# DACSEIS

IST-2000-26057

## Workpackage 9

## Variance Estimation for Change

### Deliverable 9.2

2002

**List of contributors**:

Yves Berger, Chris Skinner, University of Southampton.

**Main responsibility**:

Yves Berger, University of Southampton.

**IST–2000–26057–DACSEIS**

The DACSEIS research project is financially supported within the IST programme of the European Commission. Research activities take place in close collaboration with Eurostat.

`http://europa.eu.int/comm/eurostat/research/`

`http://www.cordis.lu/ist/`

`http://www.dacseis.de/`

# Preface

There is considerable interest in changes in estimates from surveys, for example the change in the number of unemployed or in the unemployment rate. Methods for estimating change need to take account of rotation designs and methods for handling different patterns of non-response. The use of standard software also tends to be complicated by the fact that weights change over time. This workpackage investigates alternative approaches to estimating variances of changes.

Yves Berger and Chris Skinner — Southampton, October 2002

# Contents

# List of Tables

# Chapter 1

# Introduction

Most surveys in government are continuing surveys; that is, repeated monthly, quarterly, annually or with some other fixed frequency. An important reason for doing this is to estimate the manner in which the population changes from one survey period to the next. In this case, the population is sampled repeatedly and the same study variables are measured at each occasion. For example, in many countries, labour-force surveys or business surveys are conducted monthly or quarterly to estimate the change in the number of employed or the volume of retail sales, for example. A common problem in survey sampling is to compare two estimates for the same study variable taken on two occasions and to judge whether the observed change is statistically significant. It is therefore necessary to estimate the variance of a difference. This estimation would be relatively straightforward if the target population and the survey sample remained the same from one period to the next. Unfortunately, this is almost never the case (Nordberg, 2000). Methods for coping with the complications caused by sample and population change over time are discussed in this report.

To keep notational complexity to a minimum, we restrict ourselves to change in a finite population total between two occasions. Consider two population totals for the same study variable taken on two occasions: $t = 0$ and $t = 1$.

$$\begin{aligned} \tau_0 &= \sum_{i \in U_0} y_{0;i} \\ \tau_1 &= \sum_{i \in U_1} y_{1;i} \end{aligned}$$

where $y_{0;i}$ and $y_{0;i}$ are respectively the value of the study variable at time $t = 0$ and $t = 1$. The set $U_0$ is the population frame at $t = 0$ and $U_1$ is the population frame at $t = 1$. These populations can be different (see Chapter 4).

The aim is to estimate the absolute change

$$\Delta = \tau_1 - \tau_0. \tag{1.1}$$

Suppose that $\Delta$ is estimated by

$$\widehat{\Delta} = \widehat{\tau}_1 - \widehat{\tau}_0, \tag{1.2}$$

where $\widehat{\tau}_0$ and $\widehat{\tau}_1$ are Horvitz-Thompson (HT) estimators (HORVITZ and THOMPSON, 1952) given by

$$\begin{aligned} \widehat{\tau}_0 &= \sum_{i \in s_0} \frac{y_{0;i}}{\pi_{0;i}} \\ \widehat{\tau}_1 &= \sum_{i \in s_1} \frac{y_{1;i}}{\pi_{1;i}} \end{aligned} \tag{1.3}$$

where $s_0$ and $s_1$ denote samples reported on the first and second occasion. The quantities $\pi_{0;i}$ and $\pi_{0;i}$ are the inclusion probabilities for unit $i$ on the first and second occasion. These probabilities are defined in Chapter 2. The samples $s_0$ and $s_1$ are selected according to different sampling designs. We call the sampling scheme, the sampling process that selects $s_0$ and $s_1$. In other words, the sampling scheme is a combination of two sampling designs.

Populations are often surveyed at regular intervals using the same questionnaire. With a rotating scheme, new units are selected to replace old units that have been in the sample for a specified number of surveys. WOLTER (1979) describes the various forms of rotation schemes. There can be advantage in rotating the sample. For example, a rotation reduces the sample fatigue and response burden.

The purpose of this report is to estimate the variance of (1.2) given by

$$\text{var}(\widehat{\Delta}) = \text{var}(\widehat{\tau}_1) + \text{var}(\widehat{\tau}_0) - 2\ \text{cov}(\widehat{\tau}_0, \widehat{\tau}_1). \tag{1.4}$$

under rotation schemes. Development of design variances for these estimates is complicated by the need to evaluate the design covariance. This covariance is different from zero, as $s_0$ and $s_1$ are usually overlapping set of units. In Section 2.3, we propose an estimator of (1.4) asymptotically unbiased and always positive. In this report, complete response is assumed. An application to estimation of variance due to non-response will be the topic of a forthcoming report.

The Taylor series linearisation methods can be used to estimate the variance of the estimate of alternative measure of change which are functions of the population totals at each occasion. To illustrate, consider the case where we are interested in relative change measured by the ratio of population totals of the two variables.

$$\psi = \tau_1\ \tau_0^{-1}$$

and where $\psi$ is estimated by

$$\widehat{\psi} = \widehat{\tau}_1\ \widehat{\tau}_0^{-1}.$$

We have by Taylor linearization,

$$\widehat{\psi} - \psi \approx \widetilde{\tau}_1 - \widetilde{\tau}_0$$

where $\widetilde{\tau}_1 = \tau_0^{-1}\widehat{\tau}_1$ and $\widetilde{\tau}_0 = \tau_0^{-1}\psi\widehat{\tau}_0$. A variance estimator for (1.2), can be used to derive a variance estimator of $(\widetilde{\tau}_1 - \widetilde{\tau}_0)$; after replacing $\psi$ and $\tau_0$ by $\widehat{\psi}$ and $\widehat{\tau}_0$. In other situations, one can be interested in the change of parameters more complex than a total or a ratio (ANDERSSON and NORDBERG, 1994). When these parameters are function of

totals, Taylor linearization can also be used to approximate them by an absolute change (1.2). For example, with composite estimators (Fuller and Fuller, 2001) or regression estimators, Taylor linearization can be used to approximate them by a difference of HT estimators.

In this report, we propose an estimator for the design variance of a measure of change between two occasions, when the sample is selected with a rotation sampling scheme. The variance estimator proposed is always positive and can be implemented with stratified samples, when the stratification as well as the population change over time; that is, units may move between strata, new strata may be created, new units may appear in the population and units may disappear from the population.

In Chapter 2, The methodology proposed is illustrated in the special case of a single stratum. A variance estimator is given in Section 2.3. In Chapter 3, the estimator proposed is generalized for stratified populations. The population is likely to change at the second occasion. New units may appears in the population and some units may disappear. In Chapter 4, we show how the variance estimator can be generalized to accommodate this situation. In Chapter 5, a series of Monte-Carlo simulations supports our findings.

# Chapter 2

# Case of a single stratum

In this chapter, we suppose that the populations $U_0$ and $U_1$ are equal ($U_0 = U_1 = U$) and composed of a single stratum of size $N$. Assume that $s_0$ is a $\pi$ps sample with first-order inclusion probabilities $\pi_{0;i}$. Suppose that the sample $s_1$ is selected conditionally on $s_0$ in the following way: a $\pi$ps sample of $n_{01}$ units is selected from $s_0$ with probabilities $q_{0;i}$ and a $\pi$ps sample of $n_{1|0}$ units is selected form $U/s_0$ with probabilities $q_{1;i}$. The set $U/s_0$ is the set of units not selected on the first occasion. We assume that $q_{0;i}$ and $q_{1;i}$ are given constants. With this scheme, the samples $s_0$ and $s_1$ overlap. The size $n_{01}$ is the fraction of the sample $s_1$ in the matched sample $s_0 \cap s_1$. We have $n_1 = n_{01} + n_{1|0}$ units in the sample $s_1$. Let $n_0$ denotes the size of $s_0$. The sizes $n_0$, $n_1$ and $n_{01}$ are assumed fixed. In practice $n_0$ and $n_1$ are the desired sample sizes which are usually fixed or approximately fixed. This report's results can be generalized to random sizes using Nordberg's approach (Nordberg, 2000, p. 367). As shown in Examples 1, 2 and 3, the actual sampling scheme corresponds to schemes commonly used in practice.

The probability $\widetilde{\pi}_{1;i}$ for unit $i \in U_1$ to be selected in $s_1$ conditionally on $s_0$ is

$$\widetilde{\pi}_{1;i} = q_{0;i}\delta_{0;i} + q_{1;i}(1 - \delta_{0;i}) \tag{2.1}$$

where $\delta_{0;i} = 1$ if $i \in s_0$ and $\delta_{0;i} = 0$ otherwise. The probability $\widetilde{\pi}_{1;i}$ is random, as $\widetilde{\pi}_{1;i}$ depends on $s_0$. The probability $\pi_{1;i}$ (used in (1.3)) is the design expectation of $\widetilde{\pi}_{1;i}$; that is,

$$\pi_{1;i} = \mathrm{E}(\widetilde{\pi}_{1;i}) = q_{0;i}\pi_{0;i} + q_{1;i}(1 - \pi_{0;i}).$$

As two samples ($s_0$ and $s_1$) are selected, we have two sampling designs: $p_0(s_0)$ the sampling design used at the first occasion and $p_1(s_1|s_0)$ the sampling design used at the second occasion. We call the actual sampling scheme, the combination of $p_0(s_0)$ and $p_1(s_1|s_0)$. The first-order inclusion probabilities of $p_0(s_0)$ and $p_1(s_1|s_0)$ are respectively given by $\pi_{0;i}$ and $\widetilde{\pi}_{1;i}$.

**Example 1** *Some business surveys in Statistics Canada use a rotation group sampling schemes, in which each stratum is randomly divided into mutually exclusive rotation groups of the same size. Suppose that we have a single stratum and suppose that the population is randomly divided into $P$ rotation groups (for simplicity, we assume $N/P$ is an integer).*

*For the first occasion, the first $p$ groups are selected in the sample $s_0$. On the second occasion, group 1 rotates out and group $p+1$ rotates in. It can be shown that for this scheme,*

$$\begin{array}{lll} \pi_{0;i} = pP^{-1}, & n_{01} = n_{1|0} = NP^{-1}, & q_{0;i} = n_{01}n_0^{-1} \quad \textit{and} \\ q_{1;i} = n_{1|0}(N-n_1)^{-1}. & & \end{array} \tag{2.2}$$

*Both samples $s_0$ and $s_1$ are simple random samples.*

**Example 2** *For the permanent random number (PRN) method widely used in business surveys, an independent uniformly distributed random number is assigned to every unit (*NORDBERG*, 2000). All elements in the current stratum are ordered by the size of their random number. An arbitrary starting point $\alpha$ is chosen and the first $n_0$ units having a random number, larger than $\alpha$, is included in $s_0$. For the second occasion, the values of $\alpha$ increases so that $n_0 - n_{01}$ units rotates out and $n_{1|0}$ units rotates in. In this case, (2.2) holds.*

**Example 3** *The sequential Poisson sampling scheme (*OHLSEN*, 1990) is the generalization of the PRN method for unequal probabilities. This scheme is close to the actual scheme defined in the beginning of this section. For the sequential Poisson sampling scheme, $\pi_{0;i} = \pi_{1;i}$. This feature can be approximated by assuming $q_{1;i} = \pi_{0;i}(1-q_{0;i})(1-\pi_{0;i})^{-1}$; where $q_{0;i} \propto \pi_{0;i}$.*

**Example 4** *Collocated sampling (*BREWER et al.*, 1972) is an unequal probability scheme also used in repeated business surveys. With this scheme, the sample sizes are not fixed. Nevertheless, the variability of these sizes is small and this sampling scheme should be viewed as an approximation of the actual scheme defined in the beginning of this section.*

## 2.1 A model for the sampling scheme

Let $\Omega_0$ and $\Omega_1$ denotes the sets of all possible samples $s_0$ and $s_1$ that can be selected; that is,

$$\begin{aligned} \Omega_0 &= \{s_0 : \#s_0 = n_0\} \\ \Omega_1 &= \{s_1 : \#s_1 = n_1;\ \#(s_0 \cap s_1) = n_{01}\}. \end{aligned}$$

Several schemes can be used to select $s_0$ and $s_1$. Let us approximate the actual sampling scheme by the one (i) which can select a sample $s_0 \in \Omega_0$ and a sample $s_1 \in \Omega_1$ (ii) which has the right first-order inclusion probabilities $\pi_{0;i}$ and $\widetilde{\pi}_{1;i}$ and (iii) which is such that $p_0(s_0)$ and $p_1(s_1|s_0)$ have the largest entropy. The combination of these two sampling designs is called the maximum entropy scheme (MES). The following Proposition gives an analytic expression for the MES.

**Proposition 1** *The MES is defined by the combination of*

$$p_0(s_0) = \frac{P_0(s_0)}{P_0(s_0 \in \Omega_0)}\delta\{s_0 \in \Omega_0\} \tag{2.3}$$

$$p_1(s_1|s_0) = \frac{P_1(s_1|s_0)}{P_1(s_1 \in \Omega_1|s_0)}\delta\{s_1 \in \Omega_1\} \tag{2.4}$$

*where* $\delta\{A\} = 1$ *if* $A$ *is true and* $\delta\{A\} = 0$ *otherwise. The sampling designs* $P_0(s_0)$ *and* $P_1(s_1|s_0)$ *are Poisson sampling designs defined by*

$$P_0(s_0) = \prod_{i\in s} p_{0,i} \prod_{j\notin s}(1 - p_{0,j}) \tag{2.5}$$

$$P_1(s_1|s_0) = \prod_{i\in s} \widetilde{p}_{1,i} \prod_{j\notin s}(1 - \widetilde{p}_{1,j}). \tag{2.6}$$

*The* $p_{0,i}$ *are such that the first-order inclusion probabilities of* $p_0(s_0)$ *are given by* $\pi_{0,i}$*. The* $\widetilde{p}_{1,i}$ *are such that the first-order inclusion probabilities of* $p_1(s_1|s_0)$ *are given by* $\widetilde{\pi}_{1,i}$*. The probability* $P_0(s_0 \in \Omega_0)$ *is the probability of selecting a sample* $s_0 \in \Omega_0$ *under* $P(s_0)$*. The probability* $P_1(s_1 \in \Omega_1|s_0)$ *is the probability of selecting a sample* $s_1 \in \Omega_1$ *under* $P_1(s_1|s_0)$*.*

The proof is given in Appendix A in BERGER (2004a).

With Proposition 1, we see that $p_0(s_0)$ and $p_1(s_1|s_0)$ are conditional Poisson sampling designs. In this report, the inference is made under (2.3) and (2.4) (the MES), not under (2.5) and (2.6).

The MES is an approximation of the actual scheme implemented, as the sampling designs (2.3) and (2.4) are good approximation of the $\pi$ps sampling designs generally implemented (HÁJEK, 1981 and BERGER, 1998). Moreover, the large entropy assumption implies a conservative variance estimator which overestimates the variance (HÁJEK, 1959).

HÁJEK (1964, 1981) shows that the first-order inclusion probabilities of unconditional Poisson sampling designs ($p_{0;i}$ and $\widetilde{p}_{1;i}$) are close to the inclusion probabilities ($\pi_{0;i}$ and $\widetilde{\pi}_{1;i}$) of the corresponding conditional Poisson sampling design. Thus, we can assume that

$$p_{0;i} \approx \pi_{0;i} \tag{2.7}$$

$$\widetilde{p}_{1;i} \approx \widetilde{\pi}_{1;i}. \tag{2.8}$$

Although we could use (2.7) and (2.8) to approximate $p_{0;i}$ and $\widetilde{p}_{1;i}$, we prefer to keep the notation $p_{0;i}$ and $\widetilde{p}_{1;i}$ throughout this report. A method to compute $p_{0;i}$ and $\widetilde{p}_{1;i}$ exactly can be found in HÁJEK (1981), p. 139 and in CHEN *et al.* (1994).

## 2.2 Variance approximation

In this section, we propose a simple approximation for the variance of $\widehat{\Delta}$ under the MES.

First, consider the following Poisson scheme: a sample $s_0$ is selected according to (2.5) and the sample $s_0$ is selected according to (2.6). It is important to notice that under this Poisson sampling scheme, the sizes $n_0$, $n_1$ and $n_{01}$ are random. The Poisson scheme and the MES are different.

Let us assume that the vector

$$\mathbf{u} = (\widehat{\tau}_1, \widehat{\tau}_0, n_0, n_1, n_{01})'$$

has a normal distribution under the Poisson sampling scheme; that is,

$$\mathbf{u} \sim N(\mu_{\mathbf{u}}, \mathbf{\Sigma}_{\mathbf{u}}) \tag{2.9}$$

where $\mu_{\mathbf{u}}$ and $\mathbf{\Sigma}_{\mathbf{u}}$ are respectively the mean and the variance-covariance matrix of $\mathbf{u}$ under the Poisson scheme. In Section 2.3, we give the expression of $\mathbf{\Sigma}_{\mathbf{u}}$. The normality assumption is not an assumption about the distribution of the study variable. It concerns the sampling distribution of $\mathbf{u}$ under the Poisson scheme. This assumption seems natural, as each unit is selected with independent trials under Poisson sampling.

The matrix $\mathbf{\Sigma}_{\mathbf{u}}$ can be partitioned in four parts:

$$\mathbf{\Sigma}_{\mathbf{u}} = \begin{pmatrix} \mathbf{\Sigma}_{\mathbf{yy}} & \mathbf{\Sigma}_{\mathbf{yn}} \\ \mathbf{\Sigma}'_{\mathbf{yn}} & \mathbf{\Sigma}_{\mathbf{nn}} \end{pmatrix}$$

where $\mathbf{\Sigma}_{\mathbf{yy}}$ is the variance-covariance of the vector $(\widehat{\tau}_1, \widehat{\tau}_0)'$, $\mathbf{\Sigma}_{\mathbf{nn}}$ is the variance-covariance of $(n_0, n_1, n_{01})'$ and $\mathbf{\Sigma}_{\mathbf{yn}}$ gives the covariances between $(\widehat{\tau}_1, \widehat{\tau}_0)'$ and $(n_0, n_1, n_{01})'$ under the Poisson sampling scheme.

As (2.3) and (2.4) are conditional probabilities, we see that the MES is conditional on the sizes $(n_0, n_1, n_{01})'$. Thus, the variance-covariance matrix of $(\widehat{\tau}_1, \widehat{\tau}_0)'$ under MES can be approximated by

$$\mathbf{\Sigma}_{\mathbf{yy|n}} = \mathbf{\Sigma}_{\mathbf{yy}} - \mathbf{\Sigma}_{\mathbf{yn}} \mathbf{\Sigma}^{-}_{\mathbf{nn}} \mathbf{\Sigma}'_{\mathbf{yn}} \ , \tag{2.10}$$

as $\mathbf{u}$ has a normal distribution under the Poisson scheme. Where $\mathbf{\Sigma}^{-}_{\mathbf{nn}}$ is the Moore-Penrose Inverse of $\mathbf{\Sigma}_{\mathbf{nn}}$. The diagonal components of $\mathbf{\Sigma}_{\mathbf{yy|n}}$ gives an approximation of $\text{var}(\widehat{\tau}_1)$ and $\text{var}(\widehat{\tau}_0)$ under MES. The extra-diagonal component gives an approximation of $\text{cov}(\widehat{\tau}_0, \widehat{\tau}_1)$ under MES. As the MES is an approximation of the actual scheme, these approximation should be accurate. As $\widehat{\Delta} = (-1, 1)(\widehat{\tau}_0, \widehat{\tau}_1)'$, the estimator $\widehat{\Delta}$ is a linear combination of $\widehat{\tau}_0$ and $\widehat{\tau}_1$. The variance of $\widehat{\Delta}$ is therefore approximated by

$$\text{var}_M(\widehat{\Delta}) = (-1, 1)\, \mathbf{\Sigma}_{\mathbf{yy|n}} \begin{pmatrix} -1 \\ 1 \end{pmatrix}. \tag{2.11}$$

By (2.10) $\mathbf{\Sigma}_{\mathbf{yy|n}}$ is a function of $\mathbf{\Sigma}_{\mathbf{u}}$. Thus, the approximation (2.11) depends on $\mathbf{\Sigma}_{\mathbf{u}}$. BERGER (2004a) gives an analytic expression for $\mathbf{\Sigma}_{\mathbf{u}}$.

$$\mathbf{\Sigma_u} = \breve{\mathbf{A}}'_U \, \mathbf{C}_U \, \breve{\mathbf{A}}_U \tag{2.12}$$

where $\mathbf{C}_U = \text{var}_P(\delta)$ is the variance of $\delta$ under the Poisson scheme and where

$$\breve{\mathbf{A}}_U = \begin{pmatrix} \breve{\mathbf{y}}_0 & \mathbf{0} & \breve{\mathbf{z}}_0 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \breve{\mathbf{y}}_1 & \mathbf{0} & \breve{\mathbf{z}}_0 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \breve{\mathbf{z}}_0 \end{pmatrix}$$

$\breve{\mathbf{y}}_0 = (\breve{y}_{0;1}, \ldots, \breve{y}_{0;N})$, $\breve{y}_{0;i} = y_{0;i}\pi_{0;i}^{-1}$, $\breve{\mathbf{y}}_1 = (\breve{y}_{1;1}, \ldots, \breve{y}_{1;N})$, $\breve{y}_{1;i} = y_{1;i}\pi_{1;i}^{-1}$ and $\breve{\mathbf{z}}_0$ is a $N \times 1$ vector of 1. The $(3N \times 3N)$ matrix $\mathbf{C}_U$ is given by (see Appendix B in BERGER, 2004a)

$$\mathbf{C}_U = \begin{pmatrix} \mathbf{C}_{0;0} & \mathbf{C}_{0;1} & \mathbf{C}_{0;01} \\ \mathbf{C}_{0;1} & \mathbf{C}_{1;1} & \mathbf{C}_{1;01} \\ \mathbf{C}_{0;01} & \mathbf{C}_{1;01} & \mathbf{C}_{01;01} \end{pmatrix}. \tag{2.13}$$

The sub-matrices $\mathbf{C}_{.,.}$ are $N \times N$ diagonal matrix given by

$$\begin{aligned}
\mathbf{C}_{0;0} &= \text{diag}\left\{p_{0;i}(1 - p_{0;i}) : i \in U\right\} && (2.14)\\
\mathbf{C}_{1;1} &= \text{diag}\left\{\pi_{1;i}(1 - \pi_{1;i}) : i \in U\right\} && (2.15)\\
\mathbf{C}_{01;01} &= \text{diag}\left\{q_{0;i}p_{0;i}(1 - q_{0;i}p_{0;i}) : i \in U\right\} && (2.16)\\
\mathbf{C}_{0;1} &= \text{diag}\left\{p_{0;i}(q_{0;i} - \pi_{1;i}) : i \in U\right\} && (2.17)\\
\mathbf{C}_{0;01} &= \text{diag}\left\{q_{0;i}p_{0;i}(1 - p_{0;i}) : i \in U\right\} && (2.18)\\
\mathbf{C}_{1;01} &= \text{diag}\left\{q_{0;i}p_{0;i}(1 - \pi_{1;i}) : i \in U\right\}. && (2.19)
\end{aligned}$$

## 2.3 Variance estimator

A natural estimator of (2.11) is

$$\widehat{\text{var}}_M(\widehat{\Delta}) = (-1, 1)\, \widehat{\mathbf{\Sigma}}_{\mathbf{yy|n}} \begin{pmatrix} -1 \\ 1 \end{pmatrix} \tag{2.20}$$

where

$$\widehat{\mathbf{\Sigma}}_{\mathbf{yy|n}} = \widehat{\mathbf{\Sigma}}_{\mathbf{yy}} - \widehat{\mathbf{\Sigma}}_{\mathbf{yn}} \widehat{\mathbf{\Sigma}}^{-}_{\mathbf{nn}} \widehat{\mathbf{\Sigma}}'_{\mathbf{yn}} \tag{2.21}$$

where $\widehat{\mathbf{\Sigma}}_{\mathbf{yy}}$, $\widehat{\mathbf{\Sigma}}_{\mathbf{yn}}$ and $\widehat{\mathbf{\Sigma}}_{\mathbf{nn}}$ are estimators of $\mathbf{\Sigma}_{\mathbf{yy}}$, $\mathbf{\Sigma}_{\mathbf{yn}}$ and $\mathbf{\Sigma}_{\mathbf{nn}}$. BERGER (2004a) shows that the unbiased HT estimators of $\mathbf{\Sigma_u}$ is (see Appendix D in BERGER, 2004a)

$$\widehat{\mathbf{\Sigma}}_{\mathbf{u}} = \breve{\mathbf{A}}'_s \, \breve{\mathbf{C}}_s \, \breve{\mathbf{A}}_s \tag{2.22}$$

where $\breve{\mathbf{A}}_s$ is the sub-matrix of $\breve{\mathbf{A}}_U$ composed of rows for the sample units; that is, the rows of $\breve{\mathbf{A}}_s$ are given by the row $i$ $(i = 1, .., 3N)$ of $\breve{\mathbf{A}}_U$, where the $i$-th component of $\delta$ is equal to 1. The matrix $\breve{\mathbf{C}}_s$ is a sub-matrix of $\breve{\mathbf{C}}_U$ (defined by (2.23)). The components

of $\breve{\mathbf{C}}_s$ are given by component $(i,j)$ $(i,j=1,..,3N)$ of $\breve{\mathbf{C}}_U$, where the $i$-th and the $j$-th component of $\delta$ are equal to 1. The matrix $\breve{\mathbf{C}}_U$ is defined by

$$\breve{\mathbf{C}}_U = \begin{pmatrix} \breve{\mathbf{C}}_{0;0} & \breve{\mathbf{C}}_{0;1} & \breve{\mathbf{C}}_{0;01} \\ \breve{\mathbf{C}}_{0;1} & \breve{\mathbf{C}}_{1;1} & \breve{\mathbf{C}}_{1;01} \\ \breve{\mathbf{C}}_{0;01} & \breve{\mathbf{C}}_{1;01} & \breve{\mathbf{C}}_{01;01} \end{pmatrix} \tag{2.23}$$

where

$$\begin{aligned}
\breve{\mathbf{C}}_{0;0} &= \operatorname{diag}\left\{p_{0;i}(1-p_{0;i})\pi_{0;i}^{-1} : i \in U\right\} && (2.24)\\
\breve{\mathbf{C}}_{1;1} &= \operatorname{diag}\left\{\pi_{1;i}(1-\pi_{1;i})\pi_{1;i}^{-1} : i \in U\right\} && (2.25)\\
\breve{\mathbf{C}}_{01;01} &= \operatorname{diag}\left\{q_{0;i}p_{0;i}(1-q_{0;i}p_{0;i})(q_{0;i}\pi_{0;i})^{-1} : i \in U\right\} && (2.26)\\
\breve{\mathbf{C}}_{0;1} &= \operatorname{diag}\left\{p_{0;i}(q_{0;i}-\pi_{1;i})(q_{0;i}\pi_{0;i})^{-1} : i \in U\right\} && (2.27)\\
\breve{\mathbf{C}}_{0;01} &= \operatorname{diag}\left\{q_{0;i}p_{0;i}(1-p_{0;i})(q_{0;i}\pi_{0;i})^{-1} : i \in U\right\} && (2.28)\\
\breve{\mathbf{C}}_{1;01} &= \operatorname{diag}\left\{q_{0;i}p_{0;i}(1-\pi_{1;i})(q_{0;i}\pi_{0;i})^{-1} : i \in U\right\}. && (2.29)
\end{aligned}$$

The sub-matrices of $\widehat{\mathbf{\Sigma}}_{\mathbf{u}}$ give $\widehat{\mathbf{\Sigma}}_{\mathbf{yy}}$, $\widehat{\mathbf{\Sigma}}_{\mathbf{yn}}$ and $\widehat{\mathbf{\Sigma}}_{\mathbf{nn}}$, when substituted into (2.20) give the variance estimator. The Matrix $\widehat{\mathbf{\Sigma}}_{\mathbf{u}}$ is a $5\times 5$ matrix of $5(5+1)/2 = 15$ HT totals. The expression of these totals can be derived from (2.22). >From a computational point of view, these 15 totals can be directly computed without a complete computation of $\breve{\mathbf{C}}_s$ and $\breve{\mathbf{A}}_s$.

The variance estimator (2.20) should be asymptotically unbiased. The reason is relatively simple. The variance (2.11) is a function of population totals given by the components of $\mathbf{\Sigma}_{\mathbf{u}}$. In the estimator (2.20), these totals are replaced by their unbiased HT estimator. Thus, if the actual scheme converges to the MES and if the distribution of $\mathbf{u}$ is asymptotically normal under the Poisson scheme, (2.20) is asymptotically unbiased.

Negative variance estimates of changes is a common issue when the covariance is estimated from the units of the matched sample $s_0 \cap s_1$ (HIDIROGLOU *et al*., 1995). This is due to the bias of such estimator of covariance. The estimator (2.20) does not have this drawback, as it always gives a positive estimate (see Appendix C in BERGER, 2004a)

$$\widehat{\text{var}}(\widehat{\Delta}) \geq 0. \tag{2.30}$$

As far as the finite population correction (FPC) is concerned, it is generally not clear what kind of FPC should be used for estimation of changes (HIDIROGLOU *et al*., 1995). With our approach, we have a simple interpretation of the different FPC involved for the variances and covariances. In fact, these FPC are given by the non-zero components of $\breve{\mathbf{C}}_s$. With Assumption (2.7) and (2.8 ), these components can be simplified, implying four kind of FPC:

- $(1-\pi_{0;i})$ for the variance of HT totals based on $s_0$. This FPC is equal to $(1-n_0/N)$ for simple random sampling (SRS).

- $(1-\pi_{1;i})$ for the variance of HT totals based on $s_1$. This FPC is equal to $(1-n_1/N)$ or SRS.
- $(1-q_{0;i}\pi_{0;i})$ for the variance of HT totals based on the matched sample $s_0 \cap s_1$. The quantity $q_{0;i}\pi_{0;i}$ is the probability for $i \in s_0 \cap s_1$.
- $(1-\pi_{1;i}q_{0;i}^{-1})$ for covariance between a HT total based on $s_0$ and a HT total based on $s_1$.

If the sample sizes are not fixed and if the actual sampling scheme is the Poisson sampling scheme, the transformation (2.21) is not necessary and a variance estimator is obtained by replacing $\widehat{\mathbf{\Sigma}}_{\mathbf{yy|n}}$ by $\widehat{\mathbf{\Sigma}}_{\mathbf{yy}}$ in (2.20).

# Chapter 3

# Case of stratified sampling scheme

In this chapter, we suppose that the population $U_0$ and $U_1$ are equal and composed of $H$ strata ($U_0 = U_1 = U$).

## 3.1 Same stratification

Suppose that the stratification in $U_0$ and in $U_1$ is the same; that is, the units remain in the same stratum on both occasion. Let $U_1, \ldots, U_H$ denote those strata. In this section, we show that the variance estimator (2.20) can still be used; we only need to change the vector $\mathbf{u}$ and the matrix $\breve{\mathbf{A}}_U$. The matrix $\mathbf{C}_U$ is still given by (2.13).

Consider a set of $H$ stratification variables given by the indicator variables for the strata; that is, $z_{ih} = 1$ if $i$ belongs to the $h$-th stratum and $z_{ih} = 0$ otherwise. This information is summarized in the following matrix

$$\breve{\mathbf{Z}}_0 = \{z_{ih}\}_{i=1,..,N;h=1,..,H}$$

Now, the vector $\mathbf{u}$ is composed of the different sizes of the different strata into account

$$\mathbf{u} = (\widehat{\tau}_1, \widehat{\tau}_0, n_{0;1}, \ldots, n_{0;H}, n_{1;1}, \ldots, n_{1;H}, n_{01;1}, \ldots, n_{01;H})'$$

where $n_{0;h} = \#(U_h \cap s_0)$, $n_{1;h} = \#(U_h \cap s_1)$ and $n_{01;h} = \#(U_h \cap s_1 \cap s_0)$. As the vector $\mathbf{u}$ is different, the matrix $\breve{\mathbf{A}}_U$ is also different and given by

$$\breve{\mathbf{A}}_U = \begin{pmatrix} \breve{\mathbf{y}}_0 & \mathbf{0} & \breve{\mathbf{Z}}_0 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \breve{\mathbf{y}}_1 & \mathbf{0} & \breve{\mathbf{Z}}_0 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \breve{\mathbf{Z}}_0 \end{pmatrix}. \tag{3.1}$$

The MES is still defined as in Section 2.1 with new sets $\Omega_0$ and $\Omega_1$ given by

$$\begin{aligned} \Omega_0 &= \{s_0 : \#(U_h \cap s_0) = n_{0;h};\ h = 1, .., H\} \\ \Omega_1 &= \{s_1 : \#(U_h \cap s_1) = n_{1;h};\ \#(U_h \cap s_1 \cap s_0) = n_{01;h};\ h = 1, .., H\}. \end{aligned}$$

Using the same methodology, we can derive a variance approximation (2.11) based on the matrix $\mathbf{\Sigma_u}$ defined by ( 2.12) with $\breve{\mathbf{A}}_U$ given by (3.1) and $\mathbf{C}_U$ given by (2.13). The components of $\mathbf{\Sigma_u}$ are still population totals that can be estimated easily by their corresponding HT estimators. The final variance estimator is given by (2.20).

## 3.2 Different stratification

Suppose that $U_0$ and $U_1$ are composed of different strata; for example, if the variables used to define the strata are subject to change. Suppose that $U_0$ is composed of $H_0$ strata and that $U_1$ is composed of $H_1$ strata. At the first occasion, we have a set of $H_0$ stratification variables given by the indicator variables $z_{0;ih} = 1$ if $i$ belongs to the $h$-th stratum of $U_0$ and $z_{0;ih} = 0$ otherwise. This information is summarized by the following matrix

$$\breve{\mathbf{Z}}_0 = \{z_{0;ih}\}_{i=1,..,N;h=1,..,H_0} .$$

Equivalently, the information of the stratification of $U_1$ is summarized by the following matrix

$$\breve{\mathbf{Z}}_1 = \{z_{1;ih}\}_{i=1,..,N;h=1,..,H_1} .$$

The vector $\mathbf{u}$ includes the sizes of the strata at both occasions. This implies a matrix $\breve{\mathbf{A}}_U$ is given by

$$\breve{\mathbf{A}}_U = \begin{pmatrix} \breve{\mathbf{y}}_0 & \mathbf{0} & \breve{\mathbf{Z}}_0 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \breve{\mathbf{y}}_1 & \mathbf{0} & \breve{\mathbf{Z}}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \breve{\mathbf{Z}}_0 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \breve{\mathbf{Z}}_1 \end{pmatrix}$$

with a vector $\delta$ given by $\delta = (\delta_0', \delta_1', \delta_{01}', \delta_{01}')'$ and a matrix $\mathbf{C}_U = \mathrm{var}_P(\delta)$ given by

$$\mathbf{C}_U = \begin{pmatrix} \mathbf{C}_{0;0} & \mathbf{C}_{0;1} & \mathbf{C}_{0;01} & \mathbf{C}_{0;01} \\ \mathbf{C}_{0;1} & \mathbf{C}_{1;1} & \mathbf{C}_{1;01} & \mathbf{C}_{1;01} \\ \mathbf{C}_{0;01} & \mathbf{C}_{1;01} & \mathbf{C}_{01;01} & \mathbf{C}_{01;01} \\ \mathbf{C}_{0;01} & \mathbf{C}_{1;01} & \mathbf{C}_{01;01} & \mathbf{C}_{01;01} \end{pmatrix} .$$

The components of $\mathbf{\Sigma_u}$ (defined by (2.12)) are still population totals that can be estimated by their corresponding HT estimators. The final variance estimator is given by (2.20).

# Chapter 4

# Case of different populations

Real populations are rarely static. Thus, the units making up the population contributing to $\tau_0$ will often be different from those making up the population contributing to $\tau_1$. In many cases there will be considerable overlap between the populations at the two occasions. Suppose that the populations $U_0$ and $U_1$ are different and can be divided into three non-overlapping parts. The first part consists of the units that are included in $U_0$ but not in $U_1$; that is, those units that have disappeared between time 0 and time 1. We call this group $D$ (for "death"). The second part consists of the units that are included in both populations. We call this group $P$ (for "persistent"). The third part consists of units that are included in $U_1$ but not in $U_0$. We call this group $B$ (for "birth").

Consider the working population $U = U_0 \cup U_1$. As the units of $B$ cannot be selected at the first occasion, we have $\pi_{0;i} = 0$ for $i \in B$. As the unit of $D$ cannot be selected at the second occasion, we have $q_{0;i} = q_{1;i} = 0$ for $i \in D$. The methodology of the previous sections can be used with these inclusion probabilities and the working population $U$. Depending on the stratification involved, we use the matrix $\breve{\mathbf{A}}_U$ of Section 3.1 or of Section 3.2. The variance can be still estimated by (2.20).

# Chapter 5

# Empirical Study

In this chapter, we suppose that the populations $U_0$ and $U_1$ are equal ($U_0 = U_1 = U$) and composed of a single stratum of size $N = 500$. The study variable $y_{0;i}$ is generated according to a lognormal distribution $LogN(\mu, \sigma)$ with $\mu = 3$. Different values for $\sigma$ will be considered. The $\pi_{0;i}$ are proportional to a size variable correlated with the $y_{0;i}$ with a coefficient of correlation of 0.6. As the $\pi_{0;i}$ are correlated with the $y_{0;i}$, a skewed distribution for the $y_{0;i}$ implies a skewed distribution for the $\pi_{0;i}$. The study variable $y_{1;i}$ is generated randomly from $y_{0;i}$ with a linear model that gives a correlation of 0.9 between the $y_{1;i}$ and the $y_{0;i}$. The values of $y_{1;i}$ and $y_{0;i}$ are standardized such that the mean of $y_{1;i}$ and $y_{0;i}$ are respectively given by 10 and $(10 + \Delta/N)$; and such that the standard deviation of $y_{1;i}$ and $y_{0;i}$ are 5 and 7.

At the first occasion, the S-Plus® function `sample()` is used to select $s_0$. At the second occasion, we select a simple random sample from $s_0$ and $U/s_0$. The sampling fraction of the sample from $s_0$ is 80%. At both occasion, the sample size is the same and given by $n = n_0 = n_1$. Different values for $n$ will be considered. We use the approximation (2.7) and (2.8) for $p_{0;i}$ and $\widetilde{p}_{1;i}$.

Consider the following measures of accuracy:

$$
\begin{aligned}
CV &= 100\ \mathrm{var}(\widehat{\Delta})^{1/2}|\Delta|^{-1} && (5.1)\\
CV_M &= 100\ \mathrm{var}_M(\widehat{\Delta})^{1/2}|\Delta|^{-1} && (5.2)\\
ECV &= 100\ \mathrm{E}(\widehat{\mathrm{var}}(\widehat{\Delta}))^{1/2}|\Delta|^{-1} && (5.3)\\
RRMSE &= 100\ \mathrm{MSE}\,(\widehat{\mathrm{var}}(\widehat{\Delta}))^{1/2}\mathrm{var}(\widehat{\Delta})^{-1} && (5.4)
\end{aligned}
$$

where $\mathrm{E}(\widehat{\mathrm{var}}(\widehat{\Delta}))$, $\mathrm{var}(\widehat{\Delta})$ and $\mathrm{MSE}\,(\widehat{\mathrm{var}}(\widehat{\Delta}))$ represents the empirical expectation variance and mean square error of (2.20) based on 10000 samples.

The results are shown in Table 5.1. The column $\sigma$ gives the different standard deviations of the lognormal distribution. The column $f$ gives the different sampling fractions $f = n/N$ considered. The column $\Delta/N$ gives the population values of (1.1) divided by $N$. The values of (5.1)-(5.4) are given in column 4 to 7. The column "Cov." gives the coverage of the 95% confidence interval using ( 2.20) and the quantile of a normal distribution. In addition, we present values of the misspecification effect, $meff$, ignoring the covariance.

This effect is computed by dividing $\text{var}(\widehat{\Delta})$ by $[\text{var}(\widehat{\tau}_1) + \text{var}(\widehat{\tau}_0)]$. This quantity lies between 0 and 1. A small value for $meff$ means that the covariance contributes a lot to the variance.

The coefficient of variation of $\widehat{\Delta}$, $CV$, is a decrease function of $f^{-1}$ and is inversely proportional to $\Delta$. It does not seems to depend on $\sigma$; this is probably due to the fact that $y_{1;i} - y_{0;i}$ is not as skewed as the distribution of the study variables. The small values for $meff$ means that the covariance contributes a lot in the variance. This is the consequence of the strong correlation between the $y_{1;i}$ and the $y_{0;i}$. The approximation (2.11) captures well this effect, as $CV_M$ is always close to $CV$; that is, the MES assumption gives a good approximation of the real variance. The $ECV$ are slightly smaller than the $CV$. This means that the bias of (2.20) is very small and slightly negative. The relative root mean square error ($RRMSE$) seems to be a increase function of $CV$. The coverage is good except for $f = 0.05$. This is not surprising as the sample size is only 25 in this case.

Table 5.1: Measure of accuracy (%) for the estimator of variance proposed.

| $\sigma$ | $\Delta/N$ | $f$ | $CV$ | $CV_M$ | $ECV$ | $RRMSE$ | Cov. | $meff$ |
|---|---|---|---|---|---|---|---|---|
| 0.2 | 1 | 0.05 | 89.51 | 89.41 | 83.59 | 29.35 | 0.92 | 0.35 |
| 0.2 | 1 | 0.10 | 63.46 | 62.64 | 60.79 | 20.19 | 0.93 | 0.36 |
| 0.2 | 1 | 0.15 | 50.33 | 50.43 | 49.60 | 15.72 | 0.94 | 0.37 |
| 0.2 | 1 | 0.20 | 43.37 | 43.07 | 42.60 | 12.96 | 0.94 | 0.39 |
| 0.2 | 5 | 0.05 | 17.91 | 17.97 | 16.79 | 29.74 | 0.92 | 0.36 |
| 0.2 | 5 | 0.10 | 12.71 | 12.51 | 12.18 | 20.85 | 0.94 | 0.37 |
| 0.2 | 5 | 0.15 | 10.08 | 10.12 | 9.93 | 15.67 | 0.94 | 0.37 |
| 0.2 | 5 | 0.20 | 8.69 | 8.63 | 8.52 | 12.95 | 0.94 | 0.39 |
| 0.2 | 10 | 0.05 | 8.82 | 8.83 | 8.29 | 29.96 | 0.92 | 0.36 |
| 0.2 | 10 | 0.10 | 6.16 | 6.14 | 5.96 | 20.18 | 0.94 | 0.38 |
| 0.2 | 10 | 0.15 | 4.93 | 4.92 | 4.83 | 15.51 | 0.94 | 0.39 |
| 0.2 | 10 | 0.20 | 4.27 | 4.20 | 4.14 | 13.50 | 0.94 | 0.41 |
| 0.9 | 1 | 0.05 | 86.80 | 86.20 | 81.12 | 34.63 | 0.92 | 0.37 |
| 0.9 | 1 | 0.10 | 59.43 | 60.17 | 58.52 | 23.89 | 0.93 | 0.39 |
| 0.9 | 1 | 0.15 | 48.68 | 48.76 | 47.99 | 19.25 | 0.94 | 0.39 |
| 0.9 | 1 | 0.20 | 41.26 | 41.60 | 41.09 | 15.93 | 0.95 | 0.40 |
| 0.9 | 5 | 0.05 | 16.30 | 16.43 | 15.47 | 44.03 | 0.93 | 0.37 |
| 0.9 | 5 | 0.10 | 11.51 | 11.47 | 11.18 | 31.57 | 0.94 | 0.37 |
| 0.9 | 5 | 0.15 | 9.32 | 9.28 | 9.14 | 25.36 | 0.94 | 0.40 |
| 0.9 | 5 | 0.20 | 7.93 | 7.94 | 7.86 | 21.72 | 0.94 | 0.40 |
| 0.9 | 10 | 0.05 | 8.03 | 8.11 | 7.61 | 37.92 | 0.93 | 0.39 |
| 0.9 | 10 | 0.10 | 5.61 | 5.61 | 5.45 | 26.29 | 0.94 | 0.40 |
| 0.9 | 10 | 0.15 | 4.55 | 4.55 | 4.48 | 22.02 | 0.94 | 0.40 |
| 0.9 | 10 | 0.20 | 3.89 | 3.89 | 3.85 | 18.53 | 0.95 | 0.41 |

# Chapter 6

# Conclusion

The variance estimator proposed is simple to implement, as it involves the multivariate normal theory (see (2.21)). It can be used for stratified populations. At the second occasion the population can change, as well as the stratification. The variance estimator proposed is always positive and asymptotically unbiased.

# Appendix A

# S-Plus® Functions

In this appendix, we have a series of functions that compute the variance estimator (2.20) used for the simulations.

```
# FUNCTION VAR.DIFF()
# ===================
#
# This function computes the BERGER (2004b) variances estimator
# for a pps systematic sampling design at both occasion.
#
# INPUT:
#       * Sample.0: a Nx1 vector containing the information about
#         the sample at the first occasion: the i-th component of Sample
#         is equal to 1 if the i-th unit is selected, otherwise, the i-th
#         component is equal to 0.
#       * Sample.1: a Nx1 vector containing the information about
#         the sample at the second occasion: the i-th component of Sample
#         is equal to 1 if the i-th unit is selected, otherwise, the i-th
#         component is equal to 0.
#
# IMPORTANT REMARK:
#       The function VAR.DIFF() need to be initialised by
#       Var.Vect.Delta.Sample <- INIT.VAR.DIFF.1(Vect.Pi.0,Sample.Size,
#       Percentage.Overlap)
#       Matrix.A <- INIT.VAR.DIFF.2(Vect.Y.0,Vect.Y.1,Vect.Pi.0,
#       Sample.Size,Percentage.Overlap)
#       A new call of INIT.VAR.DIFF.1() and INIT.VAR.DIFF.2() is
#       necessary if one of their parameters change.



VAR.DIFF <- function(Sample.0,Sample.1) {


# COMPUTE VECTOR DELTA
    Vect.Delta <- c(Sample.0,Sample.1,Sample.0*Sample.1)
    Logical.Vect.Delta <- (Vect.Delta == 1)

# COMPUTE Matrix.A.S
```

```
    Matrix.A.S <- Matrix.A[Logical.Vect.Delta,]


# COMPUTE MATRIX Var.Vect.Delta.S
    Var.Vect.Delta.S <- Var.Vect.Delta.Sample[Logical.Vect.Delta,
                        Logical.Vect.Delta]


# ESTIMATION OF THE VARIANCE-COVARIANCE MATRIX UNDER POISSON SAMPLING
    Matrix.Var.Poisson.S <- t(Matrix.A.S) %*% Var.Vect.Delta.S %
                            *% Matrix.A.S


# COMPUTE THE CONDITIONAL VARIANCE-COVARIANCE MATRIX
# UNDER CONDITIONAL POISSON SAMPLING
    Sub.Matrix.YY <- Matrix.Var.Poisson.S[(1:2),(1:2)]
    Sub.Matrix.XX <- Matrix.Var.Poisson.S[(3:NB.Col.Matrix.A.S),
                     (3:NB.Col.Matrix.A.S)]
    Sub.Matrix.YX <- Matrix.Var.Poisson.S[(1:2),(3:NB.Col.Matrix.A.S)]
    Matrix.Var.T.0.T.1.S <- Sub.Matrix.YY - Sub.Matrix.YX %*% %
                            GINVERSE(Sub.Matrix.XX) %*% t(Sub.Matrix.YX)


# COMPUTE THE COVARIANCE AND VARIANCES
    Var.T.0.CP.S <- Matrix.Var.T.0.T.1.S[1,1]
    Var.T.1.CP.S <- Matrix.Var.T.0.T.1.S[2,2]
    Cov.T.0.T.1.CP.S <- Matrix.Var.T.0.T.1.S[1,2]
    Var.Diff.CP.S <- Var.T.0.CP.S + Var.T.1.CP.S - 2 * Cov.T.0.T.1.CP.S


# OUTPUT
    Var.Diff.CP.S

 }
```

```
# FUNCTION INIT.VAR.DIFF.1()
# ==========================
#
# This function initialises the function VAR.DIFF() variance estimator.
#
# INPUT:
#       * Vect.Pi.0: the Nx1 vector of the first-order inclusion
#         probabilities at the first occasion.
#       * Sample.Size: the sample size, n.
#       * Percentage.Overlap: the percentage overlap (common units)
#         between the two samples.


INIT.VAR.DIFF.1 <- function(Vect.Pi.0,Sample.Size,Percentage.Overlap) {


# COMPUTE PROBABILITY IN AND OUT
    Pop.Size <- length(Vect.Pi.0)
    Sample.Size.Common <- round(Percentage.Overlap * Sample.Size,digits=0)
    Prob.In <- Sample.Size.Common / Sample.Size
```

```
    Prob.Out <- (Sample.Size - Sample.Size.Common) /
                (Pop.Size - Sample.Size)
    NB.Row <- 3 * Pop.Size
    Var.Vect.Delta <- matrix(rep(0,times=NB.Row*NB.Row),
                      nrow=NB.Row,ncol=NB.Row)


# PART CORRESPONDING TO Var(Delta.0)
    Diagonal <- as.vector(Vect.Pi.0 * (1 - Vect.Pi.0))
    Var.Vect.Delta[(1:Pop.Size),(1:Pop.Size)] <- diag(Diagonal)


# PART CORRESPONDING TO Var(Delta.1)
    A.0 <- Prob.In * Vect.Pi.0 + Prob.Out * (1 - Vect.Pi.0)
    Diagonal <- as.vector(A.0 * (1 - A.0))
    Var.Vect.Delta[((Pop.Size+1):(2*Pop.Size)),((Pop.Size+1):
             (2*Pop.Size))] <- diag(Diagonal)


# PART CORRESPONDING TO Cov(Delta.0,Delta.1)
    Diagonal <- as.vector(Prob.In * Vect.Pi.0 - Vect.Pi.0 *
                (Prob.In * Vect.Pi.0 + Prob.Out * (1 - Vect.Pi.0)))
    Var.Vect.Delta[((Pop.Size+1):(2*Pop.Size)),(1:Pop.Size)]
             <- diag(Diagonal)


# PART CORRESPONDING TO Cov(Delta.0,Delta.0.1)
    Diagonal <- as.vector(Prob.In * Vect.Pi.0 * (1 - Vect.Pi.0))
    Var.Vect.Delta[((2*Pop.Size+1):(3*Pop.Size)),(1:Pop.Size)]
             <- diag(Diagonal)


# PART CORRESPONDING TO Cov(Delta.1,Delta.0.1)
    Diagonal <- as.vector(Prob.In * Vect.Pi.0 * (1 - (Prob.In *
                Vect.Pi.0 + Prob.Out * (1 - Vect.Pi.0))))
    Var.Vect.Delta[((2*Pop.Size+1):(3*Pop.Size)),((Pop.Size+1):
            (2*Pop.Size))] <- diag(Diagonal)


# PART CORRESPONDING TO Cov(Delta.0.1,Delta.0.1)
    Diagonal <- as.vector(Prob.In * Vect.Pi.0 * (1 - Prob.In * Vect.Pi.0))
    Var.Vect.Delta[((2*Pop.Size+1):(3*Pop.Size)),((2*Pop.Size+1):
            (3*Pop.Size))] <- diag(Diagonal)


# COMPUTE THE UPPER DIAGONAL
    Diagonal <- as.vector(diag(Var.Vect.Delta))
    Var.Vect.Delta <- Var.Vect.Delta + t(Var.Vect.Delta)
    diag(Var.Vect.Delta) <- Diagonal


# INITIALISATION MATRIX Var.Vect.Delta.Sample
    Var.Vect.Delta.Sample <- matrix(rep(0,times=NB.Row*NB.Row),
                             NB.Row,NB.Row)


# PART CORRESPONDING TO Var(Delta.0)
```

```
    Diagonal <- as.vector(diag(Var.Vect.Delta[(1:Pop.Size),
                (1:Pop.Size)])) / Vect.Pi.0
    Var.Vect.Delta.Sample[(1:Pop.Size),(1:Pop.Size)] <- diag(Diagonal)


# PART CORRESPONDING TO Var(Delta.1)
    Diagonal <- as.vector(diag(Var.Vect.Delta[((Pop.Size+1):(2*Pop.Size)),
                ((Pop.Size+1):(2*Pop.Size))]))/ (Prob.In * Vect.Pi.0 +
                Prob.Out * (1 - Vect.Pi.0))
    Var.Vect.Delta.Sample[((Pop.Size+1):(2*Pop.Size)),
                ((Pop.Size+1):(2*Pop.Size))] <- diag(Diagonal)


# PART CORRESPONDING TO Var(Delta.01)
    Diagonal <- as.vector(diag(Var.Vect.Delta[((2*Pop.Size+1):
                (3*Pop.Size)), ((2*Pop.Size+1):(3*Pop.Size))])) /
                (Prob.In * Vect.Pi.0)
    Var.Vect.Delta.Sample[((2*Pop.Size+1):(3*Pop.Size)),
                ((2*Pop.Size+1):(3*Pop.Size))] <- diag(Diagonal)


# PART CORRESPONDING TO Cov(Delta.0,Delta.1)
    Diagonal <- as.vector(diag(Var.Vect.Delta[((Pop.Size+1):(2*Pop.Size)),
                (1:Pop.Size)]))/ (Prob.In * Vect.Pi.0)
    Var.Vect.Delta.Sample[((Pop.Size+1):(2*Pop.Size)),(1:Pop.Size)]
            <- diag(Diagonal)


# PART CORRESPONDING TO Cov(Delta.0,Delta.01)
    Diagonal <- as.vector(diag(Var.Vect.Delta[((2*Pop.Size+1):
                (3*Pop.Size)),(1:Pop.Size)]))/ (Prob.In * Vect.Pi.0)
    Var.Vect.Delta.Sample[ ((2*Pop.Size+1):(3*Pop.Size)),(1:Pop.Size)]
            <- diag(Diagonal)


# PART CORRESPONDING TO Cov(Delta.1,Delta.01)
    Diagonal <- as.vector(diag(Var.Vect.Delta[((2*Pop.Size+1):
                (3*Pop.Size)),((Pop.Size+1):(2*Pop.Size))])) /
                (Prob.In * Vect.Pi.0)
    Var.Vect.Delta.Sample[ ((2*Pop.Size+1):(3*Pop.Size)),
                ((Pop.Size+1):(2*Pop.Size))] <- diag(Diagonal)


# PUT EXTRA-DIAGONAL COMPONENTS
    Diagonal <- as.vector(diag(Var.Vect.Delta.Sample))
    Var.Vect.Delta.Sample <- Var.Vect.Delta.Sample +
                             t(Var.Vect.Delta.Sample)
    diag(Var.Vect.Delta.Sample) <- Diagonal


# OUTPUT
    Var.Vect.Delta.Sample

 }
```

```
# FUNCTION INIT.VAR.DIFF.2()
# ==========================
#
# This function initialises the function VAR.DIFF() variance estimator.
#
# INPUT:
#       * Vect.Y.0: the Nx1 vector of population values of the
#         survey variable for the first occasion.
#       * Vect.Y.1: the Nx1 vector of population values of the
#         survey variable for the second occasion.
#       * Vect.Pi.0: the Nx1 vector of the first-order inclusion
#         probabilities at the first occasion.
#       * Sample.Size: the sample size, n.
#       * Percentage.Overlap: the percentage overlap (common units)
#         between the two samples.
#


INIT.VAR.DIFF.2 <- function(Vect.Y.0, Vect.Y.1, Vect.Pi.0,
                            Sample.Size, Percentage.Overlap)  {

    Pop.Size <- length(Sample.0)
    Sample.Size.Common <- round(Percentage.Overlap * Sample.Size, digits=0)
    Size.Prod.Sample <- as.numeric(sum(Sample.0*Sample.1))
    Prob.In <- Sample.Size.Common / Sample.Size
    Prob.Out <- (Sample.Size - Sample.Size.Common) /
                (Pop.Size - Sample.Size)
    Vect.Pi.1 <- Prob.In * Vect.Pi.0 + Prob.Out * (1 - Vect.Pi.0)
    Vect.Y.0.Breve <- Vect.Y.0 / Vect.Pi.0
    Vect.Y.1.Breve <- Vect.Y.1 / Vect.Pi.1


# CREATION OF THE MATRIX A
    NB.Col.Matrix.A <- 5
    NB.Row <- 3 * Pop.Size
    Nb.Col <- NB.Col.Matrix.A
    Matrix.A <- matrix(rep(0,times=NB.Row*Nb.Col),nrow=NB.Row,ncol=Nb.Col)
    Vect.Z <- rep(1,times=Pop.Size)


# PUT THE VECTOR Y AT T = 0
    Matrix.A[(1:Pop.Size),1] <- Vect.Y.0.Breve


# PUT THE VECTOR Y AT T = 1
    Matrix.A[((Pop.Size+1):(2*Pop.Size)),2] <- Vect.Y.1.Breve


# PUT THE DESIGN VARIABLE AT T = 0 (CORRESPOND TO n0)
    Matrix.A[(1:Pop.Size),3] <- Vect.Z


# PUT THE DESIGN VARIABLE AT T = 1 (CORRESPOND TO n1)
    Matrix.A[((Pop.Size+1):(2*Pop.Size)),4] <- Vect.Z
```

```
# PUT THE DESIGN VARIABLE AT T = 0 (CORRESPOND TO n01)
    Matrix.A[((2*Pop.Size+1):(3*Pop.Size)),5] <- Vect.Z


# OUTPUT
    Matrix.A

  }
```

# References

**Andersson, C. and Nordberg, L. (1994)**: A method for variance estimation of non-linear function of totals in surveys - theory and a software implementation. *Journal of Official Statistics* **10**, 395–405.

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

**Berger, Y. G. (2004a)**: *Variance Estimation for Measures of Change in Probability Sampling.* To appear in the Canadian Journal of Statistics in December 2004. `http://www.mat.ulaval.ca/rcs/`.

**Berger, Y. G. (2004b)**: A simple variance estimator for unequal probability sampling without replacement. *Journal of Applied Statistics* **31**, 305–315.

**Brewer, K. R. W., Early, L. J. and Joyce, S. F. (1972)**: Selecting several samples from a single population. *Australian and New Zealand Journal of Statistics* **14**, 231–239.

**Chen, X. H., Dempster, A. P. and Liu, S. L. (1994)**: Weighted finite population sampling to maximise entropy. *Biometrika* **81**, 457–469.

**Fuller, W. A. and Fuller, J. N. K. (2001)**: A regression composite estimator with application to the Canadian Labour Force Survey. *Survey Methodology* **27**, 45–52.

**Hájek, J. (1959)**: Optimum strategy and other problems in probability sampling. *Asopis Pro Pestovani Mathematiky (Journal for Cultivation of Mathematics)* **84**, 387–423.

**Hájek, J. (1964)**: Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics* **35**, 1491–1523.

**Hájek, J. (1981)**: *Sampling from a Finite Population.* New York, Marcel Dekker.

**Hidiroglou, M. A., Särndal, C. E. and Binder, D. A. (1995)**: *Weighting and Estimation in Business Surveys.* In: Business Survey Methods, Wiley.

**Horvitz, D. G. and Thompson, D. J. (1952)**: A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47**, 663–685.

**Nordberg, L. (2000)**: On variance estimation for measure of change when samples are coordinated by the use of permanent random numbers. *Journal of Official Statistics* **16**, 363–378.

**Ohlsen, L. (1990)**: Sequential sampling from business register and its application to the Swedish Consumer Price Index. Technical report, R&D Report 1990:6, Stockholm: Statistics Sweden.

**Wolter, K. M. (1979)**: Composite estimation in finite population. *Journal of the American Statistical Association* **74**, 604–613.