Biometrika (1968), 55, 3, *p.* 547
Printed in Great Britain

A new estimation theory for sample surveys

BY H. O. HARTLEY AND J. N. K. RAO
Texas A and M University

SUMMARY

A new estimation theory for sample surveys is proposed, the basic feature of which is a special parametrization of finite populations based on the assumption that a character attached to the units is measured on a known scale with a finite set of scale points. In the class of estimators which do not functionally depend on the identification labels preattached to the units, the following results are proved: (1) For simple or stratified simple random sampling without replacement, the customary estimators are minimum variance unbiased. (2) For simple random sampling with replacement, the sample mean based only on the distinct units in the sample is the maximum-likelihood estimator of the population mean. (3) If a concomitant variable with known population mean is also observed, an approximation to the maximum-likelihood estimator of the population mean is closely related to the customary regression estimator. (4) If prior information in the form of a prior distribution is available, Bayes estimators can be derived using the complete likelihood.

1. INTRODUCTION

Historically the development of sample-survey estimation theory has progressed mainly inductively. Estimators which appeared to be reasonable have been proposed and their comparative properties carefully examined either by analytic evaluation or data-analysis. Most of these studies have been concerned with comparisons of variances and biases of the various estimators. The absence of a basic deductive estimation theory, has recently been lamented by Godambe (1955, 1965) who considers the sample-theorists' search for unbiased minimum variance (U.M.V.) estimators to be a basic 'fallacy' and proves that, under certain assumptions, such estimators do not exist. It is obvious that the question of the existence or otherwise of U.M.V. estimators depends, among other things, on the class of estimators considered; Godambe considered a class of estimators which is permitted to depend functionally on identification labels $i = 1, 2, \ldots, N$ preattached to the N units of the population and used as observables in the computation of estimators. If preassigned labels are not observable and are only conceptually attached to the units, then clearly estimators which functionally depend on the labels cannot be implemented.

In the present theory we do not wish to confine the activities of sampling from finite populations to those with labelled units only as this would exclude many important situations of finite population sampling. For example, the important area of acceptance sampling of finite lots of machine parts would be excluded since here the attachment of identifying labels is often impractical. We consider it therefore of interest to develop an estimation theory in which estimators are allowed to depend on labels only if these can be regarded as informative concomitant variables, and in the present paper we confine ourselves to estimators which do not functionally depend on the labels. Within this class of estimators, defined below, we shall be able to prove the following results.

(*a*) Certain of the estimators in current use and advocated in the literature are indeed maximum-likelihood estimators and/or U.M.V. estimators.

(*b*) If suitable prior information is available, Bayesian concepts can be adjoined to our theory using the complete likelihood. However, since we shall be concerned with a rather large number of parameters, the computation or graphical representation of the full parametric posterior distribution would appear to be difficult. Nevertheless, it is possible to use point estimation via the Bayes estimator, and thereby obtain simple results for important parametric functions such as the population mean and variance.

The basic feature of our theory is a special parametrization of finite populations of N units. Considering first the case where a single characteristic, say, y is attached to the units, we assume, with essentially no loss of generality, that this characteristic is measured on a known scale with a finite set of scale points y_t $(t = 1, 2, \ldots, T)$. A similar assumption has recently been made by Kempthorne (1965) for experimental data. Any finite population can then be completely described by the set of T non-negative integer parameters

N_t, being the number of units in the population having the characteristic y_t (1)

and satisfying the condition

$$\sum_{t=1}^{T} N_t = N. \tag{2}$$

Henceforth, sums and products over t are for $1, \ldots, T$. In certain cases we describe subsets of the target population each by their separate sets of parameters N_t, e.g. for stratification. If k characteristics are attached to the units, y_t denotes a k-element vector and T the total number of possible measurement combinations. It is clear that the present parametrization of a finite population makes no model assumptions whatsoever except that of discrete scale measurements. Although the number of parameters is very large, our estimation theory will be predominantly concerned with a few important parameter functions such as the moments $\mu_r' = N^{-1}\Sigma N_t y_t^r$. Moreover, the vast majority of the parameters N_t will usually be estimated as zero; in fact it will be seen that the number of positive parameter estimates will be at most equal to the sample size and does not depend on T or the non-observed y_t.

The task of finite population sampling will consist of

(*a*) the sample design, i.e. the procedure of drawing a sample consisting of n *distinct* units, where n may be fixed or random, and measuring the y_t for these units;

(*b*) the use of measured y_t to compute estimators of the population parameters N_t or functions thereof, such as the population mean $\bar{Y} = \Sigma(N_t/N)\, y_t$.

With regard to (*a*), we stress that there are many situations where distinct units can be sampled without identifying labels being attached to the units. For example, a foreman can sample distinct machine parts from a bin in accordance with instructions. However, preattached labels, if available, will certainly be helpful in the process of implementing a specified sampling procedure of distinct units. With regard to (*b*) we shall only consider estimators which do not directly depend on the labels and are defined as mathematical functions of the scale points y_t and of the sample frequencies n_t, i.e. the number of distinct units in the sample observed to have the scale point y_t. We shall term this restricted class of estimators scale-load estimators. We consider that the restriction to these estimators will normally result in no loss of information because

(i) there are many situations in which no labels are attached to the units and

(ii) in most situations in which labels are attached to the units it is known that they cannot be informative beyond the design stage (*a*).

In situations in which it is held that available labels are informative they can be adjoined to the scales y_t as a $(k+1)$th vector component. However, in this case difficulties of non-estimability are usually encountered which we shall not discuss in this paper.

In the present paper, except for §§ 2·3 and 3, we assume that the only prior information on the N_t that is available is given by (2). However, in § 2·3 we develop a Bayesian theory using prior distributions for the parameters and in § 3 we develop a theory of regression-type estimation assuming that y_t has two components, a target variable and a concomitant variable, and that the population mean for the concomitant variable is known.

It will be seen that with our theory every sample design used in (*a*) requires the derivation of its appropriate likelihood for the observables, n_t. In this paper, the only sampling procedures considered are simple random sampling with equal probabilities with or without replacement. Extensions to multi-stage designs, unequal probability sampling, etc., will be considered in subsequent papers.

2. Simple random sampling without replacement

2·1. U.M.V. *estimators*

A random sample of fixed size n is drawn with equal probabilities without replacement. Let n_t denote the number of units having the value y_t in the sample. Then clearly the n_t's are integers with

$$n_t \geqslant 0, \quad \Sigma n_t = n. \tag{3}$$

The likelihood $L(N_1, \ldots, N_T)$ is given by the multidimensional hypergeometric distribution

$$L(N_1, \ldots, N_T) = \Pi \binom{N_t}{n_t} \Big/ \binom{N}{n}. \tag{4}$$

We now give a proof, due to B. K. Kale, that $(n_1, \ldots, n_T)$ is complete sufficient for $(N_1, \ldots, N_T)$. Assuming that $T = m$ gives a complete family, we prove the result by induction. Consider $T = m+1$ and let $E\{g(n_1, \ldots, n_{m+1})\} = 0$ for a real-valued function g for all $N_1, \ldots, N_{m+1}$ such that $N_1 + \ldots + N_{m+1} = N$. Now, considering $N_{m+1} = 0$ and noting that $n_{m+1} \leqslant N_{m+1}$, we get $g(n_1, \ldots, n_m, 0) = 0$ for all $n_1, \ldots, n_m$ such that $n_1 + \ldots + n_m = n$, $n_i \leqslant N_i$ $(i = 1, \ldots, m)$ and $N_1 + \ldots + N_m = N$, since we have a complete family for $T = m$. Considering next $N_{m+1} = 1$ and using the above argument, we get $g(n_1, \ldots, n_m, 1) = 0$ for all $n_1, \ldots, n_m$ such that

$$n_1 + \ldots + n_m = n-1, \quad n_i \leqslant N_i \, (i = 1, \ldots, m) \quad \text{and} \quad N_1 + \ldots + N_m = N-1.$$

Continuing in this way we find that $g(n_1, \ldots, n_m, j) = 0$ for all $n_1, \ldots, n_m, j$ such that

$$n_1 + \ldots + n_m + j = n.$$

Now permuting $n_1, \ldots, n_{m+1}$ the proof is complete, noting that we have a complete family for $T = 2$.

Consequently, the moment estimators $m'_r = n^{-1} \Sigma_t y_t^r$ are U.M.V. estimators of the μ'_r, since $E(n_t/n) = N_t/N$. In particular, the sample mean $\bar{y} = m'_1$ is the U.M.V. estimator of the population mean $\bar{Y} = \mu'_1$. Further

$$s^2 = (n-1)^{-1} \{\Sigma n_t y_t^2 - (\Sigma n_t y_t)^2/n\} \tag{5}$$

is the U.M.V. estimator of

$$S^2 = N\sigma^2/(N-1). \tag{6}$$

We now consider the asymptotic distribution of the maximum-likelihood estimators as $n \to \infty$, $N \to \infty$ with $n/N \to \lambda\,(0 < \lambda < 1)$, $N_t/N \to P_t$ and $n_t/n \to p_t$. The distribution of

$$n_1, \ldots, n_{T-1}$$

may be written as

$$p(n_1, \ldots, n_{T-1}) = \Pi \left\{ \binom{N_t}{n_t} \left(\frac{n}{N}\right)^{n_t} \left(\frac{N-n}{N}\right)^{N_t - n_t} \right\} \Big/ \left\{ \binom{N}{n} \left(\frac{n}{N}\right)^{n} \left(\frac{N-n}{N}\right)^{N-n} \right\}. \tag{7}$$

Now, using the normal approximation to the binomial, it follows that

$$(n_1/n - P_1, \ldots, n_{T-1}/n - P_{T-1}) \sqrt{n}$$

has asymptotically a $(T-1)$-variate normal distribution with zero means and variance-covariance matrix (a_{ij}), where $a_{ii} = (1-\lambda)\,P_i(1-P_i)$ and $a_{ij} = -(1-\lambda)\,P_i P_j$. Consequently, $g(n_1/n, \ldots, n_{T-1}/n)$ is asymptotically normal where g is totally differentiable (Rao, 1965, p. 321). In particular, the maximum-likelihood estimators of μ'_r and σ^2 are asymptotically normal. It is also easily seen that they are asymptotically efficient.

Extension of the above results to stratified simple random sampling without replacement is straightforward. In particular, the customary estimators are U.M.V., noting that each stratum is described by its separate set of parameters N_t.

2·2. *Maximum-likelihood estimators*

Since we have already found the U.M.V. estimators, it may be argued that it is of little value to derive the maximum-likelihood estimators because the maximization of the likelihood, restricting the parameters N_t to an integral mesh, may not have particular merit for small samples. We, nevertheless, consider the maximum-likelihood estimation of the N_t for two reasons. First, if N/n is an integer the maximum-likelihood estimators are identical to the U.M.V. estimators. Secondly, in those cases where T is small and the N_t are the parameters of interest, say in the estimation of frequency distributions, maximum-likelihood estimation may have some advantage since it incorporates the information that the N_t are integral.

Case 1. *The expansion factor N/n is an integer.* In this case we have

$$N/n = r = \text{integer}. \tag{8}$$

We have to maximize $L(N_1, \ldots, N_T)$ given by (4) subject to (1)–(3) and (8). It can be shown that the solution is given by the maximum-likelihood estimators

$$\hat{N}_t = \frac{N}{n} n_t \quad (t = 1, \ldots, T); \tag{9}$$

the proof will be given elsewhere. Consequently, the U.M.V. estimators m'_r are also the maximum-likelihood estimators of the μ'_r.

Case 2. *The expansion factor N/n is not integral.* In this case the maximum-likelihood estimators $\hat{N}_t$ of N_t will be found to be rounded up or down versions of the $\hat{N}_t$ as defined by (9) and will be obtained as follows. First, for all indices t with $n_t = 0$, put $\hat{N}_t = 0$. For the remaining t, start with any set of N_t, say $N_t^{(0)}$, such that $\Sigma N_t^{(0)} = N$. We now obtain a global maximum $\hat{N}_t$ from $N_t^{(0)}$ by a series of interchanges. Starting with $N_t^{(0)}$ we step from $N_t^{(i-1)}$ to $N_t^{(i)}$

by the ith interchange defined as follows. Let θ_i denote the t-index for which $N_t^{(i-1)}/n_t$ is a maximum and τ_i the t-index for which $(N_t^{(i-1)}+1)/n_t$ is a minimum. Then if

$$\frac{N_{\theta_i}^{(i-1)}}{n_{\theta_i}} > \frac{N_{\tau_i}^{(i-1)}+1}{n_{\tau_i}}, \tag{10}$$

we clearly have $\theta_i \neq \tau_i$ and define

$$N_t^{(i)} = \begin{cases} N_{\theta_i}^{(i-1)} - 1 & \text{for} \quad t = \theta_i, \\ N_{\tau_i}^{(i-1)} + 1 & \text{for} \quad t = \tau_i, \\ N_t^{(i-1)} & \text{for} \quad t \neq \theta_i, \tau_i. \end{cases} \tag{11}$$

If (10) is satisfied then it can be shown that

$$L(N_1^{(i)}, \ldots, N_T^{(i)}) > L(N_1^{(i-1)}, \ldots, N_T^{(i-1)}). \tag{12}$$

The process is repeated until for some i (10) is *not* satisfied and then it can be shown that $N_t^{(i-1)}$ provides a global maximum of the likelihood. Because of (12) the process cannot cycle and must come to a close as $L(N_1, \ldots, N_T)$ is defined only on a finite set. The proof that the above process leads to a global maximum, and a simplified algorithm based on a particular choice of the starting set of N_t will be given elsewhere.

2·3. *Bayesian estimation*

We first consider the case when $N \to \infty$, $N_t/N \to P_t$ and n fixed, so that the likelihood is given by the multinomial distribution

$$L(P_1, \ldots, P_T) = \frac{n!}{\Pi n_t!} \Pi P_t^{n_t}. \tag{13}$$

We now assume that prior information on the parameters P_t is available in the form of a joint prior distribution on the P_t. If we can assume the conjugate prior distribution

$$\phi(P_1, \ldots, P_T) \propto \Pi P_t^{\nu_t - 1}, \tag{14}$$

where the ν_t (> 0) are constants with $\Sigma \nu_t = \nu$ (Raiffa & Schlaifer, 1961, p. 47), the joint posterior distribution of the P_t for a given set of n_t is then given by

$$\phi^*(P_1, \ldots, P_T) \propto \Pi P_t^{n_t + \nu_t - 1}. \tag{15}$$

We now derive the Bayes estimator. With a quadratic loss function, the estimator which minimizes the expected loss is given by the posterior expectation of the parameter being estimated and this is the Bayes estimator. Now, noting that (15) is nothing but a Dirichlet distribution, we get the Bayes estimator of μ'_r as

$$E'(\mu'_r) = \Sigma E'(P_t) y_t^r = \Sigma \frac{n_t + \nu_t}{n + \nu} y_t^r$$

$$= w m'_r + (1 - w) M'_r, \tag{16}$$

where the weight w is given by
$$w = \frac{n}{n + \nu} \tag{17}$$

and
$$M'_r = \nu^{-1} \Sigma \nu_t y_t^r. \tag{18}$$

In particular, for the estimation of the population mean μ_1', we have

$$E'(\mu_1') = wm_1' + (1-w)\, M_1' \tag{19}$$

which has been used in the past in situations in which n is small and M_1' represents a prior data sample mean based on a large sample size ν of related data from a population whose mean is not likely to differ to a great deal from μ_1'. We have recently had occasion to examine the properties of estimators of type (19) bartering bias ($w = 0$) against variance ($w = 1$). In general it should be noted that for the application of (16) we only need to know M_r', the prior mean of μ_r', and ν. The expected loss which the decision-maker faces by choosing the Bayes estimator of μ_r' is given by the posterior variance

$$V'(\mu_r') = (n+\nu+1)^{-1}\,[wm_{2r}' + (1-w)\, M_{2r}' - \{wm_r' + (1-w)\, M_r'\}^2]. \tag{20}$$

The Bayes estimator of $\sigma^2 = \Sigma P_t y_t^2 - (\Sigma P_t y_t)^2$ is

$$E'(\sigma^2) = (n+\nu)\, V'(\mu_2'). \tag{21}$$

We next consider the case of finite N. Now Hoadley, in an unpublished paper, has shown that a convenient prior distribution on the N_t is given by the compound multinomial distribution

$$\phi(N_1, \ldots, N_T) \propto \Pi \frac{(N_t+\nu_t-1)!}{N_t!\,(\nu_t-1)!} \quad (\nu_t > 0). \tag{22}$$

Using Hoadley's posterior moments of the N_t, we obtain the Bayes estimator of μ_r' as

$${}_N E'(\mu_r') = \left(1-\frac{n}{N}\right) E'(\mu_r') + \frac{n}{N} m_r', \tag{23}$$

where $E'(\mu_r')$ is given by (16). The posterior variance of μ_r' and the Bayes estimator of σ^2 are respectively given by

$${}_N V'(\mu_r') = \left(1-\frac{n}{N}\right)\left(1+\frac{\nu}{N}\right) V'(\mu_r') \tag{24}$$

and

$${}_N E'(\sigma^2) = \left(1-\frac{n}{N}\right)\left(1-\frac{1-w}{N}\right) E'(\sigma^2) + \frac{n}{N}\,(m_2' - m_1'^2) + \frac{n}{N}\left(1-\frac{n}{N}\right)(1-w)^2\,(m_1' - M_1')^2, \tag{25}$$

where $V'(\mu_r')$ and $E'(\sigma^2)$ are given by (20) and (21) respectively. As $N \to \infty$ with n fixed, (23), (24) and (25) respectively tend to (16), (20) and (21).

3. Estimation with concomitant variables

We now consider a situation customarily dealt with by ratio or regression method of estimation in which two variates y and x are attached to each unit and the mean of the target variate y is to be estimated utilizing the available information about x.

As before, we assume that a finite set of T distinct, known values y_t are feasible for y. Likewise, we assume that a finite set of I distinct, known values x_i are feasible for

$$x\,(x_1 < x_2 < \ldots < x_I).$$

Let N_{it} denote the number of units in the population which have x_i and y_t attached to them. We then have

$$N_{it} \geqslant 0 \quad \text{and} \quad \sum_{i=1}^{I} \sum_{t=1}^{T} N_{it} = N. \tag{26}$$

Henceforth, sums for t and i are over $1, \dots, T$ and $1, \dots, I$ respectively. A random sample of size n is drawn with equal probability and without replacement. Denote by n_{it} the number of units in the sample which have x_i and y_t attached to them. Clearly we have

$$n_{it} \geqslant 0 \quad \text{and} \quad \Sigma\Sigma n_{it} = n. \tag{27}$$

The likelihood is given by the multidimensional hypergeometric distribution

$$L(N_{11}, \dots, N_{IT}) = \prod_{i,t} \binom{N_{it}}{n_{it}} \Big/ \binom{N}{n}. \tag{28}$$

If (26) is the only information available on the N_{it}, the principle of maximum likelihood leads to the maximization of (28) subject to (26) resulting in maximum-likelihood estimators as derived in § 2·2. The amount of information about the x variable varies from case to case. However, one of the most frequent situations arising in sample surveys is one in which *only* the population mean $\bar{X}$, or a total $N\bar{X}$, of the x_i's is known, and this is the case we consider here. We assume, therefore that the parameters N_{it} are known to satisfy

$$N^{-1}\Sigma\,\Sigma N_{it} x_i = \bar{X} \tag{29}$$

and maximize (28) subject to (26) and (29).

We confine ourselves here, however, to the multinomial situation in which $N \to \infty$ and $N_{it}/N \to P_{it}$ while n is held fixed. The likelihood (28) is then replaced by

$$L(P_{11}, \dots, P_{IT}) = \frac{n!}{\prod_{i,t} n_{it}!} \prod_{i,t} P_{it}^{n_{it}} \tag{30}$$

and the restrictions (26) and (29) are replaced by

$$P_{it} \geqslant 0, \quad \Sigma\Sigma P_{it} = 1 \tag{31}$$

and

$$\Sigma\,\Sigma\, P_{it} x_i = \bar{X}. \tag{32}$$

It is easy to see that the complete sufficient statistic does not exist here because the number of 'free' parameters is $IT-2$, whereas the dimensionality of the sufficient statistic

$$(n_{11}, \dots, n_{I,T-1}) \quad \text{is} \quad IT-1.$$

Denote by $(P_{11}^*, \dots, P_{it}^*, \dots, P_{IT}^*)$ a point within the closed space defined by (31) and (32) at which $L(P_{11}, \dots, P_{IT})$ has a global maximum. Assume further that we have a sample with at least two observed x_i values, i.e. $n_{it} > 0$ for some t and at least two i-values, then no P_{it}^* can be equal to one as otherwise all other P_{it}^* would be zero and $L(P_{11}^*, \dots, P_{IT}^*) = 0$. We now show that if a particular $n_{jt} = 0$, then we must have $P_{jt}^* = 0$ unless $j = 1$, i.e. $x_j = x_1 = x_{\min}$, or $j = I$, i.e. $x_j = x_I = x_{\max}$. For suppose $P_{jt}^* > 0$ with $x_1 < x_j < x_I$; then it is always possible to find a pair (x_k, x_r) such that

$$x_k < x_j < x_r \tag{33}$$

and

$$n_{kt'} + n_{rt''} > 0 \tag{34}$$

so that we can change the three P^*-values by

$$P_{jt}^* + \delta_j, \quad P_{kt'}^* + \delta_k, \quad P_{rt''}^* + \delta_r \tag{35}$$

satisfying (32) and (33) and increasing $L(P_{11}^*, \dots, P_{IT}^*)$. The increments δ_j and δ_k need only be defined in terms of $\delta_r > 0$ by

$$\delta_j = -\delta_r \frac{x_k - x_r}{x_k - x_j}; \quad \delta_k = -\delta_r \frac{x_r - x_j}{x_k - x_j}, \tag{36}$$

where, because of $\delta_r > 0$ and (33), $\delta_j < 0$ and $\delta_k > 0$ and hence $L(P_{11}^*, \ldots, P_{IT}^*)$ increases because of (34). This contradicts the assumption that $L(P_{11}^*, \ldots, P_{IT}^*)$ was a global maximum. By a similar argument it follows that, if there is at least one n_{1t} or n_{IT} greater than zero, then for all $(1, t')$ with $n_{1t'} = 0$ we must have $P_{1t'}^* = 0$ and for all (I, t') with $n_{It'} = 0$ we must have $P_{It'}^* = 0$. Further if all $n_{1t} = n_{It} = 0$, there is at most one $P_{1t}^* > 0$ or at most one $P_{It}^* > 0$. This means that the search for a global maximum of $L(P_{11}, \ldots, P_{IT})$ can be confined to the following three cases: (*a*) For all (i, t) with $n_{it} = 0$ we can fix $P_{it}^* = 0$ except for one pair $(1, t)$ for which $P_{1t}^* = P^* > 0$; (*b*) for all (i, t) with $n_{it} = 0$ we can fix $P_{it}^* = 0$ except for one pair (I, t) for which $P_{It}^* = P^{**} > 0$; (*c*) for all (i, t) with $n_{it} = 0$ we can fix $P_{it}^* = 0$.

We now derive the Lagrangian necessary conditions for the cases (*a*), (*b*) and (*c*). For case (*a*) the Lagrangian equations are

$$P_{it}^*(\lambda + x_i) = \mu n_{it} \quad \text{for} \quad n_{it} > 0, \tag{37}$$

$$\lambda + x_1 = 0 \quad \text{for} \quad n_{1t} = 0. \tag{38}$$

Using (31) and (32) we obtain

$$P_{it}^* = \frac{n_{it}(\bar{X} - x_1)}{n(x_i - x_1)} \quad \text{for} \quad n_{it} > 0, \tag{39}$$

$$P^* = 1 - \Sigma\Sigma \frac{n_{it}(\bar{X} - x_1)}{n(x_i - x_1)} \quad (i \neq 1), \tag{40}$$

which yields a contradiction unless $P^* < 1$, i.e.

$$\Sigma\Sigma \frac{n_{it}(\bar{X} - x_1)}{n(x_i - x_1)} < 1 \quad (i \neq 1). \tag{41}$$

It will be seen below that the necessary condition (41) for a global maximum of type (*a*) to occur is usually violated for moderate n and, therefore, we do not pursue this case further.

Similarly the necessary condition for a global maximum of type (*b*) to occur is

$$\Sigma\Sigma \frac{n_{it}(x_I - \bar{X})}{n(x_I - x_i)} < 1 \quad (i \neq I). \tag{42}$$

The condition (42) will be usually violated for moderate n and, therefore, we do not pursue case (*b*) any further.

For case (*c*) the Lagrangian equations are

$$P_{it}^*(\lambda + x_i) = \mu n_{it} \quad \text{for} \quad n_{it} > 0 \tag{43}$$

and hence, summing over i and t,

$$\lambda + \bar{X} = n\mu, \tag{44}$$

from which we obtain

$$P_{it}^* = \frac{n_{it}}{n}\left(1 + \frac{x_i - \bar{X}}{n\mu}\right)^{-1}. \tag{45}$$

Summing (45) over i and t we get

$$1 = \Sigma\Sigma \frac{n_{it}}{n}\left(1 + \frac{x_i - \bar{X}}{n\mu}\right)^{-1}, \tag{46}$$

which has to be solved for $n\mu$.

It will be seen now that (46) has a finite solution $n\mu$ if (41) and/or (42) are violated; however, the solution is not a global maximum unless both (41) and (42) are violated. Writing

$$f(n\mu) = \Sigma\Sigma \frac{n_{it}}{n}\left(1 + \frac{x_i - \bar{X}}{n\mu}\right)^{-1}, \tag{47}$$

it is seen that $f(\bar{X}-x_1)$ is the left-hand side of (41) so that a violation of (41) implies

$$f(\bar{X}-x_1) \geqslant 1.$$

On the other hand, since $f(0) < 1$ it follows that there is a root $n\mu$ of (47) with

$$0 < n\mu \leqslant \bar{X}-x_1. \tag{48}$$

Similarly, it is seen that there is a root $n\mu$ of (47) with

$$\bar{X}-x_I \leqslant n\mu < 0 \tag{49}$$

if (42) is violated.

It can be shown that the probability that (41) is satisfied can be approximated by the normal tail area beyond $z_1 = CV(1)\sqrt{n}$ where $CV(1)$ is the coefficient of variation of the $x_i - x_1$. The probability that (42) is satisfied can be approximated by the normal tail area beyond $z_I = CV(I)\sqrt{n}$, where $CV(I)$ is the coefficient of variation of the $x_I - x_i$. Since these probabilities are negligible for moderate n and not too small $CV(1)$ and $CV(I)$, the conditions (41) and (42) will usually be violated. As an example, consider data of $n = 49$ cities drawn from a population of $N = 196$ cities by Cochran (1963, p. 156). Choosing $x_1 = 0$, we obtain for the right-hand side of (41) the value $3{\cdot}09 > 1$.

We do not, in this paper, investigate the properties of the maximum-likelihood estimator of $\bar{Y}$ resulting from (45) and (46), namely

$$\hat{\bar{Y}} = \Sigma\Sigma \hat{P}_{it} y_t = \Sigma\Sigma \frac{n_{it}}{n} y_t \left(1+\frac{x_i-\bar{X}}{n\mu}\right)^{-1}. \tag{50}$$

It is worth noting, however, that if all observed $y_t = \bar{y}$ (50) yields $\hat{\bar{Y}} = \bar{y}$ in analogy to the customary regression estimator.

We now develop an approximation to (50) valid for moderately large n. In most practical situations of x-distributions and samples it will be possible to solve (46) by a value $n\mu$ such that $(x_i-\bar{X})/(n\mu) \ll 1$ for all feasible x_i, so that the solution can be obtained by expanding (46) to first three terms. Hence we obtain

$$\frac{1}{n\mu} \simeq \frac{n(\bar{x}-\bar{X})}{\Sigma\Sigma\, n_{it}\,(x_i-\bar{X})^2}, \tag{51}$$

where $\bar{x} = n^{-1}\Sigma\Sigma n_{it} x_i$ is the sample mean. Using (51) and expanding (45) to the first three terms, we obtain

$$\hat{P}_{it} \simeq \frac{n_{it}}{n}\left[1+\frac{n(x_i-\bar{X})\,(\bar{X}-\bar{x})}{\Sigma\Sigma n_{it}\,(x_i-\bar{X})^2}+\frac{n^2(x_i-\bar{X})^2\,(\bar{X}-\bar{x})^2}{\{\Sigma\Sigma n_{it}(x_i-\bar{X})^2\}^2}\right]. \tag{52}$$

From (52), an approximation to $\hat{\bar{Y}} = \Sigma\Sigma \hat{P}_{it}\, y_t$ is given by

$$\hat{\bar{Y}} \simeq \bar{y} + b_1(\bar{X}-\bar{x}) + b_2(\bar{X}-\bar{x})^2, \tag{53}$$

where $\bar{y} = n^{-1}\Sigma\Sigma n_{it}\, y_t$ is the sample mean,

$$b_1 = \frac{\Sigma\Sigma n_{it} y_t (x_i-\bar{X})}{\Sigma\Sigma n_{it}(x_i-\bar{X})^2} \tag{54}$$

and

$$b_2 = \frac{n\Sigma\Sigma n_{it} y_t (x_i-\bar{X})^2}{\{\Sigma\Sigma n_{it}(x_i-)^2\}\bar{X}^2}. \tag{55}$$

If all observed $y_t = \bar{y}$, (53) reduces to $\hat{\hat{\bar{Y}}} = \bar{y}$. The contributions from the last term on the right-hand side of (53) will usually be small compared to the other terms for moderately large n, so that a further approximation to $\hat{\hat{\bar{Y}}}$ is

$$\hat{\hat{\bar{Y}}} \simeq \bar{y} + b_1(\bar{X} - \bar{x}). \tag{56}$$

Although (54) differs slightly from the customary definition of the sample regression coefficient, the above development clearly shows that, at least in large samples, the customary regression estimator is essentially the maximum-likelihood estimator.

4. Simple random sampling with replacement

Although this method is seldom used in practice, we nevertheless consider it here in view of the considerable attention it has received in recent years (e.g. Basu, 1958; Des Raj & Khamis, 1958; Pathak, 1962; Godambe, 1965) and to show that identifying labels are not informative beyond the design stage. Basu (1958) has shown that the sample mean based only on the distinct units in the sample is uniformly better than the customary sample mean based on all sample draws. Godambe (1965) has used Basu's result to emphasize the concept of identifiability of units in sample surveys.

A random sample of fixed size m is drawn with equal probability and with replacement. Let n denote the number of distinct units in the sample and n_t the number of distinct units having the value y_t in the sample. Then clearly

$$n_t \geqslant 0, \quad \Sigma n_t = n. \tag{57}$$

For fixed n, the conditional likelihood is given by the multidimensional hypergeometric distribution. Consequently the total likelihood is

$$L(N_1, \ldots, N_T) = P(n) \frac{\Pi \binom{N_t}{n_t}}{\binom{N}{n}}, \tag{58}$$

where the probability $P(n)$ is a function only of m and N. Clearly $(n_1, \ldots, n_T)$ is sufficient for $(N_1, \ldots, N_{T-1})$, but not complete sufficient, so that no U.M.V. estimator exists.

For the case of integral N/n, it follows from § 2·2 and (58) that the maximum-likelihood estimator of N_t is given by

$$\hat{N}_t = \frac{N}{n} n_t \tag{59}$$

and, hence, that of μ_r' is

$$m_r' = n^{-1} \Sigma n_t y_t^r. \tag{60}$$

In particular, the maximum-likelihood estimator of $\bar{Y}$ is identical to Basu's estimator. If $N = c \times$ least common multiple of $1, 2, \ldots, m$, where c is an integer, then N/n is an integer for *any* n with $1 \leqslant n \leqslant m$ and the estimators (59) and (60) are all maximum-likelihood estimators for *any* sampling outcome.

With the compound multinomial prior distribution (22), the Bayes estimator of μ_r', the posterior variance of μ_r' and the Bayes estimator of σ^2 are respectively given by (23), (24) and (25), where n now denotes the number of distinct units in the simple random sample drawn with replacement and the n_t are given by (57).

We wish to thank Professors D. A. S. Fraser and B. K. Kale, as well as the referee, for helpful suggestions concerning this paper. The work was supported by the U.S. Army Research Office. Sections 2·1, 2·2, and 4 of our paper have a considerable overlap with the results independently derived by Richard M. Royall. Both the contents of the present paper and Royall's paper *An old approach to finite population sampling theory* were presented at the I.M.S. Annual Meetings, Washington, D.C., December 1967.

References

Basu, D. (1958). On sampling with and without replacement. *Sankhyā* **20**, 287–94.

Cochran, W. G. (1963). *Sampling Techniques*. 2nd ed. New York: Wiley.

Des Raj & Khamis, S. H. (1958). Some remarks on sampling with replacement. *Ann. Math. Statist.* **29**, 550–7.

Godambe, V. P. (1955). A unified theory of sampling from finite populations. *J. R. Statist. Soc.* B **17**, 267–77.

Godambe, V. P. (1965). A review of the contributions towards a unified theory of sampling from finite populations. *Rev. Int. Statist. Inst.* **33**, 242–58.

Kempthorne, O. (1965). Some aspects of experimental inference. *J. Am. Statist. Ass.* **61**, 11–34.

Pathak, P. K. (1962). On simple random sampling with replacement. *Sankhyā* **24**, 287–302.

Rao, C. R. (1965). *Linear Statistical Inference and its Applications*. New York: Wiley.

Raiffa, H. & Schlaifer, R. O. (1961). *Applied Statistical Decision Theory*. Boston: Division of Research, Graduate School of Business Administration, Harvard University.

[*Received December* 1967. *Revised April* 1968]