Jour. Ind. Soc. Ag. Statistics
56(3), 2003 : 237-252

On the Estimation of Total, Mean and Distribution Function using Two-Phase Sampling: Calibration Approach

Sarjinder Singh and Sergio Martinez Puertas[1]
St. Cloud State University, St. Cloud, MN 56301, USA
(Received : May, 2001)

SUMMARY

In this paper, a general set-up for estimating population total and distribution function in two-phase sampling has been proposed. The estimators of population total considered by Hidiroglou and Sarndal ([9], [10]) and Dupont [7] in two-phase sampling are shown to be special cases of the proposed strategy. Following Singh *et al.* [23], a higher level calibration approach in two-phase sampling has also been discussed, which is in fact an extension of the recent work by Singh ([21], [22]). The statistical package, GES, developed at Statistics Canada may be further improved to handle two-phase sampling strategies using higher order calibration approach. An empirical study has also been carried out to study the performance of the proposed strategies.

Key words : Calibration approach, Estimation of totals and distribution functions, Two-phase sampling.

1. Introduction

The use of two-phase sampling is a powerful and cost-effective technique and hence has an eminent role in survey sampling. The population is represented by $\Omega = \{1, 2, \ldots, i, \ldots, N\}$. A first-phase probability sample $s_1, (s_1 \subset \Omega)$ is drawn from the population Ω using a sampling design that generates the selection probabilities π_{1i}. Given that the first sample s_1 has been drawn, the second-phase sample $s_2, (s_2 \subset s_1 \subset \Omega)$ is selected from s_1, with a sampling design with the selection probabilities $\pi_{2i} = \pi_{i|s_1}$. Evidently, the first-phase and second-phase sampling weights are defined as $d_{1i} = \frac{1}{\pi_{1i}}$ and $d_{2i} = \frac{1}{\pi_{2i}}$, respectively. The overall sampling weights for the selected i^{th} unit in the second

[1] University of Almeria, Spain

phase sample s_2 will be $d_i^* = d_{1i}d_{2i}$. Since we have considered the problem of estimation of totals and distribution functions in two-phase sampling, the following table summarizes our assumptions on the information available at estimation stage.

Table 1. Relationship between set of units and available data at each phase

Set of units	Data available
Population	$\{Z_i \mid i \in \Omega\}$, $H_Z = \sum_{i \in \Omega} h(z_i)$ $V_{HT}(\hat{H}_Z) = \frac{1}{2}\sum_{i \in \Omega}\sum_{j \in \Omega} D_{1ij}\left(\frac{h(z_i)}{\pi_{1i}} - \frac{h(z_j)}{\pi_{1j}}\right)^2$ where $D_{1ij} = (\pi_{1i}\,\pi_{1j} - \pi_{ij})$
First-phase sample	$\{(x_i, z_i) \mid i \in s_1\}\ i = 1, 2, \ldots m$
Second-phase sample	$\{(x_i, y_i, z_i) \mid i \in s_2\}\ i = 1, 2, \ldots, n$

Following mathematical notations of Rao [15], we have considered the problem of estimation of general parameters of interest

$$H_Y = \sum_{i \in \Omega} h(y_i) \text{ and } \overline{H}_Y = N^{-1} \sum_{i \in \Omega} h(y_i) \tag{1.1}$$

for a specified function h. For $h(y_i) = y_i$, H_Y and $\overline{H}_Y$ respectively give the population total and mean. Also $h(y_i) = \Delta(t - y_i)$, with $\Delta(a) = 1$ when $a \geq 0$ and $\Delta(a) = 0$ otherwise, gives the distribution function

$$\overline{H}_Y = F_y(t) = N^{-1} \sum_{i = \Omega} \Delta(t - y_i) \tag{1.2}$$

An unbiased estimator of population parameter H_Y in two-phase sampling is given by

$$\hat{H}_u = \sum_{i=1}^{n} d_{1i}d_{2i}h(y_i) \tag{1.3}$$

We have considered a new estimator of H_y in two-phase sampling, defined as

$$\hat{H}_1 = \sum_{i=1}^{n} \tilde{d}_i^* h(y_i) \tag{1.4}$$

where $\tilde{d}_i^*$ the ultimate calibrated weights, can be obtained in several ways. It is interesting to note that in multi-phase sampling, we can obtain calibrated weights at each phase separately. For example in the present situation, we are considering two phase sampling, hence two types of calibrated weights are possible.

(a) First-phase calibrated weights, say, $\tilde{d}_{1i}$, i = 1, 2, ..., m

(b) Second-phase calibrated weights, say, $\tilde{d}_i^*$, i = 1, 2, ..., n

A combination of (a) and (b) will lead to ultimate calibrated weights in two-phase sampling, which is in fact an extension of the work of Deville and Sarndal [6].

The next sections have been devoted to discuss the purpose of the first-phase and second-phase calibration at the estimation stage.

2. *First-phase Calibration*

An unbiased estimator of population parameter $H_X = \sum_{i \in \Omega} h(x_i)$ from the first-phase sample information available in Table 1 is given by

$$\hat{H}_X = \sum_{i=1}^{m} d_{1i} h(x_i) \tag{2.1}$$

A calibrated estimator of population parameter H_X is given by

$$\hat{H}_X^* = \sum_{i=1}^{m} \tilde{d}_{1i} h(x_i) \tag{2.2}$$

where $\tilde{d}_{1i}$ are the calibrated weights obtained from the first-phase sample information. Choose the first-phase calibrated weights $\tilde{d}_{1i}$ such that the chi-square distance

$$D_1 = \sum_{i=1}^{m} \frac{(\tilde{d}_{1i} - d_{1i})^2}{q_{1i} d_{1i}} \tag{2.3}$$

is minimum subject to the first-phase calibration constraint

$$\sum_{i=1}^{m} \tilde{d}_{1i} h(z_i) = H_Z \tag{2.4}$$

The choice of q_{1i} decides the form of the estimator. Then the first phase calibrated weights are given by

$$\tilde{d}_{1i} = d_{1i} + \frac{q_{1i} d_{1i} h(z_i)}{\sum_{i=1}^{m} d_{1i} q_{1i} \{h(z_i)\}^2} \left(H_Z - \sum_{i=1}^{m} d_{1i} h(z_i) \right) \tag{2.5}$$

An estimator of population parameter H_X can be obtained on substituting (2.5) in (2.2). Our objective is to obtain calibrated estimator of population parameter H_Y of the study variable Y rather than that of the population parameter H_X of the auxiliary variable X. To achieve our goal, we have to obtain new calibrated weights, called second-phase calibrated weights. We discuss here the simplest method, which results in the chain/regression type estimators for the population parameter H_Y.

3. Second-phase Calibration

An unbiased estimator of the general population parameter H_Y in two-phase sampling is defined as

$$\hat{H}_u = \sum_{i=1}^{n} d_{1i} d_{2i} h(y_i) \tag{3.1}$$

We consider another estimator of the population parameter H_Y defined as

$$\hat{H}_c = \sum_{i=1}^{n} \tilde{d}_i^* h(y_i) \tag{3.2}$$

where $\tilde{d}_i^*$ are called the second-phase calibrated weights. Let us choose the second-phase calibrated weights $\tilde{d}_i^*$ such that the chi-square function

$$D_2 = \sum_{i=1}^{n} \frac{(\tilde{d}_i^* - d_{1i} d_{2i})^2}{d_{1i} d_{2i} q_{2i}} \tag{3.3}$$

is minimum subject to the calibration constraint

$$\sum_{i=1}^{n} \tilde{d}_i^* h(x_i) = \hat{H}_X^* \tag{3.4}$$

where $\hat{H}_X^*$ is given by (2.2) after fixing first-phase calibration. The choice of q_{2i} gives different forms of estimators in two-phase sampling. Minimization of (3.3), subject to (3.4), leads to second-phase calibrated weights given by

$$\tilde{d}_i^* = d_{1i}d_{2i} + \frac{d_{1i}d_{2i}q_{2i}h(x_i)}{\sum_{i=1}^{n} d_{1i}d_{2i}q_{2i}\{h(x_i)\}^2}\left(\hat{H}_X^* - \sum_{i=1}^{n} d_{1i}d_{2i}h(x_i)\right) \tag{3.5}$$

On using (3.5) in (3.2), we get the calibrated estimator in two-phase sampling given by

$$\hat{H}_c = \sum_{i=1}^{n} d_{1i}d_{2i}h(y_i) + \frac{\sum_{i=1}^{n} d_{1i}d_{2i}q_{2i}h(x_i)h(y_i)}{\sum_{i=1}^{n} d_{1i}d_{2i}q_{2i}\{h(x_i)\}^2}\left[\hat{H}_X^* - \sum_{i=1}^{n} d_{1i}d_{2i}h(x_i)\right] \tag{3.6}$$

where

$$\hat{H}_X^* = \sum_{i=1}^{m} d_{1i}h(x_i) + \frac{\sum_{i=1}^{m} d_{1i}q_{1i}h(x_i)h(z_i)}{\sum_{i=1}^{m} d_{1i}q_{1i}\{h(z_i)\}^2}\left[H_Z - \sum_{i=1}^{m} d_{1i}h(z_i)\right] \tag{3.7}$$

The next section is devoted to discuss special casa of the estimator (3.7) available in the literature, but not discussed by Hidiroglou and Sarndal ([9], [10]).

4. Special Cases

Case 1. If $q_{1i} = \{h(z_i)\}^{-1}$ and $q_{2i} = \{h(x_i)\}^{-1}$, then the calibrated estimator $\hat{H}_c$ at (3.6) reduces to the chain ratio type estimator of population parameter H_X as

$$\hat{H}_R = \sum_{i=1}^{n} d_{1i}d_{2i}h(y_i)\left(\frac{\sum_{i=1}^{n} d_{1i}h(x_i)}{\sum_{i=1}^{n} d_{1i}d_{2i}h(x_i)}\right)\left(\frac{H_Z}{\sum_{i=1}^{m} d_{1i}h(z_i)}\right) \tag{4.1}$$

Case 2. If $q_{1i} = 1$ and $q_{2i} = 1$, then the resultant calibrated estimator (3.6) becomes

$$\hat{H}_G = \sum_{i=1}^{n} d_{1i}d_{2i}h(y_i) + \hat{\beta}_1\left[\sum_{i=1}^{m} d_{1i}h(x_i) - \sum_{i=1}^{n} d_{1i}d_{2i}h(x_i)\right] + \hat{\beta}_2\left[H_Z - \sum_{i=1}^{m} d_{1i}h(z_i)\right] \tag{4.2}$$

where $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} d_{1i} d_{2i} h(x_i) h(y_i)}{\sum_{i=1}^{n} d_{1i} d_{2i} \{h(z_i)\}^2}$ and $\hat{\beta}_2 = \hat{\beta}_1 \left[\dfrac{\sum_{i=1}^{m} d_{1i} h(x_i) h(z_i)}{\sum_{i=1}^{m} d_{1i} \{h(z_i)\}^2} \right]$

have their usual meanings. The estimator $\hat{H}_G$ can be easily named as chain regression type estimator of population parameter H_Y.

5. Conditional Variance of the Calibrated Estimator

Define V_1 and V_2 as the variance over all possible first-phase samples and for all second-phase samples from a given first-phase sample. For the given first-phase and second-phase samples, the weights $\tilde{d}_i^*, (i \in s_2)$ satisfy the calibration constraint, but rest of the weights $\tilde{d}_i^*, (i \notin s_2 \cap i \in \Omega)$ can easily be forecasted using known auxiliary information. For the given first-phase and second phase samples, we have

$$V(\hat{H}_c \mid s_1, s_2) = E_1 V_2(\hat{H}_c) + V_1 E_2(\hat{H}_c) = E_1 V_2 \left[\sum_{i=1}^{n} \tilde{d}_i^* h(y_i) \right] + V_1 E_2 \left[\sum_{i=1}^{n} \tilde{d}_i^* h(y_i) \right]$$

$$= E_1 \left[\frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} (\pi_{2i}\pi_{2j} - \pi_{2ij}) \{ \tilde{d}_i^* h(y_i) - \tilde{d}_j^* h(y_j) \}^2 \right] + V_1 \left[\sum_{i=1}^{m} \tilde{d}_i^* \pi_{2i} h(y_i) \right]$$

$$= \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \pi_{1ij} (\pi_{2i}\pi_{2j} - \pi_{2ij}) \left\{ \tilde{d}_i^* h(y_i) - \tilde{d}_j^* h(y_i) \right\}^2$$

$$+ \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} (\pi_{1i}\pi_{1j} - \pi_{1ij}) \left\{ \tilde{d}_i^* \pi_{2i} \pi_{1i} h(y_i) - \tilde{d}_j^* \pi_{2j} \pi_{1j} h(y_j) \right\}^2 \qquad (5.1)$$

In the next sections, we consider the problem of estimation of variance (5.1) using two levels of calibration.

6. Estimators of Variance: Low Level Calibration

Using the concept of two-phase sampling, an unbiased estimator of $V(\hat{H}_c \mid s_1, s_2)$ is

$$\hat{V}(\hat{H}_c \mid s_1, s_2) = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{(\pi_{2i}\pi_{2j} - \pi_{2ij})}{\pi_{2ij}} \left\{ \tilde{d}_i^* h(y_i) - \tilde{d}_j^* h(y_j) \right\}^2$$

$$+\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{(\pi_{1i}\pi_{1j}-\pi_{1ij})}{\pi_{1ij}\pi_{2ij}}\left\{\tilde{d}_i^*\pi_{2i}\pi_{1i}h(y_i)-\tilde{d}_{ji}^*\pi_{2j}\pi_{1j}h(y_j)\right\}^2 \quad (6.1)$$

Following Singh *et al.* [23], a low level calibrated estimator of variance of $\hat{H}_c$ is

$$\hat{V}_l\left(\hat{H}_c \mid s_1, s_2\right)=\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\pi_{1ij}W_{2ij}\{\tilde{d}_i^*h(y_i)-\tilde{d}_j^*h(y_i)\}^2$$

$$+\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}W_{1ij}\{\tilde{d}_i^*\pi_{2i}\pi_{1i}h(y_i)-\tilde{d}_j^*\pi_{2j}\pi_{1j}h(y_j)\}^2 \quad (6.2)$$

where $$W_{2ij}=\frac{(\pi_{2i}\pi_{2j}-\pi_{2ij})}{\pi_{1ij}\pi_{2ij}},\ W_{1ij}=\frac{(\pi_{1i}\pi_{1j}-\pi_{1ij})}{\pi_{1ij}\pi_{2ij}}$$

$$\tilde{d}_{1i}=d_{1i}+\frac{q_{1i}d_{1i}h(z_i)}{\sum_{i=1}^{m}d_{1i}q_{1i}\{h(z_i)\}^2}\left(H_Z-\sum_{i=1}^{m}d_{1i}h(z_i)\right)$$

and

$$\tilde{d}_i^*=d_{1i}d_{2i}+\frac{d_{1i}d_{2i}q_{2i}h(x_i)}{\sum_{i=1}^{n}d_{1i}d_{2i}q_{2i}\{h(x_i)\}^2}\left(\hat{H}_X^*-\sum_{i=1}^{n}d_{1i}d_{2i}h(x_i)\right)$$

A large number of estimators of variance can be shown to be special cases of the estimator considered at (6.2).

7. *Estimators of Variance: Higher Level Calibration*

Following Singh *et al.* [23], a higher order calibration estimator of the variance in two-phase sampling is given by

$$\hat{V}_h(\hat{H}_c)=\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\pi_{1ij}\Omega_{2ij}\{\tilde{d}_i^*h(y_i)-\tilde{d}_j^*h(y_j)\}^2$$

$$+\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Omega_{1ij}\{\tilde{d}_i^*\pi_{2i}\pi_{1i}h(y_i)-\tilde{d}_j^*\pi_{2j}\pi_{1j}h(y_i)\}^2 \quad (7.1)$$

where Ω_{1ij} and Ω_{2ij} are the weights such that the distance between Ω_{1ij} and W_{1ij} and that between Ω_{2ij} and W_{2ij} is minimum. Let us define two chi-square type distance functions

$$D_1 = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\frac{(\Omega_{1ij} - W_{1ij})^2}{Q_{1ij}W_{1ij}} \tag{7.2}$$

and

$$D_2 = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{(\Omega_{2ij} - W_{2ij})^2}{Q_{2ij}W_{2ij}} \tag{7.3}$$

Also let us define, the first calibration constraints as

$$\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\Omega_{1ij}\{d_{1i}h(z_i) - d_{1j}h(z_j)\}^2 = V(\hat{H}_Z) \tag{7.4}$$

where $V(\hat{H}_Z) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\pi_{1i}\pi_{1j} - \pi_{1ij})\{d_{1i}h(z_i) - d_{1j}h(z_j)\}^2$ denotes the known variance of the estimator of the auxiliary character z_i based on the assumption made in Table 1. Second calibration constraint can be defined as

$$\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Omega_{2ij}\{d_2h(x_i) - d_{2j}h(x_j)\}^2 = \hat{V}(\hat{H}_x) \tag{7.5}$$

where $\hat{V}(\hat{H}_X) = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\frac{(\pi_{1i}\pi_{1j} - \pi_{1ij})}{\pi_{1ij}}\{d_{1i}\ h(x_i) - d_{1j}h(x_j)\}^2$ denotes the estimator of variance of the estimator of population parameter of second auxiliary character x_i based on first-phase sample information. Minimization of (7.2) subject to (7.4) leads to second order calibrated weights obtained from first phase sample information given by

$$\Omega_{1ij} = W_{1ij} + \frac{Q_{1ij}W_{1ij}\{d_{1i}h(z_i) - d_1h(z_j)\}^2}{\sum_{i=1}^{m}\sum_{j=1}^{m}Q_{1ij}W_{1ij}\{d_{1i}h(z_i) - d_{1j}h(z_j)\}^4}$$
$$\left[V\left(\hat{H}_z - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}W_{1ij}\{d_{1j}h(z_i) - d_{1j}h(z_j)\}^2\right)\right] \tag{7.6}$$

The minimization of (7.3) subject to (7.5) leads to second order calibrated weights obtained from second phase sample information, given by

$$\Omega_{2ij} = W_{2ij} + \frac{Q_{2ij}W_{2ij}\{d_{2i}h(x_i) - d_{2j}h(x_j)\}^2}{\sum_{i=1}^{n}\sum_{j=1}^{n} Q_{2ij}W_{2ij}\{d_{2i}h(x_i) - d_{2j}h(x_j)\}^4}$$
$$\left[\hat{V}(\hat{H}_x) - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} W_{2ij}\{d_{2i}h(x_i) - d_{2j}h(x_j)\}^2\right] \tag{7.7}$$

Use of (7.6) and (7.7) in (7.1) leads to the higher level calibration estimator of variance in two-phase sampling. Several estimators can be shown to be special cases of the proposed higher level calibration approach.

8. Applications of the Proposed Strategy

Here we will discuss a simple case to study the performance of the higher level calibration estimators in comparison to lower level calibration estimators in two-phase sampling. The well known regression estimator of population mean in two-phase sampling is given by

$$\bar{y}_{lr} = \bar{y}_n + \beta_1(\bar{x}_m - \bar{x}_n) + \beta_2(\bar{Z} - \bar{z}_n) \tag{8.1}$$

where β_1 and β_2 are suitably chosen constants such that the variance of the estimator (8.1) is minimum. Also $\bar{y}_n = n^{-1}\sum_{i=1}^{n} y_i$, $\bar{x}_m = m^{-1}\sum_{i=1}^{m} x_i$, $\bar{x}_n = n^{-1}\sum_{i=1}^{n} x_i$, $\bar{z}_n = n^{-1}\sum_{i=1}^{n} z_i$ and $\bar{Z} = N^{-1}\sum_{i=1}^{N} z_i$ have their usual meaning.

Let us consider two super-population models defined as

$$Y_i = \beta_2^* Z_i + \eta_i \tag{8.2}$$

and

$$Y_i = \beta_1 X_i + \beta_2 Z_i + \upsilon_i \tag{8.3}$$

where η_i and υ_i are independent random errors following the assumptions of the ordinary least squares method. Under SRSWOR sampling, the variance of the estimator $\bar{y}_{lr}$ can be expressed as

$$V(\bar{y}_{lr}) = \left(\frac{1}{m} - \frac{1}{N}\right)\frac{1}{N-1}\sum_{i=1}^{N}\eta_i^2 + \left(\frac{1}{n} - \frac{1}{m}\right)\frac{1}{N-1}\sum_{i=1}^{N}\upsilon_i^2 \tag{8.4}$$

Obviously an estimator of $V(\bar{y}_{lr})$ under the concept of two super-population models is given by

$$\hat{V}_1(\bar{y}_{lr}) = \left(\frac{1}{m} - \frac{1}{N}\right)\frac{1}{n-1}\sum_{i=1}^{n}\hat{\eta}_i^2 + \left(\frac{1}{n} - \frac{1}{m}\right)\frac{1}{n-1}\sum_{i=1}^{n}\hat{\upsilon}_i^2 \tag{8.5}$$

where $\hat{\eta}_i = (y_i - \bar{y}_n) - \hat{\beta}_2^*(z_i - \bar{z}_n)$ and $\hat{\upsilon}_i = (y_i - \bar{y}_n) - \hat{\beta}_1(x_i - \bar{x}_n)$ $w = \frac{-q}{c}, q$ are the estimates of the residuals from two different super-population models.

Under the concept of low level calibration approach, we consider the following estimator

$$\hat{V}_2(\bar{y}_{lr}) = \left[\left(\frac{1}{m} - \frac{1}{N}\right)\frac{1}{n-1}\sum_{i=1}^{n}\hat{\eta}_i^2 + \left(\frac{1}{n} - \frac{1}{m}\right)\frac{1}{n-1}\sum_{i=1}^{n}\hat{\upsilon}_i^2\right]\left(\frac{\bar{x}_m}{\bar{x}_n}\right)^2\left(\frac{\bar{Z}}{\bar{z}_n}\right)^2 \tag{8.6}$$

Under the concept of higher level calibration approach, we consider the following estimator

$$\hat{V}_3(\bar{y}_{lr}) = \left[\left(\frac{1}{m} - \frac{1}{N}\right)\frac{1}{n-1}\sum_{i=1}^{n}\hat{\eta}_i^2\left(\frac{s_1^2(z)}{s_2^2(z)}\right) + \left(\frac{1}{n} - \frac{1}{m}\right)\frac{1}{n-1}\sum_{i=1}^{n}\hat{\upsilon}_i^2\left(\frac{S_z^2}{s_1^2(z)}\right)\right]$$
$$\left(\frac{\bar{x}_m}{\bar{x}_n}\right)^2\left(\frac{\bar{Z}}{\bar{z}_n}\right)^2\left(\frac{s_1^2(x)}{s_2^2(x)}\right) \tag{8.7}$$

where

$$s_1^2(x) = (m-1)^{-1}\sum_{i=1}^{m}(x_i - \bar{x}_m)^2, \; s_2^2(x) = (n-1)^{-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$$

$$s_1^2(z) = (m-1)^{-1}\sum_{i=1}^{m}(z_i - \bar{z}_m)^2, \; s_2^2(z) = (n-1)^{-1}\sum_{i=1}^{n}(z_i - \bar{z}_n)^2$$

and $s_Z^2 = (N-1)^{-1}\sum_{i=1}^{N}(Z_i - \bar{Z})^2$

have their usual meanings.

9. Empirical Study

The empirical study has been carried out on the basis of two types of data, viz., real data and artificial data. The real data have been taken from the web page www.minagricultura.gov.co and have been shown in Table 3, whereas artificial data have been generated according to the practical situations in actual practice using standard subroutine available in FORTRAN. We will first explain the steps taken for simulation while using real data as follows.

Real Data: First we consider the problem of estimation of production of the cocoa in N = 22 regions of Colombia during 2000. For the purpose of the empirical study, we assumed y = "Production of cocoa during 2000", x = "efficiency of area cultivated during 2000", and z = "Area cultivated during 2000". First we selected all possible $^{22}C_7 = 170544$ preliminary large samples of size m = 7 units. From every given preliminary large sample of m = 7 units we selected one random sample of size n = 5 by SRSWOR sampling as second-phase samples. Thus we obtained regression estimate of mean and estimates of variance considered in this paper for all possible $^{22}C_7$ ultimate samples. On the basis of sample information so obtained, the 95% coverage by the confidence intervals (CCI) were obtained by counting the number of times the true population mean $\overline{Y}$ falls in the interval given by

$$\overline{y}_{lr} + \mp t_{df_j}(0.05)\sqrt{\hat{V}_j},\ j = 1, 2, 3 \tag{9.1}$$

where $df_j = n - j$ was used for j^{th} estimator. The results so obtained are presented in Table 2.

Table 2. Results obtained from real populations for m = 7 and n = 5

Description of population	95%		
	CCI(1)	CCI(2)	CCI(3)
N = 22 y = "Production of cocoa during 2000" x = "Efficiency of area cultivated during 2000" z = "Area cultivated during 2000"	0.4644	0.5364	0.852

Artificial Data: In order to study the performance of the proposed estimators in actual practice, we generated two auxiliary characters having different amounts of correlation with study variable. The transformations used to generate different variables are given by

$$x_i = 20 + \sqrt{(1-\rho_{XY})}\,x_i^* + \rho_{XY}\frac{S_X}{S_Y}y_i^* \tag{9.2}$$

$$y_i = 10 + y_i^* \tag{9.3}$$

and

$$z_i = 20 + \sqrt{(1-\rho_{XZ})}x_i^* + \rho_{XZ}\frac{S_X}{S_Z}z_i^* \tag{9.4}$$

where x_i^*, y_i^* and z_i^* are independent beta variates generated by subroutine BETACH for a = 2.6, b = 2.3, seed1 = 1331963, seed2 =1963133, seed3 = 568798, $S_Y = 5.5$, $S_X = 1.5$ and $S_Z = 3.5$ following Bratley *et al.* [3] in FORTRAN 77 for different values of correlation coefficient ρ. The values of $\rho_{YX} = 0.95$ and $\rho_{XZ} = 0.75$ were fixed, because for the chainratio or regression type estimators, it is assumed that the variable Z is highly correlated with X and remotely correlated with Y. For a population of N = 100 units, we selected randomly 10,000 first-phase samples each of size m = 20 units. From the given first-phase sample, we selected randomly 5000 second-phase samples each of size n = 10 units. The rest of the procedure was repeated as we did for real data. Similar exercises with slight modifications in the above transformations where needed and different sample sizes were repeated for other distributions as shown in Table 4. We observe that the estimator V performs better than all the other estimators considered in the present investigation.

10. Conclusions

The proposed methodology is the generalization of the exiting methodology in the literature under the concept of two-phase sampling. The statistical package, GES, developed at Statistics Canada, can be further modified to obtain better estimators of variance of any parameter of interest in survey sampling under the concept of two-phase sampling.

Table 3. Production of the cocoa in N-22 regions of Colombia during 2000.

Region	Area cultivated	Production	Efficiency
Antioquia	4530	1501	331
Arauca	6004	3457	576
Boyaca	321	150	467
Caldas	844	383	454
Caqueta	420	219	521
Cauca	241	130	539
Cesar	2222	1061	477
Choco	1309	351	268
Cundinamarca	1104	581	526
Guainia	627	246	392
Huila	9118	3884	426
La Guajira	611	464	759
Magdalena	635	317	499
Meta	429	279	650
Narino	3950	728	184
N. Santander	11288	4610	408
Putumayo	22	4	182
Quindio	20	5	250
Risaralda	1070	450	421
Santander	40211	20547	511
Tolima	7537	4563	605
V. Cauca	154	59	383

Area in hectares (H)
Production in tons
Efficiency Kg/H

Table 4. The values of 95% confidence intervals (CI) obtained by three estimators from the different distributions

Description of various distributions used for generating populations						Sample size			
						m = 20, n = 10	m = 40, n = 10	m = 60, n = 15	m = 50, n = 20
Sr. No.	Distribution	Density function	Range	Skewed					
1	Right Triangular	$f(x) = 2(1-x)$	$0 \le x \le 1$	Positively	CI(1)	0.756	0.786	0.864	0.901
					CI (2)	0.862	0.901	0.922	0.921
					CI (3)	0.901	0.924	0.928	0.942
2	Exponential	$f(x) = e^{-x}$	$x \ge 0$	Positively	CI (1)	0.726	0.763	0.826	0.921
					CI (2)	0.812	0.852	0.874	0.954
					CI (3)	0.869	0.925	0.964	0.949
3	Chi-square at v = 6	$f(x) = \frac{1}{2^{v/2}\Gamma_{v/2}} e^{-x/2} x^{(v-2)/2}$	$x \ge 0$	Positively	CI (1)	0.658	0.695	0.742	0.896
					CI (2)	0.789	0.821	0.856	0.925
					CI (3)	0.695	0.842	0.879	0.934
4	Gamma, p = 2	$f(x) = \frac{1}{\Gamma_p} e^{-x} x^{p-1}$	$x \ge 0$	Positively	CI (1)	0.725	0.784	0.802	0.804
					CI (2)	0.765	0.796	0.826	0.834
					CI (3)	0.795	0.812	0.865	0.886
5	Log Normal	$f(x) = \frac{1}{x\sqrt{2\pi}} e^{-\{\log(x)\}^2/2}$	$x > 0$	Positively	CI (1)	0.698	0.725	0.736	0.865
					CI (2)	0.697	0.736	0.736	0.870
					CI (3)	0.645	0.736	0.738	0.887
6	Beta B(.4, 1)	$f(x) = \frac{1}{\beta(p,q)} x^{p-1}(1-x)^{q-1}$	$0 \le x \le 1$	Positively	CI (1)	0.769	0.801	0.836	0.921
					CI (2)	0.836	0.823	0.896	0.946
					CI (3)	0.845	0.863	0.921	0.965
7	B(1, .4)	$f(x) = \frac{1}{\beta(p,q)} x^{p-1}(1-x)^{q-1}$	$0 \le x \le 1$	Negatively	CI (1)	0.742	0.835	0.896	0.914
					CI (2)	0.796	0.836	0.945	0.945
					CI (3)	0.799	0.846	0.960	0.961

8	B(1.5, 2.5)	$f(x) = \frac{1}{\beta(p, q)} x^{p-1}(1-x)^{q-1}$	$0 \le x \le 1$	Positively	CI (1)	0.752	0.825	0.884	0.910
					CI (2)	0.796	0.826	0.954	0.927
					CI (3)	0.801	0.836	0.961	0.935
9	B(2.5, 1.5)	$f(x) = \frac{1}{\beta(p, q)} x^{p-1}(1-x)^{q-1}$	$0 \le x \le 1$	Negatively	CI (1)	0.768	0.825	0.885	0.914
					CI (2)	0.796	0.835	0.926	0.924
					CI (3)	0.802	0.839	0.931	0.930
10	B(2, 2)	$f(x) = \frac{1}{\beta(p, q)} x^{p-1}(1-x)^{q-1}$	$0 \le x \le 1$	Normal (Hump type)	CI (1)	0.723	0.814	0.902	0.923
					CI (2)	0.736	0.845	0.932	0.935
					CI (3)	0.745	0.896	0.936	0.954
11	B(.6, .6)	$f(x) = \frac{1}{\beta(p, q)} x^{p-1}(1-x)^{q-1}$	$0 \le x \le 1$	U-shaped (Cauldron shape)	CI (1)	0.698	0.732	0.812	0.914
					CI (2)	0.725	0.765	0.836	0.924
					CI (3)	0.755	0.801	0.864	0.945
12	Rayleigh $\alpha = 1.5$	$f(x) = 2\alpha x e^{-\alpha x^2}$	$x > 0$	Positively	CI (1)	0.836	0.856	0.895	0.921
					CI (2)	0.856	0.896	0.921	0.945
					CI (3)	0.901	0.914	0.935	0.975
13	Pareto $\alpha = 1.6$	$f(x) = \alpha / x^{\alpha+1}$	$x > 1$	Positively	CI (1)	0.821	0.851	0.862	0.921
					CI (2)	0.865	0.864	0.914	0.942
					CI (3)	0.921	0.895	0.965	0.951
14	Weibull $\alpha = 0.5$, $k = 1.2$, $\beta = 2.2$	$f(x) = kx^{\beta-1}{}_e - \alpha x^{\beta}$	$x > 0$	Positively	CI (1)	0.762	0.756	0.794	0.821
					CI (2)	0.796	0.782	0.814	0.834
					CI (3)	0.824	0.814	0.834	0.924

ACKNOWLEDGEMENTS

The authors are thankful to Professor Bhatia and learned referee to bring the original manuscript in the present form.

REFERENCES

[1] Abanihe, U.C.I. (1994). Reproductive motivation and family size preferences among Nigerian men. *Studies in Family Planning*, **25(3)**, 149-161.

[2] Ahmed, M.S. (1997). The general class of chain estimators for the ratio of two means using double sampling. *Comm. Stat.-Theory Methods*, **26(9)**, 2247-2254.

[3] Bratley, P., Fox, B.L. and Schrage, L.E. (1983). *A Guide to Simulation.* Springer -Verlag, New York.

[4] Chand, L. (1975). *Some ratio-type estimators based on two or more auxiliary variables.* Ph.D. thesis submitted to Iowa State University, Ames, Iowa.

[5] Deng, L.Y. and Wu, C.F.J. (1987). Estimation of variance of the regression estimator. *J. Amer. Statist. Assoc.*, **82**, 568-576.

[6] Deville, J.C. and Sarndal, C.E. (1992). Calibration estimators in survey sampling. *J. Amer. Statist. Assoc.*, **87**, 376-382.

[7] Dupont, F. (1995). Alternative adjustments where there are several levels of auxiliary information. *Survey Methodology*, **21**, 125-135.

[8] Gupta, R.K., Singh, S. and Mangat, N.S. (1992-93). Some chain ratio type estimators for estimating finite population variance. *Aligarh J. Statist.*, **12 & 13**, 65-69.

[9] Hidiroglou, M.A. and Sarndal, C.E. (1995). Use of auxiliary information for two-phase sampling. *Proc. Sec. Survey Res. Meth., Amer. Statist. Assoc.*, **2**, 873-878.

[10] Hidiroglou, M.A. and Sarndal, C.E. (1998). Use of auxiliary information for two-phase sampling. *Survey Methodology*, **24 (1)**, 11-20.

[11] Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.*, **47**, 663-685.

[12] Isaki, C.T. (1983). Variance estimation using auxiliary information. *J. Amer. Statist. Assoc.*, **78**, 117-123.

[13] Kiregyera, B. (1980). A chain ratio-type estimator in finite population double sampling using two auxiliary variables. *Metrika*, **27**, 217-223.

[14] Prasad, B., Singh, R.S. and Singh, H.P. (1996). Some chain ratio-type estimators for ratio of two population means using two auxiliary characters in two phase sampling. *Metron*, 95-113.

[15] Rao, J.N.K. (1994). Estimating totals and distribution functions using auxiliary information at the estimation stage. *J. Official Statist.*, **10(2)**, 153-165.

[16] Sahoo, L.N. and Swain, A.K.P.C. (1983). Chain ratio estimators. *Jour. Ind. Soc. Agril. Stat.*, **35**, 70-79.

[17] Sahoo, L.N. and Swain, A.K.P.C. (1986). Chain product estimators. *Aligarh J. Statist.*, **6**, 53-58.

[18] Singh, G.N. and Upadhyaya, L.N. (1995). A class of modified chain-type estimators using two auxiliary variables in two phase sampling. *Metron*, 117-125.

[19] Singh, S. (1991). Estimation of finite population variance using double sampling. *Aligarh J. Statist.*, **11**, 53-65.

[20] Singh, S. (1998). Adolescent childbearing in developing countries : A global review. *Studies in Family Planning*, **29(2)**, 117-136.

[21] Singh, S. (2000). Estimation of variance of regression estimator in two phase sampling. *Calcutta Statist. Assoc. Bull.*, **50**, 49-63.

[22] Singh, S. (2001). Generalized calibration approach for estimating the variance in survey sampling. *Ann. Inst. Statist. Math.* (In press)

[23] Singh, S., Horn, S. and Yu, F. (1998). Estimation of variance of the general regression estimator: Higher level calibration approach. *Survey Methodology*, **24(1)**, 41-50.

[24] Singh, S., Horn, S., Chowdhury, S. and Yu, F. (1999). Calibration of the estimators of variance. *Austr. New Zealand J. Statist.*, **41(2)**, 199-212.

[25] Singh, V.K. and Singh, G.N. (1991). Chain type regression estimator with two auxiliary variables under double sampling scheme. *Metron*, 279-283.

[26] Singh, V.K., Singh, H.P. and Singh, H.P. (1994). A general class of chain estimators for ratio and product of two means of a finite population. *Comm. Stat. - Theory Methods*, **23(5)**, 1341-1355.

[27] Srivastava, S.K. and Jhajj, H.S. (1981). A class of estimators of the population mean in survey sampling using auxiliary information. *Biometrika*, **68**, 341-343.

[28] Srivastava, S.R., Khare, B.B. and Srivastava, S.R. (1990). A generalized chain ratio estimator for mean of finite population. *Jour. Ind. Soc. Agril. Stat.*, **42**, 108-117.

[29] Upadhyaya, L.N., Khushwaha, K.S. and Singh, H.P. (1990). A modified chain ratio-type estimator in two-phase sampling using multi auxiliary information. *Metron*, 381-393.