Arijit Chaudhuri · Amitava Saha

Estimation from two-stage unequal probability sampling with missing units

Abstract A formula is presented for an unbiased estimator for the variance of an unbiased estimator of a survey population total as well as for an unbiased estimator of its variance based on sampling in two-stages following Rao et al. J Roy Stat Soc B 24: 482–491 (1962) scheme in both stages when the originally selected units in both stages cannot be fully covered in the survey but are to be randomly subsampled. The development is helpful to tackle non-responses if assumed to have occurred at random in either or both the stages.

Keywords Randomly missing units · Rao–Hartley–Cochran scheme · Two-stage sampling · Unbiased variance estimation

Mathematics Subject Classification (2000) 62 D05

1 Introduction

Stratified two-stage sampling with unit-wise varying probabilities of selection for the first stage units (fsu) and also for the second stage units (ssu) or equal selection probabilities for the latter is a common Survey Sampling practice.

Rao et al. (1962) scheme, applicable when normed size-measures for the units are priorly available, is both convenient and suitable because it is easy to implement, yields only distinct units in a sample, admits a simple unbiased estimator for a population total as well as a simple uniformly non-negative unbiased estimator for the variance of the estimator of the total. If the Rao–Hartley–Cochran (RHC) scheme is employed in both the first and the second stages of sampling then thanks to Chaudhuri et al. (2000) works, these advantages still continue to exist.

A problem arises however, if, because of resource crunches, unanticipated at the design stage but unavoidably faced when the field work for the survey is about to start, the sample-sizes in one or both the stages need to be cut down allowing the survey to be completed in time consistently with resource constraints.

It may be of interest to illustrate a practical application. The Indian Statistical Institute (ISI), Kolkata, undertook recently to assist the government in implementing 'internal audit' by sampling entries from account books in various offices in 17 districts. Initially, a sample of 79 offices spread widely in 7 sampled districts was chosen adopting the RHC technique to accomplish the task within a stipulated time. After some initial time and resources were exhausted in course of the planning, it was badly needed to cut down the sample-size to 65, consistently with the resource constraints. District-wise simple random sampling without replacement (SRSWOR) from the originally chosen RHC samples was adjudged to be feasible retaining unbiasedness in estimation of totals and of variances of estimates of totals. Of course the sub-sample of 65 out of 79 offices could be chosen with unequal probabilities as well. But to ensure simplicity this was avoided.

An extension of this principle is also feasible if one encounters 'non-respondents' in an initial RHC sample, assuming the incidences of non-responses to occur only at random. In this case, however, a more complicated probability system in the 'non-response behaviour mechanism' could be envisaged leading to complicated analytical developments. Consideration of simplicity guides us to restrict in the present work to 'equal-probability sub-sampling' alone.

In the section 2 we present the details to show the necessary changes if simple random samples (SRS) of suitable sizes be drawn without replacement (WOR) from the initially chosen samples in both the stages. By simulation we illustrate in section 3 how the efficiency level declines with increasing sub-sampling rates.

2 Sample selection and estimation

Let U denote a finite population of N units labelled $i = 1, \ldots, N$ bearing values y_i of a variable of interest y and the known normed size-measures p_i $(0 < p_i < 1, \sum p_i = 1$, writing $\sum$ for sum over i in $U)$. Let again this i, to be called fsu, in its turn be composed of ssu, M_i in number. The ssu ij has y-values y_{ij} and normed size-measures $p_{ij} (0 < p_{ij} < 1, \sum_{j=1}^{M_i} p_{ij} = 1)$, $j = 1, \ldots, M_i; i = 1, \ldots, N$.

Let the problem be to unbiasedly estimate $Y = \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij} = \sum_{i=1}^{N} y_i$ where $y_i = \sum_{j=1}^{M_i} y_{ij}$, on taking (1) a sample of n units from $U = (1, \cdots\cdots\cdots, N)$ by the RHC scheme and (2) independently choosing from each sampled fsu, i, say, again by the RHC scheme a sample of m_i ssu's, $i \in s$, writing s as the sample of the n fsu's chosen.

2.1 Estimation from first stage sampling allowing random sub-sampling

In choosing an RHC sample s of n distinct units the procedure is to first randomly divide U into n non-overlapping groups taking N_i units in the ith group choosing them subject to $\sum_n N_i = N$, writing $\sum_n$ to denote summing over the n groups. Writing Q_i as the sum of the p_i-values for the fsu's falling in the ith group thus formed, from this group one unit is to be chosen with a probability equal to this unit's p_i-value, divided by Q_i; and this is to be independently repeated for each of the n groups formed. Writing, for simplicity, p_i, y_i as the p_i-value and y-value for the unit actually chosen from the ith group, RHC's unbiased estimator for $\sum_{i=1}^N y_i$ is

$$t = \sum\nolimits_n y_i \frac{Q_i}{p_i}$$

Writing

$$B = \frac{\sum_n N_i^2 - N}{N^2 - \sum_n N_i^2},$$

RHC's unbiased estimator for the variance of t is

$$v(t) = B\left[\sum\nolimits_n Q_i \left(\frac{y_i}{p_i}\right)^2 - t^2\right]$$

Supposing inadequacy in resources to cover all the n fsu's in a field survey let an SRSWOR of m fsu's be chosen from s. Then, writing $\sum_m$ for the summation over these m sampled groups out of the original n groups, let

$$e = \frac{n}{m} \sum\nolimits_m y_i \frac{Q_i}{p_i} \tag{2.1.1}$$

be taken as an estimator for $Y = \sum_{i=1}^N y_i$.

Writing E_R, V_R as the operators for expectation, variance with respect to this SRSWOR selection we have

$$E_R(e) = t,\ V_R(e) = n^2\left(\frac{1}{m} - \frac{1}{n}\right)\frac{1}{(n-1)}\left[\sum\nolimits_n \left(y_i\frac{Q_i}{p_i} - \frac{t}{n}\right)^2\right];$$

then, an unbiased estimator of $V_R(e)$ is

$$v_R(e) = n^2\left(\frac{1}{m} - \frac{1}{n}\right)\frac{1}{(m-1)}\left[\sum\nolimits_m Q_i^2\left(\frac{y_i}{p_i}\right)^2 - m\left(\frac{e}{n}\right)^2\right] \tag{2.1.2}$$

for which $E_R v_R(e) = V_R(e)$.

Writing E_1, V_1 as the expectation, variance operators with respect to the above RHC sampling it is known that

$$V_1(t) = A\left[\sum_{i=1}^N \frac{y_i^2}{p_i} - Y^2\right],\ \text{where } A = \frac{\sum_n N_i^2 - N}{N(N-1)}.$$

Again, on writing

$$E = E_1 E_R, \quad V = V_1 E_R + E_1 V_R,$$

the variance of e is $V(e) = V_1 E_R(e) + E_1 V_R(e) = V_1(t) + E_1 E_R v_R(e)$ and an unbiased estimator for $V(e)$ is $v(e)$ given below with the following result in Theorem 1.

Theorem 1 $Ev(e) = V(e)$
where

$$v(e) = (1+B)v_R(e) + B\left[\frac{n}{m}\sum\nolimits_m Q_i\left(\frac{y_i}{p_i}\right)^2 - e^2\right] \tag{2.1.3}$$

Proof Let

$$\hat{v}(t) = B\left[\frac{n}{m}\sum\nolimits_m Q_i\left(\frac{y_i}{p_i}\right)^2 - \left(e^2 - v_R(e)\right)\right]$$

and

$$\hat{t}^2 = e^2 - v_R(e)$$

Then, $E_R(\hat{t}^2) = t^2$ because $E_R(e) = t$,

$$E\hat{v}(t) = E_1[E_R(\hat{v}(t))] = E_1 v(t) = V_1(t).$$

So, for

$$\begin{aligned} &B\left[\frac{n}{m}\sum\nolimits_m Q_i\left(\frac{y_i}{p_i}\right)^2 - \left(e^2 - v_R(e)\right)\right] + v_R(e) \\ &= (1+B)v_R(e) + B\left[\frac{n}{m}\sum\nolimits_m Q_i\left(\frac{y_i}{p_i}\right)^2 - e^2\right] = v(e) \end{aligned}$$

we have

$$\begin{aligned} Ev(e) &= E_1E_Rv(e) = E_1v(t) + E_1E_Rv_R(e) \\ &= V_1(t) + E_1V_R(e) = V(e) \\ \text{i.e. } Ev(e) &= V(e). \end{aligned}$$

□

2.2 Modification in two-stage sampling with random sub-sampling in both stages

At this stage let us suppose that for i in s, $y_i = \sum_{j=1}^{M_i} y_{ij}$, the ith fsu total is not ascertainable and needs to be estimated through sampling of ssu's from the selected fsu's in s. We suppose that every sampled fsu is sub-sampled according to RHC scheme, taking m_i ssu's from the M_i ssu's in the ith fsu. Using parallel notations as $\sum_{m_i}$ for $\sum_n$, Q_{ij} for Q_i, M_{ij} for N_i and E_2, V_2 for E_1, V_1 we may write as follows:

$$r_i = \sum\nolimits_{m_i} y_{ij}\frac{Q_{ij}}{p_{ij}}, \quad E_2(r_i) = y_i \tag{2.2.1}$$

$$V_2(r_i) = A_i\left[\sum_{j=1}^{M_i}\frac{y_{ij}^2}{p_{ij}} - y_i^2\right], \qquad A_i = \frac{\sum_{m_i} M_{ij}^2 - M_i}{M_i(M_i-1)} \tag{2.2.2}$$

$$v_2(r_i) = B_i\left[\sum\nolimits_{m_i} Q_{ij}\frac{y_{ij}^2}{p_{ij}^2} - r_i^2\right], \quad B_i = \frac{\sum_{m_i} M_{ij}^2 - M_i}{M_i^2 - \sum_{m_i} M_{ij}^2}, \tag{2.2.3}$$

then,

$$E_2 v_2(r_i) = V_2(r_i) = E_2(r_i^2) - y_i^2$$

and this leads to $\hat{y}_i^2 = r_i^2 - v_2(r_i)$ satisfying $E_2(\hat{y}_i^2) = y_i^2$.

Let again, it be impossible, with the resources at hand, to complete the survey of all the m_i ssu's, for i in s and it be decided to choose by SRSWOR method only l_i of the ssu's out of m_i for i in s. Then, with the notations $\sum_{l_i}$ paralleling $\sum_{m_i}$, E_r, V_r paralleling E_R, V_R we may write as follows:

$$g_i = \frac{m_i}{l_i} \sum\nolimits_{l_i} y_{ij} \frac{Q_{ij}}{p_{ij}}, \quad E_r(g_i) = r_i \tag{2.2.4}$$

$$V_r(g_i) = m_i^2 \left(\frac{1}{l_i} - \frac{1}{m_i}\right) \frac{1}{(m_i - 1)} \sum\nolimits_{m_i} \left[y_{ij} \frac{Q_{ij}}{p_{ij}} - \frac{\sum_{m_i} y_{ij} \frac{Q_{ij}}{p_{ij}}}{m_i} \right]^2 \tag{2.2.5}$$

$$v_r(g_i) = m_i^2 \left(\frac{1}{l_i} - \frac{1}{m_i}\right) \frac{1}{(l_i - 1)} \sum\nolimits_{l_i} \left[y_{ij} \frac{Q_{ij}}{p_{ij}} - \frac{\sum_{l_i} y_{ij} \frac{Q_{ij}}{p_{ij}}}{l_i} \right]^2; \tag{2.2.6}$$

then,

$$E_r v_r(g_i) = V_r(g_i) = E_r(g_i^2) - r_i^2;$$

hence

$$\hat{r}_i^2 = g_i^2 - v_r(g_i) \text{ satisfies } E_r(\hat{r}_i^2) = r_i^2.$$

Let

$$h = \frac{n}{m} \sum\nolimits_m g_i \frac{Q_i}{p_i} = \frac{n}{m} \sum\nolimits_m \frac{Q_i}{p_i} \left[\frac{m_i}{l_i} \sum\nolimits_{l_i} y_{ij} \frac{Q_{ij}}{p_{ij}} \right]. \tag{2.2.7}$$

Then this h is our proposed unbiased estimator for $Y = \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij}$ because for this we have the following result in Theorem 2.

Theorem 2 $E_1 E_R E_2 E_r(h) = Y$.

Proof

$$E_r(h) = \frac{n}{m} \sum\nolimits_m r_i \frac{Q_i}{p_i}$$
$$E_2 E_r(h) = \frac{n}{m} \sum\nolimits_m y_i \frac{Q_i}{p_i} = e$$
$$E_R E_2 E_r(h) = E_R(e) = t$$

So,

$$E_1 E_R E_2 E_r(h) = E_1(t) = \sum_{i=1}^{N} y_i = \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij} = Y.$$

□

Occasionally in what follows we shall write $E_{1R} = E_1 E_R$, $V_{1R} = V_1 E_R + E_1 V_R$, $E_{2r} = E_2 E_r$, $V_{2r} = V_2 E_r + E_2 V_r$.

Let us further note that

$$V_r(h) = \left(\frac{n}{m}\right)^2 \sum\nolimits_m \left(\frac{Q_i}{p_i}\right)^2 V_r(g_i)$$

and

$$v_r(h) = \left(\frac{n}{m}\right)^2 \sum\nolimits_m \left(\frac{Q_i}{p_i}\right)^2 v_r(g_i) \tag{2.2.8}$$

satisfies $E_r v_r(h) = V_r(h)$.

Now,

$$\begin{aligned} V_{2r}(h) &= V_2 E_r(h) + E_2 V_r(h) \\ &= V_2 \left[\frac{n}{m} \sum\nolimits_m r_i \frac{Q_i}{p_i}\right] + E_2 E_r\,[v_r(h)] \\ &= \left(\frac{n}{m}\right)^2 \sum\nolimits_m \left(\frac{Q_i}{p_i}\right)^2 V_2(r_i) + E_2 E_r[v_r(h)] \\ &= \left(\frac{n}{m}\right)^2 \sum\nolimits_m \left(\frac{Q_i}{p_i}\right)^2 E_2 E_r(\hat{v}_2(r_i)) + E_2 E_r[v_r(h)] \end{aligned} \tag{2.2.9}$$

where $\hat{v}_2(r_i) = B_i \left[\frac{m_i}{l_i} \sum_{l_i} Q_{ij} \frac{y_{ij}^2}{p_{ij}^2} - \hat{r}_i^2\right]$ and from (2.2.9) we get

$$v_{2r}(h) = \left(\frac{n}{m}\right)^2 \sum\nolimits_m \left(\frac{Q_i}{p_i}\right)^2 \hat{v}_2(r_i) + v_r(h). \tag{2.2.10}$$

Again we have

$$\begin{aligned} V_{2r}(h) &= E_2 E_r(h^2) - [E_2 E_r(h)]^2 \\ &= E_2 E_r(h^2) - e^2. \end{aligned} \tag{2.2.11}$$

Thus from (2.2.9) and (2.2.10) it follows that

$$\hat{e}^2 = h^2 - v_{2r}(h) \tag{2.2.12}$$

and

$$E_2 E_r(\hat{e}^2) = e^2.$$

Now

$$\hat{v}_R(e) = n^2 \left(\frac{1}{m} - \frac{1}{n}\right) \frac{1}{(m-1)} \left[\sum\nolimits_m \left(\frac{Q_i}{p_i}\right)^2 \hat{\hat{y}}_i^2 - m\frac{\hat{e}^2}{n^2}\right] \tag{2.2.13}$$

where

$$\hat{\hat{y}}_i^2 = \hat{r}_i^2 - \hat{v}_2(r_i) \tag{2.2.14}$$

and $\hat{v}_2(r_i)$ is as defined earlier.

Then we have

$$\hat{\hat{y}}_i^2 = (1 + B_i)\hat{r}_i^2 - B_i \left[\frac{m_i}{l_i} \sum\nolimits_{k} Q_{ij} \frac{y_{ij}^2}{p_{ij}} \right] \tag{2.2.15}$$

and $E_r(\hat{\hat{y}}_i^2) = \hat{y}_i^2$, $E_2 E_r(\hat{\hat{y}}_i^2) = y_i^2$.

This implies that

$$E_2 E_r \, \hat{v}_R(e) = v_R(e) = n^2 \left(\frac{1}{m} - \frac{1}{n} \right) \frac{1}{(m-1)} \left[\sum\nolimits_m \left(\frac{Q_i}{p_i} \right)^2 y_i^2 - m \left(\frac{e}{n} \right)^2 \right].$$

Now, in $v(e)$ of (2.1.3) let us replace y_i^2, e^2 and $v_R(e)$ by their respective unbiased estimators just derived above, namely, $\hat{\hat{y}}_i^2$, $\hat{e}^2$ and $\hat{v}_R(e)$ in (2.2.15), (2.2.12) and (2.2.13) respectively.

Then, let us write

$$\hat{v}(e) = (1 + B)\hat{v}_R(e) + B \left[\frac{n}{m} \sum\nolimits_m Q_i \frac{\hat{\hat{y}}_i^2}{p_i^2} - \hat{e}^2 \right].$$

Then we propose

$$v(h) = v_{2r}(h) + \hat{v}(e) \tag{2.2.16}$$

as our unbiased estimator for the variance of h because we have the following result in Theorem 3.

Theorem 3

$$E_1 E_R E_2 E_r v(h) = V(h)$$

Proof We may note, writing $V = E_{1R} V_{2r} + V_{1R} E_{2r}$ that

$$\begin{aligned} V(h) &= E_1 E_R[E_{2r}(v_{2r}(h))] + V_{1R}[E_{2r}(h)] \\ &= E_1 E_R[E_{2r}(v_{2r}(h))] + V_{1R}(e) \text{ since } E_{2r}(h) = e. \end{aligned}$$

Now by using Theorem 1 we get

$$\begin{aligned} V(h) &= E_1 E_R E_2 E_r[v_{2r}(h)] + E_1 E_R v(e) \\ &= E_1 E_R E_2 E_r[v_{2r}(h)] + E_1 E_R E_2 E_r[\hat{v}(e)] \text{ since } E_{2r}[\hat{v}(e)] = v(e) \\ &= E_1 E_R E_2 E_r[v_{2r}(h) + \hat{v}(e)] = E_1 E_R E_2 E_r[v(h)]. \end{aligned}$$

So,

$$E_1 E_R E_2 E_r v(h) = E_1 E_R[V_{2r}(h)] + V_{1R}(e) = V(h).$$

□

3 A simulation exercise to illustrate loss of accuracy in estimation due to sub-sampling

We treat $N = 14$ rural administrative blocks in a particular district in India as the fsu's. The villages within them are treated as the ssu's with their numbers M_i as illustrated in Table 1.

Using the block population according to 1991 Indian Population Census as the size-measures for an RHC sample of size $n = 6$ from these blocks the total area under cultivation, namely $Y = 50083.60$ (in hectares) is estimated, by e in (2.1.1) with $v(e)$ in (2.1.3) as the variance estimator for e.

Treating $d = (e - Y)/\sqrt{v(e)}$ as a standard normal deviate $(e - 1.96\sqrt{v(e)}, e + 1.96\sqrt{v(e)})$ is taken as a 95% confidence interval (CI) for Y. Measures of performance of e are taken as the average coefficient of variation (ACV), the actual coverage percentage (ACP) and the average length (AL) of the CI respectively, which are the average over $R = 1000$ replicated samples, of the value of $100 \times (\sqrt{v(e)}/e)$, the percent of these replicated samples for which the CI covers Y and the average, over these $R = 1000$ replicates of the lengths of these CI's. Variations due to sub-sampling (of m fsu's out of n) only of the fsu's are illustrated in Table 2. Of course the closer the ACP to 95% and the smaller the ACV and the AL the better the estimator.

In Table 3 we illustrate two sets of choices of l_i's as I and II for certain given m_i's and using the same data the performances are compared in Table 4 when e is replaced by h of (2.2.7) and $v(e)$ by $v(h)$ of (2.2.15), but keeping $m = n$, i.e., when no sub-sampling is done at the fsu level and three specified choices of l_i and m_i for $i = 1, 2, \ldots, 14$.

In Table 4 $m_i = l_i$ refers to the situation where no sub-sampling needs to be done at the ssu level.

Table 1 Composition of first stage units

Serial number of fsu's (i)	Number of ssu's (M_i)	Serial number of first stage units (fsu's) (i)	Number of second stage ssu's (M_i)
1	36	8	31
2	22	9	18
3	28	10	43
4	69	11	43
5	72	12	35
6	23	13	22
7	38	14	30

Table 2 Loss in efficacy on sub-sampling of fsu's

Sub-sampling ratio $(\frac{n-m}{n}) \times 100$	Average coefficient of variation (ACV)	Actual coverage percentage (ACP)	Average length (AL)
0	10.8	97.7	20065.6
20	13.7	90.2	21048.6
30	14.2	89.7	25845.1

Table 3 Sub-sampling from sampled ssu's

Serial number of fsu's(i)	m_i	l_i	
		I	II
1	7	6	4
2	4	3	2
3	5	4	3
4	13	11	5
5	14	12	6
6	4	3	3
7	7	6	4
8	6	5	4
9	3	2	2
10	8	7	3
11	8	7	5
12	7	6	4
13	4	3	3
14	6	5	4

Table 4 Relative efficacies on sub-sampling of ssu's: Illustrating performances corresponding to the situations in Table 3

Situation	ACV	ACP	AL
I	12.2	92.3	24865.6
II	12.5	91.1	25386.6
III (with $m_i = l_i$)	12.0	97.1	23617.9

Our Table 2 above backed by Table 1 shows that reduced sizes of the samples of fsu's due to sub-sampling gradually yield diminishing efficacies in estimation through all the three criteria. Similarly, Table 4, supported by data in Table 3 also demonstrates, as expected, similar tendencies for reduced efficacy-levels caused by increase in intensity in sub-sampling at the second stage alone as well.

Acknowledgements Helpful comments and recommendations from two referees leading to improvement on our earlier drafts are gratefully acknowledged. The research by Arijit Chaudhuri is partially supported by CSIR Grant No. 21 (0539)/02/EMR-II.

References

Chaudhuri A, Adhikary AK, Dihidar S (2000) Mean square error estimation in multi-stage sampling. Metrika 52(2):115–131

Rao JNK, Hartley HO, Cochran WG (1962) On a simple procedure of unequal probability sampling without replacement. J Roy Stat Soc B 24:482–491