer is the same as that given by the fiducial argument, but the reasoning is quite different: the present method conditions by R_n while the fiducial argument uses sufficiency to reduce consideration to $\bar{x}$.

It thus appears that multinomial approximation may be used to extend the reasoning of Section 3 to continuous observables. At this point a snag arises in connection with structures of the first kind but not, remarkably enough, in connection with structures of the second kind. The snag is that different multinomial approximations may lead in the limit to different inferences. This does not happen for location parameter situations, such as the $N(\theta, 1)$ example, but a little analysis shows that it does happen for general families $\mathfrak{F}_\theta$ with sampling represented by a structure of the first kind. On the other hand, for structures of the second kind, the fundamental quantity (5.18) does approach a common limit under a wide range of approximating conditions, i.e.,

$$(6.1) \qquad [\textstyle\sum_{u=1}^{k} \max_{i \varepsilon \Sigma} \{p_{iu}/p_{ij}\}]^{-1} \rightarrow [\int \max_{i \varepsilon \Sigma} \{f_i(x)/f_i(x_1)\} \, dx]^{-1}$$

where $(p_{i1}, p_{i2}, \cdots, p_{ik})$ approximates a continuous distribution with density $f_i(x)$. This remarkable property means that the structures of the second kind extend in an unambiguous way to yield inferences for general univariate or multivariate observables. For example, inferences about all the parameters of a multivariate normal distribution from a sample of any size are uniquely defined using a structure of the second kind.

This uniqueness property together with the ability to handle multivariate observables make the inferences based on the structures of the second kind appear very attractive to the author. These inferences have the property that upper and lower probabilities differ even with continuous observables, which is also plausible in small samples.

The new methods pay for the absence of a prior distribution by being able to specify only upper and lower posterior probabilities. If two hypotheses Σ_1 and

Σ_2 are "close" in the sense that $\bar{R}_n(\Sigma_1)$ and $\bar{R}_n(\Sigma_2)$ overlap considerably, then it becomes difficult to decide between such hypotheses because both $\bar{P}(\Sigma_1)$ and $\bar{P}(\Sigma_2)$ are close to $\bar{P}(\Sigma_1 \cup \Sigma_2)$ and there is no unambiguous division of posterior probability between them. Consider an extreme case where $\Sigma_1 = \{\theta_1\}$, $\Sigma_2 = \{\theta_2\}$ and $\mathfrak{F}_{\theta_1} \equiv \mathfrak{F}_{\theta_2}$. Here the hypotheses might fairly be judged indistinguishable, and the present methods react by finding $\bar{P}(\Sigma_1 \cup \Sigma_2) = \bar{P}(\Sigma_1) = \bar{P}(\Sigma_2)$ and $\underline{P}(\Sigma_1) = \underline{P}(\Sigma_2) = 0$. (The Bayesian would, of course, distinguish between such Σ_1 and Σ_2 on the basis of his prior distribution alone.) As illustrated in Section 4, it typically happens that the overlapping of $\bar{R}_n(\Sigma_1)$ and $\bar{R}_n(\Sigma_2)$ becomes less serious as n increases, i.e., large samples have high resolving power.

As in other theories of inference, the concept of likelihood plays a prominent role, but the interpretation of likelihood is radically changed. Here, the standard likelihood function $L(\theta)$ is proportional to $\mu^n(\bar{R}_n(\{\theta\}))$ or to $\nu_n(\bar{R}_n(\{\theta\})) = \bar{P}(\{\theta\})$. While $L(\theta)$ for each θ is the measure of a set, the sets corresponding to different θ overlap in an important way which is not defined by the function $L(\theta)$ itself. Thus, all the relevant information is not contained in $L(\theta)$. In large samples, however, "nearly" all the relevant information resides in $L(\theta)$, as illustrated in Section 4.

Noting that the sampling model specifies a measure μ over $\mathfrak{A}$ in addition to a family of distributions $\mathfrak{F}_\theta$, a reader might jump to the conclusion that μ is playing a role analogous to the prior distribution adopted by a Bayesian. Such an analogy would be specious. The measure μ simply idealizes the assertion that all samples are equally likely. As such it belongs to the category of assumption which is usually regarded as objective, in contrast to the Bayesian prior distribution which is often frankly subjective. It is not the assumption of μ which gives the present methods their distinctiveness, but rather the postulate (P2), or, more precisely, the classes of structures of the first and second kind which translate (P2) into precise models.

REFERENCES

DEMPSTER, A. P. (1963). On direct probabilities. *J. Roy. Statist. Soc. Ser. B* **20** 102–107.

DEMPSTER, A. P. (1964). On the difficulties inherent in Fisher's fiducial argument. *J. Amer. Statist. Assoc.* **59** 56–66.

DEMPSTER, A. P. (1965). On a class of mathematical structures yielding upper and lower probabilities. Unpublished research report.

TODHUNTER, I. (1865). *A History of the Mathematical Theory of Probability*. Reprinted (1949) by Chelsea, New York.