NOTICE

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely related Government procurement operation, the United States Government thereby incurs no responsibility nor any obligation whatsoever; and the fact that the government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication or otherwise as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

This report has been reviewed by the Information Office (OI) and is releasable to the National Technical Information Service (NTIS). At NTIS, it will be available to the general public, including foreign nations.

This technical report has been reviewed and is approved for publication.

H. Leon Harter

H. LEON HARTER
Project Engineer

FOR THE COMMANDER

Gerald G. Leigh

GERALD G. LEIGH, Lt. Col, USAF
Chief, Structures Division

Copies of this report should not be returned unless return is required by security considerations, contractual obligations, or notice on a specific document.

AIR FORCE - 17 MAY 76 - 400

Technical Report No.

[ACTP...] H430 649 (62D)

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: AFFDL-TR-75-145.
2. GOVT ACCESSION NO.:
3. RECIPIENT'S CATALOG NUMBER:
4. TITLE (and Subtitle): A Bayesian Approach to Two-stage Sampling.
5. TYPE OF REPORT & PERIOD COVERED: Final. 1 May 1972 - 1 January 1976
6. PERFORMING ORG. REPORT NUMBER:
7. AUTHOR(s): W. A. Ericson
8. CONTRACT OR GRANT NUMBER(s): F33615-72-C-1213
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Department of Statistics, The University of Michigan, Ann Arbor, Michigan 48109
10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: DoD Element 681304, 70710201
11. CONTROLLING OFFICE NAME AND ADDRESS: Air Force Flight Dynamics Laboratory, Air Force Wright Aeronautical Laboratories, Air Force Systems Command, Wright-Patt. AF Base, Ohio
12. REPORT DATE: February 1976
13. NUMBER OF PAGES: 26
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): distributed by National Technical Information Service, Springfield, Va., 1976. v+26pp. $3.00; $0.95 microfiche.
15. SECURITY CLASS. (of this report): Unclassified
15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:
16. DISTRIBUTION STATEMENT (of this Report): Approved for public release; distribution unlimited
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report):
18. SUPPLEMENTARY NOTES:
19. KEY WORDS (Continue on reverse side if necessary and identify by block number): Bayesian, Exchangeability, Two-stage, Sampling, Best Linear Unbiased Estimators, Weighted Least Squares, Finite Populations
20. ABSTRACT (Continue on reverse side if necessary and identify by block number):

In several previous papers, the author has shown that various standard sampling designs are optimal in a Bayesian sense under corresponding classes of prior distributions on the N-dimensional vector of unknown characteristics of the N elements of a finite population. In this manner a Bayesian interpretation of simple random sampling, stratified random sampling, and of various ratio and regression estimators have been given. In the present report this work is extended to two-stage balanced sampling.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. Additionally, a simple result on a representation of finitely exchangeable discrete random variables is given which gives a slight generalization of a seemingly little-known result of de Finetti. Also a general tie between Bayes posterior means and traditional WLSE's and BLUE's is obtained, generalizing previous results given by the author.

PREFACE

This report, prepared by Professor W. A. Ericson, is one of two parts of the final report on Contract No. F33615-72-C-1213, which was funded by the Aerospace Research Laboratories and was technically monitored by Dr. David A. Harville (until 30 June 1975) and Dr. H. Leon Harter (since 1 July 1975). The other part, prepared by Professor Bruce M. Hill and entitled "Exact and Approximate Bayesian Solutions for Inference About Variance Components and Multivariate Inadmissibility," has appeared as AFFDL-TR-75-134.

The author is grateful to H. Bühlmann for pointing out reference [11] and to R. A. Olshen for bringing reference [10] to his attention as well as reading Section 2.

TABLE OF CONTENTS

1. Review of Preliminaries

The basic model of a finite population is taken to be that which was initially put forth and subsequently used by Godambe, [1], [2] and [3] and others. The author, Ericson [4], has presented a general Bayesian view of inference regarding simple characteristics of a finite population based on this model. Essentials of that earlier work are first reviewed as a basis for the new results which are to be given in the present report. Further details may be found in the earlier papers.

1.1 The Basic Model: A finite population of N elements or units is defined by letting $\underline{N} = \{1, 2, \ldots, N\}$, the label set, and $\underline{X} \equiv (X_1, X_2, \ldots, X_N)$, where X_i is the unknown value of some characteristic possessed by the ith population element. Inference concerns the N-dimensional real vector parameter $\underline{X}$ or, more realistically, some simple function $g(\underline{X})$ of $\underline{X}$. Here the primary concern will be with μ, the population mean.

A sample of size m, s^*, is defined quite generally to be an ordered sequence of m of the population elements $i_1^*, i_2^*, \ldots, i_m^*$ ($i_j^* \in \underline{N}$, $j = 1, \ldots, m$, repetitions allowed) together with the sequence of their associated observed characteristic values $\underline{x}^* = (x_{i_1^*}, x_{i_2^*}, \ldots, x_{i_m^*})$ i.e., one observes for each i_j^* that $X_{i_j^*} = x_{i_j^*}$, $j = 1, 2, \ldots, m$.

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 $p(s^*) \geq 0$, $\sum_{s^* \in S^*} p(s^*) = 1$, where $p(s^*)$ is the probability of choosing the sample s^*. Such a definition of a sample design is extremely general, covering most known and used designs.

For any such sample $(s^*, \underline{x}^*)$ define the statistic $(s; x_i, i \in s)$ to be the set of distinct population elements $s = \{i_1, \ldots, i_n\} \subseteq \underline{N}$ 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 $s = \{i_1, \ldots, i_n\}$ define the (matrix) operator S such that $S(\underline{X}) = (X_{i_1}, \ldots, X_{i_n})$, for definiteness assuming $i_1 < i_2 < \ldots < i_n$; the complementary operator $\bar{S}$ such that $\bar{S}(\underline{X}) = (X_{j_1}, X_{j_2}, \ldots, X_{j_{N-n}})$ for all $j_i \in \underline{N} - s$; $(j_1 < j_2 < \ldots < j_{N-n})$; and the vector $\underline{x} = (x_{i_1}, \ldots, x_{i_n})$ of observed values of $S(\underline{X})$.

It is obvious under this model that for any joint prior probability distribution on the vector parameter $\underline{X}$ given by the general density $p'(X_1, \ldots, X_N)$ the posterior probability distribution of $\underline{X}$ given the sample $(s^*, \underline{x}^*)$ is precisely the same as that given only the statistic $(s, \underline{x})$. Thus by the Bayesian definition of sufficiency $(s, \underline{x})$ is a sufficient statistic, a fact previously demonstrated or noted using equivalent definitions by Basu [5] and others. It is further clear that given $(s^*, \underline{x}^*)$ the likelihood function of $\underline{X}$ is given by

$$\ell(\underline{X}; (s^*, \underline{x}^*)) = \ell(\underline{X}; (s, \underline{x})) = \begin{cases} k & \text{for } \underline{X} \mid S(\underline{X}) = \underline{x} \\ 0 & \text{otherwise} \end{cases}, \quad (1)$$

where $k > 0$ is an arbitrary constant (it, of course, being assumed that $p(s^*)$ is independent of $\underline{X}$).

Finally observe that given any joint N-dimensional prior distribution on $\underline{X}$, $p'(X_1, \ldots, X_N)$, the posterior distribution on $\underline{X}$ given a sample of the sort described above is, from (1), given by the general density

$$p(\underline{X} \mid (s, \underline{x})) = \begin{cases} k\, p'(\underline{X}) & \text{for } \underline{X} \mid S(\underline{X}) = \underline{x} \\ 0 & \text{otherwise} \end{cases}, \qquad (2)$$

or

$$p(\underline{X} \mid (s, \underline{x})) = \begin{cases} p'(\underline{X}) \,/\, p'_{S(\underline{X})}(\underline{x}) & \text{for } \underline{X} \mid S(\underline{X}) = \underline{x} \\ 0 & \text{otherwise} \end{cases}, \qquad (3)$$

where $p_{S(\underline{X})}(\underline{x}) \neq 0$ is just the marginal prior density of $S(\underline{X})$.

Even for staunch supporters of the likelihood approach to inference or of the likelihood principle the likelihood function in (1) above seems alarming, for only the likelihood function which is constant over the whole parameter space would seem less informative than (1). This has resulted in various recent investigations of methods of inference for finite population parameters which are not in harmony with the likelihood principle.

Even though the likelihood function, under the present model, only partitions the parameter space it has been shown that for various classes of prior distributions of $\underline{X}$ one obtains very reasonable posterior inferences on simple functions, $g(\underline{X})$, of $\underline{X}$ which seem to agree with the best of traditional survey sampling results when the prior is suitably diffuse. Such results are given in [4] and [6].

It is the author's view that the problem of inference in a finite problem of inference in a finite population model is well within the capability of modern Bayesian views and that the basic problem is the choice of the prior distribution, $p'(\underline{X})$. While this point of view seems satisfactory for personal inference about $g(\underline{X})$ the approach promises to be even more useful in resolving questions of optimal sample designs and sizes which formally recognize the fact that practical choice of such designs and sizes must be, and is, dictated by prior information and economic constraints.

1.2 Exchangeability and the Choice of a Prior:

The simplest useful class of prior distributions is that under which the X_i's are taken to be exchangeable or symmetrically dependent. Such a class of priors was central in the author's earlier papers and will also play an important role here. A brief review of this concept is presented for completeness.

The random variables $X_1, \ldots, X_N$ are said to be exchangeable if the joint probability distribution of each of the $N!$ permutations of the variables is the same. 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 $1 \leq r \leq N$ one's initial betting behavior regarding events defined by the unknown quantities $X_{i_1}, \ldots, X_{i_r}$ is the same for every ordered set of indices $i_1, \ldots, i_r$.

There are close ties between an objectivistic sampling distribution under simple random sampling and corresponding properties of subjectivistic exchangeable prior distributions. For it may be shown that it follows strictly from an exchangeable prior distribution on $\underline{X}$ that given the collection of the N population variate values, but not the units to which they are attached, the subjective probability that any prespecified subset of n of the population elements will assume the values given by any collection of n of the N population values is precisely the same as the objective probability that the subset was selected by simple random sampling, namely, $1/\binom{N}{n}$. These ties have been examined and a Bayesian argument for the use of random sampling has been presented by Ericson, [6].

2. A Representation of Finite Discrete Exchangeability:

It is well-known (deFinetti's Theorem) that if $X_1, X_2, \ldots$ is an infinite sequence of exchangeable random variables then the joint distribution of any subset of any n of the X_i's can be represented as a probability mixture of independent, identically distributed random variables (Hewett and Savage, [7]; Feller, [8]). It is also known (Feller, [8]) that this representation breaks down for finite sequences of exchangeable random variables. Despite this failure a useful sub-class of finite sequences of exchangeable random variables can be generated via the expression

$$p(X_1, X_2, \ldots, X_N) = \int \prod_{i=1}^{N} p(X_i \mid \xi)\, dF(\xi) . \qquad (4)$$

The class of prior distributions generated in this fashion was extensively used in the author's earlier papers.

There is an analog to de Finetti's theorem giving a representation for finite sequences of discrete random variables. This result seems to be little-known and turned out to have been first observed by de Finetti in [9]. The following version of this result, a very slight generalization of de Finetti's, deserves to be better-known and promises to be useful in characterizing exchangeable random variables having various other properties.

Let $X_1, X_2, \ldots, X_N$ be any finite sequence of discrete exchangeable random variables assuming values in $\chi = \{x_1, x_2, \ldots\}$. Let N_j equal the number of the N X_i's which equal x_j, $j = 1, 2, \ldots$; let $\underline{N} = (N_1, N_2, \ldots)$; $\eta = \{\underline{N} \mid N_j \geq 0, \sum_{j=1}^{\infty} N_j = N\}$; $\underline{X} = (X_1, X_2, \ldots, X_n)$ and $\underline{x} = (x_{i_1}, x_{i_2}, \ldots, x_{i_n})$ be any point in χ^n. With these definitions one then has the following result.

Theorem (de Finetti): If $X_1, X_2, \ldots, X_N$ is any finite sequence of exchangeable discrete random variables having probability mass function, P^*, then there exists a probability mass function $P(\underline{N})$ on η such that for all n, $0 < n \leq N$, and $\underline{x} \in \chi^n$

$$P^*(\underline{X} = \underline{x}) = \sum_{\underline{N} \in \eta} \frac{\prod_{j=1}^{\infty} (N_j)_{n_j}}{(N)_n} P(\underline{N}) \qquad (5)$$

where $(A)_b = A(A-1)\ldots(A-b+1)$ and n_j is the number of x_{i_k}'s $(k = 1, \ldots, n)$ equal to x_j.

Proof: Let $\underline{I}_1 = (1, 0, 0, \ldots)$, $\underline{I}_2 = (0, 1, 0, \ldots) \ldots$ and let $\underline{X}_i$ be an infinite dimensional random variable having range $I = \{\underline{I}_j \mid j = 1, 2, \ldots\}$ such that $X_i = x_{i_j}$ if and only if $\underline{X}_i = \underline{I}_{i_j}$. It then follows that the N $\underline{X}_i$'s are exchangeable and

$$P(\underline{X} = \underline{x}) = P(\underline{X}_1 = \underline{I}_{i_1}, \ldots, \underline{X}_n = \underline{I}_{i_n}). \qquad (6)$$

Let $\underline{n} = (n_1, n_2, \ldots)$ and $E_{\underline{n}}$ denote the event on the right side of (6). It is immediate that

$$P(E_{\underline{n}}) = \sum_{\underline{N} \in \eta} P(E_{\underline{n}} \mid \sum_1^n \underline{X}_i = \underline{n} \text{ and } \sum_{i=1}^N \underline{X}_i = \underline{N}) \, P(\sum_{i=1}^n \underline{X}_i = \underline{n} \mid \sum_{i=1}^N \underline{X}_i = \underline{N}) \, P(\underline{N}) \qquad (7)$$

Now conditional on $\sum_{i=1}^n \underline{X}_i = \underline{n}$ and $\sum_{i=1}^N \underline{X}_i = \underline{N}$ the $\underline{X}_i$'s are exchangeable and hence it follows that

$$P(E_{\underline{n}} \mid \sum_{i=1}^n \underline{X}_i = \underline{n} \text{ and } \sum_{i=1}^N \underline{X}_i = \underline{N}) = \left(\frac{n!}{\prod_{j=1}^{\infty} n_j!}\right)^{-1}. \qquad (8)$$

Also the $\underline{X}_i$'s are exchangeable conditional on $\sum_{i=1}^{N} \underline{X}_i = \underline{N}$ and thus

$$\underline{N} = E\left(\sum_{i=1}^{N} \underline{X}_i \,\Big|\, \sum_{i=1}^{N} \underline{X}_i = \underline{N}\right) = N \sum_{j=1}^{\infty} P\left(\underline{X}_k = \underline{I}_j \,\Big|\, \sum_{i=1}^{N} \underline{X}_i = \underline{N}\right) \underline{I}_j \tag{9}$$

and thus for $k = 1, 2, \ldots, N$

$$P\left(\underline{X}_k = \underline{I}_j \,\Big|\, \sum_{i=1}^{N} \underline{X}_i = \underline{N}\right) = N_j/N \ , \quad j = 1, 2, \ldots \tag{10}$$

which establishes (5) for $n = 1$. The argument is concluded by induction on n. For any n, $1 \leq n \leq N$, $\underline{n} = (n_1, n_2, \ldots)$ such that $\sum_{j=1}^{\infty} n_j = n$ suppose that

$$P\left(\sum_{i=1}^{n} \underline{X}_i = \underline{n} \,\Big|\, \sum_{i=1}^{N} \underline{X}_i = \underline{N}\right) = \prod_{j=1}^{\infty} \binom{N_j}{n_j} \Big/ \binom{N}{n} \ . \tag{11}$$

Let $\underline{n}' = (n_1', n_2', \ldots)$ and $\sum_{j=1}^{\infty} n_j' = n + 1 \equiv n'$; then

$$P\left(\sum_{i=1}^{n+1} \underline{X}_i = \underline{n}' \,\Big|\, \sum_{i=1}^{N} \underline{X}_i = \underline{N}\right) = \sum_{j=1}^{\infty} P\left(\underline{X}_{n+1} = \underline{I}_j \,\Big|\, \sum_{i=1}^{n} \underline{X}_i = \underline{n}' - \underline{I}_j, \ \sum_{i=1}^{N} \underline{X}_i = \underline{N}\right) P\left(\sum_{i=1}^{n} \underline{X}_i = \underline{n}' - \underline{I}_j \,\Big|\, \sum_{i=1}^{N} \underline{X}_i = \underline{N}\right). \tag{12}$$

But $\underline{X}_{n+1}, \ldots, \underline{X}_N$, conditional on $\sum_{i=1}^{n} \underline{X}_i = \underline{n}' - \underline{I}_j$ and $\sum_{i=1}^{N} \underline{X}_i = \underline{N}$, are exchangeable and by an argument paralleling that leading to (10) (i.e., computing the conditional expectation of $\sum_{i=n+1}^{N} \underline{X}_i$) it is seen that

$$P(\underline{X}_{n+1} = \underline{I}_j \mid \sum_{i=1}^{n} \underline{X}_i = \underline{n}' - \underline{I}_j ,\ \sum_{i=1}^{N} \underline{X}_i = \underline{N}) = \frac{N_j - n'_j + 1}{N - n' + 1} . \qquad (13)$$

Finally, from (11), (12), (13) and a straightforward computation

$$P(\sum_{i=1}^{n+1} \underline{X}_i = \underline{n}' \mid \sum_{i=1}^{N} \underline{X}_i = \underline{N}) = \sum_{j=1}^{\infty} \left[\frac{n'_j}{n'} \binom{N_j}{n'_j} / \binom{N}{n'}\right] \prod_{k \neq j=1}^{\infty} \binom{N_k}{n'_k} = \prod_{j=1}^{\infty} \binom{N_j}{n'_j} / \binom{N}{n'} ,$$

completing the demonstration.

This result says that a finite sequence of exchangeable discrete random variables can be represented as a probability mixture of ordered generalized hypergeometric distributions. While this result is only of peripheral relevance to the main theme of this paper, it promises to be of use in further characterization of the particular class of prior distributions for which the main results of this paper hold. This is being studied at present. A very special sub-case of this result was also obtained by Kendall [10].

3. Posterior Means and WLSE's or BLUE's:

Bühlmann [11] and independently Ericson [12] and [13] have presented fairly general conditions under which a posterior mean of a location parameter, Θ, (or a 'best' linear approximation to that mean) is a weighted average of the prior mean of Θ and the sample mean, $\overline{X}$, with weights inversely proportional to the prior variance of Θ and the prior expectation of the variance of $\overline{X}$ given Θ, respectively.

In this section a further generalization of this result is given which provides a similar tie between posterior means and best linear unbiased and weighted least squares estimators. This result is of some interest in itself and also will be immediately applicable in the Bayesian analysis of two stage sampling given in section 4.

Theorem 1: Suppose $\underline{X} = (X_1, X_2, \ldots, X_n)$ and ξ are any $n+1$ jointly distributed random variables such that

a) $E(\xi) = m$, $\mathrm{Var}(\xi) = v(\xi) < \infty$,

b) $\underline{V}(\underline{X})$, the variance-covariance matrix of $\underline{X}$ is nxn positive definite, and

c) $E(X_i \mid \xi) = \xi$, $i = 1, 2, \ldots, n$.

Suppose also that either:

d) $E(\xi \mid \underline{X}) = \underline{X}\,\underline{a}^t + b$, $\underline{a} \equiv (a_1, a_2, \ldots, a_n)$ or

d') $\underline{a}$ and b are chosen to minimize

$$E_{\underline{X}}[E(\xi \mid \underline{X}) - \underline{X}\,\underline{a}^t - b]^2. \tag{15}$$

Under these conditions, then,

$$\underline{X}\,\underline{a}^t + b = \frac{E(\xi)EV(\hat{\xi} \mid \xi) + \hat{\xi}\,v(\xi)}{v(\xi) + EV(\hat{\xi} \mid \xi)}, \tag{16}$$

where $\hat{\xi} \equiv \hat{\xi}(\underline{X}) = \underline{1}[\underline{V}(\underline{X})]^{-1}\underline{X}^t / \underline{1}[\underline{V}(\underline{X})]^{-1}\underline{1}^t$ $(\underline{1} = (1, 1, \ldots, 1))$ is the usual BLUE or WLSE of ξ with respect to the variance covariance matrix $\underline{V}(\underline{X})$ or $E_\xi \underline{V}(\underline{X} \mid \xi)$. Additionally, if a), b), c) and d) hold, then

$$E_{\underline{X}} V(\xi | \underline{X}) = v(\xi)(1 - v(\xi)\underline{1}[\underline{V}(\underline{X})]^{-1}\underline{1}^t) = v(\xi)\left[\frac{EV(\hat{\xi} \mid \xi)}{v(\xi) + EV(\hat{\xi} \mid \xi)}\right]. \quad (17)$$

Proof: It suffices to prove the results (16) and (17) under conditions a), b), c) and d'), for condition d) is a special case of d'). A simple computation shows that

$$E_{\underline{X},\xi}(\xi - \underline{X}\,\underline{a}^t - b)^2 = E_{\underline{X}} V(\xi | \underline{X}) + E_{\underline{X}}(E(\xi | \underline{X}) - \underline{X}\,\underline{a}^t - b)^2$$

and hence $\underline{a}$ and b can be chosen to minimize $E_{\underline{X},\xi}(\xi - \underline{X}\,\underline{a}^t - b)^2$. Computing this expectation and minimizing it by straightforward methods yields

$$\underline{a}^t = [\underline{V}(\underline{X})]^{-1} v(\xi)\underline{1}^t \quad (18)$$

and

$$b = m - m\underline{1}\,\underline{a}^t . \quad (19)$$

Hence

$$\underline{a}\,\underline{X}^t + b = m(1 - \underline{1}v(\xi)[\underline{V}(\underline{X})]^{-1}\underline{1}^t) + \underline{1}v(\xi)[\underline{V}(\underline{X})]^{-1}\underline{X}^t \quad (20)$$

is the form of the posterior expectation of ξ or of the best linear approximation to it. Now from a classical point of view note that $E(\underline{X} \mid \xi) = \xi \underline{1}$ and

$$E_{\xi}\underline{V}(\underline{X} \mid \xi) = \underline{V}(\underline{X}) - v(\xi)\,\underline{1}^t\,\underline{1}\,, \tag{21}$$

from which it follows that

$$\hat{\xi} = \frac{\underline{1}[\underline{V}(\underline{X}) - v(\xi)\underline{1}^t\underline{1}]^{-1}\underline{X}^t}{\underline{1}[\underline{V}(\underline{X}) - v(\xi)\underline{1}^t\underline{1}]^{-1}\underline{1}^t} \tag{22}$$

is the BLUE of ξ with respect to the variance-covariance matrix $E_{\xi}V(\underline{X} \mid \xi)$. Since $E_{\xi}V(\underline{X} \mid \xi)$ differs from $\underline{V}(\underline{X})$ by a matrix of constants, $\hat{\xi}$ must also be BLUE with respect to the variance-covariance matrix $\underline{V}(\underline{X})$; hence

$$\hat{\xi} = \underline{1}[\underline{V}(\underline{X})]^{-1}\underline{X}^t \,/\, \underline{1}[\underline{V}(\underline{X})]^{-1}\,\underline{1}^t \tag{23}$$

(The equivalence of (22) and (23) may also be shown by direct computation; $\hat{\xi}$ is also a WLSE of ξ.) Substituting in (20) one has $\underline{a}\underline{X}^t + b = m(1 - v(\xi)\underline{1}[V(\underline{X})]^{-1}\underline{1}^t) + \hat{\xi}v(\xi)\underline{1}[\underline{V}(\underline{X})]^{-1}\underline{1}^t$. The proof of (16) is completed by showing that

$$\underline{1}[\underline{V}(\underline{X})]^{-1}\underline{1}^t = [v(\xi) + EV(\hat{\xi} \mid \xi)]^{-1}.$$

To see this, note that

$$EV(\hat{\xi} \mid \xi) = E_{\xi}\left[\frac{\underline{1}[\underline{V}(\underline{X})]^{-1}\underline{V}(\underline{X} \mid \xi)[\underline{V}(\underline{X})]^{-1}\underline{1}^{t}}{(\underline{1}[\underline{V}(\underline{X})]^{-1}\underline{1}^{t})^{2}}\right]$$

$$= \frac{1}{\underline{1}[\underline{V}(\underline{X})]^{-1}\underline{1}^{t}} - v(\xi) ,$$

from (21) . The proof of (17) follows by a direct and simple computation.

This result gives a neat and simple interpretation to the posterior mean as a Bayes estimator of ξ ; the expression in (17) will be useful in finding optimal sample designs, as will be seen in Section 4.

4. A Two-Stage Sampling Model:

Referring to the basic model of Section 1, suppose now that the N units of the finite population are partitioned into k subsets (primary sampling units or p. s. u. 's) having M units each, and that $kM = N$. Correspondingly let X_{ij} denote the characteristic value of interest associated with the j-th unit within the i-th p. s. u. Thus

$$\underline{X} = (X_{11}, X_{12}, \ldots , X_{1M}, X_{21}, \ldots , X_{2M}, \ldots , X_{kM})$$

The following other definitions are needed:

$$\mu_i = \frac{1}{M}\sum_{j=1}^{M} X_{ij} , \quad \mu = \frac{1}{k}\sum_{i=1}^{k} \mu_i ,$$

$$\sigma_i^2 = \frac{1}{M}\sum_{j=1}^{M} (X_{ij} - \mu_i)^2 , \quad \sigma^2 = \frac{1}{Mk}\sum_{i=1}^{k}\sum_{j=1}^{M} (X_{ij} - \mu)^2 = \sigma_w^2 + \sigma_b^2 ,$$

where

$$\sigma_b^2 = \frac{1}{k}\sum_{i=1}^{k}(\mu_i - \mu)^2 \quad \text{and} \quad \sigma_w^2 = \frac{1}{k}\sum_{i=1}^{k}\sigma_i^2 .$$

All of these population parameters are assumed unknown and μ is the target of inference.

4.1 A Class of Prior Distributions for $\underline{X}$:

The main results of this paper, providing a Bayesian interpretation of some standard sampling theory results, will be demonstrated under a fairly general class, C, of prior distributions for $\underline{X}$. This class of prior distributions is briefly that under which, conditional on the p. s. u., mean and variance, the X_{ij}'s are taken to be exchangeable and independent over the p. s. u.'s and, further, the p. s. u. means and variances are exchangeable. It is thus assumed that, to a fair approximation at least, the N units of the population can be partitioned into k p. s. u.'s on the basis of prior information so that a prior distribution in the class C is appropriate. As will be seen, only four parameters need to be numerically specified although the parameter space is N-dimensional.

The two assumptions, A1 and A2, defining C will be introduced separately and consequences of each noted as results.

A1: Conditional on (μ_i, σ_i^2), the X_{ij}'s are exchangeable for $j = 1, 2, \ldots, M$ and also, for $i \neq i'$, X_{ij} and $X_{i'j'}$ are independent for all j and j'.

From this assumption and by letting $\overline{X}_i$ be the mean of any n_i

of the M X_{ij}'s in the i-th p. s. u., the following results are easily shown. The proof follows exactly that given by Ericson, [9].

Lemma 1: If the X_{ij}'s have a joint prior distribution satisfying A1 then:

$$E(X_{ij} \mid \mu_i, \sigma_i^2) = E(\overline{X}_i \mid \mu_i, \sigma_i^2) = \mu_i , \tag{24a}$$

$$V(X_{ij} \mid \mu_i, \sigma_i^2) = \sigma_i^2 , \tag{24b}$$

$$\text{Cov}(X_{ij}, X_{ij'} \mid \mu_i, \sigma_i^2) = -\frac{\sigma_i^2}{(M-1)} , \text{ and} \tag{24c}$$

$$V(\overline{X}_i \mid \mu_i, \sigma_i^2) = \frac{M - n_i}{M - 1} \frac{\sigma_i^2}{n_i} . \tag{24d}$$

The second assumption defining C is

A2: For $i = 1, 2, \ldots, k$ the ordered pairs (μ_i, σ_i^2) are exchangeable random two-tuples such that the following moments exist:

$E(\mu_i) = m$, $V(\mu_i) = v$, $\text{Cov}(\mu_i, \mu_j) = c$ and $E(\sigma_i^2) = \phi$ where $|c| < v / (k-1)$.

Loosely speaking the class, C, of prior distributions is that under which units within p. s. u.'s are exchangeable and the p. s. u.'s are exchangeable.

Using A2 and Lemma 1 the following is trivially shown:

Lemma 2: If the prior distribution of $\underline{X}$ is in the class C (satisfying

A1 and A2) then:

$$E(\mu) = E(\overline{X}_i) = E(X_{ij}) = m \tag{25a}$$

$$V(\mu) = \frac{v + (k-1)c}{k} \tag{25b}$$

$$E(\sigma_w^2) = \phi \tag{25c}$$

$$E(\sigma_b^2) = \frac{k-1}{k}(v - c) \tag{25d}$$

$$V(X_{ij}) = \phi + v \tag{25e}$$

$$Cov(X_{ij}, X_{ij'}) = v - \frac{\phi}{M-1}, \quad j \neq j' \tag{25f}$$

$$Cov(X_{ij}, X_{i'j'}) = Cov(\overline{X}_i, \overline{X}_{i'}) = c, \quad i \neq i', \text{ and} \tag{25g}$$

$$V(\overline{X}_i) = v + \frac{M - n_i}{M-1}\,\frac{\phi}{n_i} \quad . \tag{25h}$$

As the final preliminary result the following can be shown:

<u>Lemma 3:</u> If the prior distribution of $\underline{X}$ is in the class C then:

$$E(\mu_i | \mu) = E(\overline{X}_i | \mu) = \mu \tag{26a}$$

$$E_\mu V(\bar{X}_i \mid \mu) = v + \frac{M - n_i}{M - 1} \frac{\phi}{n_i} - \frac{v + (k-1)c}{k} = E(\sigma_b^2) + \frac{M - n_i}{M - 1} \frac{E(\sigma_w^2)}{n_i} \qquad (26b)$$

and

$$E_\mu \operatorname{Cov}(\bar{X}_i, \bar{X}_j \mid \mu) = -\frac{(v - c)}{k} = -\frac{E(\sigma_b^2)}{k - 1} . \qquad (26c)$$

The proof of this result follows first by observing that by A2 the μ_i's are marginally exchangeable and remain exchangeable conditional on their sum or on μ. The result then follows from Lemma 1, conditions like $E(\mu \mid \mu) = \mu$, $V(\bar{X}_i) = EV(\bar{X}_i \mid \mu) + VE(\bar{X}_i \mid \mu)$, and straightforward computations.

The basic result of this subsection is now an immediate application of Theorem 1, for, under the correspondence $\mu \leftrightarrow \xi$ and $\bar{X}_i \leftrightarrow X_i$, the conditions a), b) and c) of Theorem 1 are met. (See (25a), (25b), (26a), (25g), (25h) and note that the A2 - condition that $|c| < v/(k - 1)$ ensures that $\underline{V}(\bar{\underline{X}})$ is positive definite.) This result is

<u>Theorem 2:</u> If a prior distribution in the class, C, is assigned to $\underline{X}$ and if a sample of size n_i is drawn from the i-th p. s. u. $i = 1, 2, \ldots, r$ $(1 \le r \le k,\ 1 \le n_i \le M)$ then the posterior mean of μ (or the best linear approximation to it) is given by

$$\frac{mEV(\hat{\mu} \mid \mu) + \hat{\mu}V(\mu)}{V(\mu) + EV(\hat{\mu} \mid \mu)} , \qquad (27)$$

where $\hat{\mu} = \underline{1}\,[V(\bar{\underline{X}})]^{-1}\,\bar{\underline{X}}^t / \underline{1}\,[V(\bar{\underline{X}})]^{-1}\,\underline{1}^t$ is the BLUE or WLSE of

μ with respect to either the variance-covariance matrix of $\underline{\overline{X}} = (\overline{X}_1, \ldots, \overline{X}_r)$, $\underline{V}(\underline{\overline{X}})$, or the prior expectation of the variance-covariance matrix of $\underline{\overline{X}}$ conditional on μ, $E\underline{V}(\underline{\overline{X}} \mid \mu)$. Also, if $E(\mu \mid \underline{\overline{X}})$ is linear in $\underline{\overline{X}}$ then

$$EV(\mu \mid \underline{\overline{X}}) = V(\mu)(1 - V(\mu)\, \underline{1}[\underline{V}(\underline{\overline{X}})]^{-1}\, \underline{1}^t). \tag{28}$$

By explicitly inverting the rxr variance-covariance matrix $V(\underline{\overline{X}})$ as defined by (25g) and(25h) one finds that

$$\hat{\mu} = \sum_{i=1}^{r} v_i \overline{X}_i \Big/ \sum_{i=1}^{r} v_i \tag{29}$$

and

$$EV(\mu \mid \underline{\overline{X}}) = V(\mu)(1 - V(\mu) \frac{\sum_1^r v_i}{(1 + c \sum_1^r v_i)}) \tag{30}$$

where

$$v_i = [(v - c) + \frac{M - n_i}{M - 1} \frac{\phi}{n_i}]^{-1} = [\frac{k}{k - 1} E(\sigma_b^2) + \frac{M - n_i}{M - 1} \frac{E(\sigma_w^2)}{n_i}]^{-1}. \tag{31}$$

4.2 Discussion:

Several comments on the model and results above are appropriate here.

The class of prior distributions, C, is certainly non-empty; it contains the multivariate normal distribution, the multivariate Student distribution and others. If the X_{ij}'s are discrete random variables then it appears that the theorem of Section 2 will be useful in helping to characterize this class of priors.

The results given above only require specification of four parameters of the prior distribution as per A2, viz., m, v, c, and ϕ. This specification does not define a unique prior but suffices for determining the form of the posterior mean of μ, a natural Bayes estimator of μ, and, as shown in the next section, the optimal sample design. Clearly if one desires more by way of inference, for example the posterior variance of μ, then more detailed specification of his prior distribution will be required.

5. Optimal Sample Design:

Suppose now that it costs k_1 to sample a p. s. u. and k_2 to sample each element within a p. s. u. and that a total budget of K is available for sampling. The natural Bayesian approach to the design question is to choose the design to minimize the prior expectation of the posterior variance of μ, (28) or (30), subject to the budgetary constraint. This approach can, of course, be given a formal justification from a decision - theoretic viewpoint when estimating μ with quadratic loss.

The design problem can then be formalized as one of choosing r^o and n_i^o, $i = 1, \ldots, r^o$, to minimize

$$EV(\mu \mid \bar{\underline{X}}) = V(\mu)(1 - V(\mu) \frac{\sum_{i=1}^{r} v_i}{(1 + c \sum_{i=1}^{r} v_i)}) \qquad (32)$$

subject to the constraints:

$$1 \leq r^o \leq k$$

$$1 \leq n_i^o \leq M, \quad i = 1, \ldots, r^o \quad \text{and} \tag{33}$$

$$r^o k_1 + k_2 \sum_{i=1}^{r^o} n_i^o \leq K. \tag{34}$$

5.1 Solution to the Design Problem:

The solution to this design problem turns out to be fairly simple and will be obtained in several steps indicated as Results below.

Result 1: Minimizing (32) is equivalent to maximizing

$$\sum_{i=1}^{r} v_i = \sum_{i=1}^{r} \frac{n_i (M - 1)}{n_i [(M - 1)(v - c) - \phi] + M\phi} \tag{35}$$

This result is obvious by inspecting (32).

Result 2: If $(M - 1)(v - c) \leq \phi$, the optimal design is given by utilizing the entire budget in censusing as few p. s. u.'s as possible. More specifically, the optimum is of the form

$$n_i^o = M \qquad i = 1, \ldots, r^o$$

$$n_{r^o+1}^o = \left[\frac{K - (r^o + 1) k_1 - r^o M k_2}{k_2} \right] \tag{36}$$

where the square brackets denote the greatest integer function.

To see this result, note first that if $(M - 1)(v - c) = \phi$ then the problem is one of maximizing $\sum_{i=1}^{r} n_i$ subject to the constraints (33) and (34). The solution to this problem is obviously as in (36). It is also clear that the solution will also be given by (36) whenever it is the case that increasing the sample size from n_i to $n_i + 1$ for all $n_i = 1, 2, \ldots, M - 1$ increases $\sum_{i=1}^{r} v_i$ by more than introducing a new p. s. u. i.e., increasing n_j from zero to one. (Or equivalently, if Σv_i is increasing at an increasing rate with n_i.) Either way (by comparing the increases directly or by examining the second derivative) it follows easily that the optimum is given by (36) whenever $(M - 1)(v - c) < \phi$.

The condition that $(M - 1)(v - c) \leq \phi$ is equivalent to

$$E(\sigma_b^2) \leq \frac{k-1}{k} \frac{E(\sigma_w^2)}{M-1} \tag{37}$$

which provides an interpretation of the solution in (36) and agrees with the traditional sample survey two-stage design result by simply replacing σ_b^2 and σ_w^2 by their prior expectations. (See Cochran [14], Chapter 10.)

Result 3: If $(M - 1)(v - c) > \phi$ the optimal design takes some r^o and makes n_i^o $i = 1, \ldots, r^o$ as nearly equal as possible.

This result is obvious by the symmetry of the function (35) in n_i and the negativity of the second derivatives when $(M - 1)(v - c) > \phi$, or by a direct computation showing that Σv_i increases as one moves toward equalizing the n_i's.

In view of this result, if $(M - 1)(v - c) > \phi$ then the optimal solution will be taken to be of the form $n_i^o = n^o \quad i = 1, 2, \ldots, r^o$, and the design problem reduced to that of finding r^o and n^o. For such designs the objective function to be maximized becomes (from (35))

$$\sum_{i=1}^{r} v_i = \frac{r n (M - 1)}{n[(M-1)(v-c) - \phi] + M\phi} \tag{38}$$

A small bit of manipulation shows that the choice of r^o and n^o to maximize (38) subject to the constraints $1 \leq r^o \leq k$, $1 \leq n^o \leq M$ and $r^o k_1 + r^o n^o k_2 \leq K$ is a well-known and solved problem. First maximizing (38) is equivalent to minimizing its reciprocal, V^*, which is given by

$$V^* = \frac{v - c}{r} + \frac{M - n}{M - 1} \frac{\phi}{nr} \tag{39}$$

Substituting from (25c) and (25d), one can write this as

$$V^* = \frac{k E(\sigma_b^2)}{(k - 1)r} + \frac{M - n}{M - 1} \frac{E(\sigma_w^2)}{nr} \tag{40}$$

Finally minimizing (40) is equivalent to minimizing $V^{**} \equiv (40) - E(\sigma_b^2)/(k-1)$

or

$$V^{**} = \frac{k - r}{k - 1} \frac{E(\sigma_b^2)}{r} + \frac{M - n}{M - 1} \frac{E(\sigma_w^2)}{nr} . \quad (41)$$

This form is well-known, for if the two expectation operators are removed, then it is just the expression for the sampling variance of the standard estimator of μ from a two-stage sampling design. This means simply that under the class C of prior distributions on $\underline{X}$ the optimal Bayesian sample design is precisely equivalent to the classical optimal design where the Bayesian's prior expectations, $E(\sigma_b^2)$ and $E(\sigma_w^2)$, are substituted for the classicist's σ_b^2 and σ_w^2. Specifically, n^o and r^o may be found by applying the results of Cochran, [14] Chapter 10, for example, and making the suitable identifications between (41) above and his equation (10.8).

5.2 Discussion:

The basic result above is that, under the class of prior distributions under which the p. s. u. means and variances are viewed as exchangeable random variables and the X_{ij}'s within each p. s. u. are viewed as exchangeable, the optimal Bayesian sampling design is a balanced two-stage sample. Furthermore, the specific allocation of sampling effort as between the number of p. s. u.'s to sample and the number of units to select within each sampled p. s. u. for this Bayesian model is yielded by the traditional sample survey solution to this problem by merely substituting $E(\sigma_b^2)$ and $E(\sigma_w^2)$ for σ_b^2 and σ_w^2 in the traditional optimal allocation formulae. Looked at another way, the optimal Bayes allocation which minimizes $EV(\mu \mid \underline{X})$ is precisely that which minimizes $EV(\hat{\mu} \mid \sigma_b^2, \sigma_w^2, \mu)$

where $\hat{\mu}$ is the standard unbiased estimator of μ from a two-stage sample.

In practice there will be little difference between the Bayesian's design and the traditional solution, since implementation of the traditional solution requires guesses or estimates of σ_b^2 and σ_w^2 and, given the same prior information, these values typically will be close to the Bayesian's assessments of $E(\sigma_b^2)$ and $E(\sigma_w^2)$.

For the Bayesian solution here all that is important is r^o and n^o and not which r^o p. s. u.'s or which n^o elements within a p. s. u. are selected. (For simplicity, the first r p. s. u.'s and n units have been assumed.) This is a consequence of the exchangeability assumptions defining the distributions in C which are tantamount to the prior opinion that the p. s. u.'s and units within p. s. u.'s have been randomized. Again, in practice the Bayesian would randomize his selection either to justify further the choice of a prior distribution in C or to render his conclusions more acceptable to others. (See Ericson [4] and [6] for further discussion on this.)

Finally, the results above are easily extendable to more than two stages of sampling and the case of unequal p. s. u. sizes is presently under investigation.

REFERENCES

[1] Godambe, V. P. A Unified Theory of Sampling from Finite Populations, J. R. Statist. Soc. B, 17, 267-278, 1955.

[2] Godambe, V. P. A Review of the Contributions Towards a Unified Theory of Sampling from Finite Populations, Inter. Statist. Inst. Rev. 33, 242-258, 1965.

[3] Godambe, V. P. A New Approach to Sampling from Finite Populations I and II, J. R. Statist. Soc. B, 28, 310-328, 1966.

[4] Ericson, W. A. Subjective Bayesian Models in Sampling Finite Populations, J. Roy. Statist. Soc. B., 31, 195-233, 1969.

[5] Basu, D. On Sampling With and Without Replacement, Sankhyā 20, 287-294, 1958.

[6] Ericson, W. A. Subjective Bayesian Models in Sampling Finite Populations: Stratification. New Developments in Survey Sampling (N. L. Johnson and H. Smith, Jr., eds.) New York, John Wiley and Sons, Inc., 1969, 326-357. (Reprinted in expanded form as Technical Report #15, Department of Statistics, The University of Michigan, May, 1972.)

[7] Hewett, E. and Savage, L. J. Symmetric Measures on Cartesian Products, Trans. Amer. Math. Soc. 80, 1115-1122, 1955.

[8] Feller, W. An Introduction to Probability Theory and Its Applications, Vol. II., New York, John Wiley and Sons, Inc., 1966.

[9] de Finetti, B. La Probabilità e la Statistica nei Rapporti con l'induzione, Secondo i Diversi Punti di Vista, Centro Internazionale Matematico Estivo (C. I. M. E.), Cremonese, Rome, 1959. (Appears in translation as Chapter 9 of: de Finetti, B. Probability, Induction and Statistics, New York, John Wiley and Sons, Inc., 1972).

[10] Kendall, D. G. On Finite and Infinite Sequences of Exchangeable Events. Studia Sci. Math. Hungar., 2, 319-327, 1969.

[11] Bühlmann, H. Introductory Report Experience Rating and Credibility. The Astin Bulletin IV, Part III, July, 1967.

[12] Ericson, W. A. A Note on the Posterior Mean of a Population Mean, J. Roy. Statist. Soc. B., 31, 332-334, 1969.

[13] Ericson, W. A. On the Posterior Mean and Variance of a Population Mean, J. Amer. Statist. Assoc., 65, 649-652, 1970.

[14] Cochran, W. G. Sampling Techniques (2nd Edition), New York, John Wiley and Sons, Inc., 1963.

☆ U. S. GOVERNMENT PRINTING OFFICE: 1976 — 657-630/677