Jour. Ind. Soc. Ag. Statistics
57 (Special Volume), 2004 : 129-133

Unbiased Variance Estimation on Sub-sampling from a Varying Probability Sample

Arijit Chaudhuri
Indian Statistical Institute, Kolkata

SUMMARY

A simple procedure is presented to estimate unbiasedly a survey population total and the variance of the estimator for the total based on an unequal probability sub-sample from an initially drawn sample by Rao *et al.* (RHC [4]) scheme from the population.

Key words : Rao-Hartley-Cochran scheme, Sub-sampling, Unbiased variance estimation.

1. Introduction

Recently, Indian Statistical Institute (ISI), Kolkata, implemented an audit sampling procedure to help the internal Audit Cell of the Ministry of Finance, Government of West Bengal. For this, from a sample of districts several offices stratified by divisions like Public Works, Irrigation etc. were selected following the scheme of Rao *et al.* (RHC [4]) leaving provisions for sampling at subsequent stages from the books, pages and lines hierarchically contained therein. Previous year's budget allocations provided the size-measures.

But at the planning stage itself resource crunches dictated rather drastic cut in the realized size of the sample drawn according to the RHC scheme. This necessitated notable adjustments in the estimation procedures. In Section 2 we present a relevant theory in brief.

2. Theory of Estimation in Sub-sampling from a Sample Chosen by RHC Scheme

Let $U = (1, \ldots, i, \ldots, N)$ denote a survey population, $Y = (y_1, \ldots y_i, \ldots, y_N)$, $P = (p_1, \ldots p_i, \ldots, p_N)$ with y_i as the value of a variable y and $p_i (0 < p_i < 1, \Sigma p_i = 1)$ as the known normed size-measure for the unit i in U, writing Σ to denote summing over i in U. In order to unbiasedly estimate $Y = \Sigma y_i$, the scheme of selecting a sample of $n \; (2 \le n < N)$ units from U given by Rao *et al.* (RHC [4]) consists first in fixing n integers $N_i (i = 1, \ldots, n)$ subject to $\Sigma_n N_i = N$, dividing U into n non-overlapping groups with the i^{th} group containing N_i distinct units of U, Σ_n denoting addition over the n groups. Then writing $Q_i = p_{i1} + \ldots + p_{iN}$ as the sum of the normed size-measures of the $'N_i$ units falling in the i^{th} group it chooses from the i^{th} group unit ij with a probability $\frac{p_{ij}}{Q_i}$, $j = 1, \ldots, N_i$ and repeats this independently for each of the n groups. Based on the resulting sample denoted by s, an unbiased estimator for Y given by RHC [4] is

$$t = \Sigma_n y_i \frac{Q_i}{p_i}$$

writing for simplicity (y_i, p_i) as the y-value and normed size-measure for the unit chosen from the i^{th} group, suppressing the subscript j. RHC [4] have also given $V(t) = A\left[\Sigma \frac{y_i^2}{p_i} - Y^2\right]$ as the variance of t and $\hat{V}(t) = B\left[\Sigma_n Q_i \frac{y_i^2}{p_i^2} - t^2\right]$ as an unbiased estimator for V(t), writing $A = \frac{\Sigma_n N_i^2 - N}{N(N-1)}$ and $B = \frac{(\Sigma_n N_i^2 - N)}{(N^2 - \Sigma_n N_i^2)}$.

Suppose, to save time and resources, it is felt necessary to survey not all the n units sampled as above but to restrict the field work only to a sub-sample of m $(2 \leq m < n)$ units to be suitably selected from s. To proceed accordingly let us observe that $0 < Q_i < 1, \Sigma_n Q_i = 1$ and on writing $w_i = mQ_i$, it follows that $\Sigma_n w_i = m$ and in case

$$w_i < 1 \quad \forall i \in U \tag{2.1}$$

such a w_i subject to (2.1) may be taken as the "inclusion-probability" of any of the n units of s, say i if now selected in a sub-sample of m units out of them. First we suppose (2.1) holds. Later we shall relax this.

Case I. (2.1) holds

Here we propose drawing a sample u of m distinct units of s using Q_i for i in s as the normed size-measures of the respective units. Of course RHC scheme itself may be employed with the necessary adjustments in this context. But more generally one may employ any scheme for which w_i is achieved as the inclusion -probability of i in the sample and some numbers w_{ij} satisfying

$$0 < w_{ij} < 1, \sum_{j \neq i} w_{ij} = (m-1)w_i, \sum\sum_{i \neq j} w_{ij} = m(m-1) \tag{2.2}$$

are realized as the inclusion-probabilities of the pairs of units $i, i\ (i \neq j)$ in the sample of size m from s. Then, let us write $z_i = y_i \dfrac{Q_i}{p_i}$ and propose to employ for Y the revised estimator

$$e = \Sigma_m \frac{z_i}{w_i} \tag{2.3}$$

writing Σ_m to denote sum over the m units in the subsample u from s - this of course is nothing but the Horvitz-Thompson (HT [3]) estimator for t given s. Later we shall write $\Sigma_m \Sigma_m$ to denote sum over distinct pairs of units in u with no duplication.

Let us write (E_p, V_p), (E_R, V_R), (E, V) as the expectation, variance operators over sampling of s from U, u from s and u from U. Then further noting that

$$E = E_p E_R \text{ and } V = E_p V_R + V_p E_R$$

we get the following theorem

Theorem. (a) $E(e) = Y$

(b) $Ev(e) = V(e)$, where

$$v(e) = (1+B) v_R(e) + B\left(\Sigma_m \frac{z_i^2}{Q_i w_i} - e^2 \right)$$

and $v_R(e) = \Sigma_m \Sigma_m (w_i w_j - w_{ij}) \left(\dfrac{z_i}{w_i} - \dfrac{z_j}{w_j} \right)^2 \dfrac{I_{ij}(u)}{w_{ij}}, I_{ij}(u) = 1 \text{ if } i, j \in u, 0 \text{ else}$

Proof. (a) $E_R(e) = \Sigma_n z_i = t$ and $E(e) = E_p(t) = Y$

(b) $V(e) = E_p V_R(e) + V_p E_R(e) = E_p E_R v_R(e) + V(t)$ because $v_R(e)$ is the Yates -Grundy (YG [5]) unbiased estimator of

$$V_R(e) = \Sigma_m \Sigma_m (w_i w_j - w_{ij}) \left(\frac{z_i}{w_i} - \frac{z_i}{w_j} \right)^2$$

$$= E_p E_R v_R(e) + E_p \left[B \Sigma_n \frac{z_i^2}{Q_i} - t^2 \right]$$

$$= E_p E_R v_R(e) + E_p \left[B \left\{ E_R \Sigma_m \frac{Z_i^2}{Q_i w_i} - E_R(e^2 - v_R(e)) \right\} \right]$$

$$= E_p E_R \left[(1+B) v_R(e) + B\left(\Sigma_m \frac{z_i^2}{Q_i w_i} - e^2 \right) \right]$$

So, $v(e) = (1+B)v_R(e) + B\left(\Sigma_m \frac{z_i^2}{Q_i w_i} - e^2 \right)$ is our proposed unbiased estimator of our proposed estimator e for Y in Case I.

Note. Though numerous schemes of sampling are available in the literature to answer our need to cover Case I we recommend the application of Circular systematic sampling (CSS) with probabilities proportional to sizes (PPS) using Q_i's suitably scaled up as integers X_i with an appropriate common multiplier, applying a random rather than a constant sampling interval as a number chosen at random between 1 and $(X - 1)$ with $X = \Sigma X_i$ as described by Chaudhuri and Pal [2].

Case II. (2.1) does not hold

Here we recommend selecting u from s applying CSSPPS with a random interval using X_i's as size-measures and making $(m - 1)$ further selections of units after the first. In this case we are assured that $w_{ij} > 0$ for every i, j in s. From Chaudhuri and Pal [1] we known that $V_R(e)$ is now modified into $V'_R(e) = V_R(e) + \Sigma_m a_i \frac{z_i}{w_i}$ where $a_i = \frac{1}{w_i}\left(\sum_{j=1}^{m} w_{ij} \right) - \Sigma_n w_i$ and $v_R(e)$ into $v'_R(e) = v_R(e) + \Sigma_m a_i \frac{z_i^2}{w_i} \frac{I_i(u)}{w_i}$ writing $I_i(u) = 1$ if $i \in u$ and 0 else.

So, our Theorem yields

Corollary. (a) $E(e) = Y$ and

(b) $Ev'(e) = V'(e)$, where $V'(e) = E_p V'_R(e) + V_p E_R(e)$ and

$$V'(e) = (1+B)V'_R(e) + B\left[\Sigma_m \frac{z_i^2}{d_i w_i} - e^2 \right]$$

Proof. Easy and hence omitted.

Note. $v'(e)$ is our proposed unbiased estimator for the variance of e in Case II.

Note. Instead of CSSPPS with a random interval any general scheme may be employed covering the Case II, with no formal change in the formula for $V'_R(e)$, $v'_R(e)$, $V(e)$ and $v'(e)$.

REFERENCES

[1] Chaudhuri, A. and Pal, S. (2002). On certain alternative mean square error estimators in complex survey sampling. *J. Statist. Plann. Inf.*, **104(2)**, 363-375.

[2] Chaudhuri, A. and Pal, S. (2003). Systematic sampling: Fixed versus random sampling interval. *Pak. Jour. Stat.*, **19(2)**, 259-271.

[3] Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.*, **77**, 89-96.

[4] Rao, J.N.K., Hartley, H.O. and Cochran, W.G. (1962). On a simple procedure of unequal probability sampling without replacement. *J. Roy. Statist. Soc.*, **B24**, 482-491.

[5] Yates, F. and Grundy, P.M. (1953). Selection without replacement from within strata with probability proportional to size. *J. Roy. Statist. Soc.*, **B15**, 253-261.