VARIANCE ESTIMATION IN TWO-PHASE SAMPLING

Raghunath Arnab 1,4§, Sarjinder Singh 2, Antonio Arcos 3 and Kago Kebotsamang 4

1 University of KwaZulu-Natal, RSA
Email: arnabr@mopipi.ub.bw
2 Texas A&M University, Kingsville, USA
3 Universidad de Granada, Spain
4 University of Botswana, Botswana
§ Corresponding Author

ABSTRACT

The problem of variance estimation for two-phase sampling was considered by Särndal et al. (1992). Hidiroglou et al. (2009) studied the properties of the proposed estimators theoretically while Haziza et al. (2001) empirically studied the performances of the proposed estimators. In this paper a few alternative variance estimators have been proposed. The proposed variance estimators are found to be more efficient than the existing estimators when the study variable is proportional to the auxiliary variable. Empirical investigations also reveal that the proposed variance estimators fare better in respect of efficiency and preserving non-negativity property.

KEYWORDS

Two-phase sampling, variance estimation, varying probability sampling.

1. INTRODUCTION

The two-phase sampling is used when the variable under study \( y \) is well related to the auxiliary variable \( x \) but the information of the auxiliary variable is not available before the survey and the cost of collecting data for the auxiliary variable \( x \) is much cheaper than the study variable. Hence, in two-phase sampling, a relative large sample is initially selected from the population and only information of the auxiliary variable \( x \) is collected. Then in phase two another sample is selected either from the sample selected in the first-phase or from the entire population independently and information of the study variable \( y \) is obtained. The two-phase sampling was introduced by Neyman (1938). Recently, Särndal et al. (1992) proposed the double expansion estimator for estimating the population total when the sample selected on both phases by arbitrary sampling design. They also proposed Horvitz-Thompson (HT, 1952) and Yates-Grundy (YG, 1953) type variance estimators of the proposed double expansion (DE) estimator. Hidiroglou et al. (2009) pointed out that the proposed HT type variance estimator can be very unstable and can even take negative values. If the variance estimator becomes negative it cannot be used for measuring the magnitude of the sampling error, interval estimation and testing of hypotheses among others. Hidiroglou
Variance Estimation in Two-Phase Sampling

et al. (2009) also studied the properties of the HT and YG type variance estimators theoretically in details. Haziza et al. (2011) considered DE-estimator and Hajek (1971) estimator of the population total for the two-phase sampling. They also proposed HT and YG type variance estimators for Hajek (HA) estimator. Haziza et al. (2011) studied the properties of their variance estimators in terms of stability and non-negativity theoretically as well as through extensive simulation studies. They found that HT-type variance estimators are more prone to take negative values than YG type variance estimators.

In this present paper we have proposed few alternative variance estimators which are modifications of the variance estimators investigated by Haziza et al. (2011). The proposed estimators are found better in terms of efficiency and are less prone to take negative values.

1.1 HT and YG Variance estimators in two-phase sampling

Consider a finite population \( U = \{1, \ldots, i, \ldots, N \} \) of \( N \) identifiable units. In the first phase, a sample \( s_1 \) of size \( n_1 \) units is selected using a sampling design \( p_1 \) and information of the auxiliary variable \( x \) is observed. Let \( \pi_{ii} \) and \( \pi_{ij} (>0) \) be the inclusion probabilities for the \( i \)th and \( i \)th and \( j \)th unit \((i \neq j)\) respectively. In the second phase, a sub-sample \( s_2 \) of size \( n_2 \) is selected from \( s_1 \) by following a sampling design \( p_{2|s_1} \) with inclusion probabilities \( \pi_{2i|s_1} \) and \( \pi_{2ij|s_1} (>0) \) for the \( i \)th, \( i \)th and \( j \)th unit \((i \neq j)\) respectively.

Our objective is to estimate the finite population total \( Y = \sum_{i \in U} y_i \) using a two-phase sample \( s = (s_1, s_2) \) selected by some suitable sampling designs.

Särndal et al. (1992) proposed the following Double expansion (DE) estimator for the population total \( Y \) :

\[
\hat{Y}_{DE} = \sum_{i \in s_2} \frac{y_i}{\pi_{i|s_1}}
\]

where \( \pi_{i|s_1} = \pi_{ii} \cdot \pi_{2i|s_1} \).

Haziza et al. (2011) proposed HA type estimator of \( Y \) as follows:

\[
\hat{Y}_{HA} = \frac{N}{\hat{N}_{DE}} \hat{Y}_{DE}
\]

where \( \hat{N}_{DE} = \sum_{i \in s_2} \frac{1}{\pi_{i|s_1}} \).

The estimator \( \hat{Y}_{DE} \) is design-unbiased for \( Y \) while \( \hat{N}_{DE} \) is asymptotically design unbiased.
1.2 Double Expansion Estimator \( \hat{Y}_{DE} \)

The variance of \( \hat{Y}_{DE} \) is given by

\[
V_{DE} = V \left( \hat{Y}_{DE} \right) = V \left( E \left( \hat{Y}_{DE} \mid s_1 \right) \right) + E \left( V \left( \hat{Y}_{DE} \mid s_1 \right) \right)
\]

\[
= V \left( \sum_{i \in s_1} \frac{y_i}{\pi_{i1}} \right) + E \left( V \left( \sum_{i \in s_2} \frac{y_i}{\pi_{i1}^s} \mid s_1 \right) \right)
\]

\[
= Q_1(y) + Q_2(y) \tag{1.3}
\]

where \( Q_1(y) = V \left( \sum_{i \in s_1} \frac{y_i}{\pi_{i1}} \right) \) and \( Q_2(y) = E \left( V \left( \hat{Y}_{DE} \mid s_1 \right) \right) \).

Now

\[
Q_1(y) = \sum_{i \in U} \left( \frac{1}{\pi_{i1}} - 1 \right) y_i^2 + \sum_{i \neq j \in U} \left( \frac{\pi_{ij}}{\pi_{i1} \pi_{1j}} - 1 \right) y_i y_j
\]

\[
= Q_{1h}^U(y)
\]

For fixed sample size design \( Q_1(y) \) can be written as

\[
Q_1(y) = \frac{1}{2} \sum_{i \neq j \in U} \left( \pi_{i1} \pi_{1j} - \pi_{ij} \right) \left( \frac{y_i}{\pi_{i1}} - \frac{y_j}{\pi_{1j}} \right)^2
\]

\[
= Q_{1y}^U(y)
\]

From the expressions \( Q_{1h}^U(y) \) and \( Q_{1y}^U(y) \), HT-type and YG-type unbiased estimators of \( Q_1(y) \) are obtained as follows:

\[
\hat{Q}_{1h}^U(y) = \sum_{i \in s_2} \frac{(1 - \pi_{i1})}{\pi_{2ij1}} \left( \frac{y_i}{\pi_{i1}} \right)^2 - \sum_{i \neq j \in s_2} \frac{\Delta_{ij}}{\pi_{i1j1} \pi_{i1} \pi_{1j}} \frac{y_i}{\pi_{i1}} \frac{y_j}{\pi_{1j}}
\]

and

\[
\hat{Q}_{1y}^U(y) = \frac{1}{2} \sum_{i \neq j \in s_2} \frac{\Delta_{ij}}{\pi_{i1j1} \pi_{i1} \pi_{1j}} \left( \frac{y_i}{\pi_{i1}} - \frac{y_j}{\pi_{1j}} \right)^2 \tag{1.4}
\]

where \( \Delta_{ij} = \pi_{i1} \pi_{1j} - \pi_{ij} \) and \( \pi_{i1j1} = \pi_{2ij1} \pi_{i1} \).

Similarly, writing
\[
V \left( \hat{Y}_{DE} \mid s_1 \right) = \sum_{i \in s_1} \left( \frac{1}{\pi_{2j|i_1}} - 1 \right) \left( \frac{y_i}{\pi_{i|i_1}} \right)^2 + \sum_{i \neq j} \sum_{j \in s_1} \left( \frac{\pi_{2ij|s_1}}{\pi_{2j|i_1} \pi_{2j|s_1}} - 1 \right) \frac{y_i}{\pi_{i|i_1}} \frac{y_j}{\pi_{1|j_1}} = Q_{2h}^{ht}(y)
\]
\[
= \frac{1}{2} \sum_{i \neq j} \sum_{j \in s_1} \left( \pi_{2ij|s_1} - \pi_{2j|s_1} \right) \left( \frac{y_i}{\pi_{i|i_1} \pi_{2j|s_1}} - \frac{y_j}{\pi_{1|j_1} \pi_{2j|s_1}} \right)^2 = Q_{2r}^{yr}(y),
\]
two unbiased estimators HT-type and YG type of \( Q_2(y) \) are obtained as
\[
\hat{Q}_2^{ht}(y) = \sum_{i \in s_2} \left( 1 - \pi_{2j|i_1} \right) \left( \frac{y_i}{\pi_{i|i_1} \pi_{2j|s_1}} \right)^2 - \sum_{i \neq j} \sum_{j \in s_2} \Delta_{2ij|s_1} \frac{y_i}{\pi_{i|i_1} \pi_{2j|s_1}} \frac{y_j}{\pi_{1|j_1} \pi_{2j|s_1}}
\]
and
\[
\hat{Q}_2^{yg} = \frac{1}{2} \sum_{i \neq j} \sum_{j \in s_2} \Delta_{2ij|s_1} \left( \frac{y_i}{\pi_{i|i_1} \pi_{2j|s_1}} - \frac{y_j}{\pi_{1|j_1} \pi_{2j|s_1}} \right)^2 = \hat{Q}_1^{yr}(y) + \hat{Q}_2^{yr}(y)
\]
where \( \pi_{i|i_1} = \pi_{i|s_1} \) and \( \Delta_{2ij|s_1} = \left( \pi_{2ij|s_1} - \pi_{2j|s_1} \pi_{2i|s_1} \right) \).

From the expressions (1.3) to (1.6), unbiased estimators of the variance of the total \( \hat{Y}_{DE} \) are obtained as follows:
\[
\hat{V}_{DE}^{ht} = \hat{Q}_1^{ht}(y) + \hat{Q}_2^{ht}(y)
\]
and
\[
\hat{V}_{DE}^{yg} = \hat{Q}_1^{yr}(y) + \hat{Q}_2^{yg}(y)
\]

The HT-type and YG-type variance estimators \( \hat{V}_{DE}^{ht} \) and \( \hat{V}_{DE}^{yg} \) were proposed respectively by Särndal et al. (1992) and Hidiroglou et al. (2009) respectively.

It is well known that both the estimators \( \hat{V}_{DE}^{ht} \) and \( \hat{V}_{DE}^{yg} \) can take negative values. The estimator \( \hat{V}_{DE}^{yg} \) is always non-negative for fixed sample size designs for which \( \Delta_{1ij} \) and \( \Delta_{2ij|s_1} \) are both nonnegative. We will call such sampling designs as \( P^* \). However for such \( P^* \) designs, \( \hat{V}_{DE}^{ht} \) can take negative values.

Hidiroglou et al. (2009) and Haziza et al. (2011) studied in detail the properties of stability and non-negativity of the variance estimators \( \hat{V}_{DE}^{ht} \) and \( \hat{V}_{DE}^{yg} \). Haziza et al. (2011)
pointed out that the estimators $\hat{Q}_1^{yg}(y) = 0$ if $CV(y/\pi_1) = 0$ and $\hat{Q}_2^{yg}(y) = 0$ if $CV(y/\pi^*) = 0$ equal to zero. On the other hand, the components $\hat{Q}_1^{ht}(y)$ or $\hat{Q}_2^{ht}(y)$ becomes negative to satisfy unbiasedness condition of $\hat{V}_{DE}^{ht}$. So, in this situation $\hat{V}_{DE}^{yg}$ can take negative values. Hence if $CV(y/\pi_1)$ and/or $CV(y/\pi^*)$ is small, $\hat{V}_{DE}^{yg}$ is more stable than $\hat{V}_{DE}^{ht}$. However if $CV(y/\pi_1)$ and/or $CV(y/\pi^*)$ is large, we cannot compare stability of the variance estimators.

1.3 Hajek Estimator $\hat{Y}_{HA}$

The Hajek estimator $\hat{Y}_{HA}$ is a ratio estimator and hence to find the variance of $\hat{Y}_{HA}$, Haziza et al. (2011) considered the following approximation:

$$\hat{Y}_{HA} - Y = \frac{N}{\sum_{i \in s_2} \pi_{i|s_1}} \left( \sum_{i \in s_2} \frac{y_i}{\pi_i} - \bar{y} \sum_{i \in s_2} \frac{1}{\pi_i} \right) \approx \left( \sum_{i \in s_2} \frac{y_i - \bar{Y}}{\pi_i} \right)$$

$$= \sum_{i \in s_2} \frac{z_i}{\pi_i}$$

where $z_i = y_i - \bar{Y}$ and $\bar{Y} = Y/N$.

Under the approximation (1.9), the variance of $\hat{Y}_{HA}$ was obtained as

$$V_{HA} = V\left(\hat{Y}_{HA}\right) = V\left(\sum_{i \in s_1} \frac{z_i}{\pi_i} \right) + E\left( V\left( \sum_{i \in s_2} \frac{z_i}{\pi_i} \right) \right)$$

$$= Q_1(z) + Q_2(z)$$

(1.10)

where $Q_1(z)$ and $Q_2(z)$ are obtained by substituting $y_i = z_i$ in the expressions of $Q_1(y)$ and $Q_2(y)$ respectively.

Haziza et al. (2011) proposed the following variance estimators analogous to $\hat{V}_{DE}^{ht}$ and $\hat{V}_{DE}^{yg}$:

$$\hat{V}_{HA}^{ht} = \hat{Q}_1^{ht}(\hat{z}) + \hat{Q}_2^{ht}(\hat{z})$$

(1.11)

and

$$\hat{V}_{HA}^{yg} = \hat{Q}_1^{yg}(\hat{z}) + \hat{Q}_2^{yg}(\hat{z})$$

(1.12)

where $\hat{Q}_i(\hat{z})$ was obtained from $\hat{Q}_i(y)$ by replacing $y_i$ with
\[
\hat{y}_t = \frac{1}{N} \left( \sum_{i \in s_2} \frac{y_i}{\pi_i} \right) = \frac{1}{\sum_{i \in s_2} \pi_i} \left( \sum_{i \in s_2} \frac{y_i}{\pi_i} \right) \text{ for } t = 1, 2 \text{ and } j = ht, yg
\]

(1.13)

Here also \( \hat{V}_{HA}^{yg} \) estimator is always nonnegative for \( P^* \) sampling designs. Haziza et al. (2011) studied the properties of the estimators \( \hat{V}_{HA}^{ht} \) and \( \hat{V}_{HA}^{yg} \) but no straightforward conclusion relating to non-negativity and stability properties of the estimators were obtained.

2. PROPOSED VARIANCE ESTIMATORS

Haziza et al. (2011) use the auxiliary variable \( x \) collected in the phase one as a measure of size for the selection of sample in phase two. Since the variable \( x \) is expected to be highly related to the study variable \( y \), we can improve the above mentioned variance estimators by extracting relationship between \( x \) and \( y \) at the stages of estimation also.

The proposed improved variance estimators for \( \hat{Y}_{DE} \) are as follows:

\[
\hat{V}_{DE}^{ht} = \hat{Q}_{1}^{ht}(y, x) + \hat{Q}_{2}^{ht}(y, x)
\]

and

\[
\hat{V}_{DE}^{yg} = \hat{Q}_{1}^{yg}(y, x) + \hat{Q}_{2}^{yg}(y, x)
\]

where

\[
\hat{Q}_{1}^{ht}(y, x) = \frac{\hat{Q}_{1}^{ht}(y)}{\hat{Q}_{1}^{ht}(x)} E\left(\hat{Q}_{1}^{ht}(x) \mid s_1\right) \quad \text{and} \quad \hat{Q}_{2}^{ht}(y, x) = \frac{\hat{Q}_{2}^{ht}(y)}{\hat{Q}_{2}^{ht}(x)} E\left(\hat{Q}_{2}^{ht}(x) \mid s_1\right)
\]

\[
\hat{Q}_{1}^{yg}(y, x) = \frac{\hat{Q}_{1}^{yg}(y)}{\hat{Q}_{1}^{yg}(x)} E\left(\hat{Q}_{1}^{yg}(x) \mid s_1\right) \quad \text{and} \quad \hat{Q}_{2}^{yg}(y, x) = \frac{\hat{Q}_{2}^{yg}(y)}{\hat{Q}_{2}^{yg}(x)} E\left(\hat{Q}_{2}^{yg}(x) \mid s_1\right)
\]

is obtained from \( \hat{Q}_{r}^{t}(y) \) by replacing \( y_i = x_i \) for \( r = 1, 2 \) and \( t = ht, yg \);

\[
E\left(\hat{Q}_{r}^{ht}(x) \mid s_1\right) = E \left[ \left( \sum_{i \in s_2} \frac{1 - \pi_{1i}}{\pi_{2i}} \left( \frac{x_i}{\pi_{1i}} \right)^2 - \sum_{i \neq s_2} \frac{\Delta_{ij}}{\pi_{i}} \frac{x_i}{\pi_{1i}} \frac{x_j}{\pi_{1j}} \right) \right]
\]

\[
= \left[ \sum_{i \in s_1} \left( 1 - \pi_{1i} \right) \left( \frac{x_i}{\pi_{1i}} \right)^2 - \sum_{i \neq s_1} \frac{\Delta_{ij}}{\pi_{i}} \frac{x_i}{\pi_{1i}} \frac{x_j}{\pi_{1j}} \right]
\]
Similarly, the proposed two alternative variance estimators for \( \hat{Y}_{HA} \) are given as follows:

\[
\hat{V}_{HA}^{ht} = \hat{Q}_1^{ht}(\hat{z}, \hat{x}) + \hat{Q}_2^{ht}(\hat{z}, \hat{x})
\]

and

\[
\hat{V}_{HA}^{yg} = \hat{Q}_1^{yg}(\hat{z}, \hat{x}) + \hat{Q}_2^{yg}(\hat{z}, \hat{x})
\]

where

\[
\hat{Q}_1^{ht}(\hat{z}, \hat{x}) = \frac{\hat{Q}_1^{ht}(\hat{z})}{\hat{Q}_1^{ht}(\hat{x})} E\left(\hat{Q}_1^{ht}(\hat{x}) \mid s_1\right), \quad \hat{Q}_2^{ht}(\hat{z}, \hat{x}) = \frac{\hat{Q}_2^{ht}(\hat{z})}{\hat{Q}_2^{ht}(\hat{x})} E\left(\hat{Q}_2^{ht}(\hat{x}) \mid s_1\right),
\]

\[
\hat{Q}_1^{yg}(\hat{z}, \hat{x}) = \frac{\hat{Q}_1^{yg}(\hat{z})}{\hat{Q}_1^{yg}(\hat{x})} E\left(\hat{Q}_1^{yg}(\hat{x}) \mid s_1\right) \quad \text{and} \quad \hat{Q}_2^{yg}(\hat{z}, \hat{x}) = \frac{\hat{Q}_2^{yg}(\hat{z})}{\hat{Q}_2^{yg}(\hat{x})} E\left(\hat{Q}_2^{yg}(\hat{x}) \mid s_1\right) ; \hat{\gamma}_r(\hat{x})
\]

is obtained from \( \hat{Q}_r(y) \) by replacing

\[
y_j = \hat{y}_j = \frac{N}{\sum_{i \in s_2} \frac{1}{\pi_i^{s_1}}} \left( \frac{\sum_{i \in s_2} \frac{x_i}{\pi_i^{s_1}}}{\sum_{i \in s_1} \frac{1}{\pi_i^{s_1}}} \right) \quad \text{for} \ r = 1, 2 \ \text{and} \ t = ht, yg
\]

\[ (2.5) \]
Variance Estimation in Two-Phase Sampling

\[ E\left(\hat{Q}_1^{ht}(\hat{x}) \mid s_1\right) = \left[ \sum_{i \in s_1} (1 - \pi_{1i}) \left( \frac{\hat{x}_i}{\pi_{1i}} \right)^2 - \sum_{i \neq j, j \in s_1} \Delta_{1ij} \frac{\hat{x}_i}{\pi_{1i}} \frac{\hat{x}_j}{\pi_{1j}} \right], \]

\[ E\left[ \hat{Q}_2^{ht}(\hat{x}) \mid s_1 \right] = \left[ \sum_{i \in s_1} (1 - \pi_{2i}) \pi_{2i} \left( \frac{\hat{x}_i}{\pi_{2i}} \right)^2 - \sum_{i \neq j, j \in s_1} \Delta_{2ij} \frac{\hat{x}_i}{\pi_{2i}} \frac{\hat{x}_j}{\pi_{2j}} \right], \]

\[ E\left(\hat{Q}_1^{yg}(\hat{x}) \mid s_1\right) = \frac{1}{2} \left[ \sum_{i \neq j, j \in s_1} \sum_{j \in s_1} \Delta_{1ij} \left( \frac{\hat{x}_i}{\pi_{1i}} - \frac{\hat{x}_j}{\pi_{1j}} \right)^2 \right] \]

and

\[ E\left(\hat{Q}_2^{yg}(\hat{x}) \mid s_1\right) = \frac{1}{2} \left[ \sum_{i \neq j, j \in s_1} \sum_{j \in s_1} \Delta_{2ij} \left( \frac{\hat{x}_i}{\pi_{2i}} - \frac{\hat{x}_j}{\pi_{2j}} \right)^2 \right] \]

where

\[ \hat{x}_i = \frac{N}{\sum_{i \in s_1} \pi_{1i}} \left( x_i - \frac{1}{\sum_{i \in s_1} \pi_{1i}} \right). \]  \hspace{1cm} (2.6)

**Example 2.1.**

Consider the situation where the first phase sample \( s_1 \) is selected by SRSWOR method and the second-phase sample \( s_2 \) by arbitrary sampling designs. In this case, we have

\[ \pi_{1i} = n_1 / N, \quad \pi_{1ij} = \frac{n_1(n_1 - 1)}{N(N - 1)}, \quad \hat{Y}_{DE} = \frac{N}{n_1} \sum_{i \in s_2} \frac{y_i}{\pi_{2i}}, \]

\[ \hat{Q}_1^{ht}(y) = \frac{N - n_1}{n_1} \left[ \sum_{i \in s_2} \frac{y_i^2}{\pi_{2i}} - \frac{1}{N - 1} \sum_{i \neq j} \frac{y_i y_j}{\pi_{2i} \pi_{2j}} \right], \]

\[ \hat{Q}_2^{ht}(y) = \left( \frac{n_1}{N} \right)^2 \left[ \sum_{i \in s_2} \left(1 - \pi_{2i}\right) \left( \frac{y_i}{\pi_{2i}} \right)^2 - \sum_{i \neq j} \frac{\Delta_{2ij} y_i y_j}{\pi_{2i} \pi_{2j}} \right], \]

\[ \hat{Q}_1^{yg}(y) = \frac{1}{2n_1^2 (n_1 - 1)} \sum_{i \neq j} \frac{(y_i - y_j)^2}{\pi_{2i} \pi_{2j}}. \]
\[ \hat{Q}^y_{2g}(y) = \frac{1}{2} \left( \frac{n_1}{N} \right)^2 \sum_{i \neq j} \sum_{s_{ij}} \Delta_{2ij/s} \left( \frac{y_i - y_j}{\pi_{2i/s} - \pi_{2j/s}} \right)^2 , \]

\[ E \left( \hat{Q}^h_1(x) \mid s_1 \right) = \frac{N(N-n_1)}{n_1(n_1-1)} \sum_{i \in s_1} \left( x_i - \bar{x}_{s_1} \right)^2, \quad \bar{x}_{s_1} = \sum_{i \in s_1} x_i / n_1 , \]

\[ E \left[ \hat{Q}^h_2(x) \mid s_1 \right] = \left( \frac{N}{n_1} \right)^2 \sum_{i \in s_1} \left( 1 - \pi_{2i/s} \right) \pi_{2i/s} \left( \frac{x_i}{\pi_{2i/s}} - \frac{x_j}{\pi_{2j/s}} \right)^2 - \sum_{i \neq j} \sum_{s_{ij}} \Delta_{2ij/s} \frac{x_i}{\pi_{2i/s}} \frac{x_j}{\pi_{2j/s}} . \]

\[ E \left( \hat{Q}^y_1(x) \mid s_1 \right) = \frac{N(N-n_1)}{n_1^2(n_1-1)} \sum_{i \neq j} \left( x_i - x_j \right)^2 , \]

\[ E \left( \hat{Q}^y_2(x) \right) = \frac{1}{2} \left( \frac{N}{n_1} \right)^2 \sum_{i \neq j} \sum_{s_{ij}} \Delta_{2ij/s} \left( \frac{x_i}{\pi_{2i/s}} - \frac{x_j}{\pi_{2j/s}} \right)^2 . \]

**Example 2.2**

Consider the situation where the first phase sample \( s_1 \) is selected by Lahiri-Midzuno-Sen (1951, 1952, 1953) sampling scheme and the 2nd phase sample \( s_2 \) is selected by arbitrary sampling scheme. In Lahiri-Midzuno-Sen (LMS) sampling scheme, on the first draw one unit is selected with normed size measure \( p_i \left( p_i > 0, \sum_i p_i = 1 \right) \) attached to the \( i \)th unit and the remaining \( n_1 - 1 \) are selected from those \( N - 1 \) units which were not selected in the first draw by simple random sampling without replacement (SRSWOR) method. The first and second order inclusion probabilities for the LMS sampling design are given by \( \pi_{1i} = \frac{N-n_1}{N-1} p_i + \frac{n_1-1}{N-1} \) and \( \pi_{1ij} = \left( p_i + p_j \right) \left( \frac{N-n_1(n_1-1)}{N-1} \right) + \left( \frac{n_1-1(n_1-2)}{N-1} \right) \) respectively. The expressions of \( \hat{Q}^\theta_{j} \) and \( E \left( \hat{Q}^\theta_{j}(x) \mid s_1 \right) \), \( j = 1, 2 \); \( \theta = ht, yg \) are highly complex and they are omitted here.

**2.1 Properties of the Proposed Variance Estimators**

Since the estimators \( \hat{Q}^ht_1(y, x) \), \( \hat{Q}^ht_2(y, x) \), \( \hat{Q}^{yg}_1(y, x) \) and \( \hat{Q}^{yg}_2(y, x) \) are ratio type estimators, the proposed variance estimator \( \hat{V}^{ht}_{DE} \) and \( \hat{V}^{yg}_{DE} \) are biased estimator for \( V_{DE} \). Similarly, \( \hat{V}^{ht}_{HA} \) and \( \hat{V}^{yg}_{HA} \) are biased estimator for \( V_{HA} \). The amount of bias is expected to be less if the sample size at the second phase is large. The variance estimators \( \hat{V}^{yg}_{DE} \) and \( \hat{V}^{yg}_{HA} \) are always nonnegative if \( P^* \) design is used at both the phases of sampling. In case
the study variable is approximately proportional to the auxiliary variable \( x \). \( \hat{Q}_1^{ht}(y) \) and \( \hat{Q}_1^{ht}(x) \) are expected to have the same sign and the sign of \( \hat{Q}_1^{ht}(y, x) \) will be determined by the sign of \( E(\hat{Q}_1^{ht}(x) \mid s_1) \). Since the quantity \( E(\hat{Q}_1^{ht}(x) \mid s_1) \) is based on the larger sample of size \( n_1 \), it should have higher concentration around its expected value \( Q(x) \) (which is positive) than \( \hat{Q}_1^{ht}(y) \). Hence \( E(\hat{Q}_1^{ht}(x) \mid s_1) \) is expected to take more positive values than \( \hat{Q}_1^{ht}(y) \) and \( \hat{Q}_1^{ht}(x) \). Thus the proposed estimators \( \hat{V}_{DE}^{ht} \) and \( \hat{V}_{HA}^{ht} \) are likely to take less negative values than \( \hat{V}_{DE}^{ht} \) and \( \hat{V}_{HA}^{ht} \).

As efficiency is concerned, we show that all the proposed variance estimators \( \hat{V}_{DE}^{ht}, \hat{V}_{DE}^{yg}, \hat{V}_{HA}^{ht} \) and \( \hat{V}_{HA}^{yg} \) are more efficient than those considered by Haziza et al. (2011) if the study variable \( y \) is proportional to the auxiliary variable \( x \).

**Theorem 2.1**

If the study variable \( y \) is proportional to the auxiliary variable \( x \), then

\[
Var(\hat{V}_j^\alpha) \leq Var(\hat{V}_j^\alpha); \quad j = DE, HA \text{ and } \alpha = ht, yg
\]

**Proof:**

Let

\[
y = \beta x, \quad \hat{Q}_1^{ht}(y, x) = \beta^2 E(\hat{Q}_1^{ht}(x) \mid s_1)
\]

\[
= E(\hat{Q}_1^{ht}(y) \mid s_1)
\]

Now

\[
Var(\hat{V}_{DE}^{ht}) = Var(\hat{Q}_1^{ht}(y) + \hat{Q}_2^{ht}(y))
\]

\[
= E\{Var(\hat{Q}_1^{ht}(y) + \hat{Q}_2^{ht}(y) \mid s_1)\} + Var\{E(\hat{Q}_1^{ht}(y) + \hat{Q}_2^{ht}(y) \mid s_1)\}
\]

\[
\geq Var\{E(\hat{Q}_1^{ht}(y) + \hat{Q}_2^{ht}(y) \mid s_1)\}
\]

\[
= Var\left(\hat{Q}_1^{ht}(y, x) + \hat{Q}_2^{ht}(y, x)\right)
\]

\[
= Var(\hat{V}_{DE}^{ht})
\]

Similarly we can show \( Var(\hat{V}_{DE}^{yg}) \leq Var(\hat{V}_{DE}^{ht}) \), \( Var(\hat{V}_{HA}^{ht}) \leq Var(\hat{V}_{HA}^{ht}) \) and \( Var(\hat{V}_{HA}^{yg}) \leq Var(\hat{V}_{HA}^{ht}) \).
3. SIMULATION STUDIES

In this section, we investigated the performances of the proposed variance estimators with the existing estimators investigated by Haziza et al. (2011). For the simulation studies, we generated 8 populations of which 4 were of size 100 and the remaining were of size 200 each. Each population consisted of two variables: the variable of interest $y$ and auxiliary variable $x$ as was considered by Haziza et al. (2011). The size variable $x$ was generated from a gamma distribution with the scale parameter $\gamma = 4$ and shape parameter $\theta = 25$. For given values of $x$, the values of $y$ were generated using the model $y_i = \alpha + 2x_i + \varepsilon_i$ where $\varepsilon_i$ was assumed to be normally distributed with mean zero and variance $\sigma^2 x_i^g$. The 8 populations were generated by choosing $\alpha = 0, 0.5, \sigma^2 = 1$ and $g = 0, 0.5, 0.5$. The CV of $x$ for each of the population of size 100 was 0.2023 while for the other populations of size 200 was 0.2031. Summary statistics for the generated populations, in terms of CV, are reported in Table 3.0.

At the first phase, we selected a sample $s_1$ of size $n_1 (=20, 30, 40)$ by SRSWOR method so that $\pi_{i1} = n_1 / N$ and $\pi_{ij} = n_i(n_i - 1) / \{N(N - 1)\}$ as was selected by Haziza et al. (2011). From the selected sample $s_1$, only the values of the auxiliary variable $x$ was observed. In the second phase, a sub-sample $s_2$ of size $n_2 (=10, 15, 20)$ was selected from $s_1$ by Sampford (1967) sampling scheme using $x$ as measure of size variable. The Sampford sampling scheme is as follows:

On the first draw, the $i$th unit is selected from $s_1$ with probability $p_i(1) = x_i / \sum_{i \in s_1} x_i$. Then the remaining $(n_2 - 1)$ units are drawn with replacement from the entire sample $s_1$ with probability proportional to $\lambda_i = p_i(1) / [1 - n_1 p_i(1)]$ i.e. the probability of selecting $i$th unit at $k$th draw is $p_i(k) = \lambda_i / \sum_{j \in s_1} \lambda_j$, $k = 2, \ldots, n_2$. The selected units are accepted as a sample if all the $n_2$ units happened to be distinct, otherwise the entire selection is discarded and this process is repeated until a set of $n_2$ distinct units is obtained. Sampford (1967) has shown that the inclusion probability for the selection of $i$th unit is $\pi_{2i|s_1} = n_2 p_i(1)$ and $\Delta_{2i|s_1} = \pi_{2i|s_1} \cdot \pi_{2j|s_1} - \pi_{2ij|s_1} \geq 0$. The expression for the second order inclusion probabilities is not simple. However, approximate expression of $\pi_{2i|s_1}$ correct to $O\left(\frac{n_1^{-4}}{}\right)$, derived by Asok and Sukhatme (1976) is given for $n_2 \geq 3$ as is follows:
\[
\pi_{2ij|s_1} = n_2(n_2 - 1)p_i^{(1)}p_j^{(1)} \left[ 1 + \left\{ p_i^{(1)} + p_j^{(1)} - \sum_{j \in s_1} p_j^{2(1)} \right\} \right. \\
\left. + \left\{ 2(p_i^{2(1)} + p_j^{2(1)}) - 2 \sum_{j \in s_1} p_j^{3(1)} - (n_2 - 2)p_i^{(1)}p_j^{(1)} \right\} \right.
\\
\left. + (n_2 - 3)(p_i^{(1)} + p_j^{(1)}) \sum_{j \in s_1} p_j^{2(1)} - (n_2 - 3) \left\{ \sum_{j \in s_1} p_j^{2(1)} \right\}^2 \right\} - (n_2 - 2)p_i^{(1)}p_j^{(1)} \\
\left. + (n_2 - 3)(p_i^{(1)} + p_j^{(1)}) \sum_{j \in s_1} p_j^{2(1)} - (n_2 - 3) \left\{ \sum_{j \in s_1} p_j^{2(1)} \right\}^2 \right]\right] 
\]

(3.1)

From the selected sample \( s = (s_1, s_2) \), the estimators \( \hat{Y}_{DE}, \hat{Y}_{HA} \) of the population total \( Y \), the variance estimators \( v_i = \hat{V}_{DE}^{ht}, v_2 = \hat{V}_{DE}^{yg}, v_1 = \hat{V}_{HA}^{ht}, v_2 = \hat{V}_{HA}^{yg} \) proposed by Haziza et al. (2011) and proposed in this article \( v_3 = \hat{V}_{DE}^{yg}, v_4 = \hat{V}_{DE}^{yg}, v_3 = \hat{V}_{HA}^{ht}, v_4 = \hat{V}_{HA}^{yg} \) are computed. The selection of sample \( (s_1, s_2) \) was repeated \( R (= 100,000) \) times. The values of the estimators \( \hat{Y}_{DE}, \hat{Y}_{HA}, v_j \) and \( v_j^* \) were computed at the \( r \)th iteration and are denoted respectively by \( \hat{Y}_{DE}(r), \hat{Y}_{HA}(r), v_j(r) \) and \( v_j^*(r) \); \( j = 1, \ldots, 4 \). The variances of \( \hat{Y}_{DE} \) and \( \hat{Y}_{HA} \) are respectively given by:

\[
V_{DE} = \frac{1}{R} \sum_{r=1}^{R} \left( \hat{Y}_{DE}(r) - \frac{1}{R} \sum_{r=1}^{R} \hat{Y}_{DE}(r) \right)^2 \quad \text{and} \quad V_{HA} = \frac{1}{R} \sum_{r=1}^{R} \left( \hat{Y}_{HA}(r) - \frac{1}{R} \sum_{r=1}^{R} \hat{Y}_{HA}(r) \right)^2 
\]

(3.2)

The percentage relative biases and mean-square errors of the variance estimators \( v_j \) and \( v_j^* \) are respectively obtained as follows:

\[
\text{Bias of} \quad v_j = B(j) = \frac{1}{R} \frac{\sum_{r=1}^{R} v_j(r) - V_{DE}}{V_{DE}} \times 100; \quad \text{mean square error of} \quad v_j = M(j) = \frac{1}{R} \sum_{r=1}^{R} \left( v_j(r) - V_{DE} \right)^2 
\]

\[
\text{Bias of} \quad v_j^* = B^*(j) = \frac{1}{R} \frac{\sum_{r=1}^{R} v_j^*(r) - V_{HA}}{V_{HA}} \times 100; \quad \text{mean square error of} \quad v_j^* = M^*(j) = \frac{1}{R} \sum_{r=1}^{R} \left( v_j^*(r) - V_{HA} \right)^2 
\]
A problem of simulation studies with variance estimators is that the parameter to be estimated (the variance) is unknown. Then, one should select an estimate to measure the stability of the variance estimator. Note that we used the Monte Carlo Variance as a measure of stability of a variance estimator rather than the Monte Carlo MSE.

It was found that $v_1(v_1^*)$ is much inferior (with respect to bias and mean square error) to $v_j(v_j^*)$ for $j = 2, 3, 4$. Therefore, the relative efficiencies of $v_j$ and $v_j^*$ were computed with respect to $v_2$ and $v_2^*$ respectively. They are respectively denoted by

$$E_j = \left(\frac{v_2}{v_j}\right) \times 100 \quad \text{and} \quad E_j^* = \left(\frac{v_2^*}{v_j^*}\right) \times 100; \quad j = 2, 3, 4$$

(3.3)

The % of relative biases of the estimators are presented in Table 3.1. The estimator $v_1$ possesses a large bias for all the eight populations considered here. The estimators $v_2, v_3, v_4$ possess small and almost equal relative absolute bias in all situations while $v_1^*, v_2^*, v_3^*, v_4^*$ possess moderate bias of equal magnitudes. It is very difficult to choose any particular estimator from $v_2, v_3, v_4$ since none of them have smaller bias in all the situations. The relative bias of the estimator $v_2$ is negative which implies that it underestimates the variance while the rest estimators have positive bias which indicates overestimation of the variance.

As regards to efficiencies of the estimators $v_3, v_4$ bring enormous gain in efficiency over $v_2$. Similarly, $v_3^*, v_4^*$ are much more efficient than $v_2^*$. It is very difficult to choose between $v_3$ and $v_4$ since none is better in all the situations. The efficiencies of the estimators $v_3^*$ and $v_4^*$ are virtually the same in all situations. The relative efficiencies of the estimators are given in table 3.2. Among the seven estimators $v_1, v_2, v_3, v_4, v_1^*, v_2^*, v_3^*$ only $v_1$ and $v_3$ were found to take negative values. The proposed estimator $v_3$ was found to take negative values much less than the existing estimator $v_1$. The number of negative values taken by the estimators $v_1$ and $v_3$ are presented in table 3.3.
## Table 3.0
Coefficients of Variation for the Population

<table>
<thead>
<tr>
<th>Coefficient of variation (CV)</th>
<th>$N=100$</th>
<th></th>
<th></th>
<th>$N=200$</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\alpha = 0.0$</td>
<td>$\alpha = 5.0$</td>
<td>$\alpha = 0.0$</td>
<td>$\alpha = 5.0$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$g$</td>
<td>0.0</td>
<td>0.5</td>
<td>0.0</td>
<td>0.5</td>
<td>0.0</td>
<td>0.5</td>
</tr>
<tr>
<td>$CV(x)$</td>
<td>0.2023</td>
<td>0.2023</td>
<td>0.2023</td>
<td>0.2023</td>
<td>0.2031</td>
<td>0.2031</td>
</tr>
<tr>
<td>$CV(y)$</td>
<td>0.2026</td>
<td>0.2113</td>
<td>0.1976</td>
<td>0.1988</td>
<td>0.2034</td>
<td>0.2050</td>
</tr>
<tr>
<td>$CV(y/x)$</td>
<td>0.0044</td>
<td>0.0540</td>
<td>0.0073</td>
<td>0.0502</td>
<td>0.0054</td>
<td>0.0503</td>
</tr>
<tr>
<td>$[CV(z/x)]$</td>
<td>5.1554</td>
<td>5.2851</td>
<td>5.1609</td>
<td>5.2844</td>
<td>5.1549</td>
<td>5.2717</td>
</tr>
<tr>
<td>$\rho_{xy}$</td>
<td>0.9997</td>
<td>0.9651</td>
<td>0.9996</td>
<td>0.9710</td>
<td>0.9996</td>
<td>0.9716</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Coefficient of variation (CV)</th>
<th>$N=100$</th>
<th></th>
<th></th>
<th>$N=200$</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\alpha = 0.0$</td>
<td>$\alpha = 5.0$</td>
<td>$\alpha = 0.0$</td>
<td>$\alpha = 5.0$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$g$</td>
<td>0.0</td>
<td>0.5</td>
<td>0.0</td>
<td>0.5</td>
<td>0.0</td>
<td>0.5</td>
</tr>
<tr>
<td>$CV(x)$</td>
<td>0.2023</td>
<td>0.2023</td>
<td>0.2023</td>
<td>0.2023</td>
<td>0.2031</td>
<td>0.2031</td>
</tr>
<tr>
<td>$CV(y)$</td>
<td>0.2026</td>
<td>0.2113</td>
<td>0.1976</td>
<td>0.1988</td>
<td>0.2034</td>
<td>0.2050</td>
</tr>
<tr>
<td>$CV(y/x)$</td>
<td>0.0044</td>
<td>0.0540</td>
<td>0.0073</td>
<td>0.0502</td>
<td>0.0054</td>
<td>0.0503</td>
</tr>
<tr>
<td>$[CV(z/x)]$</td>
<td>5.1554</td>
<td>5.2851</td>
<td>5.1609</td>
<td>5.2844</td>
<td>5.1549</td>
<td>5.2717</td>
</tr>
<tr>
<td>$\rho_{xy}$</td>
<td>0.9997</td>
<td>0.9651</td>
<td>0.9996</td>
<td>0.9710</td>
<td>0.9996</td>
<td>0.9716</td>
</tr>
</tbody>
</table>
Table 3.1
Relative Percentage Bias

\[
\begin{align*}
&v_1 = \hat{V}_{DE}^ht, v_2 = \hat{V}_{DE}^yt, v_1^* = \hat{V}_{HA}^ht, v_2^* = \hat{V}_{HA}^yt, v_3 = \hat{V}_{DE}^ht, v_4 = \hat{V}_{DE}^yt, v_3^* = \hat{V}_{HA}^ht, v_4^* = \hat{V}_{HA}^yt
\end{align*}
\]

<table>
<thead>
<tr>
<th>Estimators</th>
<th>$N=100$</th>
<th>$N=200$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\alpha=0.0$</td>
<td>$\alpha=5.0$</td>
</tr>
</tbody>
</table>
|            | $g$ | $g$ | $g$ | $g$
| $v_1$ | 150.64 | 127.90 | 160.94 | 136.73 | 132.53 | 106.55 | 141.83 | 114.12 |
| $v_2$ | -1.78 | -0.86 | -1.93 | -1.09 | -4.56 | -3.15 | -4.72 | -3.45 |
| $v_3$ | 1.91 | -4.00 | 3.23 | -3.16 | 0.54 | -5.86 | 1.79 | -5.53 |
| $v_4$ | 2.24 | 4.73 | 2.31 | 4.70 | 0.79 | 3.33 | 0.89 | 3.27 |
| $v_1^*$ | 33.60 | 33.18 | 34.10 | 33.56 | 28.76 | 26.99 | 29.31 | 27.22 |
| $v_2^*$ | 29.37 | 29.19 | 29.80 | 29.49 | 24.74 | 23.32 | 25.20 | 23.47 |
| $v_3^*$ | 34.11 | 36.07 | 34.76 | 36.58 | 32.49 | 32.20 | 33.24 | 32.62 |
| $v_4^*$ | 34.12 | 36.16 | 34.76 | 36.67 | 32.50 | 32.30 | 33.25 | 32.72 |

$n_1 = 20, n_2 = 10$

| $n_1 = 30, n_2 = 15$
| $n_1 = 40, n_2 = 20$

$\hat{V}_{DE}^ht, \hat{V}_{DE}^yt, \hat{V}_{HA}^ht, \hat{V}_{HA}^yt$ are estimators for the relative percentage bias.
Table 3.2
Relative Efficiencies

\[ v_2 = \hat{V}_{DE}, v_2^* = \hat{V}_{HA}, v_3 = \hat{V}_{DE}, v_4 = \hat{V}_{HY}, v_3^* = \hat{V}_{HA}, v_4^* = \hat{V}_{HY} \]

<table>
<thead>
<tr>
<th>Estimators</th>
<th>( N=100 )</th>
<th>( N=200 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \alpha = 0.0 )</td>
<td>( \alpha = 5.0 )</td>
<td>( \alpha = 0.0 )</td>
</tr>
<tr>
<td>( g )</td>
<td>( g )</td>
<td>( g )</td>
</tr>
<tr>
<td>0</td>
<td>0.5</td>
<td>0</td>
</tr>
</tbody>
</table>

\( n_2 = 10 \) \( n_1 = 20 \)

| \( v_2 \) | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| \( v_3 \) | 231.54 | 197.67 | 224.54 | 201.50 | 206.68 | 87.27 | 202.62 |
| \( v_4 \) | 230.68 | 174.10 | 229.56 | 175.01 | 206.31 | 167.25 | 205.14 |
| \( v_2^* \) | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| \( v_3^* \) | 167.13 | 129.18 | 165.46 | 128.26 | 163.31 | 134.98 | 161.23 |
| \( v_4^* \) | 167.09 | 128.83 | 165.42 | 127.91 | 163.28 | 134.68 | 161.21 |

\( n_2 = 15 \) \( n_1 = 30 \)

| \( v_2 \) | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| \( v_3 \) | 252.46 | 191.95 | 242.16 | 198.29 | 213.88 | 178.36 | 211.01 |
| \( v_4 \) | 251.89 | 189.31 | 247.28 | 190.77 | 213.52 | 174.92 | 212.87 |
| \( v_2^* \) | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| \( v_3^* \) | 153.42 | 124.89 | 154.33 | 123.98 | 155.52 | 139.35 | 153.90 |
| \( v_4^* \) | 153.39 | 124.54 | 154.30 | 123.64 | 155.50 | 139.04 | 153.88 |

\( n_2 = 20 \) \( n_1 = 40 \)

| \( v_2 \) | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| \( v_3 \) | 269.40 | 214.62 | 249.84 | 221.96 | 217.21 | 177.69 | 216.26 |
| \( v_4 \) | 268.15 | 185.82 | 258.19 | 186.62 | 216.65 | 179.07 | 218.53 |
| \( v_2^* \) | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| \( v_3^* \) | 142.19 | 117.13 | 137.56 | 115.85 | 153.66 | 142.27 | 149.37 |
| \( v_4^* \) | 142.16 | 116.83 | 137.53 | 115.57 | 153.64 | 141.97 | 149.35 |
Table 3.3
Number of Negative Values

\( v_1 = \hat{V}_{DE}^{ht}, v_3 = \hat{V}_{DE}^{ht} \)

<table>
<thead>
<tr>
<th>Estimators</th>
<th>( N=100 )</th>
<th>( N=200 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \alpha = 0.0 )</td>
<td>( \alpha = 5.0 )</td>
</tr>
<tr>
<td>( g )</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>( v_1 )</td>
<td>32674</td>
<td>31015</td>
</tr>
<tr>
<td>( v_3 )</td>
<td>239</td>
<td>2462</td>
</tr>
<tr>
<td></td>
<td>( n_1 = 20, n_2 = 10 )</td>
<td>( n_1 = 30, n_2 = 15 )</td>
</tr>
<tr>
<td>( v_1 )</td>
<td>27492</td>
<td>26273</td>
</tr>
<tr>
<td>( v_3 )</td>
<td>165</td>
<td>1974</td>
</tr>
<tr>
<td></td>
<td>26571</td>
<td>24454</td>
</tr>
<tr>
<td></td>
<td>209</td>
<td>2335</td>
</tr>
<tr>
<td></td>
<td>27569</td>
<td>24656</td>
</tr>
</tbody>
</table>

ACKNOWLEDGEMENT

The authors wish to thank the referees for their valuable suggestions which lead to substantial improvement of the earlier version of the paper.

REFERENCES