Pak. J. Statist.
2000, Vol. 16(3), pp 217-227

ON ALTERNATIVE VARIANCE ESTIMATORS IN THREE-STAGE SAMPLING

Arijit Chaudhry, Arun Kumar Adhikary and Shankar Dihidar

Indian Statistical Institute
Calcutta, India.

Abstract

Following Raj's (1968) work on the estimation of the variance of a linear unbiased estimator of a finite population total of a real variable in multistage sampling we take interest in three alternative variance estimation formulae. In two different actual surveys carried out by us we applied two of them in three stage sampling. Being curious about their relative efficacies we undertook a simulation study. The comparative performances are reported for this numerical exercise which seems to show both of them quite competitive justifying the uses of both of them in the two actually implemented surveys. A third variance estimator is also proposed but since it is not yet put to use in an actual survey its efficacy has to be tested before it may be recommended.

Key Words

Sample survey; simulation study; three-stage sampling; unbiased variance estimation.

1. INTRODUCTION

Recently, in the Indian Statistical Institute (ISI), Calcutta, two sample surveys were implemented. One of them was to examine the nature of rural indebtedness in a given geographical area within the administrative jurisdiction of a district. The other was to investigate the growth of small-scale industries and corresponding economic well-being of the villagers in a different district. For both, administrative blocks within the district, the villages within the blocks and households in the villages were naturally considered as the first, second and third stage units while drawing a suitable sample. Moreover, recent census findings on numbers of people and numbers of industries in the villages in respective blocks permitted unequal probability sampling using varying size-measures in the first two stages. Since village-wise details of households and their compositions were unknown to start with, simple random sampling without replacement (SRSWOR) was naturally employed in the third stage of selection in both the surveys.

In the first as well as in the second stage the sample was selected following the scheme due to Rao, Hartley and Cochran (RHC, 1962). To apply this scheme a population is split up at random into as many groups as is the required size of the sample. From each group so formed, one unit is selected with a probability proportional to its known size-measure. Across the groups the selection is 'independent'. For a sample so drawn a formula for a design-unbiased estimator for the population total of any variable of interest is given by RHC. These authors also prescribe how many units are to be assigned to the respective groups mentioned above in a way so as to control the variance of the RHC estimator. From Raj (1968) it is easy to work out a formula for an unbiased estimator of the variance of the above estimator. In one of the above-mentioned surveys this option is put to use. In the other survey we employed an alternative unbiased variance estimator developed by us.

In order to evaluate how efficacious is our proposed variance estimator relative to the traditional one we found it useful to carry out a simulation study. To keep the two rival strategies closely competitive we planned the following artificial formulation exercise. We supposed to have 10 'blocks' in an imaginary district with respective numbers of villages in them between 30 to 60. Choosing 10 integers at random with replacement between 30 to 60 we assigned the chosen numbers as the respective 'block-sizes'. Choosing numbers at random with replacement between 40 and 100 we assigned the chosen numbers to be the number of household (hh) in the respective villages within the respective blocks. Choosing numbers at random with replacement between 1 and 15 we assigned these selected numbers to be the respective household sizes. Using these numbers we work out the population sizes of the respective blocks and the respective villages which we take as 'size-measures' in implementing the RHC scheme in the first two-stages. Obviously, the total population in the imaginary district is thus pre-assigned. Since the third stage units, namely the households, are selected by SRSWOR method and thus varying household sizes are not utilized in drawing the sample in the manner prescribed above the unbiased estimator for the district's total population size should not equal this parameter itself through the estimator is expected to be quite accurate. To measure this accuracy we work out the variance estimator by the 'traditional' as well as our 'proposed' method. Since the population is totally at hand we repeat the sample selection, unbiased estimation of the total population size and unbiased variance estimation by each of the two methods a very large number of times, say, R taken equal to 1000. Based on these R replicates we determine the actual percentage of the replicates for which the true known population total is covered within the confidence intervals based on the respective samples. A 100 $(1 - \alpha)$ percent confidence interval, with $\alpha \in (0, 1)$ is constructed by treating the pivotal quantity namely the ratio of 'the estimated minus the true population size' to the square root of the estimated variance of the estimate (the standard error) to be a standard normal deviate. The percentage considered above is called an "Actual Coverage Percentage" (ACP). To have an idea of the width of the confidence intervals we also calculate the average, over the R replicates, of the ratio of the estimated standard error to the estimated total. The less the value of this average coefficient of variation (ACV), the better the confidence interval. For the two

rival variance estimators the values of ACP should vary differently from 100 $(1 - \alpha)$ and the values of ACV also should vary.

From these variations we may assess the comparative efficacies of the two variance estimators, the estimator for the total itself remaining the same in the 'pivotal' mentioned above. The estimation formulae are presented in section 2 and the details of simulation results along with our recommendations in section 3 below. Our proposed alternative variance estimator seems to fare competitively with the traditional one in the light of our simulation exercise reported in what follows. This vindicates the success of both the surveys implemented by us because one of them uses one of the two variance estimators and the other employed the other one.

2. NOTATIONS AND VARIANCE ESTIMATORS

Let $U = (1,...,i,...,N)$ denote a population of N first stage units (fsu) and y a real variable with values y_i for i in U. Let p_i denote known normed size-measures for i in U. By $\sum y_i = Y$ we denote the total of y_i over i in U which we need to estimate on taking a sample from U in three stages. In the first stage a sample of n fsu's is drawn from U employing the RHC scheme. For this, U is split at random into n non-overlapping groups taking in the gth group $(g = 1,...,n)N_g$ units. Here N_g is so determined that each is an integer closest to N/n subject to $\sum_n N_g = N$. By $\sum_n$ we mean summing over the n groups. From each group so formed separately and independently one fsu is chosen with a probability proportional to its p-value. For simplicity we write ρ_i and y_i respectively for the p-value and y-value of the unit chosen from the ith group $(i = 1, ..., n)$.

If y_i values were ascertainable for the sampled fsu, then one could estimate Y unbiasedly by the RHC estimator given by

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

Here Q_i denotes the sum of the p-values over the N_i fsu's falling in the ith group formed as above.

An unbiased estimator for the variance of t is given by RHC as

$$v_1(t) = \frac{\sum_n N_i^2 - N}{N^2 - \sum_n N_i^2}\left(\sum_n y_i^2 \frac{Q_i}{p_i^2} - t^2\right) \tag{2.2}$$

If y_i's were ascertainable.

In the specific survey situation of our interest as noted earlier y_i is not ascertainable. The ith fsu is supposed to consist of M_i second stage units (ssu) and for

the $j$ht ssu in the ith fsu the known normed size-measure is p_{ij} and the unknown y-value is y_{ij} ($j = 1, \ldots, M_i; i = 1, \ldots, n$). Then y_i is the sum of the M_i values namely y_{ij}. On taking a sample of m_i ssu's from the ith fsu, if selected, applying the RHC scheme, using p_{ij}'s as normed size-measures, clearly y_i may be unbiasedly estimated by

$$x_i = \sum_{m_i} y_{ij} \frac{Q_{ij}}{p_{ij}} \tag{2.3}$$

if y_{ij}'s are ascertainable.

Here $\sum_{m_i}$ *and* Q_{ij} correspond to $\sum_n$ *and* Q_i in an obvious way.

An unbiased variance estimator for x_i is

$$v_2(x_i) = \frac{\sum_{m_i} N_{ij}^2 - M_i}{M_i^2 - \sum_{m_i} N_{ij}^2} \left(\sum_{m_i} y_{ij}^2 \frac{Q_{ij}}{p_{ij}^2} - x_i^2 \right) \tag{2.4}$$

corresponding to $v_1(t)$ for t. Here N_{ij}'s are analogous to N_i's.

For simplicity we shall write

$$A = \frac{\sum_n N_i^2 - N}{N^2 - \sum_n N_i^2} \quad \text{and} \quad A_i = \frac{\sum_{m_i} N_{ij}^2 - M_i}{M_i^2 - \sum_{m_i} N_{ij}^2}$$

Since y_{ij} is also not ascertainable, it is estimated by

$$w_{ij} = \frac{T_{ij}}{t_{ij}} \sum_k y_{ijk} \tag{2.5}$$

Here T_{ij} is the number of third stage units (tsu) in "the jth ssu of ith fsu" and t_{ij} is the number of tsu's sampled out of T_{ijk} with y_{ijk} as the value of the kth tsu out of those in T_{ij} and Σ_k is the sum over the t_{ij} sampled tsu's.

An unbiased variance estimator of w_{ij} is

$$v_3(w_{ij}) = T_{ij}^2 \left(\frac{T_{ij} - t_{ij}}{T_{ij}} \right) \frac{1}{t_{ij}} \left(\frac{1}{t_{ij} - 1} \right) \sum_k \left(y_{ijk} - \frac{w_{ij}}{T_{ij}} \right)^2 \tag{2.6}$$

At this stage, let us follow Raj (1968) in nothing the theory of estimation of a survey population total and variance estimation in multistage sampling in general.

Let $\underline{Y} = (y_1, \ldots, y_i, \ldots, y_N)$, $Y = \Sigma y_i$,

$\underline{R} = (r_1, \ldots, r_i, \ldots, r_N)$ $R = \Sigma r_i$, $\underline{V} = (V_1, \ldots, V_i, \ldots V_N)$, $\underline{v} = (v_1, \ldots, v_i, \ldots, v_N)$,

E_1, E_L the operators of expectation in the first and the later stages of sampling and V_1, V_L the corresponding variance operators. Here r_i's are estimators of y_i's obtained through sampling of the later stage units of i 'maintaining independence' across i in the selection process in subsequent stages such that

$$E_L(r_i) = y_i,\ V_L(r_i) = V_i \text{ and } E_L(v_i) = V_i;$$

Here v_i's are variance estimators 'fsu'-wise. Let E, V denote expectation and variance operators over all the sampling stages.

Let $t = t(s, \underline{Y})$ be an estimator for Y such that, presuming y_i's are ascertainable for sampled fsu's,

$$E_1(t) = Y.$$

Writing $I_{si} = 1$ if $i \in s$, 0 else, $I_{sij} = I_{si}I_{sj}$ and confining to the form of t as

$$t = \Sigma y_i b_{si} I_{si}$$

with b_{si}'s as constants free of $\underline{Y}$, is $\Sigma\Sigma$ as sum over $i \neq j$, $V_1(t) = \sum y_i^2 c_i + \sum\sum y_i y_j c_{ij}$ where

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

Let there exist constants d_{si}, d_{sij} free of $\underline{Y}$ such that

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

such that

$$E_1(d_{si} I_{si}) = c_i, E_1(d_{sij} I_{sij}) = c_{ij}.$$

Then, $E_1 v_1(t) = V_1(t)$.

$$\text{Let } e = e(s, \underline{R}) = \sum r_i b_{si} I_{si}.$$

Then,

$$E(e) = E_1 E_L(e) = E_1(t) = Y.$$

Also,

$$\begin{aligned} E_1(e) &= R,\ E_L E_1(e) = E_L R = Y = E_1 E_L(e) \\ V(e) &= E_1 E_L(e - Y)^2 = E_1 E_L[(e - E_L(e)) + E_L(e) - Y]^2 \\ &= E_1 V_L(e) + E_1 (t - Y)^2 = E_1(\sum V_i b_{si}^2 I_{si}) + V_1(t) \end{aligned}$$

Also, $$\begin{aligned} E_1 V(e - Y)^2 &= E_1[(e - E_1(e)) + (E_1(e) - Y)]^2 \\ &= V_1(e) + E_1(R - Y)^2 = \sum r_i^2 c_i + \sum\sum r_i r_j c_{ij} + (R - Y)^2 \end{aligned}$$

Then,

$$\begin{aligned} E_L E_1(e - Y)^2 &= V_1(t) + \Sigma v_i c_i + V_L(R) = V_1(t) + \sum V_i \left[E_1(b_{si}^2 I_{si})\right] \\ \text{Now, } v_1(e) &= \sum r_i^2 d_{si} I_{si} + \sum\sum r_i r_j d_{sij} I_{sij} \\ \text{Then, } E_1 v_1(e) &= v_1(t) + \Sigma V_i d_{si} I_{si} \\ \text{So, } E_1 E_L v_1(e) &= V_1(t) + \Sigma V_i E_1(d_{si} I_{si}) \\ \text{So, } v^*(e) &= v_1(e) + \sum v_i (b_{si}^2 - d_{si}) I_{si} \\ \text{Satisfies } E_1 E_L\, v^*(e) &= E_1 E_L(e - Y)^2 = V(e). \end{aligned}$$

Again,

$$\begin{aligned} E_1 v_1(e) &= \sum r_i^2 c_i + \sum\sum r_i r_j c_{ij} \\ \text{So, } E_1 E_L v_1(e) &= \sum y_i^2 c_i + \sum\sum y_i y_j c_{ij} + \sum V_i c_i = V_1(t) + \sum V_i c_i \end{aligned}$$

So,

$$v(e) = v_1(e) + \Sigma v_i b_{si} I_{si}$$

satisfies

$$E_L E_1 v(e) = V_1(t) + \sum V_i E_1(b_{si}^2 I_{si}) = E_L E_1(e - Y)^2.$$

If we assume that $E_1 E_L = E_L E_1$, then

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

So, $v(e)$ and $v^*(e)$ are both unbiased estimators for $V(e)$. The formula $v(e)$ is due to Raj (1968). The form $v^*(e)$ is similar to one due to Rao (1975) except that in Rao (1975) the form of V_i is more complicated; it is V_{si} so that it may involve units other than i in the sample s of fsu's drawn.

So, in our example we may write

$$e = \sum_n \frac{Q_i}{p_i} x_i \tag{2.7}$$

Then, from Raj (1968) we have, for e, an unbiased variance estimator

$$v(e) = v_1(t)\Big|_{y_i = x_i} + \sum_n \frac{Q_i}{p_i} v_2(x_i) \tag{2.8}$$

Letting

$$z_i = \sum_{m_i} \frac{Q_{ij}}{p_{ij}} w_{ij}, \tag{2.9}$$

From Raj (1968) one may derive for Y an unbiased estimator

$$\hat{Y} = \sum_n \frac{Q_i}{p_i} z_i \tag{2.10}$$

Then, from Raj (1968) again, one has for $\hat{Y}$ an unbiased variance estimator as

$$v = v(e)\Big|_{x_i = z_i} + \sum_n \frac{Q_i}{p_i} v(z_i) \tag{2.11}$$

writing

$$v(e) = v_1(t)\Big|_{y_i = z_i} + \sum_n \frac{Q_{ij}}{p_{ij}} v_3(w_{ij}) \tag{2.12}$$

This v may be referred to as a traditional variance estimator for $\hat{Y}$.

Though there is no compelling reason for it, the following unbiased variance estimator, say, v for $\hat{Y}$ is proposed as an alternative to v, out of curiosity and in anticipation of higher efficiency, if feasible.

Collecting the appropriate coefficient let us express $v_1(t)$ as the following quadratic from:

$$v_1(t) = \sum b_{si} y_i^2 + \sum \sum b_{sij} y_i y_j \tag{2.13}$$

writing $\sum$ as sum over the units i in the sample s of fsu's from U drawn as described above, $\sum \sum$ as the corresponding distinct sampled pairs $i, j(i \neq j)$, b_{si}'s as coefficients of y_i^2 and b_{sij} as coefficient of $y_i y_j$ in $v_1(t)$ of (2.2).

Let further,

$$v_2(t) = \sum b_{si} x_i^2 + \sum \sum b_{sij} x_i x_j \tag{2.14}$$

and

$$v_3(t) = \sum b_{si} z_i^2 + \sum \sum b_{sij} z_i z_j \tag{2.15}$$

Further, let us write

$$v_3(z_i) = \sum_{m_i} \left(\frac{Q_{ij}}{p_{ij}} \right)^2 v_3(w_{ij}) \tag{2.16}$$

$$v_{2i} = A_i \left(\sum_{mi} \frac{Q_{ij}}{p_{ij}} w_{ij}^2 - z_i^2 \right) \tag{2.17}$$

and

$$\hat{v}_2(x_i) = v_{2i} - A_i \sum_{m_i} \frac{Q_{ij}}{p_{ii}} (1 - Q_{ij}) v_3(w_{ij}) \tag{2.18}$$

Then, let

$$v = v_3(t) - \sum b_{si} v_3(z_i) + \sum_n \frac{Q_i}{p_i} \hat{v}_2(x_i) + \sum_n \left(\frac{Q_i}{p_i} \right)^2 v_3(z_i) \tag{2.19}$$

It is easy to check that v is an unbiased estimator of the variance of $\hat{Y}$ and this is our proposed alternative to v.

Remark I: Unlike v, the estimator v may take a negative value. In such a case its use is not recommended.

Remark II: In out actual survey it came out positive. The formula $v^*(e)$ is not yet known to have been put to use in practice. It may be worth trying.

3. A SIMULATION STUDY FOR v VERSUS v

In section 1 we indicated how for an imaginary district with 10 rural blocks composed of various numbers of villages with varying numbers of households (hh) with variable sizes the population figures at hh, village and block levels and hence for the entire district were generated. Some specimens are revealed in the table below.

Table 1: Showing composition of 10 blocks in a district

Serial No. of block	No. of Villages in blocks	Total population in blocks
1	39	23239
2	40	22253
3	55	32756
4	51	29074
5	60	35079
6	59	33624
7	56	31373
8	41	21435
9	33	19219
10	42	23934
Total:	476	271986

First, out of 10 blocks, 4 blocks are selected by RHC method using numbers of villages within blocks as size measures. From each selected block, a 22 percent (rounded upward to an integer) sample of villages in drawn as above by RHC method with village-population as the size measure. From each selected village a 4 percent (rounded upward to an integer) SRSWOR sample is drawn. The total district population that is Y = 271986 is required to be unbiasedly estimated using the observations in the above three stage sample pretending the values for the unsampled units at each stage to be unknown. The estimate $\hat{Y}$ in (2.10) for Y is calculated along with v in (2.11) and v in (2.19) for each of R = 1000 replicated samples drawn as above.

Next we calculate, based on these replicated values of $(\hat{Y}, v, v)$, the summary measures:

(i) ACP = (Actual coverage percentage) = the percentage of replicated samples for which $\hat{Y} - 1.96\sqrt{w}, \hat{Y} + 1.96\sqrt{w},)$ covers Y, taking w as v and v – the closer it is to 95 percent, the prescribed confidence coefficient, the better;

(ii) ACV = (Average coefficient of variance) = the average, over R replicates, of the value of $\frac{\sqrt{w}}{y}$, taking w as v and v – the smaller its value, the better.

The summarized findings, so as to compare the performances of v relative to v are presented in the table below.

Table 2: Summary of efficacy of v versus v for the first Three consecutive replicated sets

Serial No. of set of replicates	No. of replicates in the set	ACP using		ACV using		Percent of replicates in the set gives v less than v
		v	v	v	v	
1	300	94.34	92.67	5.55	5.53	54.67
2	300	95.33	95.00	5.57	5.54	58.00
3	400	97.00	96.75	5.59	5.58	54.50
Total	1000	95.70	95.00	5.57	5.55	55.60

Remark III. In each of the R = 1000 replicates v turned out to be positive.

CONCLUSION AND RECOMMENDATION

In situations similar to the ones cited above, there is not much to choose between the two variance estimators put into practice by us though the newly proposed one seems to slightly outperform the traditional one. So in practice both may be employed. The third one proposed by us namely $v^*(e)$ may also be quite competitive but we cannot claim that since we have no empirical evidence yet to support it. In a future survey we plan to try it out.

ACKNOWLEDGEMENT

The authors are grateful to a referee who comments helped them in improving upon an earlier draft.

REFERENCE

(1) Raj, D. (1968): *Sample Theory*. McGraw-Hill, N.Y.
(2) Rao, J.N.K. (1975). Unbiased variance estimation for multi-stage designs. *Sankhyä*, C, 37, 133-139.
(3) Rao, J.N.K., Hartley, H.O. and Cochran, W.G. (1962): On a simple procedure of unequal probability sampling without replacement. *Jour. Roy. Stat. Soc.* B, 24, 482-491.

An Appendix

Using the data as in Table 1 and following the sample procedure as reported in Table 2 we carried out another numerical exercise to compare the performances of the variance estimator $v^* = v^*(e)$ given on p.7 vis-à-vis v and v for the estimator v of a finite population total. The Table 3 below presents a summary.

Table 3: A summary of efficacies of v, v , v^*

Serial No. of set of replicates	No. of replicates in the set	ACP using			ACV using		
		v	v	v^*	v	v	V^*
1	300	94.67	96.33	92.33	5.64	5.63	4.88
2	300	97.00	97.67	91.00	5.59	5.57	4.82
3	400	94.75	95.25	88.75	5.66	5.63	4.87
Total	1000	94.50	95.20	91.20	5.66	5.64	4.91

Comments. The third competitor v^* proposed by us may also be treated as a variable competitor and is worth trying in practice.