Metrika (2000) 52: 115–131

Metrika

Mean square error estimation in multi-stage sampling

Arijit Chaudhuri, Arun Kumar Adhikary and Shankar Dihidar

Indian Statistical Institute, 203 Barrackpore Trunk Road, Calcutta 700035, INDIA

Received: 19 February 1999

Summary: Suppose for a homogeneous linear unbiased function of the sampled first stage unit (fsu)-values taken as an estimator of a survey population total, the sampling variance is expressed as a homogeneous quadratic function of the fsu-values. When the fsu-values are not ascertainable but unbiased estimators for them are separately available through sampling in later stages and substituted into the estimator, Raj (1968) gave a simple variance estimator formula for this multi-stage estimator of the population total. He requires that the variances of the estimated fsu-values in sampling at later stages and their unbiased estimators are available in certain 'simple forms'. For the same set-up Rao (1975) derived an alternative variance estimator when the later stage sampling variances have more 'complex forms'. Here we pursue with Raj's (1968) simple forms to derive a few alternative variance and mean square error estimators when the condition of homogeneity or unbiasedness in the original estimator of the total is relaxed and the variance of the original estimator is not expressed as a quadratic form.

We illustrate a particular three-stage sampling strategy and present a simulation-based numerical exercise showing the relative efficacies of two alternative variance estimators.

Key words: Multi-stage sampling, Survey population, Variance estimation, Varying probability sampling.

AMS Subject Classification: 62 D05

1 Introduction

Suppose a finite survey population $U = (1, \ldots, i, \ldots, N)$ has N first stage units (fsu) with values y_i $(i = 1, \ldots, N)$ of a variable y of interest. Based on a sample s of fsu's suitably taken in the first stage of sampling with a probability

$p(s)$ from U let for the population total $Y = \sum y_i = y_1 + \cdots + y_N$, an estimator be taken in the from

$$t = \sum b_{si} I_{si} y_i. \tag{1.1}$$

Here $I_{si} = 1$ if $i \in s$; 0 otherwise; later we shall also use $I_{sij} = I_{si} I_{sj}$; b_{si}'s are constants free of $\underline{Y} = (y_1, \ldots, y_i, \ldots, y_N)$. Writing E_1 as the operator for expectation over the first stage of sampling, let

$$E_1(b_{si} I_{si}) = 1 \quad \text{for every } i \text{ in } U. \tag{1.2}$$

Then, $E_1(t) = Y$ i.e. t is unbiased for Y.

Denoting by V_1 the operator for variance in the first stage of sampling and by $\sum\sum$ the sum over $i, j = 1, \ldots, N$ $(i \neq j)$, we may write, following Raj (1968),

$$V_1(t) = \sum c_i y_i^2 + \sum\sum c_{ij} y_i y_j \tag{1.3}$$

where

$$c_i = E_1(b_{si}^2 I_{si}) - 1, \quad c_{ij} = E_1(b_{si} I_{si} - 1)(b_{sj} I_{sj} - 1).$$

Let

$$v_1(t) = \sum c_{si} I_{si} y_i^2 + \sum\sum c_{sij} I_{sij} y_i y_j \tag{1.4}$$

be an unbiased estimator for $V_1(t)$ so that

$$E_1(c_{si} I_{si}) = c_i, \quad E_1(c_{sij} I_{sij}) = c_{ij}. \tag{1.5}$$

Later we shall use $\sum' \sum'$ to denote summing over $i, j = 1, \ldots, N$, without the restriction $i \neq j$.

In this 'set-up' treated by Raj (1968), let y_i-values be unobservable but sampling be carried out in one or more subsequent stages in such a way that the following conditions hold with E_L and V_L as operators respectively for expectation and variance in the 'later' stages of sampling:

(i) There exist estimators r_i for y_i such that $E_L(r_i) = y_i$;
(ii) $V_L(r_i) = V_i$;
(iii) r_i's are 'independently' distributed;
(iv) there exist estimators v_i for V_i such that $E_L(v_i) = V_i$.

Under these conditions Raj (1968) recommended for Y the multi-stage estimator

$$e = \sum b_{si} I_{si} r_i$$

which is t evaluated at "$\underline{Y}$ equal to $\underline{R} = (r_1, \ldots, r_N)$". Thus, if we write $t = t(s, \underline{Y})$, then $e = t(s, \underline{R})$. Let $E = E_1 E_L$ be the overall operator for expectation

and $V = V_1E_L + E_1V_L$ the over-all variance operator. It follows that

$$E(e) = E_1[E_L(e)] = E_1(t) = Y. \tag{1.6}$$

Thus, e is an unbiased estimator for Y. Also

$$V(e) = V_1E_L(e) + E_1V_L(e) = V_1(t) + E_1\left(\sum b_{si}^2 I_{si} V_i\right), \tag{1.7}$$

$$V_1(e) = \sum c_i r_i^2 + \sum\sum c_{ij} r_i r_j \tag{1.8}$$

which is $V_1(t)$ evaluated at $\underline{Y} = \underline{R}$, and

$$v_1(e) = \sum c_{si} I_{si} r_i^2 + \sum\sum c_{sij} I_{sij} r_i r_j \tag{1.9}$$

which is $v_1(t)$ evaluated at $\underline{Y} = \underline{R}$.

Then using (1.2), (1.3), and (1.7), for

$$v(e) = v_1(e) + \sum b_{si} I_{si} v_i \tag{1.10}$$

we have $Ev(e) = V(e)$ (1.11)

as observed by Raj (1968). This observation led Raj to recommend $v(e)$ as an unbiased estimator for $V(e)$.

We may remark that if we write $\underline{V} = (v_1, \ldots, v_N)$, then $\sum b_{si} I_{si} v_i$ in (1.10) may be expressed as $t(s, \underline{V})$. Thus, Raj's (1968) multi-stage variance estimation rule for e is

$$v(e) = v_1(t)|_{\underline{Y}=\underline{R}} + t|_{\underline{Y}=\underline{V}} = v_1(t)|_{\underline{Y}=\underline{R}} + t(s, \underline{V}). \tag{1.12}$$

It may be remarked that Durbin (1953) earlier gave a version of this rule with t as the Horvitz and Thompson's (1952) estimator, in particular.

Retaining the above set-up but with the modifications that (ii) and (iv) above are respectively replaced by (ii)′ and (iv)′ where (ii)′ $V_L(r_i) = V_{si}$ for i in s; (iv)′ there exist estimators v_{si} for V_{si} such that $E_L(v_{si}) = V_{si}$ when $i \in s$.

Rao (1975) recommended for $V(e)$ the estimator

$$v^*(e) = v_1(e) + \sum (b_{si}^2 - c_{si}) I_{si} v_{si} \tag{1.13}$$

for which he proved the unbiasedness condition

$$Ev^*(e) = V(e). \tag{1.14}$$

For the results (1.10)–(1.12) of Raj (1968) and (1.13)–(1.14) of Rao (1975) the relations (1.1)–(1.5) are all essential. If (1.1) is replaced by

$$t' = t'(s, \underline{Y}) = a_s + \sum b_{si} I_{si} y_i \tag{1.15}$$

such that $a_s \neq 0$ but $E_1(a_s) = 0$ and (1.2) is retained, that is a "non-

homogeneous'' linear unbiased estimator is tried, then (1.10), (1.12)–(1.14) need not follow. If (1.1) is retained but (1.2) is relaxed, then the probelm of variance estimation reduces to one of estimating the ''Mean Square error'' (MSE) of e, namely,

$$M(e) = E_1 E_L (e - Y)^2.$$

Then,

$$M(e) = E_1 E_L ((e - E_L e) + (E_L e - Y))^2$$

$$= E_1 V_L(e) + E_1 (t - Y)^2 = E_1 V_L(e) + M_1(t),$$

where $M_1(t) = E_1(t - Y)^2$, the MSE of t in the first stage of sampling. This problem we intend next to address.

Again, if (1.1)–(1.2) are retained and following Rao (1979) the first stage sampling variance of t is expressed as

$$V_1(t) = \frac{1}{2} \sum{}' \sum{}' d_{ij} w_i w_j \left(\frac{y_i}{w_i} - \frac{y_j}{w_j} \right)^2, \tag{1.16}$$

where

$$w_i \neq 0 \quad \text{and} \quad d_{ij} = -E_1 (b_{si} I_{si} - 1)(b_{sj} I_{sj} - 1),$$

then following Rao (1979) again an unbiased estimator of this $V_1(t)$ may be taken as

$$v_2(t) = \frac{1}{2} \sum{}' \sum{}' d_{sij} I_{sij} w_i w_j \left(\frac{y_i}{w_i} - \frac{y_j}{w_j} \right)^2, \tag{1.17}$$

such that $E_1(d_{sij} I_{sij}) = d_{ij}$, then also Raj's (1968) and Rao's (1975) methods of estimating $V(e)$ shown earlier are not immediately applicable to derive estimators for

$$V(e) = V_1 E_L(e) + E_1 V_L(e)$$

with $V_1(t)$ as in (1.16), t as in (1.1)–(1.2) unless one tediously re-expresses $v_2(t)$ in (1.17) as a quadratic form in y_i's.

Raj (1956) gave a well-known estimator for Y based on the method of sampling with probabilities proportional to sizes (PPS) without replacement (WOR) as briefly described below. Suppose there are numbers p_i $(0 < p_i < 1, \sum p_i = 1)$ called 'normed size-measures'. Then in PPSWOR sampling distinct units from $U = (1, \ldots, i, \ldots N)$ in $n(\geq 2)$ successive draws namely $i_1, \ldots, i_n$ are respectively chosen with probabilities

$$p_{i_1}, \frac{p_{i_2}}{1 - p_{i_1}}, \ldots, \frac{p_{i_n}}{1 - p_{i_1} - \cdots - p_{i_{n-1}}},$$

$$i_1, \ldots, i_n = 1, \ldots, N \quad (i_1 \neq \cdots \neq i_n, 2 \leq n < N).$$

Then, Raj's (1956) unbiased estimator for Y is

$$t_D = \frac{1}{n}\sum_{j=1}^{n} t_j, \text{ where}$$

$$t_1 = \frac{y_{i_1}}{p_{i_1}}, \quad t_j = y_{i_1} + \cdots + y_{i_{j-1}} + \frac{y_{i_j}}{p_{i_j}}(1 - p_{i_1} - \cdots - p_{i_{j-1}}), \quad j = 2, \ldots, n.$$

For this t_D a simple unbiased variance estimator given by Raj (1956) is

$$v_D = \frac{1}{n(n-1)}\sum_{j=1}^{n}(t_j - t_D)^2.$$

Throwing t_D and v_D into the forms (1.1), (1.4) respectively would be a tedious exercise needed to apply Raj's (1968) and Rao's (1975) variance estimation formulae if this strategy of Raj (1956) is to be extended to cover the multi-stage sampling situation.

Moreover, it is worthwhile to mention that in a given survey for certain variables the fsu-values may be "ascertainable" but not for some others. For example, villages may be fsu's and one may know the number of households in them classified by the occupations of their principal earners, the numbers of schools, health care centres, business establishments etc. they respectively have but one may not know the age and sex-wise distribution in the households of the villages, the extent of indebtedness of the household members, the household expenses on their necessities etc. In that case further sampling of the 'households' which are the ssu's may be needed to gather village level information. In such cases it is useful, for the sake of easy computerised processing to use a standard uni-stage variance estimator like $v_2(t)$ or v_D above to cover the 'former set' and consider its easy modification like (1.12) or (1.13) applicable to cover the 'latter set' of variables. But the approaches of Raj (1968) and Rao (1975) do not readily make it evident that it may really be always possible to do so.

Bearing these in mind we develop and present some results in the next section.

2 Developing 'Variance and MSE-estimators' in multi-stage sampling

Retaining the conditions (i)–(iv) in Raj's (1968) set-up it is possible to claim that "E_1 commutes with E_L" i.e. $E = E_1E_L = E_LE_1$. But with Rao's (1975) approach when (ii), (iv) are replaced by (ii)′, (iv)′ this cannot be the case as we shall see.

We feel it is worthwhile to verify this commutativity property with one illustration which we shall utilize in the sequel.

Let us consider a case of sampling in two stages for which a sample of n fsu's is drawn from the population of N fsu's employing the scheme due to Rao, Hartley and Cochran (RHC, 1962) using known positive normed size-measures p_i $(0 < p_i < 1, i = 1, \ldots, N; \sum p_i = 1)$. The ith fsu is supposed to consist of M_i second stage units (ssu) bearing known normed positive size-

measures g_{ij} $(j = 1, \ldots, M_i; i = 1, \ldots, N)$. From each selected fsu, say, i, a sample of m_i ssu's is then selected applying the RHC scheme again using these g_{ij}'s – the selection is done independently across the selected fsu's.

In applying the RHC scheme in the first stage n non-overlapping 'groups' are formed at random out of the N fsu's, the ith group containing N_i fsu's $(i = 1, \ldots, n)$; N_i's are chosen as integers closest to $\frac{N}{n}$ subject to $\sum_n N_i = N$; here $\sum_n$ denotes summing over the n groups. From the ith group one fsu is selected out of the N_i fsu's with a probability proportional to its p-value – this is repeated independently over the groups.

Writing $Q_i = p_{i1} + \cdots + p_{iN_i}$ and denoting for simplicity by (p_i, y_i), the p – and y-values for the fsu selected from the ith group the unbiased estimator for Y given by RHC for the single-stage sampling is

$$t_R = \sum_n \frac{Q_i}{p_i} y_i. \tag{2.1}$$

Writing $A = \dfrac{\sum_n N_i^2 - N}{N(N-1)}$, the variance of t_R is

$$V_1(t_R) = A\left(\sum \frac{y_i^2}{p_i} - Y^2\right) = \frac{1}{2} A \sum\sum p_i p_j \left(\frac{y_i}{p_i} - \frac{y_j}{p_j}\right)^2. \tag{2.2}$$

When y_i is not ascertainable, from each selected fsu, the ssu's are independetly selected by the RHC scheme again. Using somewhat obvious notations we may write the RHC estimator r_i for y_i as

$$r_i = \sum_{m_i} \frac{H_{ij}}{g_{ij}} y_{ij} \tag{2.3}$$

Here $\sum_{m_i}$ is the sum over the m_i groups into which the M_i ssu's in the ith fsu are to be split up to choose from them a sample of m_i ssu's by the RHC scheme; (g_{ij}, y_{ij}) – the known normed size-measure and the y-value for the single ssu selected from the ijth group $(j = 1, \ldots, m_i)$ corresponding to the ith fsu, H_{ij} is the sum of the normed size-measure – values over the N_{ij} ssu's taken in the ijth group, each N_{ij} taken close to $\left[\frac{M_i}{m_i}\right]$ subject to $\sum_{m_i} N_{ij} = M_i$.

By v_{ij} we shall denote the V_i-value corresponding to the jth $(j = 1, \ldots, N_i)$ fsu falling in the ith group in choosing the sample of n fsu's $(i = 1, \ldots, n)$ by the RHC scheme. Further, we shall denote by E_G the operator of expectation for a given grouping and by E_S that over formation of these n groups.

With this background we have

$$e_R = \sum_n \frac{Q_i}{p_i} r_i = t_R|_{\underline{Y}=\underline{R}}$$

for which it follows that

$$E_L(e_R) = t_R, \quad E_1(e_R) = \sum r_i = R, \text{ say,}$$

$E(e_R) = E_1E_L(e_R) = E_1(t_R) = Y$ and also,

$E(e_R) = E_LE_1(e_R) = E_L(R) = Y.$

Then,

$$V(e_R) = E_1E_L(e_R - Y)^2 = E_1V_L(e_R) + V_1E_L(e_R)$$

$$= E_1\left[\sum\nolimits_n \left(\frac{Q_i}{p_i}\right)^2 V_i\right] + V_1(t_R)$$

$$= \sum\nolimits_n E_S\Bigg[(p_{i_1} + \cdots + p_{i_{N_i}})^2 \sum_{j=1}^{N_i}\left(\frac{v_{i_j}}{p_{i_j}}\right)\frac{1}{(p_{i_1} + \cdots + p_{i_{N_i}})}\Bigg] + A\left(\sum\frac{y_i^2}{p_i} - Y^2\right)$$

$$= \sum\nolimits_n\left[E_S\left(\sum_1^{N_i} p_{i_j}\right)\left(\sum_1^{N_i}\frac{v_{i_j}}{p_{i_j}}\right)\right] + A\left(\sum\frac{y_i^2}{p_i} - Y^2\right)$$

$$= \sum\nolimits_n\left[\left(\frac{N_i}{N}\sum V_i\right) + \frac{N_i}{N}\frac{N_i - 1}{N - 1}\sum\frac{V_i}{p_i}(1 - p_i)\right] + A\left(\sum\frac{y_i^2}{p_i} - Y^2\right)$$

$$= \sum V_i + A\left(\sum\frac{V_i}{p_i} - \sum V_i\right) + A\left(\sum\frac{y_i^2}{p_i} - Y^2\right)$$

$$= A\sum\frac{V_i}{p_i} + (1 - A)\sum V_i + A\left(\sum\frac{y_i^2}{p_i} - Y^2\right). \tag{2.4}$$

On the other hand reversing the order of the operators of expectation we get

$$V(e_R) = E_LE_1(e_R - Y)^2 = E_LV_1(e_R) + V_LE_1(e_R)$$

$$= E_L\left[A\left(\sum\frac{r_i^2}{p_i} - R^2\right)\right] + V_L(R)$$

$$= A\left(\sum\frac{y_i^2}{p_i} - Y^2\right) + A\left(\sum\frac{V_i}{p_i} - \sum V_i\right) + \sum V_i$$

$$= A\sum\frac{V_i}{p_i} + (1 - A)\sum V_i + A\left(\sum\frac{y_i^2}{p_i} - Y^2\right) \tag{2.5}$$

From (2.5) and (2.4) our claim that "$E_1E_L = E_LE_1$" is verified in this case. In some other examples also we checked this to be true. So, from now on we presume that when (i)–(iv) hold good we have "$E_1E_L = E_LE_1$".

We may however note that if (ii)′ and (iv)′ hold, that is if we intend to cover Rao's (1975) approach, then we cannot use this 'commutativity'. This is because V_{si}, v_{si} depend on s and as such without operating first by E_L on terms involving v_{si} the operator E_1 cannot be applied as one may check with any of the examples treated by Rao (1975).

So, in what follows we shall throughout assume that (i)–(iv) hold and E_1 commutes with E_L. Let us present below a few results of interest in the present context.

Theorem 1. *Let (i)–(iv) hold and E_1 commute with E_L;*

$$t = t(s, \underline{Y}) \quad \textit{satisfy } E_1(t) = Y;$$

$$e = t(s, \underline{R}) \quad \textit{satisfy } E_L(e) = t.$$

Then, (a) $E_1(e) = R$, (b) $E(e) = Y$, (c) $V(e) = E_L V_1(e) + V_L E_1(e) = E_L V_1(e) + \sum V_i$; (d) If there exists any $v_1(t) = v_1(s, \underline{Y})$ satisfying $E_1 v_1(t) = V_1(t)$, then writing $v_1(e)$ for $v_1(t)$ with $\underline{Y}$ in the latter equal to $\underline{R}$ and v for t with $\underline{Y}$ in the latter replaced by $\underline{V}$, it follows that

$$v(e) = v_1(e) + v \tag{2.6}$$

satisfies $Ev(e) = V(e)$.

Proof: Easy and hence omitted.

Remark I. (2.6) is a generalization of (1.12). For example, t in Theorem 1 may be chosen as t' of (1.15), $V_1(t)$ may be as in (1.16), $v_1(e)$ may be taken in the form v_D in section 1 and in each such case the simple formula (2.6) applies.

To establish our next result let $t = \sum b_{si} I_{si} y_i$ for which $E_1(t)$ may not equal Y i.e. '(1.2) is relaxed', but let there exist $w_i (\neq 0)$ such that

$$t \text{ equals } Y \quad \text{if } y_i \propto w_i.$$

Rao (1979) has illustrated many such situations and from this source we know that we may write

$$M_1(t) = E_1(t - Y)^2 = \frac{1}{2} \sum{}' \sum{}' d_{ij} w_i w_j \left(\frac{y_i}{w_i} - \frac{y_j}{w_j} \right)^2$$

with d_{ij} as in (1.16). Then, for $e = \sum b_{si} I_{si} r_i$, we have

$$M(e) = E(e - Y)^2 = E_1 E_L [(e - E_L(e)) + (E_L(e) - Y)]^2$$

$$= E_1 V_L(e) + M_1(t)$$

and we have

Theorem 2. *With d_{sij} as in (1.17) and*

$$m_1(e) = \frac{1}{2} \sum{}' \sum{}' d_{sij} I_{sij} w_i w_j \left(\frac{r_i}{w_i} - \frac{r_j}{w_j} \right)^2, \tag{2.7}$$

an unbiased estimator for $M(e)$ *is*

$$m(e) = m_1(e) - \frac{1}{2}\sum{}' \sum{}' d_{sij} I_{sij} w_i w_j \left(\frac{v_i}{w_i^2} + \frac{v_j}{w_j^2}\right) + \sum b_{si}^2 I_{si} v_i \tag{2.8}$$

Proof: That $Em(e)$ equals $M(e)$ follows immediately.

To work out our next result, letting t be as in (1.15) allowing $a_s = 0$ or $\neq 0$ and relaxing (1.2) suppose we commit negligible errors if we ignore the discrepancies $\Delta = E_1(t) - Y$ and $d = E_1(e) - R$, for t and $e = t|_{\underline{Y}=\underline{R}}$ assuming the sample-size to be large so that for Y, R respectively, $t = t(s, \underline{Y})$, $e = t(s, \underline{R})$ may be regarded as 'asymptotically design unbiased' (ADU) and 'asymptotically design consistent' (ADC) estimators, in the sense of Brewer's (1979) asymptotic approach. Then we have the following proposition.

Proposition:

$$M(e) = E_L E_1[(e - E_1(e)) + E_1(e) - Y)]^2$$

'approximately equals'

$$E_L E_1(e - R)^2 + E_L(R - Y)^2 = E_L E_1(e - R)^2 + \sum V_i.$$

Then, if there exists a function $m_2(t)$ such that $E_1 m_2(t)$ 'approximately equals' $M_1(t)$, and $m_2(e) = m_2(t)|_{\underline{Y}=\underline{R}}$, then an "approximately unbiased" estimator for $M(e)$ is

$$v(e) = m_2(e) + a_s + \sum b_{si} I_{si} v_i = m_2(e) + t|_{\underline{Y}=\underline{Y}} \tag{2.9}$$

with a_s equal to or not equal to zero. For an illustration let x be a variable well-correlated with y having values x_i and a total X. Let $\pi_i = E_1(I_{si}) > 0$, $\pi_{ij} = E_1(I_{sij}) > 0$,

$$\Delta_{ij} = \pi_i \pi_j - \pi_{ij}, \quad R_i (> 0)$$

be freely assignable constants like

$$\frac{1}{x_i}, \frac{1}{x_i^2}, \frac{1}{\pi_i x_i}, \frac{1 - \pi_i}{\pi_i x_i}$$

etc. Then, letting

$$b_R = \frac{\sum y_i x_i R_i I_{si}}{\sum x_i^2 R_i I_{si}}, \quad e_i = y_i - b_R x_i,$$

$$B_R = \frac{\sum y_i x_i R_i \pi_i}{\sum x_i^2 R_i \pi_i}, \quad E_i = y_i - B_R x_i,$$

the well-known ADU and ADC estimator for Y based on a single-stage sam-

ple is Cassel, Särndal and Wretman's (CSW, 1976) generalized regression (GREG) estimator of the form t' as in (1.15) with $a_s = 0$ and is given by

$$t_G = \sum y_i \frac{I_{si}}{\pi_i} + \left(X - \sum x_i \frac{I_{si}}{\pi_i}\right) b_R$$

$$= \sum y_i g_{si} \frac{I_{si}}{\pi_i}, \text{ writing}$$

$$g_{si} = 1 + \left(X - \sum x_i \frac{I_{si}}{\pi_i}\right) \frac{x_i R_i \pi_i}{\sum x_i^2 R_i \pi_i}$$

From Särndal (1982) we know that its approximate MSE is given by

$$M(t_G) = \frac{1}{2} \sum \sum \Delta_{ij} \left(\frac{E_i}{\pi_i} - \frac{E_j}{\pi_j}\right)^2.$$

Särndal (1982) has also given 2 estimators for $M(t_G)$ as

$$m_k(t_G) = \frac{1}{2} \sum \sum \Delta_{ij} \frac{I_{sij}}{\pi_{ij}} \left(a_{ki} \frac{e_i}{\pi_i} - a_{kj} \frac{e_j}{\pi_j}\right)^2, \quad k = 1, 2$$

$$a_{1i} = 1, \quad a_{2i} = g_{si}.$$

Chaudhuri and Maiti (1994) considered the following versions of t_G, $M(t_G)$, $m_k(t_G)$ when the sample is chosen employing the RHC scheme. They respectively take the forms

$$t_{GR} = \sum_n \frac{Q_i}{p_i} y_i + \left(X - \sum\nolimits_n \frac{Q_i}{p_i} x_i\right) b_R = \sum\nolimits_n \frac{Q_i}{p_i} y_i h_{si}$$

$$\text{where } h_{si} = 1 + \left(X - \sum\nolimits_n \frac{Q_i}{p_i} x_i\right) \frac{x_i R_i \frac{p_i}{Q_i}}{\sum_n x_i^2 R_i \frac{p_i}{Q_i}},$$

$$M(t_{GR}) = \frac{A}{2} \sum \sum p_i p_j \left(\frac{F_i}{p_i} - \frac{F_j}{p_j}\right)^2,$$

$$F_i = y_i - C_R x_i, \quad C_R = \frac{\sum y_i x_i R_i \frac{p_i}{Q_i}}{\sum x_i^2 R_i \frac{p_i}{Q_i}},$$

$$m_k(t_{GR}) = B \sum\nolimits_n \sum\nolimits_n Q_i Q_j \left(b_{ki} \frac{e_i}{p_i} - b_{kj} \frac{e_j}{p_j}\right)^2, \quad k = 1, 2, b_{1i} = 1, b_{2i} = h_{si},$$

– a formula due to RHC (1962), $B = \dfrac{\sum_n N_i^2 - N}{N^2 - \sum_n N_i^2}$, and

$$m_k'(t_{GR}) = C \sum\nolimits_n \sum\nolimits_n \frac{Q_i Q_j}{N_i N_j} \left(b_{ki} \frac{e_i}{p_i} - b_{kj} \frac{e_j}{p_j} \right)^2, \quad k = 1, 2,$$

– a formula due to Ohlsson (1989),

$$C = \frac{\sum_n N_i^2 - N}{N(N-1)}.$$

Here $\sum_n \sum_n$ denotes summing over the distinct pairs of n groups with no overlap. Corresponding to $t_G, t_{GR}, m_k(t_G), m_k'(t_{GR})$, the multi-stage estimators, applying (2.8) are respectively

$$e_G = \sum r_i g_{si} \frac{I_{si}}{\pi_i}, \quad e_{GR} = \sum\nolimits_n \frac{Q_i}{p_i} r_i h_{si} \tag{2.10}$$

$$v_k(e_G) = m_k(e_G) + \sum v_i g_{si} \frac{I_{si}}{\pi_i}, \quad k = 1, 2 \tag{2.11}$$

writing $m_k(e_G)$ for $m_k(t_G)$ with $\underline{Y}$ replaced by $\underline{R}$,

$$v_k(e_{GR}) = m_k(e_{GR}) + \sum\nolimits_n \frac{Q_i}{p_i} v_i h_{si}, \quad k = 1, 2 \tag{2.12}$$

$$v_k'(e_{GR}) = m_k'(e_{GR}) + \sum\nolimits_n \frac{Q_i}{p_i} v_i h_{si}, \quad k = 1, 2 \tag{2.13}$$

writing $m_k(e_{GR})$ for $m_k(t_{GR})$ with $\underline{Y}$ replaced by $\underline{R}$, and $m_k'(e_{GR})$ for $m_k'(t_{GR})$ with $\underline{Y}$ replaced by $\underline{R}$.

In the next section we report certain results that we developed along the above lines as we needed them to apply in implementing two surveys in Indian Statistical Institute, Calcutta. There we actually adopted a three-stage sampling scheme in which the RHC scheme was adopted in the first two stages and a simple random sample (SRS) was taken without replacement (WOR) in the third stage.

3 Variance Estimation in a three stage sampling scheme

Suppose a sample is to be chosen in three stages. In the first two stages the selection is by adopting the RHC scheme as described in Section 2; let the ijth ssu in the ith fsu consist of T_{ij} third stage units (tsu) and a sample of t_{ij} tsu's be selected out of these T_{ij} tsu's by the SRSWOR method.

Using N_{ij}'s introduced in Section 2 and also g_{ij}, H_{ij} etc, let

$$A_i = \frac{\sum_{m_i} N_{ij}^2 - M_i}{M_i(M_i - 1)}, \quad B_i = \frac{\sum_{m_i} N_{ij}^2 - M_i}{M_i^2 - \sum_{m_i} N_{ij}^2};$$

further let

$$x_i = \sum\nolimits_{m_i} \frac{H_{ij}}{g_{ij}} y_{ij}, \quad e = \sum\nolimits_n \frac{Q_i}{p_i} x_i,$$

$$z_i = \sum\nolimits_{m_i} \frac{H_{ij}}{g_{ij}} w_{ij}, \quad X = \sum x_i, \quad Z = \sum z_i,$$

writing y_{ijk} as the y-value for the kth tsu in the ijth ssu and $\sum_k$ as sum over the t_{ij} tsu's in the sample, let

$$w_{ij} = \frac{T_{ij}}{t_{ij}} \sum\nolimits_k y_{ijk}, \quad w_i = \sum_{j=1}^{M_i} w_{ij}.$$

Then, writing $E_i, V_i, i = 1, 2, 3$, as operators for expectation and variance respectively for the ith stage of sampling we have,

$$E_2(x_i) = y_i, \quad E_1E_2(e) = Y, \quad E_3(w_{ij}) = y_{ij}, \quad E_3(z_i) = x_i.$$

Also, we shall write

$$E_{123} = E_1(E_2E_3) = E_1E_2(E_3) = E_1E_2E_3 = E,$$

the over-all operator of expectation over the three stages of sampling and V for the overall variance operator.

Next, let us write

$$u = \sum\nolimits_n \frac{Q_i}{p_i} z_i, \quad v_3(w_{ij}) = T_{ij}^2 \left(\frac{1}{t_{ij}} - \frac{1}{T_{ij}}\right) \frac{1}{(t_{ij} - 1)} \sum\nolimits_k \left(y_{ijk} - \frac{w_{ij}}{T_{ij}}\right)^2,$$

$$v_2(x_i) = B_i \left(\sum\nolimits_{m_i} \frac{H_{ij}}{g_{ij}^2} y_{ij}^2 - x_i^2\right), \quad v_2(z_i) = B_i \left(\sum\nolimits_{m_i} \frac{H_{ij}}{g_{ij}^2} w_{ij}^2 - z_i^2\right).$$

Then, we may observe the following.

$$E_2 v_2(x_i) = V_2(x_i) = A_i \left(\sum_1^{M_i} \frac{y_{ij}^2}{g_{ij}} - y_i^2\right),$$

$$V_2(z_i) = A_i \left(\sum_1^{M_i} \frac{w_{ij}^2}{g_{ij}} - w_i^2\right),$$

$$E_2 v_2(z_i) = V_2(z_i), \quad E_2(z_i) = w_i, \quad V_3 E_2(z_i) = \left(\sum_1^{M_i} V_3(w_{ij})\right)$$

$$E_3 v_3(w_{ij}) = V_3(w_{ij}), \quad E_2 E_3 \left[\sum\nolimits_{m_i} \frac{H_{ij}}{g_{ij}} v_3(w_{ij})\right] = \sum_1^{M_i} V_3(w_{ij}) = V_3 E_2(z_i)$$

$$E_1E_2E_3\left(\sum_n \frac{Q_i}{p_i}\left[v_2(z_i)+\sum_{m_i}\frac{y_{ij}}{g_{ij}}v_3(w_{ij})\right]\right)=V_{23}(Z)=E_{23}(Z-Y)^2.$$

$$E(u)=Y,$$

$$V(u)=E_{123}(u-Y)^2=E_{23}[E_1(u-Y)^2]=E_{23}[E_1(u-Z)^2+(Z-Y)^2].$$

Now, $E_1(u-Z)^2=A\left(\sum\frac{z_i^2}{p_i}-Z^2\right)$ and it follows that for

$$d_1=B\sum_n\sum_n Q_iQ_j\left(\frac{z_i}{p_i}-\frac{z_j}{p_j}\right)^2=B\left(\sum_n\frac{Q_i}{p_i^2}z_i^2-u^2\right)$$

and

$$d_2=C\sum_n\sum_n\frac{Q_iQ_j}{N_iN_j}\left(\frac{z_i}{p_i}-\frac{z_j}{p_j}\right)^2$$

$$E_1(d_1)=E_1(d_2)=E_1(u-Z)^2.$$

Then, we have

Theorem 3. *Given*

$$v_1(u)=B\left(\sum_n\frac{Q_i}{p_i^2}-u^2\right)+\sum_n\frac{Q_i}{p_i}\left(v_2(z_i)+\sum_{m_i}\frac{H_{ij}}{g_{ij}}v_3(w_{ij})\right)$$

and

$$v_2(u)=C\sum_n\sum_n\frac{Q_iQ_j}{N_iN_j}\left(\frac{z_i}{p_i}-\frac{z_j}{p_j}\right)^2+\sum_n\frac{Q_i}{p_i}\left(v_2(z_i)+\sum_{m_i}\frac{H_{ij}}{g_{ij}}v_3(w_{ij})\right),$$

it follows that $Ev_1(u)=Ev_2(u)=V(u)$.

Proof. Easy and hence omitted.

In hitting upon Theorem 3 we applied the approach of our Theorem 1. An alternative estimator for $V(u)$ is available following the approach of Raj (1968) as follows:

Let us express the RHC estimator of the variance of the RHC estimator for Y namely $t_R=\sum_n\frac{Q_i}{p_i}y_i$ in the form

$$v_1(t_R)=\sum d_{si}I_{si}y_i^2+\sum\sum d_{sij}I_{sij}y_iy_j \text{ so that}$$

$$E_1(d_{si}I_{si})=E_1\left(\sum_n\left(\frac{Q_i}{p_i}\right)^2\right)-1.$$

$$\text{Let } v_2(t_R) = \sum d_{si} I_{si} x_i^2 + \sum \sum d_{sij} I_{sij} x_i x_j$$

$$v_3(t_R) = \sum d_{si} I_{si} z_i^2 + \sum \sum d_{sij} I_{sij} z_i z_j$$

$$v_2'(z_i) = v_2(z_i) - B_i \left(\sum\nolimits_{m_i} \frac{H_{ij}(1 - H_{ij})}{g_{ij}^2} v_3(w_{ij}) \right).$$

Then we have

Theorem 4. *Given*

$$v(u) = v_3(t_R) - \sum d_{si} I_{si} v_3(z_i) + \sum\nolimits_n \frac{Q_i}{p_i} v_2'(z_i) + \sum\nolimits_n \left(\frac{Q_i}{p_i} \right)^2 v_3(z_i)$$

it follows that $Ev(u) = V(u)$.

Proof:

$$E_3 v_3(t_R) = v_2(t_R) + \sum d_{si} I_{si} V_3(z_i)$$

$$E_3 \left(\sum\nolimits_n \frac{Q_i}{p_i} v_2'(z_i) \right) = \sum\nolimits_n \frac{Q_i}{p_i} \left[B_i \left(\sum\nolimits_{m_i} \frac{H_{ij}}{g_{ij}^2} y_{ij}^2 - x_i^2 \right) \right]$$

$$\text{So } E_3 v(u) = v_2(t_R) + \sum\nolimits_n \frac{Q_i}{p_i} v_2(x_i) + \sum\nolimits_n \left(\frac{Q_i}{p_i} \right)^2 V_3(z_i)$$

$$E_2 E_3 v(u) = v_1(t_R) + \sum d_{si} I_{si} V_2(x_i) + \sum\nolimits_n \frac{Q_i}{p_i} V_2(x_i) + \sum\nolimits_n \left(\frac{Q_i}{p_i} \right)^2 E_2 V_3(z_i)$$

$$\text{So, } Ev(u) = E_1[E_2 E_3 v(u)] = E_1 v_1(t_R) + E_1 \left[\sum d_{si} I_{si} V_2(x_i) + \sum\nolimits_n \frac{Q_i}{p_i} V_2(x_i) \right] + E_1 \sum\nolimits_n \left(\frac{Q_i}{p_i} \right)^2 E_2 V_3(z_i)$$

$$= V_1(t_R) + E_1 \left[V_2 \left(\sum\nolimits_n \frac{Q_i}{p_i} x_i \right) \right] + E_1 \left[\sum\nolimits_n \left(\frac{Q_i}{p_i} \right)^2 E_2 V_3(z_i) \right]$$

This is because

$$E_1\left[\sum d_{si}I_{si}V_2(x_i)+\sum_n\left(\frac{Q_i}{p_i}\right)V_2(x_i)\right]$$

$$=E_1\left[\left(\sum_n\left(\frac{Q_i}{p_i}\right)^2V_2(x_i)-\sum_n\frac{Q_i}{p_i}V_2(x_i)\right)+\sum_n\left(\frac{Q_i}{p_i}\right)V_2(x_i)\right]$$

$$=E_1\left[\sum_n\left(\frac{Q_i}{p_i}\right)^2V_2(x_i)\right]=E_1\left[V_2\left(\sum_n\frac{Q_i}{p_i}x_i\right)\right].$$

So, finally,

$$V(u)=E_1E_2E_3(u-Y)^2=E_1[E_2E_3(u-E_{23}u)^2+(E_{23}u-Y)^2]$$

$$=E_1(t_R-Y)^2+E_1[V_{23}(u)]$$

$$=V_1(t_R)+E_1\left[V_2\left(\sum_n\frac{Q_i}{p_i}x_i\right)+E_2\left(\sum_n\left(\frac{Q_i}{p_i}\right)^2V_3(z_i)\right)\right]=Ev(u).$$

Remark II. There seems to be no guarantee that $v_1(u)$, $v_2(u)$, $v(u)$ must be non-negative for every sample of observations.

In the Section 4 below we shall numerically examine the relative efficacies of two alternative estimators $v_1(u)$ and $v(u)$ for the variance $V(u)$ through a simulation exercise.

4 Relative efficacies of two variance estimators in a three stage sampling

In two different surveys carried out at the Indian Statistical Institute, Calcutta the above-mentioned three stage sampling was implemented. In one of them the variance estimator $v_1(u)$ and in the other $v(u)$ was applied as is reported for the two surveys. To compare the relative efficacies of $v_1(u)$ and $v(u)$ we consider it useful to apply certain performance criteria which may be evaluated only if certain details are used for numerical calculations. So, we consider it appropriate to undertake a simulation study.

Let us consider certain fictitious data relating to a district composed of 10 administrative blocks. The blocks are taken as the fsu's and they are supposed to be composed of a number of villages which are the ssu's. The households (hh) in the villages are the tsu's. Some details are given in Table 1.

The number of villages in a 'block' is taken as its size-measure; using this applying RHC scheme 4 blocks are sampled. Using the number of people in a 'village' as its size-measure, from each selected block, 22 percent of the villages, rounded up to the nearest integer, is sampled applying the RHC scheme again. A 4 percent, rounded up to the higher integer, sample of households is taken by SRSWOR method from each village.

The purpose is to estimate the total population in the district. Note that though the size-measures are chosen in the manners described, the total population $Y = 271986$ will not be estimated free of error because the households are chosen by SRSWOR method. To compare $v_1(u)$ with $v(u)$ we repeat the

Table 1. Composition of 10 blocks in a district

Serial number of block	Number of villages in blocks	Total population in blocks	Serial number of blocks	Number of villages in blocks	Total population in blocks
1	39	23239	6	59	33624
2	30	22253	7	56	31373
3	55	32756	8	41	21435
4	51	29074	9	33	19219
5	60	35079	10	42	23934
Total					271986

Table 2. A summary of efficacy of $v_1(u)$ vs $v(u)$

Serial number of 'set' of replicated samples	Number of replicated samples in the 'set'	ACP using $v_1(u)$	$v(u)$	ACV using $v_1(u)$	$v(u)$	Percent of replicates giving $v_1(u)$ less than $v(u)$	$RE = 100 \times \frac{v_1(u)}{v(u)}$ for the last 10 replicates in the sets
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	300	94.34	92.67	5.55	5.53	54.67	107.67, 98.49 104.06, 100.43 111.82, 98.57 86.50, 111.00 90.53, 89.89
2	300	95.33	95.00	5.57	5.54	58.00	81.10, 92.10 93.48, 103.89 106.98, 76.06 88.88, 115.21 84.48, 94.84
3	400	97.00	96.75	5.59	5.58	54.50	102.22, 100.12 96.44, 81.83 02.83, 100.05 75.58, 97.89 123.26, 97.57
Total	1000	95.70	95.00	5.57	5.55	55.60	

drawing of the sample a total of F = 1000 times divided into 3 disjoint sets of 300, 300 and 400 replicates. Writing w for $v_1(u)$ and $v(u)$ in turn we calculate the percentage of replicates for which the intervals $(u - 1.96\sqrt{w}, u + 1.96\sqrt{w})$ cover the value of Y. Each interval has a nominal confidence coefficient of 95 per cent assuming normality. This realized per cent is called the ACP – the actual coverage percent. Also, we calculate the ACV, the average coefficient of variation. This is the average, over the F = 1000 replicates of the value of $\frac{\sqrt{w}}{u} \times 100$. This reflects the length of the confidence interval. Between $v_1(u)$ and $v(u)$ that one is preferable for which the ACP is closer to 95 per cent and the ACV is smaller. The actual simulated findings are shown below in Table 2.

Here we also indicate the values of $RE = 100 \times \frac{v_1(u)}{v(u)}$ for the last 10 replicates out of each of the 3 sets of replicates of samples numbering 300, 300 and 400 mentioned above to illustrate the efficiency of $v(u)$ relative to $v_1(u)$ – the smaller it is the better the one proposed by us relative to the one given by Raj (1968).

Remark III. Each replicate gave us positive values for $v_1(u)$ and $v(u)$.

Conclusion: The two variance estimators tried turn out quite competitive and adequately effective. In an actual survey both should be calculated and a confidence interval may be reported in terms of the one for which its length happens to be shorter. Our method at least provides a serviceable competitor against Raj's (1968).

Acknowledgment: The authors gratefully acknowledge the helpful comments on an earlier draft from a referee which led to a substantial improvement in the presentation.

References

Brewer KRW (1979) A class of robust sampling designs for large-scale surveys. Jour Amer Stat Assoc 74:911–915

Cassel CM, Särndal CE, Wretman JH (1976) Some results in generalized difference estimation and generalized regression estimation for finite populations. Biometrika 63:615–620

Chaudhuri A, Maiti T (1994) On the regression adjustment to Rao, Hartley, Cochran estimator. Jour Stat Res 29(2):71–78

Durbin J (1953) Some results in sampling theory when the units are selected with unequal probabilities. Jour Roy Stat Soc Ser B 15:262–269

Horvitz DG, Thompson DJ (1952) A generalization of sampling without replacement from a finite universe. Jour Amer Stat Assoc 47:663–685

Murthy MN (1957) Ordered and unordered estimators in sampling without replacement Sankhyá 18:379–390

Ohlsson E (1989) Variance estimation in Rao, Hartley, Cochran procedure. Sankhyá B 51:348–367

Raj Des (1956) Some estimators in sampling with varying probabilities without replacement. Jour Amer Stat Assoc 51:269–284

——— (1968) Sampling theory. McGraw Hill NY

Rao JNK (1975) Unbiased variance estimation for multi-stage designs. Sankhyá, C 37:133–139

——— (1979) On deriving mean square errors and their non-negative unbiased estimators in finite population sampling. Jour Ind Stat Assoc 17:125–136

Rao JNK, Hartley HO, Cochran WG (1962) On a simple procedure of unequal probability sampling without replacement. Jour Roy Stat Soc B24:482–491

Särndal CE (1982) Implications of survey design for generalized regression estimation of linear functions. Jour Stat Plan Inf 7:155–170