On certain alternative mean square error estimators in complex survey sampling

Arijit Chaudhuri [*], Sanghamitra Pal

Applied Statistics Unit, Indian Statistical Institute, 203 B.T. Road, Calcutta 700 035, India

Received 1 November 2000; received in revised form 14 May 2001; accepted 4 June 2001

Abstract

Rao (J. Indian Statist. Assoc. 17 (1979) 125) has given a 'necessary form' for an unbiased mean square error (MSE) estimator to be 'uniformly non-negative'. The MSE is of a homogeneous linear estimator 'subject to a specified constraint', for a survey population total of a real variable of interest. We present a corresponding theorem when the 'constraint' is relaxed. Certain results are added presenting formulae for estimators of MSEs when the variate-values for the sampled individuals are not ascertainable. Though not ascertainable, they are supposed to be suitably estimated either by (1) randomized response techniques covering sensitive issues or by (2) further sampling in 'subsequent' stages in specific ways when the initial sampling units are composed of a number of sub-units. Using live numerical data, practical uses of the proposed alternative MSE estimators are demonstrated.

MSC: 62D05

Keywords: Homogeneous linear estimator; Mean square error; Multi-stage sampling; Randomized response; Selection with unequal probabilities

1. Introduction

We consider a survey population $U=(1,\ldots,i,\ldots,N)$ of a known number N of identifiable units labelled $i=1,\ldots,N$. On it is defined a real variable of interest y with values y_i with a population total $Y=\sum y_i$, writing $\sum$ to denote sum over i in U. We suppose that a sample s is drawn from U with a probability $p(s)$ and the values y_i are ascertained for the units i in s. A homogeneous linear estimator (HLE) for Y is to be employed. For such an estimator written as

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

with b_{si}'s as constants free of $\underline{Y}=(y_1,\ldots,y_i,\ldots,y_N)$ and $I_{si}=1$ if $i\in s$; and 0 if $i\notin s$, then the MSE is

$$M(t_b)=E_1(t_b-Y)^2=\sum\sum d_{ij}\, y_i y_j. \tag{1.2}$$

Here, E_1 denotes expectation with respect to the design p corresponding to $p(s)$ above; $\sum\sum$ denotes sum over i, j in U with no restrictions:

$$d_{ij} = E_1(b_{si}I_{si} - 1)(b_{sj}I_{sj} - 1).$$

Rao (1979) considered a sub-class of estimators t_b in (1.1) for which the following "condition, say C" holds:
'If there exist w_i $(\neq 0)$ as constants free of $\underline{Y}$, then

$$M(t_b) = 0 \quad \text{if } z_i = \frac{y_i}{w_i} \text{ for every } i \text{ in } U \text{ is a constant.} \tag{1.3}$$

Under 'C' it follows that $M(t_b)$ may be written as

$$M(t_b) = -\sum{}'\sum{}' d_{ij} w_i w_j \left(\frac{y_i}{w_i} - \frac{y_j}{w_j}\right)^2, \tag{1.4}$$

where $\sum'\sum'$ denotes sum over i, j $(i < j)$ in U. He then deduced that in the class of homogeneous quadratic unbiased estimators (HQUE) of $M(t_b)$ one that may be uniformly non-negative (UNN) is 'necessarily of the form'

$$m(t_b) = -\sum{}'\sum{}' d_{sij} I_{sij} w_i w_j \left(\frac{y_i}{w_i} - \frac{y_j}{w_j}\right)^2, \tag{1.5}$$

where d_{sij}'s are constants free of $\underline{Y}$ and $I_{sij} = I_{si}I_{sj}$

$$\text{subject to } E_1(d_{sij}I_{sij}) = d_{ij} \quad \text{for every } i, j \text{ in } U. \tag{1.6}$$

In the next section, we present certain results when the 'Constraint C' is relaxed.

2. Alternative MSE estimators

Rao (1979) illustrated several classical sampling strategies for which the above theory applies. For example, by denoting $\pi_i = \sum p(s)I_{si}$, as the inclusion probability of i, the Horvitz and Thompson's (HT, 1952) estimator (HTE) given by $t_{\mathrm{H}} = \sum y_i(I_{si}/\pi_i)$, assuming $\pi_i > 0$ for every i in U, satisfies 'C' if '$v(s) = \sum I_{si}$, the number of distinct units in s, is a constant for every s with $p(s) > 0$'.

Since t_{H} is unbiased for Y its MSE equals its variance which is, writing $\pi_{ij} = \sum p(s)I_{sij}$, as given by HT (1952): $V_1(t_{\mathrm{H}}) = \sum y_i^2(1 - \pi_i)/\pi_i + 2\sum'\sum' y_i y_j(\pi_{ij} - \pi_i\pi_j)/\pi_i\pi_j$, in line with (1.2). If $p(s) > 0 \Rightarrow v(s)$ is a constant, then 'C' holds and in accordance with (1.4), $V_1(t_{\mathrm{H}})$ equals

$$V_2(t_{\mathrm{H}}) = \sum{}'\sum{}' (\pi_i\pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2.$$

Accordingly, if 'C holds', t_{H} has its unbiased variance estimator given by Yates and Grundy (YG, 1953) as $v_{\mathrm{YG}} = \sum\sum(\pi_i\pi_j - \pi_{ij})(I_{sij}/\pi_{ij})(y_i/\pi_i - y_j/\pi_j)^2$, assuming $\pi_{ij} > 0$ $\forall i, j$. This is 'UNN' if $\pi_i\pi_j \geqslant \pi_{ij}$ $\forall i, j$.

If ‘C’ does not hold, then $V_1(t_{\mathrm{H}})$ may be unbiasedly estimated by

$$v_{\mathrm{HT}} = \sum y_i^2 \, \frac{1-\pi_i}{\pi_i}\frac{I_{si}}{\pi_i} + 2\sum{}'\sum{}' y_i y_j \left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\right)\frac{I_{sij}}{\pi_{ij}}, \tag{2.1}$$

as was proposed by HT (1952).

Unfortunately, it is difficult to work out conditions for ‘UNN’ property of v_{HT}, except noting that π_i, π_{ij}’s should be such that v_{HT} must be a ‘non-negative definite’ as a quadratic form.

We intend to provide another alternative unbiased variance estimator for t_{H} when ‘C’ fails and yet it is easy to present conditions for its ‘UNN’ property. Before that we present a theorem covering the general estimator t_b for which ‘C’ fails.

Theorem 1. *Let there exist non-zero constants w_i free of $\underline{Y}$, for $i \in U$. Then, writing $z_i = y_i/w_i$, it follows that*

$$M(t_b) = -\sum{}'\sum{}' d_{ij} w_i w_j (z_i - z_j)^2 + \sum \frac{y_i^2}{w_i}\alpha_i, \quad \text{where } \alpha_i = \sum_{j=1}^{N} d_{ij} w_j. \tag{2.2}$$

Proof.

$$-\sum{}'\sum{}' d_{ij} w_i w_j (z_i - z_j)^2 = -\frac{1}{2}\sum\sum_{i\neq j} d_{ij} w_i w_j \left(\frac{y_i^2}{w_i^2} + \frac{y_j^2}{w_j^2} - \frac{2y_i y_j}{w_i w_j}\right)$$

$$= \sum\sum_{i\neq j} d_{ij} y_i y_j - \sum_i \frac{y_i^2}{w_i}\left(\sum_{j=1}^{N} d_{ij} w_j - d_{ii} w_i\right)$$

$$= \sum\sum d_{ij} y_i y_j - \sum_i \frac{y_i^2}{w_i}\alpha_i.$$ Hence the result. □

Corollary 1. *Two unbiased estimators of $M(t_b)$ are*

$$m_1(t_b) = -\sum{}'\sum{}' d_{sij} I_{sij} w_i w_j \left(\frac{y_i}{w_i} - \frac{y_j}{w_j}\right)^2 + \sum \frac{y_i^2}{w_i}\alpha_i \frac{I_{si}}{\pi_i},$$

$$m_2(t_b) = -\frac{1}{p(s)}\left[\sum{}'\sum{}' c_{sij} w_i w_j \left(\frac{y_i}{w_i} - \frac{y_j}{w_i}\right)^2 - \sum \frac{y_i^2}{w_i}\alpha_i c_{si}\right]$$

writing

$$c_{sij} = \frac{I_{sij} d_{ij}}{\sum_s I_{sij}}, \qquad c_{si} = \frac{I_{si}}{\sum_s I_{si}}.$$

Remark I. Conditions for the ‘UNN’ properties of $m_1(t_b)$ and $m_2(t_b)$ are obviously (i) $w_i w_j d_{sij} I_{sij} \leqslant 0, w_i \alpha_i I_{si} \geqslant 0$ for the former and (ii) $w_i w_j c_{sij} \leqslant 0$, $w_i \alpha_i I_{si} \geqslant 0$ for the latter.

Remark II. $\pi_i > 0 \ \forall i \to \sum_s I_{si} > 0 \forall i$ and $\pi_{ij} > 0 \ \forall i, j \Rightarrow \sum_s I_{sij} > 0 \ \forall i, j$.

This is easy to prove.

Corollary 2. *If 'C' does not hold, writing $v=\sum v(s)p(s)$, the variance of t_{H} equals, on allowing $v(s)$ to vary with $s, p(s)>0$,*

$$V_3(t_{\mathrm{H}})=\sum{}'\sum{}'(\pi_i\pi_j-\pi_{ij})\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2+\sum\frac{y_i^2}{\pi_i}\beta_i \tag{2.3}$$

writing

$$\beta_i=\left(1+\frac{1}{\pi_i}\sum_{j\neq i}\pi_{ij}-v\right),\quad i\in U.$$

Proof. Easy on recalling $v=\sum\pi_i$ and is hence omitted. □

Corollary 3. *Two unbiased estimators of $V_3(t_{\mathrm{H}})$ are*

$$v_1(t_{\mathrm{H}})=\sum{}'\sum{}'\frac{I_{sij}}{\pi_{ij}}(\pi_i\pi_j-\pi_{ij})\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2+\sum\frac{I_{si}}{\pi_i}\frac{y_i^2}{\pi_i}\beta_i \tag{2.4}$$

and

$$v_2(t_{\mathrm{H}})=\frac{1}{p(s)}\left[\sum{}'\sum{}'I_{sij}\left(\frac{\pi_i\pi_j-\pi_{ij}}{\sum_s I_{sij}}\right)\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2+\sum c_{si}\frac{y_i^2}{\pi_i}\beta_i\right]. \tag{2.5}$$

Remark III. Conditions for the 'UNN' properties of both $v_1(t_{\mathrm{H}})$ and $v_2(t_{\mathrm{H}})$ are '$\pi_i\pi_j\geqslant\pi_{ij}\ \forall i,j\ (i\neq j),\ \beta_i\geqslant 0\ \forall i$'.

In Section 3, we illustrate situations when (A) $v(s)$ varies with s when $p(s)>0$, but the (B) conditions for 'UNN' properties of $v_1(t_{\mathrm{H}})$ and $v_2(t_{\mathrm{H}})$ hold.

3. An illustrative sampling scheme for which the 'constraint C' does not hold but alternatives to HT, YG estimators have 'UNN' property

Brewer (1963) has given a sampling scheme when normed size-measures p_i ($0<p_i<1\ \forall i, \sum p_i=1$) are available for $i\in U$. Here, on the first draw, the unit i is chosen with a probability proportional to $q_i=p_i(1-p_i)/(1-2p_i)$ and leaving aside the unit i so chosen, a second unit $j(\neq i)$ is chosen in the second draw with a probability $p_j/(1-p_i)$. Writing $D=\sum(p_i/(1-2p_i))$ he showed that for this scheme the inclusion probability π_i (2) for i equals $2p_i$ and that for (i,j), denoted by π_{ij} (2) equals $(2p_ip_j/(1+D))(\frac{1}{1-2p_i}+\frac{1}{1-2P_j})$. It is further known that

$$\Delta_{ij}(2)=\pi_i(2)\pi_j(2)-\pi_{ij}(2)\geqslant 0\quad\forall i,j(i\neq j)\text{ in }U. \tag{3.1}$$

We use '2' within parentheses to emphasize that this scheme uses 2 draws. Let the sample chosen as above be augmented by adding to the 2 distinct units so drawn as above, $(r-2)$ further distinguish the units from the remaining $(N-2)$ units of U by simple random sampling (SRS) without replacement (WOR). For such a scheme introduced by Seth (1966) admitting r distinct units in each sample, the inclusion probabilities $\pi_i(r)$ for i and $\pi_{ij}(r)$ for $(i,j)(i \neq j)$, involving $r(>2)$ draws, are

$$\pi_i(r) = \frac{1}{N-2}[(r-2) + (N-r)\pi_i(2)],$$

$$\pi_{ij}(r) = \pi_{ij}(2) + \left(\frac{r-2}{N-2}\right)(\pi_i(2) + \pi_j(2) - 2\pi_{ij}(2))$$
$$+ \left(\frac{r-2}{N-2}\right)\left(\frac{r-3}{N-3}\right)(1 - \pi_i(2) - \pi_j(2) + \pi_{ij}(2)).$$

Let us slightly modify this sampling scheme of Seth (1966) by allowing $(r-2)$ to be (1) a number $(n-2)$ to be chosen with a pre-assigned probability w $(0 < w < 1)$ and (2) a number $(n-1)$ to be chosen with the complementary probability $(1-w)$. Then a sample s so drawn will have a size 'n with probability w' and '$(n+1)$ with probability $(1-w)$'. Thus, $\nu(s)$ is either n or $(n+1)$. Putting $n, (n+1)$ by turn in $\pi_i(r), \pi_{ij}(r)$ above we get the inclusion probabilities π_i^*, say, for i and π_{ij}^* for (i,j) for this sampling scheme as

$$\pi_i^* = w\pi_i(n) + (1-w)\pi_i(n+1)$$

and

$$\pi_{ij}^* = w\pi_{ij}(n) + (1-w)\pi_{ij}(n+1).$$

Then, we have

Theorem 2. $\pi_i^*\pi_j^* \geqslant \pi_{ij}^* \quad \forall i,j\ (i \neq j)$ *in* U.

Proof. On simplifications we get, using the results of this section,

$$\pi_i^*\pi_j^* = \pi_i(2)\pi_j(2) + \frac{(n-1-w)}{(N-2)}(\pi_i(2) + \pi_j(2) - 2\pi_i(2)\pi_j(2))$$
$$+ \left[\frac{(n-1-w)}{(N-2)}\right]^2 (1 - \pi_i(2) - \pi_j(2) + \pi_i(2)\pi_j(2)) \tag{3.2}$$

and

$$\pi_{ij}^* = \pi_{ij}(2) + \frac{(n-1-w)}{(N-2)}(\pi_i(2) + \pi_j(2) - 2\pi_{ij}(2))$$
$$+ \frac{(n-2)(n-1-2w)}{(N-2)(N-3)}(1 - \pi_i(2) - \pi_j(2) + \pi_{ij}(2)). \tag{3.3}$$

Subtracting (3.3) from (3.2) and using (3.1), on further simplifications the theorem is immediately proved. □

Next, let us note the

Lemma. *For any design* p

$$\sum_{s \ni i} \nu(s)p(s) = \sum_{j \neq i} \pi_{ij} + \pi_i.$$

Proof. $\pi_{ij}=\sum_s p(s)I_{sij}$. So, $\sum_{j\neq i}\pi_{ij}=\sum_s p(s)(v(s)-1)I_{si}=\sum_{s\ni i}p(s)v(s)-\pi_i$. This gives the result. □

For the scheme of sampling given by Seth (1966) as modified above, the formula for β_i turns out to be

$$\beta_i^*=1+\frac{1}{\pi_i^*}\sum_{j\neq i}\pi_{ij}^*-\sum\pi_i^*.$$

The above lemma yields

Theorem 3. $\beta_i^*>0$.

Proof. $\sum_{s\ni i}p(s)v(s)=wn\pi_i(n)+(1-w)(n+1)\pi_i(n+1)$. So,

$$\beta_i^*=\frac{1}{\pi_i^*}[wn\pi_i(n)+(1-w)(n+1)\pi_i(n+1)]-[wn+(1-w)(n+1)]$$

$$=\frac{1}{\pi_i^*}[w(1-w)(\pi_i(n+1)-\pi_i(n)]>0$$

because $\pi_i(n+1)-\pi_i(n)=(1-\pi_i(2))/(N-2)>0$. Since for this scheme $v(s)$ is either n with a probability w or $(n+1)$ with a probability $(1-w)$, Rao's (1979) 'constraint C' here is violated. Yet, for t_{H} based on this scheme, our proposed new estimators $v_1(t_{\mathrm{H}})$, $v_2(t_{\mathrm{H}})$ for $V_3(t_{\mathrm{H}})$ are 'uniformly non-negative'.

It seems to us that the celebrated generalized regression (greg) estimator or predictor given by Cassel et al. (1976) for Y needs a discussion in the present context. We take it up below.

4. MSE estimation for greg predictor

Let $\underline{X}=(x_1,\ldots,x_i,\ldots,x_N)$ with x_i as the value of an auxiliary variable x for i in U with a known total $X=\sum x_i$. Then, choosing numbers R_i suitably for example as $1/x_i$ or $1/\pi_i x_i$, or $(1-\pi_i)/\pi_i x_i$ or $1/x_i^2$ etc the greg predictor for Y is

$$t_g=\sum\frac{y_i}{\pi_i}I_{si}g_{si}=\sum\frac{y_i}{\pi_i}I_{si}+b_R(X-\sum\frac{x_i}{\pi_i}I_{si}),$$

writing

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

$$b_R=\frac{\sum y_ix_iR_iI_{si}}{\sum x_i^2R_iI_{si}}.$$

Let $B_R=\sum y_ix_iR_i\pi_i/\sum x_i^2R_i\pi_i$, $e_i=y_i-b_Rx_i$ and $E_i=y_i-B_Rx_i$.

The t_g is a biased predictor for Y but its bias may be neglected for large samples. Assuming large samples and applying Taylor series neglecting terms involving second and higher order derivatives, the following formulae for MSE of t_g and estimators of MSE are well known, especially from Särndal et al. (SSW, 1992).

$M_1 = M(t_g) = \sum E_i^2(1-\pi_i)/\pi_i + 2\sum'\sum' E_iE_j(\pi_{ij}-\pi_i\pi_j)/\pi_i\pi_j =$ variance of $\sum E_iI_{si}/\pi_i$, vide HT (1952), $M_2 = M(t_g) = \sum'\sum'(\pi_i\pi_j - \pi_{ij})(E_i/\pi_i - E_j/\pi_j)^2$, following YG (1953), applicable when '$v(s)$ is a constant for every s with $p(s) > 0$' — an example where Rao's (1979) 'constraint C' is imposed taking (i) $E_i = a\pi_i \;\forall i$ with 'a' as a constant and (ii) $v(s) = v$ for every s with $p(s) > 0$. Estimators of M_1 are well known to be

$$m_{kg} = \sum(a_{ki}e_i)^2\frac{1-\pi_i}{\pi_i}\frac{I_{si}}{\pi_i} + 2\sum'\sum'(a_{ki}e_i)(a_{kj}e_j)\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\right)\frac{I_{sij}}{\pi_{ij}},$$

$$k = 1,2, \qquad a_{1i} = 1, \qquad a_{2i} = g_{si}. \tag{4.1}$$

Estimators of M_2 are well known to be

$$r_{kg} = \sum\sum(\pi_i\pi_j - \pi_{ij})\left(\left(\frac{a_{ki}e_i}{\pi_i} - \frac{a_{kj}e_j}{\pi_j}\right)^2\right)\frac{I_{sij}}{\pi_{ij}}, \quad k = 1,2. \tag{4.2}$$

For m_{kg}, conditions for 'uniform non-negativity' are difficult to check, but they are usable even if $v(s)$ varies with s. On the contrary, r_{kg} is 'UNN' if $\pi_i\pi_j \geqslant \pi_{ij}$ for $i \neq j$, but its use is recommended if '$v(s)$ is a constant for every s with $p(s) > 0$'. If $v(s)$ varies with s then we have the following alternative approximate formula for $M(t_g)$ based on Taylor series expansion:

$$M_3 = M(t_g) = \sum'\sum'(\pi_i\pi_j - \pi_{ij})\left(\left(\frac{E_i}{\pi_i} - \frac{E_j}{\pi_j}\right)^2\right) + \sum\frac{E_i^2}{\pi_i}\beta_i \tag{4.3}$$

To observe this, let us write

$$t_1 = \sum\frac{y_i}{\pi_i}I_{si}, \quad t_2 = \sum\frac{x_i}{\pi_i}I_{si}, \quad t_3 = \sum y_ix_iR_iI_{si},$$

$$t_4 = \sum x_i^2R_iI_{si}, \quad \underline{t} = (t_1, t_2, t_3, t_4)$$

$$\theta_1 = \sum y_i = Y, \quad \theta_2 = \sum x_i = X, \quad \theta_3 = \sum y_ix_iR_i\pi_i,$$

$$\theta_4 = \sum x_i^2R_i\pi_i, \quad \underline{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4).$$

Then, $E_1(t_j) = \theta_j$, $j = 1,\ldots,4$,

$$t_g = t_1 + (X - t_2)\frac{t_3}{t_4} = f(\underline{t}), \text{ say; } f(\underline{\theta}) = Y.$$

$$\lambda_1 = \frac{\delta}{\delta t_1}f(\underline{t})|_{\underline{t}=\underline{\theta}} = 1, \quad \lambda_2 = \frac{\delta}{\delta t_2}f(\underline{t})|_{\underline{t}=\underline{\theta}} = -B_R, \quad \lambda_3 = \frac{\delta}{\delta t_3}f(\underline{t})|_{\underline{t}=\underline{\theta}} = 0,$$

$$\lambda_4 = \frac{\delta}{\delta t_4}f(\underline{t})|_{\underline{t}=\underline{\theta}} = 0.$$

Then, $t_g = f(\underline{t})$ may be approximated by

$$\begin{aligned} f(\underline{\theta}) + \lambda_1(t_1 - \theta_1) + \lambda_2(t_2 - \theta_2) &= Y + (t_1 - Y) - B_R(t_2 - \theta_2) \\ &= Y + \left(\sum E_i\frac{I_{si}}{\pi_i} - \sum E_i\right). \end{aligned}$$

Then, $M(t_g)$ is approximately equal to $M_3(t_g) = E_1[\sum E_iI_{si}/\pi_i - \sum E_i]^2$ which is the variance of $\sum E_iI_{si}/\pi_i$.

Now applying Corollary 2, we obtain analogously to (2.2),

$$M_3(t_g) = \sum{}'\sum{}'(\pi_i\pi_j - \pi_{ij})\left(\left(\frac{E_i}{\pi_i} - \frac{E_j}{\pi_j}\right)^2\right) + \sum\frac{E_i^2}{\pi_i}\beta_i.$$

So, our proposed estimators for $M_3(t_g)$ are

$$v_k(g) = \sum{}'\sum{}'(\pi_i\pi_j - \pi_{ij})\left(\left(\frac{a_{ki}e_i}{\pi_i} - \frac{a_{kj}e_j}{\pi_j}\right)^2\right)\frac{I_{sij}}{\pi_{ij}}$$
$$+\sum(a_{ki}e_i)^2\beta_i\frac{I_{si}}{\pi_i}, \quad k = 1, 2. \tag{4.4}$$

By way of justification of $v_k(g)$ we may observe that

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

may be approximated, through Taylor expansion, by E_i, $i \in U$ on approximating t_j by θ_j, $j = 2, 3, 4$. The rest follows as in Corollary 2 in Section 2.

Next, we consider the application of the above results when y_i is not directly ascertainable. This is for example, (1) when y relates to a sensitive characteristic like number of days of drunken driving, amount of tax evaded, etc., and (2) when i itself contains a large number, say, N_i of further sub-units or second-stage units of which only a sample may be observed or more generally multi-stage sampling seems feasible in a given context. We present relevant results in the next section. In these 2 cases (1) and (2), appropriate devices and designs are employed by an investigator to derive suitable estimators for y_i, $i \in s$ to be subsequently used in estimating $Y = \sum y_i$. A third possibility of a super-population model-based approach of dealing with the situation when y_i is subject to measurement and observational errors as treated by Fuller (1987) and Bolfarine and Zacks (1992) among others is not considered here.

5. Multi-stage sampling and randomized response surveys

Let E_2 denote the operator for taking expectation and V_2 that for variance with respect to either (a) randomized response (RR) technique or (b) sampling at later stages of sampling. Let 'independent' observations of r_i be available in either case along with sample-based observations v_i such that

$$\text{(i) } E_2(r_i) = y_i, \quad \text{(ii) } V_2(r_i) = V_i \quad \text{and} \quad \text{(iii) } E_2(v_i) = V_i, \quad i \in U. \tag{5.1}$$

We may further assume that

$$E_1E_2 = E_2E_1. \tag{5.2}$$

Raj (1968) and Chaudhuri (1987) considered this set-up in the contexts, respectively, of multi-stage and RR given in (5.1). Chaudhuri et al. (2000) use (5.2). Let E, V denote the over-all expectation, variance operators. Then, $E = E_1E_2 = E_2E_1$ and $V = E_1V_2 + V_1E_2$ which equals $E_2V_1 + V_2E_1$ by (5.2).

Writing any of the above-noted estimators or predictors for Y based on y_i, $i \in s$ as $t = t(s, \underline{Y})$, we may write $e = e(s, \underline{r})$, where $\underline{r} = (r_1, \ldots, r_i, \ldots, r_N)$, to denote the function t in which y_i is replaced by r_i for $i \in s$. We shall next write $R = \sum r_i$ and $\underline{V} = (v_1, \ldots, v_i, \ldots, v_N)$. Then, the MSE of e about Y will be taken as

$$M_1^*(e) = E(e - Y)^2 = E_1 E_2[(e - t) + (t - Y)]^2 = E_1 E_2 (e - t)^2 + M(t) \qquad (5.3)$$

and

$$M_2^*(e) = E(e - Y)^2 = E_2 E_1[(e - R) + (R - Y)]^2 = E_2 M(e) + \sum V_i. \qquad (5.4)$$

Remark IV. $M(e) = E_1(e - R)^2 = E_1(t - Y)^2|_{\underline{Y}=\underline{R}} = M(t)|_{\underline{Y}=\underline{R}}$, $e_b = t_b|_{\underline{Y}=\underline{R}}$, $M(e_b) = M(t_b)|_{\underline{Y}=\underline{R}} = -\sum'\sum' d_{ij} w_i w_j (r_i/w_i - r_j/w_j)^2 + \sum (r_i^2/w_i)\alpha_i$, from (2.2), $e_{\mathrm{H}} = t_{\mathrm{H}}|_{\underline{Y}=\underline{R}}$, $V(e_{\mathrm{H}}) = V(t_{\mathrm{H}})|_{\underline{Y}=\underline{R}} = \sum'\sum' (\pi_i\pi_j - \pi_{ij})(r_i/\pi_i - (r_j/\pi_j))^2 + \sum (r_i^2/\pi_i)\beta_i$, from (2.3), $e_g = t_g|_{\underline{Y}=\underline{R}}$, $M_3(e_g) = M_3(t_g)|_{\underline{Y}=\underline{R}} = \sum'\sum' (\pi_i\pi_j - \pi_{ij})((E_i(r)/\pi_i - E_j(r)/\pi_j)^2) + \sum (E_i^2(r)/\pi_i)\beta_i$, writing $E_i(r) = E_i|_{\underline{Y}=\underline{R}}$, from (4.3).

Next, let us write

$$e_i(r) = e_i|_{\underline{Y}=\underline{R}} = r_i - b_R(r) x_i, \quad \text{writing } b_R(r) = \frac{\sum r_i x_i R_i I_{si}}{\sum x_i^2 R_i I_{si}} = b_R|_{\underline{Y}=\underline{R}}. \qquad (5.5)$$

Our proposed estimators of MSEs are the following:

For $M_1^*(e_b)$, the proposed unbiased estimators are

$$m_1^*(e_b) = m_1(e_b) + \sum'\sum' d_{sij} I_{sij} w_i w_j \left(\frac{v_i}{w_i^2} + \frac{v_j}{w_i^2} \right)$$
$$- \sum \frac{v_i}{w_i} \alpha_i \frac{I_{si}}{\pi_i} + \sum v_i b_{si}^2 I_{si}, \qquad (5.6)$$

$$m_2^*(e_b) = m_2(e_b) + \frac{1}{p(s)} \left[\sum'\sum' c_{sij} w_i w_j \left(\frac{v_i}{w_i^2} + \frac{v_j}{w_j^2} \right) \right.$$
$$\left. - \sum \frac{v_i}{w_i} \alpha_i c_{si} \right] + \sum v_i b_{si}^2 I_{si} \qquad (5.7)$$

writing $m_k(e_b) = m_k(t_b)|_{\underline{Y}=\underline{R}}$, $k = 1, 2$, given in Corollary 1. It is easy to note that

$$E(m_k^*(e_b)) = M_1^*(e_b) = E_1 E_2 (e_b - t_b)^2 + M(t_b), \quad k = 1, 2.$$

For $M_2^*(e_b)$, the proposed unbiased estimators are

$$\hat{m}_k(e_b) = m_k(e_b) + \sum v_i b_{si} I_{si}, \quad k = 1, 2. \qquad (5.8)$$

It is easy to check that

$$E\hat{m}_k(e_b) = M_2^*(e_b), \quad k = 1, 2.$$

For $V_1^*(e_{\mathrm{H}}) = E_1 E_2[(e_{\mathrm{H}} - t_{\mathrm{H}})^2] + V(t_{\mathrm{H}}) = E_1[\sum V_i I_{si}/\pi_i^2] + V(t_{\mathrm{H}})$ our proposed unbiased estimators are

$$v_1^*(e_{\mathrm{H}}) = v_1(e_{\mathrm{H}}) - \sum'\sum' \frac{I_{sij}}{\pi_{ij}} (\pi_i\pi_j - \pi_{ij}) \left(\frac{v_i}{\pi_i^2} + \frac{v_j}{\pi_j^2} \right) + \sum v_i (1 - \beta_i) \frac{I_{si}}{\pi_i^2} \qquad (5.9)$$

and

$$v_2^*(e_{\mathrm{H}}) = v_2(e_{\mathrm{H}}) - \frac{1}{p(s)}\left[\sum{}'\sum{}' c_{sij}(\pi_i\pi_j - \pi_{ij})\left(\frac{v_i}{\pi_i^2} + \frac{v_j}{\pi_j^2}\right) + \sum c_{si}\frac{v_i}{\pi_i}\beta_i\right] + \sum v_i \frac{I_{si}}{\pi_i^2} \tag{5.10}$$

writing $v_k(e_{\mathrm{H}}) = v_k(t_{\mathrm{H}})|_{\underline{Y}=\underline{R}}$, $k = 1, 2$, as given in Corollary 3.

For $V_2^*(e_{\mathrm{H}}) = E_2E_1[(e_{\mathrm{H}} - R) + (R - Y)]^2 = E_2V_1(e_{\mathrm{H}}) + \sum V_i$ where $V_1(e_{\mathrm{H}}) = V_1(t_{\mathrm{H}})|_{\underline{Y}=\underline{R}} = V(t_{\mathrm{H}})|_{\underline{Y}=\underline{R}}$, our proposed unbiased estimators are

$$\hat{v}_k(e_{\mathrm{H}}) = v_k(e_{\mathrm{H}}) + \sum \frac{v_i}{\pi_i} I_{si}, \quad k = 1, 2. \tag{5.11}$$

Next, our proposed estimators for

$$M_2^*(e_g) = E_2M_3(e_g) + \sum V_i$$

are

$$m_k^*(e_g) = v_k(g)|_{\underline{Y}=\underline{R}} + \sum v_i \frac{I_{si}}{\pi_i}, \quad k = 1, 2, . \tag{5.12}$$

Here, $v_k(g)|_{\underline{Y}=\underline{R}}$ equals $v_k(g)|_{e_i=e_i(r)}$ as given in (4.4), $k = 1, 2$.

Remark V. Rao (1975), in the context of multi-stage sampling, illustrated a situation where (ii) should be replaced by (ii)′ $V_2(r_i) = V_{si}, v_i$ by v_{si} and (iii) by (iii)′ $E_2(v_{si}) = V_{si}$ when $i \in s$. Chaudhuri et al. (2000) noted that in this situation (5.2) is not applicable.

When (ii)′, (iii)′ are assumed and (5.2) is ruled out our proposals are the following: (I) Replace v_i by v_{si} in the formulae (5.6), (5.7),(5.9), (5.10); (II) Rule out the uses of (5.8), (5.11), (5.12).

Remark VI. For $M_1^*(e_g) = E_1E_2(e_g - t_g)^2 + M(t_g)$ we do not propose any estimator here because no elegant estimator seems to be available.

6. A numerical exercise on efficacy in estimation

Applying the modified sampling scheme of Seth (1966) we illustrate how the estimators given by (2.1), (2.4), (4.1) and (4.4) fare in yielding estimated coefficients of variation (CV) of estimates of totals. From SSW (1992, Appendix C, pp. 660–661) we take the first $N = 29$ municipalities as our illustrative population for which ‘size’-values are taken as size measures to apply the modified Seth (1966) scheme with $w = 0.4$ and the values of the total population in 1985 and 1995 are taken as, respectively, the values of y, the variable of interest and of x, the auxiliary correlated variable. We take $n = 9$. Table 1 presents, for 10 replicates of samples drawn as above

Table 1
Performance of $v_1(t_H)$ V Sv_{HT} and v_{kg} V Sm_{kg} in terms of the criteria a_1, a_2, b_k, c_k $(k=1,2)$ based on modified Seth (1966) scheme using SSW (1992) data

Sample number	Sample size realized	a_1	a_2	b_1	c_1	b_2	c_2
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	9		29.95		1.34		1.20
2	9		14.84		1.31		1.43
3	9		18.17	0.74	2.38	0.87	2.79
4	10		27.32		0.33		0.28
5	10		11.41		0.80		0.70
6	10		11.77	0.28	0.38	0.40	0.53
7	9		15.11		0.27		0.27
8	10		18.34		0.36		0.43
9	9		13.63		1.08		1.24
10	9		16.13	0.59	1.15	0.54	1.07

Table 2
Performance of $v_1(t_H)$ V Sv_{HT} and v_{kg} V Sm_{kg} in terms of the criteria a_1, a_2, b_k, c_k $(k=1,2)$ based on modified Seth (1966) scheme using Indian census 1991 data

Sample number	Sample size realized	a_1	a_2	b_1	c_1	b_2	c_2
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	6		20.36		7.86		7.77
2	7		17.42	0.00	8.50	0.00	7.88
3	7		12.37		12.15		10.15
4	7		18.34	6.73	6.52	9.01	8.70
5	7		9.56	3.08	3.47	4.99	5.62
6	7		18.23	6.33	10.31	5.32	8.52
7	7		27.95	1.93	10.84	2.09	11.74
8	7		12.73	10.53	10.05	11.16	10.65

from this population, the values of

$$a_1 = 100\frac{\sqrt{v_{\mathrm{HT}}}}{t_{\mathrm{H}}}, \qquad a_2 = 100\frac{\sqrt{v_1(t_{\mathrm{H}})}}{t_{\mathrm{H}}},$$

$$b_k = \frac{100\sqrt{m_{kg}}}{t_g}, \quad k=1,2, \qquad c_k = 100\frac{\sqrt{v_{kg}}}{t_g}, \quad k=1,2.$$

Note: Absence of an entry in each table below signifies 'negative' value of v_{HT} or m_{kg}. R_i is throughout taken as $(1-\pi_i)/\pi_i x_i$, $i \in U$.

We apply the same method taking $n=6$ and $w=0.4$ as before to draw samples from 23 villages in a particular district for which the household size is taken as the size measure, y as the area in hectare and x as the total population size, the source for each being the Indian population census, 1991. For 8 replicates of samples the values of a_1, a_2, b_k, c_k are presented in Table 2.

Table 3
Performance of $v_1(t_{\mathrm{H}})\ V\ Sv_{\mathrm{HT}}$ and $v_{kg}\ V\ Sm_{kg}$ via the criteria a_1, a_2, b_k, c_k $(k=1,2)$ based on PPS circular systematic samples repeated twice using data from SSW (1992)

Sample number	Realized sample size	a_1	a_2	b_1	c_1	b_2	c_2
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	8	41.61	164.24		4.68		10.38
2	9	31.04	107.24		2.85		5.12
3	10	22.19	84.80		1.54		3.19
4	10	29.09	101.60	0.26	1.73	0.57	3.82
5	10	27.77	94.01		2.39		4.42
6	7	37.07	124.40		1.43		3.73
7	10	30.82	106.58	0.09	1.49	0.16	2.76
8	10	32.56	149.02		3.11		3.36

Finally, we illustrate instead of the rather artificial sampling scheme above a realistic one which in fact is usually applied in Indian annual national sample surveys covering many socio-economic issues. This is the circular systematic sampling (CSS) with probability proportional to size (PPS) in a pre-assigned number of draws, but the entire draw is independently' repeated twice. The draw is repeated because for many pairs (i,j) the inclusion probabilities π_{ij}'s turn out to be zero in case of a 'single' draw of the sample. For the CSSPPS sampling repeated twice, each sample being of size n, the realized number of distinct units $v(s)$ varies between n and $2n$, the inclusion probabilities, say ψ_i, of i are determined in terms of the normed size measures p_i's and the inclusion probabilities ψ_{ij}'s of (i,j)'s, say, turn out positive. So, for this sampling scheme v_{HT} and $v_1(t_{\mathrm{H}})$ are competitors and so are m_{kg} vis a vis v_{kg} $(k=1,2)$. Using the same 29 values of size measures, y and x as in Table 1 and taking $n=5$ for each CSSPSS repeated twice, the similar exercise as in Tables 1 and 2 is presented for 8 replicated samples in Table 3.

7. Comments on numerical findings and recommendations

For the modified Seth (1966) scheme v_{HT} is negative throughout justifying thoroughly the introduction of $v_1(t_{\mathrm{H}})$. For PPSCSS repeated twice however this is not the case and $v_1(t_{\mathrm{H}})$ leads to loss in efficiency.

For both the schemes, both m_{kg} $(k=1,2)$ turn out negative, justifying the proposal for v_{kg} $(k=1,2)$ as their competitors. However, when they turn out positive, they often yield higher efficiencies compared to v_{kg} $(k=1,2)$.

Incidentally, for PPS CSS repeated twice, (1) $(\psi_i\psi_j - \psi_{ij})$ have variable signs but (2) $\beta_i > 0$ for every i, while (3) our proposed MSE estimators turn out positive for each sample.

Our recommendation is that when employing t_g and t_{H} based on 'varying sample size sampling schemes' one should employ, respectively, v_{kg} as possible competitors against m_{kg} $(k=1,2)$ and $v_1(t_{\rm H})$ against $v_{\rm HT}$ irrespective of whether theorems like Theorems 2 and 3 apply for the sampling scheme employed or not.

References

Bolfarine, H., Zacks, S., 1992. Prediction Theory for Finite Populations. Springer, Berlin.

Brewer, K.R.W., 1963. A model of systematic sampling with unequal probabilities. Austral. J. Statist. 5, 5–13.

Cassel, C.M., Särndal, C.E., Wretman, J.H., 1976. Some results on generalized difference estimation and generalized regression estimation for finite populations. Biometrika 63, 615–620.

Chaudhuri, A., 1987. Randomized response surveys of finite populations: a unified approach with quantitative data. J. Statist. Plann. Inference 15, 157–165.

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

Fuller, W.A., 1987. Measurement Error Models. Wiley, New York.

Horvitz, D.G., Thompson, D.J., 1952. A generalization of sampling without replacement from a finite universe. J. Amer. Statist. Assoc. 47, 663–685.

Raj, D., 1968. Sampling Theory. Mc-Graw Hill, New York.

Rao, J.N.K., 1975. Unbiased variance estimation for multi-stage designs. Sankhyá C 37, 133–139.

Rao, J.N.K., 1979. On deriving mean square errors and other non-negative unbiased estimators in finite population sampling. J. Indian Statist. Assoc. 17, 125–136.

Särndal, C.E., Swensson, B., Wretman, J., 1992. Model assisted survey sampling. Springer-Verlag, New York.

Seth, G.R., 1966. On estimators of variance of estimate of population total in varying probabilities. J. Indian Soc. Agricultural Statist. 18 (2), 52–56.

Yates, F., Grundy, P.M., 1953. Selection without replacement from within strata with probability proportional to size. J. Roy. Statist. Soc. B 15, 253–261.