



Catalogue no. 11-522-XIE

**Statistics Canada International Symposium
Series - Proceedings**

**Symposium 2003: Challenges
in Survey Taking for the Next
Decade**

2003



VARIANCE ESTIMATION IN TWO-PHASE SAMPLING

M.A. Hidirolou and J.N.K. Rao¹

ABSTRACT

Two-phase sampling is often used for estimating a population total or mean when the cost per unit of collecting auxiliary variables \mathbf{x} is much smaller than the cost per unit of measuring the characteristic of interest. In the first-phase, a large sample s_1 is drawn according to a specific sampling design $p(s_1)$ and \mathbf{x} is observed for the units $i \in s_1$. Given the first-phase sample s_1 , a second-phase sample s_2 is selected from s_1 according to a specified sampling design $\{p(s_2 | s_1)\}$ and (y, \mathbf{x}) is observed for the units $i \in s_2$. In some cases, the population totals of some components of \mathbf{x} may also be known. Two phase sampling is used for stratification at the second phase (Neyman, 1938; Rao, 1973) or both phases (Binder et al., 2000) and for regression estimation (Särndal et al., 1992, chapter 9; Hidirolou and Särndal, 1998). Horvitz-Thompson (HT) type variance estimators are used for variance estimation. However, the HT variance estimator in uni-phase sampling is known to be highly unstable and may take negative values when the units are selected with unequal probabilities. On the other hand, the Sen-Yates-Grundy (SYG) variance estimator is relatively stable and nonnegative for several unequal probability sampling designs with fixed sample sizes. In this paper, we extend the SYG variance estimators to two-phase sampling, assuming fixed first-phase sample size and fixed second phase sample size given the first-phase sample. We apply the new SYG variance estimators to two-phase sampling designs with stratification at the second phase or both phases. We also develop SYG type variance estimators of the two-phase regression estimators that make use of the first phase auxiliary data.

KEYWORDS: Double-Expansion Estimator; Ratio-Estimator; Regression Estimator; Stratification.

1. INTRODUCTION

Two-phase sampling is often used for estimating a population total or a mean when the cost per unit of collecting auxiliary data \mathbf{x} is much smaller than the cost per unit of measuring the characteristics of interest y . The sampling scheme consists of two phases. In the first-phase, a large sample s_1 of size n_1 is drawn from the universe U according to a specified sampling design with probabilities $\{p(s_1)\}$ and \mathbf{x} is observed for the sample units $i \in s_1$. Given the first-phase sample s_1 , the second-phase sample s_2 is selected from s_1 according to a specified sampling design with conditional probabilities $\{p(s_2 | s_1)\}$ and (y, \mathbf{x}) is observed for the units $i \in s_2$. In some cases, the population totals of some components \mathbf{x}_1 of \mathbf{x} may also be known.

Neyman (1938) first proposed two-phase sampling for stratification. The first-phase sample s_1 , selected by simple random sampling, is stratified on the basis of a scalar auxiliary variable x observed on the units in the context of a first-phase simple random sample s_1 of size n_1 , $i \in s_1$: $s_1 = \bigcup_g s_{1g}$, where s_{1g} is the first phase sample of random size, n_{1g} , in stratum g . In the second-phase, simple random samples s_{2g} of fixed sizes n_{2g} are drawn from the first-phase samples s_{1g} of random sizes n_{1g} , $\sum_g n_{1g} = n_1$. In the second phase, simple random samples s_{2g} of fixed sizes n_{2g} are drawn from the first-phase samples s_{1g} independently. The assumption of fixed sizes n_{2g} , however, is inconsistent with the sampling procedure because n_{2g} is bounded above by the random variable n_{1g} which varies from 0 to $\min(n_1, N_g)$, where N_g is the number of population units in stratum g . Rao (1973) proposed an alternative

¹ Mike Hidirolou, Director of Survey Methods Division, Room D141, Methodology and Statistical Development Directorate, Cardiff Road, Newport, NP9-1XG, United Kingdom; J.N.K. Rao, School of Mathematics and Statistics, Ottawa, Ontario, K1S 5B6, Canada

sample allocation that avoids the difficulty with Neyman’s method of sample allocation at the second-phase. This method selects a fixed fraction ν_g of the sample units in the first-phase, i.e., $n_{2g} = \nu_g n_{1g}$, $0 < \nu_g \leq 1$. Cochran (1977, Chapter 12) studied ratio and regression estimation for the special case of simple random sampling in both phases.

More recently, Särndal, Swensson, and Wretman (1992, Chapter 9) allowed arbitrary sampling designs in both phases. Let π_{1i} and π_{1ij} be the first-order and second-order inclusion probabilities for the first-phase sample s_1 , and $\pi_{2i|s_1}$ and $\pi_{2ij|s_1}$ be the conditional first-order and second-order inclusion probabilities for the second phase sample s_2 , given s_1 . An unbiased estimator of the population total $Y = \sum_U y_i$ is given by

$$\hat{Y}_2 = \sum_{s_2} \frac{y_i}{\pi_{1i} \pi_{2i|s_1}} = \sum_{s_2} \frac{\dot{y}_i}{\pi_{2i|s_1}} \tag{1.1}$$

where $\dot{y}_i = y_i / \pi_{1i}$ and \sum_a denotes summation over units $i \in a$. This estimator is called the “double-expansion” estimator in analogy with the “expansion” (or Horvitz-Thompson (HT)) estimator for uni-phase sampling. Särndal et al (1992) derived an unbiased estimator of the variance of \hat{Y}_2 as

$$v_{HT}(\hat{Y}_2) = \sum \sum_{s_2} \frac{\Delta_{1ij}}{\pi_{ij}^*} \dot{y}_i \dot{y}_j + \sum \sum_{s_2} \frac{\Delta_{2ij|s_1}}{\pi_{2ij|s_1} \pi_{2i|s_1} \pi_{2j|s_1}} \dot{y}_i \dot{y}_j, \tag{1.2}$$

where $\pi_{ij}^* = \pi_{1ij} \pi_{2ij|s_1}$, $\Delta_{1ij} = \pi_{1ij} - \pi_{1i} \pi_{1j}$ and $\Delta_{2ij|s_1} = \pi_{2ij|s_1} - \pi_{2i|s_1} \pi_{2j|s_1}$.

The formulae (1.1) and (1.2) may be more compactly expressed as:

$$\hat{Y}_2 = \sum_{s_2} \frac{y_i}{\pi_i^*} \tag{1.3}$$

and

$$v_{HT}(\hat{Y}_2) = \sum \sum_{s_2} \frac{\Delta_{ij}^*}{\pi_{ij}^*} y_i y_j \tag{1.4}$$

where $\pi_i^* = \pi_{1i} \pi_{2i|s_1}$ and $\Delta_{ij}^* = \pi_{ij}^* - \pi_i^* \pi_j^*$ (Särndal et al. (1992), p. 347).

The variance estimator (1.4) (or equivalently (1.2)) has the same form as the HT variance estimator in single-phase sampling. For general single-phase designs with unequal inclusion probabilities, the latter variance estimator is known to be highly unstable and may take negative values (see Rao and Singh, 1973 and Cochran 1977, Chapter 10a). On the other hand, an alternative variance estimator, known as the Sen-Yates-Grundy (SYG) variance estimator, is relatively more stable than the HT variance estimator. It is therefore useful to develop SYG-type variance estimators under two-phase sampling.

Särndal et al. (1992) extended the unbiased estimator (1.1) to incorporate auxiliary data \mathbf{x} collected in the first phase, using Generalised Regression Estimator (GREG) estimation. They also obtained a Taylor linearization variance estimator of the form (1.2). GREG estimators calibrate to the first-phase estimators of \mathbf{x} totals; that is, the GREG estimator of Y is of the form $\sum_{s_2} w_i y_i$ with weights w_i satisfying $\sum_{s_2} w_i \mathbf{x}_i = \sum_{s_1} d_{1i} \mathbf{x}_i$. Hidiroglou and Särndal (1998) proposed GREG estimators based on calibrating to first-phase estimators, which are calibrated to known \mathbf{x}_1 totals; that is, $\sum_{s_2} w_i \mathbf{x}_i = \sum_{s_1} w_{1i} \mathbf{x}_i$ and $\sum_{s_2} w_{1i} \mathbf{x}_i = \sum_U \mathbf{x}_{1i}$. They also obtained a linearization variance estimator of the form (1.2); see also Estevao and Särndal (2002).

Binder, Babyak, Brodeur, Hidiroglou, and Jocelyn (2000) simplified the HT variance estimator (1.2) when a first-phase stratified simple random sample is re-stratified, using auxiliary data, \mathbf{x} collected in the first-phase, and simple random samples are then drawn without replacement from the second-phase strata to observe y . Kott and Stukel (1997) studied a similar two-phase design except that the first-phase sample is a stratified with replacement single-stage cluster sample. They proposed a reweighted “expansion” estimator that, in general, is different from the double

expansion estimator, and obtained a jackknife variance estimator. They also demonstrated that the proposed jackknife method is not effective for the double expansion estimator. Lee and Kim (2002) also studied a similar two-phase design except that the first-phase sample is a stratified single-stage cluster sample with clusters drawn by simple random sampling without replacement. The double expansion and reweighted expansion estimators are identical for their design. Lee and Kim (2002) developed a jackknife variance estimator that takes account of the sampling fractions for the two phases.

Even though all of the available auxiliary information is used for estimation of the total, the variance estimator is usually based on second-phase sample information. It is reasonable to ask whether the auxiliary information available for the elements not included in the second-phase sample could, and if so, should, be used more extensively for variance estimation as well. Dorfman (1994), Rao and Sitter (1995), Sitter (1997) and Axelson (1998) proposed to use the first-phase auxiliary data in variance estimation.

The paper is structured as follows. In section 2, we develop a SYG-type variance estimator of \hat{Y}_2 . We apply this result in Section 3 to the two-phase design of Binder et al. (2000) and obtain a non-negative variance estimator different from the HT-variance estimator of Binder et al. (2000). In section 4, we obtain a SYG-type variance estimator for the two-phase regression estimator of total that makes use of the first-phase auxiliary data.

2. SYG TYPE VARIANCE ESTIMATOR

The estimator \hat{Y}_2 is conditionally unbiased for the single phase estimator $\hat{Y}_1 = \sum_{s_1} \dot{y}_i$, given the first-phase sample s_1 , where $\dot{y}_i = y_i/\pi_{1i} = d_{1i}y_i$; i.e., $E(\hat{Y}_2 | s_1) = \hat{Y}_1$. Hence, it is unconditionally unbiased for the total $Y = \sum_U y_i$. The variance of \hat{Y}_2 is given by

$$\begin{aligned} v(\hat{Y}_2) &= E \left[V(\hat{Y}_2 | s_1) \right] + V \left[E(\hat{Y}_2 | s_1) \right] \\ &= E \left[V(\hat{Y}_2 | s_1) \right] + V(\hat{Y}_1). \end{aligned} \tag{2.1}$$

We can estimate the conditional variance $V(\hat{Y}_2 | s_1)$ in (2.1) by using the SYG variance estimator, provided the second-phase sample size is fixed for given s_1 (see Rao 1979). The SYG variance estimator is given by

$$v(\hat{Y}_2 | s_1) = \sum_{i < j \in s_2} \sum_{i < j \in s_2} \frac{(\pi_{2i|s_1} \pi_{2j|s_1} - \pi_{2ij|s_1})}{\pi_{2ij|s_1}} \left(\frac{\dot{y}_i}{\pi_{2i|s_1}} - \frac{\dot{y}_j}{\pi_{2j|s_1}} \right)^2. \tag{2.2}$$

It is conditionally unbiased for $V(\hat{Y}_2 | s_1)$ and hence unconditionally unbiased for $E \left[V(\hat{Y}_2 | s_1) \right]$.

Turning to the second term in (2.1), we have

$$V(\hat{Y}_1) = \sum_{i < j \in U} \sum_{i < j \in U} (\pi_{1i} \pi_{1j} - \pi_{1ij}) (\dot{y}_i - \dot{y}_j)^2 \tag{2.3}$$

provided the first-phase sample size is fixed. If the y_i 's were known for all $i \in s_1$, then the SYG variance estimator of $V(\hat{Y}_1)$ is given by

$$v(\hat{Y}_1) = \sum_{i < j \in s_1} \sum_{i < j \in s_1} \frac{\pi_{1i} \pi_{1j} - \pi_{1ij}}{\pi_{1ij}} (\dot{y}_i - \dot{y}_j)^2. \tag{2.4}$$

But y_i is only known for $i \in s_2$, so we estimate (2.4) using the second phase sample s_2 to get

$$v_2(\hat{Y}_1) = \sum_{i < j \in s_2} \sum \frac{\pi_{1i} \pi_{1j} - \pi_{1ij}}{\pi_{1ij} \pi_{2ij|s_1}} (\dot{y}_i - \dot{y}_j)^2 . \tag{2.5}$$

The variance estimator (2.5) is unbiased for $V(\hat{Y}_1)$. Hence, it follows from (2.1) that a SYG-type unbiased estimator of $V(\hat{Y}_1)$ is given by

$$v_{\text{SYG}}(\hat{Y}_2) = v(\hat{Y}_2 | s_1) + v_2(\hat{Y}_1), \tag{2.6}$$

where $v(\hat{Y}_2 | s_1)$ and $v_2(\hat{Y}_1)$ are given by (2.2) and (2.5) respectively.

Drawing analogy to the single-phase SYG, Chaudhuri (1994) gave a SYG-type estimator of $V(\hat{Y}_2)$, but his formula for $v(\hat{Y}_2 | s_1)$ seems to be incorrect as it uses $(\dot{y}_i - \dot{y}_j)^2$ instead of the correct term $(\dot{y}_i / \pi_{2i|s_1} - \dot{y}_j / \pi_{2j|s_1})^2$ given in (2.2).

The HT-type variance estimator, $v_{\text{HT}}(\hat{Y}_2)$, is valid for both fixed and non-fixed sample designs, unlike the SYG-type variance estimator (2.6). However, the SYG variance estimator remains valid for many commonly used two-phase designs, and in analogy to the uni-phase case it should be more stable relative to the HT variance estimator and remain nonnegative for several well-known probability proportional to size (PPS) designs. Rao and Singh (1973) provided extensive empirical evidence on the superiority of the SYG variance estimator over the HT variance estimator for uni-phase sampling.

3. TWO-PHASE SAMPLING FOR STRATIFICATION

3.1 General Set-Up

In this section, we evaluate the SYG variance estimator (2.6) for two-phase sampling for stratification. In the first-phase, a large sample s_1 of size n_1 is drawn according to a specified design with marginal inclusion probabilities π_{1i} and joint inclusion probabilities π_{1ij} . Using inexpensive auxiliary information collected on the units $i \in s_1$, the first-phase sample s_1 is stratified into $G(s_1)$ strata, denoted as s_{1g} ($g = 1, \dots, G(s_1)$), with m_{1g} elements in stratum g , ($\sum_g m_{1g} = n_1$). In the second-phase, a probability sample s_{2g} of size m_{2g} is drawn from s_{1g} , independently across g , and the characteristic of interest, y , is recorded. Note that the number of second-phase strata $G(s_1)$ and the sample sizes m_{1g} and m_{2g} depend on s_1 , although $G(s_1)$ may be predetermined, i.e., $G(s_1) \approx G$. For notational simplicity, we suppress dependence on s_1 .

Noting that $\pi_{2ij|s_1} = \pi_{2i|s_1} \pi_{2j|s_1}$ if $i \in s_{1g}$ and $j \in s_{1\ell}$ ($g \neq \ell$), $v(\hat{Y}_2 | s_1)$ reduces to

$$v(\hat{Y}_2 | s_1) = \sum_{g=1}^G \sum_{i < j \in s_{2g}} \sum \Delta_{2ij|s_{1g}} \left(\frac{\dot{y}_i}{\pi_{2i|s_{1g}}} - \frac{\dot{y}_j}{\pi_{2j|s_{1g}}} \right)^2 \tag{3.1}$$

where

$$\Delta_{2ij|s_{1g}} = \frac{\pi_{2i|s_{1g}} \pi_{2j|s_{1g}} - \pi_{2ij|s_{1g}}}{\pi_{2ij|s_{1g}}}. \tag{3.2}$$

Expression (3.1) is valid for general second-phase sampling within strata with conditional inclusion probabilities $\pi_{2i|s_{1g}}$ and $\pi_{2ij|s_{1g}}$, provided $\sum_{s_1} \pi_{2i|s_{1g}}$ is fixed for a given s_1 . In the special case of simple random sampling within second-phase strata, we have that $\pi_{2i|s_{1g}} = m_{2g} / m_{1g}$ and $\pi_{2ij|s_{1g}} = m_{2g} (m_{2g} - 1) / [m_{1g} (m_{1g} - 1)]$, and (3.1) reduces to

$$v(\hat{Y}_2 | s_1) = \sum_{g=1}^G m_{1g}^2 \left(\frac{1 - f_{2g}}{m_{2g}} \right) \left(\frac{1}{m_{2g} - 1} \right) \sum_{i < j \in s_{2g}} (\dot{y}_i - \dot{y}_j)^2, \tag{3.3}$$

where $f_{2g} = m_{2g} / m_{1g}$.

Now, using the Lagrange identity

$$\sum_{i < j=1}^m (z_i - z_j)^2 = m \sum_{i=1}^m (z_i - \bar{z})^2, \tag{3.4}$$

expression (3.3) reduces to

$$v(\hat{Y}_2 | s_1) = \sum_{g=1}^G \left(\frac{1 - f_{2g}}{m_{2g}} \right) m_{1g}^2 \left(\frac{1}{m_{2g} - 1} \right) \hat{S}_{2g\dot{y}}^2 \tag{3.5}$$

where $\hat{S}_{2g\dot{y}}^2$ is the sample mean square of the first-phase weighted values $\dot{y}_i = y_i / \pi_i$ for $i \in s_{2g}$. The second component of the HT variance estimator (1.2), under simple random sampling within second-phase strata, agrees with (3.5); see formula (9.4.8) of Särndal et al. (1992), p. 352. The component $v_2(\hat{Y}_1)$ in the SYG variance estimator (2.6) under simple random sampling within second-phase strata reduces to

$$\begin{aligned} v_2(\hat{Y}_1) &= \sum_{g=1}^G \frac{m_{1g} (m_{1g} - 1)}{m_{2g} (m_{2g} - 1)} \sum_{i < j \in s_{2g}} \Delta_{1ij} (\dot{y}_i - \dot{y}_j)^2 + \sum_{g < \ell=1}^G \sum_{i \in s_{2g}} \sum_{j \in s_{2\ell}} \frac{m_{1g} m_{1\ell}}{m_{2g} m_{2\ell}} \Delta_{1ij} (\dot{y}_i - \dot{y}_j)^2 \\ &=: v_2^{(1)}(\hat{Y}_1) + v_2^{(2)}(\hat{Y}_1), \end{aligned} \tag{3.6}$$

where

$$\Delta_{1ij} = \frac{\pi_{1i} \pi_{1j} - \pi_{1ij}}{\pi_{1ij}}. \tag{3.7}$$

Further reduction is not possible for general first-phase designs with inclusion probabilities π_{1i} and π_{1ij} .

Example 3.1: If the first-phase sample s_1 of size n_1 is selected by simple random sampling from a population U of size N , then $\pi_{1i} = n_1 / N$, $\pi_{1ij} = n_1 (n_1 - 1) / [N(N - 1)]$, and $\Delta_{1ij} = (1 - f_1) / (n_1 - 1)$. The two-phase estimator \hat{Y}_2 reduces to $N \sum_g w_{1g} \bar{y}_{2g}$ where $\bar{y}_{2g} = m_2^{-1} \sum_{s_{2g}} y_i$. Using the Lagrange identity (3.4) and the above values of π_{1i} and Δ_{1ij} , the first term on the right side of (3.6) reduces to

$$v_2^{(1)}(\hat{Y}_1) = \frac{N^2 (1 - f_1)}{n_1} \sum_{g=1}^G w_{1g} \frac{(m_{1g} - 1)}{n_1 - 1} \hat{S}_{2g\dot{y}}^2, \tag{3.8}$$

where $\hat{S}_{2g\dot{y}}^2$ is the sample mean square of y_i for $i \in s_2$. Further, the second term on the right side of (3.6) reduces to

$$\begin{aligned}
 v_2^{(2)}(\hat{Y}_1) &= \frac{N^2(1-f_1)}{n_1^2(n_1-1)} \left[\frac{1}{2} \sum_{g=1}^G \sum_{\ell=1}^G \frac{m_{1g} m_{1\ell}}{m_{2g} m_{2\ell}} \sum_{i \in s_{2g}} \sum_{j \in s_{2\ell}} (y_i - y_j)^2 - \sum_{g=1}^G m_{1g}^2 \frac{m_{2g}-1}{m_{2g}} \hat{S}_{2gy}^2 \right] \\
 &= \frac{N^2(1-f_1)}{n_1} \left[\sum_{g=1}^G \frac{(n_1 - m_{1g})}{n_1 - 1} w_{1g} (m_{2g} - 1) \hat{S}_{2gy}^2 + \frac{n_1}{n_1 - 1} \sum_{g=1}^G w_{1g} (\bar{y}_{2g} - \bar{y}_{2a})^2 \right],
 \end{aligned}
 \tag{3.9}$$

where $\bar{y}_{2a} = \hat{Y}_2 / N = \sum_{g=1}^G w_{1g} \bar{y}_{2g}$. The sum of (3.8) and (3.9) gives $v_2(\hat{Y}_1)$ as

$$v_2(\hat{Y}_1) = \sum_{g=1}^G (1 - \delta_g) w_{1g} \hat{S}_{2gy}^2 + \frac{n_1}{n_1 - 1} \sum_{g=1}^G w_{1g} (\bar{y}_{2g} - \bar{y}_{2a})^2
 \tag{3.10}$$

where $\delta_g = \frac{1}{m_{2g}} \frac{n_1 - m_{1g}}{n_1 - 1}$.

Särndal et al. (1992) simplified the first component on the right hand side of the HT variance estimator (1.2) for the special case of simple random sampling in the first-phase (without giving details) to obtain their (9.4.12) on p. 353. This formula agrees with our $v_2(\hat{Y}_1)$ given by (3.10).

3.2 Stratified Two-Phase Sampling

Suppose the population U is stratified into H strata, U_h , with N_h elements in the h -th stratum ($\sum_{h=1}^H N_h = N$). In the first phase, we draw simple samples s_{1h} independently from the first-phase strata U_h , and observe a scalar variable, x , for $i \in s_{1h}, h = 1, \dots, H$, where the size of s_{1h} is n_{1h} ($\sum_{h=1}^H n_{1h} = n_1$). We re-stratify the first phase sample $s_1 = \bigcup_{h=1}^H s_{1h}$ into G strata \tilde{s}_{1g} of sizes m_{1g} ($\sum_{h=1}^G m_{1g} = n_1$), using the auxiliary variable x observed in the first-phase. Simple random samples, s_{2g} of size m_{2g} are then drawn independently from the second phase strata \tilde{s}_{1g} ($g=1, \dots, G$).

For the above design, $\pi_{1i} = \frac{n_{1h}}{N_h}$ if $i \in s_{1h}$ and for $i \neq j$,

$$\pi_{1ij} = \begin{cases} \frac{n_{1h}(n_{1h}-1)}{N_h(N_h-1)} & \text{if } i \neq j \in s_{1h} \\ \frac{n_{1h} n_{1k}}{N_h N_k} & \text{if } i \in s_{1h}; j \in s_{1k}; h \neq k. \end{cases}
 \tag{3.11}$$

The two-phase estimator \hat{Y}_2 reduces to

$$\hat{Y}_2 = \sum_{h=1}^H \frac{N_h}{n_{1h}} \sum_{g=1}^G \frac{m_{1g}}{m_{2g}} \sum_{i \in s_{2gh}} y_i,
 \tag{3.12}$$

where $s_{2gh} = s_{1h} \cap s_{2g}$. Note that some of the s_{2gh} 's may be empty, in which case we set $\sum_{i \in s_{2gh}} y_i$ to 0 in (3.12).

Turning to variance estimation, the component $v(\hat{Y}_2 | s_1)$ is given by (3.5) with $\dot{y}_i = y_i(N_h/n_{1h})$ if $i \in s_{1h}$. To evaluate $v_2(\hat{Y}_1)$ given by (3.6), we need the Δ_{lij} -values. Using (3.11), we have

$$\Delta_{lij} = \begin{cases} \frac{1-f_{1h}}{n_{1h}-1} & \text{if } i, j \in \tilde{s}_{2h} \\ = 0 & \text{if } i \in \tilde{s}_{2h}, j \in \tilde{s}_{2k}, h \neq k, \end{cases} \quad (3.13)$$

where $\tilde{s}_{2h} = \bigcup_g s_{2gh}$ and $f_{1h} = n_{1h}/N_h$. Substituting the above values of Δ_{lij} in (3.6), the first component $v_2^{(1)}(\hat{Y}_1)$ reduces to

$$v_2^{(1)}(\hat{Y}_1) = \sum_{g=1}^G \frac{m_{1g}(m_{1g}-1)}{m_{2g}(m_{2g}-1)} \sum_{h=A_g} \left(\frac{N_h}{n_{1h}}\right)^2 \frac{1-f_{1h}}{n_{1h}-1} \sum_{i < j \in s_{2gh}} (y_i - y_j)^2, \quad (3.14)$$

where A_g is the set of first-phase strata h with at least two units in s_{2gh} ; the remaining first-phase strata do not contribute to $v_2^{(1)}(\hat{Y}_1)$. Using Lagrange's identity (3.4), expression (3.14) reduces to

$$v_2^{(1)}(\hat{Y}_1) = \sum_{g=1}^G \frac{m_{1g}(m_{1g}-1)}{m_{2g}(m_{2g}-1)} \sum_{h \in A_g} \left(\frac{N_h}{n_{1h}}\right)^2 \frac{1-f_{1h}}{n_{1h}-1} m_{2gh}(m_{2gh}-1) \hat{S}_{2ghy}^2, \quad (3.15)$$

where m_{2gh} is the number of units in s_{2gh} and \hat{S}_{2ghy}^2 is the sample mean square of the values y_i for $i \in s_{2gh}$.

We can express $v_2^{(2)}(\hat{Y}_1)$ as

$$v_2^{(2)}(\hat{Y}_1) = \sum_{g < \ell} \sum \frac{m_{1g}m_{1\ell}}{m_{2g}m_{2\ell}} \sum_{h \in U_{2g\ell}} \left(\frac{N_h}{n_{1h}}\right)^2 \frac{1-f_{1h}}{n_{1h}-1} \left\{ \sum_{i \in s_{2gh}} \sum_{j \in s_{2h\ell}} (y_i - y_j)^2 \right\}, \quad (3.16)$$

where $U_{2g\ell}$ is the set of first-phase strata h with at least one unit in both s_{2gh} and $s_{2h\ell}$. Further simplification of (3.16) is not possible unless $m_{2gh} \geq 2$ for all (gh) . The SYG-type variance estimator, $v(\hat{Y}_2)$, is now given by the sum of (3.5), (3.15), and (3.16), and it is always nonnegative.

We now consider the special case of $m_{2gh} \geq 2$ for all (gh) . In this case, $v_2^{(1)}(\hat{Y}_1)$ is given by (3.15) with \sum_{A_g} changed to $\sum_{h=1}^H$. Further, we can write $v_2^{(2)}(\hat{Y}_1)$ as

$$\begin{aligned} v_2^{(2)}(\hat{Y}_1) &= \frac{1}{2} \sum_{g=1}^G \sum_{\ell=1}^G \frac{m_{1g}m_{1\ell}}{m_{2g}m_{2\ell}} \sum_{i \in s_{2gh}} \sum_{j \in s_{2h\ell}} (\dot{y}_i - \dot{y}_j)^2 - \sum_{g=1}^G \left(\frac{m_{1g}}{m_{2g}}\right)^2 \Delta_{lij} \sum_{i < j \in s_{2g}} \sum (\dot{y}_i - \dot{y}_j)^2 \\ &= I - II. \end{aligned} \quad (3.17)$$

Following the steps used to get (3.15), we have

$$-II = - \sum_{g=1}^G \left(\frac{m_{1g}}{m_{2g}}\right)^2 \sum_{h=1}^H \left(\frac{N_h}{n_{1h}}\right)^2 \frac{1-f_{1h}}{n_{1h}-1} m_{2gh}(m_{2gh}-1) \hat{S}_{2ghy}^2. \quad (3.18)$$

Combining (3.15) with (3.18), we get

$$v_2^{(1)}(\hat{Y}_1) - \Pi = \sum_{g=1}^G \left(\frac{m_{1g}}{m_{2g}} \right)^2 \frac{1-f_{2g}}{m_{2g}-1} \sum_{h=1}^H \left(\frac{N_h}{n_{1h}} \right)^2 \frac{1-f_{1h}}{n_{1h}-1} m_{2gh} (m_{2gh}-1) \hat{S}_{2gh}^2. \tag{3.19}$$

Turning to the term I in (3.17), we can write

$$\begin{aligned} I &= \sum_{h=1}^H \frac{N_h^2}{n_{1h}^2} \frac{1-f_{1h}}{n_{1h}-1} \left\{ \frac{1}{2} \sum_{g=1}^G \sum_{\ell=1}^G \frac{m_{1g} m_{1\ell}}{m_{2g} m_{2\ell}} \sum_{i \in s_{2gh}} \sum_{j \in s_{2\ell h}} (\dot{y}_i - \dot{y}_j)^2 \right\} \\ &= \sum_{h=1}^H \frac{N_h^2}{n_{1h}^2} \frac{1-f_{1h}}{n_{1h}-1} \hat{n}_{1h} \sum_{g=1}^G \frac{m_{1g}}{m_{2g}} \sum_{i=1}^{m_{2gh}} (y_i - \bar{y}_{ah})^2 \end{aligned} \tag{3.20}$$

where

$$\hat{n}_{1h} = \sum_{g=1}^G \frac{m_{1g}}{m_{2g}} m_{2gh} \text{ and } \bar{y}_{ah} = \hat{n}_{1h}^{-1} \left(\sum_{g=1}^G \frac{m_{1g}}{m_{2g}} \sum_{k=1}^{m_{2gh}} y_k \right). \tag{3.21}$$

The variance estimator $v(\hat{Y}_2)$ is now given by the sum of (3.50), (3.19) and (3.20).

The H-T variance estimator of Binder et al. (2000) derived from (1.2), is different from our $v(\hat{Y}_2)$, but $v(\hat{Y}_2 | s_1)$, given by (3.5), is identical to Binder et al. (2000) formula. Binder et al.'s expression corresponding to (3.19) is given by

$$\sum_{g=1}^G \left(\frac{m_{1g}}{m_{2g}} \right)^2 \frac{1-f_{2g}}{m_{2g}-1} \sum_{h=1}^H \left(\frac{N_h}{n_{1h}} \right)^2 \frac{1-f_{1h}}{n_{1h}-1} m_{2g} \left\{ (m_{2gh}-1) S_{2gh}^2 + m_{2gh} \left(1 - \frac{m_{2gh}}{m_{2g}} \right) \bar{y}_{2gh}^2 \right\}. \tag{3.19}^*$$

where \bar{y}_{2gh} is the mean of y for s_{2gh} . Binder et al.'s expression corresponding to (3.20) is given by

$$\sum_{h=1}^H \frac{N_h^2}{n_{1h}^2} \frac{1-f_{1h}}{n_{1h}-1} \left[\sum_{g=1}^G \left(\frac{m_{1g}^2}{m_{2g}} \right) \frac{1-f_{2g}}{m_{2g}-1} \frac{m_{2gh}}{m_{2g}} \sum_{i=1}^{m_{2gh}} (y_i - \bar{y}_{ah})^2 + \hat{n}_{1h} \left(\frac{\hat{n}_{1h}-1}{n_{1h}} \right) \bar{y}_{ah}^2 \right]. \tag{3.20}^*$$

The variance estimator of Binder et al. (2000) is now given by the sum of (3.5), (3.19)* and (3.20)*. Note that the term $(\hat{n}_{1h} / n_{1h}) - 1$ can either be positive or negative.

4. VARIANCE ESTIMATOR OF THE REGRESSION ESTIMATOR INCORPORATING PHASE 1 AUXILIARY DATA

Auxiliary data may be available from different sources in two-phase sampling. We consider the case where auxiliary data are available from the frame U , as well as from the first-phase sample s_1 . Auxiliary data available from U are denoted as \mathbf{x}_{1i} , whereas those available from the first phase sample s_1 are denoted as \mathbf{x}_{2i} . The auxiliary data vector $\mathbf{x}_i = (\mathbf{x}'_{1i}, \mathbf{x}'_{2i})'$ contains data from both U and s_1 . Data (y_i, \mathbf{x}'_i) are collected using the second-phase sample s_2 . The regression estimator $\hat{Y}_{2,REG}$ of the total Y that incorporates auxiliary data from both phases is given by

$$\hat{Y}_{2,REG} = \hat{Y}_2 + (\mathbf{X}_{1,1} - \hat{\mathbf{X}}_{1,1})' \hat{\mathbf{B}}_1 + (\hat{\mathbf{X}}_1 - \hat{\mathbf{X}}_2)' \hat{\mathbf{B}}_2, \tag{4.1}$$

In (4.1), $\hat{Y}_2 = \sum_{s_2} y_i / \pi_i^*$, $\mathbf{X}_{1,1} = \sum_U \mathbf{x}_{1i}$ is the sum of the auxiliary \mathbf{x}_{1i} data available from the frame U and $\hat{\mathbf{X}}_{1,1} = \sum_{s_1} \mathbf{x}_{1i} / \pi_{1i}$, $\hat{\mathbf{X}}_1 = \sum_{s_1} \mathbf{x}_i / \pi_{1i}$ and $\hat{\mathbf{X}}_2 = \sum_{s_2} \mathbf{x}_i / \pi_i^*$. The regression vectors $\hat{\mathbf{B}}_1$ and $\hat{\mathbf{B}}_2$ are estimated by

$$\hat{\mathbf{B}}_1 = \left(\sum_{s_2} \mathbf{x}_{i1} \mathbf{x}'_{i1} / \lambda_i \pi_i^* \right)^{-1} \sum_{s_2} \mathbf{x}_{i1} y_i / \lambda_i \pi_i^*,$$

and

$$\hat{\mathbf{B}}_2 = \left(\sum_{s_2} \mathbf{x}_i \mathbf{x}'_i / \lambda_i \pi_i^* \right)^{-1} \sum_{s_2} (\mathbf{x}_i y_i / \lambda_i \pi_i^*).$$

The known constants λ_{i1} and λ_i are factors that yield different forms of the regression estimator of total. For example, if auxiliary data x_i are known only for $i \in s_1$ and λ_i is proportional to x_i , then (4.1) reduces to the two-phase ratio estimator $\hat{Y}_{2,RAT} = \hat{Y}_2 (\hat{X}_1 / \hat{X}_2)$.

We proceed to obtain the estimated variance for $\hat{Y}_{2,REG}$ by first linearizing it. To this end, the difference between $\hat{Y}_{2,REG}$ and Y is

$$\hat{Y}_{2,REG} - Y = (\hat{Y}_2 - Y) + (\mathbf{X}_{1,1} - \hat{\mathbf{X}}_{1,1})' \hat{\mathbf{B}}_1 + (\mathbf{X} - \hat{\mathbf{X}}_2)' \hat{\mathbf{B}}_2 + (\hat{\mathbf{X}}_1 - \mathbf{X})' \hat{\mathbf{B}}_2. \quad (4.2)$$

The linearized version of $\hat{Y}_{2,REG} - Y$ is given by

$$\begin{aligned} \hat{Y}_{\ell,2,REG} - Y &= (\hat{Y}_2 - Y) + (\mathbf{X}_{1,1} - \hat{\mathbf{X}}_{1,1})' \mathbf{B}_1 + (\mathbf{X} - \hat{\mathbf{X}}_2)' \mathbf{B}_2 + (\hat{\mathbf{X}}_1 - \mathbf{X})' \mathbf{B}_2 \\ &= \left(\sum_{s_1} \frac{e_{1i}}{\pi_{1i}} - \sum_U e_{1i} \right) + \left(\sum_{s_2} \frac{e_{2i}}{\pi_i^*} - \sum_{s_1} \frac{e_{2i}}{\pi_{1i}} \right) \end{aligned} \quad (4.3)$$

where $e_{1i} = y_i - \mathbf{x}'_{i1} \mathbf{B}_1$, $e_{2i} = y_i - \mathbf{x}'_i \mathbf{B}_2$, with $\mathbf{B}_1 = \left(\sum_U \mathbf{x}_{i1} \mathbf{x}'_{i1} / \lambda_i \right)^{-1} \sum_U \mathbf{x}_{i1} y_i / \lambda_i$, and

$\mathbf{B}_2 = \left(\sum_U \mathbf{x}_i \mathbf{x}'_i / \lambda_i \right)^{-1} \sum_U (\mathbf{x}_i y_i / \lambda_i)$. Using (4.3), the population variance of $\hat{Y}_{\ell,2,REG}$ is of the form (2.1). That is

$$\begin{aligned} V(\hat{Y}_{\ell,2,REG}) &= E[V(\hat{Y}_{\ell,2,REG} | s_1)] + V[E(\hat{Y}_{\ell,2,REG} | s_1)] \\ &= E \left[\sum_{i < j \in s_1} \sum (\pi_{2i|s_1} \pi_{2j|s_1} - \pi_{2ij|s_1}) \left(\frac{e_{2i}}{\pi_i^*} - \frac{e_{2j}}{\pi_j^*} \right)^2 \right] + \sum_{i < j \in U} (\pi_{1i} \pi_{1j} - \pi_{1ij}) \left(\frac{e_{1i}}{\pi_{1i}} - \frac{e_{1j}}{\pi_{1j}} \right)^2 \end{aligned} \quad (4.4)$$

It follows from (4.4) that the first component of $V(\hat{Y}_{\ell,2,REG})$ is estimated by

$$v(\hat{Y}_{\ell,2,REG} | s_1) = \sum_{i < j \in s_2} \frac{\pi_{2i|s_1} \pi_{2j|s_1} - \pi_{2ij|s_1}}{\pi_{2ij|s_1}} \left(\frac{\hat{e}_{2i}}{\pi_i^*} - \frac{\hat{e}_{2j}}{\pi_j^*} \right)^2. \quad (4.5)$$

where $e_{2i} = y_i - \mathbf{x}'_i \hat{\mathbf{B}}_2$. We proceed as in Axelson (1998) to estimate the second component of (4.4). That is, we substitute $e_{1i} = e_{2i} + d_i$, where $d_i = e_{1i} - e_{2i} = \mathbf{x}'_i \mathbf{B}_2 - \mathbf{x}'_{i1} \mathbf{B}_1$, into the second component of (4.4) to obtain

$$\begin{aligned}
 V \left[E \left(\hat{Y}_{\ell,2REG} \mid s_1 \right) \right] &= \sum_{i < j \in U} \sum_{i < j \in U} (\pi_{1i} \pi_{1j} - \pi_{1ij}) \left(\frac{e_{2i} + d_i}{\pi_{1i}} - \frac{e_{2j} + d_j}{\pi_{1j}} \right)^2 \\
 &= \sum_{i < j \in U} \sum_{i < j \in U} (\pi_{1i} \pi_{1j} - \pi_{1ij}) \left(\frac{e_{2i}}{\pi_{1i}} - \frac{e_{2j}}{\pi_{1j}} \right)^2 + \sum_{i < j \in U} \sum_{i < j \in U} (\pi_{1i} \pi_{1j} - \pi_{1ij}) \left(\frac{d_i}{\pi_{1i}} - \frac{d_j}{\pi_{1j}} \right)^2 \\
 &\quad + 2 \sum_{i < j \in U} \sum_{i < j \in U} (\pi_{1i} \pi_{1j} - \pi_{1ij}) \left(\frac{d_i}{\pi_{1i}} - \frac{d_j}{\pi_{1j}} \right) \left(\frac{e_{2i}}{\pi_{1i}} - \frac{e_{2j}}{\pi_{1j}} \right).
 \end{aligned} \tag{4.6}$$

Given that $d_i = \mathbf{x}'_i \mathbf{B}_2 - \mathbf{x}'_{1i} \mathbf{B}_1$ is available for all the units in the first phase sample s_1 , an unbiased estimator of

$$\sum_{i < j \in U} \sum_{i < j \in U} (\pi_{1i} \pi_{1j} - \pi_{1ij}) \left(\frac{d_i}{\pi_{1i}} - \frac{d_j}{\pi_{1j}} \right)^2 \text{ is } \sum_{i < j \in s_1} \sum_{i < j \in s_1} \frac{\pi_{1i} \pi_{1j} - \pi_{1ij}}{\pi_{1ij}} \left(\frac{d_i}{\pi_{1i}} - \frac{d_j}{\pi_{1j}} \right)^2.$$

Substituting $\hat{d}_i = \mathbf{x}'_i \hat{\mathbf{B}}_2 - \mathbf{x}'_{1i} \hat{\mathbf{B}}_1$ into (4.6), a variance estimator for $V \left[E \left(\hat{Y}_{\ell,2REG} \mid s_1 \right) \right]$ is

$$\begin{aligned}
 v_2 \left(\hat{Y}_{\ell,2REG} \right) &= \sum_{i < j \in s_2} \sum_{i < j \in s_2} \frac{\pi_{1i} \pi_{1j} - \pi_{1ij}}{\pi_{ij}^*} \left(\frac{\hat{e}_{2i}}{\pi_{1i}} - \frac{\hat{e}_{2j}}{\pi_{1j}} \right)^2 + \sum_{i < j \in s_1} \sum_{i < j \in s_1} \frac{\pi_{1i} \pi_{1j} - \pi_{1ij}}{\pi_{1ij}} \left(\frac{\hat{d}_i}{\pi_{1i}} - \frac{\hat{d}_j}{\pi_{1j}} \right)^2 \\
 &\quad + 2 \sum_{i < j \in s_2} \sum_{i < j \in s_2} \frac{\pi_{1i} \pi_{1j} - \pi_{1ij}}{\pi_{ij}^*} \left(\frac{\hat{d}_i}{\pi_{1i}} - \frac{\hat{d}_j}{\pi_{1j}} \right) \left(\frac{\hat{e}_{2i}}{\pi_{1i}} - \frac{\hat{e}_{2j}}{\pi_{1j}} \right).
 \end{aligned} \tag{4.7}$$

A SYG-type linearization estimator of $V(\hat{Y}_{2,REG})$ is now given by $v_{SYG}(\hat{Y}_{2,REG}) \doteq v(\hat{Y}_{\ell,2REG} \mid s_1) + v_2(\hat{Y}_{\ell,2REG})$, where

$v(\hat{Y}_{\ell,2REG} \mid s_1)$ and $v_2(\hat{Y}_{\ell,2REG})$ are given by (4.5) and (4.7) respectively. As Axelson (1998) points out, the “g” adjusted forms of this variance estimator are numerous. Letting the “g” factors for phase 1 and phase 2 be

$$g_{1i} = 1 + (\mathbf{X}_1 - \hat{\mathbf{X}}_{1,1})' \left(\sum_{s_1} \frac{\mathbf{x}_{1i} \mathbf{x}'_{1i}}{\pi_{1i} \lambda_{1i}} \right)^{-1} \frac{\mathbf{x}_{1i}}{\lambda_{1i}} \quad \text{and} \quad g_{2i} = 1 + (\hat{\mathbf{X}}_1 - \hat{\mathbf{X}}_2)' \left(\sum_{s_2} \frac{\mathbf{x}_i \mathbf{x}'_i}{\pi_i^* \lambda_i} \right)^{-1} \frac{\mathbf{x}_i}{\lambda_i},$$

respectively, a possible “g” adjusted form would be to replace \hat{d}_i by $\tilde{d}_i = g_{1i} \hat{d}_i$ and \hat{e}_{2i} by $\tilde{e}_{2i} = g_{2i} \hat{e}_{2i}$.

Example 4.1: Suppose that simple random sampling without replacement is used in both phases, and that an intercept term is included in the regression. Expression (4.7) becomes

$$v_{SYG}(\hat{Y}_{2,REG}) \doteq N^2 \left[\left(\frac{1}{n_2} - \frac{1}{n_1} \right) \hat{S}_{2e_2}^2 + \left(\frac{1}{n_1} - \frac{1}{N} \right) \left(\hat{S}_{2e_2}^2 + 2 \hat{S}_{2e_2,d} \right) + \left(\frac{1}{n_1} - \frac{1}{N} \right) \hat{S}_{1d}^2 \right] \tag{4.8}$$

where

$$\hat{S}_{1d}^2 = (1/(n-1)) \sum_{s_1} \left(\hat{d}_i - \bar{\hat{d}}_1 \right)^2, \quad \hat{S}_{2e_2}^2 = (1/(n_2-1)) \sum_{s_2} \left(\hat{e}_{2i} - \bar{\hat{e}}_2 \right)^2, \quad \hat{S}_{2e_2,d} = (1/(n_2-1)) \sum_{s_2} \left(\hat{e}_{2i} - \bar{\hat{e}}_2 \right) \left(\hat{d}_i - \bar{\hat{d}}_1 \right),$$

with $\bar{\hat{d}}_1 = (1/n_1) \sum_{s_1} \hat{d}_i$; $\bar{\hat{d}}_2 = (1/n_2) \sum_{s_2} \hat{d}_i$; $\bar{\hat{e}}_2 = (1/n_2) \sum_{s_2} \hat{e}_{2i}$; $e_{1i} = (y_i - \bar{y}_{s_2}) - (\mathbf{x}_{1i} - \bar{\mathbf{x}}_{1s_2})' \hat{\mathbf{B}}_1$ and

$e_{2i} = (y_i - \bar{y}_{s_2}) - (\mathbf{x}_i - \bar{\mathbf{x}}_{s_2})' \hat{\mathbf{B}}_2$. Both $\hat{\mathbf{B}}_1$ and $\hat{\mathbf{B}}_2$ are the least squares regression estimators adjusted to remove the intercept.

The corresponding $\tilde{v}_{SYG}(\hat{Y}_{2,Reg})$ version obtained by substituting $\tilde{d}_i = g_{1i} \hat{d}_i$ for \hat{d}_i and $\tilde{e}_{2i} = g_{2i} \hat{e}_{2i}$ for \hat{e}_{2i} in $v_{SYG}(\hat{Y}_{2,Reg})$ is

$$\tilde{v}_{SYG}(\hat{Y}_{2,Reg}) \doteq N^2 \left[\left(\frac{1}{n_2} - \frac{1}{n_1} \right) \tilde{S}_{2e_2}^2 + \left(\frac{1}{n_1} - \frac{1}{N} \right) (\tilde{S}_{2e_2}^2 + 2 \tilde{S}_{2e_2,d}) + \left(\frac{1}{n_1} - \frac{1}{N} \right) \tilde{S}_{1d}^2 \right], \quad (4.9)$$

where

$$\tilde{S}_{1d}^2 = (1/(n_1 - 1)) \sum_{s_1} (\tilde{d}_i - \bar{\tilde{d}}_1)^2, \quad \tilde{S}_{2e_2}^2 = (1/(n_2 - 1)) \sum_{s_2} (\tilde{e}_{2i} - \bar{\tilde{e}}_2)^2, \quad \tilde{S}_{2e_2,d} = (1/(n_2 - 1)) \sum_{s_2} (\tilde{e}_{2i} - \bar{\tilde{e}}_2)(\tilde{d}_i - \bar{\tilde{d}}_2),$$

$$\text{with } \bar{\tilde{d}}_1 = (1/n_1) \sum_{s_1} \tilde{d}_i, \quad \bar{\tilde{d}}_2 = (1/n_2) \sum_{s_2} \tilde{d}_i \quad \text{and} \quad \bar{\tilde{e}}_2 = (1/n_2) \sum_{s_2} \tilde{e}_{2i}.$$

If scalar auxiliary information x is available from the first-phase sample, expression (4.8) reduces to the one suggested by Dorfman (1994). This expression is given by

$$v_{SYG}(\hat{Y}_{2,REG}) = N^2 \left[\left(\frac{1}{n_2} - \frac{1}{n_1} \right) \hat{S}_{2e_2}^2 + \left(\frac{1}{n_1} - \frac{1}{N} \right) \hat{S}_{2e_2}^2 + \left(\frac{1}{n_1} - \frac{1}{N} \right) \hat{S}_{1x}^2 \hat{B}_2^2 \right], \quad (4.10)$$

$$\text{where } \hat{S}_{1x}^2 = (1/(n_1 - 1)) \sum_{s_1} (x_i - \bar{x}_1)^2.$$

Incorporation of the “g” terms into (4.10) yields

$$\tilde{v}_{SYG}(\hat{Y}_{2,Reg}) = N^2 \left\{ \left(\frac{1}{n_2} - \frac{1}{N} \right) \tilde{S}_{2e_2}^2 + \left(\frac{1}{n_1} - \frac{1}{N} \right) \hat{S}_{1x}^2 \hat{B}_2^2 + 2 \left(\frac{1}{n_1} - \frac{1}{N} \right) \tilde{S}_{2e_2,x} \hat{B}_2 \right\}, \quad (4.11)$$

$$\text{where } g_{2i} = 1 + \frac{n_2}{n_2 - 1} \frac{\bar{x}_1 - \bar{x}_2}{\hat{S}_{2x}^2} (x_i - \bar{x}_2).$$

REFERENCES

- Axelsson, M. (1998). Variance estimation for the generalised regression estimator under two-phase sampling-a modified approach. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 85-89.
- Binder, D.A., Babyak, C., Brodeur, M., Hidiroglou, M.A., and Jocelyn, W. (2000). Variance Estimation for Two-phase Stratified Sampling. *Canadian Journal of Statistics*, Vol. 28, No 4, 751-764.
- Chaudhuri, A., and Roy, D. (1994). Model assisted survey-sampling strategy in two phases. *Metrika*, 41, 355-362.
- Cochran, W.G. (1977). *Sampling Techniques*, 3rd Ed., New York : John Wiley.
- Dorfman, A.H. (1994). A note on variance estimation for the regression estimator in double sampling. *Journal of the American Statistical Association*, 89, 137-140.
- Hidiroglou, M.A., and Särndal, C.-E. (1998). Use of auxiliary information for two-phase sampling. *Survey Methodology*, 24, 11-20, Vol. 24, 11-20.
- Kott, P.S. and Stukel, D.M. (1997). Can the jackknife be used with a two-phase sample? *Survey Methodology*, 23, 81-90.

- Lee, H. and Kim, J.K. (2002). Jackknife Variance Estimation for Two-Phase Samples: An Application to the Study on Personnel Need for Special Education (SPeNSE). *Proceedings of the Section on Survey Research Methods*, 2002 Joint Statistical Meetings in New York.
- Neyman, J. (1938). Contribution To The Theory of Sampling Human Populations. *Journal of the American Statistical Association*, 33, 101-116.
- Rao, J.N.K. (1973). Double Sampling for Stratification and Analytical Surveys", *Biometrika*, **60**, 125-133.
- Rao, J.N.K. (1979), "On Deriving Mean Square Errors and their Nonnegative Unbiased Estimators", *Journal of the Statistical Association*, **17**, 125-136.
- Rao, J.N.K. and Singh, M.P (1973),"On the Choice of Estimator in Survey Sampling", *Australian J. Stat.*, **15**, 95-104.
- Särndal, C.-E., Swensson, B., and Wretman, Y. (1992). *Model Assisted Survey Sampling*, New York, Springer-Verlag.
- Sitter, R.R. (1997). Variance estimation for the regression estimator in two-phase sampling. *Journal of the American Statistical Association*, 92. 780-787.