Survey Methodology, December 2001
Vol. 27, No. 2, pp. 143-154
Statistics Canada, Catalogue No. 12-001

Double Sampling

M.A. Hidiroglou [1]

Abstract

The theory of double sampling is usually presented under the assumption that one of the samples is nested within the other. This type of sampling is called two-phase sampling. The first-phase sample provides auxiliary information ($\boldsymbol{x}$) that is relatively inexpensive to obtain, whereas the second-phase sample contains the variables of interest. The first-phase data are used in various ways: (a) to stratify the second-phase sample; (b) to improve the estimate using a difference, ratio or regression estimator; or (c) to draw a sub-sample of non-respondent units. However, it is not necessary for one of the samples to be nested in the other or selected from the same frame. The case of *non-nested* double sampling is dealt with in passing in the classical works on sampling (Des Raj 1968, Cochran 1977). This method is now used in several national statistical agencies.

This paper consolidates double sampling by presenting it in a unified manner. Several examples of surveys used at Statistics Canada illustrate this unification.

Key Words: Double sampling; Auxiliary data; Regression; Optimal.

1. Introduction

The theory of double-phase sampling is usually presented under the assumption that one of the samples is nested within the other. This type of sampling is called two-phase sampling. The first-phase sample provides auxiliary information ($\boldsymbol{x}$) that is relatively inexpensive to obtain, whereas the second-phase sample contains the variables of interest. The first-phase data are used in various ways: (a) to stratify the second-phase sample; (b) to improve the estimation by using a difference, ratio or regression estimator; or (c) to draw a sub-sample of non-respondent units. Two-phase sampling is a powerful and cost-effective technique with a long history. Neyman (1938) was first to propose it. Rao (1973) studied double sampling in the context of stratification and analytic studies. Cochran (1977) presented the basic results of two-phase sampling, including the simplest regression estimators for this type of sampling design. More recent work on the subject includes that of Breidt and Fuller (1993), who developed efficient estimation methods for three-phase sampling computations using auxiliary data. Chaudhuri and Roy (1994) focused on the optimal properties of simpler but well-known regression estimators of two-phase sampling. Hidiroglou and Särndal (1998) proposed estimators based on calibration and regression for two-phase sampling to account for the availability of auxiliary data at both levels of the sampling design.

Estimation for nested and non-nested double sampling has been treated separately in the survey literature. However, it is not necessary for one of the samples to be nested within the other, or even be selected from the same survey frame. This case will be termed *non-nested* double sampling. It has been briefly discussed in such classical books on sampling such as Des Raj (1968) and Cochran (1977). This method is used in several statistical agencies. For example, at Statistics Canada, the Canadian Survey of Employment, Payrolls and Hours (SEPH) is using this sampling procedure (Rancourt and Hidiroglou 1998). In this survey, two independent samples are drawn from two different frames, which nevertheless represent the same universe. The auxiliary data ($\boldsymbol{x}$), which includes the number of employees and the total amount of payrolls are obtained from a sample selected from a Canada Customs and Revenue Agency administrative data file. These same variables, together with the variables of interest (y), the number of hours worked by employees and summarised earnings, are collected from a sample drawn from the Statistics Canada Business Register. Another example described by Deville (1999) is the case of a household survey conducted at INSEE.

A single estimator can represent the overall estimation process, and the only difference is with respect to variance estimation. This paper is structured as follows. Part 2 sets out the notation. Part 3 describes how the double sampling procedures can be obtained from a single estimator. In Part 4, the estimated variance for the nested and non-nested calibration estimator is presented. Several practical examples are provided in Part 5. Finally, Part 6 contains a brief summary.

2. Notation

2.1 Nested Case

The population is represented by $U = \{1, ..., k, ..., N\}$. First, a probability sample $s_1 (s_1 \subseteq U)$ is selected from population U using a sampling design with inclusion

1. M.A. Hidiroglou, Business Survey Methods Division, R.H. Coats Building, 11th Floor, Section A, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6. E-mail: hidirog@statcan.ca.

probability of $\pi_{1k} = P(k \in s_1)$ for the k^{th} sampled unit in s_1. Given s_1, a second sample s_2 ($s_2 \subseteq s_1 \subseteq U$) is drawn from s_1 using a sample design with conditional inclusion probability $\pi_{2k|s_1} = P(k \in s_2 \mid s_1)$ for the k^{th} sampled unit in s_2. Note that the probabilities are conditional since it is assumed that s_1 is known. Figure 1 displays an example of nested sampling.

We assume that $\pi_{1k} > 0$ for all values $k \in U$ and that $\pi_{2k|s_1} > 0$ for all values $k \in s_1$. The weight of a sampled unit k will be denoted by $w_{1k} = 1/\pi_{1k}$ for the first-phase sample and $w_{2k} = 1/\pi_{2k|s_1}$ for the second phase sample. The overall sampling weight of a selected second-phase unit, $k \in s_2$, will therefore be $w_k^* = w_{1k}\, w_{2k}$.

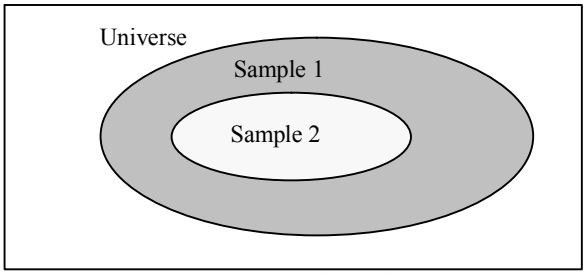

Figure 1. Nested Samples.

Let $\boldsymbol{x}$ denote the auxiliary data vector available with the first-phase sample, and $\boldsymbol{x}_k$ the value for unit k. We proceed as in Hidiroglou and Särndal (1998), that is, we divide $\boldsymbol{x}_k$ into two parts $\boldsymbol{x}_{1k}$ and $\boldsymbol{x}_{2k}$. The values of the data vector $\boldsymbol{x}_{1k}$ as assumed to be known for the entire population U, while the values of data vector $\boldsymbol{x}_{2k}$ are only known for the first-phase sample s_1.

2.2 Non-Nested Case

It is possible for the two samples to be drawn independently from the same frame or even from different (but equivalent) frames. Figures 2 and 3 provide examples of these non-nested cases.

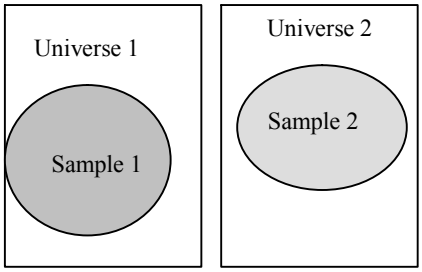

Figure 2. Two independent samples selected from different sample frames.

The non-nested case represented by Figure 3 is not considered in this paper. This case can be complicated for arbitrary sampling plans because it is necessary to compute joint inclusion probabilities between the two samples s_1 and s_2. This computation is simpler when the two samples s_1 and s_2 have been selected using a simple sampling design such as simple random sampling (with or without replacement). It is then possible to use Tam's results (1984) to obtain the required joint selection probabilities for the computation of the estimated variance for a given estimator of the total $Y = \sum_U y_k$.

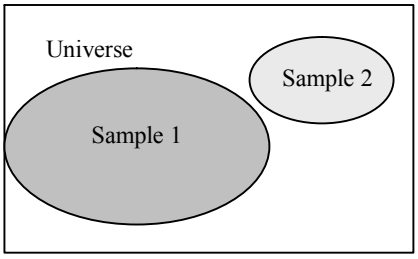

Figure 3. Two samples drawn independently from the same sample frame.

For the case that we will study, we assume that samples s_1 and s_2 are drawn independently from two different frames $U_1 = \{1, \ldots, k, \ldots, N_1\}$ and $U_2 = \{1, \ldots, k, \ldots, N_2\}$ (see Figure 2). The inclusion probabilities of a sampled unit k are respectively $\pi_{1k}^{(1)} = P(k \in s_1) > 0$ and $\pi_{2k}^{(2)} = P(k \in s_2) > 0$ for samples s_1 ($s_1 \subseteq U_1$) and s_2 ($s_2 \subseteq U_2$). The weight of unit k is $w_{1k}^{(1)} = 1/\pi_{1k}^{(1)}$ for the first sample s_1 and $w_{2k}^{(2)} = 1/\pi_{2k}^{(2)}$ for the second sample s_2. The superscripts (1) and (2) are used to differentiate between the selection probabilities of the samples drawn in the nested case. The sampling units may differ between the two frames, but these frames represent the same coverage. Examples of such sampling procedures were mentioned in the introduction and more details are provided in the second example given in section 5.3.

Let $\boldsymbol{x}_k = (\boldsymbol{x}_{1k}', \boldsymbol{x}_{2k}')'$, be an auxiliary data vector. We assume that $\boldsymbol{x}_{1k}^{(1)}$ is known for all units belonging to frame U_1, while $\boldsymbol{x}_k^{(1)}$ is only known for sample s_1. We collect $y_k^{(2)}$, $\boldsymbol{x}_k^{(2)}$ from sample s_2. The $\boldsymbol{x}$ data collected for corresponding units in samples s_1 and s_2 may differ. The degree in difference between the data values will vary according to the complexity of the sampling unit, and how much these units differ in concept between the two sampling frames. For « simpler » units the data reported for « similar » units in s_1 and s_2 should be equal or almost equal. Departures in the data similarity for the same units in s_1 and s_2 would most likely be due to the different questionnaire wording or due to different respondents filling in the questionnaires. Nevertheless, we assume that $\boldsymbol{X}_1 = \sum_{U_1} \boldsymbol{x}_{1k}^{(1)} = \sum_{U_2} \boldsymbol{x}_{1k}^{(2)}$ since U_1 and U_2 have the same coverage.

3. Optimal Estimator for Nested and Non-Nested Samples

In both cases, nested and non-nested, the objective is to estimate the population total $Y = \sum_U y_k$ where y_k represents the value of unit $k \in U$. An unbiased estimator of Y is $\hat{Y}_{\text{HT}} = \sum_{s_2} w_k^* y_k$, where $w_k^* = w_{1k}\, w_{2k}$ for the nested case and $w_k^* = w_{2k}^{(2)}$ for the non-nested case.

The sampling weight of a unit is modified by multiplying it by the calibration factor obtained using the various levels of the auxiliary data (universe, first-phase sample). The product is called a "calibration weight". Table 1 summarises the available data for the nested and non-nested cases, corresponding to Figures 1 and 2.

Table 1
Data Available for the Population and Samples

Set of Elements	Nested Case	Non-Nested Case
Population	$\boldsymbol{x}_{1k}$: known for $k \in U$	$\boldsymbol{x}_{1k}^{(1)}$: known for $k \in U_1$
First sample	$\boldsymbol{x}_k$: observed for $k \in s_1$	$\boldsymbol{x}_k^{(1)}$: observed for $k \in s_1$
Second sample	y_k, $\boldsymbol{x}_k$: observed for $k \in s_2$	$y_k^{(2)}$, $\boldsymbol{x}_k^{(2)}$: observed for $k \in s_2$

The following regression estimator is used to estimate the population total Y for nested and non-nested samples:

$$\hat{\hat{Y}}_{\text{REG}} = \hat{\hat{Y}}_{\text{HT}} + (\boldsymbol{X}_1 - \hat{\boldsymbol{X}}_1)' \boldsymbol{B}_1 + (\hat{\boldsymbol{X}} - \hat{\hat{\boldsymbol{X}}})' \boldsymbol{B}. \quad (3.1)$$

The various totals corresponding to the auxiliary data $\boldsymbol{x}$ and y – variable of interest given in equation (3.1) are provided in Table 2.

It is assumed that the variances, $V(\hat{\hat{Y}}_{\text{HT}})$, and covariances $\text{Cov}(\hat{\boldsymbol{X}}, \hat{\boldsymbol{X}}')$, $\text{Cov}(\hat{\boldsymbol{X}}_1, \hat{\boldsymbol{X}}')$, $\text{Cov}(\hat{\boldsymbol{X}}_1, \hat{\hat{\boldsymbol{X}}}')$, $\text{Cov}(\hat{\hat{Y}}_{\text{HT}}, \hat{\boldsymbol{X}}')$ and $\text{Cov}(\hat{\hat{Y}}_{\text{HT}}, \hat{\hat{\boldsymbol{X}}}')$, are known or estimable.

To simplify the notation, we drop the superscripts for the remainder of this section. The estimation of the parameters, $\boldsymbol{B}$ and $\boldsymbol{B}_1$ as well as of their associated variance, reflect that we have sampled differently for the nested and non-nested cases. The estimators of $\boldsymbol{B}$ and $\boldsymbol{B}_1$ are obtained by minimising the variance of $\hat{\hat{Y}}_{\text{REG}}$. This variance is:

$$\begin{aligned} V(\hat{\hat{Y}}_{\text{REG}}) &= V(\hat{\hat{Y}}_{\text{HT}}) + \boldsymbol{B}_1' V(\hat{\boldsymbol{X}}_1) \boldsymbol{B}_1 + \boldsymbol{B}' V(\hat{\boldsymbol{X}} - \hat{\hat{\boldsymbol{X}}}) \boldsymbol{B} \\ &\quad - 2 \text{Cov}(\hat{\hat{Y}}_{\text{HT}}, \hat{\boldsymbol{X}}_1') \boldsymbol{B}_1 + 2 \text{Cov}(\hat{\hat{Y}}_{\text{HT}}, (\hat{\boldsymbol{X}} - \hat{\hat{\boldsymbol{X}}})') \boldsymbol{B} \\ &\quad - 2 \boldsymbol{B}_1' \text{Cov}(\hat{\boldsymbol{X}}_1, (\hat{\boldsymbol{X}} - \hat{\hat{\boldsymbol{X}}})') \boldsymbol{B}. \end{aligned} \quad (3.2)$$

Deriving (3.2) with respect to $\boldsymbol{B}$ and $\boldsymbol{B}_1$, we obtain the following two equations:

$$\begin{aligned} &V(\hat{\boldsymbol{X}} - \hat{\hat{\boldsymbol{X}}}) \boldsymbol{B} + \text{Cov}((\hat{\boldsymbol{X}} - \hat{\hat{\boldsymbol{X}}}), \hat{\hat{Y}}_{\text{HT}}) \\ &\qquad - \text{Cov}((\hat{\boldsymbol{X}} - \hat{\hat{\boldsymbol{X}}}), \hat{\boldsymbol{X}}_1') \boldsymbol{B}_1 = \boldsymbol{0} \end{aligned} \quad (3.3)$$

and

$$\begin{aligned} &- \text{Cov}(\hat{\boldsymbol{X}}_1, (\hat{\boldsymbol{X}} - \hat{\hat{\boldsymbol{X}}})') \boldsymbol{B} - \text{Cov}(\hat{\boldsymbol{X}}_1, \hat{\hat{Y}}_{\text{HT}}) \\ &\qquad + V(\hat{\boldsymbol{X}}_1) \boldsymbol{B}_1 = \boldsymbol{0}. \end{aligned} \quad (3.4)$$

Solving the system of equations (3.3) and (3.4), we obtain the required parameters $\boldsymbol{B}$ and $\boldsymbol{B}_1$. That is:

$$\boldsymbol{B} = \boldsymbol{T}^{-1} \boldsymbol{H} \quad (3.5)$$

where

$$\begin{aligned} \boldsymbol{T} &= V(\hat{\boldsymbol{X}} - \hat{\hat{\boldsymbol{X}}}) \\ &\quad - (\text{Cov}(\hat{\boldsymbol{X}}_1, (\hat{\boldsymbol{X}} - \hat{\hat{\boldsymbol{X}}})'))' V^{-1}(\hat{\boldsymbol{X}}_1) (\text{Cov}(\hat{\boldsymbol{X}}_1, (\hat{\boldsymbol{X}} - \hat{\hat{\boldsymbol{X}}})')), \\ \boldsymbol{H} &= (\text{Cov}((\hat{\hat{\boldsymbol{X}}} - \hat{\boldsymbol{X}}), \hat{\hat{Y}}_{\text{HT}})) \\ &\quad + (\text{Cov}(\hat{\boldsymbol{X}}_1, (\hat{\boldsymbol{X}} - \hat{\hat{\boldsymbol{X}}})'))' V^{-1}(\hat{\boldsymbol{X}}_1) \text{Cov}(\hat{\boldsymbol{X}}_1, \hat{\hat{\boldsymbol{Y}}}_{\text{HT}}) \end{aligned}$$

and

$$\boldsymbol{B}_1 = \boldsymbol{T}_1^{-1} \boldsymbol{H}_1 \quad (3.6)$$

where

$$\boldsymbol{T}_1 = V(\hat{\boldsymbol{X}}_1),$$

and

$$\boldsymbol{H}_1 = \text{Cov}(\hat{\boldsymbol{X}}_1, \hat{\hat{Y}}_{\text{HT}}) + \text{Cov}(\hat{\boldsymbol{X}}_1, (\hat{\boldsymbol{X}} - \hat{\hat{\boldsymbol{X}}})')' \boldsymbol{B}.$$

Table 2
Sum of the Auxiliary Data $\boldsymbol{x}$ and y for Nested and Non-Nested Cases

Set of Elements	Nested Case	Non-nested Case
Population	$\boldsymbol{X}_1 = \sum_U \boldsymbol{x}_{1k}$	$\boldsymbol{X}_1 = \sum_{U_1} \boldsymbol{x}_{1k}^{(1)}$
First sample	$\hat{\boldsymbol{X}}_1 = \sum_{s_1} w_{1k} \boldsymbol{x}_{1k}$; $\hat{\boldsymbol{X}} = \sum_{s_1} w_{1k} \boldsymbol{x}_k$	$\hat{\boldsymbol{X}}_1 = \sum_{s_1} w_{1k} \boldsymbol{x}_{1k}^{(1)}$; $\hat{\boldsymbol{X}} = \sum_{s_1} w_{1k} \boldsymbol{x}_k^{(1)}$
Second sample	$\hat{\hat{\boldsymbol{X}}}_1 = \sum_{s_2} w_k^* \boldsymbol{x}_{1k}$; $\hat{\hat{\boldsymbol{X}}} = \sum_{s_2} w_k^* \boldsymbol{x}_k$	$\hat{\hat{\boldsymbol{X}}}_1 = \sum_{s_2} w_{2k} \boldsymbol{x}_{1k}^{(2)}$; $\hat{\hat{\boldsymbol{X}}} = \sum_{s_2} w_{2k} \boldsymbol{x}_k^{(2)}$
	$\hat{\hat{Y}}_{\text{HT}} = \sum_{s_2} w_k^* y_k$	$\hat{\hat{Y}}_{\text{HT}} = \sum_{s_2} w_{2k}^* y_k^{(2)}$

Result 1: An optimal regression estimator for the nested and non-nested samples is:

$$\hat{\bar{Y}}_{\text{OPT}} = \hat{\bar{Y}}_{\text{HT}} + (\boldsymbol{X}_1 - \hat{\boldsymbol{X}}_1)' \hat{\boldsymbol{B}}_{1,\text{OPT}} + (\hat{\boldsymbol{X}} - \hat{\bar{\boldsymbol{X}}})' \hat{\bar{\boldsymbol{B}}}_{\text{OPT}} \quad (3.7)$$

where

$$\hat{\bar{\boldsymbol{B}}}_{\text{OPT}} = \hat{\boldsymbol{T}}^{-1} \hat{\boldsymbol{H}} \quad (3.8)$$

and

$$\hat{\bar{\boldsymbol{B}}}_{1,\text{OPT}} = \hat{\boldsymbol{T}}_1^{-1} \hat{\boldsymbol{H}}_1. \quad (3.9)$$

$\hat{\boldsymbol{T}}_1$, $\hat{\boldsymbol{H}}_1$, $\hat{\boldsymbol{T}}$ and $\hat{\boldsymbol{H}}$ are the estimated values of $\boldsymbol{T}_1$, $\boldsymbol{H}_1$, $\boldsymbol{T}$ and $\boldsymbol{H}$, and they are obtained using a framework leading to the inference based on the sampling design. These values are dependent on the sample selection scheme. The population variance of $\hat{\bar{Y}}_{\text{OPT}}$ and its associated estimated variance depend on whether or not the samples are nested or non-nested. Since the regression vectors are optimal, it follows that the regression estimator $\hat{\bar{Y}}_{\text{OPT}}$ is also optimal. The optimal form has been discussed by Montanari (1987, 1998, and 2000) for the case of a single phase sampling design.

3.1 The Case of Nested Double Sampling

The theory for this case is developed using a conditional approach. Suppose that two parameters are given by $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, and that they are estimated by $\hat{\bar{\boldsymbol{\theta}}}_1$ and $\hat{\bar{\boldsymbol{\theta}}}_2$ from sample s_2. If we condition on the realised samples s_1, then the following well-known results hold:

(i) The expectation of $\hat{\bar{\boldsymbol{\theta}}}$ is $E(\hat{\bar{\boldsymbol{\theta}}}) = E_1 E_2 (\hat{\bar{\boldsymbol{\theta}}} \mid s_1)$, where E_2 denotes the expectation of $\hat{\bar{\boldsymbol{\theta}}}$ given s_1.

(ii) The variance of $\hat{\bar{\boldsymbol{\theta}}}$ is

$$V(\hat{\bar{\boldsymbol{\theta}}}) = E_1 V_2 (\hat{\bar{\boldsymbol{\theta}}} \mid s_1) + V_1 E_2 (\hat{\bar{\boldsymbol{\theta}}} \mid s_1). \quad (3.10)$$

(iii) The covariance between $\hat{\bar{\boldsymbol{\theta}}}_1$ and $\hat{\bar{\boldsymbol{\theta}}}_2$ is:

$$\text{Cov}(\hat{\bar{\boldsymbol{\theta}}}_1, \hat{\bar{\boldsymbol{\theta}}}_2') = E_1 \text{Cov}_2 ((\hat{\bar{\boldsymbol{\theta}}}_1, \hat{\bar{\boldsymbol{\theta}}}_2') \mid s_1) + \text{Cov}_1 (E_2(\hat{\bar{\boldsymbol{\theta}}}_1 \mid s_1), E_2(\hat{\bar{\boldsymbol{\theta}}}_2' \mid s_1)).$$

The various components of $\hat{\boldsymbol{T}}$, $\hat{\boldsymbol{H}}$, $\hat{\boldsymbol{T}}_1$ and of $\hat{\boldsymbol{H}}_1$ will be estimated assuming an arbitrary sampling design with a non-fixed sample size. The case of a fixed size sampling design follows easily as it is a special case of the arbitrary sampling design. Using expressions (i)–(iii), we can re-express the terms defining parameter $\boldsymbol{B}$ as:

$$\text{Cov}(\hat{\boldsymbol{X}}, \hat{\bar{\boldsymbol{X}}}') = \text{Cov}(\hat{\boldsymbol{X}}, \hat{\boldsymbol{X}}') = V(\hat{\boldsymbol{X}});$$

$$\text{Cov}(\hat{\bar{Y}}_{\text{HT}}, \hat{\bar{\boldsymbol{X}}}') = \text{Cov}(\hat{\bar{Y}}_{\text{HT}}, \hat{\boldsymbol{X}}');$$

$$V(\hat{\boldsymbol{X}} - \hat{\bar{\boldsymbol{X}}}) = E_1 \left[\sum\sum_{s_1} c_{2k\ell|s_1} \boldsymbol{x}_k \boldsymbol{x}_\ell' \right];$$

$$\text{Cov}\left[\hat{\boldsymbol{X}}_1, (\hat{\boldsymbol{X}} - \hat{\bar{\boldsymbol{X}}})' \right] = \boldsymbol{0};$$

and

$$\text{Cov}(\hat{\bar{\boldsymbol{X}}}, \hat{\bar{\boldsymbol{Y}}}_{\text{HT}}) = \text{Cov}(\hat{\mathbf{X}}, \hat{Y}_{\text{HT}}) + E_1 \left[\sum\sum_{s_1} c_{2k\ell|s_1} \boldsymbol{x}_k \boldsymbol{x}_\ell' \right]; \quad (3.11)$$

where $c_{2k\ell|s_1} = (\pi_{2k\ell|s_1} - \pi_{2k|s_1} \pi_{2\ell|s_1}) / \pi_k^* \pi_\ell^*$ and $\hat{Y}_{\text{HT}} = \sum_{s_1} y_k / \pi_{1k}$. The inclusion probabilities in these expressions are $\pi_{2k\ell|s_1} = \Pr(k, \ell \in s_2 \mid s_1)$ and $\pi_k^* = \pi_{1k} \pi_{2k|s_1}$.
We can express $\boldsymbol{B}$ more simply as:

$$\boldsymbol{B} = \left[E_1 \left(\sum\sum_{s_1} c_{2k\ell|s_1} \boldsymbol{x}_k \boldsymbol{x}_\ell' \right) \right]^{-1} E_1 \left[\sum\sum_{s_1} c_{2k\ell|s_1} \boldsymbol{x}_k \, y_\ell \right] \quad (3.12)$$

and the corresponding optimal estimator is given by:

$$\hat{\bar{\boldsymbol{B}}}_{\text{OPT}} = \left[\sum\sum_{s_2} \hat{c}_{2k\ell|s_1} \boldsymbol{x}_k \boldsymbol{x}_\ell' \right]^{-1} \left[\sum\sum_{s_2} \hat{c}_{2k\ell|s_1} \boldsymbol{x}_k y_\ell \right] \quad (3.13)$$

where $\hat{c}_{2k\ell|s_1} = c_{2k\ell|s_1} / \pi_{2k\ell|s_1}$.

The optimal regression estimator $\hat{\bar{\mathbf{B}}}_{1,\text{OPT}}$, is given by (3.9) with

$$\hat{\boldsymbol{T}}_1 = \hat{\boldsymbol{V}}(\hat{\boldsymbol{X}}_1)$$

and

$$\hat{\mathbf{H}}_1 = \hat{\text{C}}\text{ov}(\hat{\mathbf{X}}_1, \hat{\bar{Y}}_{\text{HT}}) + \hat{\text{C}}\text{ov}(\hat{\mathbf{X}}_1, \hat{\mathbf{X}}') \hat{\bar{\mathbf{B}}}_{\text{OPT}} - \hat{\text{C}}\text{ov}(\hat{\mathbf{X}}_1, \hat{\bar{\mathbf{X}}}') \hat{\bar{\mathbf{B}}}_{\text{OPT}}.$$

Each component defining $\hat{\mathbf{T}}_1$ and $\hat{\mathbf{H}}_1$ is estimated as follows. We first estimate $V(\hat{\boldsymbol{X}}_1) = \sum\sum_{s_1} c_{1k\ell} \boldsymbol{x}_{1k} \boldsymbol{x}_{1\ell}'$ by

$$\hat{V}(\hat{\boldsymbol{X}}_1) = \sum\sum_{s_1} \hat{c}_{1k\ell} \boldsymbol{x}_{1k} \boldsymbol{x}_{1\ell}' \quad (3.14)$$

where $c_{1k\ell} = (\pi_{1k\ell} - \pi_{1k} \pi_{1\ell}) / (\pi_{1k} \pi_\ell)$ and $\hat{c}_{1k\ell} = c_{1k\ell} / \pi_{1k\ell}$.

Next, since

$$\begin{aligned}\mathrm{Cov}(\hat{\boldsymbol{X}}_1, \hat{\hat{Y}}_{\mathrm{HT}}) &= E_1\,\mathrm{Cov}_2\,[(\hat{\boldsymbol{X}}_1, \hat{\hat{Y}}_{\mathrm{HT}}) \mid s_1] \\ &\quad + \mathrm{Cov}_1[E_2\,(\hat{\boldsymbol{X}}_1 \mid s_1), E_2\,(\hat{\hat{Y}}_{\mathrm{HT}} \mid s_1)] \\ &= \mathrm{Cov}_1\,(\hat{\boldsymbol{X}}_1, \hat{Y}_{\mathrm{HT}}) \\ &= \sum\sum\nolimits_{U_1} c_{1k\ell}\,\mathbf{x}_{1k}\,y_{1\ell} \qquad (3.15)\end{aligned}$$

we estimate $\mathrm{Cov}(\hat{\boldsymbol{X}}_1, \hat{\hat{Y}}_{\mathrm{HT}})$ by

$$\hat{\mathrm{C}}\mathrm{ov}(\hat{\boldsymbol{X}}_1,\ \hat{\hat{Y}}_{\mathrm{HT}}) = \sum\sum\nolimits_{s_2} c^*_{1k\ell}\,\boldsymbol{x}_{1k}\,y_\ell \qquad (3.16)$$

where

$$c^*_{1k\ell} = c_{1k\ell}/\pi^*_{k\ell}, \pi^*_{k\ell} = \pi_{1k\ell}\,\pi_{2k\ell|s_1},$$

$$\pi_{1k\ell} = \Pr\,(k, \ell \in s_1),$$

$$\pi_{2k\ell|s_1} = \Pr\,(k, \ell \in s_1)$$

and

$$\pi^*_k = \pi_{1k}\pi_{2k|s_1}.$$

Similarly,

$$\hat{\mathrm{C}}\mathrm{ov}(\hat{\boldsymbol{X}}_1,\ \hat{\hat{\boldsymbol{X}}}') = \sum\sum\nolimits_{s_2} c^*_{1k\ell}\,\boldsymbol{x}_{1k}\,y'_\ell \qquad (3.17)$$

and

$$\hat{\mathrm{C}}\mathrm{ov}(\hat{\boldsymbol{X}}_1,\ \hat{\boldsymbol{X}}') = \sum\sum\nolimits_{s_1} \hat{c}_{1k\ell}\,\boldsymbol{x}_{1k}\,x'_\ell. \qquad (3.18)$$

Hence, in the case of nested double sampling the optimal estimator of $\boldsymbol{B}_1$ is given by:

$$\hat{\boldsymbol{B}}_{1,\mathrm{OPT}} = (\hat{V}\,(\hat{\boldsymbol{X}}_1))^{-1}\begin{bmatrix}\hat{\mathrm{C}}\mathrm{ov}(\hat{\boldsymbol{X}}_1,\ \hat{\hat{Y}}_{\mathrm{HT}}) \\ \quad + (\hat{\mathrm{C}}\mathrm{ov}(\hat{\boldsymbol{X}}_1, \hat{\boldsymbol{X}}') \\ \quad - \hat{\mathrm{C}}\mathrm{ov}(\hat{\boldsymbol{X}}_1, \hat{\hat{\boldsymbol{X}}}'))\,\hat{\hat{\boldsymbol{B}}}_{\mathrm{OPT}}\end{bmatrix} \qquad (3.19)$$

where the components of $\hat{\boldsymbol{B}}_{1,\mathrm{OPT}}$ have been defined by expressions (3.14) – (3.18).

The optimal form of estimators $\hat{\boldsymbol{B}}_{1,\mathrm{OPT}}$ and $\hat{\hat{\boldsymbol{B}}}_{\mathrm{OPT}}$ has its advantages and disadvantages. One of the biggest advantages of the optimal form, as reported by Cassady and Valliant (1993), Rao (1994), and Montanari (2000), is that it has good conditional inference properties (by conditioning on the auxiliary variable $\boldsymbol{x}$). As Montanari (2000) observed, the asymptotic optimality of $\hat{Y}_{\mathrm{OPT}}$ is strictly a property based on the sampling design and achieved conditionally on the finite population. The biggest disadvantage of the optimal estimator is that it requires the computation of joint inclusion probabilities.

We can, however, use the optimal form, and express it more simply for several sampling designs. For sampling designs where the sample selection is with unequal probability and without replacement, we can bypass the computation of the joint probability by approximating the exact variance. Several authors, including Hartley and Rao (1962), Deville (1999), Berger (1998), Rósen (2000) and Brewer (2000) proposed such approximating procedures. Recently, Tillé (2001) proposed the following approximation for the estimated variance of $\hat{Y}_{\mathrm{HT}} = \sum_s y_k/\pi_k$ in the context of single-phase sampling, where

$$\begin{aligned}\hat{V}(\hat{Y}_{\mathrm{HT}}) &= \sum\nolimits_s \frac{c_k}{\pi_k^2}(y_k - y_k^*)^2 \\ &= \sum\nolimits_s c_k\left(\frac{y_k}{\pi_k} - \tilde{y}\right)^2. \qquad (3.20)\end{aligned}$$

Here, c_k is the variable used as the approximation, $y_k^* = \pi_k \sum_s c_\ell\, y_\ell/\pi_\ell / \sum_s c_\ell$, $\tilde{y} = y_k^*/\pi_k$, and π_k is the probability of selection of a given unit k. Tillé (2001) provided several examples of the c_k values for various sampling schemes.

This formula is exact in the case of a stratified simple sampling design drawn without replacement in each stratum U_h $(h = 1, ..., L)$ of population U. Let k be a sampled unit in sample s_h from stratum U_h, then $c_k = n_h/(n_h - 1)\,(1 - n_h/N_h)$ if $k \in U_h$ and 0 otherwise, and $\pi_k = n_h/N_h$ if $k \in U_h$ and 0 otherwise. This gives us the exact estimated variance, $\hat{V}\,(\hat{Y}_{\mathrm{HT}}) = \sum_{h=1}^{L} N_h^2(1 - n_h/N_h)\,\sum_{s_h}(y_k - \bar{y}_h)^2/n_h(n_h - 1)$. The formula is also exact in the case of a stratified sampling design where the sample is selected with replacement. Here $c_k = 1$ for all units belonging to stratum U_h and zero otherwise. Using this approximation, the double sums appearing in $\hat{\hat{\boldsymbol{B}}}_{\mathrm{OPT}}$ and $\hat{\boldsymbol{B}}_{1,\mathrm{OPT}}$ can be expressed as simple sums. Hidiroglou and Särndal (1998) bypassed the problem of double sums in estimating $\boldsymbol{B}$ and $\boldsymbol{B}_1$ by proposing the GREG estimator, $\hat{\hat{Y}}_{\mathrm{GREG}}$, for a nested two-phase sampling design. Their estimator is given by:

$$\hat{\hat{Y}}_{\mathrm{GREG}} = \hat{Y}_{\mathrm{HT}} + (\boldsymbol{X}_1 - \hat{\boldsymbol{X}}_1)'\,\hat{\boldsymbol{B}}_{1,\mathrm{GREG}} + (\hat{\boldsymbol{X}} - \hat{\hat{\boldsymbol{X}}})'\,\hat{\hat{\boldsymbol{B}}}_{\mathrm{GREG}}$$

where

$$\hat{\hat{\boldsymbol{B}}}_{\mathrm{GREG}} = \left(\sum\nolimits_{s_2} \frac{\tilde{w}_{1k}\,w_{2k}\boldsymbol{x}_k\,\boldsymbol{x}_k'}{\sigma_{2k}^2}\right)^{-1}\sum\nolimits_{s_2}\frac{\tilde{w}_{1k}\,w_{2k}\boldsymbol{x}_k\,y_k}{\sigma_{2k}^2}, \qquad (3.21)$$

$$\begin{aligned}\hat{\boldsymbol{B}}_{1,\mathrm{GREG}} &= \left(\sum\nolimits_{s_1}\frac{w_{1k}\,\boldsymbol{x}_{1k}\,\boldsymbol{x}_{1k}'}{\sigma_{1k}^2}\right)^{-1} \\ &\quad \left\{\sum\nolimits_{s_2}\frac{w_k^*\boldsymbol{x}_{1k}\,y_k}{\sigma_{1k}^2} + \sum\nolimits_{s_1}\frac{w_{1k}\,\boldsymbol{x}_{1k}\,\boldsymbol{x}_k'}{\sigma_{1k}^2}\hat{\hat{\boldsymbol{B}}}_{\mathrm{GREG}}\right. \\ &\quad \left. - \sum\nolimits_{s_2}\frac{w_k^*\,\boldsymbol{x}_{1k}\,\boldsymbol{x}_k'}{\sigma_{1k}^2}\hat{\hat{\boldsymbol{B}}}_{\mathrm{GREG}}\right\} \qquad (3.22)\end{aligned}$$

with $\{\sigma_{1k}^2 : k \in s_1\}$ and $\{\sigma_{2k}^2 : k \in s_2\}$ being predetermined positive factors.

Estimators $\hat{\tilde{\boldsymbol{B}}}_{\text{GREG}}$ and $\hat{\boldsymbol{B}}_{1,\text{GREG}}$ can be justified either by assuming different regression models for each phase or by using two successive calibrations. For the calibration approach, calibration weights $\tilde{w}_{1k}$ associated with the first-phase are first obtained, and they satisfy the calibration equation $\sum_{s_1} \tilde{w}_{1k}\, \boldsymbol{x}_{1k} = \sum_U \boldsymbol{x}_{1k}$. These calibration weights can be expressed as the product of sample weights w_{1k} and a calibration factor g_{1k} where:

$$g_{1k} = 1 + \left(\sum_U \boldsymbol{x}_{1k} - \sum_{s_1} w_{1k}\, \boldsymbol{x}_{1k}\right)' \left(\sum_{s_1} w_{1k}\, \frac{\boldsymbol{x}_{1k}\, \boldsymbol{x}_{1k}'}{\sigma_{1k}^2}\right)^{-1} \frac{\boldsymbol{x}_{1k}}{\sigma_{1k}^2} \qquad (3.23)$$

for $k \in s_1$.

The first-phase calibration weights $\tilde{w}_{1k}$ are then used as initial weights to compute the overall calibration weights $\tilde{w}_k^*$. These overall calibration weights satisfy the second-phase calibration equation $\sum_{s_2} \tilde{w}_{1k}^*\, \boldsymbol{x}_k = \sum_{s_1} \tilde{w}_{1k}\, \boldsymbol{x}_k$. The estimator of the total, $\hat{\tilde{Y}}_{\text{GREG}}$, can be expressed as the sum of the product of the overall calibration weight $\tilde{w}_k^*$ and the associated $y-$value, that is $\hat{\tilde{Y}}_{\text{GREG}} = \sum_{s_2} \tilde{w}_k^*\, y_k$. The calibrated overall weights can be expressed as $\tilde{w}_k^* = w_k^*\, g_k^*$, where $g_k^* = g_{1k}\, g_{2k}$. Here, g_{1k} is given by (3.23), while g_{2k} is equal to

$$g_{2k} = 1 + \left(\sum_{s_1} \tilde{w}_{1k}\, \boldsymbol{x}_k - \sum_{s_2} \tilde{w}_{1k}\, w_{2k}\, \boldsymbol{x}_k\right)' \left(\sum_{s_1} \frac{\tilde{w}_{1k}\, w_{2k}\, \boldsymbol{x}_k\, \boldsymbol{x}_k'}{\sigma_{2k}^2}\right)^{-1} \frac{\boldsymbol{x}_k}{\sigma_{2k}^2} \qquad (3.24)$$

for $k \in s_2$.

Comment: The estimators of $\hat{\boldsymbol{B}}_{1,\text{GREG}}$ (3.21) and $\hat{\tilde{\boldsymbol{B}}}_{\text{GREG}}$ (3.22) correspond to Hidiroglou and Särndal's (1998) *additive case* and have the same form as the optimal regression estimators $\hat{\boldsymbol{B}}_{1,\text{OPT}}$ (3.8) and $\hat{\tilde{\boldsymbol{B}}}_{\text{OPT}}$ (3.9). Indeed, the components of the estimator of $\boldsymbol{B}$ are obtained by respectively estimating $\boldsymbol{T}$ by $(\sum_{s_2} w_{1k} w_{2k}\, \boldsymbol{x}_k\, \boldsymbol{x}_k' / \sigma_{2k}^2)$ and $\boldsymbol{H}$ by $\sum_{s_2} w_{1k} w_{2k}\, \boldsymbol{x}_k\, y_k / \sigma_{2k}^2$. The second terms of $\boldsymbol{H}$ and $\boldsymbol{T}$ are exactly equal to zero. Similarly, to estimate $\boldsymbol{B}_1$, the component $\boldsymbol{T}_1$ is estimated by $\sum_{s_1} w_{1k}\, \boldsymbol{x}_{1k}\, \boldsymbol{x}_{1k}' / \sigma_{1k}^2$, while $\boldsymbol{H}_1$ is estimated by

$$\sum_{s_2} \frac{w_k^*\, \boldsymbol{x}_{1k}\, y_{2k}}{\sigma_{1k}^2} + \left(\sum_{s_1} \frac{w_{1k}\, \boldsymbol{x}_{1k} \boldsymbol{x}_k'}{\sigma_{1k}^2} - \sum_{s_2} \frac{w_k^*\, \boldsymbol{x}_{1k}\, \boldsymbol{x}_k'}{\sigma_{1k}^2}\right)' \hat{\tilde{\boldsymbol{B}}}_{\text{GREG}}.$$

The estimated variance of $\hat{\tilde{Y}}_{\text{GREG}} = \hat{\tilde{Y}}_{\text{HT}} + (\boldsymbol{X}_1 - \hat{\boldsymbol{X}}_1)'\, \hat{\boldsymbol{B}}_{1,\text{GREG}} + (\hat{\boldsymbol{X}} - \hat{\tilde{\boldsymbol{X}}})'\, \hat{\tilde{\boldsymbol{B}}}_{\text{GREG}}$ is presented in Hidiroglou and Särndal (1998).

Comment: The efficiency of the GREG, as stated in Särndal, Swensson and Wretman (1992), requires that the proposed model be correct. Furthermore, if the sample size is large enough, optimal estimators are more efficient (Rao 1994) than the GREG. However, if the sample size is relatively small, one disadvantage of the optimal form OPT is that it is generally less stable and more complex to compute than the GREG. Furthermore, an additional consequence of a relatively small sample size, as reported by Särndal (1996), and illustrated by simulation by Montanari (2000), is that if the sample size is relatively small, then the optimal form is not significantly more efficient than the GREG. It is even possible for the estimated variance to be greater than that associated with the GREG.

3.2 The Case of Non-nested Double Sampling

Deville (1999) considered the non-nested case (Figure 2) by assuming that $\boldsymbol{x}_{2k}$ is known for s_1 and s_2. The optimal regression estimator is:

$$\hat{\tilde{Y}}_{\text{OPT}} = \hat{\tilde{Y}}_{\text{HT}} + (\hat{\boldsymbol{X}}_2 - \hat{\tilde{\boldsymbol{X}}}_2)'\, \hat{\tilde{\boldsymbol{B}}}_{2,\text{OPT}} \qquad (3.25)$$

where $\hat{\tilde{Y}}_{\text{HT}} = \sum_{s_2} w_{2k}\, y_k^{(2)}$, $\hat{\boldsymbol{X}}_2 = \sum_{s_1} w_{1k}\, \boldsymbol{x}_{2k}^{(1)}$, $\hat{\tilde{\boldsymbol{X}}}_2 = \sum_{s_2} w_{2k}\, \boldsymbol{x}_{2k}^{(2)}$. The optimal estimator for $\boldsymbol{B}_2 = (\sum_{U_2} \boldsymbol{x}_{2k}\, \boldsymbol{x}_{2k}')^{-1} \sum_{U_2} \boldsymbol{x}_{2k}\, y_k$ is $\hat{\tilde{\boldsymbol{B}}}_{2,\text{OPT}} = (\hat{V}(\hat{\boldsymbol{X}}_2) + \hat{V}(\hat{\tilde{\boldsymbol{X}}}_2))^{-1} \hat{\text{Cov}}(\hat{\tilde{Y}}_{\text{HT}}, \hat{\boldsymbol{X}}_2')$ if the two sampling frames U_1 and U_2 are independent. The form of the variance and of the covariance terms defining $\hat{\tilde{\boldsymbol{B}}}_{2,\text{OPT}}$ depends on the sampling design of s_1 and s_2.

The accuracy of the estimator of $\boldsymbol{X}_2$ can be improved by minimising the variance of $\tilde{\boldsymbol{X}}_2 = \boldsymbol{A}_2\, \hat{\boldsymbol{X}}_2 + (\boldsymbol{I} - \boldsymbol{A}_2)\, \hat{\tilde{\boldsymbol{X}}}_2$ yielding, $\boldsymbol{A}_2 = (V(\hat{\boldsymbol{X}}_2) + V(\hat{\tilde{\boldsymbol{X}}}_2))^{-1}\, V(\hat{\boldsymbol{X}}_2)$. Assuming that $V(\hat{\boldsymbol{X}}_2)$ is approximately a multiple of $V(\hat{\tilde{\boldsymbol{X}}}_2)$, that is $V(\hat{\boldsymbol{X}}_2) \doteq \alpha_2\, V(\hat{\tilde{\boldsymbol{X}}}_2)$, we obtain $\boldsymbol{A}_2 \doteq \boldsymbol{I}/(1+\alpha_2)$ where $\boldsymbol{I}$ is the identity matrix has the same dimension as the covariance matrix $V(\hat{\boldsymbol{X}}_2)$. The optimal value of α_2 is obtained by minimising the variance of $\tilde{\boldsymbol{X}}_2$. A sub-optimal but adequate choice, suggested by Deville (1999), for α_2 is $\alpha_2 = n_1/(n_1+n_2)$, where n_1 and n_2 are the respective sizes of samples s_1 and s_2. Note that Korn and Graubart (1999) also made the same suggestion in the context of combining two totals estimated from two different sources. Substituting $\tilde{\boldsymbol{X}}_2$ in place of $\hat{\boldsymbol{X}}_2$ in expression (3.25), yields

$$\tilde{\boldsymbol{X}}_2 - \hat{\boldsymbol{X}}_2 = (\hat{\boldsymbol{X}}_2 - \hat{\tilde{\boldsymbol{X}}}_2)/(1+\alpha_2). \qquad (3.26)$$

The estimator of the population total Y, is:

$$\tilde{\tilde{Y}}_{\text{OPT}} = \hat{\tilde{Y}}_{\text{HT}} + (\tilde{\boldsymbol{X}}_2 - \hat{\tilde{\boldsymbol{X}}}_2)'\, \tilde{\tilde{\boldsymbol{B}}}_{2,\text{OPT}} \qquad (3.27)$$

where

$$\tilde{\tilde{\boldsymbol{B}}}_{2,\text{OPT}} = -[\hat{V}(\tilde{\boldsymbol{X}}_2 - \hat{\tilde{\boldsymbol{X}}}_2)]^{-1}\, \hat{\text{Cov}}\,(\hat{\tilde{Y}}_{\text{HT}}, (\tilde{\boldsymbol{X}}_2 - \hat{\tilde{\boldsymbol{X}}}_2)'). \qquad (3.28)$$

If (3.26) is substituted in (3.28), we can re-express $\tilde{\tilde{\boldsymbol{B}}}_{2,\text{OPT}}$ as:

$$\tilde{\tilde{\boldsymbol{B}}}_{2,\text{OPT}} = [\hat{V}(\hat{\tilde{\boldsymbol{X}}}_2)]^{-1}\, \hat{\text{Cov}}\,(\hat{\tilde{Y}}_{\text{HT}}, \hat{\tilde{\boldsymbol{X}}}_2). \qquad (3.29)$$

Comment: We see that $\hat{\tilde{Y}}_{\text{OPT}}$ (3.25) is exactly equal to $\tilde{\tilde{Y}}_{\text{OPT}}$ (3.27). This implies that there was no advantage in using a better estimator of $\boldsymbol{X}_2$ to estimate Y. However, the estimator $\tilde{\tilde{\boldsymbol{B}}}_{2,\text{OPT}}$ associated with $\tilde{\tilde{Y}}_{\text{OPT}}$ looks more like a traditional regression estimator than the regression estimator $\hat{\tilde{\boldsymbol{B}}}_{2,\text{OPT}}$ associated with $\hat{\tilde{Y}}_{\text{OPT}}$.

Note that the GREG estimator for the case where $\tilde{\boldsymbol{X}}_2$ is used instead of $\tilde{\boldsymbol{X}}_2$ is:

$$\tilde{\tilde{Y}}_{\text{GREG}} = \hat{\tilde{Y}}_{\text{HT}} + (\tilde{\boldsymbol{X}}_2 - \hat{\tilde{\boldsymbol{X}}}_2)' \tilde{\tilde{\boldsymbol{B}}}_{2,\text{GREG}} \tag{3.30}$$

where

$$\tilde{\tilde{\boldsymbol{B}}}_{2,\text{GREG}} = \left(\sum_{s_2} w_{2k} \frac{\boldsymbol{x}_k^{(2)} \boldsymbol{x}_k^{\prime(2)}}{\sigma_{2k}^2} \right)^{-1} \sum_{s_2} w_{2k} \frac{\boldsymbol{x}_k^{(2)} y_k^{(2)}}{\sigma_{2k}^2}.$$

Furthermore, if we also know $\boldsymbol{x}_{1k}^{(1)}$ for $k \in U_1$ where $\boldsymbol{X}_1 = \sum_{U_1} \boldsymbol{x}_{1k}^{(1)}$, we can consider the regression estimator

$$\tilde{\tilde{Y}}_{\text{OPT}} = \hat{\tilde{Y}}_{\text{HT}} + (\boldsymbol{X}_1 - \tilde{\boldsymbol{X}}_1)' \tilde{\boldsymbol{B}}_{1,\text{OPT}} + (\tilde{\boldsymbol{X}} - \hat{\tilde{\boldsymbol{X}}})' \tilde{\tilde{\boldsymbol{B}}}_{\text{OPT}}. \tag{3.31}$$

We obtain $\tilde{\boldsymbol{X}}$ by minimising the linear combination $\boldsymbol{A}\hat{\boldsymbol{X}} + (\boldsymbol{I} - \boldsymbol{A})\hat{\tilde{\boldsymbol{X}}}$ and $V(\hat{\boldsymbol{X}}) \doteq \alpha V(\hat{\tilde{\boldsymbol{X}}})$. The difference between $\tilde{\boldsymbol{X}}$ and $\tilde{\tilde{\boldsymbol{X}}}$ can be re-expressed as

$$\tilde{\boldsymbol{X}} - \hat{\tilde{\boldsymbol{X}}} = (\hat{\boldsymbol{X}} - \hat{\tilde{\boldsymbol{X}}})/(1 + \alpha). \tag{3.32}$$

Given that s_1 and s_2 are independent samples, it can be shown that:

$$\tilde{\tilde{\boldsymbol{B}}}_{\text{OPT}} = [\hat{V}(\hat{\tilde{\boldsymbol{X}}})]^{-1} \hat{\text{Cov}}(\hat{\tilde{\boldsymbol{X}}}, \hat{\tilde{Y}}_{\text{HT}}) \tag{3.33}$$

and that

$$\tilde{\boldsymbol{B}}_{1,\text{OPT}} = [\hat{V}(\hat{\tilde{\boldsymbol{X}}}_1)]^{-1} [\hat{\text{Cov}}(\hat{\tilde{\boldsymbol{X}}}_1, \hat{\tilde{Y}}_{\text{HT}})]. \tag{3.34}$$

The components of $\tilde{\tilde{\boldsymbol{B}}}_{\text{OPT}}$ are estimated by:

$$\hat{V}(\hat{\tilde{\boldsymbol{X}}}) = \sum\sum_{s_2} \hat{c}_{2k\ell}\, \boldsymbol{x}_k^{(2)} \boldsymbol{x}_\ell^{\prime(2)} \tag{3.35}$$

and

$$\hat{\text{Cov}}(\hat{\tilde{\boldsymbol{X}}}, \hat{\tilde{Y}}_{\text{HT}}) = \sum\sum_{s_2} \hat{c}_{2k\ell}\, \boldsymbol{x}_k^{(2)} y_\ell^{(2)} \tag{3.36}$$

whereas the components of $\tilde{\tilde{\boldsymbol{B}}}_{1,\text{OPT}}$ are estimated by:

$$\hat{V}(\hat{\tilde{\boldsymbol{X}}}_1) = \sum\sum_{s_2} \hat{c}_{2k\ell}\, \boldsymbol{x}_{1k}^{(2)} \boldsymbol{x}_{1\ell}^{\prime(2)} \tag{3.37}$$

and

$$\hat{\text{Cov}}(\hat{\tilde{\boldsymbol{X}}}_1, \hat{\tilde{Y}}_{\text{HT}}) = \sum\sum_{s_2} \hat{c}_{2k\ell}\, \boldsymbol{x}_{1k}^{(2)} y_\ell^{(2)} \tag{3.38}$$

where

$$\hat{c}_{2k\ell} = \frac{\pi_{2k\ell} - \pi_{2k}\,\pi_{2\ell}}{(\pi_{2k\ell})\,(\pi_{2k}\,\pi_{2\ell})}.$$

Approximation (3.20) can also be used to estimate the terms (3.35)–(3.38). The corresponding GREG which bypasses the computation of joint selection probabilities is given by:

$$\tilde{\tilde{Y}}_{\text{GREG}} = \hat{\tilde{Y}}_{\text{HT}} + (\boldsymbol{X}_1 - \tilde{\boldsymbol{X}}_1)' \tilde{\boldsymbol{B}}_{1,\text{GREG}} + (\tilde{\boldsymbol{X}} - \hat{\tilde{\boldsymbol{X}}})' \tilde{\tilde{\boldsymbol{B}}}_{\text{GREG}} \tag{3.39}$$

where $\boldsymbol{X}_1 = \sum_{U_1} \boldsymbol{x}_{1k}^{(1)}$, $\hat{\boldsymbol{X}}_1 = \sum_{s_1} w_{1k}\, \boldsymbol{x}_{1k}^{(1)}$, $\hat{\boldsymbol{X}} = \sum_{s_1} w_{1k}\, \boldsymbol{x}_k^{(1)}$ and $\hat{\tilde{\boldsymbol{X}}} = \sum_{s_2} w_{2k}\, \boldsymbol{x}_k^{(2)}$.

GREG-type regression estimators in equation (3.39) are estimated by

$$\tilde{\boldsymbol{B}}_{1,\text{GREG}} = \left(\sum_{s_2} w_{2k} \frac{\boldsymbol{x}_{1k}^{(2)} \boldsymbol{x}_{1k}^{\prime(2)}}{\sigma_{1k}^2} \right)^{-1} \sum_{s_2} w_{2k} \frac{\boldsymbol{x}_{1k}^{(2)} y_k^{(2)}}{\sigma_{1k}^2} \tag{3.40}$$

and

$$\tilde{\tilde{\boldsymbol{B}}}_{\text{GREG}} = \left(\sum_{s_2} w_{2k} \frac{\boldsymbol{x}_k^{(2)} \boldsymbol{x}_k^{\prime(2)}}{\sigma_{2k}^2} \right)^{-1} \sum_{s_2} w_{2k} \frac{\boldsymbol{x}_k^{(2)} y_k^{(2)}}{\sigma_{2k}^2}. \tag{3.41}$$

4. Estimator of the Variance for the Optimal Regression Estimator

4.1 Nested Double Sampling

Recall that the optimal regression estimator of Y is given by

$$\hat{\tilde{Y}}_{\text{OPT}} = \hat{\tilde{Y}}_{\text{HT}} + (\boldsymbol{X}_1 - \hat{\boldsymbol{X}}_1)' \hat{\boldsymbol{B}}_{1,\text{OPT}} + (\hat{\boldsymbol{X}} - \hat{\tilde{\boldsymbol{X}}})' \hat{\tilde{\boldsymbol{B}}}_{\text{OPT}}. \tag{4.1}$$

To obtain the estimated variance of (4.1), we re-express the terms associated with the y – variable within $\hat{\tilde{\boldsymbol{B}}}_{\text{OPT}}$ and $\hat{\boldsymbol{B}}_{1,\text{OPT}}$ as a simple sums instead of double sums. Montanari (1998) described this algebra for an arbitrary single-phase sampling design. Following Montanari (1998), and adapting the single-phase algebra to double sampling, we obtain:

$$\begin{aligned} \hat{\tilde{\boldsymbol{B}}}_{\text{OPT}} &= \left[\sum\sum_{s_2} \hat{c}_{2k\ell|s_1} \boldsymbol{x}_k \boldsymbol{x}_\ell' \right]^{-1} \left[\sum\sum_{s_2} \hat{c}_{2k\ell|s_1} \boldsymbol{x}_k y_\ell \right] \\ &= \left[\sum\sum_{s_2} \hat{c}_{2k\ell|s_1} \boldsymbol{x}_k \boldsymbol{x}_\ell' \right]^{-1} \left[\sum_{s_2} \frac{a_{2k}}{\pi_k^*} y_k \right] \end{aligned} \tag{4.2}$$

where

$$a_{2k} = \frac{1 - \pi_{2k|s_1}}{\pi_k^*} \boldsymbol{x}_k + \sum_{\substack{\ell \neq k \\ \ell \in s_2}} \frac{(\pi_{2k\ell|s_1} - \pi_{2k|s_1}\pi_{2\ell|s_1})}{\pi_{2k\ell|s_1}\pi_\ell^*} \boldsymbol{x}_\ell.$$

We approximate $\hat{\boldsymbol{B}}_{1,\text{OPT}}$ given by (3.15) by $[\hat{V}(\hat{\tilde{\boldsymbol{X}}}_1)]^{-1} [\hat{\text{Cov}}(\hat{\tilde{\boldsymbol{X}}}_1, \hat{\tilde{Y}}_{\text{HT}})]$, and hence,

$$\hat{\boldsymbol{B}}_{1,\mathrm{OPT}} \doteq [\hat{V}(\hat{\boldsymbol{X}}_1)]^{-1}\left[\hat{\mathrm{Cov}}(\hat{\boldsymbol{X}}_1, \hat{Y}_{\mathrm{HT}})\right] = \left[\sum\sum_{s_1} \hat{c}_{1k\ell}\, \boldsymbol{x}_{1k}\, \boldsymbol{x}'_{1\ell}\right]^{-1}\left[\sum_{s_1} \frac{a_{1k}}{\pi_{1k}}\, y_k\right] \quad (4.3)$$

where

$$a_{1k} = \frac{1-\pi_{1k}}{\pi_{1k}}\boldsymbol{x}_{1k} + \sum_{\substack{\ell\neq k\\ \ell\in s_1}} \frac{(\pi_{1k\ell} - \pi_{1k}\ \pi_{1\ell})}{\pi_{1\ell}\pi_{1k\ell}}\boldsymbol{x}_{1\ell}.$$

By substituting (4.2) and (4.3) in (4.1), and by subtracting the population total Y, we get:

$$\hat{\ddot{Y}}_{\mathrm{OPT}} - Y \doteq \left(\sum_{s_1} g_{1k}\frac{y_k}{\pi_{1k}} - \sum_U y_k\right) + \left(\sum_{s_2} g_{2k}\frac{y_k}{\pi^*_k} - \sum_{s_1}\frac{y_k}{\pi_{1k}}\right) \quad (4.4)$$

where

$$g_{1k} = 1 + (\boldsymbol{X}_1 - \hat{\boldsymbol{X}}_1)'\,(\hat{V}(\hat{\boldsymbol{X}}_1))^{-1}\, a_{1k} \ \text{ for } k \in s_1 \quad (4.5)$$

and

$$g_{2k} = 1 + (\hat{\boldsymbol{X}} - \hat{\hat{\boldsymbol{X}}})'\,(\hat{V}(\hat{\hat{\boldsymbol{X}}}))^{-1}\, a_{2k} \ \text{ for } k \in s_2. \quad (4.6)$$

Result 2: The estimated variance of $\hat{\ddot{Y}}_{\mathrm{OPT}}$ defined by equation (4.1) is:

$$\hat{V}(\hat{\ddot{Y}}_{\mathrm{OPT}}) = \sum\sum_{s_2} c^*_{1k\ell}\, g_{1k}\, g_{1\ell}\, e_{1k}\, e_{1\ell} + \sum\sum_{s_2} c^*_{2k\ell}\, g_{2k}\, g_{2\ell}\, e_{2k}\, e_{2\ell} \quad (4.7)$$

where

$$c^*_{1k\ell} = \frac{(\pi_{1k\ell} - \pi_{1k}\ \pi_{1\ell})}{\pi^*_{k\ell}\ \pi_{1k}\ \pi_{1\ell}};$$

$$c^*_{2k\ell} = \frac{(\pi_{2k\ell|s_1} - \pi_{2k|s_1}\ \pi_{2\ell|s_1})}{\pi_{2k\ell|s_1}\ \pi^*_k\ \pi^*_\ell};$$

$$e_{1k} = y_k - \boldsymbol{x}'_{1k}\ \hat{\boldsymbol{B}}_{1,\mathrm{OPT}};$$

and

$$e_{2k} = y_k - \boldsymbol{x}'_k\ \hat{\hat{\boldsymbol{B}}}_{\mathrm{OPT}}.$$

4.2 Non-Nested Double Sampling

We obtain the estimated variance of $\tilde{\ddot{Y}}_{\mathrm{OPT}}$ by using the following approximation.

$$\tilde{\ddot{Y}}_{\mathrm{OPT}} = \hat{\hat{Y}}_{\mathrm{HT}} + (\boldsymbol{X}_1 - \tilde{\boldsymbol{X}}_1)'\,\tilde{\boldsymbol{B}}_{1,\mathrm{OPT}} + (\tilde{\boldsymbol{X}} - \hat{\hat{\boldsymbol{X}}})'\,\tilde{\tilde{\boldsymbol{B}}}_{\mathrm{OPT}} = \ddot{Y}_{\mathrm{OPT}} + O_p(n_1^{-1/2}) \quad (4.8)$$

where

$$\ddot{Y}_{\mathrm{OPT}} = \hat{\hat{Y}}_{\mathrm{HT}} + (\boldsymbol{X}_1 - \tilde{\boldsymbol{X}}_1)'\,\boldsymbol{B}_{1,\mathrm{OPT}} + (\tilde{\boldsymbol{X}} - \hat{\hat{\boldsymbol{X}}})'\,\boldsymbol{B}_{\mathrm{OPT}}. \quad (4.9)$$

Decomposing $\ddot{Y}_{\mathrm{OPT}}$ into more elementary components, we have that:

$$\ddot{Y}_{\mathrm{OPT}} = \hat{\hat{Y}}_{\mathrm{HT}} + \left(\boldsymbol{X}_1 - \frac{\hat{\boldsymbol{X}}_1 + \alpha\,\hat{\hat{\boldsymbol{X}}}_1}{1+\alpha}\right)'\boldsymbol{B}_{1,\mathrm{OPT}} + \frac{(\tilde{\boldsymbol{X}} - \hat{\hat{\boldsymbol{X}}})'}{1+\alpha}\,\boldsymbol{B}_{\mathrm{OPT}}$$

$$= \left(\hat{\hat{Y}}_{\mathrm{HT}} - \frac{1}{1+\alpha}(\alpha\,\hat{\boldsymbol{X}}'_1\,\boldsymbol{B}_{1,\mathrm{OPT}} + \hat{\hat{\boldsymbol{X}}}'\,\boldsymbol{B}_{1,\mathrm{OPT}})\right) + \left(\boldsymbol{X}'_1\,\boldsymbol{B}_{1,\mathrm{OPT}} - \frac{1}{1+\alpha}(\hat{\boldsymbol{X}}'_1\boldsymbol{B}_{1,\mathrm{OPT}} - \hat{\hat{\boldsymbol{X}}}'\boldsymbol{B}_{\mathrm{OPT}})\right). \quad (4.10)$$

The variance of $\ddot{Y}_{\mathrm{OPT}}$ is:

$$V(\ddot{Y}_{\mathrm{OPT}}) = V\left(\hat{\hat{Y}}_{\mathrm{HT}} - \frac{1}{1+\alpha}(\alpha\,\hat{\boldsymbol{X}}'_1\,\boldsymbol{B}_{1,\mathrm{OPT}} + \hat{\hat{\boldsymbol{X}}}'\,\boldsymbol{B}_{\mathrm{OPT}})\right) + \frac{1}{(1+\alpha)^2}\begin{bmatrix} \alpha\,\boldsymbol{B}'_{1,\mathrm{OPT}}\,V(\hat{\boldsymbol{X}}_1)\,\boldsymbol{B}_{1,\mathrm{OPT}} \\ +\boldsymbol{B}'_{\mathrm{OPT}}\,V(\hat{\hat{\boldsymbol{X}}})\,\boldsymbol{B}_{\mathrm{OPT}} \\ +2\alpha\,(\boldsymbol{B}'_{\mathrm{OPT}}\,V(\hat{\hat{\boldsymbol{X}}})\,\tilde{\boldsymbol{B}}'_{1,\mathrm{OPT}} \\ +\mathrm{Cov}(\hat{\boldsymbol{X}}_1, \hat{\hat{\boldsymbol{X}}}'))\,\boldsymbol{B}_{\mathrm{OPT}} \end{bmatrix}. \quad (4.11)$$

Result 3: The estimated variance of $\tilde{\tilde{Y}}_{\mathrm{OPT}}$, $\hat{V}(\tilde{\tilde{Y}}_{\mathrm{OPT}})$, defined by equation (4.8) is approximately equal to:

$$\hat{V}\left(\hat{\hat{Y}}_{\mathrm{HT}} - \frac{1}{1+\alpha}\,(\alpha\,\hat{\boldsymbol{X}}'_1\,\tilde{\boldsymbol{B}}_{1,\mathrm{OPT}} + \hat{\hat{\boldsymbol{X}}}'\,\tilde{\tilde{\boldsymbol{B}}}_{\mathrm{OPT}})\right) + \frac{1}{(1+\alpha)^2}\begin{bmatrix} \alpha\,\tilde{\boldsymbol{B}}'_{1,\mathrm{OPT}}\,\hat{V}(\hat{\boldsymbol{X}}_1)\,\tilde{\boldsymbol{B}}_{1,\mathrm{OPT}} \\ +\,\tilde{\tilde{\boldsymbol{B}}}'_{\mathrm{OPT}}\,\hat{V}(\hat{\hat{\boldsymbol{X}}})\,\tilde{\tilde{\boldsymbol{B}}}_{\mathrm{OPT}} \\ +\,2\alpha\,(\tilde{\tilde{\boldsymbol{B}}}'_{\mathrm{OPT}}\,\hat{V}(\hat{\hat{\boldsymbol{X}}})\,\tilde{\boldsymbol{B}}_{1,\mathrm{OPT}} \\ +\,\hat{\mathrm{Cov}}(\hat{\boldsymbol{X}}_1, \hat{\hat{\boldsymbol{X}}}')\,\tilde{\tilde{\boldsymbol{B}}}_{\mathrm{OPT}}) \end{bmatrix}. \quad (4.12)$$

Computation of the first term of (4.12) is based on the residuals $y_k - (\alpha\,\boldsymbol{x}'_{1k}\,\tilde{\boldsymbol{B}}_{1,\mathrm{OPT}} + \boldsymbol{x}'_k\,\tilde{\tilde{\boldsymbol{B}}}_{\mathrm{OPT}})/(1+\alpha)$. The computation of the other terms of (4.12) is mainly based on the estimated variances of $\hat{X}_1$ and of $\hat{X}$, as well as on their estimated covariances. We can use the approximation of the variance, as described by Tillé (2001), and suitably adapt it to estimate the required covariances.

5. Some Specific Examples

Three traditional examples for double sampling are presented for the two cases (nested and non-nested). Furthermore, we briefly describe how two major business surveys carried out by Statistics Canada use double sampling.

5.1 Nested Sampling

Example 1: Let us assume that a simple random sample s_1 of size n_1 is selected from a population U of size N. The sample is stratified into L strata s_{1h} each of size n_{1h}. Random samples s_{2h} of size n_{2h} are then selected without replacement in each stratum s_{1h}. The estimator of the total is $\hat{\hat{Y}}_{\text{EXP}} = N\sum_{h=1}^{L} p_{1h}\,\bar{y}_{2h} = N\,\bar{y}_{2,st}$, where $p_{1h} = n_{1h}/n_1$. Using (4.7), we can show that the estimated variance of $\hat{\hat{Y}}_{\text{EXP}}$, $\hat{V}(\hat{\hat{Y}}_{\text{EXP}})$, consists of the sum of $\hat{V}_1(\hat{\hat{Y}}_{\text{EXP}})$ and $\hat{V}_2(\hat{\hat{Y}}_{\text{EXP}})$ corresponding to the first and second phases of the sampling design. Thus:

$$\hat{V}(\hat{\hat{Y}}_{\text{EXP}}) = \hat{V}_1(\hat{\hat{Y}}_{\text{EXP}}) + \hat{V}_2(\hat{\hat{Y}}_{\text{EXP}})$$

where

$$\hat{V}_1(\hat{\hat{Y}}_{\text{EXP}}) = N^2\frac{(1-f_1)}{n_1}\sum_{h=1}^{L} p_{1h}\left[\begin{array}{l}(1-a_h)\;\hat{S}^2_{2yh}\\ \\ +\dfrac{n_1}{n_1-1}(\bar{y}_{2h}-\bar{y}_{2,st})^2\end{array}\right];$$

$$\hat{V}_2(\hat{\hat{Y}}_{\text{EXP}}) = N^2\sum_{h=1}^{L}\frac{(1-f_{2h})}{n_{2h}}p_{1h}^2\,\hat{S}^2_{2yh};$$

and

$$a_h = \frac{(n_1-n_{1h})}{n_{2h}(n_1-1)};\; f_1 = \frac{n_1}{N};\; f_{2h} = \frac{n_{2h}}{n_{1h}};$$

$$\hat{S}^2_{2yh} = \frac{1}{n_{2h}-1}\sum_{s_{2h}}(y_k-\bar{y}_{2h})^2;$$

$$\bar{y}_{2h} = \frac{1}{n_{2h}}\sum_{s_{2h}} y_k$$

and

$$\bar{y}_{2,st} = \sum_{h=1}^{L} p_{1h}\,\bar{y}_{2h}.$$

Example 2: Let us assume that, for the sampling design described in Example 1, we also have auxiliary data, $\boldsymbol{x}_k$, available in the first phase s_1. If we assume that the slopes $(\boldsymbol{\beta}_h)$ vary among the strata, we can assume that the following model $y_k = \boldsymbol{x}_k'\,\boldsymbol{\beta}_h + \varepsilon_k$ holds, where $E(\varepsilon_k) = 0$, $E(\varepsilon_k^2) = \sigma_k^2$, $k \in s_{1h}, h = 1, \ldots, L$, and $E(\varepsilon_k\varepsilon_\ell) = 0$ for $k \neq \ell$, for $k, \ell \in s_{1h}$ $h = 1, \ldots, L$. This model gives us a separate regression estimator, that is,

$$\hat{\boldsymbol{Y}}_{\text{SEP, REG}} = \sum_{h=1}^{L}\frac{N}{n_1}\frac{n_{1h}}{n_{2h}}\sum_{s_2} g_{2k}\,y_k$$

where

$$g_{2k} = 1 + \left(\sum_{s_{1h}}\boldsymbol{x}_k' - \sum_{s_{2h}}\frac{n_{1h}}{n_{2h}}\boldsymbol{x}_k'\right)\left(\sum_{s_{2h}}\frac{n_{1h}}{n_{2h}}\frac{\boldsymbol{x}_k\boldsymbol{x}_k'}{\sigma_k^2}\right)^{-1}\frac{\boldsymbol{x}_k}{\sigma_k^2}$$

if $k \in s_{2h}$. In each stratum h, the slopes $\boldsymbol{\beta}_h$ are estimated as

$$\hat{\hat{\boldsymbol{B}}}_{2h} = \left(\sum_{s_{2h}}\frac{n_{1h}}{n_{2h}}\frac{\boldsymbol{x}_k\boldsymbol{x}_k'}{\sigma_k^2}\right)^{-1}\left(\sum_{s_{2h}}\frac{n_{1h}}{n_{2h}}\frac{\boldsymbol{x}_k y_k}{\sigma_k^2}\right).$$

The variance of $\hat{Y}_{\text{SEP,REG}}$ is estimated as being the sum of the variance components of each phase. These components are $\hat{V}_1(\hat{\hat{Y}}_{\text{EXP}})$ and $\hat{V}_2(\hat{Y}_{\text{SEP, REG}})$, where $\hat{V}_1(\hat{\hat{Y}}_{\text{EXP}})$ was defined in example 1. Variance $\hat{V}_2(\hat{Y}_{\text{SEP, REG}})$ is obtained by replacing variable y_k by $e_k = g_k(y_k - \boldsymbol{x}_k'\,\hat{\hat{\boldsymbol{B}}}_h)$ in $\hat{V}_2(\hat{\hat{Y}}_{\text{EXP}})$. The estimated variance of $\hat{Y}_{\text{SEP, REG}}$ is therefore:

$$\hat{V}(\hat{Y}_{\text{SEP, REG}}) = \frac{N^2(1-f_1)}{n_1}\sum_{h=1}^{L} p_{1h}\left[\begin{array}{l}(1-a_h)\,\hat{S}^2_{2yh}\\ \\ +\dfrac{n_1}{n_1-1}(\bar{y}_{2h}-\bar{y}_{2,st})^2\end{array}\right]$$

$$+\sum_{h=1}^{L}\frac{N^2(1-f_{2h})}{n_{2h}}p_{1h}^2\;\hat{S}^2_{2eh}$$

where

$$\hat{S}^2_{2eh} = \sum_{s_{2h}}\frac{(e_k-\bar{e}_h)^2}{n_{2h}-1}$$

and

$$\hat{S}^2_{2yh} = \frac{1}{n_{2h}-1}\sum_{s_{2h}}(y_k-\bar{y}_{2h})^2.$$

5.2 Non-Nested Sampling

These two examples are taken from Des Raj (1968, pages 142–149). We are using them to illustrate the results of sections 3 and 4. We consider two different sampling designs.

With the first sampling design, we assume that: (i) the first sample s_1 of size n_1 is selected with a simple random sampling design without replacement from population U; and (ii) the second sample s_2 of size n_2 is selected either by using measurements of size x_i found in the first sample s_1 (nested case) or by selecting it independently (non-nested case) from the first sample s_1 in a manner proportional to

size x_i (known for all units of the population). The resulting estimator is

$$\hat{\hat{Y}}_{\text{EPTAR}} = \frac{N}{n_1}\frac{\sum_{s_1} x_i}{n_2}\sum_{s_2}\frac{y_i}{x_i}.$$

For the second sampling design, we assume that the two samples s_1 and s_2 have been selected using a simple random sampling design without replacement. Here again, we examine the nested and non-nested cases. We assume that we find the auxiliary observation x_i for any unit selected in the first sample s_1. The estimator is $\hat{\hat{Y}}_{\text{RAT}} = (N/n_1\sum_{s_1} x_i)(\sum_{s_2} y_i/\sum_{s_2} x_i) = \hat{X}\hat{\hat{R}}$. Table 3 summarizes these two sampling designs, as well as this corresponding estimators with their estimated variances for the nested and non-nested cases.

The undefined terms in Table 3 are given by $p_{1i} = x_i/\sum_{s_1} x_i$; $p_i = x_i/\sum_U x_i$; $V(\hat{Y}_p) = 1/n_1\sum_U p_i(y_i/p_i - Y)^2$; $S^2_{y-Rx} = (N-1)^{-1}\sum_U (y_i - Rx_i)^2$; $f_2 = n_2/N$ $f_1 = n_1/N$, and $R = Y/X$.

Table 3 shows that there is little difference in the variances between the nested and non-nested cases. For $\hat{\hat{Y}}_{\text{EPTAR}}$, the variance will be smaller for the nested case if the coefficient of variation (CV) of variable y is smaller that that of variable x. For $\hat{\hat{Y}}_{\text{RAT}}$, the variance will be smaller for the nested case if $\rho\ \text{CV}(\bar{y}) < \text{CV}(\bar{x})$ where ρ is the correlation between y and x.

5.3 Two Statistics Canada Surveys

Several Statistics Canada surveys use double sampling. We will illustrate the ideas presented in this paper using two business surveys. These surveys are the Quarterly Retail Commodity Survey (QRCS) and the Survey of Employment, Payrolls and Hours (SEPH). The Quarterly Retail Commodity Survey uses nested double sampling, whereas the Survey of Employment, Payrolls and Hours (SEPH) uses non-nested double sampling.

The Quarterly Retail Commodity Survey: The purpose of the (QRCS) is to obtain detailed information on retail commodity sales on a quarterly basis. The RCS is a sub-sample of the Monthly Survey of Retail Trade (MRTS), a monthly survey. The MRTS measures mainly sales by trade group (group of three or four-digit codes of the 1980 Standard Industrial Classification (SIC)), by province and for certain census metropolitan areas (CMA). The target population is statistical companies with statistical locations identified on the Business Register and which are active in the retail trade. About 16,000 companies are interviewed each month. The population is stratified by province, territory, certain CMA and by trade group.

The MRTS is stratified in H strata, based on size (2–3 groups), geography (10 provinces, 2 territories) and industry (16 main groups). This sample is restratified independently for the QRCS. The QRCS stratification differs from the MRTS geographically, by size and by industry. A sub-sample is selected using the "new" stratification of the MRTS sample. The QRCS estimate is based on a double-ratio estimator that uses auxiliary data (sales) from the MRTS. The second-phase sampling unit (QRCS) remains the statistical company. The first-phase sample is restratified by trade group, by province and by size based on the most recent information from the MRTS. For stratification purposes, each company is assigned a province and a dominant trade group based on the one that generates the most sales. The two-phase estimator is used by the MRTS. Binder, Babyak, Brodeur, Hidiroglou, and Jocelyn (2000) derived a variance estimator that took into account the sampling design and the estimation method. They expressed variance estimators of the total as simple sums of appropriate residual terms for the case of the ratio estimator.

The results of Binder *et al.* (2000) can be adapted to incorporate the optimal regression estimator in each phase. We assume that the auxiliary information $(\boldsymbol{x}_{1k})$ is known at

Table 3
Two Sampling Designs with Nested and Non-Nested Samples

	Sampling design 1	Sampling design 2
Sampling design	$N \to n_1$ (SRSWOR) $n_1 \to n_2$ (PPSWOR)	$N \to n_1$ (SRSWOR) $n_1 \to n_2$ (SRSWOR)
Estimator	$\hat{\hat{Y}}_{\text{EPTAR}} = \frac{N}{n_1}\sum_{s_2}\frac{y_i}{n_2 p_{1i}}$	$\hat{\hat{Y}}_{\text{RAT}} = \sum_{s_1}\frac{\sum_{s_2} y_i}{\sum_{s_2} x_i} = \hat{X}\hat{\hat{R}}$
Variance Nested	$\frac{N^2(1-f_1)}{n_1}S_y^2 + \frac{V(\hat{Y}_p)}{n_2}$	$\frac{N^2(1-f_1)}{n_1}(2RS_{xy} - R^2S_x^2) + N^2\frac{(1-f_2)}{n_2}S^2_{y-Rx}$
Unnested	$\frac{N^2(1-f_1)}{n_1}R^2S_x^2 + \frac{V(\hat{Y}_p)}{n_2}\left[1 + \frac{1}{n_1}(1-f_1)\frac{S_x^2}{\bar{X}^2}\right]$	$\frac{N^2(1-f_1)}{n_1}R^2S_x^2 + N^2\frac{(1-f_2)}{n_2}S^2_{y-Rx}$

the level of population U, either for each unit $k \in U$ or for the total $\boldsymbol{X}_{1k} = \sum_U \boldsymbol{x}_{1k}$. The QRCS sampling design can be formally stated as follows. The population is stratified in H strata U_h; $h = 1, ..., H$, and simple random samples without replacement s_{1h}, of size n_{1h}, are selected in each stratum U_h. The $\boldsymbol{x}_k$ variable is observed for each unit belonging to s_1. The resulting first-phase sample, $s_1 = \boldsymbol{U}_{h=1}^{H} s_{1h}$, is then stratified in strata s_{1g}, $g = 1, ..., G$. The stratification of s_1 is independent of the stratification of the universe U. A simple random sample s_{2g} of size n_{2g} is then selected from each stratum s_{1g}, $g = 1, ..., G$. We observe $(y_k, \boldsymbol{x}_k')$, where $\boldsymbol{x}_k = (\boldsymbol{x}_{1k}', \boldsymbol{x}_{2k}')'$ for each unit belonging to sample $s_2 = U_{g=1}^{G} s_{2g}$. We assume that models $y_k = \boldsymbol{x}_{1k}' \boldsymbol{\beta}_1 + \varepsilon_{1k}$ and $y_k = \boldsymbol{x}_k' \boldsymbol{\beta} + \varepsilon_{2k}$ hold for s_1 and s_2 respectively. For each of these models $\varepsilon_{1k} \sim (0, \sigma_1^2 z_{1k})$ and $\varepsilon_{2k} \sim (0, \sigma_2^2 z_{2k})$ or z_{1k} and z_{1k} are known positive factors. If $z_{1k} \neq 1$ or $z_{1k} \neq 1$ for all units $k \in U$, the data can be standardized by dividing them either by $\sqrt{z_{1k}}$ or $\sqrt{z_{2k}}$. The resulting optimal regression estimator for the total Y is given by:

$$\tilde{\hat{Y}}_{\mathrm{OPT}} = \hat{\hat{Y}}_{\mathrm{HT}} + (\boldsymbol{X}_1 - \tilde{\boldsymbol{X}}_1)' \tilde{\boldsymbol{B}}_{1,\mathrm{OPT}} + (\hat{\boldsymbol{X}} - \tilde{\boldsymbol{X}})' \tilde{\tilde{\boldsymbol{B}}}_{\mathrm{OPT}}$$

where the components of $\tilde{\hat{Y}}_{\mathrm{OPT}}$ were defined in Section 3.1. The simplified form (without double sums) of the variance of $\tilde{\hat{Y}}_{\mathrm{OPT}}$ is:

$$\hat{V}(\tilde{\hat{Y}}_{\mathrm{OPT}}) = \sum_{h=1}^{H} N_h^2 (1 - f_{1h}) \frac{\hat{S}_{1h}^2}{n_{1h}} + \sum_{g=1}^{G} n_{1g}^2 (1 - f_{2g}) \frac{\hat{S}_{2g}^2}{n_{2g}} + \sum_{h=1}^{H} \sum_{g=1}^{G} \frac{N_h^2 (1 - f_{1h}) n_{2g}^2 (1 - f_{2g})}{n_{1h}^2 (n_{1h} - 1)} \frac{\hat{S}_{2hg}^2}{n_{2h}}$$

where the variances are defined by

$$\hat{S}_{1h}^2 = \frac{1}{n_{1h} - 1} \left\{ \sum_{g=1}^{G} \sum_{k=1}^{n_{2gh}} \frac{n_{1g}}{n_{2g}} \tilde{e}_{1k}^2 - \frac{1}{n_{1h}} \left(\sum_{g=1}^{G} \sum_{k=1}^{n_{2gh}} \frac{n_{1g}}{n_{2g}} \tilde{e}_{1k} \right)^2 \right\};$$

$$\hat{S}_{2hg}^2 = \frac{1}{n_{2hg} - 1} \sum_{k=1}^{n_{2hg}} (\tilde{e}_{1k} - \bar{\tilde{e}}_{1(hg)})^2$$

and

$$\hat{S}_{2g} = \frac{1}{n_{2g} - 1} \sum_{k=1}^{n_{2g}} (\tilde{e}_{2k} - \bar{\tilde{e}}_{2h})^2.$$

The means in these estimated variances are

$$\bar{\tilde{e}}_{1(hg)} = \frac{1}{n_{2hg}} \sum_{k=1}^{n_{2hg}} \tilde{e}_{1k}, \quad \bar{\tilde{e}}_{1(hg)} = \frac{1}{n_{2hg} 1} \sum_{k=1}^{n_{2hg}} \tilde{e}_{1k}$$

and

$$\bar{\tilde{e}}_{2h} = \frac{1}{n_{2g}} \sum_{k=1}^{n_{2g}} \tilde{e}_{2k}.$$

Here, n_{2hg} is the number of units selected in sample s_2 belonging to the intersection of strata U_h and s_{1g}. Also, the required residuals are $\tilde{e}_{1k} = g_{1k} (y_k - \boldsymbol{x}_{1k}' \tilde{\boldsymbol{B}}_{1,\mathrm{OPT}})$ and $\tilde{e}_{2k} = g_{2k} (y_k - \boldsymbol{x}_k' \tilde{\tilde{\boldsymbol{B}}}_{\mathrm{OPT}})$. The adjustment factors g_{1k} and g_{2k} are as defined in Section 4.1.

The Survey of Employment, Payrolls and Hours: The objective of this survey is to obtain estimates of the number of paid employees, the average weekly payroll and other related variables using various combinations of industry and province. This survey was recently redesigned to use administrative data for all businesses included in the survey universe. The survey produces estimates based on both the administrative data (ADMIN sample) and data directly obtained by a survey known as the Business Payroll Survey (BPS).

The ADMIN sample s_1 consists of some 200,000 units selected from universe U_1 of the pay deduction accounts to obtain the administrative data. The sampling design for this sample is stratified Bernoulli (by region), and the sampling rate varies between 10% to 100% amongst the different strata (region). The size of the sample represents approximately 20% of the total number of pay deduction accounts. Only two variables represented as $(\mathbf{x}_{1k}^{(1)})$ are available from the administrative source: these are the number of paid employees and the gross monthly payroll.

The BPS sample s_2 consists of approximately 10,000 establishments drawn from the Business Register U_2. The BPS collects the same two variables as the administrative source, namely, the number of paid employees and the gross monthly payroll denoted as $(\mathbf{x}_{1k}^{(2)})$, several other variables $(\mathbf{x}_{2k}^{(2)})$ of interest defined by type of employee (employees paid by the hour, salaried, active owners, other employees), and variables of interests, such as the number of paid hours and weekly earnings, $(y_k^{(2)})$. More information on the BPS is provided in Rancourt and Hidiroglou (1998).

The BPS is stratified by industry type, geographic region and size (varying from two to three groups based on the number of employees). These strata were designed to take into account the different regression models between $y_k^{(2)}$ and $\boldsymbol{x}_k^{(2)}$. The resulting estimated regression coefficients are used to predict $\hat{\boldsymbol{y}}_k$ for each sampled administrative record. There are two steps involved in the estimation of the total for a given variable of interest. First, the sampling weights $w_k^{(1)}$ associated with the administrative data are calibrated using known regional population counts, N_i, for regions U_{1i}, $i = 1, ..., I$. The adjusted weight of a sample unit k belonging to region U_{1i} is $\tilde{w}_k^{(1)} = w_k^{(1)} g_{1i}$, where $g_{1i} = N_i / \sum_{s_{1i}} w_k^{(1)}$ and $s_{1i} = s_1 \cap U_{1i}$. Second, $y_k^{(2)}$ is regressed on $\boldsymbol{x}_k^{(2)}$ using subsets $s_{2,j}$, $j = 1, ..., J$, of the s_2 sample. The $s_{2,j}$

subsets, classified by industry, region and sometimes size, are formed in advance to obtain the best possible regression fits. For each subset $s_{2,j}$, the estimated regression vectors $\hat{\boldsymbol{B}}_j$ are obtained as:

$$\hat{\hat{\boldsymbol{B}}}_j = \left(\sum_{s_{2,j}} w_k^{(2)} \boldsymbol{x}_k^{(2)} \boldsymbol{x}_k'^{(2)} / \hat{\sigma}_k^2 \right)^{-1} \sum_{s_{2,j}} w_k^{(2)} \boldsymbol{x}_k^{(2)} y_k^{(2)} / \hat{\sigma}_k^2 ; \quad j = 1, \ldots, J$$

where $w_k^{(2)}$ is the sampling weight for each sampled establishment, and $\hat{\sigma}_k^2$ are known positive factors that control the impact of outliers or define the required estimator. For example, if $\hat{\sigma}_k^2$ is proportional to one of the components of $\boldsymbol{x}_k^{(2)}$, we obtain the ratio estimator. The estimator of total for a variable y is therefore $\hat{Y} = \sum_{j=1}^{J} \sum_{s_{1,j}} \tilde{w}_k^{(1)} \boldsymbol{x}_k'^{(1)} \hat{\boldsymbol{B}}_j$, where $s_{1,j}$ is a partition of s_1 corresponding to the subsets defining $s_{2,j}$. SEPH is an example of a non-nested double sampling sampling design. More details of the SEPH redesign are available in Hidiroglou (1995) and Hidiroglou, Latouche, Armstrong and Gossen (1995).

6. Conclusion

Nested and non-nested double sampling are usually treated separately in the literature. Given that the population total Y is of interest, and that there is auxiliary information available, this paper has unified the estimation procedures for these two sampling methods using an optimal regression approach. Also, for the nested case, the procedure has been linked to the GREG procedure proposed by Hidiroglou and Särndal (1998). For the non-nested case, the method used by Deville (2000) has been extended when there are also auxiliary data at the population level. Lastly, practical examples were provided to illustrate this theory.

References

Berger, Y. (1998). Rate of convergence for asymptotic variance for the Horvitz-Thompson estimator. *Journal of Statistical Planning and Inference*, 74, 149-168.

Binder, D.A., Babyak, C., Brodeur, M., Hidiroglou, M.A. and Jocelyn, W. (2000). Variance estimation for two-phase stratified sampling. *The Canadian Journal of Statistics*, 28, 4, 751-764.

Breidt, J., and Fuller, W.A. (1993). Regression weighting for multiphase samples. *Sankhyā*, 55, 297-309.

Brewer, K. (2000). Deriving and estimating an approximate variance for the Horvitz-Thompson estimator using only first order inclusion probabilities. In the *Proceedings of the Second International Conferences on Establishment Surveys*. Buffalo, New York, 1417-1422.

Cassady, R.J., and Valliant, R. (1993). Conditional properties of post-stratified estimation under normal theory. *Survey Methodology*, 19, 183-192.

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 & Sons, Inc.

Des Raj (1968). *Sampling Theory*. TMH Edition.

Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey methodology*, 25, 193-204.

Deville, J.-C. (1999). Simultaneous calibrating of several surveys. *Proceedings*: *Symposium 1999, Combining Data from Different Sources*, 207-212.

Hartley, H.O., ad Rao, J.N.K. (1962). Sampling with unequal probabilities and without replacement. *Annals of Mathematical Statistics*, 33, 350-374.

Hidiroglou, M.A. (1995). Sampling and estimation for stage one of the canadian survey of employment, payrolls and hours survey redesign. *Proceedings of The Survey Methods Section*, Statistical Society of Canada, 123-128.

Hidiroglou, M.A., Latouche, M., Armstrong, B. ad Gossen, M. (1995). Improving survey information using administrative records: The case of the canadian employment survey. *Proceedings of the 1995 Annual Research Conference*. U.S. Bureau of the Census, 171-197.

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

Korn, E.L., and Graubard, B.I. (1999). *Analysis of Health Surveys*. Wiley series in probability and Statistics.

Montanari, G.E. (1987). Post-sampling efficient prediction in large-scale surveys. *International Statistical Review,* 55, 191-202.

Montanari, G.E. (1998). On regression estimation of finite population means. *Survey Methodology*, 24, 69-77.

Montanari, G.E. (2000). Conditioning on auxiliary variables means in finite population inference. *Australian New Zealand Journal of Statistics*, 42, 407-421.

Neyman, J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, 33, 101-116.

Rancourt, E., and Hidiroglou, M.A. (1998). Use of administrative records in the Canadian survey of employment, payrolls and hours. *Proceedings of the Survey Methods Section*, 39-47.

Rao, J.N.K. (1973). On double sampling for stratification and analytic surveys . *Biometrika*, 60, 125-133.

Rao, J.N.K. (1994). Estimation of totals and distribution functions using auxiliary information at the estimation stage. *Journal of Official Statistics*, 10, 153-166.

Rósen, B. (2000). A user's guide to pareto πps sampling. In the *Proceedings of the Second International Conference on Establishment Surveys*, Buffalo, New York, 289-294.

Särndal, C.-E. (1996). Efficient estimators with simple variances in unequal probability sampling. *Journal of the American Statistical Association,* 91, 1289-1300.

Särndal, C.-E., Swensson, B. and Wretman, Y. (1992). *Model assisted survey sampling*. New York, Springer-Verlag.

Tam, S.M.(1984). On covariances from nested samples. *The American Statistician*. 38, 288-289.

Tillé, Y. (2001). *Théorie des Sondages* : *Échantillonnage et estimation en population finies*. Dumond.