Survey Methodology, December 1984
Vol. 10, No. 2, pp. 181-185
Statistics Canada

Unbiased Estimation of Domain Parameters in Sampling without Replacement

ARIJIT CHAUDHURI AND RAHUL MUKERJEE[1]

ABSTRACT

A finite population of size N is supposed to contain M (unknown) units of a specified category A (say) constituting a domain with mean μ. A procedure which involves drawing units using simple random sampling without replacement till a preassigned number of members of the domain is reached is proposed. An unbiased estimator of μ is also derived. This is seen to be superior to the corresponding possibly biased estimator based on a comparable SRSWOR scheme with a fixed number of draws. The proposed scheme is also shown to admit unbiased estimators of M and the domain total T.

KEY WORDS: Domain estimation; Simple random sampling without replacement.

1. INTRODUCTION

In large scale sample surveys, utilization of available resources and consideration for efficiency often demand realization in a sample of adequate representation from a specified category (A, say) of members with required characteristics. For example, clients and users of survey data may insist on estimates from a sample with a specified (m, say) number: (1) of farmers (i) using a particular fertilizer, (ii) employing a particular irrigation and cultivation technique and (iii) ready to respond truthfully to queries made; (2) of manufacturers using iron and steel with a specific purpose; (3) of household members with a requisite academic qualification, etc. While designing a sampling plan for the purpose, in spite of careful efforts, it is often possible that 'frames' may not be accurately constructed. The faulty list may be supposed to include N units which are well in excess over the M genuine units of the required A-category. Hence arises a problem of sampling to yield estimators for the mean, (and also total and size), of the domain of A-members. A solution to this problem is attempted below using 'inverse' SRSWOR scheme. Inverse sampling plans with replacement are, however, available in the literature (vide Haldane 1945, Sampford 1962 among others) for estimating the proportion $f = M/N$ of domain elements. Domain estimators for μ are also given by Rao (1975) but they are ratio estimators and are not unbiased. The proposed inverse SRSWOR scheme is seen to admit an unbiased estimator of μ which is more efficient than the corresponding possibly biased estimator based on a comparable SRSWOR scheme with the fixed number of draws.

2. A METHOD OF SAMPLING AND ESTIMATION

The population $I_N = (1, ..., j, ..., N)$ is supposed to consist of N units labelled $1, ..., j, ..., N$ and valued $y_1, ..., y_j, ..., y_N$. Of them some M (unknown) units possess certain exclusive features to constitute a class or domain, say A. In practice, some idea about M is usually available and let the parameter space for M be $\mathcal{M} = \{r, r+1, ..., R\}$, where $r(\geq 1)$ and $R(\leq N)$ are known. In almost all real life situations r will be much greater than 1 and R much less than N.

[1] Arijit Chaudhuri and Rahul Mukerjee, Indian Statistical Institute, 203 Barrackpore Trunk Road, Calcutta 700 035, India.

Writing $X_1, ..., X_i, ..., X_M$ as the y_j values for the M units of A, estimators are required for $\mu = (\Sigma_1^M X_i)/M$, and perhaps also for M and $T = \Sigma_1^M X_i$, from a sample containing a preassigned number, say m ($\leq r$), of units of A. The expressions for the variances of these estimators, presented later as functions of m, may be employed for an appropriate choice of m. For convenience, we shall write $X_{M+1} = \ldots = X_N = 0$ for the 'non-A' units of I_N.

Let units be chosen in successive draws by SRSWOR till exactly m units of A are realized. The number of draws, u, is then a random variable with a probability distribution $P_M(\cdot)$ (say, depending on the unknown parameter M) which is given by

$$P_M(u=n) = \frac{\binom{M}{m-1}\binom{D}{n-m}}{\binom{N}{n-1}} \cdot \frac{M-m+1}{N-n+1} = g_{Mn} \text{ (say) } (m \leq n \leq D+m), \tag{2.1}$$

where $D = N - M$. To avoid trivialities, hereafter, we shall make the reasonable assumption that m $\geq$ 2. Then the following results hold for the above inverse sampling scheme.

Lemma 2.1. Every parametric function $f(M)$ is unbiasedly estimable.

Proof. Let $h(u)$, if available be an unbiased estimator (UE) of $f(M)$. Then

$$f(M) = \sum_{n=m}^{D+m} h(n)\, g_{Mn},\ r \leq M \leq R. \tag{2.2}$$

If the above system of $R - r + 1$ equations in $N - r + 1$ unknowns $h(m), ..., h(N - r + m)$ be written in matrix notation, then the fact that $g_{Mn} > 0$ ($m \leq n \leq D + m, r \leq M \leq R$) implies that the resulting coefficient matrix is of full row rank. This guarantees the existence of a solution and completes the proof.

Remark. In particular, if $R = N$ then the number of equations in (2.1) equals the number of unknowns. As such the coefficient matrix becomes nonsingular and every parametric function $f(M)$ becomes uniquely unbiasedly estimable.

Corollary 2.1. A UE of M based on u is $\hat{M}(u) = N(m-1)/(u-1) = \hat{M}$ (say).

Proof. First observe that the assumption $m \geq 2$ ensures that $u > 1$ with probability 1 (whatever be M) so that $\hat{M}(u)$ is well defined. Now

$$E\left(\frac{1}{u-1}\right) = \sum_{n=m}^{D+m} \frac{1}{n-1} g_{Mn}$$

$$= \frac{M!\, D!}{(M-m)!\,(m-1)!\, N!} \sum_{n=m}^{D+m} \frac{(n-2)!\,(N-n)!}{(n-m)!\,(D-n+m)}$$

$$= \frac{M!\, D!}{(M-m)!\,(m-1)!\, N!} \cdot \frac{(M-m)!\,(m-2)!\,(N-1)!}{(M-1)!\, D!} = \frac{M}{N(m-1)},\ \forall\, M \in \mathcal{M}.$$

Hence the result.

Remark. The relation (2.1) and Lemma 2.1 may be employed to find $V_M(\hat{M})$ and a UE of this variance. The resulting algebraic expression, although straightforward to evaluate numerically in any practical situation, are somewhat involved and will not be presented here.

In the following, $S^2 = (M-1)^{-1} \Sigma_1^M (x_i - \mu)^2$, $q(u)$ and $\ell(u)$ are any UE's for M^{-1} and M^2 respectively (available by (2.2) above) Σ' denotes summation over the A-units included in the sample, $\bar{x} = m^{-1}\Sigma' X_i$, $Z = m^{-1}\Sigma' X_i^2$ and $s^2 = (m-1)^{-1}\Sigma'(X_i - \bar{x})^2$.

Theorem 2.1. A UE of μ is $\bar{x}$ with $V_M(\bar{x}) = S^2(1/m - 1/M)$. A UE of $V_M(\bar{x})$ is given by $v(\bar{x}) = s^2(m^{-1} - q(u))$.

Proof. Easy and hence omitted.

Theorem 2.2. (i) A UE of T is $\hat{T} = \hat{M}\bar{x}$ with

$$V_M(\hat{T}) = S^2(1/m - 1/M)\, E_M(\hat{M}^2) + \mu^2 V_M(\hat{M}).$$

(ii) $v(\hat{T}) = \hat{T}^2 - [\ell(u)(Z - s^2) + \hat{M}s^2]$ is a UE of $V_M(\hat{T})$.

Proof. The proof of (i) is easy and hence omitted. To prove (ii) note that

$$E\,[\,\{\ell(u)(Z - s^2) + \hat{M}s^2\}\,|\,u\,]$$

$$= \ell(u)(M^{-1}\sum_1^M X_i^2 - S^2) + \hat{M}S^2 = \ell(u)(\mu^2 - M^{-1}S^2) + \hat{M}S^2.$$

Hence

$$E_M v(\hat{T}) = E_M(\hat{T}^2) - [M^2(\mu^2 - M^{-1}S^2) + MS^2] = E_M(\hat{T}^2) - T^2 = V_M(\hat{T}).$$

3. COMPARISON WITH SRSWOR WITH A FIXED NUMBER OF DRAWS

In this section, first it will be shown that if one insists on unbiased estimation of μ then our strategy will be superior to the one based on SRSWOR with a fixed number of draws. Secondly, this superiority will be demonstrated even when biased estimators are allowed.

Let d be a fixed (somehow) number of draws in SRSWOR sampling, $\hat{s}$ a sample so drawn, $\hat{s} \cap A$ the set of A-units in $\hat{s}$ and C the cardinality of $\hat{s} \cap A$. We will use, for this scheme also previous notations P_M, E_M, V_M to imply phenomena relevant here. Then for such a sampling we have:

Theorem 3.1. μ admits a UE if and only if $d \geq N - r + 1$.

Proof. Let $d \geq N - r + 1$. Then $P_M[c = 0] = 0$, $\forall\, M \in \mathcal{M}$ and $\hat{\mu} = c^{-1}\Sigma' X_i$ is a UE of μ.

To prove the necessity it will be enough to show that if $d = N - r$, then μ does not admit a UE. For this the following notations will be used. let $j_1, \ldots, j_d$ be d distinct increasingly ordered units out of $1, \ldots, N$, constituting the elements of $\hat{s}$ and such that some k of them ($0 \leq k \leq d$), say $i_1, \ldots, i_k$ (increasingly ordered) belong to A. Then we write $\hat{s} = (j_1, \ldots, j_d)$, $\hat{s}' = (i_1, \ldots, i_k) = \hat{s} \cap A$ (so that $k = 0 \Rightarrow \hat{s}' = \Phi$ and $k = d \Rightarrow \hat{s}' = \hat{s}$) and $X(\hat{s}') = (X_{i_1}, \ldots, X_{i_k})$, a sequence of X_i values for the units in $\hat{s}'$. Then if there exists a UE for μ, say t, we may write $t = t(X(\hat{s}')|\hat{s})$ such that

$$E_M(t) = \mu,\ \forall\, X_1, \ldots, X_M,\ \forall\, M \in \mathcal{M}. \tag{3.1}$$

For $0 \leq k \leq d$, let $t_k = \Sigma_k t(X(\hat{s}')|\hat{s})$, Σ_k being sum over all samples with exactly k A-units. Clearly t_o is free from X_i's.

If $d = N - r$, then $\mathcal{M} = \{N - d, N - d + 1, \ldots, N\}$. Suppose $M = N - d + j$ ($0 \leq j \leq d$). Then the A-units may be chosen in $\binom{N}{M} = \binom{N}{d-j}$ ways. Accordingly $\binom{N}{d-j}$ equations are involved in (3.1). Summing over the number of ways of choosing the A-units, (3.1) yields

$$\sum_{w=j}^{d} a_{jw} t_w = \frac{\binom{N}{d}\binom{N-1}{d-j} T}{N - d + j}, \tag{3.2}$$

where, for $0 \leq j \leq w \leq d$, $a_{jw} = \binom{N-d}{w-j}$ if $N - d \geq w - j$; 0, otherwise. From (3.2) the solutions for the t_w's may be obtained as

$$t_w = \binom{N}{d}\binom{d}{w}\frac{T}{N}, \quad 0 \leq w \leq d, \tag{3.3}$$

and the validity of (3.3) follows from the fact that

$$\sum_{w=j}^{d}\binom{N-d}{w-j}\binom{d}{w} = \binom{N}{d-j}.$$

In particular, (3.3) yields $t_o = N^{-1}\binom{N}{d}T$. But then t_o is not free from the X_i's implying a contradiction, proving the necessity and completing the proof.

Thus with a fixed size (d) SRSWOR scheme, for unbiased estimation of μ we need $d \geq N - r + 1$ which may become too large (especially if r is small) making the scheme operationally inconvenient. Even if $d \geq N - r + 1$, the fixed size SRSWOR scheme together with the UE $\hat{\mu} = c^{-1}\Sigma' X_i$ can be seen to be less efficient than the strategy described in the preceding section when compared at equal level of cost of inspection.

To elaborate, suppose $d \geq N - r + 1$ and note that

$$V_M(\hat{\mu}) = S^2\left[E_M(1/c) - 1/M\right]. \tag{3.4}$$

For our inverse sampling scheme, by (2.1) the expected number of draws is given by $m(N+1)/(M+1)$ and, to make our scheme comparable to a fixed size (d) scheme, this should equal d i.e. one should have $m = d(M+1)/(N+1)$, in which case Theorem 2.1 yields

$$V_M(\bar{x}) = S^2\left[\frac{N+1}{d(M+1)} - \frac{1}{M}\right]. \tag{3.5}$$

Since

$$E_M(c^{-1}) > [E_M(c)]^{-1} = \frac{N}{dM} > \frac{N+1}{d(M+1)},$$

it follows that (3.4) is greater than (3.5), proving our assertion.

It is also interesting to compare our strategy with the fixed size scheme when a possibly biased estimator of μ is allowed in the latter. In fixed size (d) SRSWOR scheme, consider the usual (ratio) estimator of μ given by [vide e.g. Rao (1975)]

$$\mu^* = c^{-1}\Sigma' X_i \quad \text{if } c > 0$$
$$= 0 \quad \text{if } c = 0$$

The bias in μ^* equals $-\mu P_M(c = 0)$ (observe that if $d \geq N - r + 1$, then $P_M(c = 0) = 0$, $\forall M \in \mathcal{M}$ and μ^* reduces to the UE $\hat{\mu}$ defined earlier) and it can be shown that

$$MSE_M(\mu^*) = S^2\sum_{a \geq 1}(1/a - 1/M)P_M(c = a) + \mu^2 P_M(c = 0). \tag{3.6}$$

A straightforward analytic comparison between (3.5) and (3.6) is difficult but as numerical examples including the two cited below suggest, in most practical situations (3.5) will be smaller than (3.6), indicating the superiority of our strategy even when a possibly biased estimator is allowed in the fixed size scheme.

Example 3.1. The following data relate to the aggregate percentage of marks of all the students who passed the Bachelor of Statistics Examination of the Indian Statistical Institute (ISI) during the last five academic years ended 1984[1].

[1] The data are obtained from the office of the ISI Dean of Studies to whom the authors are grateful for granting an access to them.

68	80	80	72	87	71	55	75	85	52	82
76	73	54	57	51	56	48	73	54	76	69
87	81	68	74	58	56	71	66	69	81	59
65	83	79	72	50	44	65	61	57	50	73
85	87	64	70	48	58	61	53	56	62	61
74	62	56	62	58	58	66	70	80	74	80

Suppose it is desired to estimate from a sample the mean score of those students who obtained a first class (i.e. sixty percent or above). Then $N = 66$, $M = 44$, $\mu = 73.1818$, $S^2 = 61.6871$. For a fixed size SRSWOR scheme with $d = 10$, (3.6) equals 9.5967. The comparable m in our inverse sampling strategy is $d(M + 1)/(N + 1) = 6.72$ and with this $= 6, 7$, (3.5) equals 8.8792, 7.4105 and the resulting gains in efficiency, compared to the fixed size scheme, are 8.08 and 29.50 percent respectively.

Example 3.2. As a somewhat less traditional example, consider the problem of estimating the mean of the prime numbers among the first sixty natural numbers. The $N = 60$, $M = 18$, $\mu = 24.5$, $S^2 = 350.1471$. For a fixed size SRSWOR scheme with $d = 7$, the value of (3.6) is 205.4654. The comparable inverse sampling strategy requires $m = d(M + 1)/(N + 1) = 2.18$ and with this $= 2$, (3.5) equals 155.6209, indicating a gain in efficiency by 32.03 percent.

4. CONCLUDING REMARKS

In this paper we have considered the estimation problem for a single domain. In large scale surveys estimators are often required for several domains in which case the present procedure may be modified as follows.

Let there be t domains, the domain sizes M_k being unknown, having respective parameter spaces $\mathcal{M}_k = \{r_k, r_{k+1}, ..., R_k\}$, where r_k and R_k are known ($1 \leq k \leq t$). Let μ_k and S_k^2 and denote the population mean and variance of the study variate in the kth domain. The sampling scheme may be inverse generalized hypergeometric, i.e. inverse SRSWOR may be continued till at least $m_1, m_2 ..., m_t$ ($m_k \leq r_k$ for each k) units of the 1st, 2nd, ..., tth domains are realized. For each k, clearly the number of units, say ξ_k, in the sample from the kth domain is now a random variable, with $P_M(\xi_k \geq m_k) = 1$ (where $m = (M_1, ..., M_t)'$, P_M, E_M the corresponding probability and expectation operator), since even when the quota for kth domain is filled up, sampling may have to be continued to fill up those for the other domains thus possibly including in the sample some additional units from the kth domain. The mean, $\bar{x}_k$, of the units in the sample from the kth domain is a UE of $\mu_k$with a variance $S^2 [E_M(1/\xi_k) - 1/M_k]$, $1 \leq k \leq t$.

In this set-up also numerical investigations (records omitted since they seem uninteresting in the present context) suggest that the inverse sampling strategy will be more efficient than the one based on fixed size SRSWOR when compared at the same level of cost. For multidomain situations, however, the detailed algebraic expressions become somewhat involved so that an analytic comparison along the line of the preceding section becomes difficult and hence not reported here.

ACKNOWLEDGEMENT

The authors are thankful to the referee for his highly constructive suggestions that helped improvement on an earlier draft.

REFERENCES

Haldane, J.B.S. (1945). On a method of estimating frequencies. *Biometrika,* 33, 222-225.

Rao, J.N.K. (1975). Analytical studies of sample survey data. *Survey Methodology,* 1, 1-76.

Sampford, M.R. (1962). Methods of cluster sampling with and without replacement for clusters of unequal sizes. *Biometrika,* 49, 27-40.