DACSEIS

IST-2000-26057

Workpackage 9

Variance Estimation for Change

Deliverable 9.1

2003

List of contributors:

Yves Berger and Chris Skinner, University of Southampton.

Main responsibility:

Yves G. 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 estimation of changes between two waves of a survey (Smith *et al*., 2003), for example the change in the number of unemployed or in the unemployment rate. In this report, we consider three variance estimators of change: the Kish variance estimator (Kish, 1965), the Tam variance estimator (Tam, 1984) and a novel variance estimator (Berger, 2004). We compare these estimators by a simulation-based approaches based on the 2000 Finish Labour Force survey. Variance estimation of change needs to take account of rotation schemes. Three rotations will be considered: rotation with simple random sampling, rotation group sampling and rotation with systematic sampling.

Yves Berger and Chris Skinner — Southampton

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Most surveys 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 population parameters change from one survey period (or wave) to the next. Estimates of change are usually computed from estimates of variance and covariance between cross-sectional estimators at different waves (see (2.4)). Variance and covariance estimation would be relatively straightforward if the sample remained the same from one wave to the next. Unfortunately, this is rarely the case, as samples at different waves are usually overlapping sets of units. In Section 4.1 and 4.2, we define standard estimators and in Section 4.3, we propose a novel estimator for the variance of change. We compare these three variance estimators via a series of simulation based on the Finish Labour Force survey universe (see workpackage 3).

Chapter 2

Measure of change

Suppose, we wish to estimate the absolute change

$$\Delta = \tau_1 - \tau_0$$

between two population totals at wave $t = 0$ and wave $t = 1$

$$\tau_0 = \sum_{i \in U} y_{0;i}$$

$$\tau_1 = \sum_{i \in U} y_{1;i}$$

where $y_{0;i}$ and $y_{1;i}$ are respectively values of a study variable at time $t = 0$ and $t = 1$. The set U is a population frame assumed to be the same at both wave. Suppose that Δ is estimated by

$$\widehat{\Delta} = \widehat{\tau}_1 - \widehat{\tau}_0. \tag{2.1}$$

where $\widehat{\tau}_0$ and $\widehat{\tau}_1$ are Horvitz-Thompson (HT) estimators (Narain, 1951; Horvitz and Thompson, 1952) given by

$$\widehat{\tau}_0 = \sum_{i \in s_0} \frac{y_{0;i}}{\pi_{0;i}} \tag{2.2}$$

$$\widehat{\tau}_1 = \sum_{i \in s_1} \frac{y_{1;i}}{\pi_{1;i}} \tag{2.3}$$

where s_0 and s_1 denote two overlapping samples reported at $t = 0$ and $t = 1$. The quantities $\pi_{0;i}$ and $\pi_{1;i}$ are the inclusion probabilities of unit i at $t = 0$ and $t = 1$.

The variance of (2.1) is given by

$$\text{var}(\widehat{\Delta}) = \text{var}(\widehat{\tau}_1) + \text{var}(\widehat{\tau}_0) - 2\,\text{cov}(\widehat{\tau}_1, \widehat{\tau}_0) \tag{2.4}$$

Standard estimators can be used to estimate $\text{var}(\widehat{\tau}_0)$ and $\text{var}(\widehat{\tau}_1)$. The covariance can be estimated from the correlation estimated from the matched sample (see Section 4.1). However, if a large correlation is slightly over-estimated, the resulting estimator can under-estimate (2.4) significantly (see Section 4.1). In Section 4.3, we propose an alternative estimator for the covariance.

Chapter 3

Rotation schemes considered

Let n_0, n_1 and n_{01} denote respectively the size of the sample s_0 at $t = 0$, the sample s_1 at $t = 1$ and the common sample $s_{01} = s_0 \cap s_1$. We assume that the sample sizes n_0, n_1 and n_{01} are fixed. This assumption holds for most rotating sampling schemes. However, there exists rotating scheme with random sizes (Brewer *et al*, 1972; Ohlsen, 1990). We will not consider rotation schemes with random sizes. In this case, we suggest using Nordberg's approach (Nordberg, 2000, page 367). Unless otherwise stated, the sizes n_0, n_1 and n_{01} are assumed fixed.

A rotation scheme is often characterized by the fraction of the matched sample $g = n_{01}/n_0$. For example, with the Canadian Labour Force survey $g = 0.8$ and for the Finish Labour Force survey $g = 0.6$.

3.1 Rotation with simple random sampling

Assume that s_0 is a probability sample without replacement with first-order inclusion probabilities $\pi_{0;i}$. Suppose that s_1 is a simple random sample without replacement (srswor) of n_{01} units selected without replacement from s_0 combined with a srswor of $n_{0|1}$ units selected without replacement from U/s_0; where U/s_0 is the set of units not selected at $t = 0$ and $n_{0|1} = n_1 - n_{01}$. Tam (1984) studied the efficiency of this scheme. Under this scheme, the matched samples s_{01} contains n_{01} units, $n_1 = n_{01} + n_{0|1}$ and $g = n_{01}/n_0$.

The inclusion probability $\pi_{1;i}$ is

$$\pi_{1;i} = g\pi_{0;i} + q(1 - \pi_{0;i}),$$

where $q = n_{0|1}/(N - n_0)$ denotes the probability for a non-sampled unit to join the sample at $t = 1$.

3.2 Rotation group sampling

Suppose that the population is randomly divided into P mutually exclusive rotation groups of same size. At $t = 0$, the first p groups are selected ($p < P$). At $t = 1$,

group 1 rotates out and group $p+1$ rotates in. s_0 and s_1 are therefore simple random samples without replacement (srswor) with $\pi_{0;i} = p/P$, $n_0 = pN/P$, $n_{01} = (p-1)N/P$ and $n_{1|0} = N/P$. This implies $g = (p-1)/p$. Business surveys in Statistics Canada are selected with rotation group sampling.

3.3 Rotation with systematic sampling

Suppose that a different systematic sample is selected at each wave and each systematic sample is retained in the survey for p consecutive occasions. The same number of units rotates in and out and it provides a fixed degree of overlap $g = (p-1)/p$ between s_0 and s_1. This method is used by the British Labour Force Survey (HOLMES and SKINNER, 2000). The systematic sampling (sys) design (MADOW and MADOW, 1944) is widely used by statistical offices due to its simplicity and efficiency (e.g. IACHAN, 1982). For example, it can take into account of the hidden stratification in the population (WOLTER, 1985).

3.4 Stratified rotating scheme

For stratified population, we select a stratified sample at $t = 0$ and at $t = 1$, within each stratum we select the second sample according to a rotating scheme; that is, we have a stratum-by-stratum rotation scheme. For the simulation, stratified rotation schemes will be considered. As far as the rotation with systematic sampling is concerned, it is not necessary to select systematic samples stratum-by-stratum, as the systematic sampling design takes the stratification into account. With the rotation with systematic sampling, we select systematic samples across strata.

Chapter 4

Variance estimation

In the following section, we define three estimators for the variance (2.4). These estimators can be computed using the `Splus` library `rot` available at

`http://www.socstats.soton.ac.uk/staff/berger/change.html`

This library is also available for R.

A variance estimator of change can have a large negative bias when the correlation is strong. This can be interpreted the following way. Let $\widehat{\mathrm{var}}(\widehat{\Delta})$ and $\widehat{\rho}$ be any estimator for (2.4) and ρ. Assuming $\mathrm{var}(\widehat{\tau}_0) = \mathrm{var}(\widehat{\tau}_1) = v^2$, we have $\mathrm{E}(\widehat{\mathrm{var}}(\widehat{\Delta})) \approx 2v^2(1 - \mathrm{E}(\widehat{\rho}))$. Let $RB(\widehat{\rho})$ be the relative bias of $\widehat{\rho}$, i.e $RB(\widehat{\rho}) = (\mathrm{E}(\widehat{\rho}) - \rho)\rho^{-1}$. As $\mathrm{E}(\widehat{\rho}) = \rho\,(RB(\widehat{\rho}) + 1)$, the relative bias of $\widehat{\mathrm{var}}(\widehat{\Delta})$, i.e. $RB(\widehat{\mathrm{var}}(\widehat{\Delta})) = (\mathrm{E}(\widehat{\mathrm{var}}(\widehat{\Delta})) - \mathrm{var}(\widehat{\Delta}))\mathrm{var}(\widehat{\Delta})^{-1}$ is approximately

$$RB(\widehat{\mathrm{var}}(\widehat{\Delta})) \approx RB(\widehat{\rho})\frac{\rho}{\rho - 1} \tag{4.1}$$

Thus, a small positive relative bias for $\widehat{\rho}$ implies a large negative relative bias for $\widehat{\mathrm{var}}(\widehat{\Delta})$ for large correlations often observed in practice.

4.1 Kish's variance estimator

A naïve way of estimating a covariance would be to estimate it from the correlation (Kish, 1965, Holmes and Skinner, 2000)

$$\widehat{\rho}^{(m)} = \frac{\widehat{\mathrm{cov}}_{01}(\widehat{\tau}_0^{(m)}, \widehat{\tau}_1^{(m)})}{\left(\widehat{\mathrm{var}}_{01}(\widehat{\tau}_0^{(m)})\widehat{\mathrm{var}}_{01}(\widehat{\tau}_1^{(m)})\right)^{1/2}} \tag{4.2}$$

(the superscript m means "matched") where

$$\widehat{\mathrm{cov}}_{01}(\widehat{\tau}_0^{(m)}, \widehat{\tau}_1^{(m)}) = \frac{1}{2}\left(\widehat{\mathrm{var}}_{01}(\widehat{\tau}_0^{(m)}) + \widehat{\mathrm{var}}_{01}(\widehat{\tau}_1^{(m)}) - \widehat{\mathrm{var}}_{01}(\widehat{\tau}_2^{(m)})\right)$$

$$\widehat{\text{var}}_{01}(\widehat{\tau}_\ell^{(m)}) = \frac{n_{01}}{n_{01}-1}\sum_{i\in s_{01}}\left(\check{y}_{\ell;i} - \frac{1}{n_{01}}\sum_{j\in s_{01}}\check{y}_{\ell;j}\right)^2 \quad (\ell = 0,\ 1,\ 2) \tag{4.3}$$

$$\widehat{\tau}_\ell^{(m)} = \sum_{i\in s_{01}}\check{y}_{\ell;i} \quad (\ell = 0, 1)$$

$\check{y}_{0;i} = y_{0;i}\pi_{0;i}^{-1}$, $\check{y}_{1;i} = y_{1;i}\pi_{1;i}^{-1}$ and $\check{y}_{2;i} = \check{y}_{1;i} - \check{y}_{0;i}$. $\widehat{\text{var}}_{01}(\widehat{\tau}_\ell^m)$ is the standard variance estimator based on with-replacement sampling (HANSEN and HURWITZ, 1943). An estimator for the covariance is given by the Kish estimator (KISH, 1965, pp. 457)

$$\widehat{\text{cov}}_K(\widehat{\tau}_0, \widehat{\tau}_1) = \left(\widehat{\text{var}}_K(\widehat{\tau}_0)\widehat{\text{var}}_K(\widehat{\tau}_1)\right)^{1/2}\widehat{\rho}^m \tag{4.4}$$

where $\widehat{\text{var}}_K(\widehat{Y}_0)$ and $\widehat{\text{var}}_K(\widehat{Y}_1)$ are the standard with replacement variance estimator.

$$\widehat{\text{var}}_K(\widehat{\tau}_\ell) = \frac{n_\ell}{n_\ell - 1}\sum_{i\in s_\ell}\left(\frac{y_{\ell;i}}{\pi_{\ell;i}} - \frac{1}{n_\ell}\sum_{j\in s_\ell}\frac{y_{\ell;j}}{\pi_{\ell;j}}\right)^2 \quad (\ell = 0,\ 1)$$

The resulting estimator for (2.4) is

$$\widehat{\text{var}}_K(\widehat{\Delta}) = \widehat{\text{var}}_K(\widehat{\tau}_0) + \widehat{\text{var}}_K(\widehat{\tau}_1) - 2\,\widehat{\text{cov}}_K(\widehat{\tau}_0, \widehat{\tau}_1). \tag{4.5}$$

Furthermore, as (4.2) is a biased estimate for the correlation ρ between (2.2) and (2.3), (4.5) can over-estimate significantly the variance. BERGER (2004) shows that the Kish estimator can have a large negative bias when the correlation is strong (see *10). Note that a positive bias for (4.2) is not uncommon in practice, as the correlation tends to be stronger among units from the common sample s_{01}.

4.2 Tam's variance estimator

Assuming simple random sampling, TAM (1984) used the same methodology as Kish, for estimating the covariance, but after removing the assumption of a large population. He obtained the same expression for the sampling variance, but with finite population correction (FPC). The covariance estimator proposed by Tam is given by

$$\widehat{\text{cov}}_T(\widehat{\tau}_0,\ \widehat{\tau}_1) = \left(1 - \frac{n_0\ n_1}{N\ n_{01}}\right)\widehat{\text{cov}}_K(\widehat{\tau}_0, \widehat{\tau}_1)$$

where $\widehat{\text{cov}}_K(\widehat{\tau}_0, \widehat{\tau}_1)$ is given by (4.4). $\text{var}(\widehat{\tau}_0)$ and $\text{var}(\widehat{\tau}_1)$ can be estimated with the modified Hansen-Hurwitz variance estimator (HANSEN and HURWITZ, 1943) given by

$$\widehat{\text{var}}_T(\widehat{\tau}_\ell) = \left(1 - \frac{n_\ell}{N}\right)\widehat{\text{var}}_K(\widehat{\tau}_\ell) \quad (\ell = 0,\ 1) \tag{4.6}$$

The resulting estimator for (2.4) is

$$\widehat{\text{var}}_T(\widehat{\Delta}) = \widehat{\text{var}}_T(\widehat{\tau}_0) + \widehat{\text{var}}_T(\widehat{\tau}_1) - 2\,\widehat{\text{cov}}_T(\widehat{\tau}_0, \widehat{\tau}_1). \tag{4.7}$$

4.3 Berger's variance estimator

This estimator is based on the conditional Poisson approach (Hájek, 1964, and Hájek, 1981). Let us assume that the sample sizes n_0, n_1 and n_{01} are fixed. First, we estimate the variances and covariances unconditionally by assuming that s_0 and s_1 are two independent Poisson samples. In order to capture the fixed sizes, we will derive the conditional variance and covariance given numbers of units caught in s_0, s_1 and s_{01}.

Assumes that the following vector has a multivariate normal distribution

$$\mathbf{u} = (\widehat{\tau}_0, \widehat{\tau}_1, n_0, n_1, n_{01})^T \sim N(\mu_{\mathbf{u}}, \mathbf{\Sigma_u})$$

with respect to a Poisson sampling scheme, where s_0 and s_1 are selected independently with Poisson sampling with inclusion probabilities $\pi_{0;i}$ and $\pi_{1;i}$. The variance-covariance matrix $\mathbf{\Sigma_u}$ is defined as

$$\mathbf{\Sigma_u} = \begin{pmatrix} \mathbf{\Sigma}_{\tau\tau} & \mathbf{\Sigma}_{\tau\mathbf{n}} \\ \mathbf{\Sigma}_{\tau\mathbf{n}}^T & \mathbf{\Sigma_{nn}} \end{pmatrix}$$

$\mathbf{\Sigma}_{\tau\tau}$ is the variance-covariance of the vector $\tau = (\widehat{\tau}_0, \widehat{\tau}_1)^T$. $\mathbf{\Sigma_{nn}}$ is the variance-covariance matrix of

$$\mathbf{n} = (n_0, n_1, n_{01})^T. \tag{4.8}$$

$\mathbf{\Sigma}_{\tau\mathbf{n}}$ is the covariance between τ and $\mathbf{n}$.

In order to capture the fixed sizes, we will derive the conditional variance and covariance given numbers of units caught in s_0, s_1 and s_{01}; that is, the variance of τ conditionally on $\mathbf{n}$:

$$\mathbf{\Sigma}_{\tau\tau|\mathbf{n}} = \mathbf{\Sigma}_{\tau\tau} - \mathbf{\Sigma}_{\tau\mathbf{n}}\mathbf{\Sigma}_{\mathbf{nn}}^{-1}\mathbf{\Sigma}_{\tau\mathbf{n}}^T = \begin{pmatrix} \mathrm{var}(\widehat{\tau}_0|\mathbf{n}) & \mathrm{cov}(\widehat{\tau}_0, \widehat{\tau}_1|\mathbf{n}) \\ \mathrm{cov}(\widehat{\tau}_0, \widehat{\tau}_1|\mathbf{n}) & \mathrm{var}(\widehat{\tau}_1|\mathbf{n}) \end{pmatrix}$$

which can be estimated by

$$\widehat{\mathbf{\Sigma}}_{\tau\tau|\mathbf{n}} = \widehat{\mathbf{\Sigma}}_{\tau\tau} - \widehat{\mathbf{\Sigma}}_{\tau\mathbf{n}}\widehat{\mathbf{\Sigma}}_{\mathbf{nn}}^{-1}\widehat{\mathbf{\Sigma}}_{\tau\mathbf{n}}^T$$

Where

$$\widehat{\mathbf{\Sigma}}_{\mathbf{u}} = \begin{pmatrix} \widehat{\mathbf{\Sigma}}_{\mathbf{yy}} & \widehat{\mathbf{\Sigma}}_{\mathbf{yn}} \\ \widehat{\mathbf{\Sigma}}_{\mathbf{yn}}^T & \widehat{\mathbf{\Sigma}}_{\mathbf{nn}} \end{pmatrix},$$

is the standard estimator of $\mathbf{\Sigma_u}$ under Poisson Sampling. The diagonal components give $\widehat{\mathrm{var}}(\widehat{\tau}_0|\mathbf{n})$ and $\widehat{\mathrm{var}}(\widehat{\tau}_1|\mathbf{n})$. The extra-diagonal component gives $\widehat{\mathrm{cov}}(\widehat{\tau}_0, \widehat{\tau}_1|\mathbf{n})$.

The estimator for the covariance is

$$\widehat{\mathrm{cov}}_B(\widehat{\tau}_0, \widehat{\tau}_1) = \left(\widehat{\mathrm{var}}_B(\widehat{\tau}_0)\widehat{\mathrm{var}}_B(\widehat{\tau}_1)\right)^{1/2} \widehat{\rho}_n$$

where

$$\widehat{\rho}_n = \frac{\widehat{\mathrm{cov}}_B(\widehat{\tau}_0, \widehat{\tau}_1|\mathbf{n})}{\left(\widehat{\mathrm{var}}(\widehat{\tau}_0|\mathbf{n})\widehat{\mathrm{var}}(\widehat{\tau}_1|\mathbf{n})\right)^{1/2}}.$$

Estimators for $\text{var}(\widehat{\tau}_\ell)$ and $(\ell = 0,\ 1)$ can be derived in similar way by estimating the conditional variance $\text{var}(\widehat{\tau}_\ell | n_\ell)$. The resulting variance estimators is the Hájek variance estimator (HÁJEK, 1964) given by

$$\widehat{\text{var}}_B(\widehat{\tau}_\ell) = \sum_{i \in s_\ell} (1 - \pi_{\ell;i}) \left(\frac{y_{\ell;i}}{\pi_{\ell;i}} - \frac{1}{d_\ell} \sum_{j \in s_\ell} (1 - \pi_{\ell;j}) \frac{y_{\ell;j}}{\pi_{\ell;j}} \right)^2 \quad (\ell = 0,\ 1)$$

where $d_\ell = \sum_{i \in s_\ell} (1 - \pi_{\ell;i})$.

The resulting estimator for (2.4) is

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

A series of simulations in BERGER (2004) shows that $\widehat{\text{var}}_B(\widehat{\Delta})$ is more accurate than $\widehat{\text{var}}_K(\widehat{\Delta})$ and $\widehat{\text{var}}_T(\widehat{\Delta})$.

4.4 Variance estimation with stratified rotating schemes

The variance estimators (4.3) and (4.6) can be easily generalised for stratified sampling scheme. It is only necessary to estimate the stratum-by-stratum variance. Thus, the (4.5) and (4.7) estimators can be easily generalised for stratification. The Berger estimator can be generalised for stratification (Berger, 2003). It is only necessary to include all the fixed stratum sizes in the vector $\mathbf{n}$; that is, instead of (4.8), we consider

$$\mathbf{n} = (n_{0;1}, n_{1;1}, n_{01;1}, \ldots, n_{0;h}, n_{1;h}, n_{01;h}, \ldots, n_{0;H}, n_{1;H}, n_{01;H})^T,$$

where $n_{0;h}$ is the number of units in the h-th stratum at $t = 0$, $n_{1;h}$ is the number of units in the h-th stratum at $t = 1$ and $n_{01;h}$ is the number of units in the common sample of the h-th stratum.

BERGER (2004) also shows that the Berger estimator can be extended for temporal stratification; that is, when stratification changes between waves; that is, when units move between strata and new strata are created. The `Splus` library `rot` can take this kind of stratification into account. Nevertheless, temporal stratification will not be considered in the simulation study.

Chapter 5

The Finish LFS pseudo-universe

The Finnish Labour Force Survey (FLFS) is a systematic sample from the Central Population Register. The individuals from this register are the sampling units. Based on these data, a pseudo Universes has been created (see workpackage 3). The number of individuals in the pseudo universe is 3 900 000.

The variables available are given in Table 5.1. Data are available for two waves (February 2000 and May 2000); that is, we know the values of the variables "lfstat"; "isced" at both waves. The variables "region", "gender" and "age" are assumed to be the same at both waves. The variable "region" specifies the stratification. The variables "region", "gender" and "age" will be used for the response mechanism. The variables "lfstat" and "isced" will be the variables of interest; that is, for example, we will estimate the change in the total number of employed, unemployed, individuals with upper secondary education, etc ...

Table 5.1: Variables included in the Finnish pseudo universes.

Region:			
	0.Uusimaa		3.Mid-Finland
	1.Southern Finland		4.Northern Finland
	2.Eastern Finland		5.Aland
gender:			
	0...male		
	1...female		
labour force characteristics (lfstat) (Feb & May 2000):			
	0...employed		4...disabled
	1...unemployed		5...pensioners
	2...conscripts		6...domestic work
	3...students		7...others
highest level of education (isced) (Feb & May 2000):			
0...no answer		4...5A-programmes	
1...upper secondary education		5...second stage	
2...post secondary		6...level unknown	
3...5B-programmes			
age:			
0...15-19	3...30-34	6...45-49	9...60-64
1...20-24	4...35-39	7...50-54	10..65-69
2...25-29	5...40-44	8...55-59	11..70-74

Chapter 6

The response mechanism

In Chapter 7, we will analyse the sampling distribution of the variance estimators among 1000 samples randomly selected according to the three rotation schemes described in Chapter 3. For each sample selected, total non-response will be generated randomly according to probabilities of response assumed unknown for the inference. Thus, these probabilities will not be used for variance and point estimation. The response mechanism gives a response rate between 85% and 88% on both waves.

Kaja Sõstra and Seppo Laaksonen (Statistics Finland) have fitted a model for the response based on the FLFS data. This response mechanism is described in Section 6.1 and Section 6.2.

6.1 Response Mechanism for the first wave and non-overlapping part of the second wave

Based on the FLFS data, the response probability for a unit of the first wave and non-overlapping part of the second wave is given by

$$pr = \frac{\exp(\text{logit})}{1 + \exp(\text{logit})} \tag{6.1}$$

where

$$\begin{aligned}\text{logit} = {} & 2.6349 - 0.2114\ (\text{gender} = 0) + 0.2715\ (\text{age10} = 1) - 0.3406\ (\text{age10} = 2) \\ & - 0.3827\ (\text{age10} = 3) - 0.2664\ (\text{age10} = 4) - 0.1789\ (\text{age10} = 5) \\ & - 0.5342\ (\text{region2} = 0) - 0.1062\ (\text{region2} = 1) + 0.1633\ (\text{region2} = 2) \\ & + 0.1606\ (\text{region2} = 3) - 0.6218\ (\text{isced2} = 0) - 0.2429\ (\text{isced2} = 1) \\ & + 0.1037\ (\text{isced2} = 3)\end{aligned}$$

where "age10" is the 10-year age group variable; that is, age10 is the integer part of (age+2)/2). "region2" is the variable "region" with the region 1 and 5 grouped as a single

region (region 1). isced2 = 0 if isced = 0; isced2 = 1 if isced = 1 or 2, isced2 = 3 if isced = 3 and isced2 = 4 if isced = 4 or 5. "(gender = 0)", "(age10 = 1)",... are dummy variable. For example, (gender = 0) = 1 if gender = 0 and 0 otherwise.

For each unit $i \in s_0 \cup (s_1/s_{01})$, we compute pr given by (6.1). Let $unif$ be a random number generated from the uniform distribution $U(0,1)$. If $i \in s_0$ and $unif < pr$, the unit is a respondent at $t = 0$. If $i \in s_1$ and $unif < pr$, the unit is a respondent at $t = 1$.

6.2 Response Mechanism for the overlapping part of the second wave

The response mechanism for the overlapping part of the second wave uses conditional response probabilities based on the FLFS data.

For each unit $i \in s_{01}$, we compute pr_2 given by

$$pr_2 = \frac{\exp(\text{logit}_2)}{1 + \exp(\text{logit}_2)}$$

where

$$\text{logit}_2 = 3.5071 - 0.3849\ (\text{gender} = 0).$$

Let $unif_2$ be a random number generated from the uniform distribution $U(0,1)$. If $i \in s_{01}$ is a respondent at $t = 0$ and if $unif_2 < pr_2$, i is also a respondent at $t = 1$. If $i \in s_{01}$ is not a respondent at $t = 0$ and if $unif_2 < 0.182$, i is a respondent at $t = 1$.

Chapter 7

Monte-Carlo simulation

We propose three series of simulations. For the first series, we select 1000 samples with a rotation with simple random sampling. For the second series, we select 1000 samples with the rotation group sampling scheme. For the third series, we select 1000 samples with a rotation with systematic sampling. These rotating sampling schemes with rotation groups and with simple random sampling are stratified according to the variable "region". $n_0 = 10\ 000$ units are selected at $t = 0$ according to a stratified simple random sampling with proportional allocation. As the overlapping part between two sequential quarters in the Finnish LFS is 60%, we consider $g = 0.6$. For each sample selected, we compute $\widehat{\Delta}$ and the three estimators for the variance proposed: $\widehat{\text{var}}_K(\widehat{\Delta})$ ("Kish"), $\widehat{\text{var}}_T(\widehat{\Delta})$ ("Tam") and $\widehat{\text{var}}_B(\widehat{\Delta})$ ("Berger").

A set of respondents will be selected according to the response mechanism described in Chapter 6. For simplicity, we will denote by s_0 the set of respondents at $t = 0$ and s_1 the set of respondents at $t = 1$. The total non-response is taken into account by multiplying the first inclusion probabilities $\pi_{0;i}$ and $\pi_{1;i}$ by the observed response rate at $t = 0$ and $t = 1$. Due to non-response, the sizes n_0, n_1 and n_{01} are now random. However, as shown in Table 7.1, the response rate is high (between 85% and 88%), we will assume that the sizes are fixed and we will estimate the variance with the estimator described in Chapter 4.

Table 7.1: The minimum and maximum response rates at wave 1 and 2 for three rotation schemes.

Rotation with...	Response rates			
	Wave 1		Wave 2	
	Min	Max	Min	Max
SRSWOR	0.85	0.87	0.85	0.87
systematic sampling	0.86	0.88	0.85	0.88
rotation group sampling	0.86	0.88	0.85	0.87

We are interested in the change in total number of employed, unemployed, conscripts, students, disabled, pensioners, persons performing domestic work, individuals with other

labour force status, individuals with upper secondary education, post secondary non-tertiary education, 5B-programmes education, 5A-programmes education, second stage of tertiary education, level unknown or individuals with no answer for the level of education. In this report, we only consider the change in the number of unemployed and students and pensioners. The simulation results for the other variables will be available in workpackage 1.

7.1 Change in the number of unemployed

The change the number of unemployed is -26.8% of the total number of unemployed at $t = 0$; that is, $\tau_1 = \tau_0(1 - 0.268)$. The change in the number of unemployed is therefore relatively large.

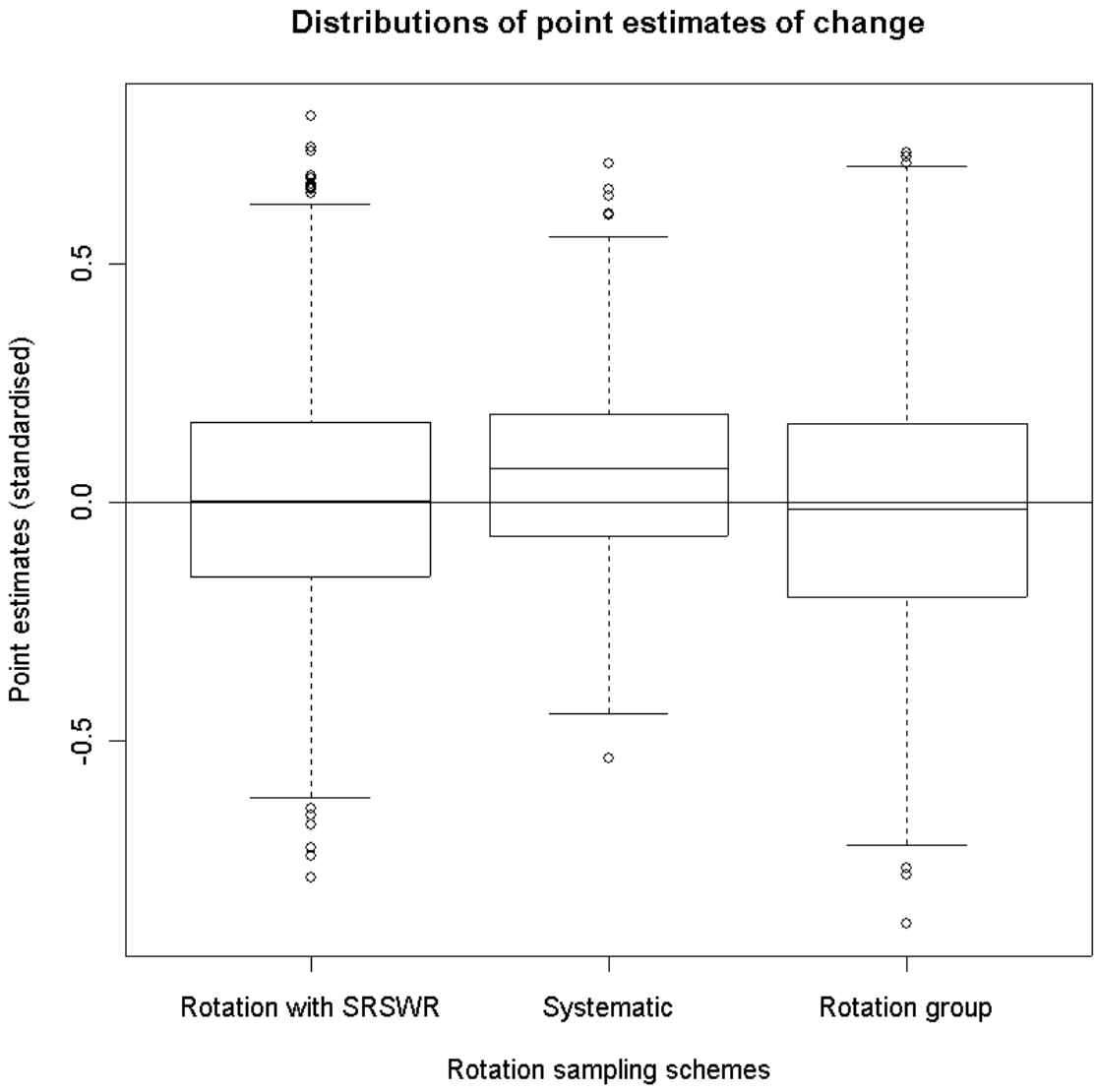

Figure 7.1: Sampling distribution of $\widehat{\Delta}_{st}$ for the three methods of sampling. τ = total number of unemployed.

In Figure 7.1, we have the empirical sampling distribution of $\widehat{\Delta}_{st} = (\widehat{\Delta} - \Delta)/\Delta$ for the three methods of sampling. The rotation with systematic sampling seems to give better point estimates.

In Figure 7.2, we have the empirical sampling distribution of $\widehat{\text{var}}(\widehat{\Delta})_{st} = (\widehat{\text{var}}(\widehat{\Delta}) - \text{var}(\widehat{\Delta}))/\text{var}(\widehat{\Delta})$ when $\widehat{\text{var}}(\widehat{\Delta})$ equals $\widehat{\text{var}}_K(\widehat{\Delta})$, $\widehat{\text{var}}_T(\widehat{\Delta})$ and $\widehat{\text{var}}_B(\widehat{\Delta})$. $\text{var}(\widehat{\Delta})$ is the empirical variance of $\widehat{\Delta}$. $\widehat{\text{var}}_B(\widehat{\Delta})$ gives more accurate variance estimators especially with systematic sampling. As expected (see Section 4.1), $\widehat{\text{var}}_K(\widehat{\Delta})$, $\widehat{\text{var}}_T(\widehat{\Delta})$ have a large negative bias. This is probably due to the fact that the correlation is over-estimated. Simulation in Berger (2004) also shows that $\widehat{\text{var}}_K(\widehat{\Delta})$ has a large negative bias. $\widehat{\text{var}}_K(\widehat{\Delta})$ and $\widehat{\text{var}}_T(\widehat{\Delta})$ seems to have a very similar distribution. This is due to the fact that the FPC is negligible (1 – 10 000 / 39 000 000 = 0.999). Thus, the difference in accuracy between $\widehat{\text{var}}_K(\widehat{\Delta})$ and $\widehat{\text{var}}_B(\widehat{\Delta})$ is not due to the FPC.

In Table 7.2, we have the relative bias ($RB = Bias(\widehat{\text{var}}(\widehat{\Delta}))\text{var}(\widehat{\Delta})^{-1}$) and the relative mean square error ($RMSE = MSE(\widehat{var}(\widehat{\Delta}))\text{var}(\widehat{\Delta})^{-2}$) of $\widehat{\text{var}}_K(\widehat{\Delta})$, $\widehat{\text{var}}_T(\widehat{\Delta})$ and $\widehat{\text{var}}_B(\widehat{\Delta})$. We have also the 95% confidence interval (CI) coverage using the normal assumption. $\widehat{\text{var}}_B(\widehat{\Delta})$ gives the best CI coverage.

The median of the time (in seconds) required for the computation of one variance estimate is given in Table 7.2. A 2GHz Pentium 4 CPUs with 1Gb of RAM has been used. Only 15 seconds is necessary for $\widehat{\text{var}}_K(\widehat{\Delta})$ and $\widehat{\text{var}}_T(\widehat{\Delta})$. $\widehat{\text{var}}_B(\widehat{\Delta})$ is a computing intensive estimator. It takes 44 minutes (2640 seconds) to compute one estimate. Note that this is the times required for the computation of the variance-covariance matrix of all the survey variables. Thus, 44 minutes is necessary to compute the variance of change between all the variables.

Table 7.2: Relative bias (RB), the relative mean square error (RMSE) and 95% confidence coverage of the variance estimator considered. "Time (med.)" is the median of the time for computing one estimate. τ= total number of unemployed.

Rotation with...		Kish	Tam	Berger
SRSWOR	RB	-0.32	-0.32	0.02
	RMSE	0.12	0.12	0.03
	Coverage	0.90	0.90	0.94
	Time (med.)	14 sec	14 sec	2732 sec
systematic sampling	RB	-0.22	-0.22	0.06
	RMSE	0.05	0.05	0.00
	Coverage	0.90	0.90	0.95
	Time (med.)	15 sec	15 sec	2636 sec
rotation group sampling	RB	-0.40	-0.41	-0.06
	RMSE	0.18	0.18	0.02
	Coverage	0.88	0.88	0.94
	Time (med.)	15 sec	15 sec	2634 sec

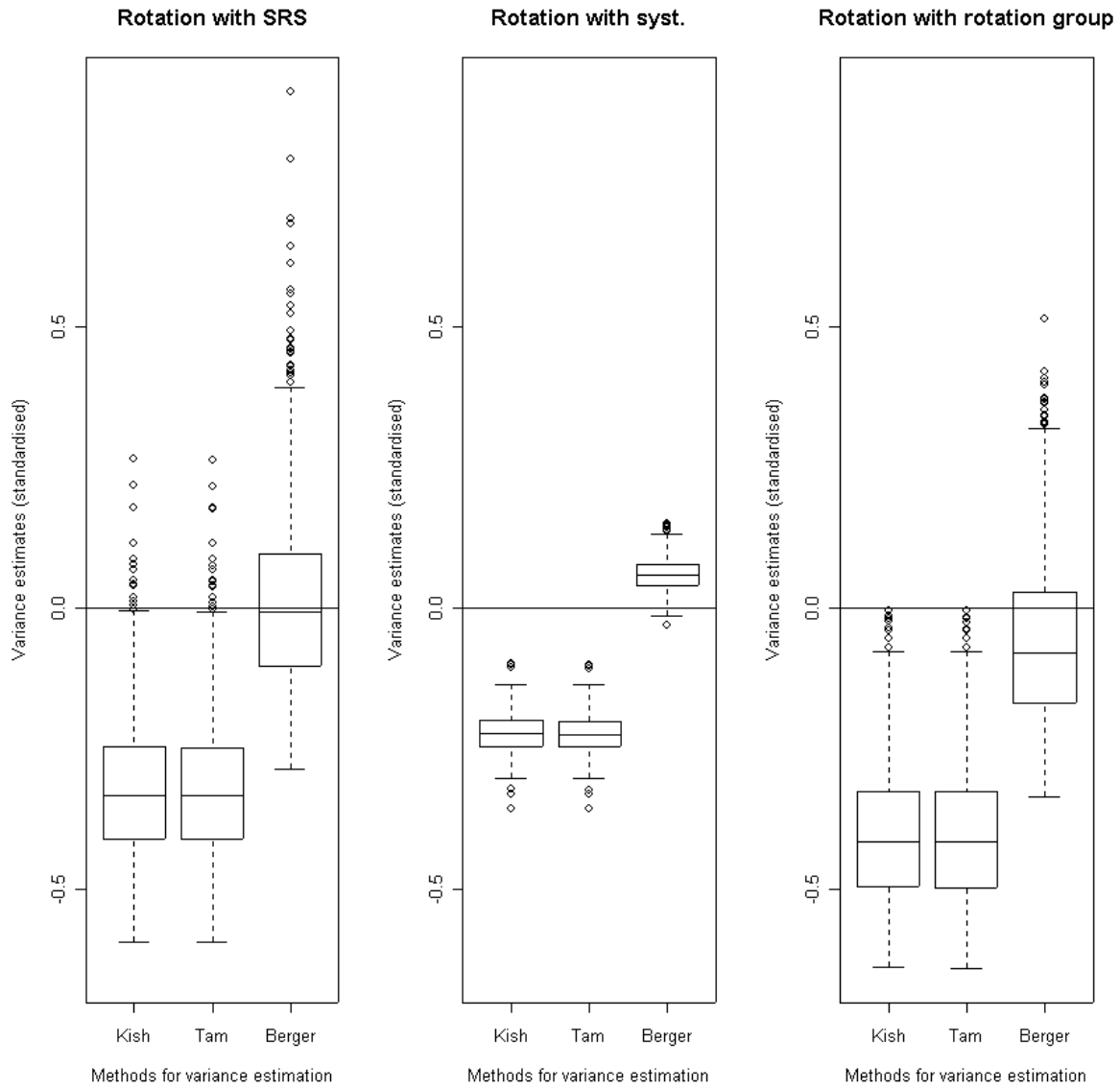

Figure 7.2: Empirical sampling distribution of $\widehat{\text{var}}(\widehat{\Delta})_{st}$ for the Kish, Tam and Berger variance estimators. $\tau=$ total number of unemployed.

7.2 Change in the number of Students

The change in the number of students is 41.4% of the total number of students at $t = 0$; that is, $\tau_1 = \tau_0(1+0.414)$. This change is therefore large. The empirical sampling distribution of $\widehat{\Delta}_{st}$ is Figure 7.3 shows that $\widehat{\Delta}$ is roughly unbiased and that systematic sampling gives better point estimates. The observed bias might be due to the non-response. We can draw the same conclusion as Section 7.1. However, the coverage in Table 7.3 is not as good as in Table 7.2.

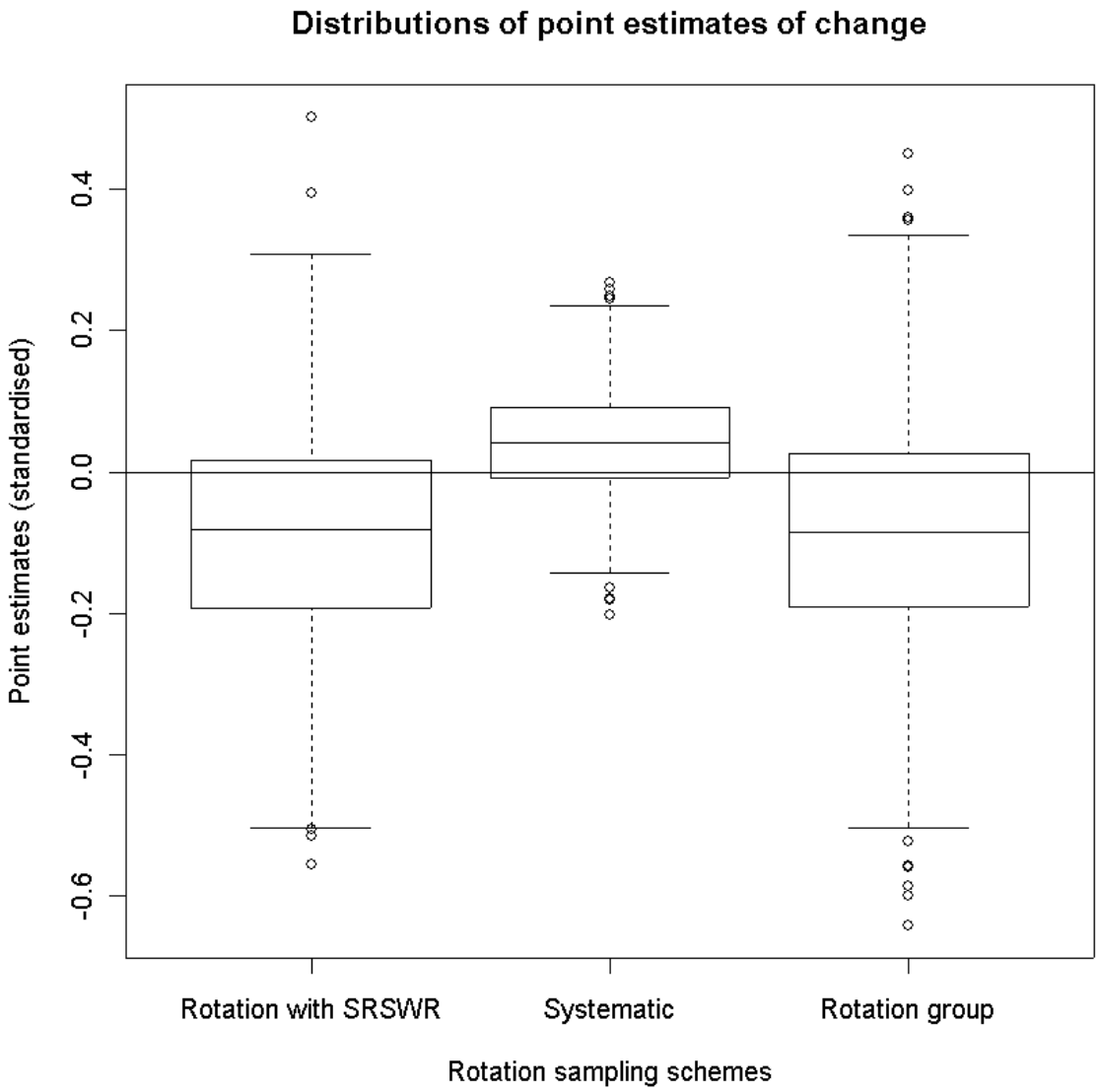

Figure 7.3: Sampling distribution of $\widehat{\Delta}_{st}$ for the three methods of sampling. $\tau=$ total number of students.

Table 7.3: Relative bias (RB), the relative mean square error (RMSE) and 95% confidence coverage of the variance estimator considered. "Time (med.)" is the median of the time for computing one estimate. τ= total number of students.

Rotation with...		Kish	Tam	Berger
SRSWOR	RB	-0.79	-0.79	-0.02
	RMSE	0.63	0.63	0.03
	Coverage	0.56	0.56	0.91
	Time (med.)	14 sec	14 sec	2732 sec
systematic sampling	RB	-0.36	-0.36	0.15
	RMSE	0.13	0.13	0.02
	Coverage	0.83	0.83	0.95
	Time (med.)	15 sec	15 sec	2636 sec
rotation group sampling	RB	-0.81	-0.81	0.01
	RMSE	0.66	0.66	0.03
	Coverage	0.54	0.54	0.92
	Time (med.)	15 sec	15 sec	2634 sec

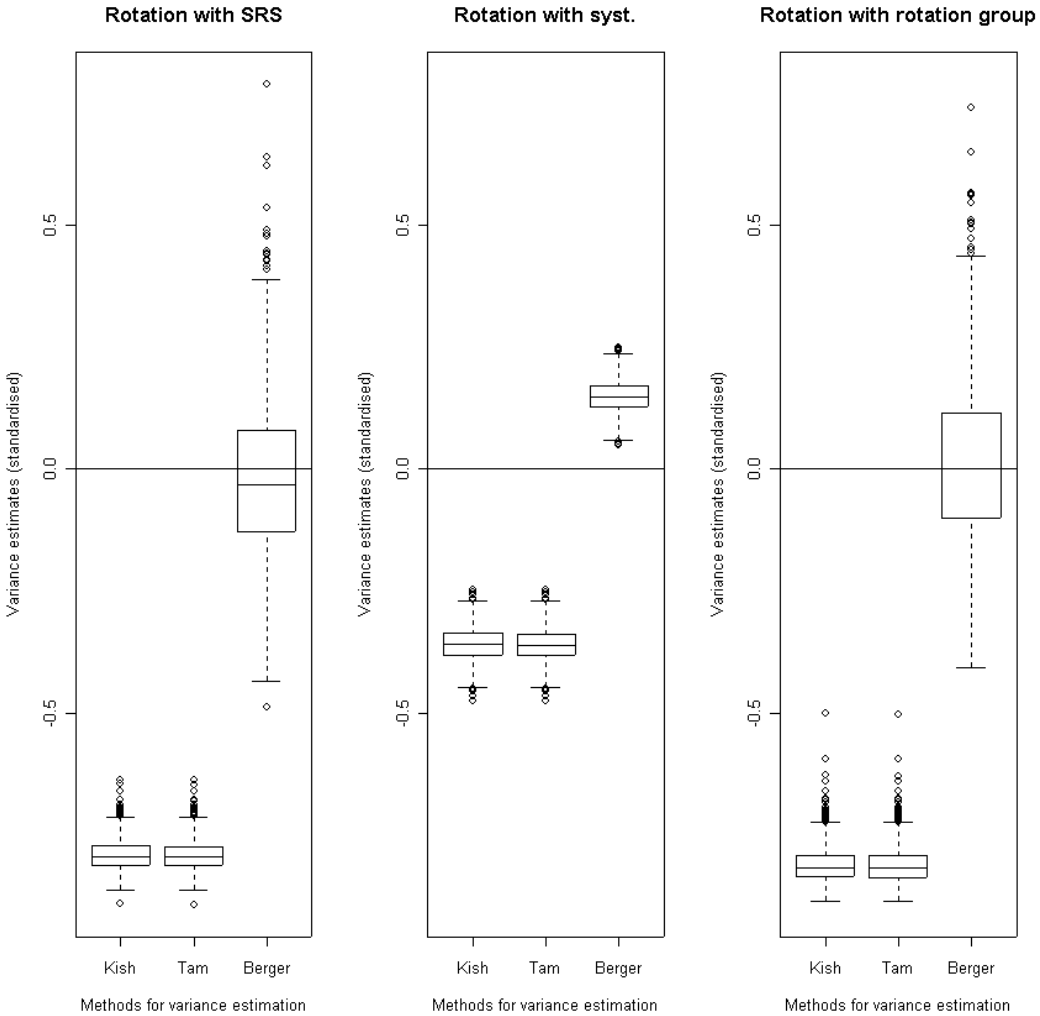

Figure 7.4: Empirical sampling distribution of $\widehat{\text{var}}(\widehat{\Delta})_{st}$ for the Kish, Tam and Berger variance estimators. τ= total number of students.

7.3 Change in the number of pensioners

The change in the number of pensioners is -0.4% of the total number of pensioners at $t = 0$; that is, $\tau_1 = \tau_0(1 - 0.004)$. This change is therefore negligible. The empirical sampling distribution of $\widehat{\Delta}_{st}$ is Figure 7.5 shows that $\widehat{\Delta}$ is roughly unbiased and that systematic sampling gives better point estimates. We can draw the same conclusion as Section 7.1. However, $\widehat{\text{var}}_K(\widehat{\Delta})$, $\widehat{\text{var}}_T(\widehat{\Delta})$ gives a very bad inference, as they under-estimate the variance significantly. Thus, they are not recommended when the change is negligible.

As the change in the number of pensioners is negligible, H_0: $\Delta = 0$ is true. Thus, we should not reject H_0, if we use a standard t test. In Table 7.5, we have the observed type I error α_{obs} for this test when the critical value is 1.96; that is, when reject H_0 if $|\widehat{\Delta}(\widehat{\text{var}}(\widehat{\Delta}))^{-1/2}| > 1.96$. We see that if we use $\widehat{\text{var}}_B(\widehat{\Delta})$, α_{obs} is close to 5%. With rotation with systematic sampling, α_{obs} is smaller as the variance is slightly over-estimated in this case (see Table 7.4). $\widehat{\text{var}}_K(\widehat{\Delta})$ and $\widehat{\text{var}}_T(\widehat{\Delta})$ gives a α_{obs} between 57% and 70%. Thus, if we base the inference on $\widehat{\text{var}}_K(\widehat{\Delta})$ or $\widehat{\text{var}}_T(\widehat{\Delta})$, we would reject H_0 with a very large probability although H_0 is true.

Table 7.4: Relative bias (RB), the relative mean square error (RMSE) and 95% confidence coverage of the variance estimator considered. "Time (med.)" is the median of the time for computing one estimate. τ= total number of pensioners.

Rotation with...		Kish	Tam	Berger
SRSWOR	RB	-0.95	-0.95	-0.01
	RMSE	0.90	0.90	0.03
	Coverage	0.33	0.33	0.94
	Time (med.)	14 sec	14 sec	2732 sec
systematic sampling	RB	-0.91	-0.91	0.12
	RMSE	0.83	0.83	0.02
	Coverage	0.43	0.43	0.96
	Time (med.)	15 sec	15 sec	2636 sec
rotation group sampling	RB	-0.96	-0.96	-0.04
	RMSE	0.91	0.91	0.03
	Coverage	0.30	0.30	0.94
	Time (med.)	15 sec	15 sec	2634 sec

Table 7.5: Observed type I error (α_{obs}) for the test H_0: $\Delta = 0$ H_1: $\widehat{\Delta} \neq 0$, if we use a standard t test; that is, we reject H_0 if $|\widehat{\Delta}(\widehat{\text{var}}(\widehat{\Delta}))^{-1/2}| > 1.96$.

Rotation with...	Observed type I error		
	Kish	Tam	Berger
SRSWOR	0.66	0.66	0.06
systematic sampling	0.57	0.57	0.03
rotation group sampling	0.70	0.70	0.06

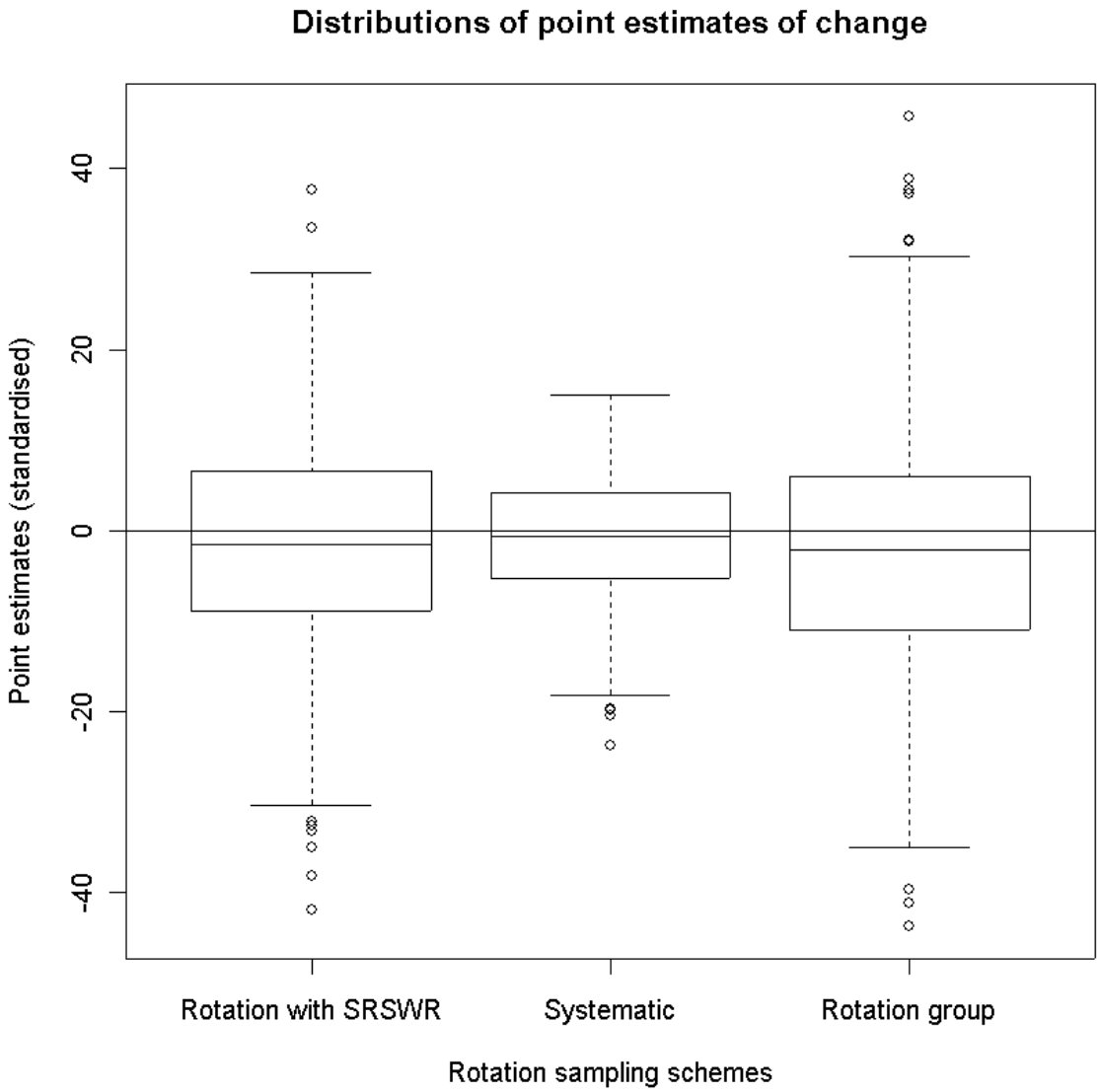

Figure 7.5: Sampling distribution of $\widehat{\Delta}_{st}$ for the three methods of sampling. τ= total number of pensioners.

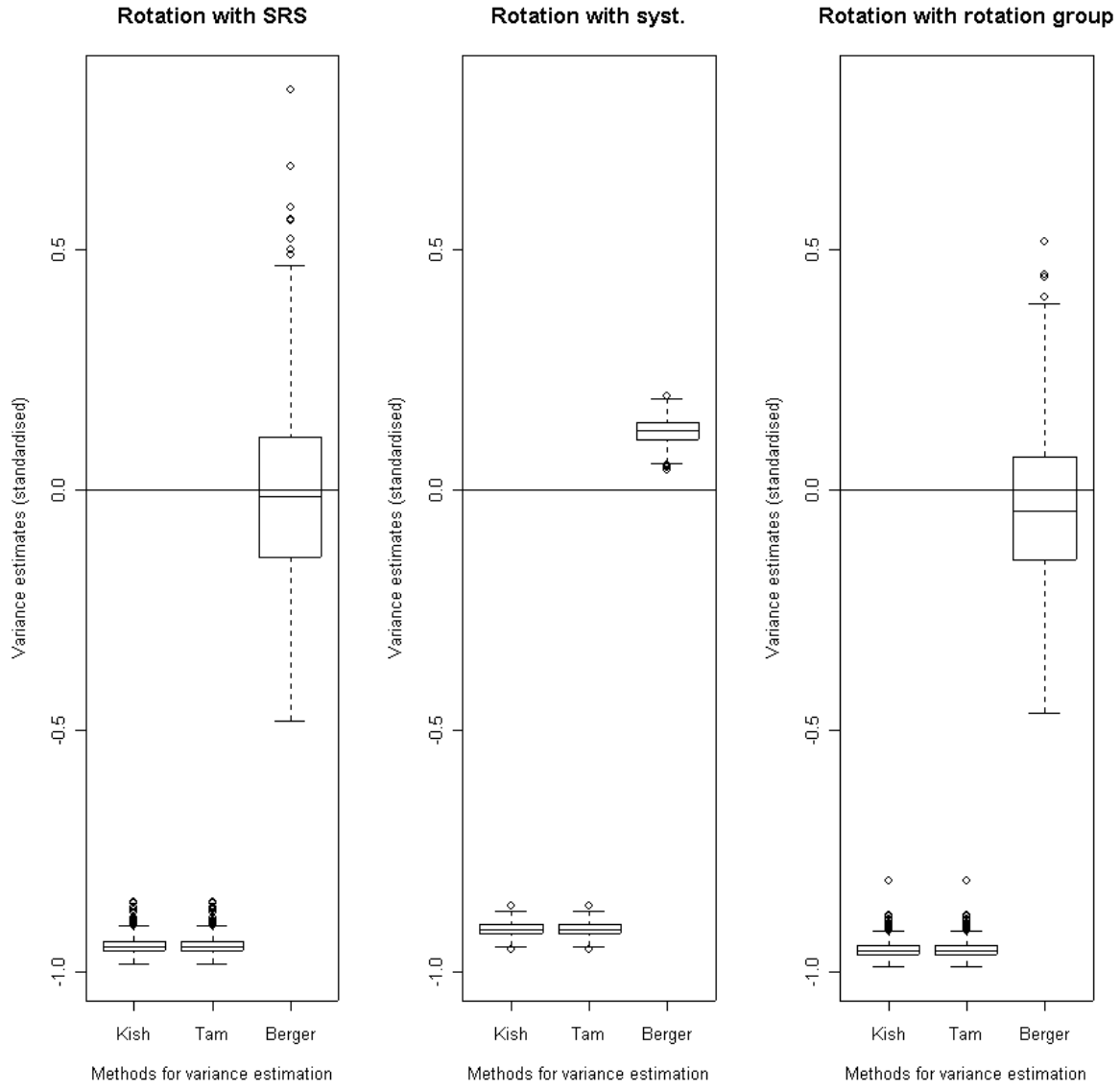

Figure 7.6: Empirical sampling distribution of $\widehat{\text{var}}(\widehat{\Delta})_{st}$ for the Kish, Tam and Berger variance estimators. τ= total number of pensioners.

Chapter 8

Generalisation to variance estimation for change in functions of totals

It is possible to extend the proposed estimator for variance estimation of change between estimators which are function of P totals, say $\widehat{\theta}_0 = f(\widehat{\tau}_{01}, \cdots, \widehat{\tau}_{0P})$ and $\widehat{\theta}_1 = f(\widehat{\tau}_{11}, \cdots, \widehat{\tau}_{1P})$. Linearization (e.g. Andersson and Nordberg, 1994) consists on approximating $\widehat{\theta}_1 - \widehat{\theta}_0$ by the first term of its Taylor series; that is,

$$\widehat{\theta}_1 - \widehat{\theta}_0 = \theta_1 - \theta_0 + \text{grad}(\tau)^T(\widehat{\tau} - \tau)$$

where $\widehat{\tau} = (\widehat{\tau}_{01}, \cdots, \widehat{\tau}_{0P}, \widehat{\tau}_{11}, \cdots, \widehat{\tau}_{0P})^T$, $\tau = (\tau_{01}, \cdots, \tau_{0P}, \tau_{11}, \cdots, \tau_{0P})^T$ and $\text{grad}(\tau)$ is the gradient of $\widehat{\theta}_1 - \widehat{\theta}_0$ at τ. An estimator for the variance of $\widehat{\theta}_1 - \widehat{\theta}_0$ is therefore

$$\widehat{\text{var}}(\widehat{\theta}_1 - \widehat{\theta}_0) = \text{grad}(\widehat{\tau})^T \widehat{\text{var}}(\widehat{\tau})\ \text{grad}(\widehat{\tau})$$

where $\widehat{\text{var}}(\widehat{\tau})$ is an estimator for the variance-covariance matrix of $\widehat{\tau}$. The estimators of Chapter 3 can be used to compute $\widehat{\text{var}}(\widehat{\tau})$. This variance-covariance matrix has to be non-negative definite if we want to guarantee positive estimates for the variance. Berger (2004) showed that the estimator of Section 4.3 always gives a non-negative definite matrix $\widehat{\text{var}}(\widehat{\tau})$.

Chapter 9

Conclusion

We have compared three estimators for the variance of change. We recommend the estimator of Section 4.3, as it gives accurate variance estimates. Furthermore, it still gives unbiased estimates when the change is negligible. The other estimator proposed have a large negative bias which can be explain by an overestimation of the correlation (see Section 4.1). As far as the method of rotation is concerned, the series of simulation suggests that the rotation with systematic sampling gives more precise point estimates and better variance estimates.

Acknowledgements

The author is grateful to Laila Habib Al Ajmi who did her Msc dissertation on variance for change.

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. (2004): *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/`.

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.

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.

Hansen, M. H. and Hurwitz, W. N. (1943): On the theory of sampling from a finite population. *Annals of Mathematical Statistics* **14**, 333–362.

Holmes, D. J. and Skinner, C. J. (2000): Variance estimation for labour force survey estimates of level and change. Technical report, Government Statistical Service Methodology Series. 21.

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.

Iachan, R. (1982): Systematic sampling: A critical review. *International Statistical Review* **50**, 293–303.

Kish, L. (1965): *Survey Sampling.* New York: John Wiley.

Madow, L. H. and Madow, W. G. (1944): On the theory of systematic sampling. *Annals of Mathematical Statistics* **15**, 1–24.

Narain, R. D. (1951): On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics* **3**, 169–174.

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.

Smith, P., Pont, M. and Jones, T. (2003): Developments in business survey methodology in the Office for National Statistics, 1994-2000. *Journal of the Royal Statistical Society, Series D* **52**, 1–30.

Tam, S. M. (1984): On covariances from overlapping samples. *The American Statistician* **38**, 288–289.

Wolter, K. M. (1985): *Introduction to Variance Estimation.* New York: Springer-Verlag.