Subjective Bayesian Models in Sampling Finite Populations
Author(s): W. A. Ericson
Source: *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 31, No. 2 (1969), pp. 195-233
Published by: Blackwell Publishing for the Royal Statistical Society
Stable URL: http://www.jstor.org/stable/2984206
Accessed: 20/10/2008 13:45

Subjective Bayesian Models in Sampling Finite Populations

By W. A. Ericson

University of Michigan

[Read before the Royal Statistical Society at a meeting organized by the Research Methods Section on Wednesday, March 12th, 1969, Professor R. L. Plackett in the Chair]

Summary

A general and basic model for inference about characteristics of a finite population of distinguishable elements is presented from a subjectivistic–Bayesian point of view. A subjectivist analogue to simple random sampling, based on the notion of exchangeable random variables, is discussed and the inputs and assumptions underlying the model are shown to involve nothing more than is required for inference under Bayesian models for infinite populations. The model is illustrated by a number of particular examples including one based on the multinomial distribution which incorporates a prior distribution representing an extreme position of initial ignorance. Inferences under this particular model are shown to agree closely in several respects with usual "classical" results. Finally, an extension of the results is presented involving the use of concomitant measurements, and under this Bayesian model several common ratio and regression estimators are shown to arise as means of posterior distributions.

1. Preliminaries

Following several recent writers, Godambe (1955), Hájek (1959), Godambe (1965), and others, we define a finite population of N distinguishable elements labelled by the integers $1,2,\ldots,N$. We let $\mathscr{N} = \{1,2,\ldots,N\}$, the label set, and $\mathbf{X} \equiv (X_1, X_2, \ldots, X_N)$, where X_i is the unknown value of some characteristic possessed by the ith population element. The unknown X_i can be taken as vector valued, a case of considerable practical importance, though only the scalar case will be treated here. Inference concerns the N-dimensional real vector parameter $\mathbf{X}$ or, more realistically, some simple function $g(\mathbf{X})$ of $\mathbf{X}$ say. Here we will be mainly concerned with the simple functions $T = \sum_1^N X_i$, $\mu = T/N$, and $\sigma^2 = 1/N \sum_1^N (X_i - \mu)^2$, the population total, mean, and variance.

A *sample* of size m, s^* is defined quite generally to be an ordered sequence of m of the population elements $i_1^*, i_2^*, \ldots, i_m^*$ ($i_j^* \in \mathscr{N}$, $j = 1, \ldots, m$, repetitions allowed) *together* with the sequence of their associated observed characteristic values $\mathbf{x}^* = (x_{i^*(1)}, x_{i^*(2)}, \ldots, x_{i^*(m)})$, i.e. we observe for each i_j^* that $X_{i^*(j)} = x_{i^*(j)}$, $j = 1, 2, \ldots, m$. While it is therefore assumed that if element j is included in the sample then the value of X_j becomes known with certainty the model is being extended to weaken this restriction and thereby incorporate response error, bias, and non-response.

A *sample design* is then defined by some countable set S^* of ordered sequences, s^*, together with a probability measure assigned by choosing a function $p(s^*) \geqslant 0$, $\sum_{s^* \in S^*} p(s^*) = 1$, where $p(s^*)$ is the probability of choosing the sample s^*. That any such sample design may be implemented by an element by element sampling procedure has been demonstrated by Hanurav (1962).

For any such sample $(s^*, \mathbf{x}^*)$ we define the statistic $(s, \mathbf{x})$ to be the set of indices of *distinct* population elements, $s = \{i_1, \ldots, i_n) \subseteq \mathscr{N}$, included in the observed sequence s^* together with the observed values x_j of X_j, $j \in s$. For notational convenience, given any sample s^* containing the n distinct units $s = \{i_1, \ldots, i_n\}$, we define the (matrix) operator $\mathbf{S}$ such that $\mathbf{S}(\mathbf{X}) = (X_{i(1)}, \ldots, X_{i(n)})$ (for definiteness we assume $i_1 < i_2 < \ldots < i_n$); the complementary operator $\bar{\mathbf{S}}$ such that

$$\bar{\mathbf{S}}(\mathbf{X}) = (X_{j(1)}, X_{j(2)}, \ldots, X_{j(N-n)})$$

for all $j_i \in \mathscr{N} - s$, $(j_1 < j_2 < \ldots < j_{N-n})$; and the vector $\mathbf{x} = (x_{i(1)}, \ldots, x_{i(n)})$ of observed values of $\mathbf{S}(\mathbf{X})$.

It is obvious under this model that for any joint prior probability distribution of the vector parameter $\mathbf{X}$ given by the general density $p'(X_1, \ldots, X_N)$ the posterior probability distribution of $\mathbf{X}$ given the sample $(s^*, \mathbf{x}^*)$ is precisely the same as that given only the statistic $(s, \mathbf{x})$. It then follows immediately from the Bayesian definition of sufficiency that $(s, \mathbf{x})$ is a sufficient statistic, a fact previously demonstrated or noted using the more usual (and equivalent) definitions by Basu (1958), Hájek (1959), and others.

It also follows that given $(s^*, \mathbf{x}^*)$ the likelihood function of $\mathbf{X}$ (the likelihood function being unique only up to an arbitrary positive multiplicative constant) is given by

$$l\{\mathbf{X};\, (s^*, \mathbf{x}^*)\} = l\{\mathbf{X};\, (s, \mathbf{x})\} = \begin{cases} kp(s^*), & \text{for } \mathbf{X} \,|\, S(\mathbf{X}) = \mathbf{x}, \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

where $k > 0$ is an arbitrary constant.

Hence given a joint N-dimensional prior on $\mathbf{X}$, with density $p'(\mathbf{X})$, posterior distribution of $\mathbf{X}$ conditional on a sample of the sort described above has a density given by

$$p\{\mathbf{X} \,|\, (s, \mathbf{x})\} \propto \begin{cases} p(s^*)\, p'(\mathbf{X}) & \text{for } \mathbf{X} \,|\, S(\mathbf{X}) = \mathbf{x}, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

(Here and in the sequel, for notational convenience and where it will not lead to confusion, we will use the same symbol, for instance, $\mathbf{X}$, to indicate both a random variable and its generic value.)

Finally, if $p(s^*)$ is independent of $\mathbf{X}$ (the case to be assumed here) then this posterior density is given by

$$p\{\mathbf{X} \,|\, (s, \mathbf{x})\} = \begin{cases} p'(\mathbf{X}) / p'_{\mathbf{S}(\mathbf{X})}(\mathbf{x}), & \text{for } \mathbf{X} \,|\, \mathbf{S}(\mathbf{X}) = \mathbf{x}, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

where $p'_{\mathbf{S}(\mathbf{X})}(\mathbf{x}) \neq 0$ is just the marginal prior density of $\mathbf{S}(\mathbf{X})$.

Several writers, notably Godambe (1966), have viewed the likelihood function, (1), as being almost, if not completely, uninformative and have adopted principles of inference eschewing the likelihood principle. Our view here, quite to the contrary, is that when reasonable prior distributions are introduced, their revision by sample data can lead to meaningful and useful inferences on those functions of $\mathbf{X}$ which are typically of interest. No new principles of inference are necessary.

In this paper we examine several instances of what we consider to be the simplest useful class of prior distributions—those reflecting exchangeable or symmetrically dependent opinions regarding the X_j's. Such priors seem, in some sense, to yield a subjectivist analogue to inference under simple random sampling. Some basic notions and discussion of subjective Bayesian views in sampling are given in the next section while specific examples and results are given in Sections 3 and 4. Section 5 presents some results on the use of auxiliary information in setting priors.

Before proceeding we note the following earlier relevant work. One of the earliest published examples of a posterior distribution of a mean or total of a finite population appears in a paper of Karl Pearson (1928). In this paper Pearson gives the posterior distribution of the number or proportion of elements possessing some attribute using a "diffuse" prior, by essentially normalizing the hypergeometric likelihood function. Pearson's result is a special case of the results of Section 4. More recently, Aggarwal (1959, 1966) has used some normal distribution priors, mainly as a technique for generating minimax estimators in finite populations. Some of his Bayes estimators are special cases of those given in Section 3.

Hill (1968) has considered the posterior distributions of means and percentiles of finite and infinite populations using quite a different model from that of the present paper. Also, independently, Roberts (1967), using a model similar to that given below, has obtained a few of the results given in Sections 3.1 and 3.2.

Other writers, notably Cochran (1939, 1946), Godambe (1955 and later), Hájek (1959), have used prior distributions (or their equivalents) in sampling theory. These uses have been almost exclusively in discussion of optimum design strategy and not with a view to obtaining useful posterior distributions.

2. EXCHANGEABLE PRIORS

The view taken here is a purely subjectivistic Bayesian one which essentially views statistical inference as a process of revision, by relevant evidence, of one's degrees of belief or ignorance as measured by subjective probability. That is, in the model outlined in the preceding section, $p'(\mathrm{X})$ represents one's initial betting odds on the X_i's in the subjective probability sense of de Finetti, Savage, and others. Perhaps the simplest class of prior distributions is that given by taking $p'(\mathrm{X}) = \prod_1^N p_i'(X_i)$, that is, by viewing the X_i's as independent *a priori*. The meaning and consequences of such a prior are clear—given a sample we learn for certain that $X_i = x_i$ for $i \in s$ but such information does not alter our opinions in any way regarding the non-sampled X_i's. It seems doubtful that there are many, if any, real problems in which one's real opinions would be so expressible.

The simplest useful class of prior distributions in this situation would seem to be that in which the X_i's are viewed as exchangeable random variables. This notion was introduced by de Finetti and discussions are available in de Finetti (1937) and briefly in Feller (1966). De Finetti's monograph is available in English translation and with new notes in Kyburg and Smokler (1964).) The notion is one of symmetry: random variables $X_1, \ldots, X_N$ are said to be exchangeable if each of the $N!$ permutations, $X_{i(1)}, \ldots, X_{i(N)}$, has the same joint probability distribution. Exchangeability thus expresses the prior knowledge that while the units of the finite population are identifiable by their labels (here the integers $1, 2, \ldots, N$) there is no information carried by these labels regarding the associated X_i's; that is, under exchangeability, given $1 \leqslant r \leqslant N$, one's initial betting behaviour regarding events defined by the unknown quantities $X_{i(1)}, \ldots, X_{i(r)}$ is invariant over the ordered sets of indices $i_1, \ldots, i_r$.

Alternatively, exchangeability is akin to viewing the finite population as being effectively randomized. I believe that the notion of exchangeability and exchangeable prior distributions very closely approximates the real opinions of thoughtful "classical" practitioners in many situations where they deem simple random sampling to be appropriate.

Assuming that the X_i's are viewed as exchangeable, one still faces the problem of assigning the N-dimensional joint prior distribution, $p'(\mathbf{X})$. One may be aided in this task by noting that a wide class of prior distributions $p'(\mathbf{X})$ representing exchangeability may be generated by viewing the X_i's as independent, identically distributed *conditional* on some real or hypothetical parameter $\boldsymbol{\theta} = (\theta_1, ..., \theta_k)$ with general density $p(X_i \mid \boldsymbol{\theta})$ and where $\boldsymbol{\theta}$ is assigned the probability distribution function $F(\boldsymbol{\theta})$. The joint prior density $p'(\mathbf{X})$ is then taken as the marginal distribution of $\mathbf{X}$ given by the mixture

$$p'(\mathbf{X}) = \int_{\boldsymbol{\theta}} \prod_{i=1}^{N} p(X_i \mid \boldsymbol{\theta})\, dF(\boldsymbol{\theta}). \tag{4}$$

The generation of a joint prior distribution by this approach is, barring differences in probabilistic interpretation, equivalent to viewing the finite population as a sample from an infinite superpopulation having unknown parameter $\boldsymbol{\theta}$. This notion has been used previously in sampling theory. (See Cochran, 1939, 1946, for example.) Here, unlike earlier uses of the superpopulation concept, the parameter $\boldsymbol{\theta}$ is itself assigned a subjective probability distribution. This "superpopulation" notion seems more generally reasonable from the present viewpoint than under objectivist interpretations of probability.

2.1. *Posterior Distribution*

Combining a prior distribution of the sort generated by (4) with the likelihood function (1) resulting from a sample $(s^*, \mathbf{x}^*)$ which yields the sufficient statistic $(s, \mathbf{x})$ we find that the posterior distribution of $\mathbf{X}$ is given by the density

$$p\{\mathbf{X} \mid (s, \mathbf{x})\} \propto \begin{cases} \displaystyle\int_{\boldsymbol{\theta}} \prod_{i \notin s} p(X_i \mid \boldsymbol{\theta}) \prod_{i \in s} p(x_i \mid \boldsymbol{\theta})\, dF(\boldsymbol{\theta}), & \text{for } \mathbf{X} \mid \mathrm{S}(\mathbf{X}) = \mathbf{x}, \\ 0, & \text{elsewhere.} \end{cases} \tag{5}$$

Noting that the posterior distribution of the "superpopulation" parameter $\boldsymbol{\theta}$, is proportional to $\prod_{i \in s} p(x_i \mid \boldsymbol{\theta})\, dF(\boldsymbol{\theta})$, the posterior on $\mathbf{X}$ may be expressed as

$$p\{\mathbf{X} \mid (s, \mathbf{x})\} = \begin{cases} \displaystyle\int_{\boldsymbol{\theta}} \prod_{i \notin s} p(X_i \mid \boldsymbol{\theta})\, dF(\boldsymbol{\theta} \mid \mathbf{x}), & \mathbf{X} \mid \mathrm{S}(\mathbf{X}) = \mathbf{x}, \\ 0, & \text{otherwise,} \end{cases} \tag{6}$$

where $F(\boldsymbol{\theta} \mid \mathbf{x})$ is the posterior distribution function of $\boldsymbol{\theta}$. Thus it is seen that one's posterior opinions regarding the unobserved X_i's are also exchangeable but now with a density yielded by a mixing distribution equal to the usual posterior on the parameters, $\boldsymbol{\theta}$. Note that under this formulation both the prior and posterior distributions of $\mathbf{X}$ can be viewed as predictive distributions of the unobserved components of $\mathbf{X}$ under the parametric model specified by $p(X_i \mid \boldsymbol{\theta})$ and the prior $F(\boldsymbol{\theta})$. The notion of predictive distributions has been used and discussed by Roberts (1965).

2.2. *Discussion*

Several points should be noted explicitly. First, the basic finite population model which yields the likelihood function given in (1) above is general and virtually free from any subjective element. From a Bayesian point of view the subjectivity enters solely through the choice of the prior, $p'(\mathbf{X})$. When one's prior knowledge regarding the X_i's is such that they may be viewed as exchangeable then, as indicated in expression (4), the assessment of such a prior may be thought of as involving (a) the choice of a parametric family of distributions, $p(X_i \mid \boldsymbol{\theta})$, and (b) the choice of a prior distribution of $\boldsymbol{\theta}$. The *same* two specifications, where *both* typically involve some degree of subjectivity, are precisely the ones required for a Bayesian approach in most parametric inference problems.

The second point is that the basic notions given above can be extended in several directions to give formal expression for a much wider class of realistic prior knowledge. Two fairly obvious extensions are to cases where the elements of the population can be partitioned into k subsets such that within each such subset one views the X_i's as exchangeable and to cases where one's opinions regarding the X_i's are exchangeable conditional on some known auxiliary measures y_i, $i = 1, \ldots, N$. These extensions have obvious connections with "classical" notions of stratification and regression and ratio estimation. Some specific results along the latter lines will be given in Section 5. Stratification will be dealt with in a subsequent paper.

A final point, to be noted explicitly here and used in the sequel, concerns intimate similarities between subjective prior knowledge expressible by exchangeable prior distributions of $\mathbf{X}$ and objective distributions induced by simple random sampling. Recall that under the general model put forth above the units of the population are distinguished by an associated tag (name, address, etc., of the element)—here taken to be coded by the integers $1, \ldots, N$. Thus X_i is the (unknown) variate value possessed by the population element tagged by the integer i (or with name, address, etc., coded by i).

From an objectivist point of view if a simple random sample of n of the N population elements is drawn then each of the $\binom{N}{n}$ subsets, $s = \{i_1, \ldots, i_n\} \subseteq \mathcal{N}$ has probability $p(s) = 1 \Big/ \binom{N}{n}$. Thus conditional on $\mathbf{X} = \mathbf{x}$ the probability that the n sample units assume the values $x_{i(1)}, \ldots, x_{i(n)}$ is given by

$$p(x_{i(1)}, \ldots, x_{i(n)} \mid \mathbf{X} = \mathbf{x}) = p(s) = 1 \Big/ \binom{N}{n}.$$

This distribution, of course, yields the sampling distributions and moments of functions of the sample observations upon which standard sampling theory inferences are based.

From a Bayesian view suppose that a joint prior distribution, $p'(\mathbf{X})$, has been assessed which reflects exchangeable prior knowledge of the X_i's. Let s' be any prespecified subset of n of the population elements—for definiteness we will suppose $s' = \{1, 2, \ldots, n\}$. Then the marginal prior distribution of X_i, $i \in s'$ is the same as that for any other subset of n elements, that is,

$$p'(X_1, X_2, \ldots, X_n) = p'(X_{i(1)}, \ldots, X_{i(n)}) \quad \text{for every } \{i_1, \ldots, i_n\} \subseteq \mathcal{N}.$$

This in turn implies, by simply looking at mixtures, that one's prior distribution of $X_1, \ldots, X_n$ is precisely the same as that on the variate values attached to n units selected by simple random sampling (or indeed by any fixed size probability sampling design). Further we note the property of such priors that given $\mathbf{X} = \mathbf{x}$ then

$$p'(X_1, \ldots, X_n \mid \mathbf{X} = \mathbf{x}) = \begin{cases} 1, & \text{if } X_i = x_i, \quad i = 1, \ldots, n, \\ 0, & \text{otherwise.} \end{cases}$$

While this stands in contrast to the objectivist's sampling distribution there is another aspect of an exchangeable prior which is tantamount to the sampling distribution. Let $\mathbf{X}^* = (X_1^*, \ldots, X_N^*)$ denote the collection of values of $\mathbf{X}$ without their distinguishing labels (or under an arbitrary labelling). Note that $\mathbf{X}^* = \mathbf{x}^*$ whenever $\mathbf{X}$ equals any permutation of the co-ordinates of $\mathbf{x}^* = (x_1^*, \ldots, x_N^*)$. It then follows immediately from the exchangeability of $p'(\mathbf{X})$ that conditional on $\mathbf{X}^* = \mathbf{x}^*$ the X_i's are still exchangeable random variables and hence

$$p'(\mathbf{X} \mid \mathbf{X}^* = \mathbf{x}^*) = 1/N!$$

for $\mathbf{X}$ equal to any permutation of the co-ordinates of $\mathbf{x}^*$; and for any n co-ordinates of $\mathbf{x}^*$, $x_{i^*(1)}, \ldots, x_{i^*(n)}$

$$p'(X_1 = x_{i^*(1)}, \ldots, X_n = x_{i^*(n)} \mid \mathbf{X}^* = \mathbf{x}^*) = \sum p'(\mathbf{X} \mid \mathbf{X}^* = \mathbf{x}^*) = (N-n)!/N!$$

Finally, if we ignore the ordering by letting $\mathbf{X}_n = \{x_{i^*(1)}, \ldots, x_{i^*(n)}\}$ denote the event that $x_{i^*(1)}, \ldots, x_{i^*(n)}$ are the values assumed by $X_1, \ldots, X_n$ *in any order* then

$$p'(\mathbf{X}_n = \{x_{i^*(1)}, \ldots, x_{i^*(n)}\} \mid \mathbf{X}^* = \mathbf{x}^*) = 1 \bigg/ \binom{N}{n}.$$

Thus it follows strictly from an exchangeable prior on $\mathbf{X}$ that given the collection, $\mathbf{X}^*$, of the N population variate values, but not knowing the units to which they are attached, the probability that any prespecified subset of n of the population *units* will assume any subset of these N values is precisely the same as if the subset were drawn by simple random sampling from the N elements of $\mathbf{X}^*$.

As immediate consequences of this result we have that, if the prior on $\mathbf{X}$ is exchangeable and if s denotes any specified set of n of the population elements, and letting $\bar{X}$ and s_n^2 be the mean and variance of these X_i's ($i \in s$), then

$$E(\bar{X} \mid \mu) = E(\bar{X} \mid \mu, \sigma^2) = \mu, \tag{7}$$

$$V(\bar{X} \mid \mu, \sigma^2) = \frac{N-n}{N-1} \frac{\sigma^2}{n}, \tag{8}$$

and

$$E(s_n^2 \mid \mu, \sigma^2) = \sigma^2 \frac{n}{N-1}, \tag{9}$$

where μ and σ^2 are the finite population mean and variance, respectively, as defined earlier. These may also be seen directly by observing that conditional on μ (or on μ and σ^2) the X_i's remain exchangeable random variables and hence have common means, variances, and covariances, and using the conditions:

$$E(\mu \mid \mu) = \mu,$$

$$V(\mu \mid \mu, \sigma^2) = 0,$$

and

$$E(s_N^2\,|\,\mu,\sigma^2) = \sigma^2\frac{N}{N-1}.$$

2.3. *A Result on the Posterior Mean of* μ

In this section we show that under reasonably general conditions the posterior expectation of μ, $E(\mu\,|\,(s,\mathbf{x}))$ is a weighted average of $\bar{x}$ and the prior expectation of μ, $E(\mu)$, with weights respectively inversely proportional to the prior variance of μ, $V(\mu)$, and the prior expectation of the conditional "sampling" variance of $\bar{X}$. This appealing form follows as a corollary of a similar result for "infinite" populations given by Ericson (1969).

By way of preliminaries observe that under the model above the posterior expectation of $\mu = \sum_1^N X_i/N$ is given by

$$E\{\mu\,|\,(s,\mathbf{x})\} = \frac{1}{N}[n\bar{x}+(N-n)\,E\{X_i\,|\,(s,\mathbf{x})\}],$$

and since conditional on $\boldsymbol{\theta}$ the X_i's are independent identically distributed it follows that

$$E\{X_i\,|\,(s,\mathbf{x})\} = E\{\mu(\boldsymbol{\theta})\,|\,(s,\mathbf{x})\},$$

where $\mu(\boldsymbol{\theta}) \equiv E(X_i\,|\,\boldsymbol{\theta})$, the common superpopulation mean. Thus we have

$$E\{\mu\,|\,(s,\mathbf{x})\} = \frac{1}{N}[n\bar{x}+(N-n)\,E\{\mu(\boldsymbol{\theta})\,|\,(s,\mathbf{x})\}]. \tag{10}$$

Assuming the existence of the expectations used below and that $V\{\mu(\boldsymbol{\theta})\} > 0$ we have the following theorem.

Theorem. If $E\{\mu(\boldsymbol{\theta})\,|\,(s,\mathbf{x})\} = \alpha\bar{x}+\beta$ where α and β are independent of the x_i's then

$$E\{\mu\,|\,(s,\mathbf{x})\} = \frac{\bar{x}V(\mu)+m'E_\mu\,V(\bar{X}\,|\,\mu)}{V(\mu)+E_\mu\,V(\bar{X}\,|\,\mu)}, \tag{11}$$

where $V(\mu)$ is the prior variance of μ, $m' \equiv E(\mu)$, the prior mean of μ and $E_\mu\,V(\bar{X}\,|\,\mu)$ is the prior expectation of the conditional variance of $\bar{X}$ given μ.

Proof. It was shown by the author (1969) that under the conditions of the present theorem

$$E\{\mu(\boldsymbol{\theta})\,|\,(s,\mathbf{x})\} = \frac{\bar{x}V\{\mu(\boldsymbol{\theta})\}+E\{\mu(\boldsymbol{\theta})\}\,E_{\boldsymbol{\theta}}\,V(\bar{X}\,|\,\boldsymbol{\theta})}{V\{\mu(\boldsymbol{\theta})\}+E_{\boldsymbol{\theta}}\,V(\bar{X}\,|\,\boldsymbol{\theta})}. \tag{12}$$

Now note that

$$m' \equiv E(\mu) = E_{\boldsymbol{\theta}}\,E(\mu\,|\,\boldsymbol{\theta}) = E\{\mu(\boldsymbol{\theta})\},$$

and by the conditional independence of the X_i's

$$V(\mu) = E_{\boldsymbol{\theta}}\,V(\mu\,|\,\boldsymbol{\theta})+V_{\boldsymbol{\theta}}\,E(\mu\,|\,\boldsymbol{\theta}) = \frac{1}{N}E_{\boldsymbol{\theta}}\,V(X_i\,|\,\boldsymbol{\theta})+V\{\mu(\boldsymbol{\theta})\}$$

and

$$E_{\boldsymbol{\theta}}\,V(\bar{X}\,|\,\boldsymbol{\theta}) = \frac{1}{n}E_{\boldsymbol{\theta}}\,V(X_i\,|\,\boldsymbol{\theta}).$$

Substituting these results in (12) and simplifying we find

$$E\{\mu\,|\,(s,\mathbf{x})\} = \frac{\bar{x}V(\mu)+m'\{(N-n)/Nn\}\,E_{\boldsymbol{\theta}}\,V(X_i\,|\,\boldsymbol{\theta})}{V(\mu)+\{(N-n)/Nn\}\,E_{\boldsymbol{\theta}}\,V(X_i\,|\,\boldsymbol{\theta})}.$$

The asserted result then follows using (8) and noting that

$$V(\bar{X}\,|\,\mu) = E_{\sigma^2|\mu}\,V(\bar{X}\,|\,\mu,\sigma^2)$$

and

$$E_\mu\,V(\bar{X}\,|\,\mu) = E_{\mu,\sigma^2}\,V(\bar{X}\,|\,\mu,\sigma^2) = \frac{N-n}{(N-1)n}E(\sigma^2)$$

$$= \frac{N-n}{(N-1)n}E_{\boldsymbol{\theta}}\,E(\sigma^2\,|\,\boldsymbol{\theta}) = \frac{N-n}{nN}E_{\boldsymbol{\theta}}\,V(X_i\,|\,\boldsymbol{\theta}). \qquad (13)$$

It will be noted that in all of the examples discussed in Sections 3 and 4 below, as well as many others, the condition of the theorem holds, and thus the form (11) holds for a variety of distribution assumptions for finite populations. The usefulness of this theorem is that the condition $E\{\mu(\boldsymbol{\theta})\,|\,(s,\mathbf{x})\} = \alpha\bar{x}+\beta$ is very easily verified, while to show (11) directly has been found often to require considerable manipulation. The last form in (13) also provides an easy way to evaluate $E_\mu\,V(\bar{X}\,|\,\mu)$ rather than using the conditional distribution of $\bar{X}$ given μ.

2.4. *The Role of Randomization*

A few fragmentary comments are in order on the subtle subject of the role of randomization in inference and particularly as it pertains to the specific model outlined above. These comments are in the spirit of general comments on randomization in Bayesian inference given by Savage (1962).

The first point to be re-emphasized here is that under the present model and any prior on $\mathbf{X}$ the form of the posterior conditional on the sample sufficient statistic $(s,\mathbf{x})$ is given by (3) no matter how the sample was selected. (Provided, of course, that the measure $p(s^*)$ is independent of $\mathbf{X}$). The problem of an initial choice of sample design still remains and one might, for example, adopt a criterion such as choosing the design to minimize the prior expectation of the posterior variance of μ. The choice of design under this model will quite realistically depend on one's prior knowledge of the population as reflected here in $p'(\mathbf{X})$ as well as on the economics of implementing various alternative designs.

It is also immediately evident by symmetry that if one's initial knowledge regarding the X_i's is truly reflected by an exchangeable prior on $\mathbf{X}$ then, economics aside, there are no *a priori* grounds for preferring a sample consisting of any particular subset of n of the N units over any other subset of the same size. Under this state of prior knowledge and for the purposes of one's own personal inference, *ceteris paribus*, one may feel rather indifferent regarding randomization and depending on its cost may have preferences against it. When observational costs are introduced there would still remain the question of economic sample size.

Randomization seems compelling because the real world differs from the idealization of the preceding paragraph in at least two important respects. First, sample information is seldom obtained solely for one's own personal inference purposes. Clearly even if one has truly exchangeable opinions regarding the X_i's the use of randomization may more than offset its cost through the increased utility of the resulting sample to others.

Second, it would appear that one's true prior opinions are rarely *exactly* represented by an exchangeable prior distribution of $\mathbf{X} = (X_1, \ldots, X_N)$. Note, however, that for *any* prior on $\mathbf{X}$, $p_0'(\mathbf{X})$ say, if the original identifiability of the population units is to be destroyed by a replacement of their labels (integers $1, \ldots, N$) by one of the $N!$ permutations of these integers chosen with equal probability then the prior distribution of the resulting $\mathbf{X}' = (X_1', \ldots, X_N')$ (under the random labelling) is given by the *exchangeable* prior $p'(\mathbf{X}')$. ($p'(\mathbf{X}')$ being merely the equally weighted mixture of $p_0'(\mathbf{X})$ over all $N!$ permutations of labels.) In view of this, if one's initial opinions regarding $\mathbf{X}$ were roughly exchangeable there may be considerable economy of thought and effort through symmetry and little loss in directly assessing the exchangeable prior $p'(\mathbf{X}')$ rather than agonizing in detail over the contingencies needed to assess $p'(\mathbf{X})$. In this case the model discussed above is exactly applicable for inference regarding $\mathbf{X}'$ and functions of $\mathbf{X}'$, and as pointed out above one would be indifferent among samples consisting of any subset of n of the randomly relabelled units. Finally, for inference regarding permutation–invariant functions of $\mathbf{X}$ (mean, total, variance, percentiles, etc.) this procedure is tantamount to taking $p'(\mathbf{X}) = p'(\mathbf{X}')$, that is, the exchangeable prior on the original identifiable units, and then selecting the sample of n by simple random sampling.

3. Results for Specific Distributions

In this section we give illustrative results for two special cases obtained by taking $p(X_i \mid \boldsymbol{\theta})$ to be normal with either known or unknown variance and taking $dF(\boldsymbol{\theta})$ to be either natural conjugate or diffuse. The case where $p(X_i \mid \boldsymbol{\theta})$ is binomial is a special case of results given in Section 4. Similar results are readily obtained under various other distributional assumptions. In carrying out this program much of the notation follows that of Raiffa and Schlaifer (1961). The reader will also find that a good deal of the required distribution theory used below is given in their monograph.

3.1. *Normal–Superpopulation Variance Known*

As a first example which may not be useful in characterizing real prior opinion but which sets the pattern for many of the results persisting under different assumptions, suppose $p(X_i \mid \boldsymbol{\theta})$ is taken as a normal density with unknown mean θ and *known* variance v and suppose further that $dF(\theta)$ is taken as the conjugate prior—normal with mean m' and variance v'. It follows readily that the prior on $\mathbf{X}$ is an N-dimensional symmetric normal distribution with common means, m', common variances, $v+v'$, and common covariances, v'. The prior distribution on $\mu = N^{-1} \sum_1^N X_i$ is thus normal with mean $E(\mu) = m'$ and variance

$$V(\mu) = v' + v/N. \tag{14}$$

As one expects under such a model, the prior distribution of $\sigma^2 = \sum_1^N (X_i - \mu)^2/N$, the variance of the finite population, is such that $N\sigma^2/v$ has a χ^2 distribution on $N-1$ degrees of freedom—reflecting little prior uncertainty regarding σ^2.

Given a sample consisting of n distinct units and observed variate values $\mathbf{x}$, the posterior on $\mu(\boldsymbol{\theta})$, here equal to θ, is, by well-known normal distribution theory, normal with mean

$$E\{\theta \mid (s, \mathbf{x})\} \equiv m'' = \frac{\bar{x}v' + m'v/n}{v' + v/n} \tag{15}$$

and variance

$$v'' = \frac{1}{n/v+1/v'} = \left(\frac{v/n}{v'+v/n}\right) v', \tag{16}$$

that is, the posterior mean of the superpopulation mean θ is a weighted average of the sample mean, $\bar{x}$, and the prior mean, m', with weights inversely proportional to the variances $V(\bar{X}|\theta)$ and $V(\theta)$, while the posterior variance of θ is the reciprocal of the sum of these weights. These forms carry over to the finite population in a very natural way.

From the form (6) it is clear that the posterior distribution of **X** is an N-dimensional singular normal distribution with all its probability concentrated in the subspace where $\mathbf{S}(\mathbf{X}) = \mathbf{x}$, or equivalently the posterior distribution on the unobserved co-ordinates $\bar{\mathbf{S}}(\mathbf{X})$ is $(N-n)$-dimensional normal with common means, m'', common variances and covariances given by $v+v''$ and v'' respectively, where v'' is as in (16).

It then follows immediately that the finite population mean μ, which we write in the form

$$\mu = N^{-1}\left(n\bar{x} + \sum_{i\notin s} X_i\right),$$

being a linear combination of normal random variables has a normal posterior distribution with mean

$$E\{\mu|(s,\mathbf{x})\} = N^{-1}\{n\bar{x}+(N-n)\,m''\}$$

and variance

$$V\{\mu|(s,\mathbf{x})\} = N^{-2}\{(N-n)(v+v'')+(N-n)(N-n-1)\,v''\}.$$

Substituting from (14), (15), and (16) these quantities may be rewritten as

$$E\{\mu|(s,\mathbf{x})\} = \frac{n(Nv'+v)\,\bar{x}+(N-n)\,vm'}{N(nv'+v)} \tag{17}$$

and

$$V\{\mu|(s,\mathbf{x})\} = \frac{N-n}{N^2}\,\frac{v(Nv'+v)}{(nv'+v)} = \frac{N-n}{N}\,\frac{v/n}{v'+v/n}\,V(\mu). \tag{18}$$

This last expression for the posterior variance is instructive, for the first factor is a finite population correction factor and the second factor is just that by which the data reduce the prior variance of the superpopulation mean (compare with (16)). It is also clear from the theorem of Section 2.3 and the linearity of $E\{\theta|(s,\mathbf{x})\}$ in $\bar{x}$ (formula (15)) that the weights in (17) have the interpretation given in (11). This is also directly verified using (14) and from (13) noting that $E_\mu\, V(\bar{X}|\mu) = \{(N-1)/nN\}v$.

Alternatively, under the present model, it is easily verified that $\bar{X}_n$, the mean of any n of the X_i's, and μ have a bivariate normal distribution. From this and the sufficiency of $\bar{X}_n$ it follows that the distribution of $\bar{X}_n$ given μ is $N[\mu,\{(N-n)/N\}(v/n)]$, while the marginal on μ is $N\{m',(Nv'+v)/N\}$. Thus using standard normal prior-to-posterior results it follows that the posterior distribution of μ given $(s,\mathbf{x})$ is normal with mean and variance given by (15) and (16), with v' replaced by $(Nv'+v)/N$ and v by $(N-n)\,v/N$. It also follows immediately that in the present case

$$V(\bar{X}|\mu) = E_\mu\, V(\bar{X}|\mu).$$

Finally, note that if the prior on θ is taken as degenerate uniform on the whole real line than the posterior on the finite population mean μ is normal with mean $\bar{x}$ and variance $\{(N-n)/N\}(v/n)$.

3.2. *Normal–Superpopulation Mean and Variance Unknown*

Suppose now, somewhat more realistically, that the finite population is viewed as a sample from a normal superpopulation with unknown mean θ and unknown variance $1/h$, thus we here take $\boldsymbol{\theta} = (\theta, h)$. Hence

$$p(\mathbf{X}\,|\,h,\theta) \propto h^{\frac{1}{2}N}\exp(-\tfrac{1}{2}hN\sigma^2)\exp(-\tfrac{1}{2}hN(\mu-\theta)^2\} = h^{\frac{1}{2}N}\exp\left\{-\tfrac{1}{2}h\sum_{1}^{N}(X_i-\theta)^2\right\}, \quad (19)$$

where again μ and σ^2 are the mean and variance of the finite population. Note that here σ^2 is not equal to $1/h$.

Suppose further that $F(\boldsymbol{\theta})$ is taken to be non-degenerate normal-gamma with density

$$f(\theta,h) \propto \exp\{-\tfrac{1}{2}hn'(\theta-m')^2\}h^{\frac{1}{2}\delta(n')}\exp(-\tfrac{1}{2}hv'v')h^{\frac{1}{2}v'-1}, \quad 0<h<\infty, \quad -\infty<\theta<\infty,$$
$$v', n', v' > 0. \quad (20)$$

(See Raiffa and Schlaifer, 1961, for further discussion.)

By defining $\delta(n') = 0$ if $n' = 0$ and $\delta(n') = 1$ for $n' > 0$ various degenerate priors can be had as special cases of this density: for example, taking $n' = v' = 0$ (20) reduces to

$$f(\theta,h) \propto h^{-1}, \quad 0<h<\infty, \quad -\infty<\theta<\infty. \quad (21)$$

This is merely the often used prior discussed by Jeffreys (1948) and Savage (1961), which seems a successful representation of vagueness in that it is uniform in θ and in the logarithm of the variance, h^{-1}.

It is shown in the Appendix that under this model the prior distribution of the finite population mean, μ, is "Student" that is, μ is distributed like

$$m' + t_{v'}\left\{\frac{(N+n')v'}{n'N}\right\}^{\frac{1}{2}}, \quad (22)$$

where $t_{v'}$ is a standard t random variable on v' degrees of freedom. Thus the prior mean and variance of μ are m' and $v'v'(N+n')/n'N(v'-2)$ respectively.

The main result concerning the posterior distribution of μ is given in the following:

Theorem. (a) With the joint prior on **X** implied by (19) and (20) and given the sample $(s, \mathbf{x})$ with mean $\bar{x}$ and variance s^2 the posterior distribution of μ is "Student", that is, μ is distributed like the quantity

$$\frac{V(\mu)\,\bar{x} + V(\bar{X}\,|\,\mu)\,E(\mu)}{V(\mu) + V(\bar{X}\,|\,\mu)} + t_{v''}\left[\frac{N-n}{N}\,\frac{V\{\theta\,|\,(s,\mathbf{x})\}}{V(\theta)}\,V(\mu)\,\frac{v''-2}{v''}\right]^{\frac{1}{2}}, \quad (23)$$

where $v'' = v'+n$, $V(\theta\,|\,(s,\mathbf{x}))$ and $V(\theta)$ are the posterior and prior variances of θ, $V(\bar{X}\,|\,\mu)$ is the prior variance of the mean of a sample of size n given the finite population mean, $V(\mu)$ and $E(\mu)$ are the prior variance and mean of μ, respectively, and $t_{v''}$ is a standard t random variable on v'' degrees of freedom.

(b) Under the prior on **X** generated by (19) and (21) the posterior distribution of μ given $(s, \mathbf{x})$ is such that μ is distributed like the quantity

$$\bar{x} + t_{n-1}\left(\frac{N-n}{N}\,\frac{s^2}{n}\right)^{\frac{1}{2}}. \quad (24)$$

A proof of this result, along with a number of alternative expressions, explicit formulae, and some results on the prior and posterior distributions of the finite population variance, σ^2, are given in the Appendix. We note here merely that the posterior mean of μ is again of the weighted average form given in equation (11), and where, in this particular instance, $V(\bar{X}|\mu)$ is independent of μ. Also from (23) it follows that the posterior variance of μ, $\{(N-n)/N\}[V(\theta|(s,\mathbf{x})\}/V(\theta)]\,V(\mu)$, is simply the prior variance reduced by the same two factors as in the preceding example (compare (18)). Finally, for a diffuse prior on the superpopulation parameters the posterior mean and variance of μ are seen to be $\bar{x}$ and $\{(N-n)/N\}\{(n-1)/(n-3)\}s^2/n$ respectively, in close agreement with that which one might expect from traditional sampling theory.

4. Extreme Prior Vagueness and the Multinomial

Analyses of the sort given in the preceding section can, of course, be carried through under various alternative assumptions regarding $p(X_i|\boldsymbol{\theta})$, for example, by assuming the superpopulations to be Poisson, gamma, binomial (a special case of the results given below), etc. Under these models one's prior uncertainty is represented in terms of his imperfect knowledge regarding $\boldsymbol{\theta}$ and, to a lesser extent especially for large finite populations, by viewing the population as a sample from a superpopulation. While such models may often adequately approximate one's prior uncertainty regarding the unknown $\mathbf{X}$, nevertheless they do assume strong prior knowledge regarding the shape of the finite population distribution by taking the superpopulation form as known. Weaker prior knowledge of the shape of the finite population can be modelled by taking the $p(X_i|\boldsymbol{\theta})$ in (4) to be a member of a more flexible family of distributions, for example, the power distributions used by Box and Tiao (1962). Other approaches might take $p'(\mathbf{X})$ as a more complicated mixture. An alternative approach, based on the multinomial distribution and incorporating extreme vagueness regarding the shape of the finite population, is developed in this section. Special cases of the results below are the early result of Pearson (1928) suitable for dichotomous populations (using (4) with a binomial superpopulation), and a generalization using more general beta priors on θ than the uniform one used by Pearson.

4.1. *The Basic Model*

Suppose that each X_i can only assume one of the finite set of numerical values $\mathscr{Y} = \{y_1, y_2, \ldots, y_k\}$, $y_1 < y_2 < \ldots < y_k$, where k may be an extremely large integer having no relation whatever to N, the population size. This assumption clearly recognizes the inherent discreteness of almost all observation due to limitations of measuring instruments, etc. Suppose that the probability that X_i equals y_j is p_j, that is,

$$P\{X_i = y_j|\mathbf{p}\} = p_j, \quad j = 1, \ldots, k, \quad \sum_{j=1}^{k} p_j = 1, \tag{25}$$

and where $\mathbf{p} = (p_1, \ldots, p_{k-1})$. Here $\mathbf{p}$ is assumed unknown and plays the role of $\boldsymbol{\theta}$ in (4). The X_i's are assumed independent and identically distributed with the distribution (25). Thus for any $\mathbf{y} = (y_{i(1)}, \ldots, y_{i(N)})$, $y_{i(j)} \in \mathscr{Y}$, a joint prior on $\mathbf{X}$ can be given by

$$p'(\mathbf{X} = y) = \int_{\mathbf{p}} \prod_{j=1}^{N} p_{i(j)} f'(\mathbf{p})\, d\mathbf{p}, \tag{26}$$

where $f'(\mathbf{p})$ is a prior density on the superpopulation parameter $\mathbf{p}$. The class of prior distributions of $\mathbf{X}$ of the form (26) again reflects exchangeability regarding the X_i's.

We shall find it more convenient in dealing with certain aspects of this model to consider it in slightly different terms. Let N_j be the unknown number of the N population elements for which $X_i = y_j, j = 1, \ldots, k$. In this notation $\mu = N^{-1}\sum_{j=1}^{k} y_j N_j$ and $\sigma^2 = N^{-1}(\sum_{j=1}^{k} y_j^2 N_j - N\mu^2)$. In this manner inferences regarding these and other symmetric functions of the unknown finite population X_j's are expressible in terms of the unknown N_j's. From (26) it follows that the joint prior distribution on the N_j's is given in non-singular form by

$$p'(\mathbf{N}) = \int_{\mathbf{p}} \frac{\Gamma(N+1)}{\prod_{i=1}^{k-1}\Gamma(N_i+1)\,\Gamma\left(N-\sum_{1}^{k-1} N_i+1\right)} \prod_{i=1}^{k-1} p_i^{N_i}\left(1-\sum_{1}^{k-1} p_i\right)^{N-\Sigma_1^{k-1}N_i} f'(\mathbf{p})\,d\mathbf{p}, \qquad (27)$$

where $\mathbf{N} = (N_1, N_2, \ldots, N_{k-1})$, $\mathbf{p} = (p_1, p_2, \ldots, p_{k-1})$, and the integral is over the simplex $\{\mathbf{p}\,|\,0 \leqslant p_i \leqslant 1, \sum_1^{k-1} p_i \leqslant 1\}$.

To summarize briefly and in slightly different form: Under the present model both the real finite population and the hypothetical "superpopulation" are defined by unknown distributions over the k distinct and ordered values $y_1, \ldots, y_k$. The real finite population is defined by the unknown distribution function (d.f.)

$$F_X(x) = P(X \leqslant x) = \begin{cases} 0, & x < y_1, \\ \sum_{\mathscr{I}(x)} N_i/N, & y_1 \leqslant x \leqslant y_k, \\ 1, & x > y_k, \end{cases}$$

where $\mathscr{I}_x = \{i\,|\,y_i \leqslant x\}$. In addition this finite population is assumed to have been generated by N independent observations on a random variable having unknown distribution function defined by

$$F_Z(z) = P(Z \leqslant z) = \begin{cases} 0, & z < y_1, \\ \sum_{\mathscr{I}(z)} p_i, & y_1 \leqslant z \leqslant y_k, \\ 1, & z > y_k, \end{cases}$$

where $\mathscr{I}(z) = \{i\,|\,y_i \leqslant z\}$.

Choice of $f'(\mathbf{p})$

To this point the model outlined above seems straightforward and realistic. The difficulty, however, is in choosing appropriate and useful prior distributions for the parameter $\mathbf{p}$. Pending further study and only as a tentative and convenient approximation in certain special cases, we will take the prior on $\mathbf{p}$ to be a $(k-1)$-dimensional Dirichlet distribution, (see Wilks, 1962) with density

$$f'(\mathbf{p}) = \frac{\Gamma(\epsilon)}{\prod_1^{k-1}\Gamma(\epsilon_i)\,\Gamma\left(\epsilon-\sum_1^{k-1}\epsilon_i\right)} \prod_{i=1}^{k-1} p_i^{\epsilon_i-1}\left(1-\sum_1^{k-1} p_i\right)^{\epsilon-\Sigma_i^{k-1}\epsilon_i-1}, \qquad (28)$$

for parameter values $\epsilon_i > 0$, $i = 1, \ldots, k$, $(\epsilon = \sum_1^k \epsilon_i)$. It is well known that this distribution has means, variances, and covariances given by

$$E(p_i) = \epsilon_i/\epsilon, \qquad i = 1, \ldots, k,$$

$$V(p_i) = \frac{\epsilon_i(\epsilon - \epsilon_i)}{\epsilon^2(\epsilon+1)}, \qquad i = 1, \ldots, k,$$

$$\operatorname{cov}(p_i, p_j) = \frac{-\epsilon_i \epsilon_j}{\epsilon^2(\epsilon+1)}, \qquad i \neq j = 1, \ldots, k.$$

It then follows immediately that the mean and variance of the distribution function $F_Z(z)$ of the "superpopulation" are given by

$$E\{F_Z(z)\} = \begin{cases} 0, & z < y_k, \\ \dfrac{1}{\epsilon} \sum\limits_{\mathscr{I}(z)} \epsilon_i, & y_1 \leqslant z \leqslant y_k, \\ 1, & z > y_k, \end{cases}$$

where $\mathscr{I}_z$ was defined above, and

$$V\{F_Z(z)\} = \begin{cases} 0, & z < y_1 \quad \text{or} \quad z > y_k, \\ \dfrac{1}{\epsilon+1} E\{F_Z(z)\}\,[1 - E\{F_Z(z)\}], & y_1 \leqslant z \leqslant y_k. \end{cases}$$

It then follows that by the choice of the ϵ_i's (up to an arbitrary positive multiplicative constant) the prior expectation of the d.f., $F_Z(z)$, can assume almost any desired shape, while by choosing that constant so that $\epsilon = \sum_1^k \epsilon_i$ is small, the variance of $F_Z(z)$ can be made large—representing vagueness regarding this unknown "superpopulation" d.f. It is in fact only for such small values of ϵ_i and ϵ that we believe such a Dirichlet prior is at all tenable in this setting and even then it is only tentatively put forth to represent an initial state of extreme vagueness.

Before discussing this further we note that the resulting prior on **N** is given, using (27) and (28), as

$$p'(\mathbf{N}) = f_{\mathrm{DM}}^{(k-1)}(\mathbf{N} \mid N, \epsilon, \boldsymbol{\epsilon}) \equiv$$

$$\frac{\Gamma(N+1) \prod\limits_{i=1}^{k-1} \Gamma(N_i + \epsilon_i)\, \Gamma\left(N - \sum\limits_1^{k-1} N_i + \epsilon - \sum\limits_1^{k-1} \epsilon_i\right) \Gamma(\epsilon)}{\prod\limits_{i=1}^{k-1} \Gamma(N_i+1)\, \Gamma\left(N - \sum\limits_1^{k-1} N_i + 1\right) \Gamma(N+\epsilon) \prod\limits_{i=1}^{k-1} \Gamma(\epsilon_i)\, \Gamma\left(\epsilon - \sum\limits_1^{k-1} \epsilon_i\right)},$$

$$N_i = 0, 1, \ldots, N, \quad \sum_{i=1}^{k-1} N_i \leqslant N, \tag{29}$$

and where $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_{k-1})$.

This joint distribution is merely a straightforward generalization of what Raiffa and Schlaifer (1961) have termed the beta–binomial distribution. It might analogously be termed the Dirichlet–multinomial distribution; Mosimann (1962) has termed it the compound multinomial distribution. We will adopt the notation $f_{\mathrm{DM}}^{(k-1)}(\mathbf{N}\,|\,N,\epsilon,\boldsymbol{\epsilon})$ to denote the $(k-1)$-dimensional Dirichlet-multinomial distribution whose density is given in (29). It is also easily seen that this joint distribution arises by: assuming that the N_i's are independent Poisson random variables with parameters λ_i, $i=1,\ldots,k$, assuming that the λ_i's are independent gamma random variables with parameters ϵ_i and common scale parameter α, and then finding the joint marginal distribution of the N_i's conditioned on $\sum N_i = N$. Thus this prior has the property that the only relationship among the N_i's is due to the constraint that they sum to N. When the finite population is viewed microscopically as envisaged under the present model (i.e. when k is huge and the collection of y_j's represents all possible observations on any X_i to the practical limitations of one's measuring ability), such a feature of one's prior opinions does seem approximately realistic. However, this property persists even when large numbers of adjacent cell frequencies, N_j, are grouped. This class of priors thus seems incapable of giving formal expression to prior knowledge which is characterized by vagueness concerning the shape of the finite population and a belief that the grouped frequencies, at least, are "smooth".

We proceed to examine some consequences arising from the adoption of such a Dirichlet-multinomial prior with emphasis on the case where the ϵ_i's and their sum are chosen to be small. In such a case this prior seems to represent an extreme position of prior vagueness, not even incorporating prior belief in the "smoothness" of grouped frequencies. As will be seen below, even though in any sample from the finite population almost all cells have an observed frequency of zero, seemingly realistic and useful posterior inferences can be made about those properties of the population typically of interest in sampling. Such inferences depend little on the actual choice of the ϵ_i's, providing they and their sum are small. However, since there are aspects of this class of priors which are not at all realistic, it is to be expected that certain features of the resulting posteriors will be disturbing.

Properties of the prior on N

Before proceeding with the analysis of aspects of the posterior distributions it is useful for future reference to catalogue some properties of the prior distribution of several characteristics of the finite population. Since, given $\mathbf{p}$, the N_i's are multinomially distributed while the p_i's have means and variances given in the preceding section, it follows readily that

$$E(\mu) = \sum_{1}^{k} y_j\,\epsilon_j/\epsilon \tag{30}$$

and

$$V(\mu) = \frac{N+\epsilon}{N(\epsilon+1)}\left\{\sum_{1}^{k} y_i^2\cdot\epsilon_i/\epsilon - \left(\sum_{1}^{k} y_i\,\epsilon_i/\epsilon\right)^2\right\}, \tag{31}$$

where again μ is the finite population mean.

Similarly, the prior expectation of the population variance, σ^2, is given by

$$E(\sigma^2) = \frac{\epsilon(N-1)}{N(\epsilon+1)}\left\{\sum_{1}^{k} y_i^2\,\epsilon_i/\epsilon - \left(\sum_{1}^{k} y_i\,\epsilon_i/\epsilon\right)^2\right\}. \tag{32}$$

Also if ξ_π is defined as the πth percentile of the finite population, that is, $\xi_\pi = y_j$ if j is the smallest integer for which $\sum_{i=1}^{j} N_i/N \geqslant \pi$, then since $\xi_\pi \leqslant y_j$ whenever $\sum_{i=1}^{j} N_i \geqslant N\pi$ the prior distribution function of ξ_π is given by

$$p'(\xi_\pi \leqslant y_j) = \begin{cases} \displaystyle\sum_{c_j=\{N\pi\}}^{N} \binom{N}{c_j} \frac{\Gamma(\epsilon)\Gamma\left(c_j+\sum_1^j \epsilon_i\right)\Gamma\left(N+\epsilon-c_j-\sum_1^j \epsilon_i\right)}{\Gamma\left(\sum_{i=1}^{j}\epsilon_i\right)\Gamma\left(\epsilon-\sum_{i=1}^{j}\epsilon_i\right)\Gamma(N+\epsilon)}, & j = 1,\ldots,k-1, \\ 1, & j = k, \end{cases} \tag{33}$$

where $\{N\pi\}$ denotes the smallest integer not less than $N\pi$. This follows since, given $\mathbf{p}$, $\sum_{i=1}^{j} N_i \equiv c_j$ has a binomial distribution with parameters N and $\sum_{i=1}^{j} p_i$ while, by properties of the Dirichlet distribution on $\mathbf{p}$, $\sum_{i=1}^{j} p_i$ has a beta distribution. The successive terms in the sum (33) are merely the terms of the beta-binomial distribution. For further discussion of this distribution see Raiffa and Schlaifer (1961).

There is a further interesting and useful property implicit in the use of the prior (29) which we have assumed. If s denotes *any* subset of n of the N distinct population elements and if n_j denotes the number of those X_i's ($i \in s$) equal to y_j then the joint prior distribution on $\mathbf{n} = (n_1, \ldots, n_{k-1})$ *given* $\mathbf{N}$ is simply the generalized hypergeometric

$$p'(\mathbf{n}\,|\,\mathbf{N}) = \frac{\prod_{i=1}^{k-1}\binom{N_i}{n_i}\binom{N-\sum_1^{k-1} N_i}{n-\sum_1^{k-1} n_i}}{\binom{N}{n}}, \quad n_i = 0, \ldots, N_i, \quad \sum_1^{k-1} n_i \leqslant n. \tag{34}$$

This result follows readily from (26) since the joint distribution of $\mathbf{N}$ and $\mathbf{n}$ is given, in singular form, by

$$\begin{aligned} p'(\mathbf{n},\mathbf{N}) &= \int_{\mathbf{p}} \frac{n!}{\prod_1^k n_i!} \prod_1^k p_i^{n_i} \frac{(N-n)!}{\prod_{i=1}^{k}(N_i-n_i)!} \prod_1^k p_i^{N_i-n_i} f'(\mathbf{p})\,d\mathbf{p} \\ &= \frac{n!}{\prod_1^k n_i!} \frac{(N-n)!}{\prod_1^k (N_i-n_i)!} \frac{\Gamma(\epsilon)\prod_1^k \Gamma(N_i+\epsilon_i)}{\prod_1^k \Gamma(\epsilon_i)\,\Gamma(N+\epsilon)}. \end{aligned}$$

The result (34) then follows by dividing this expression by $p'(\mathbf{N})$ (formula (29)).

This conditional "sampling" distribution is a property of the exchangeable prior distribution and does not, in the usual sense, depend on random sampling. We record, for future reference, some well-known properties of this conditional distribution:

$$E(n_i\,|\,\mathbf{N}) = \frac{n}{N} N_i, \tag{35}$$

$$V(n_i\,|\,\mathbf{N}) = \frac{(N-n)\,nN_i(N-N_i)}{N^2(N-1)}, \tag{36}$$

and

$$\operatorname{cov}(n_i, n_j \mid \mathbf{N}) = \frac{-(N-n)\,nN_i N_j}{N^2(N-1)}. \tag{37}$$

4.2. *Posterior Inference*

We now proceed to develop some important aspects of the posterior distribution of several characteristics of the finite population typically of interest in survey work. Given any sample, no matter how selected, consisting of n distinct population units with their associated observed variate values, that is, given $(s, \mathbf{x})$, it follows immediately from (5) and (26) that the posterior distribution of $\mathbf{X}$ is given by

$$p\{\mathbf{X} = \mathbf{y} \mid (s, \mathbf{x})\} \propto \begin{cases} \displaystyle\int_{\mathbf{p}} \prod_{j \notin s} p_{i(j)} \prod_{j=1}^{k} p_j^{n_j} f'(\mathbf{p})\, d\mathbf{p}, & \mathbf{y} \mid \mathbf{S}(\mathbf{y}) = \mathbf{x}, \\ 0, & \text{otherwise}, \end{cases} \tag{38}$$

where $\mathbf{y} = (y_{i(1)}, \ldots, y_{i(N)})$ and n_j $(j = 1, \ldots, k)$ is the number of the n observed x_i's equal to y_j. Combining the second product in the integral above with the Dirichlet density $f'(\mathbf{p})$ it is clear that the posterior distribution on the unsampled units is of the same form as the prior, (29). It is immediately evident that the posterior distribution on the parameter $\mathbf{p}$ is of the same form as the prior—a $(k-1)$-dimensional Dirichlet with the ϵ_i's replaced by $(\epsilon_i + n_i)$'s.

Thus letting $M_j = N_j - n_j$ be the number of unsampled X_i's which are equal to y_j, $j = 1, \ldots, k$, $\mathbf{n} = (n_1, \ldots, n_{k-1})$, and $M = N - n$ it follows that the posterior distribution of $\mathbf{M} = (M_1, \ldots, M_{k-1})$ is also $(k-1)$-dimensional Dirichlet–multinomial with parameters M, $\epsilon + n$, and $\boldsymbol{\epsilon} + \mathbf{n}$, that is, has a density given by

$$p\{\mathbf{M} \mid (s, \mathbf{x})\} = f_{\mathrm{DM}}^{(k-1)}(\mathbf{M} \mid M, \epsilon + n, \boldsymbol{\epsilon} + \mathbf{n}) \tag{39}$$

in the notation of formula (29). The posterior distribution of N is then immediately obtainable by substitution.

This posterior distribution, of course, is exactly that resulting by taking (34) as a likelihood function using the observed $\mathbf{n}$, with the prior (29): that is, (39) is, as a function of $\mathbf{N}$, proportional to the product of (34) and (29). Also it should be emphasized that *a posteriori* the M_j's are distributed multinomially given $\mathbf{p}$ while $\mathbf{p}$ has the $(k-1)$-dimensional Dirichlet distribution defined by the parameters $\epsilon_i + n_i$.

The posterior mean of μ

Given any sample $(s, \mathbf{x})$ with observed sample mean, $\bar{x}$, the posterior distribution of the finite population mean, μ, has a mean $E\{\mu \mid (s, \mathbf{x})\}$ which is again of the weighted average form displayed in (11): also, for all intents and purposes, $E\{\mu \mid (s, \mathbf{x})\} \doteq \bar{x}$.

We note first that

$$E\{M_i \mid (s, \mathbf{x})\} = \frac{(N-n)(\epsilon_i + n_i)}{\epsilon + n}, \tag{40}$$

$$V\{M_i \mid (s, \mathbf{x})\} = \frac{(N-n)(N+\epsilon)(\epsilon_i + n_i)(\epsilon + n - \epsilon_i - n_i)}{(\epsilon + n)^2(\epsilon + n + 1)} \tag{41}$$

and

$$\operatorname{cov}\{M_i, M_j \mid (s, \mathbf{x})\} = \frac{-(N-n)(N+\epsilon)(\epsilon_i + n_i)(\epsilon_j + n_j)}{(\epsilon + n)^2(\epsilon + n + 1)}. \tag{42}$$

These formulae follow most easily by conditioning on $\mathbf{p}$ initially, using moments of the multinomial and Dirichlet distributions. Since $\mu = N^{-1}\sum_1^k y_i(M_i+n_i)$ and since $n^{-1}\sum_1^k n_i y_i = n^{-1}\sum_{i\in s} x_i = \bar{x}$, it follows readily using (40) and (30) that

$$E\{\mu\,|\,(s,\mathbf{x})\} = \frac{n(N+\epsilon)\,\bar{x}+(N-n)\,\epsilon E(\mu)}{N(\epsilon+n)}. \tag{43}$$

It is then clear that if ϵ is small, as assumed here, then

$$E\{\mu\,|\,(s,\mathbf{x})\} \doteq \bar{x}; \tag{44}$$

even if ϵ is near unity the relative weight assigned to $E(\mu)$ is the typically small fraction $(N-n)/N(n+1)$.

It is here again easily verified that the condition of the theorem of Section 2.3 holds with $\boldsymbol{\theta} = \mathbf{p}$ under the present model, and hence the posterior mean, (43), can be recast in the form (11). We then have

$$\mu(\mathbf{p}) = E(X_i\,|\,\mathbf{p}) = E_{\mathbf{N}|\mathbf{p}}E(X_i\,|\,\mathbf{N},\mathbf{p}) = E_{\mathbf{N}|\mathbf{p}}\sum_1^k y_i N_i/N = \sum_1^k y_i p_i, \tag{45}$$

and using the fact that $\mathbf{p}$ has, *a posteriori*, a Dirichlet distribution it follows that

$$E\{\mu(\mathbf{p})\,|\,(s,\mathbf{x})\} = \sum_1^k y_i \frac{n_i+\epsilon_i}{n+\epsilon} = \frac{n\bar{x}+\epsilon E(\mu)}{n+\epsilon}, \tag{46}$$

which is linear in $\bar{x}$, as required.

We can thus deduce by equating (43) with (11) and using (31) and (32) that

$$E\{V(\bar{X}\,|\,\mu)\} = \frac{N-n}{N-1}\,\frac{E(\sigma^2)}{n}, \tag{47}$$

where $E(\sigma^2)$ is as given in (32).

The posterior variance of μ

Several results can now be demonstrated concerning the variance of the posterior distribution μ, $V(\mu\,|\,(s,\mathbf{x}))$. Clearly

$$V\{\mu\,|\,(s,\mathbf{x})\} = \frac{1}{N^2}\,V\left(\sum_{j=1}^k M_j y_j\,|\,(s,\mathbf{x})\right),$$

and using the results (41) and (42) we find the expression

$$V\{\mu\,|\,(s,\mathbf{x})\} = \frac{(N-n)(N+\epsilon)}{N^2}\left\{\sum_1^k y_i^2\frac{(\epsilon_i+n_i)(\epsilon+n-\epsilon_i-n_i)}{(\epsilon+n)^2(\epsilon+n+1)} - \sum_{i\neq j} y_i y_j \frac{(\epsilon_i+n_i)(\epsilon_j+n_j)}{(\epsilon+n)^2(\epsilon+n+1)}\right\}. \tag{48}$$

The first result, in analogy with results given in Section 3, is that if $\mu(\mathbf{p}) \equiv \sum_1^k y_i p_i$, the superpopulation mean, then using (31), and the facts that the posterior variance of $\mu(\mathbf{p})$ is merely the quantity in curly brackets in (48) and that the prior variance of $\mu(\mathbf{p})$ is the same quantity with $n_i = n = 0$ for all i, it follows from (48) that

$$V\{\mu\,|\,(s,\mathbf{x})\} = \frac{N-n}{N}\left[\frac{V\{\mu(\mathbf{p})\,|\,(s,\mathbf{x})\}}{V\{\mu(\mathbf{p})\}}\right]V(\mu). \tag{49}$$

The other, more interesting, results come from re-expressing (48) in a more enlightening form. By letting $s^2 = (n-1)^{-1}(\sum_{i=1}^{k} y_i^2 n_i - n\bar{x}^2)$, the observed sample variance, then after some algebraic manipulation it may be verified that (48) is equivalent to

$$V\{\mu|\,(s,\mathbf{x})\} = \frac{(N-n)(N+\epsilon)(n-1)s^2}{N^2(n+\epsilon)(n+\epsilon+1)} + \frac{\epsilon(\epsilon+1)(N-n)\,V(\mu)}{N(n+\epsilon)(n+\epsilon+1)} + \frac{n\epsilon(N-n)(N+\epsilon)\{\bar{x}-E(\mu)\}^2}{N^2(n+\epsilon)^2(n+\epsilon+1)}. \qquad (50)$$

From this form it follows that if the prior parameter ϵ is chosen small (approaching zero) then

$$V\{\mu|(s,\mathbf{x})\} \doteq \frac{N-n}{N}\,\frac{n-1}{n+1}\,\frac{s^2}{n} \doteq \frac{N-n}{N}\,\frac{s^2}{n}. \qquad (51)$$

Even if ϵ is near unity the first term in (50) may be an adequate approximation if n is large, for the latter two terms of (50) are of $O(1/n^2)$. Thus under the extreme diffuseness of prior knowledge captured in this model the posterior distribution of the finite population mean, μ, has approximate mean $\bar{x}$ and variance given by (51). This provides a Bayesian interpretation for the usual unbiased estimates in traditional sample survey theory. Conclusions like this, especially regarding the mean, under an interesting alternative Bayesian model have been obtained earlier by Hill (1969).

The final result concerns the exact form (50). At first glance it appears similar to that obtained for the posterior variance of μ assuming a normal superpopulation with unknown mean and variance (Section 3.2). On closer examination it turns out that $V\{\mu|(s,\mathbf{x})\}$, (50), is of precisely the same form as under that normal distribution model! This may be seen by comparing (50) with the result (A21) of the Appendix and using (A7) and taking $\epsilon = n'$ and $\epsilon+3 = \nu'$. This coincidence is being studied further.

The posterior mean of σ^2

It is also interesting to note how the data change the expectation of the variance, σ^2, of the finite population. Since $\sigma^2 = N^{-1}(\sum_1^k y_i^2 N_i - N\mu^2)$ it follows that the posterior expectation of σ^2 is given by

$$E\{\sigma^2|(s,\mathbf{x})\} = \frac{1}{N}\left[\left[\sum_1^k y_i^2\,[n_i + E\{M_i|(s,\mathbf{x})\} - NV\{\mu|(s,\mathbf{x})\} - NE^2\{\mu|(s,\mathbf{x})\}]\right]\right]. \qquad (52)$$

Substituting from (40), (43), and (50) and after some further algebra using (32) and (33) it may be verified that

$$E\{\sigma^2|(s,\mathbf{x}) = \frac{1}{N}\left[\frac{(N+\epsilon)\{N\epsilon+n(N+1)\}(n-1)s^2}{N(n+\epsilon)(n+\epsilon+1)} + \frac{(N-n)(\epsilon+1)\{Nn+\epsilon(N-1)\}E(\sigma^2)}{(N-1)(n+\epsilon)(n+\epsilon+1)} + \frac{(N-n)(N+\epsilon)(n\epsilon)\{\bar{x}-E(\mu)\}^2}{N(n+\epsilon)(n+\epsilon+1)}\right]. \qquad (53)$$

Using (32) it follows that if ϵ is very small then

$$E\{\sigma^2|(s,\mathbf{x})\} \doteq \frac{N+1}{N}\,\frac{n-1}{n+1}\,s^2, \qquad (54)$$

where s^2 is, as before, the observed sample variance. For ϵ near unity the two other terms in (53) play a more important role than in $V\{\mu|(s,\mathbf{x})\}$, for here they are each of $O(1/n)$.

Here too it may be verified that the expression (53) is formally identical to that (formula (A26) of the Appendix) under the normal model, by making the correspondence $\epsilon = n'$ and $\epsilon+3 = \nu'$.

Posterior distribution of percentiles

In this section we obtain quite easily the exact posterior distribution of ξ_π, the πth percentile of the finite population. It is then briefly demonstrated that standard confidence intervals for percentiles agree approximately with posterior probability intervals under the model being treated here. An alternative subjectivistic approach to the distribution of percentiles has been given by Hill (1968).

To obtain the posterior distribution of ξ_π, as defined below equation (32), we note that $\xi_\pi \leqslant y_j$ whenever $\sum_1^j (M_i+n_i) \geqslant N\pi$ and thus

$$p\{\xi_\pi \leqslant y_j|(s,\mathbf{x})\} = p\left\{\sum_1^j M_i \geqslant N\pi - \sum_1^j n_i \,\Big|\, (s,\mathbf{x})\right\}.$$

As with the prior distribution on ξ_π the posterior distribution of $\sum_{i=1}^j M_i$ is given by the beta-binomial distribution:

$$\begin{aligned} p\left\{\sum_1^j M_i = u \,\Big|\, (s,\mathbf{x})\right\} &= \int_0^1 \binom{N-n}{u} w^u (1-w)^{N-n-u} \frac{\Gamma(n+\epsilon)\, w^{\Sigma_1{}^j(n_i+\epsilon_i)-1}\,(1-w)^{n+\epsilon-\Sigma_1{}^j(n_i+\epsilon_i)-1}}{\Gamma\{\sum_1^j (n_i+\epsilon_i)\}\,\Gamma\{n+\epsilon-\sum_1^j (n_i+\epsilon_i)\}}\, dw \\ &= \binom{N-n}{u} \frac{\Gamma(n+\epsilon)\,\Gamma\{u+\sum_1^j (n_i+\epsilon_i)\}\,\Gamma\{N+\epsilon-u-\sum_1^j (n_i+\epsilon_i)\}}{\Gamma\{\sum_1^j (n_i+\epsilon_i)\}\{\Gamma n+\epsilon-\sum_1^j (n_i+\epsilon_i)\}\,\Gamma(N+\epsilon)}, \\ & \qquad\qquad u = 0, 1, \ldots, N-n. \qquad (55) \end{aligned}$$

It follows that

$$p\{\xi_\pi \leqslant y_j|(s,\mathbf{x})\} = \begin{cases} 0, \quad \text{if } \{N\pi\} - \sum\limits_1^j n_i > N-n, \quad j = 1, \ldots, k, \\ \sum\limits_{u=\{N\pi\}-\Sigma_1{}^j n_i}^{N-n} p\left\{\sum\limits_1^j M_i = u \,\Big|\, (s,\mathbf{x})\right\}, & 0 < \{N\pi\} - \sum_1^j n_i \leqslant N-n, \\ & j = 1, \ldots, k-1, \\ 1, & \text{if } j = k \quad \text{and/or} \quad \{N\pi\} - \sum\limits_1^j n_i \leqslant 0, \end{cases} \qquad (56)$$

where again $\{N\pi\}$ denotes the smallest integer not less than $N\pi$, and where the terms in the sum are given by (55). This expression of course yields the posterior distribution on ξ_π for any configuration of sample observations. It is clear from (55) and (56) that with the extremely diffuse prior (including no smoothness beliefs) under this model, that is, ϵ_i and ϵ small, most of the posterior probability is unrealistically concentrated on those values, y_i, observed in the sample. Nonetheless, posterior probability is attached to *intervals* in approximate agreement with standard confidence intervals.

To show briefly the relationship between inferences based on the above posterior distribution and standard confidence intervals for percentiles we assume, for simplicity, that the n sample observations are distinct, in other words that n_j is either zero or one for all j. We denote by $x_{(i)}$ the ith sample order statistic, thus $x_{(1)} < x_{(2)} < \ldots < x_{(n)}$; and suppose also that $x_{(i)} = y_l < y_m = x_{(j)}$ for any $i<j$. We finally approximate the distribution in (56) by taking $\epsilon_i = 0$, $i = 1, \ldots, k$, in (55).

Under these assumptions it is first clear that if $n_j = 0$ then $p\{\xi_\pi = y_j \,|\, (s, \mathbf{x})\} = 0$. Additionally if $n_j = 1$ then using (55) and (56) and for notational convenience letting

$$v_j \equiv \sum_{i=1}^{j-1} n_i, \quad \left(v_j + 1 = \sum_{i=1}^{j} n_i\right),$$

one has

$$p(\xi_\pi = y_j \,|\, (s,\mathbf{x})) = \frac{1}{\binom{N-1}{n-1}} \left\{ \sum_{u=\{N\pi\}-v_j-1}^{N-n} \binom{u+v_j}{v_j} \binom{N-u-v_j-2}{n-v_j-2} \right.$$

$$\left. - \sum_{n=\{N\pi\}-v_j}^{N-n} \binom{u+v_j-1}{v_j-1} \binom{N-u-v_j-1}{n-v_j-1} \right\}.$$

Using the well-known identity $\binom{n}{r} = \binom{n-1}{r-1} + \binom{n-1}{r}$, this expression reduces to the hypergeometric term

$$p\{\xi_\pi = y_j \,|\, (s,\mathbf{x})\} = \frac{\binom{\{N\pi\}-1}{\sum_1^{j-1} n_i} \binom{N-1-(\{N\pi\}-1)}{n-1-\sum_1^{j-1} n_i}}{\binom{N-1}{n-1}}, \tag{57}$$

whenever this expression is defined and zero otherwise. Using these results it follows that

$$p\{x_{(i)} < \xi_\pi \leqslant x_{(j)} \,|\, (s,\mathbf{x})\} = \sum_{j=l+1}^{m} p\{\xi_\pi = y_j \,|\, (s,\mathbf{x})\}$$

$$= \sum_{v=i}^{j-1} \frac{\binom{\{N\pi\}-1}{v} \binom{N-1-(\{N\pi\}-1)}{N-1-v}}{\binom{N-1}{n-1}}. \tag{58}$$

Finally, if N is large relative to n and since $(\{N\pi\}-1)/(N-1) \doteq \pi$, using the binomial approximation to the hypergeometric we have

$$p\{x_{(i)} < \xi_\pi \leqslant x_{(j)} \,|\, (s,\mathbf{x})\} \doteq \sum_{v=i}^{j-1} \binom{n-1}{v} \pi^v (1-\pi)^{n-1-v}$$

$$= I_\pi(i, n-i) - I_\pi(j, n-j), \tag{59}$$

where $I_\pi(u,v)$ is the usual incomplete beta function. This expression is then recognized as approximately the confidence coefficient attached to the statement that ξ_π is trapped within the random interval $\{x_{(i)}, x_{(j)}\}$ (see Wilks, 1948).

5. SOME RESULTS ON THE USE OF AUXILIARY MEASUREMENTS

In this section we give a simple extension of some of the earlier results to indicate how the basic model may be extended to incorporate *a priori* knowledge of some concomitant measurements $y_1, y_2, \ldots, y_N$ and one's prior knowledge regarding the relation of the unknown X_i's to these values, by using them to help assess the requisite N-dimensional prior distribution of **X**. Under various assumed relationships between y and **X** some commonly used ratio and regression estimators turn out to be the means of the posterior distribution of μ under diffuse priors.

5.1. *Regression Model*

In the following it is assumed that the y_i's are known positive values associated with the distinguishable population elements. Let $z_i = g(y_i)$, $i = 1, \ldots, N$, where g is some pre-specified positive valued function. We consider the model obtained by assuming that given y_i, α, and h, the X_i's are independent normally distributed with means αy_i and variances z_i/h, $i = 1, \ldots, N$. The parameter (α, h) is assumed unknown and assigned a normal-gamma prior.

We will be interested in three special cases obtained by taking z_i to be y_i^2, 1, and y_i respectively. Each of these cases is equivalent to a regression through the origin with three different assumptions regarding the error variances.

Letting $v = 1/h$ note that these three cases are respectively equivalent to:

(a) $$X_i/y_i = \alpha + \epsilon_i, \quad \text{where} \quad \epsilon_i \sim N(0, v), \tag{60}$$

that is, it is assumed that the unknown ratios X_i/y_i, given y_i, are equal to an unknown constant, α, plus a normally distributed error having unknown variance, v.

(b) $$X_i = \alpha y_i + \epsilon_i, \quad \text{where} \quad \epsilon_i \sim N(0, v), \tag{61}$$

a regression through the origin with constant unknown error variance.

(c) $$X_i/y_i = \alpha + \epsilon_i, \quad \text{where} \quad \epsilon_i \sim N(0, v/y_i), \tag{62}$$

that is, the ratios X_i/y_i equal an unknown constant plus a normal error having an unknown variance proportional to y_i^{-1}.

Before proceeding one further comment seems appropriate. Although it is assumed here that the investigator can, on the basis of his prior information and knowledge, adequately approximate his prior distribution by appropriately choosing z_i, this is not completely required. He can let the data (sample) point the way by his choosing only a family of functions, say $z_i = y_i^r$, where r is unknown and then assigning a joint prior distribution on (α, h, r). The basic form of the analysis to derive the posterior distribution of μ remains unchanged. This remark is in the spirit of the important paper of Box and Cox (1964.)

5.2. *Analysis of Model*: *Prior to Posterior*

Under this model for setting a joint prior on **X**

$$p(\mathbf{X} \mid \alpha, h) \propto \prod_{i=1}^{N} \left(\frac{h}{z_i}\right)^{\frac{1}{2}} \exp\left\{-\frac{1}{2}\frac{h}{z_i}(X_i - \alpha y_i)^2\right\}, \tag{63}$$

the conditioning on the y_i's being implicit. It is assumed that a normal-gamma prior distribution is assigned (α, h) having density

$$f'(\alpha, h) \propto \exp\{-\tfrac{1}{2}hn'(\alpha - \bar{\alpha}')^2\}\, h^{\frac{1}{2}\delta(n')} \exp(-\tfrac{1}{2}h\nu' v')\, h^{\frac{1}{2}\nu' - 1}, \tag{64}$$

where $\delta(n') = 0$ if $n' = 0$ and one otherwise. The "diffuse" prior,

$$f'(\alpha, h) \propto h^{-1}, \tag{65}$$

will be considered as a special case obtained by putting $n' = \nu' = 0$ in (64). The joint prior distribution of $\mathbf{X}$ can then be obtained by multiplying (63) by (64) and integrating out (α, h).

Given α and h, $\mu = \sum_1^N X_i/N$ is clearly normally distributed with mean $\alpha \sum_1^N y_i/N$ and variance $\sum_1^N z_i/(N^2 h)$, and thus has density

$$p(\mu \mid h, \alpha) \propto (N^2 h/z)^{\frac{1}{2}} \exp\{-\tfrac{1}{2}(N^2 h/z)(\mu - \alpha\bar{y})^2\},$$

where $z = \sum_{i=1}^N z_i$ and $\bar{y} = \sum_{i=1}^N y_i/N$. Multiplying by $f'(\alpha, h)$ and integrating out (α, h) we find, after some simplification and assuming $n' > 0$, $\nu' > 0$, that the prior density of μ is given by

$$p'(\mu) \propto \left\{\nu' + \frac{n'N^2}{v'(n'z + y^2)}(\mu - \bar{\alpha}'\bar{y})^2\right\}^{-\{(\nu'+1)/2\}},$$

where $y = N\bar{y} = \sum_1^N y_i$. From well-known properties of the "Student" distribution it follows that

$$E(\mu) = \bar{\alpha}'\bar{y}$$

and

$$V(\mu) = \frac{\nu' v'}{\nu' - 2}\,\frac{n'z + y^2}{n'N^2}, \quad \nu' > 2.$$

To obtain the posterior distribution of μ given the sample $(s, \mathbf{x})$ we recall the fact that $\mu = (n\bar{x}_s + (N-n)\mu_{\bar{s}})/N$ where $\mu_{\bar{s}} = \sum_{i \notin s} X_i/(N-n)$ and $\bar{x}_s$ is the sample mean. Proceeding conditionally we have that, given $(s, \mathbf{x})$, α and h, $\mu_{\bar{s}}$ has a normal density

$$p(\mu_{\bar{s}} \mid (s, \mathbf{x}), \alpha, h) \propto \left\{\frac{(N-n)^2 h}{z_{\bar{s}}}\right\}^{\frac{1}{2}} \exp\left\{-\frac{1}{2}\frac{h(N-n)^2}{z_{\bar{s}}}(\mu_{\bar{s}} - \alpha\bar{y}_{\bar{s}})^2\right\}, \tag{66}$$

where $z_{\bar{s}} = \sum_{i \notin s} z_i$ and $\bar{y}_{\bar{s}} = \sum_{i \notin s} y_i/(N-n)$.

Next the posterior distribution of (α, h) given $(s, \mathbf{x})$ has density

$$f\{\alpha, h \mid (s, \mathbf{x})\} \propto h^{n/2} \exp\left\{-\tfrac{1}{2}h \sum_{i \in s}\left(\frac{x_i - \alpha y_i}{\sqrt{z_i}}\right)^2\right\}$$
$$\exp\{-\tfrac{1}{2}hn'(\alpha - \bar{\alpha}')^2\}\, h^{\frac{1}{2}\delta(n')} \exp(-\tfrac{1}{2}h\nu' v')\, h^{\frac{1}{2}\nu' - 1}.$$

This, after some simplification, can be written as

$$f\{\alpha, h \mid (s, \mathbf{x})\} \propto \exp\{-\tfrac{1}{2}hn''(\alpha - \bar{\alpha}'')^2\}\, h^{\frac{1}{2}\delta(n'')} \exp\{-\tfrac{1}{2}h\nu'' v''\}\, h^{\frac{1}{2}\nu'' - 1}, \tag{67}$$

where

$$n'' = \sum_{i \in s} y_i^2/z_i + n', \tag{68}$$

$$\bar{\alpha}'' = \frac{1}{n''}\left(\sum_{i \in s} x_i y_i/z_i + n'\bar{\alpha}'\right), \tag{69}$$

$$\nu'' v'' = \left\{\nu' v' + \frac{n'}{n''}\sum_{i \in s}\left(\frac{x_i - \bar{\alpha}' y_i}{\sqrt{z_i}}\right)^2 + \frac{1}{n''}\sum_{i \in s} y_i^2/z_i \sum_{i \in s} x_i^2/z_i - \frac{1}{n''}\left(\sum_{i \in s} x_i y_i/z_i\right)^2\right\}, \tag{70}$$

$$\nu'' = n + \nu' + \delta(n') - \delta(n''), \tag{71}$$

and $\delta(n'') = 0$ if $n'' = 0$ and unity otherwise.

Writing the distribution in this fashion it is immediately obvious that it is still of the normal–gamma family. Thus the derivation of the posterior distribution of $\mu_{\bar{s}}$ is analogous to that of the prior on μ. The result is again a "Student" density,

$$f\{\mu_{\bar{s}} | (s,\mathbf{x})\} \propto \left\{\nu'' + \frac{n''(N-n)^2}{v''(n'' z_{\bar{s}} + y_{\bar{s}}^2)} (\mu_{\bar{s}} - \bar{\alpha}'' \bar{y}_{\bar{s}})^2\right\}^{-\frac{1}{2}(\nu''-1)},$$

where $y_{\bar{s}} = (N-n)\bar{y}_{\bar{s}} = \sum_{i \notin s} y_i$; and hence the posterior mean and variance of $\mu_{\bar{s}}$ are given by

$$E\{\mu_{\bar{s}} | (s,\mathbf{x})\} = \bar{\alpha}'' \bar{y}_{\bar{s}}$$

and

$$V\{\mu_{\bar{s}} | (s,\mathbf{x})\} = \frac{\nu'' v''}{\nu''-2} \frac{n'' z_{\bar{s}} + y_{\bar{s}}^2}{n''(N-n)^2}, \quad \nu'' > 2.$$

It now follows immediately that the posterior distribution of μ given $(s,\mathbf{x})$ is like that of the quantity

$$\frac{n}{N}\bar{x}_s + \bar{\alpha}'' \frac{N-n}{N} \bar{y}_{\bar{s}} + t_{\nu''} \left\{\frac{v''(n'' z_{\bar{s}} + y_{\bar{s}}^2)}{n'' N^2}\right\}^{\frac{1}{2}}, \tag{72}$$

where $t_{\nu''}$ is a standard t random variable on ν'' degrees of freedom. Thus

$$E\{\mu | (s,\mathbf{x})\} = \frac{n}{N}\bar{x}_s + \frac{N-n}{N} \bar{\alpha}'' \bar{y}_{\bar{s}} \tag{73}$$

and

$$V\{\mu | (s,\mathbf{x})\} = \frac{\nu'' v''}{\nu''-2} \frac{(n'' z_{\bar{s}} + y_{\bar{s}}^2)}{n'' N^2}. \tag{74}$$

The posterior distribution of the finite population mean, μ, with only diffuse prior information on (α, h) is immediately obtainable from the above results simply by letting $\nu' = n' = \delta(n') = 0$. It then follows that given the n distinct sample elements, $(s,\mathbf{x})$, μ is distributed like a linear function of a standard t random variable on $n-1$ degrees of freedom. This situation will be discussed explicitly for the three special cases mentioned earlier.

5.3. *Special Cases of Diffuseness*

Several interesting results obtain by reconsidering the three cases introduced in Section 5.1 with only vague prior information on α and h, as represented by the improper prior density of (65). Under each of these models the posterior "Student" distribution of μ turns out to be centred on (has mean equal to) a familiar and natural "classical" estimator. This then gives some formal subjective Bayesian interpretation or justification for these estimators and some insight into the sort of prior distributions which might result in their use. Exact Bayesian credible intervals on μ are immediately obtainable using (72). In each case below we merely state the posterior mean and variance of μ obtained from (73) and (74) using the definitions (67)–(71) and taking $\nu' = n' = \delta(n') = 0$.

In the first case (a) $(z_i = y_i^2)$ one finds

$$E\{\mu | (s,\mathbf{x})\} = \frac{1}{N}(n\bar{x}_s + \bar{r} y_{\bar{s}}) = \frac{1}{N}\{n(\bar{x}_s - \bar{r}\bar{y}_s) + \bar{r} y\}, \tag{75}$$

where $\bar{r} = \sum_{i\in s} r_i/n$, $y = \sum_1^N y_i$ and $r_i = x_i/y_i$. Also

$$V\{\mu|(s,\mathbf{x})\} = \left(\frac{n-1}{n-3}\right)\frac{s_r^2}{n}\frac{n\sum_{i\in s}y_i^2+y_{\bar{s}}^2}{N^2}, \tag{76}$$

where $s_r^2 = \sum_{i\in s}(r_i-\bar{r})^2/(n-1)$. This posterior mean is a natural average of ratios estimator.

In case (b) ($z_i = 1$) one finds

$$E\{\mu|(s,\mathbf{x})\} = \frac{1}{N}\{n\bar{x}_s+\hat{\alpha}(y-n\bar{y}_s)\}, \tag{77}$$

where $\hat{\alpha} = \sum_s x_i y_i/\sum_s y_i^2$ is the usual least-squares estimator of α, and $\bar{y}_s = \sum_s y_i/n$. Also

$$V\{\mu|(s,\mathbf{x})\} = \frac{n-1}{n-3}s_{\hat{\alpha}}^2\frac{(N-n)\sum_s y_i^2-y_{\bar{s}}^2}{N^2}, \tag{78}$$

where

$$s_{\hat{\alpha}}^2 = \frac{\sum_s y_i^2\sum_s x_i^2-(\sum_s x_i y_i)^2}{(n-1)(\sum y_i^2)^2}$$

is also the usual "classical" estimator of the variance of $\hat{\alpha}$.

Finally, in the last model (c) ($z_i = y_i$) one finds

$$E\{\mu|(s,\mathbf{x})\} = \frac{\bar{x}_s}{\bar{y}_s}\bar{y}, \tag{79}$$

the usual ratio estimator, and

$$V\{\mu|(s,\mathbf{x})\} = \frac{1}{N^2}\frac{s^2}{y_s}(y_{\bar{s}}y), \tag{80}$$

where

$$s^2 = \frac{\sum_s y_i\sum_s x_i^2/y_i-(\sum_s x_i)^2}{(n-3)\sum_s y_i},$$

y and y_s are the totals over the population and sample respectively of the y_i's, and $y_{\bar{s}} = y-y_s$.

ACKNOWLEDGEMENTS

I am deeply grateful for the many useful comments on earlier drafts of this paper by G. R. Antelman, V. P. Godambe, J. W. Pratt, H. V. Roberts, J. Sedransk, and a referee.

This research was partially sponsored by the U.S. Bureau of the Census and by the National Science Foundation under Grants NSF-GP-6008 and NSF-GP-7350.

REFERENCES

AGGARWAL, O. P. (1959). Bayes and minimax procedures in sampling from finite and infinite populations, I. *Ann. Math. Statist.*, **30**, 206–218.

—— (1966). Bayes and minimax procedures for estimating the arithmetic mean of a population with two-stage sampling. *Ann. Math. Statist.*, **37**, 1186–1195.

BASU, D. (1958). On sampling with and without replacement. *Sankhyā*, **20**, 287–294.

BOX, G. E. P. and COX, D. R. (1964). An analysis of transformations (with discussion). *J. R. Statist. Soc.* B, **26**, 211–252.

BOX, G. E. P. and TIAO, G. C. (1962). A further look ot robustness via Bayes's theorem. *Biometrika*, **49**, 419–432.

COCHRAN, W. G. (1939). The use of the analysis of variance in enumeration by sampling. *J. Amer. Statist. Ass.*, **34**, 492–510.

—— (1946). Relative accuracies of systematic and stratified random samples for a certain class of population. *Ann. Math. Statist.*, **17**, 164–179.

DE FINETTI, B. (1937). La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, **7**, 1–68.

ERICSON, W. A. (1969). A note on the posterior mean of a population mean. *J. R. Statist. Soc.* B, **31**, 332–334.

FELLER, W. (1966). *An Introduction to Probability Theory and its Applications*, Vol. II. New York: Wiley.

GODAMBE, V. P. (1955). A unified theory of sampling from finite populations. *J. R. Statist. Soc.* B, **17**, 267–278.

—— (1965). A review of the contributions towards a unified theory of sampling from finite populations. *Inter. Statist. Inst. Rev.*, **33**, 242–258.

—— (1966). A new approach to sampling from finite populations, I and II. *J. R. Statist. Soc.* B, **28**, 310–328.

HÁJEK, J. (1959). Optimum strategy and other problems in probability sampling. *Časopis Pest. Mat.*, **84**, 387–423.

HANURAV, T. V. (T. V. HANUMANTHA RAO) (1962). An existence theorem in sampling theory. *Sankhyā*, **24**, 327–330.

HILL, B. M. (1968). Posterior distribution of percentiles: Bayes's theorem for sampling from a population. *J. Amer. Statist. Ass.*, **63**, 677–691.

—— (1969). Foundations for the theory of least squares. *J. R. Statist. Soc.* B, **31**, 89–97.

JEFFREYS, H. (1948). *Theory of Probability*, 2nd ed. Oxford: Clarendon Press.

KYBURG, H. E. and SMOKLER, H. E. (Ed.) (1964). *Studies in Subjective Probability*. New York: Wiley.

MOSIMANN, J. E. (1962). On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. *Biometrika*, **49**, 65–82.

PEARSON, K. (1928). On a method of ascertaining limits to the actual number of marked members in a population of given size from a sample. *Biometrika*, **20**, 149–174.

RAIFFA, H. and SCHLAIFER, R. O. (1961). *Applied Statistical Decision Theory*. Boston: Graduate School of Business Administration, Harvard University.

ROBERTS, H. V. (1965). Probabilistic prediction. *J. Amer. Statist. Ass.*, **60**, 50–62.

—— (1967). Statistical inference and decision, Chapter 14: Sampling from finite populations. (Private communication.)

SAVAGE, L. J. (1961). The subjective basis of statistical practice. (Private communication.)

—— (1962). *The Foundations of Statistical Inference*. London: Methuen.

WILKS, S. S. (1948). Order statistics. *Bull. Amer. Math. Soc.*, **54**, 6–50.

—— (1962). *Mathematical Statistics*. New York: Wiley.

APPENDIX

We give here a proof of the theorem of Section 2.3, as well as a number of related results under the normal distribution model of that section.

Prior on (μ, σ^2)

We begin by obtaining the prior distribution on **X**. Multiplying (19) by (20) and integrating out θ and h one finds that the prior on **X** is "Student" with density

$$p'(\mathbf{X}) \propto \{\nu' + (\mathbf{X}-\mathbf{m}')\,\mathbf{H}(\mathbf{X}-\mathbf{m}')^t\}^{-\frac{1}{2}(N+\nu')}, \quad -\infty < \mathbf{X} < \infty, \tag{A1}$$

where the superscript t denotes transpose, $\mathbf{m}' = (m', \ldots, m')$ and

$$\mathbf{H} = \frac{1}{\nu'(N+n')}\begin{bmatrix} N+n'-1 & -1 & \ldots & -1 \\ -1 & N+n'-1 & \ldots & -1 \\ \vdots & \vdots & & \vdots \\ -1 & -1 & \ldots & N+n'-1 \end{bmatrix}. \tag{A2}$$

The prior distribution of μ can be readily deduced from this distribution, but since we wish the prior on σ^2 as well we take the following seemingly easier approach. Note that given h and θ the joint distribution of μ and σ^2 has density

$$f(\mu,\sigma^2|h,\theta)\propto h^{\frac{1}{2}}\exp\{-\tfrac{1}{2}hN(\mu-\theta)^2\}(\sigma^2)^{\frac{1}{2}\nu-1}h^{\frac{1}{2}\nu}\exp(-\tfrac{1}{2}N\sigma^2), \tag{A3}$$

where $\nu\equiv N-1$. This follows since given h and θ, $N\sigma^2 h\sim\chi^2_\nu$ independently of $(hN)^{\frac{1}{2}}(\mu-\theta)$ which is distributed as $N(0,1)$. Multiplying (A3) by (20) and integrating out h and θ one finds that the joint prior on μ and σ^2 has density

$$f(\mu,\sigma^2)\propto(N\sigma^2)^{\frac{1}{2}\nu-1}\left\{N\sigma^2+\nu'v'+\frac{n'N}{N+n'}(\mu-m')^2\right\}^{-\frac{1}{2}\{\nu+\nu'+\delta(n')\}}. \tag{A4}$$

This may be integrated with respect to μ and σ^2 as shown in Raiffa and Schlaifer (1961) and yields the result that the prior distribution of μ has a "Student" density given by

$$f(\mu)\propto\left\{v'+\frac{n'N}{(n'+N)v'}(\mu-m')^2\right\}^{-\frac{1}{2}\{\nu'+\delta(n')\}},\quad -\infty<\mu<\infty. \tag{A5}$$

Expression (22) then follows immediately. Hence

$$E(\mu)=m' \tag{A6}$$

and

$$V(\mu)=\frac{\nu'v'}{\nu'-2}\left(\frac{1}{n'}+\frac{1}{N}\right)=\left(\frac{1}{n'}+\frac{1}{N}\right)\frac{N}{N-1}E(\sigma^2), \tag{A7}$$

the last equality follows from (A9) below.

Similarly, by integrating μ out of the joint density (A4) one finds that the prior density of σ^2 is given by

$$f(\sigma^2)\propto\frac{(N\sigma^2)^{\frac{1}{2}\nu-1}}{(N\sigma^2+\nu'v')^{\frac{1}{2}\{\nu+\nu'+\delta(n')-1\}}},\quad 0<\sigma^2<\infty, \tag{A8}$$

i.e. $N\sigma^2/(N\sigma^2+\nu'v')$ has a prior beta distribution, or $N\sigma^2$ has what Raiffa and Schlaifer term an "inverted beta-2" distribution. The prior mean and variance of σ^2 are given by

$$E(\sigma^2)=\frac{\nu'v'}{\nu'-2}\,\frac{N-1}{N},\quad \nu'>2, \tag{A9}$$

and

$$V(\sigma^2)=\frac{2(\nu+\nu'-2)(\nu'v')^2}{\nu(\nu'-2)(\nu'-4)}\left(\frac{N-1}{N}\right)^2,\quad \nu'>4. \tag{A10}$$

Posterior on μ

Turning to the posterior distribution of μ we proceed as follows: given the sample $(s,\mathbf{x})$ with observed mean $\bar{x}$ and variance $s^2=(n-1)^{-1}\sum_{i\in s}(x_i-\bar{x})^2$, the posterior distribution of the superpopulation parameters (with the prior given in (20)) is also normal-gamma with density given by

$$f\{h,\theta|(s,\mathbf{x})\}\propto\exp\{-\tfrac{1}{2}hn''(\theta-m'')^2\}h^{\frac{1}{2}}\exp(-\tfrac{1}{2}\nu''v''h)h^{\frac{1}{2}\nu''-1}, \tag{A11}$$

where

$$m'' = \frac{n'm' + n\bar{x}}{n' + n}, \tag{A12}$$

$$n'' = n' + n, \tag{A13}$$

$$\nu''v'' = \nu'v' + (n-1)s^2 + \frac{nn'}{n'+n}(\bar{x} - m')^2, \tag{A14}$$

and

$$\nu'' = \nu' + \delta(n') + n - 1. \tag{A15}$$

From well-known properties of the normal-gamma distribution it follows that the posterior expectation of $\theta = E(X_i | \theta, h)$ is m'' which, being linear in $\bar{x}$, implies by the theorem of Section 2.3 that the posterior mean of μ must be of the weighted average form displayed in (11). By the same argument used in obtaining (A1), the posterior distribution of **X** is of the same form as the prior on **X** and is a degenerate N-dimensional "Student" distribution concentrated in the subspace of dimension $N-n$ where $\mathbf{S}(\mathbf{X}) = \mathbf{x}$. Equivalently, the posterior distribution of the $N-n$ unobserved X_j's is a non-singular $N-n$ dimensional "Student" distribution having a density of precisely the same form as the prior on **X**, (A1), obtained by replacing **X** by $\bar{\mathbf{S}}(\mathbf{X})$, N by $(N-n)$, and the primed parameters by the double-primed parameters given in (A12)–(A15). The derivation of the posterior distribution of μ is analogous to that of the prior. Let

$$\mu_{\bar{s}} = \frac{1}{N-n} \sum_{i \notin s} X_i$$

and

$$\sigma^2_{\bar{s}} = \frac{1}{N-n-1} \sum_{i \notin s} (X_i - \mu_{\bar{s}})^2$$

be the mean and variance of the unobserved X_i's, i.e. of the elements of $\bar{\mathbf{S}}(\mathbf{X})$. It is then clear that

$$f\{\mu_{\bar{s}}, \sigma^2_{\bar{s}} | h, \theta, (s, \mathbf{x})\} \propto h^{\frac{1}{2}} \exp\{-\tfrac{1}{2}(N-n)(\mu_{\bar{s}} - \theta)^2\}$$
$$(\sigma^2_{\bar{s}})^{\frac{1}{2}(N-n-1)-1} h^{\frac{1}{2}(N-n-1)} \exp\{-\tfrac{1}{2}(N-n-1)\sigma^2_{\bar{s}} h\}. \tag{A16}$$

Multiplying this expression by the posterior density of (h, θ) given $(s, \mathbf{x})$, (A11), and integrating out (h, θ) we find, in direct analogy with (A4), that

$$f\{\mu_{\bar{s}}, \sigma^2_{\bar{s}} | (s, \mathbf{x})\} \propto \{(N-n-1)\sigma^2_{\bar{s}}\}^{\frac{1}{2}(N-n-1)-1}$$
$$\left\{(N-n-1)\sigma^2_{\bar{s}} + \nu''v'' + \frac{n''(N-n)}{n''+(N-n)}(\mu_{\bar{s}} - m'')^2\right\}^{-\frac{1}{2}(N-n+\nu'')}. \tag{A17}$$

Integrating out $\sigma^2_{\bar{s}}$, one finds

$$f\{\mu_{\bar{s}} | (s, \mathbf{x})\} \propto \left\{\nu'' + \frac{n''(N-n)}{(n''+N-n)v''}(\mu_{\bar{s}} - m'')^2\right\}^{-\frac{1}{2}(\nu''+1)}, \tag{A18}$$

or the posterior distribution of $\mu_{\bar{s}}$ is like that of $m'' + t_{\nu''}\{(N-n+n'')v''/(N-n)n''\}^{\frac{1}{2}}$, where $t_{\nu''}$ is a standard t random variable on ν'' degrees of freedom.

Since $\mu = \{n\bar{x}+(N-n)\,\mu_{\bar{s}}\}/N$, it follows readily that the posterior distribution of μ given $(s, \mathbf{x})$ is like that of

$$\frac{n}{N}\bar{x}+\frac{N-n}{N}m''+t_{\nu''}\left\{\frac{(N-n+n'')\,(N-n)\,v''}{n''N^2}\right\}^{\frac{1}{2}}, \tag{A19}$$

and thus using (A12)–(A15)

$$E\{\mu\,|\,(s,\mathbf{x})\} = \frac{n}{N}\bar{x}+\frac{N-n}{N}m'' = \frac{n(N+n')\,\bar{x}+(N-n)\,n'm'}{N(n'+n)}, \tag{A20}$$

and

$$V\{\mu\,|\,(s,\mathbf{x})\} = \frac{N-n+n''}{n''N^2}\,\frac{(N-n)\,\nu''\,v''}{\nu''-2}$$

$$= \frac{N-n}{N^2}\,\frac{N+n'}{n+n'}\,\frac{\{\nu'v'+(n-1)\,s^2+nn'(\bar{x}-m')^2/(n+n')\}}{\nu'+\delta(n')+n-3}. \tag{A21}$$

Since the condition of the theorem of Section 2.3 holds here, we may identify (A20) and (11). Making use of the fact that $V(\mu)$ has the form in (A7) we deduce that

$$EV(\bar{X}\,|\,\mu) = \frac{N-n}{nN}\,\frac{\nu'v'}{\nu'-2}. \tag{A22}$$

Further in this case it follows from the fact that the joint prior on X is "Student" (formula (A1)) that the conditional distribution of $\bar{X}$ given μ has density

$$f(\bar{X}\,|\,\mu) \propto \left\{\nu'+\frac{nN}{v'(N-n)}(\bar{X}-\mu)^2\right\}^{-\frac{1}{2}\{\nu'+\delta(n')\}} \tag{A23}$$

and hence it follows (Raiffa and Schlaifer, 1961) that

$$V(\bar{X}\,|\,\mu) = \frac{N-n}{nN}\,\frac{\nu'v'}{\nu'-2} = EV(\bar{X}\,|\,\mu),$$

and thus the posterior mean of μ, (A20), may be recast as

$$E\{\mu\,|\,(s,\mathbf{x})\} = \frac{V(\mu)\,\bar{x}+V(\bar{X}\,|\,\mu)\,m'}{V(\mu)+V(\bar{X}\,|\,\mu)}. \tag{A24}$$

Finally, it may be observed, using (A7), (A21) and known properties of the normal-gamma distribution, that

$$V\{\mu\,|\,(s,\mathbf{x})\} = \frac{N-n}{N}\left\{\frac{\nu''v''}{n''(\nu''-2)}\,\frac{n'(\nu'-2)}{\nu'v'}\right\}V(\mu) = \frac{N-n}{N}\,\frac{V\{\theta\,|\,(s,\mathbf{x})\}}{V(\theta)}\,V(\mu). \tag{A25}$$

Part (a) of the theorem is thus established using these results and (A19).

In the special case of the diffuse prior on (θ, h) given by (21), all of the above results hold merely by taking $n' = \nu' = \delta(n') = 0$. This yields part (b) of the theorem.

Posterior on σ^2

While the posterior distribution of σ^2 is relatively messy, low moments may be found. For example, the posterior expectation of σ^2 may be found as follows: First we note that

$$N\sigma^2-(n-1)\,s^2 = \sum_{i\notin s}(X_i-\mu_{\bar{s}})^2+\frac{n(N-n)}{N}(\bar{x}-\mu_{\bar{s}})^2,$$

where $\mu_{\bar{s}}$ is the mean of the $N-n$ unobserved X_i's. Taking the expectation of this expression, first conditional on θ, h, and $(s, \mathbf{x})$ we find

$$E\{N\sigma^2-(n-1)s^2 \,|\, \theta, h, (s,\mathbf{x})\} = \frac{(N-n)(N-1)}{N}h^{-1}+\frac{n(N-n)}{N}(\bar{x}-\mu)^2.$$

Then taking the expectation of this quantity with respect to the posterior distribution of (θ, h) given $(s, \mathbf{x})$, using known properties of that normal-gamma distribution, and after some manipulation we have

$$\begin{aligned} E\{\sigma^2 \,|\, (s,\mathbf{x})\} = {} & \frac{(n-1)s^2}{N}\left[1+\frac{(N-n)\{N(n+n')-n'\}}{N(n+n')\{\nu'+\delta(n')+n-3\}}\right] \\ & +\frac{(\nu'-2)(N-n)\{N(n+n')-n'\}E(\sigma^2)}{N(N-1)(n+n')\{\nu'+\delta(n')+n-3\}} \\ & +\frac{nn'(N-n)}{N^2(n+n')^2}\frac{[n'\{\nu'+\delta(n')+n-4\}+N(n+n')]}{\{\nu'+\delta(n')+n-3\}}\{\bar{x}-E(\mu)\}^2. \qquad \text{(A26)} \end{aligned}$$

For the diffuse prior obtained by taking $\nu' = n' = \delta(n') = 0$

$$E\{\sigma^2 \,|\, (s,\mathbf{x})\} = \left(\frac{n-1}{n-3}\frac{N-3}{N}\right)s^2.$$

Discussion on Professor Ericson's Paper

Professor M. R. Sampford (University of Liverpool): It gives me great pleasure to see Professor Ericson here, and to have the opportunity of proposing the vote of thanks on his extremely interesting paper. I find myself in a slight difficulty here. By Society tradition the proposer of the vote of thanks is expected to be kind to the speaker: I thus find myself, a non-Bayesian, having to be kind to a Bayesian! Perhaps we should have a film of this rare and interesting event!

As I have said, I am not myself a convinced "Bayesian"—or, at least, not a convinced practitioner of Bayesian methods. I agree, of course, that prior beliefs cannot safely be ignored in the interpretation of experimental or survey results—no "thoughtful classical practitioner", to borrow a happy phrase from our speaker, can reasonably think otherwise. Further, I accept that the use of Bayes's theorem provides an apparently less arbitrary method of taking account of such beliefs than the classical "method" of viewing with extreme suspicion, and in sufficiently desperate situations repeating, any experiment whose results conflict too violently with one's prior notions. However, I remain unconvinced of the practical value of the Bayesian argument—at least, in the field of biological and agricultural statistics that I know best.

My reasons for remaining unconvinced about Bayesian methods in survey analysis are rather different from my reasons relating to experimentation. It is arguable that experimentation is a private activity (I am well aware that this is an oversimplification!), so that if the experimenter wishes to interpret his results subjectively, he has every right to do so, and no one should be able to quarrel with his interpretation. The difficulty here, in my experience, is that most biological experiments involve so much extraneous, uncontrollable variation that an honest prior must be so diffuse as to be of little value (and even then one must be prepared for results in a particular experiment to be so anomalous as to suggest that the experimental conditions achieved were quite different from those envisaged, so that the experimenter may, quite validly, choose to ignore them as being unrelated to his intentions, rather than to use them to modify, however slightly, his prior beliefs).

From a practical viewpoint, on the other hand, there is often a great deal more justification for adopting Bayesian methods in analysing survey data. We rather seldom find ourselves sampling from a completely unknown population; at worst we have a general idea of the sort of variability to be expected, while at best we may have either a quite accurately determined prior, based on extensive observation (as in acceptance sampling), or (as in some surveys) a great deal of information about individual units. The choice of a realistic prior may thus not present any great difficulty. However, survey sampling is an essentially public activity, and one must therefore question the social validity of any single subjective interpretation of a set of survey results, when such an interpretation may directly affect the actions, and possibly the whole lives, of so many different subjects. I think Professor Ericson agrees with me here, in that he recognizes the fact that the results of the survey need to be available, so that the user may, if he wishes, reject any published subjective conclusions, and draw his own.

I would like to comment on one particular topic in detail— that of randomization. We classical practitioners are, of course, by now quite used to hearing Bayesians blaspheming this central article of our faith, and there are no doubt others like myself present who feel that, while he is prepared to contemplate the possibility of random sampling, Professor Ericson is not beyond hope of salvation. He could have been, and I feel he should have been, more explicit on this topic, and in choosing to comment on his remarks, I have in mind that, as far as I can make out, several recent disputes in the field of sampling theory have arisen largely, if not entirely, because neither of the participants was clear what the other was trying to do.

I am a classical practitioner, and I hope a thoughtful one, and I am willing to agree, if I may quote from p. 198, that "the notion of exchangeability very closely approximates [my] opinions in many situations where [I] deem simple random sampling to be appropriate", but *only if* I am allowed to sample *at random*, or, in other words, if I am allowed to ensure exchangeability by randomizing my population unit labels before sampling. Professor Ericson allows me to do this, on the grounds that my true opinions will not in general be exactly exchangeable, but only roughly exchangeable, in which case he appears to feel that randomization provides a convenient simplification. This seems to me to be a dangerously casual way of disposing of one of the basic features of statistical variation—namely, non-independence.

In many finite populations the units follow some systematic arrangement, either spatial or temporal, and one cannot discount the possibility that pairs of neighbouring units will be more similar than pairs of remote units—in other words, that one's population displays some measure of clustering, or clumping. Stratification will, of course, go some way towards removing this effect, but often only part of the way. In fact, it is hard to envisage a situation in which I would be willing to accept Professor Ericson's arguments, and follow his methods, without prior randomization. It seems to me that, for most practical applications, the element of randomization must be regarded as an intrinsic part of this method, though he would possibly not agree. (Of course, the situation in which one has individual priors for all units is quite different: if one intends to adopt a Bayesian argument, randomization is then not merely unnecessary under the circumstances, but downright misguided—one should obviously expend one's sampling effort on the units with the most diffuse priors.)

Having spent most of my time explaining that, in the light of my own experiences, his paper seems a somewhat academic exercise, I must now conclude by saying that I found it (as such) interesting and attractive, and most enjoyable to read and to listen to, and that I admire the plausible simplicity of many of his results, and the way in which he has presented such a "solid" paper in such a short space of time. I can even envisage some situations in which I might be moved to adopt his method! It gives me great pleasure to propose the vote of thanks.

Dr A. Scott: (London School of Economics): As Godambe has been pointing out for a number of years, there are special difficulties involved in the formulation of a satisfactory theory of inference when the population consists of a finite set of identifiable elements. Professor Ericson is to be congratulated on his careful account of how these difficulties can be overcome in a Bayesian approach.

The two-stage approach to generating the prior distribution by setting up a parametric superpopulation and then choosing a prior distribution for the parameters of the superpopulation brings out very clearly the natural correspondence between sample survey work and the mainstream of statistical theory. This is an attractive feature and perhaps might help to end the comparative neglect of the survey field by most mathematical statisticians. Some common ratio and regression estimates have been shown to arise as special cases of ordinary regression theory, and the extension of the work in Section 3 to stratified sampling can be achieved easily using results for the one-way analysis of variance.

An important topic not covered in the paper is that of multi-stage sampling which is a natural analogue of the random effects model on the analysis of variance. Smith and myself have recently done some work exploiting this correspondence.

As Professor Ericson remarks, the concept of a superpopulation has been widely used in survey work, both for comparing the performance of estimates and particularly in analytic surveys to make comparisons between the characteristics of sub-groups. Once the superpopulation model has been set up, it seems more meaningful to work with the parametric likelihood generated by the model instead of the rather sterile finite population likelihood given in (1).

A model similar to that of Section 4, in which the X's assume only a finite set of values and the element labels carry no information, has recently been considered by Hartley and Rao (1968). Their approach is a more conventional one based on simple random sampling which leads to the generalized hypergeometric distribution of (34). Given the close relationship between exchangeability and simple random sampling and the intuitive interpetation of exchangeability as an effective pre-randomization, it seemed rather surprising that the result of (34) should depend on the particular prior distribution chosen. In fact, it is easy to see that the result is much more general and holds for any prior distribution. This seems more natural and so, in a sense, the approach is similar to the approach of Hartley and Rao. They are mainly concerned with finding UMV and maximum-likelihood estimates but do have a section on a Bayesian approach with a prior distribution for the hypergeometric parameter N and the prior they choose is also of the Dirichlet multinomial type. They derive expressions for the posterior mean and variance of the moments of the finite population, so they are actually dealing with things rather similar to Professor Ericson. I think that their results are equivalent to those given in (43) for the posterior mean of the finite population, and in (48) for the variance of the finite population.

A number of people working on problems of sampling inspection have also been interested in Bayes's estimates for the hypergeometric distribution, most notably Hald (1960), who gives a number of interesting results including some for the Pólya distribution which contains the Dirichlet multinomial as a special case. These results are for $k \leqslant 2$ but the extension to $k > 2$ is fairly simple.

A final comment on the area of estimation: it seems to me that if it is expected that works such as these are going to make an impact on people who actually carry out surveys, there needs to be work done on making the technical details more palatable. The problems of estimation seem relatively simple when compared with the problems of design. There has never been a satisfactory Bayesian approach to design problems, and Professor Ericson has made an important first step on the problem in the discussion in Section 2 and he hints he can suggest good reasons for a Bayesian to draw a stratified random sample when his opinions about the elements within a stratum are roughly exchangeable. I look forward with interest to the paper he promises on the subject. This is only the very simplest case, although it is an important subject, and there is a long way to go. Variable probability sampling methods have been used widely without any noticeable dissatisfaction amongst users. It seems a pity there has been no discussion at all from a Bayesian viewpoint so far on some of the more complex design problems—perhaps the author has some comments on this.

I would like to thank him very much for an important attack on a difficult problem, and I therefore have much pleasure in seconding the vote of thanks.

The vote of thanks was put to the meeting and carried unanimously.

Professor M. STONE (University College London): A little mystery has been cleared up by the reference to Karl Pearson in tonight's paper. For some time, newcomers in our Department at University College have been intrigued by the source of a large framed likelihood function on one of the walls.

May I join in the surprising reference game by mentioning Pearson and Sukhatme (1935), who derive (24) in fiducial and confidence interval terms?

Professor Ericson's message is beautifully clear—anything that classical statistics can do, Bayesian statistics can do better. I want to fight a small rearguard action in the context of Karl Pearson's example, which may be formulated as the problem of estimation of the total number of affected individuals in a labelled population of size N when we are allowed to take a sample of size n.

On the question of randomization, Professor Ericson suggests that we randomize when it does *not* matter, that is, when prior distribution is nearly exchangeable. The prior distribution then becomes perfectly exchangeable and may be employed without much average loss of information relative to the direct use of the nearly exchangeable prior.

This suggestion is in apparent contrast to Fisher's principle which is to randomize when it *is* likely to matter and when it is not known which systematic design will provide proper safeguards. But harmony prevails when the insightful Bayesian remarks that exchangeability is the Bayesian equivalent of Fisher's conditions.

But the harmony is disrupted when it comes to analysis. Suppose, in our example, the sample of size n were obtained by random sampling and that all n individuals were found to be affected. Let H_1 and H_2 denote the two extreme hypotheses that, in addition to the observed data, all the unobserved individuals are affected and unaffected respectively. For the Bayesian, the posterior odds for H_1 against H_2 equal the prior odds. Your bets would be the same whether you were to make a conditional bet on H_1 or H_2 either before or after the data.

The Fisherian approach is, as one might expect, full of ambiguity. The likelihood ratio is unity so randomization tests of H_1 and H_2 might be envisaged:

$$P(\text{number affected in } \textit{random} \text{ sample} \geqslant \text{number affected in actual sample} \mid H_1) = 1,$$

$$P(\text{number affected in } \textit{random} \text{ sample} \geqslant \text{number affected in actual sample} \mid H_2) = 1\Big/\binom{N}{n}.$$

So the *data* would speak against H_2 in favour of H_1, the remarkable nature of the actual sample in relation to H_2 being overlooked. Now I think that there is something to be said, once we are off the Bayesian turf and into the area of scientific research, for using randomization in this way to punish hypotheses such as H_2 for their *ad hoc* appearance.

Professor D. V. LINDLEY (University College London): Tonight's paper is an important contribution to our understanding of both Bayesian methods and sampling from a finite population. In particular it highlights the significant role played by a prior distribution even when it superficially appears uninformative. It also demonstrates a substantial difficulty in the Bayesian approach, namely the expression of these prior ideas in a realistic and yet tractable form: in particular I notice that Ericson experiences the same difficulty as myself in the unsuitability of the Dirichlet prior in adequately conveying the full impact of the initial knowledge. A suitable prior is desperately needed here, for example, in problems of medical diagnosis.

Tonight I should like to describe briefly a situation in which the exchangeable prior is unsuitable and the prior knowledge is of a different type. The work is being done jointly with P. Brown. A chemical can be described by a sequence of n, say, zeros and ones: a one in the rth place indicating that a particular chemical bond is present, a zero similarly signifying its absence. Conversely each sequence describes a possible chemical (Table 1).

TABLE 1

Chemical	*Activities*		
	α	β	γ
000	0	0	1
001	0	1	0
010	0	0	1
100	0	0	0
011	0	0	0
101	0	0	0
110	1	0	0
111	1	1	0

Consider the finite population of 2^n chemicals: these are Ericson's N distinguishable elements and they may easily be enumerated by translating the sequence into binary arithmetic. Any one of these chemicals can be tested for biological activity: in the simplest situation the result can be a zero or a one corresponding to the absence or presence of activity. These are the characteristics, X_i, of the paper. To illustrate consider the simplest non-trivial case, $n = 3$. Table 1 enumerates the eight chemicals and three possible types of biological activity. The first, designated α, is chemically easily explained by activity being caused by the simultaneous presence of the first two chemical bonds. However β has no such simple chemical explanation. Nevertheless, α and β are permutations one of another. Thus we see that exchangeability is unreasonable here because the chemist has *a priori* reasons for thinking that a simple explanation such as is associated with α is more probable than a complex, artificial one of the β type.

What we have done is to express our ideas in terms of *keys*. Thus α-activity is associated with a two-factor key: β-activity is not. We can, hopefully, express our prior beliefs about the various types of key, recognizing that for a large value of n many keys may operate simultaneously. Within a key some sort of exchangeability may be assumed: thus γ lists a two-factor key depending on the absence of the first and third bonds, and has the same prior probability as α. The posterior distribution has the same form as Ericson's equation (3), namely zero for many keys, with the others all increased by the same factor. Far from being unreasonable, this seems to capture the essence of the problem.

The practical object of this exercise is to help the scientist decide which chemical to investigate with the greatest chance of biological activity. That is, to find s, $1 \leqslant s \leqslant N$, with maximum $P(X_s = 1)$. This has previously been done by some form of cluster analysis: the device of keys gives an alternative approach.

Mr T. M. F. SMITH (University of Southampton): I would like to add my congratulations to Professor Ericson for his interesting paper. Buried deep in the mathematics are some useful results for practical as well as theoretical statisticians. The key to Professor Ericson's approach is to consider the finite population as a sample from an infinite superpopulation. In addition to this, as a Bayesian he requires the specification of a prior distribution, and this assumes, to quote the paper, "strong prior knowledge regarding the shape of the finite population distribution by taking the superpopulation form as known". In the paper this problem is tackled by setting up a multinomial model, which entails some rather complicated mathematics.

Many survey practitioners find the concept of prior distributions hard to swallow while accepting the usefulness of the superpopulation, as is evidenced by the quotations in the paper referring to the use of superpopulations. An alternative to using prior distributions within the superpopulation framework is to adopt an approach similar to that of the Gauss–Markov theorem and to consider only those estimates which are linear functions of the observations. Taking the superpopulation model, with similar assumptions to Professor Ericson's, and assuming only that the mean and variance are finite, Dr Scott and myself have shown that the simple random sampling mean, $\bar{x}$, has minimum mean-squared error among the class of all linear estimates with bounded mean-squared error. As in this paper the result is independent of the sampling design.

The assumption of bounded mean-squared error is an assertion of vagueness about the possible values of the parameter on a par with the use of diffuse prior distributions in the Bayesian model, and it is not surprising that for both the normal and the multinomial models considered in the paper diffuse prior knowledge leads to the simple random sampling mean, $\bar{x}$, as the estimate. Survey practitioners in many situations use $\bar{x}$ and hence are acting as if they had diffuse prior knowledge.

If we look at the estimate suggested by Professor Ericson, given in (11), this requires the subjective value of the prior mean of μ. Every user of a sample survey would have to insert his own subjective prior value; in other words, every user would use a different estimate in this particular situation. In my experience those carrying out survey research and actually doing the estimation procedure are not the same people as those who actually use the estimates as the end-product. You employ market researchers to carry out your survey and they put the data on a computer and publish the estimates. It seems to me that if you wish to have an estimate which is useful to a large number of users, the only sensible assumption to make about prior knowledge is to assume diffuse prior knowledge and hence to use $\bar{x}$ as the estimate. In this sense Professor Ericson has reinforced what is often standard survey sampling practice. But it is not always standard practice because sample survey practitioners sometimes use a design dependent estimate, which may differ from $\bar{x}$, and if the assumptions of this paper are satisfied they would be better to use $\bar{x}$. So this paper is also a contribution towards changing standard practice.

Professor D. F. Kerridge (University of Aberdeen): I want to say something in defence of the Bayesian point of view. To start with, let us look at the basic problem.

The practical user who first meets the problem of sampling from a finite population cannot see the theoretical difficulties involved. Imagine a simple finite population consisting of 1,001 animals, of which 1,000 are mice weighing 1 oz., and one is an elephant, weighing 10 tons. Take a random sample of 10 from this population, and you will see that there is a 99 per cent probability that you get an average weight of 1 oz. and just 1 per cent probability that you get an average of 1 ton. Since the true average is a hundredth of a ton, neither answer is very helpful.

Clearly, therefore, if you are to make sense of an inference, you have to impose some constraints on your problem, such as "no elephants allowed". There are two ways of doing this. The frequentist introduces an imaginary superpopulation and says "my sample of animals is drawn from a superpopulation of mice", or perhaps one containing a certain proportion of elephants. The alternative is to introduce Bayesian prior probabilities.

What strikes me as strange about the frequentist approach is their willingness to admit this superpopulation at all because on standard frequentist principles, as applied rigidly in, for example, confidence interval estimation, there is only one true mean, not a distribution of means. Although this works in sampling from an infinite population, it fails to produce sensible results with a finite population, so the frequentist has to admit the imaginary superpopulation. I cannot see how this is very different from using a prior distribution. It expresses the same knowledge, the same story as to how the population of the animals arose, and it in fact expresses a strong opinion about the presence or absence of elephants. If that is not Bayesian prior probability, I do not know what is.

The following contributions were received in writing, after the meeting:

Professor V. P. GODAMBE (University of Waterloo, Ontario): I congratulate Professor Ericson for his excellent paper which sets out in many details the implications of the Bayesian viewpoint for survey-sampling.

I entirely agree with Professor Ericson (Section 2.2) that the likelihood function (1) of Section 1, which I (1966, 1968, 1969) have been advocating is "general and virtually free from any *subjective* element". This likelihood function is, however, *individualistic* in contrast to the conventional (Royall, 1968) likelihood function which is *frequency* oriented. This can be further elaborated as follows. The conventional frequency statement

$$P(A \mid B) = r, \tag{I}$$

asserts that "in the class B of the individuals (i, say) the proportion of the individuals possessing the property A is r". It is important to note that the statement (I) does not speak about any particular individual i. On the other hand, de Finetti's (1937) probability theory replaces statement (I) by

$$P(i \text{ is an } A \mid i \text{ is a } B) = r, \tag{II}$$

which is an assertion about the individual i. Now let X_i be the variate value associated with the individual i, in the population of N individuals, $i = 1, ..., N$. Further we write $\mathbf{X} = (X_1, ..., X_N)$. The *conventional* (Royall, 1968) statement,

$$P(X = \alpha \mid \mathbf{X}) = r, \tag{III}$$

(analogous to the statement (I) above), asserts that "the proportion of X_i's in $\mathbf{X}$ which are equal to α is r". Further interpreting randomization more or less the same way as in the Section 2.4, if $\mathbf{X}$ is unknown, the statement (III) defines the likelihood function for $\mathbf{X}$ given that the randomly drawn X (we assume for simplicity that only one draw was made) was equal to α. This likelihood function, I would say, is *frequency* oriented as it is based on the probability statement (III) which (analogous to the statement (I)) says nothing at all about any specific individual i of the population. However, one can replace the probability statement (III) by

$$P(i \text{ and } X_i = \alpha \mid \mathbf{X}) = 1/N \quad \text{or} \quad 0 \quad \text{according as } X_i = \alpha \text{ or } \neq \alpha. \tag{IV}$$

This statement is clearly analogous to the statement (II) above and it gives the likelihood function which is a special case of (1) of Section 1. This likelihood function, I say, is *individualistic* as it is derived from the probability statement (IV) which, unlike (III), speaks about the specific individual i. Yet, of course, the likelihood function (1) of Section 1 is virtually objective as Professor Ericson correctly points out. The two statements (III) and (IV) above can imply surprisingly different logical conclusions. This is best illustrated by the demonstration (Godambe, 1955) of the non-existence of UMV estimation.

Now I come to the second point. Again I believe Professor Ericson has rightly noted (Section 2.2) the *similarities* of a subjective exchangeable prior and objective distributions induced by simple random sampling. The more I meditate about it the more I realize the profundity of the issue involved here. A very commonly shared feeling about *randomization* is that it provides some protection to the *inference* in case the underlying *assumptions* go wrong. In other words, randomization renders inference more *robust*. Possibly one explanation of this robustness, as pointed out by Professor Ericson (Section 2.4), is that randomization safeguards the assumption of the exchangeability of the subjective prior distribution. A second explanation which utilizes the *similarities*, referred to above, of a subjective prior and the sampling distribution obtained by randomization is as follows. Suppose on the basis of some subjective prior distribution ξ the statistician makes a posterior probability statement (about the unknown finite population mean μ),

$$P_\xi(\alpha_1 > \mu > \alpha_2 \mid \text{data}) = r, \tag{V}$$

α_1, α_2, r being some specified numbers. Of course almost never can the statistician be *sure* that some specific prior distribution ξ *truly* represents his prior knowledge. However, in some situations he can be fairly *certain* that his true prior distribution belongs to a certain class C of prior distributions. On the other hand, the statistician may consider some specified ξ in C as the most plausible or even most convenient distribution to work with. Hence he would assert the statement (V) above but would like his statement to be protected in some sense in case ξ is not the true prior. This, sometimes, can be done by introducing appropriate randomization, i.e. appropriate sampling design. For instance, the above-referred-to class C of priors may imply (or approximately) that the finite population $\mathbf{X} = (X_1, \ldots, X_N)$ belongs to a *subset B* of the Euclidean N-space. With the appropriate sampling design then one may (in a frequency sense) be able to assert

$$P(\alpha_1 > \mu > \alpha_2 \mid \mathbf{X}) = r \text{ for all } \mathbf{X} \in B. \qquad \text{(VI)}$$

Now the validity of (V) is independent of sampling design, however, the validity of (VI) essentially depends upon the appropriate sampling design. Hence if the statistician due to incorrect analysis of his prior knowledge assumes ξ to be his prior distribution and hence makes the implied incorrect assertion (V), numerically his assertion would still be true in frequency sense as (VI), provided he had used some appropriate sampling design. Thus, the frequency assertion (VI), though weak for a single case situation, still renders the Bayesian assertion (V) considerably robust. This robustness, as we have noted, depends upon the use of some appropriate sampling design. Interestingly, however, the appropriateness or otherwise of a sampling design is determined by the prior distribution ξ. I have worked out in detail some illustrative examples which I hope to be able to publish soon. Earlier comparisons of frequency and Bayes inferences are due to Bartholomew (1965).

Professor L. Kish (University of Michigan): This is a substantial contribution to justify and explain in Bayesian terms the role of randomization in sampling from a finite population of distinguishable elements. The need to find that role had recently become a dilemma and embarrassment for thoughtful Bayesians. It was perhaps only in less thoughtful and more playful moods that the need was denied.

Exchangeability (or something much like it) can become a strong link between Bayesian and "classical" theory. To assess exchangeability each must ultimately utilize mixtures of empirical evidence and of judgement, and randomization serves as means to that end. First, note that randomizing operations and tables of random numbers must ultimately be accepted or rejected by (personal) judgement. Second, a vast portion of research evidence, even today and tomorrow, must be accepted, analysed and inferred from mere assumptions of exchangeability and without actual randomization.

On the other hand, randomization is not merely one but *the primary* means to exchangeability. The statements in Section 2.4 bear stronger emphasis and elucidation. First, without randomization the existing order of labels of real populations never justifies exact exchangeability, only approximations, with varying and vague risks. Second, statistics in research and in much else must be public, and so must be exchangeability. Third, statistics and sampling cannot exist without economics; the costs of randomization and its alternatives must be considered when choosing between them.

Real statistics for the real world is the motivation, I think, for the Bayesian departure from the clear simple world of sampling distributions. It is a difficult path, as this paper shows. In survey sampling we must face further complexities due to stratification, unequal selection probabilities, and especially cluster and multi-stage selection. Can these be best handled with complex prior distributions, or with modifications of the likelihood function?

To be useful for practice and teaching the theory must be simplified somehow. I even wonder if exchangeability can some day yield a new simple and adequate theory of (or similar to) sampling distributions. The problems and difficulties discussed here are not confined to survey sampling, but exist also for experimental design and other forms of statistical investigations. None of these can be satisfied with merely the means of simple random samples.

The author replied briefly at the meeting and subsequently more fully in writing, as follows:

First, I would like to thank the discussants for their comments.

Some issues raised by several speakers might be clarified somewhat if I briefly indicated some of the motivation underlying this work. Generally, the impetus arose from a desire to find a Bayesian framework or model within which to consider problems of inference and design in sampling finite populations. Such a model had to yield "sensible" answers in the simplest cases (those treated in the present paper) and yet be potentially flexible enough to encompass more complex practical situations. These criteria led to the general avoidance of the class of priors which take the components of $\mathbf{X}$ to be independent. Professor Stone's example seems to be the result of such a prior; when the prior on $\mathbf{X}$ is taken as in (26) with a rectangular distribution on p the likelihood ratio and the posterior odds in favour of his H_1 and against H_2 are simply $\binom{N}{n}$. They also resulted in the retention of the unit labels in the basic model put forth in the first three pages of the paper, even though under an exchangeable prior the labels are totally irrelevant for most practical purposes.

I felt that in many situations more informative inferential statements were possible from finite population samples than those which traditionally characterize this area. I also suspect that these include cases which are "public" in the spirit of comments by Professors Sampford and Kish and Mr Smith. The Bayesian approach seems suited to providing this as well as a natural approach for formal consideration of "non-sampling" errors, e.g. response error, bias and non-response. Some work by Mr C. T. Tharakan is in progress on these topics. No matter what point of view one takes on the inference question, the Bayesian approach seems ideal for handling questions of optimal design and sample size, questions where decisions must be made and where prior information and economic considerations clearly dominate. While the present paper has concentrated more on simple inference problems, some aspects of optimal stratified sample design have been treated in Ericson (1969a).

My comments on randomization are not, in any sense, intended to be complete but rather indicative of one possible factor contributing to a Bayesian role for randomization—a part which becomes less important as prior information increases.

In reply to Professor Sampford, I do feel that when my prior is roughly exchangeable, randomization provides a convenient simplification without much loss of information. On the other hand, if I had information leading me to believe that the unit values followed some systematic pattern I would then certainly exploit this by incorporating it into my prior distribution. In many cases this would lead to a non-exchangeable prior and would result in some restricted role for randomization. In any event, under the general model put forth there is *always* a (marginal) prior on each X_i and a reasonable optimal design criterion would certainly lead one to sample those units effecting the greatest expected reduction in posterior variance. However, there may be many equivalent best designs under this criterion and randomizing among them would seem judicious.

I certainly agree with Dr Scott that equation (34) is independent of $f'(\mathbf{p})$. The return reference provided by Professor Stone is equally appreciated. I tend to agree more with Professor Kerridge than with Mr Smith on the place of the superpopulation concept. From my point of view it often serves as a convenient artifact, but only as a halfway house in establishing an exchangeable prior.

Finally, I look forward to seeing more details on the example put forth by Professor Lindley and on Professor Godambe's robustness studies.

References in the Discussion

Bartholomew, D. J. (1965). A comparison of Bayesian and frequentist inferences. *Biometrika*, **52**, 19–35.

Ericson, W. A. (1969a). Subjective Bayesian Models in Sampling Finite Populations: Stratification. to appear in the *Proceedings of a Symposium on the Foundations of Survey Sampling* held in April, 1968, at Chapel Hill, to be published by John Wiley and Sons.

Godambe, V. P. (1968). Bayesian sufficiency in survey-sampling. *Annals of the Institute of Statistical Mathematics*, **20**, 363–373.

—— (1969). The fiducial argument with application to survey-sampling. *J. Royal Stat. Soc.* B, **31**, (in press).

Hald, A. (1960). The compound hypergeometric distribution and a system of single sampling inspection plans based on prior distributions and costs. *Technometrics*, **2**, 275–340.

Hartley, H. O. and Rao, J. N. K. (1968). A new estimation theory for sample surveys. *Biometrika*, **55**, 547–557.

Pearson, E. S. and Sukhatme, A. V. (1935). An illustration of the use of fiducial limits in determining the characteristics of a sampled batch. *Sankhya*, **2**, 13–32.

Royall, R. (1968). An old approach to finite population sampling theory. *J. Amer. Stat. Assoc.*, **63**, 1269–1279.