Chaudhuri, Arijit; Samaddar, Sonakhya

Article

Estimating the population mean using a complex sampling design dependent on an auxiliary variable

Statistics in Transition new series (SiTns)

Provided in Cooperation with:
Polish Statistical Association

Suggested Citation: Chaudhuri, Arijit; Samaddar, Sonakhya (2022) : Estimating the population mean using a complex sampling design dependent on an auxiliary variable, Statistics in Transition new series (SiTns), ISSN 2450-0291, Sciendo, Warsaw, Vol. 23, Iss. 1, pp. 39-54, https://doi.org/10.2478/stattrans-2022-0003

STATISTICS IN TRANSITION new series, March 2022
Vol. 23, No. 1, pp. 39–54, DOI 10.2478/stattrans-2022-0003
Received – 03.12.2019; accepted – 31.03.2021

sciendo

Estimating the population mean using a complex sampling design dependent on an auxiliary variable

Arijit Chaudhuri[1], **Sonakhya Samaddar**[2]

ABSTRACT

In surveying finite populations, the simplest strategy to estimate a population total without bias is to employ Simple Random Sampling (SRS) with replacement (SRSWR) and the expansion estimator based on it. Anything other than that including SRS Without Replacement (SRSWOR) and usage of the expansion estimator is a complex strategy. We examine here (1) if from a complex sample at hand a gain in efficiency may be unbiasedly estimated comparing the "rival population total-estimators" for the competing strategies and (2) how suitable model-expected variances of rival estimators compete in magnitude as examined numerically through simulations.

Key words: Des Raj and symmetrized Des Raj estimator and associated variance, Hansen-Hurwitz estimation and variance, Hartley-Ross, Horvitz-Thompson, Lahiri-Midzuno-Sen, Murthy, Rao-Hartley-Cochran procedures vis-a-vis SRSWOR and SRSWR.
AMS Subject classification: 62 DO5.

1. Introduction

Stratifed SRSWOR is supposed to outperform unstratified SRSWOR because the conventional unbiased estimator of the population mean in the former has a variance as a function of the 'Within Sum of Squares' contrasted with the latter involving the 'Total Sum of Squares' if the strata are well constructed and maybe, effectively controlled Between strata variability. Using the survey data from a stratified SRSWOR it is well known vide Cochran (1977) and JNK Rao (1961) how the gain in stratification may duly be estimated vis-a-vis unstratified SRSWOR.

It is our interest to extend this approach covering a few competitive pairs of strategies in each of which it is difficult to work out plausible variance formulae in closed form illustrated in Section 2 below.

Covering pairs of sampling strategies for estimating population totals when variance formulae are available for unbiased estimators, we intend to examine how more complicated complex strategies may be justified from the efficiency gaining point of view vis-a-vis SRSWR and SRSWOR as the basic procedures by postulating simplified regression models thereby working out their model-based expected values of the variances of rival unbiased estimators for the population total.

Details are given in the Section 3 below.

[1]Indian Statistical Institute, Kolkata, India. The corresponding author.
E-mail: arijitchaudhuri1@rediffmail.com.ORCID: https://orcid.org/0000-0002-4305-7686.
[2]Indian Statistical Institute, Kolkata, India. The corresponding author. E-mail: sonakhya003@gmail.com.
ORCID: https://orcid.org/0000-0002-9462-0520.

A comparative study by simulations is presented in Section 4. Comments are also stated there.

To our knowledge the literature covers no follow-up of JNK Rao's (1961) approach treating any other strategies. Our Section 2 below is a novel exercise removing this deficiency taking account of several worthy alternatives. Secondly, considering a simple special case of Fairfield Smith's (1938) popular super-population model and bringing several useful and popular sampling strategies under this umbrella we, as a novelty, study, by simulation, how the numerical model-expected design variances of unbiased estimators of finite population totals (or means) for complex and simple sampling strategies fare among each other.

2. Estimating Gain in Efficiency

2.1. (PPSWOR, Des Raj Estimator) strategy versus (SRSWOR, Expansion Estimator)

Suppose y is a variable of interest taking values y_i for the respective units i of a finite population $U = (1,\ldots,i,\ldots,N)$, with a total $Y = \sum_{i=1}^{N} y_i$.

Let positive values x_i of another positively correlated variable x be all known for the units i of U, with a total $X = \sum_{i=1}^{N} x_i$ and $p_i = \frac{x_i}{X}$ be the unit-wise normed size measures. $\overline{X}$, $\overline{Y}$ denote the population means of x and y.

Probability proportional to size measures x_i (PPS) without replacement (PPSWOR) sample selection method is implemented by selecting a number, say, n($\geq$2) units from U ordered as the $1^{st}, 2^{nd}, \ldots, n^{th}$, namely $i_1, i_2, \ldots, i_j, \ldots, i_n$ with respective probabilities

$$p_{i1}, \frac{p_{i2}}{1-p_{i1}}, \ldots, \frac{p_{ij}}{1-p_{i1}-\ldots-p_{ij-1}}$$

$j = 1,2,\ldots,n.$

Then, Des Raj's unbiased estimator for Y is

$$t_D = \frac{1}{n}(t_1 + t_2 + \ldots + t_n)$$

with

$$\begin{aligned} t_1 &= \frac{y_{i1}}{p_{i1}}, \\ t_2 &= y_{i1} + \frac{y_{i2}}{p_{i2}}(1-p_{i1}), \ldots, \\ t_j &= y_{i1} + y_{i2} + \ldots + \frac{y_{ij}}{p_{ij}}(1-p_{i1}-p_{i2}-\ldots-p_{ij-1}), j = 1,2,\ldots,n. \end{aligned}$$

The formula for the exact variance of t_D is given by Roychoudhury (1957). But its closed form expression is pretty complicated. Nevertheless, an unbiased estimator for $V(t_D)$

is given by Des Raj(1956) as

$$v(t_D) = \frac{1}{2n^2(n-1)} \sum_{j=1}^{n} \sum_{k=1,k\neq j}^{n} (t_j - t_k)^2$$

which is pretty simple in form.

Suppose a PPSWOR sample chosen as above is at hand as $s = (i_1, i_2, \ldots, i_n)$ along with the values $y_{i1}, y_{i2}, \ldots, y_{in}$.

Suppose we consider a comparable strategy composed of an SRSWOR sample s_{WOR} of size n and the expansion estimator based on it as

$$N\overline{y} = \frac{N}{n} \sum_{i \in s_{WOR}} y_i$$

with variance

$$V_{SWOR}(N\overline{y}) = \frac{(N-n)N^2}{Nn(N-1)} \sum_{i=1}^{N} (y_i - \overline{Y})^2$$

where $\overline{y}$ denotes the sample mean.

Then, an unbiased estimator for this is derived as follows: We have

$$V(t_D) = E(t_D^2) - Y^2$$

So an unbiased estimator for Y^2 is

$$\hat{Y^2} = t_D^2 - v(t_D) \qquad \ldots(2.1)$$

Also an unbiased estimator for $\sum_1^N y_i^2$ is $t_D(y^2)$, which is t_D as above with every y in t_D replaced by corresponding y^2. So, an unbiased estimator for $V(N\overline{y})$ is

$$v_1 = \left(\frac{1}{n} - \frac{1}{N}\right) \frac{N^2}{N-1} \left[t_D(y^2) - \frac{\hat{Y^2}}{N} \right] \qquad \ldots(2.2)$$

with $\hat{Y^2}$ as given in (2.1).

Then $G_1 = v_1 - v(t_D)$ unbiasedly estimates gain in efficiency of (PPSWOR,t_D) over (SRSWOR,$N\overline{y}$).

2.2. (PPSWOR, Symmetrized Des Raj Estimator) versus (SRSWOR, Expansion Estimator)

Given the ordered sample as in section (2.1) as $s = (i_1, i_2, \ldots, i_n)$ and Des Raj's estimator $t_D = t_D(s)$ based on this ordered s, let s^* be the set of all samples obtained by permuting the n units in s in all possible $n!$ ways and

$$p(s^*) = \sum_{s \to s^*} p(s),$$

writing $\sum_{s\to s^*}$ to denote the sum over all possible samples in the set s^*. Then

$$t^*_{SD} = t^*_{SD}(s') = \frac{\sum_{s\to s^*} p(s)t_D(s)}{\sum_{s\to s^*} p(s)} = t^*_{SD}(s),$$

say, for any member s' in set s^* is the 'Symmetrized Des Raj' estimator for Y. Also, it is well known, vide (Chaudhuri(2010), p19) that

$$V(t^*_{SD}(s)) = V(t_D(s)) - E(t_D - t^*_{SD})^2.$$

Hence, an unbiased estimator for $V(t^*_{SD})$ is

$$v(t^*_D) = v(t_D) - (t_D - t^*_{SD})^2.$$

If the survey data $(s', y_i | i \in s')$ are at hand, an unbiased estimate for $V_{SWOR}(N\bar{y})$ follows as

$$v_2 = \left(\frac{1}{n} - \frac{1}{N}\right) \frac{N^2}{(N-1)} \left[t^*_{SD}(y^2) - \frac{(\hat{Y}^2)'}{N}\right] \qquad \ldots(2.3)$$

Writing $t^*_{SD}(y^2)$ as $t^*_{SD}(s')$ with each y_i in $t^*_{SD}(s')$ replaced by y_i^2 and

$$(\hat{Y}^2)' = (t^*_{SD}(s'))^2 - v_2(t^*_{SD}) \qquad \ldots(2.4)$$

2.3. (Lahiri-Midzuno-Sen sampling with Ratio Estimator) versus (SRSWOR, Expansion Estimator)

Lahiri-Midzuno-Sen's (1951, 1952, 1953) or LMS sample is selected by choosing on the first draw from U a unit i with selection probability p_i followed by an SRSWOR in (n-1) draws from the remaining (N-1) units excluding the first chosen unit i from U.

Then, $t_R = X\frac{\bar{y}}{\bar{x}}$ is the exact unbiased ratio estimator for Y based on such a sample, with $\bar{x}$ denoting the sample mean of x.

Vide Chaudhuri(2010) an exactly unbiased estimator of variance of t_R is

$$v(t_R) = \sum^{N} \sum_{i<j=1}^{N} a_{ij} \frac{I_{sij}}{\sum_{i\in s} p_i} \left(\frac{N-1}{n-1} - \frac{1}{\sum_{i\in s} p_i}\right);$$

here s is the LMS sample of size n, and

$$I_{sij} = \begin{cases} 1 & i,j \in s \\ 0 & otherwise \end{cases}$$

and

$$a_{ij} = p_i p_j \left(\frac{y_i}{p_i} - \frac{y_j}{p_j}\right)^2;$$

also $\pi_i =$ the inclusion probability of i in an LMS sample is given by

$$\pi_i = \frac{N-n}{N-1} p_i + \frac{n-1}{N-1}$$

and

$$\pi_{ij} = \frac{N-n}{(N-1)(N-2)}(p_i + p_j) + \frac{(n-2)(n-1)}{(N-1(N-2))}$$

is the inclusion probability of i and j in an LMS sample of size n. So, an unbiased estimator of $V(N\overline{y})$ from the LMS sample is

$$\hat{V}_{SWOR}(N\overline{y}) = v_3 = \left(\frac{1}{n} - \frac{1}{N}\right) \frac{N^2}{(N-1)} \left[\sum_{i\in s} \frac{y_i^2}{\pi_i} - \frac{1}{N}(t_R^2 - v(t_R))\right]$$

because

$$V(t_R) = E(t_R^2) - Y^2$$

and Y^2 is unbiasedly estimated by $t_R^2 - v(t_R)$. So, $v_3 - v(t_R)$ unbiasedly estimates the gain in efficiency of (LMS,t_R) over (SRSWOR,$N\overline{y}$)

2.4. (SRSWOR,Hartley-Ross estimator) versus (SRSWOR, Expansion estimator)

Based on an SRSWOR s of size n an unbiased estimator for Y given by Hartley and Ross (1954) is

$$\hat{Y}_{HR} = N\left[\overline{r} + \left(\frac{N-1}{N}\right)\left(\frac{n}{n-1}\right)\frac{1}{\overline{\overline{X}}}(\overline{y} - \overline{rx})\right] = N\left[\overline{r} + c(\overline{y} - \overline{rx})\right],$$

say, writing $\overline{r} = \frac{1}{n}\sum_{i\in s}\frac{y_i}{xi}$ and $\overline{x}$, $\overline{y}$ are sample means of x and y and $\overline{\overline{X}}$ is the population mean of x.

An unbiased estimator for $V(\hat{Y}_{HR})$ is given by

$$v(\hat{Y}_{HR}) = (\hat{Y}_{HR})^2 - \left[\frac{N}{n}\sum_{i\in s} y_i^2 + \frac{N(N-1)}{n(n-1)}\sum\sum_{i\neq j\in s} y_i y_j\right]$$

because for SRSWOR $\pi_i = \frac{n}{N}\ \forall i$ and $\pi_{ij} = \frac{n(n-1)}{N(N-1)}\ \forall i \neq j$

An unbiased estimator for $V(N\overline{y})$ from an SRSWOR s of size n is

$$v_4 = \left(\frac{1}{n} - \frac{1}{N}\right)\frac{N^2}{(n-1)}\sum_{i\in s}(y_i - \overline{y})^2.$$

So $v_4 - v(\hat{Y}_{HR})$ tells us how much we may gain in efficiency on using $\hat{Y}_{HR}$ rather than $N\overline{y}$.

In Section 3 below we consider situations when for complex surveys variances of unbiased estimators for Y have manageably elegant forms.

3. How under a simple model expected variances fare relative to each other

Model

We assume that

$$y_i = \beta x_i + \varepsilon_i,\ i \in U = (1,2,3,\ldots,N)$$

Here β is an arbitrary unknown constant which determines y's dependence on x's. $x_i's$ are auxilliary variables which are known for all population units. $\varepsilon_i's$ are error terms with zero mean and some common variance τ^2, which is also not fixed.
Every expectation that we have taken in the upcoming Sections 3.1 to 3.9 are based on the above mentioned model.

This model is a simple special case of the well-known popular Fairfield Smith's (1938) super-population model under which the model-variance of ε_i is $\tau^2 x_i^\gamma$ for i=1,2,...,N. In the literature most strategies are treated utilizing this model and the literature on comparison among model expected variances of design- unbiased estimators of finite population totals (or means) is rather vast. But in this paper we may draw attention to the following few, namely the text by Sarndal, Swensson and Wretman (1992) and a few papers in peer-reviewed journals namely by JNK Rao and Bayless, D.L. (1969), JNK Rao and Bayless, D.L. (1970), TJ Rao (1967) and Chaudhuri and Arnab (1979). The last-mentioned paper, Chaudhuri and Arnab (1979), is worthy of attention because in it, expressing Model- (Fairfield Smith's)- expected variances of ratio estimator based on LMS scheme by E_1, that of Rao-Hartley-Cochran estimator by E_2 and that of Horvitz-Thompson estimator based on an IPPS sample by E_3, it is shown that
(i) $E_1 < E_2 < E_3$ if $\gamma < 1$,
(ii) $E_1 > E_2 > E_3$ if $\gamma > 1$ and
(iii) $E_1 = E_2 = E_3$ if $\gamma = 1$.

3.1. Strategy 1: (SRSWR, Expansion Estimator)

For SRSWR in n draws from population of size N the expansion estimator $N\bar{y}$ is unbiased for $Y = \sum_{i=1}^{N} y_i$ with variance

$$V(N\bar{y}) = \frac{N^2}{n}\sigma^2 = \frac{N}{n}\sum_{i}^{N}(y_i - \overline{Y})^2; \quad \sigma^2 = \frac{1}{N}\sum_{1}^{N}(y_i - \overline{Y})^2.$$

Under a model its expected value is

$$\mathscr{E}(V(N\bar{y})) = \frac{N}{n}\mathscr{E}\left[\sum_{i=1}^{N}(y_i - \overline{Y})^2\right]$$

$\mathscr{E}$ denotes generically a model-based expectation operator.

Then

$$\mathscr{E}(V(N\bar{y})) = \frac{N(N-1)}{n}\left[\tau^2 + \beta^2 S_{xx}\right] = (srswr)$$

where

$$S_{xx} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{X})^2.$$

3.2. Strategy 2: (SRSWOR,$N\overline{y}$)

We have in this case

$$V(N\overline{y}) = \frac{N^2(N-n)}{Nn(N-1)}\sum_{i=1}^{N}(y_i - \overline{Y})^2.$$

And thus,

$$\mathscr{E}(V(N\overline{y})) = \frac{N(N-n)}{n}(\tau^2 + \beta^2 S_{xx}) = (srswor).$$

3.3. Strategy 3: (PPSWR, Hansen-Hurwitz Estimator t_{HH})

The Hansen Hurwitz estimator (1943) is given by

$$t_{HH} = \frac{1}{n}\sum_{r=1}^{n}\frac{y_r}{p_r},$$

with
y_r=y-value for the unit chosen on r-th draw
p_r=probability of the unit being chosen on r-th draw

$$V(t_{HH}) = \frac{1}{n}\left[\sum_{1}^{N}\frac{y_i^2}{p_i} - Y^2\right]$$

with

$$\mathscr{E}(V(t_{HH})) = \frac{\tau^2}{n}\left(N\overline{X}\sum_{1}^{N}\frac{1}{x_i} - N\right) = (ppswr).$$

3.4. Strategy 4: (PPSWR, Horvitz-Thompson Estimator

For PPSWR sampling in n draws the inclusion-probabilities are

$$\pi_i = 1 - (1-p_i)^n$$

$$\pi_{ij} = 1 - (1-p_i)^n - (1-p_j)^n + (1-p_i-p_j)^n.$$

Following Chaudhuri and Pal (2003) the Horvitz & Thompson's (1952) estimator (HTE), $t_{HT} = \sum_{i\in s}\frac{y_i}{\pi_i}$ based on a PPSWR sample s in n draws has the variance

$$V(t_{HT})_{PPS} = \sum_{i<j}^{N}\sum^{N}(\pi_i\pi_j - \pi_{ij})\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 + \sum_{1}^{N}\frac{y_i^2}{\pi_i}\alpha_i$$

with $\alpha_i = 1 + \frac{1}{\pi_i}\sum_{j\neq i}\pi_{ij} - \sum_1^N \pi_i$. Then

$$\mathscr{E}(V(t_{HT})_{PPS}) = \beta^2\left[\sum_1^N \frac{x_i^2}{\pi_i} + \sum\sum_{i\neq j} x_i x_j \frac{\pi_{ij}}{\pi_i\pi_j} - X^2\right] + \tau^2\left(\sum_1^N \frac{1}{\pi_i} - N\right) + \beta^2\sum_1^N \alpha_i\frac{x_i^2}{\pi_i} + \tau^2\sum_1^N \frac{\alpha_i}{\pi_i} = (ppswrht).$$

3.5. Strategy 5: (SRSWR,HTE)

For SRSWR in n draws the inclusion-probabilities are

$$\pi_i = 1 - (\frac{N-1}{N})^n$$

$$\pi_{ij} = 1 - 2(\frac{N-1}{N})^n + (\frac{N-2}{N})^n.$$

For the HTE based on SRSWR in n draws the variance is

$$V(t_{HT})_{SRS} = \sum_1^N\sum_{i<j}^N(\pi_i\pi_j - \pi_{ij})\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 + \sum_1^N \frac{y_i^2}{\pi_i}\alpha_i$$

with

$$\mathscr{E}(V(t_{HT})_{SRS}) = \beta^2\left[\sum_1^N \frac{x_i^2}{\pi_i} + \sum\sum_{i\neq j} x_i x_j \frac{\pi_{ij}}{\pi_i\pi_j} - X^2\right] + \tau^2\left(\sum_1^N \frac{1}{\pi_i} - N\right) + \beta^2\sum_1^N \alpha_i\frac{x_i^2}{\pi_i} + \tau^2\sum_1^N \frac{\alpha_i}{\pi_i} = (srswrht).$$

3.6. Strategy 6: (SRSWR, N times the mean of the sampled distinct units only

From Chaudhuri (2010, pp. 35-36)we know that the sample mean of the distinct units in a sample s chosen by SRSWR in n-draws is unbiased for the population mean $\overline{Y}$ and the expansion estimator given by N multiplied by this mean $\bar{y}_d$, say, $\hat{Y}_d = N\bar{y}_d$ has the variance

$$V(\hat{Y}_d) = N^2\left[\frac{1}{N}\sum_{i=1}^N\left(\frac{j}{N}\right)^{n-1} - \frac{1}{N}\right]S^2$$

writing $S^2 = \frac{1}{(N-1)}\sum_{i=1}^N(y_i - \overline{Y})^2$ and so

$$\mathscr{E}(V(\hat{Y}_d)) = N^2\left[\frac{1}{N}\sum_1^N\left(\frac{j}{N}\right)^{n-1} - \frac{1}{N}\right](\tau^2 + \beta^2 S_{xx}) = (srswrd)$$

writing $S_{xx} = \frac{1}{N-1}\sum_1^N (x_i - \overline{X})^2$.

3.7. Strategy 7: (Rao-Hartley-Cochran Sampling, Rao-Hartley-Cochran Estimator)

A sample of size n by Rao-Hartley-Cochran (RHC(1962)) scheme is taken by choosing from the population first a sample of N_1 units by SRSWOR, then a sample of size N_2 from the remaining $(N - N_1)$ units of the population and successively and similarly, finally an SRSWOR of size N_n keeping $N_1 + N_2 + \ldots + N_n = N$ and for the sake of efficiency taking $N_i = \left[\frac{N}{n}\right]$ for $i = 1, 2, \ldots, k$ and the last (n-k) of these N_iFLs as $\left[\frac{N}{n}\right] + 1$ with the restriction $N_1 + N_2 + \ldots + N_n = N$. Such a choice is uniquely possible. For the parts of the population so constructed, the values of p_i are noted and

$$Q_i = p_{i1} + \ldots + p_{iN_i}$$

for the i-th pair or group is noted.

Then, writing $\sum_n$ as the sum over these n pairs or groups and $\sum_n \sum_n$ as the sum over the distinct pairs of these groups follows the RHC's unbiased estimator for Y as

$$t_{RHC} = \sum_n \frac{y_{ij}}{p_{ij}} Q_i$$

on taking independently across these n groups just one unit say labelled ij from the i-th group denoting the associated y value as y_{ij}. Then, it follows that

$$V(t_{RHC}) = \frac{\sum_n N_i^2 - N}{N(N-1)} \sum_n \sum_n p_i p_j \left(\frac{y_i}{p_i} - \frac{y_j}{p_j}\right)^2.$$

Also,

$$\mathscr{E}(V(t_{RHC})) = \frac{\left(\sum_n N_i^2 - N\right)}{N-1} \tau^2 \left(\overline{X} \sum_1^N \frac{1}{x_i} - 1\right) = (rhc).$$

3.8. Strategy 8: (An Inclusion Probability Proportional to size (IPPS or πPS) sampling, Horvitz-Thompson Estimator

The Horvitz Thompson estimator based on a sample s of size (of distinct unit) n is $t_{HT} = \sum_{i \in s} \frac{y_i}{\pi_i}$ and $\pi_i = np_i$, $i \in Population$

$$V(t_{HT}) = \sum \sum_{i<j} (\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2$$

and

$$\mathscr{E}(V(t_{HT})) = \frac{N\tau^2}{n} \left(\overline{X} \sum_1^N \frac{1}{x_i} - n\right) = (ippsht).$$

3.9. Strategy 9: Lahiri-Midzuno-Sen (LMS) sampling Scheme, Horvitz-Thompson Estimator (HTE)

For the sample s of size n for the LMS scheme, the HT estimator is $t_{HT} = \sum_{i\in s} \frac{y_i}{\pi_i}$ with the variance

$$V(t_{HT}) = \sum\sum_{i<j} (\pi_i\pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2$$

with

$$\pi_i = \frac{(N-n)}{(N-1)} p_i + \frac{(n-1)}{(N-1)}$$

and

$$\pi_{ij} = \frac{(N-n)(n-1)}{(N-1)(N-2)}(p_i + p_j) + \frac{(n-1)(n-2)}{(N-1)(N-2).}$$

It follows that

$$\mathscr{E}(V(t_{HT})_{LMS}) \;=\; \beta^2 \left[\sum_1^N \frac{x_i^2}{\pi_i} + \sum\sum_{i\neq j} x_i x_j \frac{\pi_{ij}}{\pi_i\pi_j} - X^2\right] \;+\; \tau^2 \left(\sum_1^N \frac{1}{\pi_i} - N\right) \;=\; (lmsht).$$

4. A numerical study by Simulation

4.1. Simple Model yielding x,y values

Model: Let $y_i = \beta x_i + \varepsilon_i \qquad i \in U = (1,2\ldots,N)$ with β an arbitrarily chosen positive constant; x_iFLs are independently generated from distribution function

$$F(x) = 1 - e^{-\frac{1}{10}x}, \qquad x > 0.$$

The choice of x was from such a distribution mainly because we wanted to use positive values of explanatory variables keeping in mind the application of such a model in real life. The mean of 10 was taken to choose values with considerably moderate values.

ε_iFLs are independently randomly generated from the Normal distribution $N(0,1)$ for $i = 1,2,\ldots,N$.

Also, we take $\beta = 2.3$, 1.6, and 3.6 and $N = 23$. Using these different values of β we generated three sets of values which shall be treated as population. The generated values with $\beta = 2.3$ are reported in Table 1.

Table 1: Table of Population values

Sl.No	x	y	Sl.No	x	y
1	7.55	18.15	13	12.38	28.41
2	11.82	27.25	14	44.24	100.37
3	1.46	1.36	15	10.55	23.84
4	1.40	3.84	16	10.35	23.42
5	4.36	9.97	17	18.76	43.09
6	28.94	66.42	18	6.55	16.16
7	12.30	26.81	19	3.37	8.51
8	5.40	11.93	20	5.88	13.37
9	9.57	22.41	21	23.65	54.13
10	1.47	4.74	22	6.42	15.46
11	13.91	31.88	23	2.94	7.32
12	7.62	17.91			

Throughout this paper we shall use only these three sets of (x,y)-values whenever needed for illustrations as the finite population values of x and y. For the material presented in sections 2.1-2.4 we intend to use the values generated as given above to illustrate the realized magnitudes of estimated gains in efficiencies of pairs of competing strategies. For this, from the population of N=23 sets of (x,y)-values samples of size n=7 are chosen by appropriately defined procedures. The findings are presented in Section 4.2 below in specified tables.

In order to present numerical illustrations for the materials covered in Sections 3.1 through 3.9 we use the x-values of three generated populations mentioned above for the population of size N=23 but β values are differently taken and ε_i's are supposed to have a constant model variance τ^2 which are variably taken for illustration. On every occasion a sample of size n is illustrated with the value 7 but y-values are accordingly supposed to be generated yielding the specified model-expected variances illustrated in tables in Section 4.3 below.

4.2. Numerical study of material in Sections 2.1-2.4

For the purpose of presentation of materials covered in Sections 2.1-2.4, we have chosen 10 separate and independent samples each of size 7 from different populations generated as mentioned above. We worked on deriving the estimated variances and tabulated them below side by side.

Table 2: Estimated variances given in Sections 2.1-2.4 for samples of size n=7 from the (x,y)

$v(t_D)$	v_1	$v(t^*_{SD})$	v_2	$v(t_R)$	v_3	$v(\hat{Y}_{HR})$	v_4
i	ii	iii	iv	v	vi	vii	viii
1806.07	4908.88	1718.08	2418.30	325.73	15196.67	5219.82	25418.88
1.76	3291.07	1.31	395.99	31.21	7845.60	24429.38	18072.16
7.57	12883.59	7.54	1097.78	509.82	2496.31	10429.85	45723.08
3.58	19646.11	2.72	1854.53	35.15	3269.31	5805.09	6658,41
39.77	6404.84	39.74	2163.54	6.49	411.98	22844.20	4321.82
315.56	4776.37	315.49	4321.30	39.93	2237.48	11985.97	5815.61
9.85	6366.44	9.18	3045.95	18.25	1476.29	860.27	2640.17
324.63	3685.89	324.49	5052.97	117.28	4474.86	4470.89	7616.78
3152.35	1301.71	3147.20	8209.40	8.76	344.86	35946.81	6572.70
15.29	5249.95	11.14	2583.30	239.06	21144.43	2014.46	42327.04

Comments

From the values of the estimated variances we may say that (i) (PPSWOR, t_D) is substantially more gainful in efficiency over (ii) (SRSWOR, Expansion Estimator) both of which are much inferior to (iii) (PPSWOR, Symmetrized Des Raj Estimator). For the sample size n=7 we had to obtain $7! = 5040$ Des Raj estimates. But with powerful statistical software it did not cost us much time.

Compared to (vi) (SRSWOR,$N\bar{y}$) the strategy (v) (LMS, Ratio Estimator) is enormously more gainful as it should be because a size variable is employed. Compared to (viii) (SRSWOR,$N\bar{y}$), (vii) (SRSWOR, Hartley-Ross Estimator) is also more gainful, but presumably because an auxiliary size-measure is employed.

4.3. Numerical study of material in Sections 3.1-3.9

Using the values of x_i from three different populations and variously choosing β and τ^2 explained in Section 4.1, we present below in Table 3, 4 and 5 the values of the model expected variances of various unbiased estimators for a finite population total of a variable y of interest based on samples taken according to various schemes.

Table 3: Ten values for each of the model-expected variances for first population

	SRSWR	*SRSWOR*	*PPSWR*	*PPSWR*	*SRSWR*	SRSWR	*RHC*	*IPPS*	*LMS*
	$N\bar{y}$	$N\bar{y}$	t_{HH}	*HTE*	*HTE*	$N\bar{y}_d$	*RHC*	*HTE*	*HTE*
(β,τ^2)	$(srswr)$	$(srswor)$	$(ppswr)$	$(ppswrht)$	$(srswrht)$	$(srswrd)$	(rhc)	$(ippsht)$	$(lmsht)$
(0.1,5)	434.55	316.03	815.49	1051.10	448.74	388.17	609.19	716.92	298.45
(0.1,10)	795.98	578.89	1630.98	2073.27	805.91	711.03	1218.4	1433.83	564.18
(0.2,2)	437.05	317.85	326.20	524.57	509.11	390.41	243.68	286.77	237.13
(0.2,10)	1015.33	738.42	1630.98	2160.04	1080.6	906.97	1218.4	1433.83	662.31
(0.5,5)	2189.4	1592.29	815.49	1745.31	2646.19	1955.74	609.19	716.92	1083.45
(1.5,2)	16596.34	12070.07	326.16	6917.12	20744.01	14825.12	243.68	286.77	7465.73
(2.5,5)	46060.8	33498.76	815.49	19100.64	57582.57	41145.02	609.19	716.92	20708.61
(2.5,10)	46422.23	33671.62	1630.98	20122.81	57939.74	41467.87	1218.34	1433.83	20974.35
(2.5,25)	47506.51	34550.19	4077.44	23189.33	59011.27	42436.44	3046	3584.59	21771.56
(3,2)	65951.66	47964.85	326.19	26441.87	82547.43	58913.06	243.68	286.767	29544.03

Table 4: Ten values for each of the model-expected variances for second population

	SRSWR	*SRSWOR*	*PPSWR*	*PPSWR*	*SRSWR*	SRSWR	*RHC*	*IPPS*	*LMS*
	$N\bar{y}$	$N\bar{y}$	t_{HH}	*HTE*	*HTE*	$N\bar{y}_d$	*RHC*	*HTE*	*HTE*
(β,τ^2)	$(srswr)$	$(srswor)$	$(ppswr)$	$(ppswrht)$	$(srswrht)$	$(srswrd)$	(rhc)	$(ippsht)$	$(lmsht)$
(0.1,5)	421.76	306.74	827.45	1045.82	436.46	376.75	618.14	728.88	298.43
(0.1,10)	783.19	569.59	1654.91	2063.89	793.64	699.61	1236.28	1457.76	557.62
(0.2,2)	385.92	280.67	330.98	518.18	460.03	344.73	247.26	291.55	215.02
(0.2,10)	964.20	701.24	1654.91	2147.11	1031.51	861.30	1236.27	1457.76	639.33
(0.5,5)	1869.84	1359.89	827.45	1711.52	2339.40	1670.29	618.14	728.88	946.09
(1.5,2)	13720.30	9978.4	330.98	6648.23	17982.88	12256.02	247.26	291.55	6234.14
(2.5,5)	38071.79	27688.57	827.45	18354.19	49912.75	34008.62	618.14	728.88	17287.6
(2.5,10)	38433.21	27951.43	1654.91	19372.27	50269.93	34331.48	1236.27	1457.76	17552.79
(2.5,25)	39517.50	28740	4137.27	22426.51	51341.45	35300.05	3090.69	3644.41	18348.37
(3,2)	54447.49	39598.17	330.98	25371.23	71502.90	48636.65	247.26	291.55	24618.34

Table 5: Ten values for each of the model-expected variances for third population

	SRSWR	*SRSWOR*	*PPSWR*	*PPSWR*	*SRSWR*	SRSWR	*RHC*	*IPPS*	*LMS*
	$N\bar{y}$	$N\bar{y}$	t_{HH}	*HTE*	*HTE*	$N\bar{y}_d$	*RHC*	*HTE*	*HTE*
(β,τ^2)	$(srswr)$	$(srswor)$	$(ppswr)$	$(ppswrht)$	$(srswrht)$	$(srswrd)$	(rhc)	$(ippsht)$	$(lmsht)$
(0.1,5)	433.82	315.50	2637.29	3664.84	441.64	387.52	1970.15	2538.72	298.03
(0.1,10)	795.25	578.36	5274.58	7307.63	798.82	710.38	3940.30	5077.44	565.04
(0.2,2)	434.14	315.74	1054.91	1545.29	480.75	387.81	788.06	1015.49	230.89
(0.2,10)	1012.43	736.31	5274.59	7373.77	1052.24	904.38	3940.31	5077.44	658.11
(0.5,5)	2171.27	1579.11	2637.29	4193.92	2468.95	1939.54	1970.15	2538.72	1042.59
(1.5,2)	16433.17	11951.39	1054.92	6417.29	19148.91	14679.36	788.06	1015.49	7087.09
(2.5,5)	45607.53	33169.11	2637.29	17421.07	53151.72	40740.12	1970.15	2538.72	19656.68
(2.5,10)	45968.96	33431.97	5274.58	21063.87	53508.89	41062.98	3940.30	5077.44	19923.69
(2.5,25)	47053.24	34220.54	13186.47	31992.25	54580.42	42031.55	9850.76	12693.61	20724.72
(3,2)	65298.96	47490.15	1054.92	21297.84	76167.01	58330.01	788.06	1015.49	28027.93

Comments

Among the strategies (srswr) (SRSWR, $N\bar{y}$), (srswor) (SRSWOR, $N\bar{y}$) and (srswrd) (SRSWR,$N\bar{y}_d$), as anticipated (srswor) fares best and (srswrd) in between the other two. Moreover (srswrht) (SRSWR, t_{HT}) fares worse than all these three. Between the (ppswr) (PPSWR, t_{HH}) and the (rhc) RHC strategy, the latter performs better for every (β, τ^2) pair as it should. Interestingly, (ppswrht) (PPSWR, t_{HT}) fares worse than both.

(ippsht) (IPPS, t_{HT}) strategy fares competitively against (lmsht) (LMS, t_{HT}), the latter poorer as τ^2 is taken higher. Interestingly, they are found competitive against all four strategies with equal probability sampling making no use of size measures.

Most interestingly, the (rhc) RHC strategy fares by far the best among all the strategies under our competition here for almost all choices of our (β, τ^2)'s.

5. Discussions

The present work has clearly two distinct aspects. One of them is extending the well-known approach of comparing the classical stratified sampling strategy with the corresponding unstratified one in terms of the two variance estimates from the stratified sample at hand on identifying the Des Raj estimator combined with PPSWOR, the symmetrized Des Raj estimator with PPSWOR, the Hartley-Ross estimator based on SRSWOR and the mean of values of distinct sample units in SRSWR versus the over-all sample mean in SRSWR and the expansion estimator in SRSWOR as the situations when easy variance estimator formulae are easy to derive. The other being on observing that the first aspect can be impressively clarified through simulated illustrations, following it through an appeal to simulated illustrations also to compare model-based expectations of exact design variances of pairs of well-known unbiased estimators of population totals citing several complex and simple sampling strategies.

6. Conclusions

(i) In the first case in Section 5, as expected, the complex strategies numerically outperform the respective simpler ones. Thus, it is vindicated that to go for the complex alternatives is lucrative rather than to remain complacent about the simpler alternatives.
(ii) In the second case of Section 5, for various alternative pairs relative performances of model-expected variances are comparatively demonstrated in Section 4.3. The Rao-Hartley-Cochran (1962) strategy for our illustrated numerical situation is demonstrated to fare as the most effective strategy. But from this it cannot be claimed that one should always go for this in practice. Respective performances are well illustrated in Section 4.3 of course. We cannot make general conclusions beyond our illustrated example of course.

Acknowledgement

The authors gratefully acknowledge two referee's recommendations that led to this improved version over an earlier one.

References

Bayless, D. L. and Rao, J. N. K., (1970). *An empirical study of stabilities of estimators and variance estimators in unequal probability sampling (n=3or 4)*, Jour. Amer. Stat.Assoc. 65, pp. 1645–1667.

Chaudhuri, A., (2010). *Essentials of survey sampling*, Prentice Hall of India, New Delhi.

Chaudhuri, A. and Arnab, R., (1979). *On the relative efficiencies of sampling strategies under a super-population model*, Sankhya, Ser. C. 41, pp. 40–43.

Cochran, W., G., (1977). *Sampling Techniques.* John Wiley and Sons. New York.

Des Raj, (1956). *Some estimators in sampling with varying probabilities without replacement*, Jour. Amer. Stat. Assocn. 51, pp. 269–284.

Hansen, M. H. and Hurwitz, W. N., (1943). *On the theory of sampling from finite populations*, Ann. Math. Stat, 14, pp. 333–362.

Hartley, H. O. and Ross, A., (1954). *Unbiased ratio estimators*, Nature, 174, pp. 270–271.

Horvitz, D. G. and Thompson, D. J., (1952). *A generalization of sampling without replacement from a finite universe*, Jour. Amer. Stat. Assoc., 47, 663–685.

Lahiri, D. B., (1951). *A method of sample selection providing unbiased ratio estimates*, Bull. Int. Stat. Inst. 33(2), pp. 133–140.

Midzuno, H., (1952). *An Outline of the theory of sampling systems*, Annals. Inst. Stat. Math, 1, pp. 149–156.

Murthy, M. N., (1957). *Ordered and unordered estimators in sampling without replacement*, Sankhy$\bar{a}$, 18, pp. 379–390.

Rao, J. N. K., (1961). *On the estimate of variance in unequal probability sampling*, Annals. Inst. Stat. Math, 13, pp. 57–60.

Rao, J.N.K. and Bayless, D. L., (1969). *An empirical study of the stabilities of estimators and variance estimators in unequal probability of two units per stratum*, Jour. Amer. Stat. Assoc. 64, pp. 540–549.

Rao, J. N. K., Hartley, H. O., Cochran, N.G., (1962). *On a simple procedure of unequal probability sampling without replacement*, Jour. Roy. Stat. Soc. B. 24, pp. 482–491.

Rao, T. J., (1967). *On the choice of a strategy for a ratio method of estimation*, Jour. Roy.Stat.Soc. B. 29, pp. 392–397.

Roychoudhury, D. K., (1957). *Unbiased sampling design using information provided by linear function of auxiliary variate*, Chapter 5, thesis for Associateship of Indian Statistical Institute, Kolkata.

Sarndal. C. E., Swensson, B and Wretman, J., (1992). *Model Assisted Survey Sampling*, Springer Verlag, Heidelberg.

Sen, A. R., (1953). *On the estimator of the variance in sampling with varying probabilities*, J. Ind. Soc. Agri. Stat. 5(2), pp. 119–127.

Smith, H. F., (1938). *An empirical law describing heterogeneity in the yields of agricultural crops*, Jour. Agri. Sci. 28, pp. 1–23.