J. Ind. Soc. Agril. Statist.
59(2), 2005 : 115–117

On Sub-sampling by Rao-Hartley-Cochran (RHC) Scheme from an Initial RHC Sample to Save Resources in Estimating a Survey Population Total and the Variance of the Estimator of the Total

Arijit Chaudhuri and Uppala Srinivas
Indian Statistical Institute, Kolkata
(Received : June, 2004)

SUMMARY

Rao *et al.* (1962) have given a well-known method of selecting a sample of distinct units from a finite survey population admitting known positive normed size-measures of units, unbiasedly estimating the population total and also positively the variance of this estimator. Once an RHC sample is drawn, if the sample-size is judged too high before the actual survey is undertaken, then a sub-sample of a manageably reduced size may be appropriate under a constrained budget. We present revised estimation and variance estimation procedures if the sub-sample is again chosen following the RHC scheme with necessary adjustments.

Key words : Estimation of population total, Unequal probability sampling and sub-sampling, Variance estimation.

1. INTRODUCTION

Chaudhuri (2003) treated a practical problem of requiring to choose a sub-sample from an initial sample with a total size deemed feasible under a budget within a stipulated time, but later judged too high, before undertaking the survey. In this follow-up we present formulae for an unbiased estimator of the population total along with a positive-valued unbiased estimator for its variance when the initial sample is chosen following Rao-Hartley-Cochran (RHC (1962)) scheme utilizing certain known positive size-measures when the sub-sample is chosen therefrom adopting the same RHC scheme with necessary adjustments. Chaudhuri (2004) presented the theories for 2 alternative methods of drawing the sample from an initial RHC sample.

2. METHOD OF SAMPLING AND ESTIMATION

Let U = (1,2, . . . , i, . . ., N) denote a survey population of N units labeled i with values y_i, ($i \in U$) on a real variable y of interest and known x_i, (>0, $i \in U$) as size-measures, $Y = \sum y_i$ and $X = \sum x_i$ denoting their respective totals. By $p_i = x_i / X$ we mean the normed size-measures of the units and our immediate objective is to choose from U a sample s with a pre-assigned number of n distinct units in it, ascertain y_i for $i \in s$, use the survey data $d = (s, y_i / i \in s)$ to unbiasedly estimate Y employing a suitable statistic $t = t(d)$ providing in addition a positive valued unbiased estimator for $V_1(t)$, the variance of t.

A standard popular procedure for this is provided by Rao, Hartley and Cochran (RHC (1962)). Their sample selection scheme enjoins division of U into n random groups of N_i units (i = 1,2, . . ., n), $N_i \geq 1$, $N_1 + \ldots + N_i \ldots + N_n = N$, and selection of one unit ij from the i^{th} group with probability p_{ij} / Q_i, taking $Q_i = p_{i1} + \ldots + p_{ij} + \ldots + p_{iN_i}$ and repeating this independently over i = 1,2, . . . , n. Their unbiased estimator for Y is

$$t_{RHC} = \sum_n y_i \frac{Q_i}{p_i} = \sum_n z_i \text{ , say}$$

writing $\sum_n$ for sum over the n groups and (y_i, p_i) the y-value and normed size-measure for the unit chosen from the i^{th} group, and $z_i = y_i Q_i / p_i$, i = 1,2, . . ., n. Then they show that

$$V_1(t_{RHC}) = A\left(\sum_i \frac{y_i^2}{p_i} - Y^2\right)$$

$$= A\left[\sum\sum_{i<j} p_i p_j \left(\frac{y_i}{p_i} - \frac{y_j}{p_j}\right)^2\right] \qquad (2.1)$$

and

$$v(t_{RHC}) = B\left(\sum_n Q_i \frac{y_i^2}{p_i^2} - t_{RHC}^2\right) \qquad (2.2)$$

is an unbiased estimator for $V_1(t_{RHC})$, writing

$$A = \frac{\sum_n N_i^2 - N}{N(N-1)} \text{ and } B = \frac{\sum_n N_i^2 - N}{N^2 - \sum_n N_i^2}$$

Suppose limited resources do not permit us to survey all the units in s but a sub-sample u of size m (< n) suitably chosen from s may be surveyed to provide an alternative unbiased estimator for Y along with an unbiased estimator thereof. Let, from s of size n, the sub-sample u of size m be again chosen employing the RHC technique. Noting that $Q_1 + Q_2 + \ldots + Q_i + \ldots + Q_n$ equals 1, the Q_i's may be treated as normed size-measures of the units in s in implementing the selection of u. Writing $\sum_m$ as sum over the m groups into which s is now to be divided at random, q_j ($j = 1, \ldots, m$) as the sum of the Q_i-values falling in the j^{th} group of n_j units ($j=1, \ldots, m$), $n_1 + n_2 + \ldots + n_m = n$, (y_j, p_j) as the y-value and the original p_i value for the unit chosen from the j^{th} group now formed with a probability Q_{jk}/q_j say, noting $q_j = Q_{j1} + \ldots + Q_{jn}$, our proposed revised estimator for Y is

$$e_{RHC} = \sum_m \frac{y_j}{p_j} q_j \qquad (2.3)$$

Let E_1, E_2 and V_1, V_2 be the operators for expectation and variance for the selections of s, u respectively and E, V the over-all operators over selection of s followed by that of u and $E_2(\cdot/G)$, $V_2(\cdot/G)$ be the expectation, variance operators conditionally on the groups formed as fixed while selecting u and E_G, V_G the same over formation of these groups. Then, we have

Theorem 1. $E(e_{RHC}) = Y$

Proof.
$$E_2(e_{RHC}/G) = \sum_m \left[\sum_{k=1}^{n_j} \frac{y_{jk}}{p_{jk}} q_j \frac{Q_{jk}}{q_j}\right]$$

$$= \sum_n z_i = t_{RHC}$$

$$E_2(e_{RHC}) = t_{RHC}$$

$$E(e_{RHC}) = E_1(t_{RHC}) = Y$$

Now we work out $V(e_{RHC})$ through the following steps

$$V_2(e_{RHC}/G)$$

$$= \sum_m \left[\sum_{k<l}^{n_j} \sum^{n_j} \frac{Q_{jk}}{q_j} \frac{Q_{jl}}{q_j} \left(\frac{z_{jk}}{Q_{jk}/q_j} - \frac{z_{jl}}{Q_{jl}/q_j}\right)^2\right]$$

$$E_G V_2(e_{RHC}/G)$$

$$= \left(\frac{\sum_m n_j^2 - n}{n(n-1)}\right) \sum_{k<l}^{n} \sum^{n} Q_k Q_l \left(\frac{z_k}{Q_k} - \frac{z_l}{Q_l}\right)^2$$

$$= a\left[\sum_n \frac{z_j^2}{Q_j} - t_{RHC}^2\right]$$

$$a = \frac{\sum_m n_j^{\,2} - n}{n(n-1)}$$

$$V_2(e_{RHC}) = V_G[E_2(e_{RHC}/G)] + E_G[V_2(e_{RHC}/G)]$$

$$= a\left[\sum_n \frac{z_j^2}{Q_j} - t_{RHC}^2\right], \text{ because } V_G(t_{RHC}) = 0$$

Finally, we have

Theorem 2. For $v(e_{RHC}) = (1+B)v_2(e_{RHC})$

$$+ B\left(\sum_m \frac{y_j^2}{p_j^2} q_j - e_{RHC}^2\right)$$

$E(v(e_{RHC})) = V(e_{RHC})$ with $v_2(e_{RHC})$ given below as (2.4)

Proof. Let $v_2(e_{RHC})$ satisfy

$$E_2(v_2(e_{RHC})) = V_2(e_{RHC})$$

Then

$$E_2(v_2(e_{RHC})) = a\left[\sum_n \frac{z_j^2}{Q_j} - t_{RHC}^2\right]$$

$$= a\left[E_2\left(\sum_m \frac{z_j^2}{Q_j^2} q_j\right) - E_2\left(e_{RHC}^2 - v_2(e_{RHC})\right)\right]$$

or

$$(1-a)E_2(v_2(e_{RHC})) = aE_2\left[\sum_m \frac{z_j^2}{Q_j^2} q_j - e_{RHC}^2\right]$$

or

$$v_2(e_{RHC}) = \frac{a}{1-a}\left[\sum_m \frac{z_j^2}{Q_j^2} q_j - e_{RHC}^2\right] \quad (2.4)$$

or

$$v_2(e_{RHC}) = b\left[\sum_m \frac{y_j^2}{p_j^2} q_j - e_{RHC}^2\right]$$

where $b = \frac{a}{1-a} = \frac{\sum_m n_j^2 - n}{n^2 - \sum_m n_j^2}$

satisfies $E_2 v_2(e_{RHC}) = V_2(e_{RHC})$

Now, $V(e_{RHC}) = E_1 V_2(e_{RHC}) + V_1 E_2(e_{RHC})$

$$= E_1 E_2(v_2(e_{RHC})) + V_1(t_{RHC})$$

$$= E_1 E_2(v_2(e_{RHC}))$$

$$+ BE_1\left(E_2 \sum \frac{y_j^2}{p_j^2} q_j - E_2\left[e_{RHC}^2 - v_2(e_{RHC})\right]\right)$$

(using (2.1) and (2.2))

$$= E_1 E_2\left[(1+B)v_2(e_{RHC}) + B\left(\sum_m \frac{y_j^2}{p_j^2} q_j - e_{RHC}^2\right)\right]$$

3. A FINAL REMARK

It is trivially simple to illustrate agricultural, industrial or other relevant situations where the theories developed here may be gainfully applied. A curious reader may consult Chaudhuri (2003, 2004) for practical illustrations.

It is pertinent here to refer to the Report on Audit Sampling by Indian Statistical Institute, Kolkata, December 2003, for the Department of Public Works, Public Health, Engineering and Irrigation & Waterways of Government of West Bengal. It contains a discussion on how a sample of 79 first stage units (fsu) out of 236 fsu's, namely Divisional Offices in 7 districts of West Bengal chosen by Rao-Hartley-Cochran (RHC (1962)) scheme using the previous year's budget-allocated expenditures, had to be reduced to only 65 to save time and cost when the survey was well in progress strata-wise.

In the Agricultural Surveys by the West Bengal Government State Agricultural Statistics Authority (SASA) for crop estimation 20 percent of the 344 Mouzas in the state are required to be covered each year but at least 4 percent shortfall is encountered in practice. Error appraisals may utilize the technique presented in this paper as it essentially applies beyond RHC scheme as well.

ACKNOWLEDGEMENTS

The authors are happy to note certain comments from a referee leading to this revision on an earlier draft.

REFERENCES

Chaudhuri, A. (2003). Estimation from an under-covered sample in a complex survey for auditing. *Cal. Stat. Assoc. Bull.*, **54**, 213-214.

Chaudhuri, A. (2004). Unbiased variance estimation on sub-sampling from a varying probability sample. *J. Ind. Soc. Agril. Statist.*, **57** (Special Volume), 129-133.

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.