Université de Neuchâtel
Institut de Statistique

# Thèse

présentée à la Faculté des Sciences Economiques
pour l'obtention du grade de Docteur en Statistique

par

Lionel Qualité

# Unequal probability sampling and repeated surveys

Acceptée sur proposition du jury composé de :

| | | | | |
|---|---|---|---|---|
| Pr. | Catalin | STARICA | Univ. Neuchâtel | Président du jury |
| Pr. | Yves | TILLÉ | Univ. Neuchâtel | Directeur de thèse |
| M. | Jean-Claude | DEVILLE | Crest-Ensai | Rapporteur |
| Pr. | Wayne A. | FULLER | Iowa State University | Rapporteur |
| Pr. | Chris J. | SKINNER | Univ. Southampton | Rapporteur |

Thèse soutenue le 23 juin 2009.

Université de Neuchâtel

**IMPRIMATUR POUR LA THESE**

Unequal probability sampling and repeated surveys

**Lionel QUALITÉ**

---

UNIVERSITE DE NEUCHATEL

FACULTE DES SCIENCES ECONOMIQUES

La Faculté des sciences économiques,
sur le rapport des membres du jury

Prof. Yves Tillé (directeur de thèse, Université de Neuchâtel)
M. Jean-Claude Deville (Ecole Nationale de la Statistique
et de l'Analyse de l'Information, Rennes, France),
Prof. Chris Skinner (Université of Southampton, Grande-Bretagne)
Prof. Wayne Fuller (Iowa State University, USA)
Prof. Catalin Starica (Université de Neuchâtel)

Autorise l'impression de la présente thèse.

Neuchâtel, 2 juillet 2009

Le doyen

Kilian Stoffel

■ Téléphone : +41 32 718 15 00 ■ Fax : +41 32 718 15 01 ■ E-mail : secretariat.seco@unine.ch ■ www.unine.ch

...

# Acknowledgements

Let Jean-Claude Deville and Yves Tillé be thanked here for the opportunity, and the privilege, they gave me by allowing me to work with them. They have taught me literally all I know on the subject of survey sampling, of research in general, and have always been of great support during the past seven years. They told me what to look for, where to look for it, and where to look better. I also want to thank all the team of the Section of Methods of the Swiss Federal Statistical Office, and particularly Philippe Eichenberger, Paul-andré Salamin and Jean-Pierre Renfer. It is partly thanks to them that I could spend some years in Neuchâtel and prepare this document. I sincerely thank Pr. Catalin Starica who accepted to be the president of this jury, Pr. Wayne Fuller and Pr. Chris Skinner who accepted to review this document. It is a great honor for me. Thanks also go to my colleagues and former colleagues, in the university and in the French national institute of statistics and economic studies. They made my days at work easier. Finally, I want to thank my whole family and my friends.

Neuchâtel, le 4 mai 2009.

# Contents

# List of Figures



# List of Tables

# List of Algorithms

# Résumés

Mots-clés : Entropie maximale, Coordination d'échantillons, Variance, Plan systématique.

Keywords: Maximal entropy, Sample coordination, Variance, Systematic sampling design.

Résumé : Sondage à probabilités inégales et enquêtes répétées
Ce document est constitué de deux parties. Dans la première partie, nous nous intéressons à certains plans de sondage à probabilités inégales, et dans la deuxième partie nous étudions le problème des enquêtes répétées. Bien que les sujets développés dans ces deux parties semblent entièrement différents, ils sont en fait reliés. La première partie est principalement consacrée à l'étude des propriétés de deux plans de sondage de taille fixe. Dans un premier chapitre, il est démontré que le plan de sondage à entropie maximale et de taille fixe est plus efficace que le sondage avec remise. Dans le second chapitre, nous montrons que le sondage systématique est un plan à support minimal. Nous donnons aussi quelques résultats sur la variance de l'estimateur de Horvitz-Thompson pour les plans à entropie maximale et pour les plans à support minimal. La deuxième partie débute par une étude de cas sur l'estimation de précision des évolutions dans le panel suisse sur la valeur ajoutée. Dans le chapitre suivant, nous proposons un estimateur de covariance pour les panels rotatifs à probabilités inégales. Enfin, nous présentons un système de coordination d'échantillons poissoniens développé pour l'Office Fédéral de la Statistique Suisse.

Abstract: Unequal probability sampling and repeated surveys
This document is divided into two parts. The first part revolves around the properties of some unequal probability survey sampling designs, and the second part deals with repeated surveys. While the topics developed in these two parts appear to be largely different, they are in fact related. The first part is devoted to the study of properties of two sampling designs with fixed size. In a first chapter we show that maximum entropy sampling with fixed size is more efficient than sampling with replacement. In a second chapter we prove that systematic sampling is a minimum support design. We also give some results on the variance of the Horvitz-Thompson estimator for maximum entropy and for minimum support designs. The second part begins with a case study of the estimation of variance of evolutions in the Swiss panel on value added. In a second chapter, we give covariance estimators for rotating panels with unequal inclusion probabilities. Finally, we describe a coordination method of maximum entropy samples that was developed for the Swiss Federal Statistical Office.

# Introduction

This document is divided into two parts. In the first part we explore properties of unequal probability survey sampling designs that have a maximal or minimal entropy. The second part deals with the estimation of precision in repeated surveys and with coordination methods. While the topics developed in these two parts appear to be largely different, they are in fact related.

The first part is devoted to the study of properties of two sampling designs with fixed size. In Chapter 1, which is the reprint of Qualité (2008), we prove that maximum entropy sampling with fixed size, first introduced as rejective sampling by Hájek (1964) and made practical by Chen et al. (1994), is uniformly more efficient than sampling with replacement. For a given set of inclusion probabilities there is no sampling design without replacement that, associated with the Horvitz-Thompson estimator, gives a uniformly lower variance than every other sampling designs without replacement. Sampling with replacement associated with the Hansen-Hurwitz estimator, on the other hand, can be a less efficient strategy than sampling without replacement associated with the Horvitz-Thompson estimator. It can naturally be used as an indicator of performance and a lower bound of acceptability for other sampling designs. Gabler (1984) gave sufficient conditions on the first and second order inclusion probabilities under which a sampling design can be proved to be more efficient than sampling with replacement. We show that maximum entropy sampling with fixed size satisfies these conditions. This property has not been proven to hold for many sampling designs since their second order inclusion probabilities are most of the time too complex to obtain this kind of non asymptotical result.

The motivation for this paper came from an entirely different problem, that was presented to me by Jean-Claude Deville. He had noticed that, if the total of a variable was to be estimated on two non overlapping samples coming from a simple random sampling design, the covariance of these estimators was non-positive. He wondered whether that was still the case for maximum entropy sampling with fixed size. It appears that, in a sense that will be explained in Chapter 5, this is the case exactly for sampling designs that are more efficient than sampling with replacement.

Chapter 2 is a reprint of Pea et al. (2007). It gives some new results on systematic sampling and on sampling designs that give a positive probability of selection to a small number of samples. Necessary and sufficient conditions for a sampling design to have a minimum support, in a sense explained in this paper, are given, and it is proven that systematic sampling satisfies these conditions. The paper also contains a sampling algo-

rithm that shares some properties with systematic sampling while avoiding its drawbacks. Our present interest in systematic sampling comes from the fact that this design has desirable properties for longitudinal surveys.

Chapter 3 is a collection of results and reflexions motivated by these two papers and that were not included in them by lack of space, consistency, or time to polish them. These papers dealt with sampling designs that are almost opposites on the scale of entropy, and it was interesting to compare their properties. After a short comment on the efficiency of systematic sampling in the simplest possible case, there is a rapid exploration of the entropy of minimum sampling designs and answers to some of the questions that could be raised by the second paper. This chapter is completed by some results on the dispersion of eigenvalues of the variance matrix of sampling designs. It starts with the observation that, in some cases, maximum entropy sampling minimizes this dispersion for a given set of inclusion probabilities. We show that, for unequal probability sampling designs, this dispersion cannot be null. In the case of sampling with fixed size, one eigenvalue is always null, but the dispersion of the other eigenvalues is also positive except when the inclusion probabilities are all equal. After we gave necessary conditions for a sampling design to be a minimum, or a maximum of this dispersion, we ascertain that maximum entropy sampling with fixed size does not result in a minimal dispersion of the eigenvalues. Minimum support designs on the other hand are good candidates for being sampling designs with a maximum dispersion.

The second part is devoted to repeated survey sampling and estimation of the variance of evolutions in simple cases. Chapter 4 is a reprint of Qualité & Tillé (2008). It consists in a study of the precision of longitudinal estimators of the Swiss survey on value added. The sampling design is very simple as it is a stratified panel, with non-response that we consider uniform within strata for each sampling occasion. The estimators of totals on the other hand make use of several techniques: robustification, calibration, and the 'surprise poststratum' method of Hidiroglou & Srinath (1981). This paper is a sort of case study where we take into account every aspect of variance estimation with such a complex estimator. We also compare the performance of the naive estimator of evolution with the estimator on the matched part of the samples. The later is, of course, more precise, but still not practical for a statistical institute that needs estimators of evolutions consistent with the estimators of levels.

Chapter 5 is a short exploration of covariance estimation for repeated surveys with unequal probabilities in a simple case: a first sampling phase with unequal probabilities gives a sample that is split into two overlapping samples using simple random sampling. This is an adequate description of the sampling design resulting from uniform non-response in a panel, or of a rotating panel with unequal probabilities. Berger (2004b) gave estimators for repeated sampling designs, that were based on asymptotic normality and a high entropy assumption, using the same approach as found in Hájek (1964). In this paper we give an exact expression that depends on the variance-covariance operator of the first phase sampling design, and give applications and estimators for some sampling designs.

We complete this chapter by a short discussion on a topic that was raised by the work in Chapters 4 and 5. The estimation of covariance in overlapping samples is discussed, under the light of an argument given in Berger (2004b) that Kish (1965)'s ratio estimator may lead to dramatic bias.

In Chapter 6, we present a simple extension of the coordination method of Brewer et al. (1972) for Poisson samples. This work should allow the Swiss Federal Statistical Office to organize and coordinate all its business surveys. It gives optimal coordination in some sense, and most aspects of business sample coordination in a dynamic population are dealt with. Other existing coordination methods do not have such flexibility and are not able to manage both positive and negative coordinations that are required for rotating panels.

# Part I

# Unequal probability sampling

# 1 A comparison of conditional Poisson sampling versus unequal probability sampling with replacement

**Abstract**

The variance of the Horvitz-Thompson estimator for a fixed size Conditional Poisson sampling scheme without replacement and with unequal inclusion probabilities is compared to the variance of the Hansen-Hurwitz estimator for a sampling scheme with replacement. We show, using a theorem by S. Gabler, that the sampling design without replacement is more efficient than the sampling design with replacement.[1]
MSC: 62D05

**Keywords:** Conditional Poisson sampling, Efficiency, Gabler's conditions, Survey sampling

## 1.1 Introduction

Conditional Poisson sampling, also called rejective sampling or maximum entropy sampling, was first introduced by Hájek (1964). It is a fixed size sampling design, without replacement, on a finite population, with unequal inclusion probabilities among the units of the population. It was called rejective sampling because Hájek's implementation amounts to drawing samples with a Poisson sampling design, which has random size, until one draws a sample that has the desired size. One can also obtain the Conditional Poisson design by drawing samples, with replacement, using a multinomial sampling design and rejecting the samples which hold some unit of the population more than once.

[1] This chapter is a reprint of: Qualité, L. (2008). A comparison of conditional Poisson sampling versus unequal probability sampling with replacement. *Journal of Statistical Planning and Inference* **138**, 1428–1432.

Chen et al. (1994) proposed a draw-by-draw algorithm which made this sampling design practical for medium sized populations. They gave some properties of this design, including the fact that it respects the Sen (1953) and Yates & Grundy (1953) conditions, providing a non negative variance estimator for the Horvitz-Thompson estimator. Other non-rejective implementations of Conditional Poisson sampling can be found in Traat et al. (2004) and Tillé (2006).

Chen et al. (1994) also asserted that Conditional Poisson sampling has a smaller variance than sampling with replacement with unequal inclusion probabilities. Their justification at that time was not entirely satisfactory as they argued it was implied by the Sen-Yates-Grundy conditions, but Gabler (1984) had already given a counter-example that shows that these conditions are not sufficient. At the same time, Gabler gave sufficient conditions for a fixed size sampling design without replacement to be more efficient than with replacement. He had already pointed out that Sampford's design enjoys this property (Gabler, 1981), and since then it has been proven for other sampling designs such as Chao's design (see Sengupta, 1989; Chao, 1982) and the elimination method (see Tillé, 1996; Deville & Tillé, 1998). We prove here that Gabler's conditions are satisfied under the Conditional Poisson sampling design, and thus that this design is more efficient than sampling with replacement, as could be expected.

## 1.2 Definitions, Conditional Poisson sampling

**Definition 1.1** *A sampling design without replacement is a probability law $P(s)$ on the subsets or samples $s$ of a finite population $U$. It is said to have a fixed size $n$ when all the samples which have a strictly positive probability contain $n$ units of the population.*

We can define the first order inclusion probabilities $\pi_k = P(k \in s)$, and the second order inclusion probabilities $\pi_{k\ell} = P(k, \ell \in s)$. If all the $\pi_k$ are positive, a natural and unbiased estimator of the total $Y = y_1 + \ldots + y_N$ was proposed by Horvitz & Thompson (1952)

$$\widehat{Y}_{HT} = \sum_{k \in s} \frac{y_k}{\pi_k}.$$

Its variance is

$$\operatorname{var}(\widehat{Y}_{HT}) = \sum_{k \in U} \frac{1 - \pi_k}{\pi_k} y_k^2 + \sum_{k \in U} \sum_{\substack{\ell \in U \\ \ell \neq k}} \frac{\pi_{k\ell} - \pi_k \pi_\ell}{\pi_k \pi_\ell} y_k y_\ell.$$

**Definition 1.2** *Poisson sampling is the sampling scheme without replacement and with unequal inclusion probabilities where each unit $k$ is selected independently from the others with probability $\tilde{\pi}_k$.*

Its law is given by

$$P_{\mathcal{P}}(s) = \prod_{k \in s} \tilde{\pi}_k \prod_{k \notin s} (1 - \tilde{\pi}_k) = \frac{\prod_{k \in s} w_k}{\sum_{s \subset U} \prod_{k \in s} w_k},$$

where $w_k = \tilde{\pi}_k/(1-\tilde{\pi}_k)$ if all the inclusion probabilities are smaller than one. Unfortunately it is not a fixed size sampling design.

**Definition 1.3** *Conditional Poisson sampling or rejective sampling is obtained by conditioning a Poisson sampling design on the size $n$ of the desired samples.*

Its law is given by

$$P_{\mathcal{CP}}(s) = \frac{\prod_{k\in s} w_k}{\sum_{\substack{s\subset U \\ |s|=n}} \prod_{k\in s} w_k}, \text{ if } |s| = n,$$

where $(w_k)_{k\in U}$ is a set of non negative real numbers uniquely determined by the inclusion probabilities $(\pi_k)_{k\in U}$ up to a positive factor (see Chen et al., 1994). It is the fixed size sampling design without replacement that has maximum entropy for a given set of inclusion probabilities $(\pi_k)_{k\in U}$.

**Definition 1.4** *The multinomial sampling scheme (see Hansen & Hurwitz, 1943) is obtained by $n$ independent draws from the population $U$. At each draw a unit $k$ is selected with probability $p_k$, and the number $n_k$ of times it is selected in the sample $s$ is incremented by one.*

The sampling law can be written

$$P_{\mathcal{MULT}}(s) = \frac{n!}{n_1!\dots n_N!}\prod_{k=1}^{N} p_k^{n_k}.$$

It is a fixed size sampling design with replacement as a unit may be selected more than once. In this setting the expected number of times a unit $k$ is selected is given by $\pi_k^* = np_k$. It is different from its first order inclusion probability $1-(1-p_k)^n$, $k \in U$.

An unbiased estimator was proposed by Hansen & Hurwitz (1943), if the $(p_k)_{k\in U}$ are positive,

$$\widehat{Y}_{HH} = \frac{1}{n}\sum_{k\in U} n_k \frac{y_k}{p_k} = \sum_{k\in U} n_k \frac{y_k}{\pi_k^*}.$$

Its variance is

$$\operatorname{var}(\widehat{Y}_{HH}) = \sum_{k\in U} \pi_k^* \left(\frac{y_k}{\pi_k^*} - \frac{Y}{n}\right)^2.$$

## 1.3 Superiority of sampling without replacement

For some sampling schemes without replacement and with unequal inclusion probabilities $(\pi_k)_{k\in U}$, it has been proven that the Horvitz-Thompson estimator always has a smaller variance than the Hansen-Hurwitz estimator for the multinomial design with parameter $(p_k)_{k\in U}$ such that $\pi_k^* = np_k = \pi_k$ for all $k \in U$. For these schemes,

$$\operatorname{var}(\widehat{Y}_{HT}) \leq \operatorname{var}(\widehat{Y}_{HH}),\ \forall \mathbf{Y} \in \mathbb{R}^N. \tag{1.1}$$

This is also the case for several sampling designs when $n = 2$ (see Brewer & Hanif, 1983). For the special case of Conditional Poisson sampling

where the inclusion probabilities are equal, and we are looking at simple random sampling schemes with and without replacement, the inequality (1.1) is easy to prove (see Cassel et al., 1977). In the general case, Gabler gave sufficient conditions for (1.1) to hold. These conditions are given in Theorem 1.1.

**Definition 1.5** *A sampling design is said to be connected if for any two units $i$ and $j$ there exists a sequence of units $i_1, \ldots, i_m$ such that all the probabilities $\pi_{ii_1}, \pi_{i_1i_2}, \ldots, \pi_{i_mj}$ are positive.*

**Theorem 1.1** (Gabler, 1984) *For a connected fixed size sampling design with positive inclusion probabilities $\pi_k$, $k \in U$, if*

$$\sum_{i \in U} \min_j \frac{\pi_{ij}}{\pi_j} \geq n - 1, \tag{1.2}$$

*we have* $\mathrm{var}(\widehat{Y}_{HT}) \leq \mathrm{var}(\widehat{Y}_{HH})$, *for all* $\mathbf{Y} \in \mathbb{R}^N$.

We will also need the following lemma.

**Lemma 1.1** *If $N \geq 3$ and we have a Conditional Poisson sampling design with size $2 \leq n \leq N$, if $i, j, k$ are different units of $U$ with $\pi_j \leq \pi_k$ then $\pi_{ij}/\pi_j \leq \pi_{ik}/\pi_k$.*

*Proof.* If $R_p^A = \sum_{\substack{s \subset U \setminus A \\ |s| = p}} \prod_{k \in s} w_k$, with $R_n = R_n^{\emptyset}$, $R_0^A = 1$ and $R_p^A = 0$ if $|U \setminus A| < p$, then we have (Chen et al., 1994)

$$\pi_j = \frac{w_j \cdot R_{n-1}^j}{R_n};\ \pi_{ij} = \frac{w_i w_j \cdot R_{n-2}^{ij}}{R_n},$$

and

$$\frac{\pi_{ik}}{\pi_k} - \frac{\pi_{ij}}{\pi_j} = w_i \left( \frac{R_{n-2}^{ik}}{R_{n-1}^k} - \frac{R_{n-2}^{ij}}{R_{n-1}^j} \right).$$

Using the property that $R_n^A = w_k R_{n-1}^{A \cup \{k\}} + R_n^{A \cup \{k\}}$ if $k \notin A$, we obtain the following equations depending on $n$: If $n = 2$,

$$\frac{\pi_{ik}}{\pi_k} - \frac{\pi_{ij}}{\pi_j} = w_i \left( \frac{1}{R_{n-1}^k} - \frac{1}{R_{n-1}^j} \right) = \frac{w_i (w_k - w_j)}{R_{n-1}^k R_{n-1}^j}.$$

If $n \geq 3$,

$$\frac{\pi_{ik}}{\pi_k} - \frac{\pi_{ij}}{\pi_j} = \frac{w_i (w_k - w_j)}{R_{n-1}^k R_{n-1}^j} \left[ \left( R_{n-2}^{ijk} \right)^2 - R_{n-1}^{ijk} R_{n-3}^{ijk} \right].$$

We only have to remark that $\pi_k \geq \pi_j \Leftrightarrow w_k \geq w_j$ (see Chen et al., 1994) and to prove that $\left( R_{n-2}^{ijk} \right)^2 \geq R_{n-1}^{ijk} R_{n-3}^{ijk}$. Actually, in this case all the weights $w_\ell$ are positive and by Newton's inequalitity (see for example Hardy et al., 1956),

$$\left( R_{n-2}^{ijk} \right)^2 \geq \frac{N - n + 3}{N - n + 2} \cdot \frac{n - 1}{n - 2} R_{n-1}^{ijk} R_{n-3}^{ijk}.$$

□

**Proposition 1.1** *Conditional Poisson sampling satisfies the conditions of Gabler's theorem in 1.1 and thus is more efficient than sampling with replacement.*

*Proof.* The cases $N \leq 2$ or $n < 2$ are easily proven to satisfy inequality 1.1. In the case of Conditional Poisson sampling, if $n > 1$ and every $\pi_k$ is positive, then every second order inclusion probability $\pi_{k\ell}$ is also positive. Henceforth, Conditional Poisson designs are connected. We only have to prove condition 1.2 to have the conclusion. If $2 \leq n \leq N$, we can without loss of generality suppose that $0 < \pi_1 \leq \ldots \leq \pi_N$. Then, using lemma 1.1, if $i \neq 1$, $\min_j \pi_{ij}/\pi_j = \pi_{i1}/\pi_1$, and $\min_j \pi_{1j}/\pi_j = \pi_{12}/\pi_2$. Hence

$$\sum_{i \in U} \min_j \frac{\pi_{ij}}{\pi_j} = \frac{\pi_{12}}{\pi_2} + \sum_{i=2}^{N} \frac{\pi_{i1}}{\pi_1} = n - 1 + \frac{\pi_{12}}{\pi_2} > n - 1. \tag{1.3}$$

□

## 1.4 Discussion

Conditional Poisson sampling combined with the Horvitz-Thompson estimator is a more efficient strategy than multinomial sampling with the Hansen-Hurwitz estimator. This result gives a convenient upper bound, that does not depend on the second order inclusion probabilities, for the variance of Conditional Poisson sampling. However, it still requires to compute the exact first order inclusion probabilities. Since those were difficult to calculate until recently, in many cases the target inclusion probabilities $\pi_k^* = np_k^*$ of the rejective procedure with multinomial design (see Brewer & Hanif, 1983, procedure 14) were used instead. The estimator $\widehat{Y}^* = \sum_{k \in s} \frac{y_k}{\pi_k^*}$ does not verify the same inequality as can be seen in the following example:

Let $N = 4$, $n = 2$, $\pi_1^* = 0.2$, $\pi_2^* = 0.4$, $\pi_3^* = 0.6$, $\pi_4^* = 0.8$ and $y_1 = 1$, $y_2 = y_3 = y_4 = -1$ then $\operatorname{var}(\widehat{Y}^*) \approx 9.085$ and $\sum_{k=1}^{4} \pi_k^* \left( \frac{y_k}{\pi_k^*} - \frac{Y}{n} \right)^2 \approx 8.417$.

## Acknowledgment

The author is grateful to Jean-Claude Deville and Yves Tillé for their guidance, and to the referees and an associate editor for their constructive advices.

# 2 Systematic sampling is a minimum support design

**Abstract**

In order to select a sample in a finite population of $N$ units with given inclusion probabilities, it is possible to define a sampling design on at most $N$ samples that have a positive probability of being selected. Designs defined on minimal sets of samples are called minimum support designs. It is shown that, for any vector of inclusion probabilities, systematic sampling always provides a minimum support design. This property makes it possible to extensively compute the sampling design and the joint inclusion probabilities. Random systematic sampling can be viewed as the random choice of a minimum support design. However, even if the population is randomly sorted, a simple example shows that some joint inclusion probabilities can be equal to zero. Another way of randomly selecting a minimum support design is proposed, in such a way that all the samples have a positive probability of being selected, and all the joint inclusion probabilities are positive.[1]

**Keywords:** Minimum support design, systematic sampling, unequal probability sampling, survey sampling

## 2.1 Introduction

The support of a sampling design is the set of samples that have a positive probability of being selected. For any vector of inclusion probabilities, Wynn (1977) has proved the existence of at least one fixed size sampling design with support not larger than the population size. However, Wynn's result is not constructive. Deville & Tillé (1998) have proposed a general method for constructing minimum support designs. This minimum support design procedure is a particular case of the family of splitting procedures. Several procedures that provide minimum support designs had already been proposed by Jessen (1969) (see also Brewer & Hanif, 1983,

[1] This chapter is a reprint of: Pea, J., Qualité, L. & Tillé, Y. (2007). Systematic sampling is a minimum support design. *Computational Statistics and Data Analysis* **51**, 5591–5602.

procedure 35 and 36, pages 42-43) in a more restrictive context. Systematic sampling was first proposed by Madow (1949) and was developed by Connor (1966); Gray (1971); Brewer (1963); Pinciaro (1978); Hidiroglou & Gray (1980). It is an unequal probability sampling design with fixed sample size. Its main drawback is that it may result in null joint inclusion probabilities for some couples of units in the population. A random sorting of the population is usually recommended before applying systematic sampling. Unfortunately, this is not sufficient. A simple example given in Brewer & Hanif (1983) shows that some joint inclusion probabilities may still be equal to zero. This problem is an important drawback of random systematic sampling and is the consequence of some samples having a null probability of being selected.

We show that the systematic sampling design is a minimum support design. Hence, random systematic sampling can be viewed as a method where a minimum support design is chosen at random. We introduce a fast algorithm that computes this design and allows deriving the joint inclusion probabilities. We also show that any sample can be included in a minimum support design with given first-order inclusion probabilities. We propose a procedure whereby a minimum support design is randomly chosen and whereby all the possible samples have a positive probability of being selected. This method, called 'random minimum support design', is a good alternative to random systematic sampling.

The paper is organized as follows. After a definition of the basic concepts and of the notations in Section 2.2, minimum support designs are presented. In Section 2.3, we demonstrate that the size of the support of a systematic design is at most equal to the population size, and give some examples. In Section 2.4, we show that systematic sampling provides a minimum support design. This result leads us to propose a simple method for the computation of the joint inclusion probabilities in Section 2.5. Finally, in Section 2.6, we discuss a method to randomly select a minimum support design in such a way that all the samples have a non-null probability of being drawn. A short discussion of these methods is given in Section 2.7.

## 2.2 Basic concepts and notations

Consider a finite population of size $N$. Let $U = \{1, \ldots, k, \ldots, N\}$ be the set of labels of the units in the population. A sample (without replacement) is a subset of the population and will be represented by a vector $\mathbf{s} = (s_k)_{1 \leq k \leq N}$ defined by

$$s_k = \begin{cases} 1 & \text{if } k \text{ is in the sample,} \\ 0 & \text{otherwise.} \end{cases}$$

A support $\mathcal{Q}$ is a set of samples. The full support is the set of all the possible samples under some constraints, e.g. $\mathcal{S} = \{0,1\}^N$ when there are no constraints. The full support corresponding to the samples of fixed size $n$ is defined by $\mathcal{S}_n = \{\mathbf{s} \in \mathcal{S} : \sum_{k \in U} s_k = n\}$. The size of the support $\mathcal{Q}$ is the number of samples it contains, i.e. $\operatorname{card} \mathcal{Q}$.

A sampling design $p(\cdot)$ is a probability distribution on a support $\mathcal{Q} \subset \mathcal{S}$, so that for all $s \in \mathcal{Q}$, $p(\mathbf{s}) > 0$ and

$$\sum_{\mathbf{s}\in\mathcal{Q}} p(\mathbf{s}) = 1.$$

The first-order inclusion probability $\pi_k$ is the probability of selecting the unit $k$ in the sample, and $\boldsymbol{\pi} = (\pi_k)_{1\leq k\leq N}$ is the inclusion probability vector. It can be derived from the sampling design as follows:

$$\boldsymbol{\pi} = \sum_{\mathbf{s}\in\mathcal{Q}} \mathbf{s}p(\mathbf{s}).$$

When the design has a fixed sample size $n$, then

$$\sum_{k\in U} \pi_k = n.$$

The joint inclusion probability $\pi_{k\ell}$ is the probability of selecting the units $k$ and $\ell$ together in the sample, and $\pi_{kk} = \pi_k$. The matrix of joint inclusion probabilities is given by

$$\boldsymbol{\Pi} = \sum_{\mathbf{s}\in\mathcal{Q}} \mathbf{s}\mathbf{s}'p(\mathbf{s}).$$

The joint inclusion probabilities are usually needed to compute the variance of the estimators under the sampling design.

**Definition 2.1** *A sampling design $p_0(\cdot)$ with inclusion probabilities $(\pi_k)_{1\leq k\leq N}$ is said to be defined on a minimum support $\mathcal{Q}_0$ if, for every $\mathcal{Q} \subset \mathcal{Q}_0$ with $\mathcal{Q} \neq \mathcal{Q}_0$, there is no design $p(\cdot)$ with support $\mathcal{Q}$ and with $\sum_{\mathbf{s}\in\mathcal{Q}} s_k p(\mathbf{s}) = \pi_k$, $k = 1, \ldots, N$.*

Wynn (1977) has shown that, for any given first-order inclusion probabilities, it is always possible to define a fixed size design on at most $N$ samples $\mathbf{s}_i$ such that $p(\mathbf{s}_i) > 0$. Using Carathéodory's theorem as in Wynn (1977), one can prove that, if $\mathcal{Q}_0$ is the support of a minimum support design, then $\operatorname{card}\mathcal{Q}_0 \leq N$. Moreover, when the design does not have a fixed size, at most $N + 1$ samples are needed. Deville & Tillé (1998) have given a way to implement fixed size minimum support designs.

## 2.3 Systematic sampling

Let $\boldsymbol{\pi}$ be a vector of inclusion probabilities and $V_k$ be the cumulated inclusion probabilities, given by

$$V_k = \begin{cases} \sum_{\ell=1}^{k} \pi_\ell, & \text{if } k = 1, \ldots, N, \\ 0 & \text{if } k = 0. \end{cases}$$

In a fixed size design, $V_N = n$. The selection procedure for systematic sampling is given in Algorithm 1.

**Example 2.1** Suppose that $N = 6$, $n = 3$, $\boldsymbol{\pi} = (0.2, 0.7, 0.8, 0.5, 0.4, 0.4)$, and $u = 0.47$. The development given in Table 2.1 shows that $\mathbf{s} = \{0, 1, 1, 0, 1, 0\}$.

**Algorithm 1**: Systematic sampling

1. Generate $u$, a uniformly distributed random number in $[0,1)$.
2. For $k = 1, \ldots, N$,

$$s_k = \begin{cases} 1 & \text{if there exists an integer } j > 0 \\ & \text{such that } V_{k-1} \leq u + j - 1 < V_k, \\ 0 & \text{otherwise.} \end{cases}$$

Table 2.1 – *Example of Systematic Sampling*

| $k$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $\pi_k$ | 0.2 | 0.7 | 0.8 | 0.5 | 0.4 | 0.4 |
| $V_k$ | 0.2 | 0.9 | 1.7 | 2.2 | 2.6 | 3 |
| $j$ | - | 1 | 2 | - | 3 | - |
| $u+j-1$ | - | 0.47 | 1.47 | - | 2.47 | - |
| $s_k$ | 0 | 1 | 1 | 0 | 1 | 0 |

The following result provides a way to compute the size of the support in systematic sampling.

**Result 2.1** *For the systematic sampling, the size of the support is equal to the number of distinct $r_k$, where $r_k = V_k(mod 1)$ and $V_k = \sum_{\ell=1}^{k} \pi_\ell, k = 1, \ldots, N$.*

*Proof.* Let $D$ be the number of distinct $r_k$'s. Let $r_{(k)}$ be the sequence of the ordered $r_k$ without repetition, so that $r_{(1)} < r_{(2)} < \ldots < r_{(D-1)} < r_{(D)} < 1$. Since $r_N = 0$, it follows that $r_{(1)} = 0$. Let $u$ be a uniformly distributed random variable in $[0,1)$. The $r_{(k)}$'s divide $[0,1)$ in $D$ non-overlapping intervals

$$\left[0, r_{(2)}\right), \left[r_{(2)}, r_{(3)}\right), \ldots, \left[r_{(D-1)}, r_{(D)}\right), \left[r_{(D)}, 1\right).$$

Each interval $\left[r_{(k)}, r_{(k+1)}\right)$, in which $u$ can fall, corresponds to one and only one sample. Hence, the size of the support is equal to the number of distinct values of $r_k$. □

**Corollary 2.1** *The size of the support of the systematic sampling design is not larger than $N$.*

**Example 2.2** Suppose that $N = 6, n = 3$ and that the $\pi_k$'s, $V_k$'s and $r_k$'s are as in Table 2.2.

Table 2.2 – *Support of Example 2.2*

| $k$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $\pi_k$ | 0.3 | 0.8 | 0.4 | 0.7 | 0.6 | 0.2 |
| $V_k$ | 0.3 | 1.1 | 1.5 | 2.2 | 2.8 | 3 |
| $r_k$ | 0.3 | 0.1 | 0.5 | 0.2 | 0.8 | 0 |

Since all the $r_k$'s are distinct, the size of the support is equal to $N = 6$. The sampling design is given in Table 2.3.

Table 2.3 – *Sampling design of Example 2.2*

| $k$ | $\mathbf{s}_1$ | $\mathbf{s}_2$ | $\mathbf{s}_3$ | $\mathbf{s}_4$ | $\mathbf{s}_5$ | $\mathbf{s}_6$ | $\pi_k$ |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0.3 |
| 2 | 1 | 0 | 0 | 1 | 1 | 1 | 0.8 |
| 3 | 0 | 1 | 1 | 1 | 0 | 0 | 0.4 |
| 4 | 1 | 1 | 0 | 0 | 1 | 1 | 0.7 |
| 5 | 0 | 0 | 1 | 1 | 1 | 0 | 0.6 |
| 6 | 0 | 0 | 0 | 0 | 0 | 1 | 0.2 |
| $p(\mathbf{s}_i)$ | 0.1 | 0.1 | 0.1 | 0.2 | 0.3 | 0.2 | $\sum_U \pi_k = 3$ |

The sample $\mathbf{s}_i$, for $i = 1, \ldots, N$, is obtained when the uniform random number $u$ takes a value in $\left[r_{(i)}, r_{(i+1)}\right)$. Thus, the probability of selecting the sample $\mathbf{s}_i$ is given by $p(\mathbf{s}_i) = r_{(i+1)} - r_{(i)}$, with $r_{(7)} = 1$.

**Example 2.3** Suppose that $N = 6, n = 3$ and that the $\pi_k$'s, $V_k$'s and $r_k$'s are as in Table 2.4.

Table 2.4 – *Support of Example 2.3*

| $k$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $\pi_k$ | 0.2 | 0.7 | 0.3 | 0.6 | 0.4 | 0.8 |
| $V_k$ | 0.2 | 0.9 | 1.2 | 1.8 | 2.2 | 3 |
| $r_k$ | 0.2 | 0.9 | 0.2 | 0.8 | 0.2 | 0 |

There are only 4 distinct values for the $r_k$'s: 0, 0.2, 0.8, 0.9, therefore, the size of the support is equal to 4. The sampling design is given in Table 2.5.

Table 2.5 – *Sampling design of Example 2.3*

| $k$ | $\mathbf{s}_1$ | $\mathbf{s}_2$ | $\mathbf{s}_3$ | $\mathbf{s}_4$ | $\pi_k$ |
|---|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 0 | 0.2 |
| 2 | 0 | 1 | 1 | 0 | 0.7 |
| 3 | 1 | 0 | 0 | 1 | 0.3 |
| 4 | 0 | 1 | 0 | 0 | 0.6 |
| 5 | 1 | 0 | 1 | 1 | 0.4 |
| 6 | 0 | 1 | 1 | 1 | 0.8 |
| $p(\mathbf{s}_i)$ | 0.2 | 0.6 | 0.1 | 0.1 | $\sum_U \pi_k = 3$ |

## 2.4 Systematic sampling is a minimum support design

In order to show that systematic sampling is always defined on a minimum support, we first prove Lemma 2.1.

**Lemma 2.1** *Let $p(\cdot)$ be a sampling design defined on a support $\mathcal{Q} = \{\mathbf{s}_1, \ldots, \mathbf{s}_q\}$, and $\boldsymbol{\pi}$ be its vector of inclusion probabilities. Let $\mathbf{S}$ be the matrix defined by $\mathbf{S} = (\mathbf{s}_1, \ldots, \mathbf{s}_q)$; $\mathrm{Ker}(\mathbf{S}) = \{\mathbf{u} \in \mathbb{R}^q : \mathbf{S}\mathbf{u} = \mathbf{0}\}$ and $\mathbf{1}_q^\perp = \{\mathbf{u} \in \mathbb{R}^q : \sum_{i=1}^q u_i = 0\}$. Then:*

1. *$p(\cdot)$ is a minimum support design if and only if $\mathrm{Ker}(\mathbf{S}) \cap \mathbf{1}_q^\perp = \{\mathbf{0}\}$,*
2. *if $p(\cdot)$ is a fixed size design then it is a minimum support design if and only if the samples in $\mathcal{Q}$ are linearly independent, i.e. $\mathrm{Ker}(\mathbf{S}) = \{\mathbf{0}\}$.*

*Proof.* The case $q = 1$ is trivial, we will now assume that $q \geq 2$.

1. - *If $\mathrm{Ker}(\mathbf{S}) \cap \mathbf{1}_q^\perp = \{\mathbf{0}\}$ then $p(\cdot)$ is a minimum support design.* We will prove the contrapositive version: "if $p(\cdot)$ is not a minimum support design then $\mathrm{Ker}(\mathbf{S}) \cap \mathbf{1}_q^\perp \neq \{\mathbf{0}\}$". Suppose that $p(\cdot)$ is not a minimum support design, then there exists a sampling design $p^\star(\cdot)$ defined on a strict subset of $\mathcal{Q}$ and that has the same inclusion probabilities. Let $\mathbf{p}$ and $\mathbf{p}^\star$ be the vectors of probabilities of the designs $p(\cdot)$ and $p^\star(\cdot)$ on the support $\mathcal{Q}$, i.e $\mathbf{p} = \big(p(\mathbf{s}_1), \ldots, p(\mathbf{s}_q)\big)'$ and $\mathbf{p}^\star = \big(p^\star(\mathbf{s}_1), \ldots, p^\star(\mathbf{s}_q)\big)'$. Note that $\mathbf{p} \neq \mathbf{p}^\star$ since $\mathbf{p}$ has no null coordinates whereas $\mathbf{p}^\star$ does. Moreover, we have

$$\mathbf{S}\mathbf{p} = \boldsymbol{\pi}, \mathbf{S}\mathbf{p}^\star = \boldsymbol{\pi}. \tag{2.1}$$

Hence $\mathbf{p} - \mathbf{p}^\star$ is a non null vector in $\mathrm{Ker}(\mathbf{S})$. As $\mathbf{p}$ and $\mathbf{p}^\star$ both sum to 1, $\mathbf{p} - \mathbf{p}^\star$ is also in $\mathbf{1}_q^\perp$, and thus $\mathrm{Ker}(\mathbf{S}) \cap \mathbf{1}_q^\perp \neq \{\mathbf{0}\}$. This completes the first part of the proof.

- *If $p(\cdot)$ is a minimum support design then* $\mathrm{Ker}(\mathbf{S}) \cap \mathbf{1}_q^\perp = \{\mathbf{0}\}$. By contrapositive: assume that $\mathrm{Ker}(\mathbf{S}) \cap \mathbf{1}_q^\perp \neq \{\mathbf{0}\}$, we need to prove that $p(\cdot)$ is not a minimum support design. Let $\boldsymbol{\lambda}$ be a non null vector in $\mathrm{Ker}(\mathbf{S}) \cap \mathbf{1}_q^\perp$ and $\mathbf{p}$ be the vector of probabilities of the sampling design $p(\cdot)$. Then, for any real number $\mu$, we have

$$\begin{aligned} \mathbf{S}(\mathbf{p} + \mu\boldsymbol{\lambda}) &= \boldsymbol{\pi}, \\ \sum_{i=1}^q p(\mathbf{s}_i) + \mu\lambda_i &= 1. \end{aligned}$$

Since, for each $i \in \{1, \ldots, q\}$, $0 < p(\mathbf{s}_i) < 1$, there exist $\mu_i^a$ and $\mu_i^b$ with $\mu_i^a < 0 < \mu_i^b$ such that $\mu \in [\mu_i^a, \mu_i^b] \Leftrightarrow p(\mathbf{s}_i) + \mu\lambda_i \in [0, 1]$ (if $\lambda_i = 0$, $\mu_i^a = -\infty$ and $\mu_i^b = +\infty$). As $\boldsymbol{\lambda} \neq \mathbf{0}$, there exists

$$\mu = \max_{i=1}^q \mu_i^a \in \mathbb{R}.$$

Then $\mathbf{p} + \mu\boldsymbol{\lambda}$ is a vector of probabilities with at least one null coordinate. Hence it defines a sampling design on a support strictly included in $\mathcal{Q}$, and with inclusion probabilities $\boldsymbol{\pi}$. It follows that $p(\cdot)$ is not a minimum support design and the proof of the first statement in Lemma 2.1 is complete.

2. If $p(\cdot)$ is a fixed size design and $\mathbf{1}_N$ is the vector $(1,\ldots,1)'$ in $\mathbb{R}^N$, then $\mathbf{1}_N'\mathbf{S} = n\mathbf{1}_q'$, hence $\mathrm{Ker}(\mathbf{S}) \subset \mathbf{1}_q^{\perp}$. In that case, $\mathrm{Ker}(\mathbf{S}) \cap \mathbf{1}_q^{\perp} = \mathrm{Ker}(\mathbf{S})$ and the second statement in the lemma is a direct consequence of the first statement.

□

**Result 2.2** *Systematic sampling is a minimum support design.*

*Proof.* In order to apply Lemma 2.1, we need to show that the samples in the support of a systematic sampling are linearly independent. Result 2.1 shows that the size of the support is equal to $N$ if all the $r_i$ are distinct and is smaller than $N$ if at least two units have the same $r_i$.

*Case 1* The size of the support is equal to $N$.

All the $r_{(i)}, i = 1,\ldots,N$, are distinct. Let $\mathbf{S}$ denote the matrix of the $N$ possible samples, i.e. $\mathbf{S} = (\mathbf{s}_1,\ldots,\mathbf{s}_N)$. The samples are ordered according to the $N$ intervals $\left[0, r_{(2)}\right), \left[r_{(2)}, r_{(3)}\right), \ldots, \left[r_{(N)}, 1\right)$. We want to show that $\mathbf{S}$ is a full rank matrix. Consider the matrix $\mathbf{A}$ defined by $\mathbf{A} = (\mathbf{a}_1,\ldots,\mathbf{a}_{N-1},\mathbf{s}_N)$, where $\mathbf{a}_i = \mathbf{s}_i - \mathbf{s}_{i+1}, i = 1,\ldots,N-1$,

$$\mathbf{A} = \mathbf{S}\begin{pmatrix} 1 & 0 & \ldots & \ldots & 0 \\ -1 & 1 & 0 & & \vdots \\ 0 & -1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 & 0 \\ 0 & \ldots & 0 & -1 & 1 \end{pmatrix}.$$

Since $\mathbf{A}$ is the product of $\mathbf{S}$ with an invertible matrix, $\mathbf{A}$ and $\mathbf{S}$ have the same rank.

The last column of $\mathbf{A}$ is equal to $\mathbf{s}_N$. If $1 \leq i \leq N-1$, let us consider the unit denoted $k$ that corresponds to $r_{(i+1)}$, that is to say $r_k = r_{(i+1)}$. Necessarily, $k \leq N-1$ since $r_N = r_{(1)}$. When $u$ goes from the interval $\left[r_{(i)}, r_{(i+1)}\right)$ to the interval $\left[r_{(i+1)}, r_{(i+2)}\right)$, with $r_{(N+1)} = 1$, the unit $k$ is removed from the sample $\mathbf{s}_i$ and is replaced by the unit $k+1$, to give the sample $\mathbf{s}_{i+1}$. Hence, the column $\mathbf{a}_i = (a_{\ell i})$ of $\mathbf{A}$ is given by

$$a_{\ell i} = \begin{cases} 1 & \text{if } \ell = k, \\ -1 & \text{if } \ell = k+1, \\ 0 & \text{otherwise.} \end{cases}$$

Now, consider the $N \times N$ full rank matrix $\mathbf{T} = \left(t_{ij}\right)$, where

$$t_{ij} = \begin{cases} 1 & \text{if } i \geq j, \\ 0 & \text{if } i < j. \end{cases}$$

We obtain that

$$\mathbf{TA} = \begin{pmatrix} \mathbf{B} & \mathbf{c} \\ \mathbf{0} & n \end{pmatrix},$$

where $n$ is the sample size and $\mathbf{c}$ is a column vector of $N-1$ positive integers. Indeed, the last column of $\mathbf{TA}$ is equal to $\mathbf{Ts}_N$. It follows that $\mathbf{c}_k$ is the number of units in $\{1,\dots,k\}$ that are selected in $\mathbf{s}_N$, and the last coefficient in this column is the size of the sample $\mathbf{s}_N$.

The matrix $\mathbf{B}$ is a $(N-1)\times(N-1)$ permutation matrix that gives the correspondence between $\left(r_{(2)},\dots,r_{(N)}\right)$ and $(r_1,\dots,r_{N-1})$, and finally $\mathbf{0}$ is a row vector of zeros in $\mathbb{R}^{N-1}$. Indeed, if $1\leq i\leq N-1$, the $i^{th}$ column of $\mathbf{TA}$ has only one non null coefficient, equal to 1. It is on the row $\ell$ such that $a_{\ell i}=1$, i.e. $r_\ell = r_{(i+1)}$. It follows that the last row of the $N-1$ first columns of $\mathbf{TA}$ is null, as $r_N = r_{(1)}$ and there is no $i\in\{1,\dots,N-1\}$ such that $r_N = r_{(i+1)}$. It also follows that the matrix $\mathbf{B}=\left(b_{ij}\right)$ has exactly one non null coefficient in each column and in each row. This coefficient is equal to 1 and $b_{ij}=1$ if and only if $r_i = r_{(j+1)}$, which completes the proof that $\mathbf{B}$ is a permutation matrix and that $\mathbf{B}\left(r_{(2)},\dots,r_{(N)}\right)' = (r_1,\dots,r_{N-1})'$.

Therefore, the matrix $\mathbf{TA}$ has full rank, and so do $\mathbf{A}$ and $\mathbf{S}$.

**Remark 2.1** *This part of the proof is easier to understand if we examine an example. If we take the data of Example 2, all the $r_k$ are distinct. Thus the size of the support is equal to $N$. We get*

$$\mathbf{S}=\begin{pmatrix}1&1&1&0&0&0\\1&0&0&1&1&1\\0&1&1&1&0&0\\1&1&0&0&1&1\\0&0&1&1&1&0\\0&0&0&0&0&1\end{pmatrix},\mathbf{A}=\begin{pmatrix}0&0&1&0&0&0\\1&0&-1&0&0&1\\-1&0&0&1&0&0\\0&1&0&-1&0&1\\0&-1&0&0&1&0\\0&0&0&0&-1&1\end{pmatrix},$$

$$\mathbf{T}=\begin{pmatrix}1&0&0&0&0&0\\1&1&0&0&0&0\\1&1&1&0&0&0\\1&1&1&1&0&0\\1&1&1&1&1&0\\1&1&1&1&1&1\end{pmatrix}\ \textit{and}\ \mathbf{TA}=\begin{pmatrix}0&0&1&0&0&0\\1&0&0&0&0&1\\0&0&0&1&0&1\\0&1&0&0&0&2\\0&0&0&0&1&2\\0&0&0&0&0&3\end{pmatrix}.$$

*In this example, one can observe that* $\mathbf{T}$ *and* $\mathbf{TA}$ *are full rank matrices. Thus,* $\mathbf{A}$ *is full rank and* $\mathbf{S}$ *has the same rank as* $\mathbf{A}$.

*Case 2* The size of the support is smaller than $N$

If all the $r_i$, $i=1,\dots,N-1$, are not distinct, then by passing from some sample $\mathbf{s}_i$ to the sample $\mathbf{s}_{i+1}$, more than one unit can be removed from $\mathbf{s}_i$ and thus more than one unit can enter $\mathbf{s}_{i+1}$. If, from $\mathbf{s}_i$ to $\mathbf{s}_{i+1}$, more than one unit is replaced, it is possible to construct one or several phantom samples $\mathbf{s}_i', \mathbf{s}_i'', \dots$, so that only one unit is replaced by passing from a sample to the next one. If the creation of phantom samples is repeated every time that two $r_i$ are equal, it is possible to complete the support in order to obtain a sequence of $N$ samples that have the same properties as in Case 1. These samples are thus linearly independent, and the subset of samples actually in the support is also linearly independent.

**Remark 2.2** *Again, the proof is easier to understand if we examine an example. If we take the data of Example 3, the $r_i$ are not distinct: 0, 0.2, 0.2, 0.2, 0.8, 0.9. The sampling design is given in Table 2.6.*

Table 2.6 – *Sampling design of Example 2.3*

| $k$ | $\mathbf{s}_1$ | $\mathbf{s}_2$ | $\mathbf{s}_3$ | $\mathbf{s}_4$ | $\pi_k$ |
|---|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 0 | 0.2 |
| 2 | 0 | 1 | 1 | 0 | 0.7 |
| 3 | 1 | 0 | 0 | 1 | 0.3 |
| 4 | 0 | 1 | 0 | 0 | 0.6 |
| 5 | 1 | 0 | 1 | 1 | 0.4 |
| 6 | 0 | 1 | 1 | 1 | 0.8 |
| $p(\mathbf{s}_i)$ | 0.2 | 0.6 | 0.1 | 0.1 | $\sum_U \pi_k = 3$ |

*Now it is possible to add two phantom samples $\mathbf{s}_1'$ and $\mathbf{s}_1''$ so that only one out of the 3 selected units is replaced by passing from $\mathbf{s}_1$ to $\mathbf{s}_1'$, from $\mathbf{s}_1'$ to $\mathbf{s}_1''$, and from $\mathbf{s}_1''$ to $\mathbf{s}_2$. These phantom samples are presented in Table 2.7.*

Table 2.7 – *Sampling design of Example 2.3 with two phantom samples*

| $k$ | $\mathbf{s}_1$ | $\mathbf{s}_1'$ | $\mathbf{s}_1''$ | $\mathbf{s}_2$ | $\mathbf{s}_3$ | $\mathbf{s}_4$ | $\pi_k$ |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0.2 |
| 2 | 0 | 1 | 1 | 1 | 1 | 0 | 0.7 |
| 3 | 1 | 1 | 0 | 0 | 0 | 1 | 0.3 |
| 4 | 0 | 0 | 1 | 1 | 0 | 0 | 0.6 |
| 5 | 1 | 1 | 1 | 0 | 1 | 1 | 0.4 |
| 6 | 0 | 0 | 0 | 1 | 1 | 1 | 0.8 |
| $p(\mathbf{s}_i)$ | 0.2 | 0 | 0 | 0.6 | 0.1 | 0.1 | $\sum_U \pi_k = 3$ |

*We can define the matrix $\tilde{\mathbf{S}}$ by*

$$\tilde{\mathbf{S}} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}$$

*With the same arguments as used in the previous case, it is possible to show that $\tilde{\mathbf{S}}$ has full rank. Hence, the sampling design has a minimum support.*

□

## 2.5 Remarks on systematic sampling

The previous results can be used to compute the joint inclusion probabilities. Since the size of the support is small (at most equal to $N$), a simple way, presented in Algorithm 2, consists in directly computing the support.

**Algorithm 2**: Computation of the joint inclusion probabilities

Compute $r_{(i)}$, $i = 1, \ldots, D$; $r_{(D+1)} = 1$.
For $i = 1, \ldots, D$

- Compute the sampling design $p(\mathbf{s}_i) = r_{(i+1)} - r_{(i)}$.
- Compute $u_i = (r_{(i+1)} + r_{(i)})/2$.
- Compute $\mathbf{s}_i = (s_{1i}, \ldots, s_{Ni})'$ by

$$s_{ki} = \begin{cases} 1 & \text{if there exists an integer } j > 0 \\ & \text{such that } V_{k-1} \leq u_i + j - 1 < V_k, \\ 0 & \text{otherwise.} \end{cases}$$

Finally, the matrix of joint inclusion probabilities is given by

$$\boldsymbol{\Pi} = \sum_{i=1}^{D} \mathbf{s}_i \mathbf{s}_i' p(\mathbf{s}_i).$$

The validity of Algorithm 2 derives from the fact that, at each step $i$, a sample $\mathbf{s}_i$ is calculated. This sample corresponds to the situation where the random number $u$ of Algorithm 1 'falls' in $[r_{(i)}, r_{(i+1)})$. Any number between $r_{(i)}$ and $r_{(i+1)}$ can be used to determine this sample. In Algorithm 2, we use $(r_{(i+1)} + r_{(i)})/2$. The probability of selecting the sample $\mathbf{s}_i$ is the length of the interval $[r_{(i)}, r_{(i+1)})$ and thus $p(\mathbf{s}_i) = r_{(i+1)} - r_{(i)}$.

This method is used in the version 0.2 of the 'R' sampling package (see Tillé & Matei, 2005). Note that Algorithm 2 works even if all the $r_k$'s are not distinct. The computation of the joint inclusion probabilities of Examples 2.2 and 2.3 of Section 2.2 are given in Tables 2.8 and 2.9.

Table 2.8 – *Joint inclusion probabilities of Example 2.2*

| $k\backslash\ell$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | 0.3 | 0.1 | 0.2 | 0.2 | 0.1 | 0 |
| 2 | | 0.8 | 0.2 | 0.6 | 0.5 | 0.2 |
| 3 | | | 0.4 | 0.1 | 0.3 | 0 |
| 4 | | | | 0.7 | 0.3 | 0.2 |
| 5 | | | | | 0.6 | 0 |
| 6 | | | | | | 0.2 |

The joint inclusion probabilities of Examples 2.2 and 2.3 present the main drawback of systematic sampling: some of them are equal to zero. A currently advocated solution to this drawback is called 'random systematic sampling' and consists in randomly sorting the population before applying systematic sampling. However, Brewer & Hanif (1983, Procedure 2, page 22) have given an example where, for any permutation of the pop-

Table 2.9 – *Joint inclusion probabilities of Example 2.3*

| $k\backslash\ell$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | 0.2 | 0 | 0.2 | 0 | 0.2 | 0 |
| 2 | | 0.7 | 0 | 0.6 | 0.1 | 0.7 |
| 3 | | | 0.3 | 0 | 0.3 | 0.1 |
| 4 | | | | 0.6 | 0 | 0.6 |
| 5 | | | | | 0.4 | 0.2 |
| 6 | | | | | | 0.8 |

ulation, the joint inclusion probability of the two units with the smallest $\pi_k$'s is equal to zero. Here is another example:

**Example 2.4** Suppose that $N = 5, n = 2$ and $\pi_1 = \pi_2 = 0.25, \pi_3 = \pi_4 = \pi_5 = 0.5$. The $\pi_k$'s, $V_k$'s and $r_k$'s are given in Table 2.10.

Table 2.10 – *Values of $\pi_k$, $V_k$ and $r_k$ for Example 2.4*

| $k$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $\pi_k$ | 0.25 | 0.25 | 0.5 | 0.5 | 0.5 |
| $V_k$ | 0.25 | 0.5 | 1 | 1.5 | 2 |
| $r_k$ | 0.25 | 0.5 | 0 | 0.5 | 0 |

Independently from the order of the units, the probability of jointly selecting the two units that have their inclusion probability equal to 0.25 is always null. The joint inclusion probabilities of this example are given in Table 2.11.

Table 2.11 – *Joint inclusion probabilities of Example 2.4*

| $k\backslash\ell$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 0.25 | 0 | 0 | 0.25 | 0 |
| 2 | | 0.25 | 0 | 0.25 | 0 |
| 3 | | | 0.5 | 0 | 0.5 |
| 4 | | | | 0.5 | 0 |
| 5 | | | | | 0.5 |

Note that there are 10=$\binom{5}{2}$ possibilities to arrange the '0.25' among the 5 units. We will not carry out the computation of the joint inclusion probabilities for the 10 cases as it can be seen that, with systematic sampling, only the relative positions, under cyclical order, of the units matter. So, in order to cover all the situations, we just need to examine the first two cases: $\boldsymbol{\pi} = (0.25, 0.25, 0.5, 0.5, 0.5)'$ and $\boldsymbol{\pi} = (0.25, 0.5, 0.25, 0.5, 0.5)'$.

The joint inclusion probabilities for the first case have already been given in Table 2.11 while those for the second case are given in Table 2.12.

In both situations, it is impossible to jointly select the two units that have the smallest inclusion probabilities. Since it is valid for the 10 possible

Table 2.12 – *Joint inclusion probabilities for the case* $\boldsymbol{\pi} = (0.25, 0.5, 0.25, 0.5, 0.5)'$

| $k\backslash\ell$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 0.25 | 0 | 0 | 0.25 | 0 |
| 2 | | 0.5 | 0 | 0.25 | 0.25 |
| 3 | | | 0.25 | 0 | 0.25 |
| 4 | | | | 0.5 | 0 |
| 5 | | | | | 0.5 |

permutations, the joint inclusion probability of the two units that have the smallest inclusion probabilities is equal to zero.

We have seen that systematic sampling is a minimum support design. Random systematic sampling consists in randomly choosing a minimum support design, this choice being given by the $N!$ permutations of the population. It may happen that some samples will always have a null probability of being selected. We show in the next section that it is always possible to include any sample in a minimum support design. That means that random systematic sampling does not randomly select a design in the set of all the minimum support designs, but rather in a restricted subset of this set. In the next section, we propose an alternate procedure for randomly selecting a minimum support design whereby all the possible samples have a positive probability of being obtained.

## 2.6 Minimum support design

A minimum support design procedure was already proposed by Jessen (1969) (see also Brewer & Hanif, 1983, procedures 35 and 36, pp. 42-43). Hedayat et al. (1989, Theorem 2) have also proposed the method of emptying boxes, but its implementation is limited to inclusion probabilities that can be written as rational numbers. The minimum support procedure described in this section is a particular case of the splitting method proposed by Deville & Tillé (1998) (see also Tillé, 2006). A formal description of this procedure is presented in Algorithm 3.

This algorithm provides a design that respects the inclusion probabilities $\boldsymbol{\pi}$, and that has a support of at most $N$ linearly independent samples. This property derives from the fact that, at each step, the procedure chooses one sample that is independent from all the other remaining possible samples or keeps going on. Indeed, at each step, either a group of units is selected and completes the sample, or the procedure follows through with at least one unit in this group that will not be selected in any of the possible remaining samples or, alternately, one unit outside of this group that is part of all the remaining samples. Note that the set $D_t$ can be randomly chosen or not at each step of Algorithm 3. That means that any sample $\mathbf{s}$ can be included in a minimum support design in this procedure, it just needs to be selected at the first step $D_0$. In Table 2.13, with the minimum support design procedure, we have constructed a sampling design

**Algorithm 3**: Minimum support procedure

Set $\boldsymbol{\pi}(0) = \boldsymbol{\pi}$;
For $t = 0, 1, 2 \ldots$, and until obtaining a sample, i.e. a vector $\boldsymbol{\pi}(t)$ with all its coordinates in $\{0, 1\}$, do

1. define
   $A_t = \{k : \pi_k(t) = 0\}$, $B_t = \{k : \pi_k(t) = 1\}$, and
   $C_t = \{k : 0 < \pi_k(t) < 1\}$,

2. select a subset $D_t$ of $C_t$ such that $\text{card}D_t = n - \text{card}B_t$ ($D_t$ can be randomly selected or not),

3. define

$$\pi_k^a = \begin{cases} 0 & \text{if } k \in A_t \cup (C_t \backslash D_t), \\ 1 & \text{if } k \in B_t \cup D_t, \end{cases}$$

$$\alpha(t) = \min\{1 - \max_{k \in (C_t \backslash D_t)} \pi_k(t), \min_{k \in D_t} \pi_k(t)\},$$

   and

$$\pi_k^b = \begin{cases} 0 & \text{if } k \in A_t, \\ 1 & \text{if } k \in B_t, \\ \dfrac{\pi_k(t)}{1 - \alpha(t)} & \text{if } k \in (C_t \backslash D_t), \\ \dfrac{\pi_k(t) - \alpha(t)}{1 - \alpha(t)} & \text{if } k \in D_t, \end{cases}$$

4. select $\boldsymbol{\pi}(t+1) = \begin{cases} \boldsymbol{\pi}^a & \text{with probability } \alpha(t), \\ \boldsymbol{\pi}^b & \text{with probability } 1 - \alpha(t). \end{cases}$

that includes the two units with the smallest inclusion probabilities of the Example 2.4 in Section 2.5.

Table 2.13 – *Example of a minimum support design including the two units with the smallest* $\pi_k$

| $k$ | $\mathbf{s}_1$ | $\mathbf{s}_2$ | $\mathbf{s}_3$ | $\mathbf{s}_4$ | $\pi_k$ |
|---|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 0 | 0.25 |
| 2 | 1 | 0 | 0 | 0 | 0.25 |
| 3 | 0 | 1 | 1 | 0 | 0.5 |
| 4 | 0 | 1 | 0 | 1 | 0.5 |
| 5 | 0 | 0 | 1 | 1 | 0.5 |
| $\mathbf{s}$ | 0.25 | 0.25 | 0.25 | 0.25 | 2 |

Since Algorithm 3 allows one to include any sample in a minimum support design, an interesting alternative procedure to random systematic sampling could consist in using Algorithm 3 and randomly selecting at each step $D_t$ by simple random sampling. This new procedure could be

called 'random minimum support design'. Since at the first step each sample can be selected, the 'random minimum support design' ensures that all the samples have a positive probability of being selected. All the joint inclusion probabilities are thus strictly positive, which is not the case with random systematic sampling.

## 2.7 Discussion

The entropy of a sampling design is defined as

$$I(p) = -\sum_{\mathbf{s}} p(\mathbf{s}) \log p(\mathbf{s}).$$

The entropy of a design (or of a probability mass function) is large when the distribution is strongly spread on its support, and when the support is large. Since systematic sampling has a minimal support, the mass probability is concentrated on a small set of samples. Therefore, systematic sampling usually has a small entropy.

Even with random systematic sampling, some samples are never selected, so the support of random systematic sampling is not maximal. The minimum support design procedure described in Section 2.6 allows any sample to be included in the support of the sampling design. The random minimum support design, for example, has a maximal support, and is likely to have a higher entropy than random systematic sampling. High entropy is an important property when developing asymptotic arguments for central limit theorems (see Brewer & Donadio, 2003; Berger, 1998).

If the units of the population are ordered according to an auxiliary variable correlated to the variable of interest, systematic sampling provides a gain in accuracy. The variance is difficult to estimate, but there are good conservative estimates of it (see Iachan, 1982). However, if the units are not ordered, systematic sampling should be avoided, especially in small populations, even if the population is randomly sorted before the selection of the sample.

The implementation of random systematic sampling and of the random minimum support design is available in the package 'sampling' of the 'R' language. The following code allows a sample of 20 municipalities to be selected from the 44 municipalities of the province of Luxembourg in Belgium.

```
> library("sampling");data(belgianmunicipalities)
> attach(belgianmunicipalities)
> Tot=Tot04[Province==8]
> name=Commune[Province==8]
> pik=inclusionprobabilities(Tot,20)
> # selection of a sample with the random minimal support method
> as.vector(name[UPminimalsupport(pik)==1])
 [1] "Arlon"              "Attert"             "Aubange"
 [4] "Martelange"         "Bastogne"           "Fauvillers"
 [7] "Houffalize"         "Vielsalm"           "Marche-en-Famenne"
[10] "Nassogne"           "Herbeumont"         "Libin"
[13] "Paliseul"           "Wellin"             "Libramont-Chevigny"
[16] "Etalle"             "Meix-devant-Virton" "Musson"
```

```
[19] "Saint-Léger"        "Virton"
> # selection of a sample with the random systematic sampling
> as.vector(name[UPrandomsystematic(pik)  ==1])
 [1] "Arlon"              "Aubange"            "Bastogne"
 [4] "Vielsalm"           "Gouvy"              "Durbuy"
 [7] "Marche-en-Famenne"  "Bertrix"            "Bouillon"
[10] "Léglise"            "Neufchâteau"        "Saint-Hubert"
[13] "Wellin"             "Libramont-Chevigny" "Florenville"
[16] "Meix-devant-Virton" "Musson"             "Saint-Léger"
[19] "Virton"             "Habay"
```

The municipalities are selected with inclusion probabilities proportional to their number of inhabitants. The package is very easy to use, and we suggest using the random minimum support design rather than systematic sampling.

## Acknowledgements

The authors are grateful for the numerous suggestions made by Alina Matei. The authors would like to thank three reviewers for their very positive comments that allowed us to significantly improve the presentation of this paper. This work was supported in part by the Swiss National Science Foundation, (grant FN 205121-105187/1), when Johan Pea was a researcher at the University of Neuchâtel. The opinions expressed in this paper are those of the authors and not necessarily those of the institutions where they work.

# 3 Minimum entropy, maximum entropy and equal treatment of variables

**Abstract**

We discuss some natural questions inspired by Chapters 1 and 2. In Chapter 1, it is proven that conditional Poisson sampling associated with Horvitz-Thompson estimators is uniformly more efficient than sampling with replacement. In Chapter 2, we proved that systematic sampling is a minimum support design, and stated without proof that minimum support designs should have a low entropy. We begin this chapter by a short example in Section 3.1.1 that shows that systematic sampling is not uniformly more efficient than sampling with replacement. Then, in Section 3.1.2, we study the entropy of minimum support designs. Finally, all these considerations on the entropy of sampling designs led to the inevitable question: why are we interested in that quantity when, in practice, all that is of interest is the variance of Horvitz-Thompson estimators? The natural argument is that, by using a maximal randomization under some constraints (size, inclusion probabilities), we should achieve a maximal equity of treatment for the variables once the information in the constraints is removed (that is to say, equity for some regression residuals). In Section 3.2, we study the dispersion of variances of Horvitz-Thompson estimators for variables that lie on the unit sphere of three natural metrics. We show that, in general, this dispersion cannot be null for fixed size designs with unequal inclusion probabilities, and thus that there cannot be a strict equal treatment of variables (orthogonal to the vector of inclusion probabilities). We also show that conditional Poisson sampling does not give a minimal dispersion.

## 3.1 Considerations on entropy

### 3.1.1 Entropy and superiority over sampling with replacement

There is no sampling design without replacement uniformly superior to others with the same inclusion probabilities. One way to judge whether a fixed size design with unequal probability allows to estimate totals for all variables with a reasonable accuracy is to compare it to sampling with

replacement. We have seen in Chapter 1 that maximum entropy sampling with fixed size is more efficient than sampling with replacement. This property has also been proven to hold for other sampling designs that can be regarded as high entropy designs. All the quantities involved being continuous functions of the sampling distribution, this property will automatically hold for sampling designs close enough to the maximum entropy design. This result is not constructive, we remain unable to specify explicitly what close enough stands for, or a lower bound on the entropy of a sampling design that ensures that it is more efficient than sampling with replacement.

We may examine the case of systematic sampling, which can be viewed as a low entropy design, and compare it to sampling with replacement. The variance of the Horvitz-Thompson estimator under $\pi ps$ systematic sampling is complex as it depends on the order in the population. Nevertheless, when all units have the same inclusion probability and when $N/n$ is an integer, we obtain simple expressions. Let us note $V_{\mathcal{HH}}$ the variance of the Hansen-Hurwitz estimator of total of a variable $Y$ under sampling with replacement, and target inclusion probabilities $\pi_k^\star = n/N$, for all $k \in U$, and $V_{SYST}$ the variance of the Horvitz-Thompson estimator under systematic sampling with replacement. Then

$$V_{\mathcal{HH}} = \frac{N(N-1)}{n} S_Y^2, \text{ where } S_Y^2 = \frac{1}{N-1} \sum_{k \in U} (y_k - \bar{Y})^2,$$

and $\bar{Y}$ is the mean of $Y$. In this case, systematic sampling results in exactly $K = N/n$ samples $\mathbf{s}_1, \ldots, \mathbf{s}_K$ having a non-null probability of being selected. If we note

$$\begin{aligned} S_k^2(y) &= \frac{1}{n-1} \sum_{i \in s_k} (y_i - \bar{Y}_k)^2, \; k = 1, \ldots, K, \\ \text{and } S_{Y_K}^2 &= \frac{1}{K-1} \sum_{k=1}^{K} (\bar{Y}_k - \bar{Y})^2, \end{aligned}$$

with $\bar{Y}_k$ the mean of $Y$ in the sample $\mathbf{s}_k$, we have:

$$V_{SYST} = Nn(K-1) S_{Y_K}^2.$$

Using the relation

$$(N-1) S_Y^2 = n(K-1) S_{Y_K}^2 + (n-1) \sum_{k=1}^{K} S_k^2(y),$$

we obtain

$$V_{SYST} - V_{\mathcal{HH}} = (n-1) \left( V_{\mathcal{HH}} - N \sum_{k=1}^{K} S_k^2(y) \right).$$

Now this quantity is not always positive, and we know that systematic sampling can be optimal, for the estimation of the total of a variable $Y$ linked to the inclusion probabilities by the superpopulation model given

in Hájek (1959). And indeed, when the dispersion in the samples is such that $S_Y^2 = S_k^2(y) = S^2$ for all $k$, the last equation yields

$$V_{SYST} - V_{\mathcal{HH}} = -\frac{N(n-1)}{n} S^2.$$

When the mean of $Y$ is equal in each cluster to $\bar{Y}$, systematic sampling gives a perfect estimator. But when on the other hand $S_k^2(y) = 0$ for all $k$, then $V_{SYST} = nV_{\mathcal{HH}}$. Consequently, systematic sampling fails to be an acceptable design for all variables.

### 3.1.2 Entropy and size of the support

It is stated in Chapter 2 that minimum support designs usually have a small entropy. we shall try to give a more precise meaning to this general statement. We begin with some simple geometrical considerations. If $N$ is the size of the population $U$, the maximal support of a sampling design on $U$ is the set $\mathcal{S}$ of vertices of $[0,1]^N$ (we will use the notation $\mathbf{S}$ for the matrix that has these vertices as columns) and all sampling designs can be represented by $2^N$-vectors $\mathbf{p} = (p_1, \ldots, p_{2^N})'$, with $0 \leq p_1, \ldots, p_{2^N} \leq 1$ and $\sum_i p_i = 1$. Sampling designs that have a given vector of inclusion probabilities $\boldsymbol{\pi} = \mathbf{S} \cdot \mathbf{p} = (\pi_1, \ldots, \pi_N)'$ form a convex polytope, noted $\mathcal{C}_\pi$, of $\mathbb{R}^{2^N}$. If $\boldsymbol{\pi}$ is a $0,1-$vector, $\mathcal{C}_\pi$ is degenerate, but if the inclusion probabilities are not all equal to 0 or 1, this polytope holds more than one sampling design vector. Furthermore, if $\sum_k \pi_k = n$, we can consider $\mathcal{C}_\pi^n$ the intersection of $\mathcal{C}_\pi$ with the polytope of sampling designs with fixed size $n$, and $\mathcal{S}_n$ the subset of all samples of size $n$. The entropy of a sampling design $\mathbf{p} \in \mathcal{C}_\pi$ is defined as:

$$\mathcal{H} = \begin{pmatrix} \mathcal{C}_\pi & \longrightarrow & \mathbb{R} \\ \mathbf{p} = (p_1, \ldots, p_{2^N})' & \mapsto & -\sum_i p_i \log p_i \end{pmatrix},$$

where $0 \log 0 = 0$. Maximum entropy sampling designs in $\mathcal{C}_\pi$ (resp. $\mathcal{C}_\pi^n$) have a support equal to $\mathcal{S}$ (resp. $\mathcal{S}_n$). At last, we get to the point:

**Proposition 3.1**

1. *Minimum support designs are the extremal points of $\mathcal{C}_\pi$, or of $\mathcal{C}_\pi^n$ if we consider fixed size designs only.*
2. *Entropy reaches its local minima in $\mathcal{C}_\pi$ or $\mathcal{C}_\pi^n$ at minimum suport designs, and a design that is a global minimum of entropy is also a minimum support design.*

A proof of 3.1 is given in the appendix, Section A.1, page 115.

**Remark 3.1**

- *Minimum support designs do not all have the same size of support, even if we restrict ourselves to fixed sample size designs, and they can have different entropy. For example, if $N = 4$, $n = 2$, $\pi_1 = 1/3$, $\pi_2 = 2/3$, $\pi_3 = 2/5$, $\pi_4 = 3/5$, we consider the sampling design $p_1(\cdot)$ given by systematic sampling with the initial order on the population and the sampling, and represented by a vector $\mathbf{p}_1$, design $p_2(\cdot)$ given by systematic sampling when*

*units* 2 *and* 3 *are permuted and represented by* $\mathbf{p}_2$. *We have that:*

$$\begin{aligned} p_1(\{1,3\}) = \frac{5}{15}, \quad p_1(\{2,3\}) = \frac{1}{15}, \quad p_1(\{2,4\}) &= \frac{9}{15}, \\ p_2(\{1,2\}) = \frac{5}{15}, \quad p_2(\{3,2\}) = \frac{1}{15}, \quad p_2(\{3,4\}) &= \frac{5}{15}, \\ p_2(\{2,4\}) &= \frac{4}{15}, \\ \mathcal{H}(\mathbf{p}_1) \approx 0.85, \quad \mathcal{H}(\mathbf{p}_2) &=\approx 1.27. \end{aligned}$$

- *Not all minimum support sampling designs can be obtained with systematic sampling or algorithm* 3. *Formally, let* $C$ *be the set of minimum support designs,* $C^1$ *be the set of sampling designs given by systematic sampling for all* $N!$ *permutations of the population and* $C^2$ *be the set of sampling designs given by the minimum support algorithm* 3 *for any choice of the sets* $D_t$. *Then, in general,* $C^1$ *and* $C^2$ *are strict subsets of* $C$, $C^1 \backslash C^2 \neq \emptyset$, *and* $C^2 \backslash C^1 \neq \emptyset$. *The fact that algorithm* 3 *can reach sampling designs that cannot be obtained with systematic sampling and any order on the population is the reason for which we advocate the use of this algorithm, and we gave in example* 2.4 *a case that proves that there is no reason for* $C^2$ *to be a subset of* $C^1$. *On the other hand, there are designs in* $C^1$ *that are not in* $C^2$. *For example, if* $N = 7$, $n = 2$, $\pi_1 = 0.2$, $\pi_2 = 0.2$, $\pi_3 = 0.2$, $\pi_4 = 0.5$ $\pi_5 = 0.15$, $\pi_6 = 0.2$, $\pi_7 = 0.55$, *systematic sampling with the initial order in the population has the following support:*

$$\mathbf{S}_p = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix},$$

*and this support cannot be obtained with algorithm* 3. *Indeed, algorithm* 3 *is such that, at each step, a decision is made that forces one unit to be in all remaining samples or in none of them. So, if one of the columns of* $\mathbf{S}_p$ *was the first sample* $D_0$ *that can be selected by the algorithm, there would be at least one of the units in this sample that would not be selected again or one unit outside of* $D_0$ *that would be part of all remaining samples. That is not the case since every unit is selected in exactly two samples here: no unit is elected in six samples and no unit is selected only once.*

- *My initial idea was that all minimum support designs are local minima of the entropy, however the proof of this statement still eludes me.*

## 3.2 Dispersion of the variance of Horvitz-Thompson estimators

Among other motivations for using high entropy sampling, one is that it is by definition a method that gives a sampling distribution with a max-

imal dispersion under the constraints it has been imparted. In that way, we hope that it manages not to *disfavor* or on the contrary to *favor* the estimation of a variable *more than necessary*. It is viewed as a safeguard against unwanted artefacts that could be introduced with other sampling algorithms (e.g. systematic sampling). Ideally, we would want totals of all variables in $\mathbb{R}^N$ or in a linear subspace, and having the same norm for some given metric, to be estimated with the same precision. A related problem, is to find a design that minimizes the maximum variance of estimators for such variables. Gabler and Stenger gave several results on minimax strategies in survey sampling (on this subject, see Wynn, 1976; Stenger, 1979; Gabler, 1990). They derived in Stenger & Gabler (1996), for any given metric on the space of centered interest variables, a lower bound for the 'minimax value' of fixed size designs. They also computed second order inclusion probabilities of a sampling design that would reach this bound. In general, however, the existence of a sampling design that attains this bound cannot be guaranteed.

We give here a different approach as we restrict ourselves to the Horvitz-Thompson estimator. We give some results on the minimaxity of maximum entropy sampling designs, and also study a simpler, if less useful, property than the minimax property. We consider the dispersion of the eigenvalues of the variance operator associated with a sampling design with given inclusion probabilities and try to minimize or maximize this dispersion. We show that it is impossible to treat equally all variables orthogonal to the inclusion probabilities in a fixed size design when these inclusion probabilities are not all equal. We give additional results on maximum entropy sampling, in the case of random size sampling and fixed size sampling, minimum support designs, and on dispersions in different metrics.

### 3.2.1 Definitions - Notations

Let $p(\cdot)$ be a sampling design without replacement, with inclusion probabilities $\pi_k, k \in U$, covariance matrix $\mathbf{\Delta}$,

$$\begin{aligned} \mathbf{\Delta} &= [\Delta_{k\ell}],\ k,\ell = 1,\dots,N,\ \text{where} \\ \Delta_{k,\ell} &= \pi_{k,\ell} - \pi_k\pi_\ell \text{ if } k \neq \ell,\ \text{and } \Delta_{k,k} = \pi_k(1-\pi_k), \end{aligned}$$

support $\mathcal{Q}_p = \{s_1,\dots,s_P\}$ and support matrix $\mathbf{S}_p = (\mathbf{s}_1|\dots|\mathbf{s}_P)$. Let also $\mathbf{V}_p$ denote the covariance matrix of the Horvitz-Thompson estimator for $p(\cdot)$, $\mathbf{V}_p = \mathbf{\Phi}^{-1}\mathbf{\Delta}\mathbf{\Phi}^{-1}$, with $\mathbf{\Phi}$ the diagonal matrix that has coefficients $\pi_k$ on its diagonal,

$$\mathbf{\Phi} = \begin{pmatrix} \pi_1 & & \\ & \ddots & \\ & & \pi_N \end{pmatrix},$$

and

$$\mathbf{V}_p = \begin{pmatrix} \frac{1-\pi_1}{\pi_1} & \frac{\pi_{1,2}-\pi_1\pi_2}{\pi_1\pi_2} & \cdots & \frac{\pi_{1,N}-\pi_1\pi_N}{\pi_1\pi_N} \\ \frac{\pi_{1,2}-\pi_1\pi_2}{\pi_1\pi_2} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ \frac{\pi_{1,N}-\pi_1\pi_N}{\pi_1\pi_N} & \cdots & \cdots & \frac{1-\pi_N}{\pi_N} \end{pmatrix}.$$

Let $\widehat{Y}$ denote the Horvitz-Thompson estimator of the total of a variable $Y = (y_1, \ldots, y_N)'$. Its variance is given by

$$\mathrm{var}_p(\widehat{Y}) = Y'\mathbf{V}_p Y.$$

If $\lambda_1 \leq \ldots \leq \lambda_N$ are the eigenvalues of $\boldsymbol{V}_p$ and $(X_1, \ldots, X_N)$ is an orthonormal basis of $\mathbb{R}^N$ such that $X_i$ is an eigenvectors of $\boldsymbol{V}_p$ for the eigenvalue $\lambda_i$, $i = 1, \ldots, N$, then

$$\mathrm{var}_p(\widehat{X}_1) \leq \ldots \leq \mathrm{var}_p(\widehat{X}_N).$$

Let $\mathbf{D}$ be a symmetric positive matrix of size $N$ and note $\|\cdot\|_{\mathbf{D}}$ the associated metric on $\mathbb{R}^N$, defined by:

$$\|Y\|_{\mathbf{D}}^2 = Y'\mathbf{D}Y, \; Y \in \mathbb{R}^N.$$

We consider the variance of Horvitz-Thompson estimators of totals of variables that lie in the unit sphere of $\|\cdot\|_{\mathbf{D}}$.

**Definition 3.1** *Let $r(p, \mathbf{D})$ be the maximum variance of the Horvitz-Thompson estimator of the total of a variable that lies in the unit sphere of $\|\cdot\|_{\mathbf{D}}$.*

$$r(p, \mathbf{D}) = \max\left\{Y'\mathbf{V}_p Y \mid Y'\mathbf{D}Y \leq 1\right\}.$$

*A sampling design that minimizes $r(p, \mathbf{D})$ is called a minimax sampling design for the metric $\mathbf{D}$.*

Naturally, $r(p, \mathbf{D})$ is the largest eigenvalue of the matrix $\mathbf{M}$ defined by

$$\mathbf{M} = \mathbf{D}^{-\frac{1}{2}}\mathbf{V}_p\mathbf{D}^{-\frac{1}{2}}.$$

If $\lambda_1, \ldots, \lambda_N$ are the eigenvalues of $\mathbf{M}$, and if $\lambda_1 \leq \ldots \leq \lambda_N$, we have that:

$$\lambda_1 \|Y\|_{\mathbf{D}}^2 \leq \mathrm{var}(\widehat{Y}) \leq \lambda_N \|Y\|_{\mathbf{D}}^2.$$

There is no point in considering all possible metrics $\mathbf{D}$. For example, if $\mathbf{V}_p$ is non degenerate and $\mathbf{D} = \mathbf{V}_p$, $\mathbf{M}$ is the identity matrix $\mathbf{I}$. If $\mathbf{V}_p$ could be any symmetric positive or semi-definite matrix, and if we knew how to obtain a design with that covariance matrix, we would stop here. Unfortunately, this is not the case (see for example Sinha, 1973; Gabler & Schweigkoffer, 1990). An interesting property would be to estimate with the same precision all variables that lie on the unit sphere of some metric. For this we will need the following definition.

**Definition 3.2** *Let $d(p, \mathbf{D})$ be the sum of squares of the eigenvalues of $\mathbf{M}$,*

$$d(p, \mathbf{D}) = \left(\lambda_1^2 + \ldots + \lambda_N^2\right).$$

*When, $\mathbf{D}$ is diagonal, the sum of the eigenvalues of $\mathbf{M}$ is determined by the first order inclusion probabilities, and we will consider their dispersion:*

$$\delta(p, \mathbf{D}) = \frac{1}{N} d(p, \mathbf{D}) - \frac{1}{N^2} \left(\lambda_1 + \ldots + \lambda_N\right)^2.$$

When $\delta(p, \mathbf{D}) = 0$, all variables that lie on the unit sphere of $\|\cdot\|_{\mathbf{D}}$ have their totals estimated with the same precision.

In minimax problems, the object of interest is the maximum norm of $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_N)'$. Results on $\|\boldsymbol{\lambda}\|_\infty$ are of course much more interesting than results on $\|\boldsymbol{\lambda}\|_2$ as they give an upper bound for the risk $r(p, \mathbf{D})$. But they are also much more difficult to obtain since it is, in general, not possible to extract the exact value of the largest eigenvalue of $\mathbf{M}$ from its coefficients. Later on, we will use the relation $\|\boldsymbol{\lambda}\|_2^2 = \mathrm{Tr}(\mathbf{M}^2)$ to derive properties of sampling designs that minimize $d(p, \mathbf{D})$.

We are particularly interested in three diagonal metrics: $\mathbf{D} = \mathbf{I}$, $\mathbf{D} = \boldsymbol{\Psi}$, where $\boldsymbol{\Psi}$ is the matrix with $(1 - \pi_1)/\pi_1, \ldots, (1 - \pi_N)/\pi_N$ on its diagonal and 0 elsewhere,

$$\boldsymbol{\Psi} = \begin{pmatrix} \frac{1-\pi_1}{\pi_1} & & \\ & \ddots & \\ & & \frac{1-\pi_N}{\pi_N} \end{pmatrix}.$$

This metric allows to compare the variance of Horvitz-Thompson estimators under a sampling design with the variance under Poisson sampling with the same inclusion probabilities. In that case, we will note $\widetilde{\boldsymbol{V}}_p = \mathbf{M}$. Remark that

$$\widetilde{\boldsymbol{V}}_p = \begin{pmatrix} 1 & \frac{\pi_{1,2} - \pi_1\pi_2}{[\pi_1(1-\pi_1)\pi_2(1-\pi_2)]^{\frac{1}{2}}} & \cdots & \frac{\pi_{1,N} - \pi_1\pi_N}{[\pi_1(1-\pi_1)\pi_N(1-\pi_N)]^{\frac{1}{2}}} \\ \frac{\pi_{1,2} - \pi_1\pi_2}{[\pi_1(1-\pi_1)\pi_2(1-\pi_2)]^{\frac{1}{2}}} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ \frac{\pi_{1,N} - \pi_1\pi_N}{[\pi_1(1-\pi_1)\pi_N(1-\pi_N)]^{\frac{1}{2}}} & \cdots & \cdots & 1 \end{pmatrix},$$

is the matrix of correlation of indicators $I_k(s)$,

$$I_k(s) = \begin{cases} 1 \text{ if } k \in s, \\ 0 \text{ otherwise} \end{cases} \qquad k = 1, \ldots, N,\ s \in \mathcal{Q}_p.$$

And for fixed size designs, we will also consider the case $\mathbf{D} = \boldsymbol{\Phi}^{-1}$, that allows to compare the variance of the Horvitz-Thompson estimator with the Hansen-Hurwitz formula of variance. In section 3.2.2 we give results on $d(p, \mathbf{D})$ and $r(p, \mathbf{D})$ for random size maximum entropy sampling designs and a choice of $\mathbf{D}$. In section 3.2.3, we give results for fixed size sampling designs. We also give necessary conditions for a sampling design to be an extremum of $d(p, \mathbf{D})$ and prove that conditional Poisson sampling is not a minimum of $d(p, \mathbf{D})$ nor of $r(p, \mathbf{D})$ for all three considered metrics.

### 3.2.2 Dispersion for random size designs

We consider the case where $\mathbf{D}$ is the identity matrix $\mathbf{I}$, so that we compare $\text{var}(\widehat{Y})$ with $\|Y\|_2$. We have

$$d(p,\mathbf{I}) = \lambda_1^2 + \ldots + \lambda_N^2 = \sum_{k=1}^{N}\left(\frac{1-\pi_k}{\pi_k}\right)^2 + \sum_{k=1}^{N}\sum_{\ell\neq k}\left(\frac{\pi_{k,\ell}-\pi_k\pi_\ell}{\pi_k\pi_\ell}\right)^2,$$

where $\lambda_1,\ldots,\lambda_N$ are the eigenvalues of $\mathbf{V}_p$. We can consider the dispersion of the eigenvalues:

$$\delta(p,\mathbf{I}) = \frac{1}{N}d(p,\mathbf{I}) - \frac{1}{N^2}\left(\lambda_1 + \ldots + \lambda_N\right)^2.$$

When this dispersion $\delta(p,\mathbf{I})$ is null, all variables $Y$ that have the same 2-norm are estimated with the same precision. When this dispersion can not be null, as with unequal probability sampling designs, we can still look for designs that minimize this dispersion. The next proposition states that maximum entropy sampling is one family of designs that minimize $\delta(p,\mathbf{I})$.

**Proposition 3.2** *Maximum entropy sampling designs minimize the dispersion $\delta(p,\mathbf{D})$ in the following cases:*

1. *$\delta(p,\mathbf{I})$ is null for the Bernoulli sampling design,*
2. *for a given vector of inclusion probabilities, Poisson sampling design minimizes $\delta(p,\mathbf{I})$.*
3. *Poisson sampling design also minimizes $r(p,\mathbf{I})$.*

A proof of 3.2 is given in appendix, Section A.2, page 116.

**Remark 3.2** *We are looking at properties that depend only on first and second order inclusion probabilities and which can not completely characterize a sampling design, except in very special cases, for example for fixed size sampling of size 2. In general there exist several sampling designs that have a same given set of first and second order inclusion probabilities when the number of samples in the maximal support $\mathcal{S} = \{s \subset U\}$ or $\mathcal{S}_n = \{s \subset U \text{ s.t. } |s| = n\}$ exceeds $N(N+1)/2$ (on this subject, see Wynn, 1977).*

With unequal probability sampling, $\delta(p,\mathbf{I})$ can not be null, but there is another measure of dispersion that can be. Let $\widetilde{\boldsymbol{V}}_p$ be the matrix given by $\widetilde{\boldsymbol{V}}_p = \boldsymbol{\Psi}^{-\frac{1}{2}}\boldsymbol{V}_p\boldsymbol{\Psi}^{-\frac{1}{2}}$. We have:

$$\text{var}_p(\widehat{Y}) = Y'\boldsymbol{\Psi}^{\frac{1}{2}}\widetilde{\boldsymbol{V}}_p\boldsymbol{\Psi}^{\frac{1}{2}}Y,$$

and in the case of Poisson sampling, $\widetilde{\boldsymbol{V}}_p$ is the identity matrix. All the variables that have the same norm in the metric induced by $\boldsymbol{\Psi}$,

$$\|Y\|_{\boldsymbol{\Psi}}^2 = Y'\boldsymbol{\Psi}Y = \sum_{k=1}^{N} y_k^2\frac{1-\pi_k}{\pi_k},$$

have their total estimated with the same performance under Poisson sampling. We can thus consider another measure of dispersion, $\delta(p, \mathbf{\Psi})$ defined to be equal to the dispersion of the eigenvalues of $\widetilde{\boldsymbol{V}}_p$:

$$\delta(p, \mathbf{\Psi}) = \frac{1}{N} d(p, \mathbf{\Psi}) - \frac{1}{N^2} \left(\tilde{\lambda}_1 + \ldots + \tilde{\lambda}_N\right)^2,$$

where $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_N$ are the eigenvalues of $\widetilde{\boldsymbol{V}}_p$. The dispersion $\delta(p, \mathbf{\Psi})$ is minimal exactly when $d(p, \mathbf{\Psi})$ is minimal. Proposition 3.2 translates in term of $\delta(p, \mathbf{\Psi})$ and $r(p, \mathbf{\Psi})$ to:

**Proposition 3.3**

1. *For a given set of inclusion probabilities Poisson sampling design minimizes $d(p, \mathbf{\Psi})$ and $\delta(p, \mathbf{\Psi})$, and is such that $\delta(p, \mathbf{\Psi}) = 0$.*
2. *Poisson sampling design $\mathcal{P}(\cdot)$ also minimizes $r(p, \mathbf{\Psi})$ in the set of sampling designs with given inclusion probabilities $(\pi_1, \ldots, \pi_N)'$, and $r(\mathcal{P}, \mathbf{\Psi}) = 1$.*

**Remark 3.3**

- *When $p(\cdot)$ is a sampling design with equal inclusion probabilities, $\boldsymbol{V}_p$ and $\widetilde{\boldsymbol{V}}_p$ are proportional, and so are $\delta(p, \mathbf{I})$ and $\delta(p, \mathbf{\Psi})$. In general, when $\boldsymbol{V}_p$ is not diagonal, there is no obvious reason for $\delta(p, \mathbf{I})$ and $\delta(p, \mathbf{\Psi})$ to satisfy a monotony property. For example $\delta(p_1, \mathbf{I}) \leq \delta(p_2, \mathbf{I})$ does not necessarily imply that $\delta(p_1, \mathbf{\Psi}) \leq \delta(p_2, \mathbf{\Psi})$.*
- *The minimality of $\delta(p, \mathbf{\Psi})$ and $\delta(p, \mathbf{I})$ is obtained with Poisson sampling, but also with any other sampling design such that $\pi_{k,\ell} = \pi_k \pi_\ell$, for all $1 \leq k \neq \ell \leq N$.*
- *Poisson sampling design $\mathcal{P}(\cdot)$ minimizes $r(p, \mathbf{D})$ for any diagonal metric $\mathbf{D}$.*

### 3.2.3 Dispersion for fixed size designs

If $p(\cdot)$ is a fixed size sampling design, there is one direction that is naturally advantaged: the estimator of the total of $\boldsymbol{\pi}$ is always equal to its true value $n$. Hence the smallest eigenvalue of $\boldsymbol{V}_p$ is null. If $t_Y$ is the total of $Y$, and we note $\check{Y} = (\pi_k t_Y / \sum \pi_i)$, $k = 1, \ldots, N$, and $Z = Y - \check{Y}$, then we have $\mathrm{var}(\widehat{Y}) = \mathrm{var}(\widehat{Z})$. In this case, as in Cheng & Li (1983), it is natural to compare $\mathrm{var}(\widehat{Y})$ with $\|Z\|^2_{\mathbf{D}}$. We still have:

$$r(p, \mathbf{D}) = \max \left\{ Y' \mathbf{V}_p Y = Z' \mathbf{V}_p Z \mid Z' \mathbf{D} Z \leq 1 \right\}.$$

Indeed, $r(p, \mathbf{D})$ is also equal to the largest eigenvalue of the matrix $\check{\mathbf{M}}$ defined as the restriction of $\mathbf{M} = \mathbf{D}^{-\frac{1}{2}} \mathbf{V}_p \mathbf{D}^{-\frac{1}{2}}$ to the orthogonal of $\mathbf{D}^{\frac{1}{2}} \boldsymbol{\pi}$. In fact, instead of $Z = Y - \check{Y}$, we can use any variable of the form $Y - a\boldsymbol{\pi}$, and have the same value for $r(p, \mathbf{D})$. For example, we can consider

$$\widetilde{Z} = Y - \boldsymbol{\pi} \frac{\sum_{k \in U} \pi_k y_k}{\sum_{k \in U} \pi_k^2}.$$

Naturally, the eigenvalues of $\check{\mathbf{M}}$ are the 'other' eigenvalues of $\mathbf{M}$ (there could still be a zero in the spectrum of $\check{\mathbf{M}}$, if the sampling design is balanced on some other variable, but with an order of multiplicity decreased

by one). If $\mathbf{D} = \mathbf{I}$, then

$$r(p, \mathbf{I}) = \sup_{Y \neq 0} \frac{\text{var}(\widehat{Y})}{S_y^2},$$

is, up to a multiplicative coefficient, the design effect of $p(\cdot)$. Another metric of interest is $\mathbf{D} = \mathbf{\Phi}^{-1}$. The variance of the Hansen-Hurwitz estimator of $Y$ under the with-replacement sampling design, with parameters $p_k = \pi_k / n$, is equal to:

$$\text{var}_{HH}(\widehat{Y}) = \sum_{k \in U} \pi_k \left( \frac{y_k}{\pi_k} - \frac{Y}{n} \right)^2 = \left\| Y - \check{Y} \right\|^2_{\mathbf{\Phi}^{-1}},$$

and the main result of Chapter 1 is that, if $\mathcal{CP}(\cdot)$ is the conditional Poisson sampling design, $r(\mathcal{CP}, \mathbf{\Phi}^{-1}) \leq 1$.

The value of $d(p, \mathbf{D})$ is also obtained as the 2-norm of the eigenvalues of $\check{\mathbf{M}}$. Instead of dispersions $\delta(p, \mathbf{I})$ and $\delta(p, \mathbf{\Psi})$ that will always be positive, we can define:

$$\tilde{\delta}(p, \mathbf{D}) = \frac{1}{N-1} d(p, \mathbf{D}) - \frac{1}{(N-1)^2} \left( \lambda_2 + \ldots + \lambda_N \right)^2,$$

the dispersion of the eigenvalues of $\check{\mathbf{M}}$. $\tilde{\delta}(p, \mathbf{D})$ can be null, and is linked to $\delta(p, \mathbf{D})$ by:

$$\delta(p, \mathbf{D}) = \left( 1 - \frac{1}{N} \right) \tilde{\delta}(p, \mathbf{D}) + \frac{1}{N(N-1)} \left( \lambda_2 + \ldots + \lambda_N \right)^2.$$

Maximum entropy sampling without replacement with fixed size and equal inclusion probabilities is equal to simple random sampling design. It has a uniform variance operator for all centered variables. All the non-null eigenvalues of its variance operator for the Horvitz-Thompson estimator are equal, that is to say: $\tilde{\delta}(p, \mathbf{I}) = 0$. This last property can not be obtained with an unequal probability sampling with fixed size as stated in proposition 3.4.

**Proposition 3.4** *The following statements hold:*

1. *simple random sampling minimizes the dispersion $\delta(p, \mathbf{I})$ and has a null dispersion $\tilde{\delta}(p, \mathbf{I})$,*
2. *if the inclusion probabilities are not all equal and $N \geq 3$, $\tilde{\delta}(p, \mathbf{I})$ is positive,*
3. *$\tilde{\delta}(p, \mathbf{\Psi})$ can only be null in degenerate cases where $\pi_k$ takes only two different values that sum to one, and $\widetilde{\mathbf{V}}_p$ is the same as in simple random sampling.*

A proof of 3.4 is given in appendix, Section A.3, page 117.

**Remark 3.4** *Simple random sampling also minimizes $r(p, \mathbf{I})$ (see for example Gabler, 1990).*

Proposition 3.4 proves that if the $\pi_k$ are not all equal, no sampling design with fixed size and inclusion probabilities $\pi_k$ has a variance matrix that is proportional to the orthogonal projector on the orthogonal of $\boldsymbol{\pi}$. It also states that $\widetilde{\boldsymbol{V}}_p$ can only be proportional to the orthogonal projector on the orthogonal of the vector $\boldsymbol{\tau} = \boldsymbol{\Psi}^{\frac{1}{2}}\boldsymbol{\pi}$,

$$\boldsymbol{\tau} = \left(\pi_1^{\frac{1}{2}}(1-\pi_1)^{\frac{1}{2}}, \ldots, \pi_N^{\frac{1}{2}}(1-\pi_N)^{\frac{1}{2}}\right)',$$

in very special cases. An unequal probability sampling design will thus never give equally precise Horvitz-Thompson estimates for all the normed variables in the orthogonal of the privileged direction, using the standard metric or the $\boldsymbol{\Psi}$ metric. We may still inquire which sampling designs minimize dispersions $\tilde{\delta}(p, \mathbf{I})$, $\tilde{\delta}(p, \boldsymbol{\Psi})$ and $\tilde{\delta}(p, \boldsymbol{\Phi}^{-1})$, for a given vector of inclusion probabilities, and whether maximum entropy sampling with fixed size is one of them. The answer, as could be expected is negative, as we will see shortly. We will also see that maximum entropy sampling with fixed size and unequal inclusion probabilities does not in general minimize $r(p, \mathbf{I})$, $r(p, \boldsymbol{\Psi})$ or $r(p, \boldsymbol{\Phi}^{-1})$.

The following proposition gives necessary conditions for a sampling design with given inclusion probabilities to be a local extremum of $d(p, \mathbf{D})$, and thus, when $\mathbf{D}$ is diagonal, of $\delta(p, \mathbf{D})$. It is not specific to fixed size sampling designs, but also gives necessary conditions to minimize $\tilde{\delta}(p, \mathbf{D})$ when $\mathbf{D}$ is diagonal and $p(\cdot)$ is a fixed size sampling design.

**Proposition 3.5** *For a sampling design to be a local extremum of $d(p, \mathbf{D})$, it must satisfy the necessary condition that there exists a $(N+1)-$vector $(\lambda_0, \ldots, \lambda_N)'$ solution of the linear system of $P$ equations:*

$$\sum_{k \in s_i} \sum_{\ell \in s_i} \frac{a_{k,\ell}}{\pi_k \pi_\ell} = \lambda_0 + \sum_{k \in s_i} \lambda_k, \tag{3.1}$$

*for all $s_i$, $i = 1, \ldots, P$, where $a_{k,\ell}$ are coefficients of the matrix $\mathbf{D}^{-\frac{1}{2}}\mathbf{M}\mathbf{D}^{-\frac{1}{2}}$.*

A proof of 3.5 is given in appendix, Section A.4.1, page 121.

**Remark 3.5** *The condition in Proposition 3.5 can be stated shortly as*

$$\mathrm{Diag}(\mathbf{S}_\mathrm{p}' \boldsymbol{\Omega} \mathbf{S}_\mathrm{p}) \in \mathrm{Im}(\widetilde{\mathbf{S}}_\mathrm{p}'),$$

*where $\boldsymbol{\Omega} = \boldsymbol{\Phi}^{-1}\mathbf{D}^{-1}\mathbf{V}_p\mathbf{D}^{-1}\boldsymbol{\Phi}^{-1}$, $\mathrm{Diag}(\cdot)$ is the operator that returns as a vector the diagonal of a matrix, and $\widetilde{\mathbf{S}}_p$ is the matrix $\mathbf{S}_p$ augmented with a line of 1.*

We give some direct and more explicit applications of Proposition 3.5 in Corollary 3.1.

**Corollary 3.1**

1. *For a sampling design with given inclusion probabilities to be a local extremum of the dispersion $\delta(p, \mathbf{I})$, it must satisfy the necessary condition: there exists a $(N+1)-$vector $(\lambda_0, \ldots, \lambda_N)'$ such that*

$$\sum_{k \in s_i} \sum_{\ell \in s_i} \left(\frac{\pi_{k,\ell}}{\pi_k \pi_\ell} - 1\right) \frac{1}{\pi_k \pi_\ell} = \lambda_0 + \sum_{k \in s_i} \lambda_k,$$

*for all* $s_i$, $i = 1, \ldots, P$.

2. *For a sampling design to be a local extremum of the dispersion* $\delta(p, \mathbf{\Psi})$ *under the same conditions, it must satisfy the necessary condition: there exists a* $(N+1)-$*vector* $(\tilde{\lambda}_0, \ldots, \tilde{\lambda}_N)'$ *such that*

$$\sum_{k \in s_i} \sum_{\ell \in s_i} \left( \frac{\pi_{k,\ell} - \pi_k \pi_\ell}{\pi_k (1 - \pi_k) \pi_l (1 - \pi_l)} \right) = \tilde{\lambda}_0 + \sum_{k \in s_i} \tilde{\lambda}_k,$$

*for all* $s_i$, $i = 1, \ldots, P$.

3. *For a sampling design to be a local extremum of the dispersion* $\delta(p, \mathbf{\Phi}^{-1})$ *under the same conditions, it must satisfy the necessary condition: there exists a* $(N+1)-$*vector* $(\lambda_0, \ldots, \lambda_N)'$ *such that*

$$\sum_{k \in s_i} \sum_{\ell \in s_i} \Delta_{k,\ell} = \tilde{\lambda}_0 + \sum_{k \in s_i} \tilde{\lambda}_k,$$

*for all* $\mathbf{s}_i$, $i = 1, \ldots, P$.

4. *Let us consider a metric* $m_f$ *defined by*

$$\|Y\|^2_{m_f} = \sum_{k \in U} \left( \frac{y_k}{f(\pi_k)} \right)^2,$$

*where* $f(\cdot)$ *is a positive function on* $[0, 1]$, *and note* $\mathbf{D}_f$ *the associated diagonal matrix. We get that a sampling design that is a local extremum of* $\delta(p, \mathbf{D}_f)$ *must satisfy the necessary condition: there exists a* $(N+1)-$*vector* $(\lambda_0, \ldots, \lambda_N)'$ *such that*

$$\sum_{k \in s_i} \sum_{\ell \in s_i} \frac{\Delta_{k,\ell}}{\pi_k \pi_\ell} \frac{f^2(\pi_k)}{\pi_k} \frac{f^2(\pi_\ell)}{\pi_\ell} = \lambda_0 + \sum_{k \in s_i} \lambda_k,$$

*for all* $s_i$, $i = 1, \ldots, P$.

Finally, we can use statements 1, 2 and 3 of Corollary 3.1 to observe on numeric examples that maximum entropy sampling with fixed size is not a local extremum of $d(p, \mathbf{D})$ or $\tilde{\delta}(p, \mathbf{D})$, $\mathbf{D} = I, \mathbf{\Psi}, \mathbf{\Phi}^{-1}$. But we get in return that minimum support designs are natural candidates to be local extrema.

**Corollary 3.2**

1. *Maximum entropy sampling with fixed size usually does not satisfy the preceding conditions and thus is not always a minimum of* $\tilde{\delta}(p, \mathbf{D})$, $\mathbf{D} = I, \mathbf{\Psi}, \mathbf{\Phi}^{-1}$.

2. *Maximum entropy sampling with fixed size is also not always a minimum of* $r(p, \mathbf{D})$, $\mathbf{D} = I, \mathbf{\Psi}, \mathbf{\Phi}^{-1}$.

3. *A minimum support design, such as systematic sampling, always satisfies the preceding conditions and thus may be a local extremum of* $d(p, \mathbf{D})$, $\delta(p, \mathbf{D})$, *or* $\tilde{\delta}(p, \mathbf{D})$ *for any metric* $\mathbf{D}$.

A proof of 3.2 is given in appendix, Section A.4.2, page 122.

## 3.3 Conclusion

These negative results show that good properties of Poisson sampling do not extend to fixed size sampling with maximum entropy. They also show that it is not possible to have an equal treatment of variables orthogonal to $\boldsymbol{\pi}$ using the most natural metrics. Poisson sampling design minimizes $r(p, \mathbf{D})$ and $d(p, \mathbf{D})$ for all diagonal metrics $\mathbf{D}$ thanks to the fact that it is a design for which indicator variables $I_k(\cdot)$, $k = 1, \ldots, N$, are not pairwise correlated. The preceding results on conditional Poisson sampling not minimizing $d(p, \mathbf{D})$ with $\mathbf{D} = \mathbf{I}$, $\boldsymbol{\Psi}$, or $\boldsymbol{\phi}^{-1}$ can also be interpreted from this point of view. They can be translated to: conditional Poisson sampling is not a sampling design that minimizes

$$\sum_{k \in U} \sum_{\ell \neq k} w_k w_\ell \Delta_{k,\ell}^2,$$

with weights $w_k$, $k \in U$ equal to $\pi_k^{-2}$, $[\pi_k(1 - \pi_k)]^{-1}$ or $\pi_k^{-1}$.

We gave in Corollary 3.1 necessary conditions for a sampling design to minimize one of these quantities. Unlike in Stenger & Gabler (1996), we do not compute the second order inclusion probabilities of that sampling design, but we know that a sampling design exists that satisfies these conditions and minimizes $d(p, \mathbf{D})$, with $\mathbf{D} = \mathbf{I}$, $\boldsymbol{\Psi}$, or $\boldsymbol{\phi}^{-1}$.

Conditional Poisson sampling is also not a sampling design that minimizes the maximum risk $r(p, \mathbf{D})$ for this choice of metrics $\mathbf{D}$, and using the Horvitz-Thompson estimator. With a diagonal metric, and the parameter space of centered variables, Stenger & Gabler (1996) found out that the Lahiri-Midzuno-Sen sampling design (see Midzuno, 1950; Brewer & Hanif, 1983) could be used to obtain a minimax strategy. This sampling algorithm, however, requires severe conditions on the inclusion probabilities. The problem of finding a sampling design that gives a minimum risk for any set of inclusion probabilities is still unsolved.

# Part II

# Repeated survey sampling

# 4

# Variance estimation of changes in repeated surveys and its application to the Swiss survey of value added

**Abstract**

We propose a method for estimating the variance of estimators of changes over time, a method that takes account of all the components of these estimators: the sampling design, treatment of non-response, treatment of large companies, correlation of non-response from one wave to another, the effect of using a panel, robustification, and calibration using a ratio estimator. This method, which serves to determine the confidence intervals of changes over time, is then applied to the Swiss survey of value added.[1]

**Keywords:** Covariance, Stratified sampling, Panel

## Introduction

In longitudinal surveys, the precision of changes over time depends directly on the rate of overlap of the samples. We begin by reviewing known results for disjoint simple designs (on this subject, see Kish, 1965; Sen, 1973; Wolter, 1985; Laniel, 1988; Hidiroglou et al., 1995; Holmes & Skinner, 2000; Nordberg, 2000; Fuller & Rao, 2001; Berger, 2004b). Next, we calculate the variance of such changes for simple designs in which the samples overlap. When the sampling ratios are very low, most of these results are well known and are described, for example, in Caron & Ravalet (2000). Results that take account of finite population corrections can be seen in Tam (1984).

[1] This chapter is a reprint of: Qualité, L. & Tillé, Y. (2008). Variance estimation of changes in repeated surveys and its application to the swiss survey of value added. *Survey Methodology* **34**, 173–181.

We precisely calculated the variances of estimators for a larger class of sampling designs with a finite population. Finite population corrections can play a major role in business surveys, since large companies are sometimes selected with very high probabilities of inclusion. The calculations become much more complicated with a finite population for the following reason: if the size of the population is finite, two disjoint samples are not independent. If the population is infinite, two independent samples are disjoint. Several estimators are examined: the difference of the cross-sectional estimators; the difference estimated solely on the common portion; and relative changes. The calculations become even more complex when the population is dynamic (with births, deaths, changes of structure). The theory that we develop below is limited to the case in which the population does not change over time.

In the first part, we describe the two-dimensional simple random sampling design (on this subject, see Goga, 2003) and we give the corresponding Horvitz-Thompson estimators. We calculate the variance of the estimator of changes that is based on this sampling design. In a second part, we give the variance of other simple estimators: the relative change or the totals quotient, and the difference estimator based on the overlap of the samples. We then describe how these results adapt to the presence of ignorable non-response and the use of more complex estimators, which introduce weights modified to obtain calibrated estimators, or variables modified by a robustification procedure.

These results for simple designs are easy to generalize to stratified designs, provided that companies do not change strata from one wave to the next. Lastly, we apply this method to the Swiss survey of value added, taking all components of the survey into account: stratification, the panel effect, non-response, correlation between non- responses from one wave to the next, calibration using a ratio estimator, and robustification.

## 4.1 Estimation of the difference in simple designs

Let there be a population $U = \{1, ..., k, ..., N\}$ of size $N$ in which two samples are taken: $s_1$ and $s_2$ of respective sizes $n_1$ and $n_2$. These samples may have a common portion (see Figure 4.1).

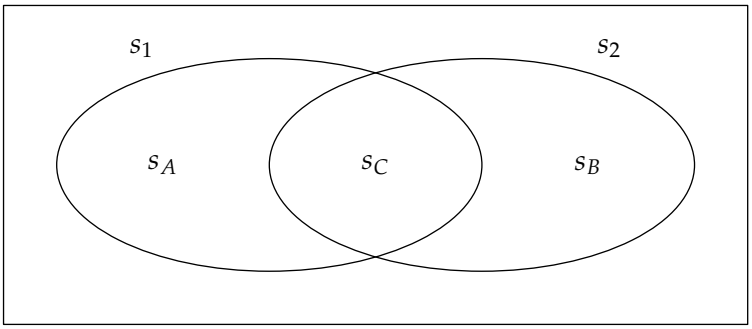



Figure 4.1 – *Overlapping samples*

Assume that $s_1$ and $s_2$ are samples taken according to a simple design

without replacement, and sizes $n_1$ et $n_2$ are therefore not random. Samples $s_1$ and $s_2$ can be broken down into three parts $s_A = s_1 \backslash s_2$, $s_B = s_2 \backslash s_1$, and $s_C = s_1 \cap s_2$. Let $n_A = |s_A|$, $n_B = |s_B|$, $n_C = |s_C|$, $n_1 = n_A + n_C$, $n_2 = n_B + n_C$. The sizes of $s_A$, $s_B$, et $s_C$, may be random. This design generalizes the following hypothetical cases:

- If samples $s_1$ and $s_2$ are selected independently, $n_C$ is then a random variable;
- If sample $s_1$ is selected first, and sample $s_2$ is selected in the complement of $s_1$ in $U$, then $s_C$ is empty and $n_C = 0$;
- If sample $s_1$ is selected first, and sample $s_2$ consists of the union of a subsample of fixed size of $s_1$ and a sample of fixed size of the complement of $s_1$ in $U$, then $n_C$ is not random, and the situation is the same as in case A of Tam (1984).

We make the additional hypothesis that conditional on $n_A$, $n_B$, and $n_C$, samples $s_A$, $s_B$, and $s_C$, are simple, without replacement and of fixed size. They come from the following sampling design:

**Definition 4.1** *Two-dimensional simple fixed-size sampling design* $(n_A, n_B, n_C)$*:*

$$p_{simpl}(s_1, s_2 | n_A, n_B, n_C) = \begin{cases} \frac{n_A! n_B! n_C! (N - n_A - n_B - n_C)!}{N!} & si\ n_A = |s_A|, \\ & n_B = |s_B|, n_C = |s_C| \\ 0 & otherwise, \end{cases}$$

*where* $s_A = s_1 \backslash s_2$, $s_B = s_2 \backslash s_1$ *and* $s_C = s_1 \cap s_2$ *(on this subject, see Goga, 2003).*

The law for drawing the pair $(s_1, s_2)$, which we do not know in general, is thus assumed to be of the form

$$p(s_1, s_2) = p_{simpl}(s_1, s_2 | n_A, n_B, n_C) \Pr(|s_1 \cap s_2| = n_C).$$

Let there be two variables $x$ and $y$ whose values, taken on the units of $U$ are denoted respectively $x_k$ and $y_k$, $k \in U$. Variables $x$ and $y$ may represent the same variable measured at two different times. Also assume that $x$ can be observed only for $s_1$ and $y$ for $s_2$. The objective is to estimate the totals

$$X = \sum_{k \in U} x_k \text{ and } Y = \sum_{k \in U} y_k,$$

as well as the difference $Y - X$. The Horvitz-Thompson estimators of $X$ and $Y$ are given by

$$\widehat{X}_1 = \frac{N}{n_1} \sum_{k \in s_1} x_k \text{ and } \widehat{Y}_2 = \frac{N}{n_2} \sum_{k \in s_2} y_k.$$

### 4.1.1 Natural estimation of the difference

#### Variance of the estimation of the difference

A first approach for estimating $\Delta = Y - X$ is to use the difference of the cross-sectional estimators $\widehat{\Delta} = \widehat{Y}_2 - \widehat{X}_1$ which is an unbiased estimator conditional on $n_C$ according to the following simple design:

$$\mathrm{E}(\widehat{\Delta}|n_C) = Y - X,$$

and is therefore also unbiased under design $p$ unconditional on $n_C$.

**Proposition 4.1** *The variance of $\widehat{\Delta}$ is:*

$$\begin{aligned}\mathrm{var}(\widehat{\Delta}) = N^2\left(\frac{1}{n_1}-\frac{1}{N}\right)S_x^2 + N^2\left(\frac{1}{n_2}-\frac{1}{N}\right)S_y^2 \\ -2N^2\left(\frac{E(n_C)}{n_1 n_2}-\frac{1}{N}\right)S_{xy},\end{aligned} \tag{4.1}$$

*where*

$$S_x^2 = \frac{1}{N-1}\sum_{k\in U}(x_k-\overline{X})^2, \quad S_y^2 = \frac{1}{N-1}\sum_{k\in U}(y_k-\overline{Y})^2,$$

$$S_{xy} = \frac{1}{N-1}\sum_{k\in U}(x_k-\overline{X})(y_k-\overline{Y}).$$

The demonstration of (4.1) is appended.

#### Specific cases and precision gain

Result (4.1) can be used to deal directly with the following specific cases of co-ordination:

- if the two samples form a panel, $n_C = n_1 = n_2$, then

$$\mathrm{var}(\widehat{\Delta}) = N^2\left(\frac{1}{n_C}-\frac{1}{N}\right)(S_x^2 + S_y^2 - 2S_{xy}).$$

- if the samples are disjoint (also see Ardilly & Tillé, 2003, pages 24-28), $n_C = 0$, and

$$\mathrm{var}(\widehat{\Delta}) = N^2\left(\frac{1}{n_1}-\frac{1}{N}\right)S_x^2 + N^2\left(\frac{1}{n_2}-\frac{1}{N}\right)S_y^2 + 2NS_{xy}.$$

  Surprisingly, the covariance does not depend on the sizes of the samples. It is negative if $x$ and $y$ are positively correlated, and it becomes negligible in relation to the variance terms when the size of the population is large;

- if $q$ is the set rate of overlap of the two samples and $n_1 = n_2 = n$, we are back to case A developed by Tam (1984). We then obtain $n_C = qn$, and

$$\mathrm{var}(\widehat{\Delta}) = N^2\left(\frac{1}{n}-\frac{1}{N}\right)(S_x^2 + S_y^2) - 2N^2\left(\frac{q}{n}-\frac{1}{N}\right)S_{xy}.$$

- if the two samples are independent, $\mathrm{E}(n_C) = n_1 n_2 / N$, and we have

$$\mathrm{var}_{IND}(\widehat{\Delta}) = N^2 \left(\frac{1}{n_1} - \frac{1}{N}\right) S_x^2 + N^2 \left(\frac{1}{n_2} - \frac{1}{N}\right) S_y^2.$$

If the size of the population is large and if the variables $x$ and $y$ have dispersions that are close to one another, the gain (or loss) of precision due to co-ordination in relation to the selection of two samples independently is

$$G = \frac{\mathrm{var}\left(\widehat{\Delta}\right)}{\mathrm{var}_{IND}\left(\widehat{\Delta}\right)} \approx 1 - \rho q, \tag{4.2}$$

where $\rho$ is the coefficient of correlation between $x$ and $y$, $\rho = S_{xy}/S_x S_y$ and $q$ is the overlap rate, $q = 2\mathrm{E}(n_C)/(n_1 + n_2)$. Expression (4.2) provides a simple multiplicative coefficient serving to take account of the effect of correlation and overlap.

**Estimation of the variance of $\widehat{\Delta}$**

To estimate the variance, two cases must be considered:

- if $\mathrm{E}(n_C)$ is known, which may be the case (for example, when the two samples are known to be independent), then

$$\widehat{\mathrm{var}}(\widehat{\Delta}) = N^2 \left(\frac{1}{n_1} - \frac{1}{N}\right) s_{x1}^2 + N^2 \left(\frac{1}{n_2} - \frac{1}{N}\right) s_{y2}^2 - 2N^2 \left(\frac{\mathrm{E}(n_C)}{n_1 n_2} - \frac{1}{N}\right) s_{xyC}. \tag{4.3}$$

  where

$$s_{x1}^2 = \frac{1}{n_1 - 1} \sum_{s_1} (x_k - \bar{x}_1)^2, \quad s_{y2}^2 = \frac{1}{n_2 - 1} \sum_{s_2} (y_k - \bar{y}_2)^2,$$

  and

$$s_{xyC} = \frac{1}{n_C - 1} \sum_{s_C} (x_k - \bar{x}_C)(y_k - \bar{y}_C).$$

  This estimator is unbiased, but it can sometimes take on negative values;

- if $\mathrm{E}(n_C)$ is not known, the only information concerning co-ordination is $n_C$.

$$\widehat{\mathrm{var}}(\widehat{\Delta}) = N^2 \left(\frac{1}{n_1} - \frac{1}{N}\right) s_{x1}^2 + N^2 \left(\frac{1}{n_2} - \frac{1}{N}\right) s_{y2}^2 - 2N^2 \left(\frac{n_C}{n_1 n_2} - \frac{1}{N}\right) s_{xyC}, \tag{4.4}$$

  This estimator is unbiased conditional on $n_C$ and is therefore also unconditionally unbiased. It can also sometimes take on negative values. We will see further on that in some applications involving non- response, $\mathrm{E}(n_C)$ is not known.

To use estimator (4.3), it is necessary to have at least two units in the overlap of the samples ($n_C \geq 2$), unless $\mathrm{E}(n_C) = n_1 n_2/N$. If $\mathrm{E}(n_C) = n_1 n_2/N$, which is the case when the two samples are independent, the third term of estimator (4.3) is nil. As to estimator (4.4) it is not defined when $n_C = 1$, unless $n_1 n_2 = N$.

### 4.1.2 Estimation using the common portion

The difference can also be estimated using only the common portion of the sample, which yields the estimator

$$\widehat{\Delta}_C = N(\overline{y}_C - \overline{x}_C),$$

with $\overline{y}_C = \frac{1}{n_C}\sum_{k\in s_C} y_k$ and $\overline{x}_C = \frac{1}{n_C}\sum_{k\in s_C} x_k$. This estimator is unbiased unconditionally and conditionally on $n_C$.

#### Estimation of the variance of $\widehat{\Delta}_C$

The conditional variance of $\widehat{\Delta}_C$ is equal to

$$\mathrm{var}(\widehat{\Delta}_C|n_C) = N^2\left(\frac{1}{n_C} - \frac{1}{N}\right)(S_y^2 + S_x^2 - 2S_{xy}).$$

The unconditional variance is equal to

$$\mathrm{var}(\widehat{\Delta}_C) = N^2\left[E\left(\frac{1}{n_C}\right) - \frac{1}{N}\right](S_y^2 + S_x^2 - 2S_{xy}).$$

This unconditional variance may be difficult to calculate when $n_C$ is random.

#### Comparison of the variances of $\widehat{\Delta}$ and $\widehat{\Delta}_C$

If we want to compare the two estimators of the difference, we can calculate

$$\begin{aligned}\mathrm{var}(\widehat{\Delta}) - \mathrm{var}(\widehat{\Delta}_C) = N^2\left[\frac{1}{n_1} - \mathrm{E}\left(\frac{1}{n_C}\right)\right]S_y^2 + N^2\left[\frac{1}{n_2} - \mathrm{E}\left(\frac{1}{n_C}\right)\right]S_x^2 \\ -2N^2\left[\frac{\mathrm{E}(n_C)}{n_1 n_2} - \mathrm{E}\left(\frac{1}{n_C}\right)\right]S_{xy}.\end{aligned}$$

Si $n_1 = n_2 = n$, $S_x^2 = S_y^2 = S^2$, et $\mathrm{E}(1/n_C) \approx 1/\mathrm{E}(n_C)$, then we obtain

$$\begin{aligned}\mathrm{var}(\widehat{\Delta}) - \mathrm{var}(\widehat{\Delta}_C) &\approx \frac{1}{qn}\left[q-1\right]2N^2S^2 - 2\frac{1}{qn}\left[q^2-1\right]\rho N^2S^2 \\ &= \frac{2N^2S^2}{qn}(1-q)\left[\rho(1+q)-1\right],\end{aligned}$$

where $q = 2\mathrm{E}(n_C)/(n_1+n_2)$ is the overlap rate. The estimator $\widehat{\Delta}_C$ is therefore more precise than $\widehat{\Delta}$ if

$$\rho \geq \frac{1}{1+q}.$$

For example, if $q = 0.7$,, it is preferable to use only the common portion once $\rho \geq 1/(1+0.7) \approx 0.588$ (on this subject, see Caron & Ravalet, 2000, p. 346). In cases where the overlap is sizable and the correlation is high, the estimator based on the difference of the cross-sectional estimators is therefore not very relevant.

## 4.2 Taking unit non-response into account

Non-response is considered to be independent of the selection design. According to the model, each unit decides randomly whether or not to respond, and the probabilities of response are equal between units. This is the most elementary model. However, if a unit does not respond in the first wave, it is highly probable that it will also not respond in the second wave. The model takes this dependency into account by considering separately four cases:

- the unit responds in both the first wave and the second;
- the unit responds in the first wave but not in the second;
- the unit does not respond in the first wave but it responds in the second;
- the unit responds in neither the first wave nor the second.

Non-response is commonly modelled by a multivariate Bernoullian design, which means that the probability of responding is the same for all statistical units and also that one unit decides to respond independently of the response of the other units. The non-response design is as follows:

$$q(r_A, r_B, r_C, r_D) = \phi_A^{\text{card } r_A} \phi_B^{\text{card } r_B} \phi_C^{\text{card } r_C} \phi_D^{\text{card } r_D},$$

where $r_A, r_B, r_C, r_D \subset U$, and $r_A, r_B, r_C, r_D$ are mutually exclusive, and where

- $\phi_A$ is the probability of responding in wave 1, but not in wave 2,
- $\phi_B$ is the probability of responding in wave 2, but not in wave 1,
- $\phi_C$ is the probability of responding in both wave 1, and wave 2,
- $\phi_D$ is the probability of responding in neither wave 1, nor wave 2.

The modelled non-response phase thus consists in selecting four disjoint samples according to Bernoullian designs with different intensities. Since it is assumed to be independent of the sampling design, conditional on the sample sizes observed, the design resulting from the selection and the non-response is a simple multivariate design. If inference is conducted conditional on the sample sizes, the estimation of probabilities $\phi_A$, $\phi_B$, $\phi_C$, $\phi_D$ is not necessary and an unbiased inference can be conducted, as if dealing with a simple design. The theory of the preceding section therefore applies directly to the respondents, and all the information on the overlap

of the two samples is found in $|s_C|$, regardless of whether this overlap is due to the design or to the link that exists between non-responses in the two waves. Note that even if the model is fairly simple, it takes account of the fact that if a unit has not responded in one wave, it will probably be less likely to respond in the following wave. Also, this model will be applied in relatively small, homogeneous strata.

## 4.3 Other measures of changes over time

The measurement of change over time is not always expressed in terms of differences. Such change is often measured in the form of a quotient or a relative difference. We therefore consider the following three measures:

- the difference $\widehat{\Delta} = \widehat{Y}_2 - \widehat{X}_1$,
- the relative change $\widehat{\Delta}_R = (\widehat{Y}_2 - \widehat{X}_1)/\widehat{X}_1 = \widehat{Y}_2/\widehat{X}_1 - 1$,
- the quotient $\widehat{Q} = \widehat{Y}_2/\widehat{X}_1$.

The variance of $\widehat{\Delta}$ may be expressed simply as a function of the estimators of variance of $\widehat{Y}_2$ and $\widehat{X}_1$ and the estimator of their covariance (see expression 4.4). The variance of $\widehat{\Delta}_R$ is equal to the variance of $\widehat{Q}$. They may be approached and then estimated using a residuals technique (on this subject, see Woodruff, 1971; Binder & Patak, 1994; Deville & Särndal, 1992; Deville, 1999),

$$\widehat{\text{var}}(\widehat{\Delta}_R) = \widehat{\text{var}}(\widehat{Q}) = \frac{1}{\widehat{X}_1^2}\left[\widehat{\text{var}}(\widehat{Y}_2) + \widehat{Q}^2\widehat{\text{var}}(\widehat{X}_1) - 2\widehat{Q}\widehat{\text{cov}}(\widehat{X}_1, \widehat{Y}_2)\right].$$

This variance can thus be simply estimated once we have estimators of $\text{var}(\widehat{Y}_2)$, $\text{var}(\widehat{X}_1)$ and $\text{cov}(\widehat{X}_1, \widehat{Y}_2)$.

## 4.4 Ratio estimation and robustification

Two techniques are commonly used for estimating the results of sample surveys: the use of a ratio estimator to calibrate on the total of a dummy (auxiliary) variable, and robustification of the estimators. These techniques must be taken into account in determining the precision of the final results.

### 4.4.1 Calibration

If an estimator is calibrated on known totals, the variance may be estimated simply by a residuals technique (see Woodruff, 1971; Binder & Patak, 1994; Deville & Särndal, 1992; Deville, 1999). For example, if $\mathbf{z}_{k1}$ and $\mathbf{z}_{k2}$ are column vectors of dummy variables on which the estimators $\widehat{X}_{1Cal}$ and $\widehat{Y}_{2Cal}$ are calibrated in waves 1 and 2, then the variances can be estimated by a residuals technique: $\text{var}(\widehat{X}_{1Cal}) \approx \text{var}(\widehat{E}_1)$ and

$\text{var}(\widehat{Y}_{2Cal}) \approx \text{var}(\widehat{E}_2)$, where $\widehat{E}_1$ and $\widehat{E}_2$ are Horvitz-Thompson estimators of the totals of the residuals, with the latter being given for a simple design and for the generalized regression estimator by:

$$e_{k1} = x_k - \mathbf{z}'_{k1}\widehat{\mathbf{B}}_1,$$
$$e_{k2} = y_k - \mathbf{z}'_{k2}\widehat{\mathbf{B}}_2,$$

with

$$\widehat{\mathbf{B}}_1 = \left(\sum_{k\in s_1} q_{k1}\mathbf{z}_{k1}\mathbf{z}'_{k1}\right)^{-1} \sum_{k\in s_1} q_{k1}\mathbf{z}_{k1}x_{k1},$$
$$\widehat{\mathbf{B}}_2 = \left(\sum_{k\in s_2} q_{k2}\mathbf{z}_{k2}\mathbf{z}'_{k2}\right)^{-1} \sum_{k\in s_2} q_{k2}\mathbf{z}_{k2}y_{k2},$$

where $q_{kj}$, $j = 1,2$, is a coefficient that serves to take account of possible heteroscedasticity.

In the case of a sampling design with unequal probabilities, e.g., a stratified sampling design such as in the Swiss survey of value added, the residuals are obtained by using a weighted regression. It is sufficient to replace $\widehat{\mathbf{B}}_1$ and $\widehat{\mathbf{B}}_2$ respectively by

$$\widehat{\mathbf{B}}_1 = \left(\sum_{k\in s_1} \frac{q_{k1}\mathbf{z}_{k1}\mathbf{z}'_{k1}}{\pi_{k1}}\right)^{-1} \sum_{k\in s_1} \frac{q_{k1}\mathbf{z}_{k1}x_{k1}}{\pi_{k1}}, \text{ and} \tag{4.5}$$
$$\widehat{\mathbf{B}}_2 = \left(\sum_{k\in s_2} \frac{q_{k2}\mathbf{z}_{k2}\mathbf{z}'_{k2}}{\pi_{k2}}\right)^{-1} \sum_{k\in s_2} \frac{q_{k2}\mathbf{z}_{k2}y_{k2}}{\pi_{k2}}, \tag{4.6}$$

where $\pi_{kj}$ is the probability of inclusion of unit $k$ in the sample for wave $j$, $j = 1,2$.

### 4.4.2 Robustification

It is often useful to apply a robustification technique which offers a way to treat outliers. Simply consider that outliers have been detected and the weights of the individuals whose values are considered outliers have been modified by a factor $u_{kj}(s)$ in wave $j$. This factor is included between 0 and 1 and is equal to 1 for units that have values considered normal. The variance of the robustified estimator can be approached by advancing the classical hypothesis that weights $u_{kj}(s)$ depend only slightly on the sample $s$ that was drawn (see Hulliger, 1999). All that is needed, then, is to replace the variables $x_k$ and $y_k$ observed by $u_{k1}x_k$ and $u_{k2}y_k$ in the variance estimators.

By bringing together all the components of the mean square error of a change over time so as to take account of all components of that variance - namely the design, the panel effect, non-response, calibration and robustification - we obtain, for the relative change in a stratum,

$$\widehat{\text{EQM}}(\widehat{\Delta}_R) = \widehat{\text{EQM}}(\widehat{Q}) = \frac{1}{\widehat{X}_1}\left[\widehat{\text{var}}(\widehat{EU}_1) + \widehat{Q}^2\widehat{\text{var}}(\widehat{EU}_2) - 2\widehat{Q}\widehat{\text{cov}}(\widehat{EU}_1, \widehat{EU}_2)\right], \tag{4.7}$$

where

$$\widehat{X}_1 = \frac{N}{m_1}\sum_{R_1} x_k, \qquad \widehat{Y}_2 = \frac{N}{m_2}\sum_{R_2} y_k, \qquad \widehat{Q} = \frac{\widehat{Y}_2}{\widehat{X}_1},$$

$$eu_{k1} = u_{k1}x_k - u_{k1}\mathbf{z}'_{k1}\widehat{\mathbf{B}}_1,$$
$$eu_{k2} = u_{k2}y_k - u_{k2}\mathbf{z}'_{k2}\widehat{\mathbf{B}}_2,$$

$$\widehat{EU}_j = \frac{N}{m_j}\sum_{R_j} eu_{kj}, \qquad \overline{EU}_j = \frac{\widehat{EU}_j}{N}, \qquad j = 1,2,$$

$$\widehat{\mathbf{B}}_1 = \left(\sum_{k\in D_1} \frac{q_{k1}u_{k1}^2\mathbf{z}_{k1}\mathbf{z}'_{k1}}{\pi_{k1}}\right)^{-1} \sum_{k\in D_1} \frac{q_{k1}u_{k1}^2\mathbf{z}_{k1}x_k}{\pi_{k1}},$$
$$\widehat{\mathbf{B}}_2 = \left(\sum_{k\in D_2} \frac{q_{k2}u_{k2}^2\mathbf{z}_{k2}\mathbf{z}'_{k2}}{\pi_{k2}}\right)^{-1} \sum_{k\in D_2} \frac{q_{k2}u_{k2}^2\mathbf{z}_{k2}y_k}{\pi_{k2}}.$$

$$\widehat{\text{var}}(\widehat{EU}_j) = N^2\left(\frac{1}{m_j} - \frac{1}{N}\right)\frac{1}{m_j - 1}\sum_{R_j}(eu_{kj} - \overline{EU}_j)^2, j = 1,2,$$

$$\widehat{\text{cov}}(\widehat{EU}_1, \widehat{EU}_2) = N^2\left(\frac{m_C}{m_1 m_2} - \frac{1}{N}\right)\frac{1}{m_C - 1}\sum_{R_C}(eu_{k1} - \overline{EU}_1)(eu_{k2} - \overline{EU}_2).$$

$R_1$ and $R_2$ designate the set of respondents in the first and the second waves in the stratum, $m_1 = |R_1|$, $m_2 = |R_2|$, $R_C = R_1 \cap R_2$ and $m_C = |R_1 \cap R_2|$. $D_1$ and $D_2$ are the sets of respondents in the two waves in the domain in which the calibration was carried out.

## 4.5 The Swiss survey of value added

### 4.5.1 Description of survey

The Swiss survey of value added is a survey of companies, conducted annually. Its purpose is to provide estimators of the main parameters of output in Switzerland: the value of gross output, the amount of intermediate consumption, the value added created by companies, and the cost of labour. The sampling design used is a stratified sampling of companies. In 1999, a sample of 11,210 companies (employing at least two persons) was selected and surveyed. This sample was run again in 2000 and 2001. Over that period, then, this is a panel survey. In the absence of a business register making it possible to identify births and deaths, the population of companies was considered constant during this period. The only adjustment to the annual data is made using a ratio estimation on the total of full-time equivalents (FTEs) per activity domain, available from an external source.

Stratification is based on the first two digits of the Nomenclature Générale des Activités économiques (general classification of economic activities) (NOGA2) and the size of the company (see Renfer, 2000). Dans

In each activity stratum, three size strata are created: small companies employing 2- 19 persons in FTE, medium-size companies, from 20 to $M$ FTE, and large companies of more than $M$ FTE. The stratum containing large companies is a take-all stratum, while small and medium-size companies are selected randomly with different sampling rates. The boundary $M$ is chosen differently in each activity stratum in order to obtain optimum precision. In these three waves, approximately 6,000 establishments responded. The response rate for large companies, which all had to be surveyed, was close to 71% and was higher than the rate for small and medium-size companies. It was decided after the fact to treat some very large companies separately according to the "surprise" stratum methodology of Hidiroglou & Srinath (1981), Considering that the response rate for the largest companies may well be better because they have an administrative structure better suited to responding to the survey questions. If they were assigned a weight equal to that of other large companies, this would introduce a bias as well as excessive variability. The "surprise" poststrata contain the 5% largest companies in the survey file. The latter were then considered as having, in effect, all been surveyed, and they received a weight of 1. No other treatment (calibration, robustification) was applied to them. The take-some strata consisting of small, medium-size and large companies were updated and some strata (size classes) containing few companies were later collapsed. If we accept the hypothesis that the very large companies were all taken, then the resulting estimator is unbiased and the variance related to very large companies is nil. We can therefore calculate only the variance in the other, updated strata.

During the survey, companies were again asked their category of economic activity. The estimates are based on these reported NOGA2s not on the NOGA2s in the sample frame. A calibration on the number of full-time equivalents (FTEs) provided by the business register is then conducted using a quotient estimator for the "reported" NOGA2 domains.

Finally, a robustification technique was used to lop the distribution of certain variables in the sample of small, medium-size and large companies (see Hulliger, 1999; Peters et al., 2001). The weights of establishments whose values are considered outliers were modified by a factor $u_{kj}(s)$ included between 0 and 1. This factor is equal to 1 for companies that have values that are considered normal.

### 4.5.2 Variance of the change in value added

The objective is to estimate correctly the variance of estimators of change in value added (see Renfer, 2000; Peters et al., 2001). In computing variances according to the hypothesis of independence of the samples, we largely overestimate the variance of changes, because the "value added" variables in times $t_1$ and $t_2$ are positively correlated. Correctly taking account of all aspects of the sampling design and the adjustment should provide better variance estimates. The study focuses on the 1999, 2000 and 2001 waves of the survey. Between these three dates, the raw sample was not modified. The fact that the sample remained fixed should make it

possible to reliably estimate changes, but a response rate hovering around 50% may cause us to lose the benefit of the panel, if the number of respondents common to successive waves is low. The case of change between two survey waves where the sample has been updated, and where there are therefore two different raw samples and reference populations, is an entirely different problem.

In the present case, the fact that low variances were obtained can be attributed to the combined effect of several factors:

1. *Optimal design*: The sampling design was optimized. According to the optimal stratification, large companies have higher probabilities of inclusion. The stratum of companies contributing the most to value added is a take-all stratum. For this reason, the cross-sectional estimators have a low variance.

2. *High response fraction*: In the take-all stratum of large companies, the response rate approaches 70%. The finite population correction $(N - n)/N$ can therefore divide the variance by 3 compared to the case of an infinite population.

3. *Panel effect*: The sample is a panel, which is the best strategy for estimating changes over time.

4. *Correlation of non-response*: The non-response in one wave is strongly related to the previous wave and therefore does not greatly degrade the panel.

5. *Correlation of variables between waves*: The value added variables in times $t$ and $t + 1$ are highly correlated, since they are the same variable estimated at two different points in time.

6. *Calibration*: The estimators are calibrated in the strata on a variable related to the variable of interest; the variance of the estimators can then be written as a residual variance.

Of the 11,210 companies selected in 1999, approximately 5,200 responded in 1999 and 2000, and 5,300 responded in the 2000 and 2001 waves. Thus the size of the panel is relatively modest, and the treatment of non-response will therefore have a major impact on the results. To make variance estimates, we have assumed that non- response is ignorable (missing completely at random) within the take-some strata.

In each wave, estimates are made in the reported NOGA2 domains. This implies the possibility of a change of domain on the part of companies, and it is necessary to try to factor this into longitudinal estimates. We decided to ignore the impact of these changes initially, and to consider for the estimation of covariance that the domains are fixed and given by the value reported in the first of the two consecutive waves. This simplification is not inappropriate, since only 30 companies changed domain between 1999 and 2000, and only 25 did so between 2000 and 2001, representing respectively less than 0.5% and 0.2% of the FTEs in the sample.

Calibration is carried out each year, and it can be taken into account using a residuals technique. As with estimating the variance of the cross-sectional estimators, robustification is taken into account by reweighting the survey variables.

With realistic assumptions, all components of the variance may be taken into account by means of the general expression (4.7). This expression is applied within each stratum and it covers all the components of the survey of value added: the panel effect, non-response, stratification, calibration and robustification. The estimators for the survey of value added are ratio estimators, and in this case the calculation of residuals is simplified. This is because in the case of the ratio, the regression coefficients given in (4.5) and (4.6) are calculated having only one dummy variable, and therefore $\mathbf{z}_{kj} = z_{kj}$ is scalar. Also, we take $q_{kj} = 1/z_{kj}$, for $j = 1, 2$, and with robustification taken into account, we thus obtain:

$$eu_{k1} = u_{k1}x_k - \widehat{B}_1 u_{k1} z_{k1},$$
$$eu_{k2} = u_{k2}y_k - \widehat{B}_2 u_{k2} z_{k2},$$

where

$$\widehat{B}_1 = \frac{\sum_{D_1} u_{k1}x_k/\pi_{k1}}{\sum_{D_1} u_{k1}z_{k1}/\pi_{k1}},$$
$$\widehat{B}_2 = \frac{\sum_{D_2} u_{k2}y_k/\pi_{k2}}{\sum_{D_2} u_{k2}z_{k2}/\pi_{k2}}.$$

### 4.5.3 Variance estimation of changes

We made estimates of the standard deviations of changes in gross output values and value added figures calculated by the Swiss Federal Statistical Office. These estimates take into consideration all the aspects described above. We compared them with the estimated standard deviations that would have been obtained under the assumption that the draws for the different waves are independent. Over the various activity strata, the standard deviations that take account of the correlation between the survey waves are 41% lower than those based on the assumption of independence. This makes it possible to have much smaller confidence intervals than those calculated before this study, which were more quickly obtained but less precise. However, the gain is not the same in all activity strata. The following tables show standard deviations (SDs), calculated for the five largest activity strata (NOGA), of changes over time in the value of gross output ($\Delta OV$) and of value added ($\Delta VA$) between 1999 and 2000. The standard deviation that would have been obtained by ignoring the correlation between samples ($SD_{ind}$) is also included in the tables, along with the "gain" in precision realized by taking this correlation into account.

### Acknowledgments

This study was carried out under an agreement between the University of Neuchâtel and the Swiss Federal Statistical Office. The findings published

Table 4.1 – *Change in gross output value between 1999 and 2000 and standard deviations (in billions of Swiss francs)*

| Stratum | $\Delta OV$ | $SD_{ind}$ | SD | Gain (%) |
|---|---|---|---|---|
| 1 | 3.31 | 2.35 | 0.87 | 63 |
| 2 | -0.77 | 4.38 | 1.98 | 55 |
| 3 | 3.07 | 2.11 | 0.94 | 56 |
| 4 | 4.33 | 1.10 | 1.00 | 09 |
| 5 | -0.09 | 0.81 | 0.53 | 35 |

Table 4.2 – *Change in value added between 1999 and 2000 standard deviations (in billions of Swiss francs)*

| Stratum | $\Delta VA$ | $SD_{ind}$ | SD | Gain (%) |
|---|---|---|---|---|
| 1 | 1.96 | 0.91 | 0.32 | 65 |
| 2 | 0.68 | 2.99 | 1.04 | 65 |
| 3 | 1.90 | 1.47 | 0.72 | 51 |
| 4 | 0.36 | 0.47 | 0.45 | 05 |
| 5 | -0.36 | 0.59 | 0.43 | 27 |

in this article are those of the authors alone and in no case do they commit the Federal Statistical Office. We wish to thank Paul-André Salamin for his contribution to this study.

## Appendix : Demonstration of proposition 4.1

It is well known that

$$\text{var}(\widehat{X}_1) = N^2 \left( \frac{1}{n_1} - \frac{1}{N} \right) S_x^2$$

and

$$\text{var}(\widehat{Y}_2) = N^2 \left( \frac{1}{n_2} - \frac{1}{N} \right) S_y^2.$$

It is thus sufficient to calculate $\text{cov}(\widehat{X}_1, \widehat{Y}_2)$. We note

$$\begin{aligned} \overline{x}_A &= \frac{1}{n_A} \sum_{k \in s_A} x_k, & \overline{x}_C &= \frac{1}{n_C} \sum_{k \in s_C} x_k, \\ \overline{y}_B &= \frac{1}{n_B} \sum_{k \in s_B} y_k, & \overline{y}_C &= \frac{1}{n_C} \sum_{k \in s_C} y_k, \\ \overline{x}_1 &= \frac{n_A \overline{x}_A + n_C \overline{x}_C}{n_1}, & \overline{y}_2 &= \frac{n_B \overline{y}_B + n_C \overline{y}_C}{n_2}, \end{aligned}$$

and therefore $\widehat{X}_1 = N\overline{x}_1$ and $\widehat{Y}_2 = N\overline{y}_2$. We must still calculate

$$\text{cov}(\overline{x}_1, \overline{y}_2) = \text{E}\, \text{cov}(\overline{x}_1, \overline{y}_2 | n_A, n_B, n_C) + \text{cov}\left[ E(\overline{x}_1 | n_A, n_B, n_C), E(\overline{y}_2 | n_A, n_B, n_C) \right].$$

Since $\overline{x}_1$ and $\overline{y}_2$ are unbiased conditional on $n_A, n_B$ and $n_C$,

$$\operatorname{cov}\left[E(\overline{x}_1|n_A, n_B, n_C), E(\overline{y}_2|n_A, n_B, n_C)\right] = \operatorname{cov}\left(\overline{X}, \overline{Y}\right) = 0.$$

We therefore obtain

$$\operatorname{cov}(\overline{x}_1, \overline{y}_2) = \mathrm{E}\operatorname{cov}(\overline{x}_1, \overline{y}_2|n_A, n_B, n_C).$$

Conditional on $n_A, n_B$, and $n_C$, we are in case A of Tam (1984, theorem 1). The conditional variance is equal to:

$$\operatorname{cov}(\overline{x}_1, \overline{y}_2|n_A, n_B, n_C) = \left(\frac{n_C}{n_1 n_2} - \frac{1}{N}\right) S_{xy}$$

and therefore

$$\operatorname{cov}(\overline{x}_1, \overline{y}_2) = \left(\frac{\mathrm{E}(n_C)}{n_1 n_2} - \frac{1}{N}\right) S_{xy}.$$

Now

$$\operatorname{cov}(\widehat{X}_1, \widehat{Y}_2) = N^2 \operatorname{cov}(\overline{x}_1, \overline{y}_2),$$

enabling us to obtain result (4.1).

# 5 Covariance of Horvitz-Thompson estimators in repeated surveys with unequal inclusion probabilities

**Abstract**

The covariance of Horvitz-Thompson estimators based on two overlapping samples with unequal probabilities is computed for a special class of bidimensional sampling designs. A family of estimators of this covariance is derived, based on well known variance approximations. These estimators can be used to estimate the variance of evolutions in rotating panels, or in panels with uniform non-response at each wave. [1]

**Keywords:** Rotating panels, Variance estimation, Overlapping samples

## Introduction

If two samples are drawn from the same population using simple random sampling designs, the covariance of Horvitz-Thompson estimators based on these samples depends on the overlap of the samples and on the correlation of the observed variables. Explicit formulae for this covariance, conditional or unconditional on the size of the overlap, are available, for example in Tam (1984); Qualité & Tillé (2008). For unequal probability sampling designs, estimators of covariance are usually not practical as they require the use of second order inclusion probabilities of the joint sampling design. Berger (2004b) derived estimators that use only the first order inclusion probabilities for maximum entropy sampling designs with fixed sizes and fixed size overlap. These estimators are based on a normality assumption, in the same way as Hájek (1964) derived a variance estimator for the rejective sampling design.

[1] AMS 2000 subject classification 62D05.

In this paper, we use another approach to derive an explicit formula for the covariance of Horvitz-Thompson estimators without making assumptions on the entropy of the sampling design or on the normality of the estimators. We consider sampling designs that are essentially applicable to obtain rotating panels and/or to model uniform wave non-response in a panel. The expressions we get for the covariance involves the variance-covariance operator of the overall sampling design of the panel, and other than that, only uses the first order inclusion probabilities. In the case of simple random sampling, the covariance depends on the correlation of interest variables and this correlation must be estimated on the overlap of the samples where both variables are observed. In the case of unequal probabilities, we will also need to be able to estimate the variance-covariance operator on the overlap of the samples.

In Section 5.1, we give some definitions and notations. In Section 5.2, we recall well known results in the case of simple random sampling. In Section 5.3, we give our new results in the case of unequal probability sampling designs with and without replacement, and in Section 5.4 we derive estimators for some specific sampling designs. Finally, in Section 5.5, we provide some simulation results.

## 5.1 Definitions

**Definition 5.1** *A sampling design without replacement is a probability law $P(\cdot)$ on the subsets or samples $s$ of a finite population $U = \{1 \dots N\}$. It is said to have a fixed size $n$ when all the samples that receive a positive probability contain exactly $n$ units of the population.*

We can define the first order inclusion probabilities $\pi_k = P(s \ni k)$, and the second order inclusion probabilities $\pi_{k\ell} = P(s \ni k, \ell)$. If all the $\pi_k$ are positive, a natural and unbiased estimator of the total $Y = y_1 + \dots + y_N$ was proposed by Narain (1951); Horvitz & Thompson (1952):

$$\widehat{Y}_{HT} = \sum_{k \in s} \frac{y_k}{\pi_k}.$$

Its variance is

$$\operatorname{var}(\widehat{Y}_{HT}) = \sum_{k \in U} \frac{1 - \pi_k}{\pi_k} y_k^2 + \sum_{k \in U} \sum_{\substack{\ell \in U \\ \ell \neq k}} \frac{\pi_{k\ell} - \pi_k \pi_\ell}{\pi_k \pi_\ell} y_k y_\ell.$$

A natural extension of definition (5.1) for the selection of multiple samples is to consider probability laws on a product space. These laws will be called multidimensional sampling designs.

**Definition 5.2** *A bidimensional sampling design is a probability law $P(s_1, s_2)$ on the couples $(s_1, s_2)$ of samples of $U$.*

We can define the marginal sampling designs $P_1(s_1) = P(s_1, \cdot)$ and $P_2(s_2) = P(\cdot, s_2)$. We can also define the marginal inclusion probabilities $\pi_k^1 = P_1(s_1 \ni k)$ and $\pi_k^2 = P_2(s_2 \ni k)$. This definition allows for the

most general sampling designs for two samples in the same population. Estimation using the intersection of these two samples also requires the parameters $\pi_k^{1,2} = P(s_1 \cap s_2 \ni k)$. Variance estimation usually depends on second order inclusion probabilities $\pi_{k,\ell}^1$, $\pi_{k,\ell}^2$, but also $\pi_{k,\ell}^{1,2} = P(s_1 \cap s_2 \ni k, \ell)$. In the rest of this paper, we will use the notation $s_{12} = s_1 \cap s_2$.

## 5.2 Simple random sampling

Fixed size bidimensional simple random sampling can be defined as a uniform probability law on all pairs of samples which respect some size constraints. In order to select rotating samples, one can use a uniform probability law on all the pairs of samples $(s_1, s_2)$ such that $s_1$ has size $n_1$, $s_2$ has size $n_2$, and $s_1 \cap s_2$ holds $n_{12}$ units (on this subject, see Goga, 2003). In this case, the variance and covariance of the Horvitz-Thompson estimators are given by (see Tam, 1984; Qualité & Tillé, 2008):

$$\begin{aligned}
\operatorname{var}(\widehat{X}_1) &= N^2 \left( \frac{1}{n_1} - \frac{1}{N} \right) S_x^2, \\
\operatorname{var}(\widehat{Y}_2) &= N^2 \left( \frac{1}{n_2} - \frac{1}{N} \right) S_y^2, \\
\operatorname{cov}(\widehat{X}_1, \widehat{Y}_2) &= N^2 \left( \frac{n_{12}}{n_1 n_2} - \frac{1}{N} \right) S_{xy}.
\end{aligned}$$

The variance of the difference estimator $\widehat{Y}_2 - \widehat{X}_1$ is equal to:

$$\operatorname{var}(\widehat{Y}_2 - \widehat{X}_1) = \operatorname{var}(\widehat{Y}_2) + \operatorname{var}(\widehat{X}_1) - 2\operatorname{cov}(\widehat{Y}_2, \widehat{X}_1).$$

This variance is minimal when $\operatorname{cov}(\widehat{Y}_2, \widehat{X}_1)$ is maximal, that is to say, if $X$ and $Y$ are positively correlated, when $n_{12} = \operatorname{Min}(\mathrm{n}_1, \mathrm{n}_2)$. We can remark that the covariance between the estimators on two non-overlapping samples is independent from the size of the samples. It is equal to $-NS_{xy}$ and has signature $(0, 1, N-1)$.

Parameters $S_x^2$ and $S_y^2$ are usually estimated respectively by the empirical variance $s_{x,s_1}^2$ in sample $s_1$ and $s_{y,s_2}^2$ in sample $s_2$. $S_{xy}$ may be estimated using observations on $s_{12}$ only, by the empirical covariance $s_{xy,s_{12}}$ or, if estimated standard errors $s_{x,s_{12}}$, $s_{y,s_{12}}$ on $s_{12}$ are positive, by $\rho_{xy,s_{12}} s_{x,s_1} s_{y,s_2}$ where

$$\rho_{xy,s_{12}} = \frac{s_{xy,s_{12}}}{s_{x,s_{12}} s_{y,s_{12}}},$$

is the correlation observed on $s_{12}$.

## 5.3 Unequal probability sampling

We derive variance-covariance formulae for bidimensional sampling designs that can be described as follows. Consider a first phase sampling design with fixed size and unequal inclusion probabilities $\pi_1^s, \ldots, \pi_N^s$. Define the bidimensional sampling design $p(s_1, s_2)$ given by a second sampling phase that consists in bidimensional simple random sampling with

fixed size $(n_1, n_{12}, n_2)$ in the first phase sample. This sampling design adequately models the result of uniform wave non-response in a panel observed on two occasions. The size $n_{12}$ of the overlap allows to take into account the correlation between non-response at the different waves of the panel. The obtained covariances are conditional on the observed sizes $n_1$, $n_2$ and $n_{12}$, but it is possible to take their expectation and obtain unconditional covariances. The considered family of sampling designs also describes rotating panels as they are usually defined on a static population. An overall sample can be selected, with unequal inclusion probabilities (e.g. proportional to a known size variable), and then split into different parts that will be surveyed so as to obtain a rotation of the sample. In the case of maximum entropy sampling designs, and still on a static population, this type of two-phase sampling designs is equivalent to the procedure where a first sample is selected, then on the following occasion a random part of this sample is discarded and replaced with a non-overlapping sample drawn with inclusion probabilities proportional to the same size variable.

Conditionally on the first selection phase we are looking at simple random sampling, and we can use the formulae of Section 5.2. This property greatly simplifies the second order inclusion probabilities $\pi^{1,2}_{k,\ell}$, and the obtention of a variance equation. This setting implies that inclusion probabilities for both marginal sampling designs are proportional:

$$\pi^1_k = \frac{n_1}{n}\pi^s_k, \text{ and } \pi^2_k = \frac{n_2}{n}\pi^s_k.$$

This restriction is not present in Berger (2004b), but is satisfied in the applications proposed in both our papers. We should also note that estimation of the evolution of a same variable observed at two different times is generally much more precise when conducted on the matched sample $s_{12}$ than using estimators $\widehat{X}_1$ and $\widehat{Y}_2$, but in many cases the published evolutions must be consistent with the estimations of levels. Hence the need for covariance estimators on samples that have only a partial overlap.

**Proposition 5.1** *If the first phase sampling design is a fixed size design of size $n$, with or without replacement, and with inclusion probabilities $\pi^s_k$, $k \in U$, if the covariance operator associated with this design is denoted* $\mathrm{cov}_s$*, and if the second phase sampling is a bidimensional simple random sampling with fixed size $(n_1, n_{12}, n_2)$ in the first phase sample, let*

$$A(X,Y) = n\sum_{k\in U}\pi^s_k\left(\frac{x_k}{\pi^s_k}-\frac{X}{n}\right)\left(\frac{y_k}{\pi^s_k}-\frac{Y}{n}\right) - \mathrm{cov}_s(\widehat{X}_s,\widehat{Y}_s), \quad (5.1)$$

*and*

$$B(X,Y) = \sum_{k\in U}\pi^s_k\left(\frac{x_k}{\pi^s_k}-\frac{X}{n}\right)\left(\frac{y_k}{\pi^s_k}-\frac{Y}{n}\right) - \mathrm{cov}_s(\widehat{X}_s,\widehat{Y}_s), \quad (5.2)$$

*where $\widehat{X}_s$ and $\widehat{Y}_s$ are the Horvitz-Thompson estimators of totals $X$ and $Y$ on the*

*first phase sample s. Then,*

$$\begin{aligned}
\operatorname{var}(\widehat{\bar{X}}_1) &= \frac{n}{n-1}\left[\frac{1}{n_1}A(X,X) - B(X,X)\right], && (5.3)\\
\operatorname{var}(\widehat{\bar{Y}}_2) &= \frac{n}{n-1}\left[\frac{1}{n_2}A(Y,Y) - B(Y,Y)\right], && (5.4)\\
\operatorname{cov}(\widehat{\bar{X}}_1, \widehat{\bar{Y}}_2) &= \frac{n}{n-1}\left[\frac{n_{12}}{n_1 n_2}A(X,Y) - B(X,Y)\right]. && (5.5)
\end{aligned}$$

A proof of proposition 5.1 is given in appendix, page 5.6. In expressions 5.1 and 5.2, we recognize the variance of the Hansen-Hurwitz estimator of a total for the sampling design with replacement and with parameter $p_k = \pi_k/n$, $k \in U$ (see Hansen & Hurwitz, 1943). Thus, we can rewrite

$$\begin{aligned}
A(X,Y) &= n \cdot \operatorname{cov}_{wr}(\widehat{X}_{HH,s}, \widehat{Y}_{HH,s}) - \operatorname{cov}_s(\widehat{X}_s, \widehat{Y}_s),\\
B(X,Y) &= \operatorname{cov}_{wr}(\widehat{X}_{HH,s}, \widehat{Y}_{HH,s}) - \operatorname{cov}_s(\widehat{X}_s, \widehat{Y}_s),
\end{aligned}$$

where $\widehat{X}_{HH,s}$, $\widehat{Y}_{HH,s}$ are Hansen-Hurwitz estimators of the total of $X$ and $Y$ and $\operatorname{cov}_{wr}$ is the variance-covariance operator of the with-replacement sampling design.

**Remark 5.1**

- *Expressions 5.1 and 5.2 are valid for any $n \geq n_1 + n_2 - n_{12}$, the first phase sample just needs to be big enough to hold samples $s_1$ and $s_2$;*
- *$A$, $B$ and $\operatorname{cov}(\widehat{\bar{X}}_1, \widehat{\bar{Y}}_2)$ are symmetric. That is to be noted for it is, in general, not the case with bidimensional sampling designs;*
- *When the first phase is a simple random sampling design, or when $n = N$, expressions 5.3, 5.4 and 5.5 are the usual formulae for bidimensional simple random sampling, $nB/(n-1)$ becomes $-NS_{xy}$ and $nA/(n-1)$ becomes $N^2 S_{xy}$;*
- *Using Cauchy-Schwarz inequality, we get that $A(X,Y)$ is non-negative. It has signature $(p,q,0)$ where $p+q=N$ and $q \geq 1$. Best precision will therefore be obtained when $n_{12}$ is maximum if variables $X$ and $Y$ are equal or if variables $(x_k/\pi_k^s)_{k\in U}$ and $(y_k/\pi_k^s)_{k\in U}$ are positively and sufficiently correlated;*
- *$B(X,Y)$ has a similar signature $(p',q',0)$ only when the first phase sampling design is more efficient than sampling with replacement and with the same selection probabilities. This property has been proven to hold for surprisingly few sampling designs: simple random sampling, Sampford's design (see Gabler, 1981; Sampford, 1967), Chao's design (see Sengupta, 1989; Chao, 1982), the elimination method (see Tillé, 1996; Deville & Tillé, 1998), and maximum entropy sampling (see Qualité, 2008). We can also remark that in the case of non-overlapping samples, the covariance does not depend on the size of the samples as in the case of simple random sampling. It is non-positive only if $B$ is non-negative.*

- *In every cases that we will consider, estimators $\widehat{X}_{HH,s}$ and $\widehat{X}_s$ (resp. $\widehat{Y}_{HH,s}$ and $\widehat{Y}_s$) have the same expression. Either the first phase sampling is with replacement and the associated estimator is the Hansen-Hurwitz estimator, or the first phase sampling is without replacement and, as $\pi_k^s = np_k$, Horvitz-Thompson and Hansen-Hurwitz estimators are the same. So, in the rest of this paper, we will simply use the notation $\widehat{X}_s$ (resp. $\widehat{Y}_s$) for both estimators.*

### Application to sampling with replacement

When the first phase sample is drawn with unequal probability sampling with replacement, the variances and covariance expressions simplify greatly as

$$B(X,Y) = 0$$

and

$$A(X,Y) = (n-1)\cdot \mathrm{cov}_{wr}(\widehat{X}_s, \widehat{Y}_s).$$

Thus in this case we have:

$$\begin{aligned}
\mathrm{var}(\widehat{X}_1) = &\quad \tfrac{n}{n_1}\mathrm{cov}_{wr}(\widehat{X}_s,\widehat{X}_s) &&= \tfrac{1}{n_1}\textstyle\sum_{k\in U} p_k \left(\frac{x_k}{p_k} - X\right)^2,\\
\mathrm{var}(\widehat{Y}_2) = &\quad \tfrac{n}{n_2}\mathrm{cov}_{wr}(\widehat{Y}_s,\widehat{Y}_s) &&= \tfrac{1}{n_2}\textstyle\sum_{k\in U} p_k \left(\frac{y_k}{p_k} - Y\right)^2,\\
\mathrm{cov}(\widehat{X}_1,\widehat{Y}_2) = &\quad \tfrac{nn_{12}}{n_1n_2}\mathrm{cov}_{wr}(\widehat{X}_s,\widehat{Y}_s) &&= \tfrac{n_{12}}{n_1n_2}\textstyle\sum_{k\in U} p_k \left(\frac{x_k}{p_k} - X\right)\left(\frac{y_k}{p_k} - Y\right),
\end{aligned}$$

where $p_k = \pi_k^s/n$. The variance of the difference estimator $\widehat{Y}_2 - \widehat{X}_1$ becomes:

$$\begin{aligned}
\mathrm{var}(\widehat{Y}_2 - \widehat{X}_1) = \sum_{k\in U} p_k \Bigg[ &\frac{1}{n_1}\left(\frac{x_k}{p_k} - X\right)^2 + \frac{1}{n_2}\left(\frac{y_k}{p_k} - Y\right)^2 \\
&- 2\frac{n_{12}}{n_1n_2}\left(\frac{x_k}{p_k} - X\right)\left(\frac{y_k}{p_k} - Y\right)\Bigg]. \qquad (5.6)
\end{aligned}$$

In many cases sampling without replacement is more efficient than sampling with replacement and, when it is the case, equation (5.6) gives a rough majoration of the variance of the difference estimator.

## 5.4 Estimation

We are confronted with the usual difficulty of covariance estimation on different samples: the couple of variables $(X,Y)$ is observed only on the subsample $s_{12}$. Thus we do not have simple substitution estimators for $A(X,Y)$ and $B(X,Y)$, even when there exists a simple expression that gives an estimator of the operator $\mathrm{cov}_s(\cdot,\cdot)$. Our aim is to estimate the covariances $\mathrm{cov}_{wr}(\widehat{X}_s,\widehat{Y}_s)$ and $\mathrm{cov}_s(\widehat{X}_s,\widehat{Y}_s)$. This can only be done when $n_{12} \geq 2$, and from now on we will assume that it is the case.

### 5.4.1 Estimation of the with-replacement covariance $\mathrm{cov}_{wr}(\cdot,\cdot)$

Thanks to its simple form, we can give various estimators for $\mathrm{cov}_{wr}(\widehat{X}_s, \widehat{Y}_s)$. The first and most obvious estimator is the simple expansion estimator:

$$\widehat{\mathrm{cov}}_{wr,a}(\widehat{X}_s, \widehat{Y}_s) = \frac{n}{n_{12}-1} \sum_{k\in s_{12}} \left(\frac{x_k}{\pi_k^s} - \frac{\widehat{X}_{12}}{n}\right)\left(\frac{y_k}{\pi_k^s} - \frac{\widehat{Y}_{12}}{n}\right), \tag{5.7}$$

where

$$\widehat{X}_{12} = \frac{n}{n_{12}} \sum_{k\in s_{12}} \frac{x_k}{\pi_k^s}, \text{ and } \widehat{Y}_{12} = \frac{n}{n_{12}} \sum_{k\in s_{12}} \frac{y_k}{\pi_k^s}.$$

$\widehat{X}_{12}$ and $\widehat{Y}_{12}$ are not the best available estimators of $X$ and $Y$, since these variables are observed on larger samples, but it is generally bad practice to use $\widehat{X}_1$ and $\widehat{Y}_2$ at this stage in an estimation of covariance based on $s_{12}$. Indeed, using these estimators instead of $\widehat{X}_{12}$ and $\widehat{Y}_{12}$ would lead to an estimator that can take negative values even when variables $X$ and $Y$ are equal.

Using the same idea as for the estimation of $S_{xy}$, we can also estimate a correlation coefficient

$$R_{xy} = \frac{\sum_{k\in U} \pi_k^s \left(\frac{x_k}{\pi_k^s} - \frac{X}{n}\right)\left(\frac{y_k}{\pi_k^s} - \frac{Y}{n}\right)}{\left[\sum_{k\in U} \pi_k^s \left(\frac{x_k}{\pi_k^s} - \frac{X}{n}\right)^2 \sum_{k\in U} \pi_k^s \left(\frac{y_k}{\pi_k^s} - \frac{Y}{n}\right)^2\right]^{\frac{1}{2}}}$$

on the subsample $s_{12}$, by

$$\rho_{w_{12}}^{xy} = \frac{\sum_{k\in s_{12}} \left(\frac{x_k}{\pi_k^s} - \frac{\widehat{X}_{12}}{n}\right)\left(\frac{y_k}{\pi_k^s} - \frac{\widehat{Y}_{12}}{n}\right)}{\left[\sum_{k\in s_{12}} \left(\frac{x_k}{\pi_k^s} - \frac{\widehat{X}_{12}}{n}\right)^2 \sum_{k\in s_{12}} \left(\frac{y_k}{\pi_k^s} - \frac{\widehat{Y}_{12}}{n}\right)^2\right]^{\frac{1}{2}}},$$

if this quantity is defined, and consider the ratio type estimator

$$\widehat{\mathrm{cov}}_{wr,b}(\widehat{X}_s, \widehat{Y}_s) = \rho_{w_{12}}^{xy} \sigma_{w_1}^x \sigma_{w_2}^y, \tag{5.8}$$

where

$$\sigma_{w_1}^x = \left[\frac{n}{n_1-1} \sum_{k\in s_1} \left(\frac{x_k}{\pi_k^s} - \frac{\widehat{X}_1}{n}\right)^2\right]^{\frac{1}{2}}$$

and

$$\sigma_{w_2}^y = \left[\frac{n}{n_2-1} \sum_{k\in s_2} \left(\frac{y_k}{\pi_k^s} - \frac{\widehat{Y}_2}{n}\right)^2\right]^{\frac{1}{2}}.$$

### 5.4.2 Estimation of the design covariance $\mathrm{cov}_s(\cdot,\cdot)$

Estimation of $\mathrm{cov}_s(\widehat{X}_s, \widehat{Y}_s)$ using subsample $s_{12}$ depends on the expression of the operator $\mathrm{cov}_s$. If the first phase sampling design is a maximum entropy design with fixed size several simple approximations of $\mathrm{cov}_s$ are available that use only the first order inclusion probabilities (on this subject, see Hájek, 1964; Deville, 1993, 1999; Brewer, 2002; Brewer & Donadio, 2003). Their performances have been studied in Berger (2004a); Matei & Tillé (2005). These approximations are still valid for other sampling designs that have high entropy but are not strictly maximum entropy designs (see for example Berger, 2005).

These estimators have a common composition: they consist in a simple sum over the observed sample of factors depending on $x_k$, $y_k$ and inclusion probabilities $\pi_k^s$. They also present the advantage of giving non-negative estimations of variance. These estimators can easily be adapted to suit our needs. If $\widehat{\mathrm{var}}_\lambda(\cdot)$ is an estimator of $\mathrm{var}_s(\cdot)$ that can be written as

$$\widehat{\mathrm{var}}_\lambda(\widehat{X}_s) = f(n)\frac{1}{n-1}\sum_{k\in s} w_k \left(\frac{x_k}{\pi_k^s} - \frac{\sum_{k\in s} w_k \frac{x_k}{\pi_k^s}}{\sum_{k\in s} w_k}\right)^2,$$

where $w_k$ are non-negative numbers, and if $s_a$ is a simple random subsample of $s$, of size $n_a > 1$, $\mathrm{var}(\widehat{X}_s)$ can be estimated using data observed in $s_a$ only by:

$$\widehat{\mathrm{var}}_{\lambda,s_a}(\widehat{X}_s) = f(n)\frac{1}{n_a-1}\sum_{k\in s_a} w_k \left(\frac{x_k}{\pi_k^s} - \frac{\sum_{k\in s_a} w_k \frac{x_k}{\pi_k^s}}{\sum_{k\in s_a} w_k}\right)^2.$$

This estimator should not be confused with an estimator of $\mathrm{var}(\widehat{X}_{s_a})$. At this stage, we want to estimate $\mathrm{var}_s(\cdot)$, this is why weights $\pi_k^s$ are not replaced with $\pi_k^{s_a}$ and coefficients $w_k$ are not modified.

An estimator in this family, which performs adequately (see Hájek, 1964; Deville, 1993; Matei & Tillé, 2005), and has a very simple expression is:

$$\widehat{\mathrm{cov}}_{HD}(\widehat{X}_s, \widehat{Y}_s) = \frac{n}{n-1}\sum_{k\in s}(1-\pi_k^s)\left(\frac{x_k}{\pi_k^s} - \sum_{k\in s} a_k^s \frac{x_k}{\pi_k^s}\right)\left(\frac{y_k}{\pi_k^s} - \sum_{k\in s} a_k^s \frac{y_k}{\pi_k^s}\right), \tag{5.9}$$

where

$$a_k^s = \frac{1-\pi_k^s}{\sum_{k\in s}(1-\pi_k^s)}.$$

If $n_{12} > 0$, a simple expansion can be used to derive an estimator based on $s_{12}$ only:

$$\widehat{\mathrm{cov}}_{s,a}(\widehat{X}_s, \widehat{Y}_s) = \frac{n}{n_{12}-1}\sum_{k\in s_{12}}(1-\pi_k^s)\left(\frac{x_k}{\pi_k^s} - \sum_{k\in s_{12}} a_k^{s_{12}} \frac{x_k}{\pi_k^s}\right)\left(\frac{y_k}{\pi_k^s} - \sum_{k\in s_{12}} a_k^{s_{12}} \frac{y_k}{\pi_k^s}\right), \tag{5.10}$$

where

$$a_k^{s_{12}} = \frac{1-\pi_k^s}{\sum_{k\in s_{12}}(1-\pi_k^s)}.$$

And we can consider the ratio type estimator:

$$\widehat{\mathrm{cov}}_{s,b}(\widehat{X}_s, \widehat{Y}_s) = \rho_{s_{12}}^{xy} \sigma_{s_1}^{x} \sigma_{s_2}^{y}, \tag{5.11}$$

where

$$\begin{aligned}
\rho_{s_{12}}^{xy} &= \frac{\sum_{k\in s_{12}} (1-\pi_k^s)\left(\frac{x_k}{\pi_k^s} - \sum_{k\in s_{12}} \frac{a_k^{s_{12}} x_k}{\pi_k^s}\right)\left(\frac{y_k}{\pi_k^s} - \sum_{k\in s_{12}} \frac{a_k^{s_{12}} y_k}{\pi_k^s}\right)}{\left[\sum_{k\in s_{12}} (1-\pi_k^s)\left(\frac{x_k}{\pi_k^s} - \sum_{k\in s_{12}} \frac{a_k^{s_{12}} x_k}{\pi_k^s}\right)^2 \sum_{k\in s_{12}} (1-\pi_k^s)\left(\frac{y_k}{\pi_k^s} - \sum_{k\in s_{12}} \frac{a_k^{s_{12}} y_k}{\pi_k^s}\right)^2\right]^{\frac{1}{2}}}, \\
\sigma_{s_1}^{x} &= \left[\frac{n}{n_1 - 1} \sum_{k\in s_1} (1-\pi_k^s)\left(\frac{x_k}{\pi_k^s} - \sum_{k\in s_1} \frac{a_k^{s_1} x_k}{\pi_k^s}\right)^2\right]^{\frac{1}{2}}, \\
\sigma_{s_2}^{y} &= \left[\frac{n}{n_2 - 1} \sum_{k\in s_2} (1-\pi_k^s)\left(\frac{y_k}{\pi_k^s} - \sum_{k\in s_2} \frac{a_k^{s_2} y_k}{\pi_k^s}\right)^2\right]^{\frac{1}{2}}, \\
a_k^{s_1} &= \frac{1-\pi_k^s}{\sum_{k\in s_1}(1-\pi_k^s)}, \\
a_k^{s_2} &= \frac{1-\pi_k^s}{\sum_{k\in s_2}(1-\pi_k^s)},
\end{aligned}$$

assuming that $\rho_{s_{12}}^{xy}$ is defined.

We will give simulation outputs for estimators 5.10 and 5.11 in Section 5.5, and some theoretical results, in Section 5.4.3, on the estimators of $\mathrm{cov}(\widehat{X}_1, \widehat{Y}_2)$ obtained when we use an estimator of $\mathrm{cov}_s(\cdot,\cdot)$ of the general form

$$\mathrm{cov}_s(\cdot,\cdot) = \rho_{s_{12}}^{xy} \sigma_{s_1}^{x} \sigma_{s_2}^{y}, \tag{5.12}$$

where $\rho_{s_{12}}^{xy}$, $\sigma_{s_1}^{x}$ and $\sigma_{s_2}^{y}$ derive from a suitable non-negative estimator $\widehat{\mathrm{var}}_\lambda(\cdot)$.

### 5.4.3 Estimators of covariance

We can derive two estimators of covariance from Equation 5.5. A first one is based only on data observed on $s_{12}$ and uses estimators 5.7 and 5.10:

$$\widehat{\mathrm{cov}}_a(\widehat{X}_1, \widehat{Y}_2) = \frac{n}{n-1}\left[\frac{n_{12}}{n_1 n_2} \widehat{A}_a(X,Y) - \widehat{B}_a(X,Y)\right], \tag{5.13}$$

where

$$\widehat{A}_a(X,Y) = n \cdot \widehat{\mathrm{cov}}_{wr,a}(\widehat{X}_s, \widehat{Y}_s) - \widehat{\mathrm{cov}}_{s,a}(\widehat{X}_s, \widehat{Y}_s),$$

and

$$\widehat{B}_a(X,Y) = \widehat{\mathrm{cov}}_{wr,a}(\widehat{X}_s, \widehat{Y}_s) - \widehat{\mathrm{cov}}_{s,a}(\widehat{X}_s, \widehat{Y}_s).$$

The second one uses ratio-type estimators 5.8 and 5.11 (or 5.8 and 5.12):

$$\widehat{\mathrm{cov}}_b(\widehat{X}_1, \widehat{Y}_2) = \frac{n}{n-1}\left[\frac{n_{12}}{n_1 n_2} \widehat{A}_b(X,Y) - \widehat{B}_b(X,Y)\right], \tag{5.14}$$

where

$$\widehat{A}_b(X,Y) = n \cdot \widehat{\mathrm{cov}}_{wr,b}(\widehat{X}_s, \widehat{Y}_s) - \widehat{\mathrm{cov}}_{s,b}(\widehat{X}_s, \widehat{Y}_s),$$

and

$$\widehat{B}_b(X,Y) = \widehat{\mathrm{cov}}_{wr,b}(\widehat{X}_s, \widehat{Y}_s) - \widehat{\mathrm{cov}}_{s,b}(\widehat{X}_s, \widehat{Y}_s).$$

We will see in Section 5.5 that, on simulations, estimator 5.14 performs at least as well as estimator 5.13, and can be much better when the variables of interest are strongly correlated and samples $s_1$ and $s_2$ have a relatively small overlap. Moreover, we have the nice property that estimator 5.14 generally provides non-negative estimations of variance for any linear combination of $\widehat{X}_1$ and $\widehat{Y}_2$. Sufficient conditions for this property to hold are given in the following proposition.

**Proposition 5.2** *Let $\widehat{\mathrm{var}}(\alpha\widehat{X}_1 + \beta\widehat{Y}_2)$ be an estimation of variance obtained from equation 5.5 with estimator 5.14 or with estimators 5.8 and 5.12. Then, any one of the following conditions is sufficient to have that $\widehat{\mathrm{var}}(\alpha\widehat{X}_1 + \beta\widehat{Y}_2) \geq 0$ for any $\alpha,\ \beta \in \mathbb{R}$.*

1. *The estimated correlation $\rho^{xy}_{s_{12}}$ as in 5.11, or 5.12, is such that:*

$$|\rho^{xy}_{s_{12}}| \leq \left(1 - \frac{1}{n_1}\right)^{\frac{1}{2}} \left(1 - \frac{1}{n_2}\right)^{\frac{1}{2}} \left(1 - \frac{n_{12}}{n_1 n_2}\right)^{-1}.$$

2. *To have the result for any couple of variables, if $n_1, n_2, n_{12}$ are not all equal, it is necessary to impose an additional condition on the design variance estimator: it must satisfy the inequality*

$$\widehat{\mathrm{var}}_{\lambda,s_i}(\cdot) \leq \gamma \cdot \widehat{\mathrm{var}}_{wr,s_i}(\cdot),\ i = 1,2, \tag{5.15}$$

*for some positive number $\gamma$ that depends on $n,\ n_1,\ n_2$ and $n_{12}$. It is hard to obtain a simple lower bound for $\gamma$ that is pertinent with every possible choice of these parameters. We begin with two simple cases:*

- *Samples size and overlap are such that $\frac{n_1 n_2}{n} \leq n_{12} < \min(n_1, n_2)$, then*

$$\gamma_1 = n \cdot \left[\frac{\min(n_1,n_2) - n_{12}}{\max(n_1,n_2) - n_{12}}\right]$$

  *gives the result.*

- *Samples size and overlap are such that $n = \max(n_1, n_2)$, and thus $n_{12} = \min(n_1, n_2)$, then we can use*

$$\gamma_2 = \max(n_1, n_2).$$

*In the general case, when $n > \max(n_1, n_2)$, then*

$$\gamma_3 = \frac{[n_1 n_2 (n - n_1)(n - n_2)]^{\frac{1}{2}} - |n_{12} n - n_1 n_2|}{n_1 n_2 - n_{12} - [n_1 n_2 (n_1 - 1)(n_2 - 1)]^{\frac{1}{2}}},$$

*gives the result.*

A proof of Proposition 5.2 is given in appendix.

**Remark 5.2**

- *The upper bound in Condition 1 of Proposition 5.2 tends rapidly to* 1 *when* $n_1$ *and* $n_2$ *are large enough. This condition will thus be satisfied in most cases excepted if the variables of interest are very strongly correlated. The worst case is when the variables are equal. In that case the variance of the difference is small and it is natural to require stronger assumptions on* $\widehat{\text{var}}_\lambda(\cdot)$.
- *The first case, with* $\gamma_1$*, is well suited to the case of panel non-response: if there is a non-response rate equal to* $r$ *at each wave,* $n_1 = n_2 = rn$*, the conditions translate to:*
  - $n_{12} \geq r(rn)$*, which is usually the case because respondents at the first wave tend to have a better response rate at the second wave than non-respondents at the first wave;*
  - $\widehat{\text{var}}_{\lambda,s_i}(\cdot) \leq n \cdot \widehat{\text{var}}_{wr,s_i}(\cdot)$*, which should also be the case except with very poor variance estimators.*
- $\gamma_3$ *can be smaller than* 1*, but after some algebra we can prove that it is possible only when* $\min(n_1, n_2) \leq 3$.
- *None of these bounds are optimal.*

It can be difficult to actually prove that the inequality 5.15 holds, but it should be the case with most reasonable variance estimators and good sampling designs. Indeed, for high entropy sampling designs, $\text{var}_s(\cdot) \leq \text{var}_{wr}(\cdot)$, so a good variance estimator $\widehat{\text{var}}_{\lambda,s_i}(\cdot)$ will not exceed by too much $\widehat{\text{var}}_{wr,s_i}(\cdot)$. Lower bounds $\gamma_1$, $\gamma_2$ and $\gamma_3$ usually take relatively large values and thus these conditions will always be satisfied in real situations. For example, with some simple algebra, we can show that Estimator 5.9 satisfies the inequality:

$$\widehat{\text{var}}_{HD,s_a}(\cdot) \leq 2\left(1 - \min_{k \in s_a} \pi_k^s\right) \widehat{\text{var}}_{wr,s_a}(\cdot).$$

Hence in most cases, when this estimator is deemed suitable, it can be used in equation 5.14 to provide a non-negative variance estimator.

## 5.5 Simulations

Simulations have been conducted using the 'R' package 'sampling'. They give an idea of the performance of estimators 5.7, 5.8, 5.10 and 5.11. We used the dataset 'belgianmunicipalities' and selected samples $s$ of 169 units, with unequal inclusion probabilities proportional to their size in 2004. For this selection we used maximum entropy with fixed size design, and Tillé's design (see Tillé, 1996). Then we selected simple random subsamples $s_1$, $s_2$ with different sizes: $a = |s_1 \backslash s_{12}|$, $b = |s_{12}|$ and $c = |s_2 \backslash s_{12}|$. We computed correlations for two pairs of variables: `Women03` and `Women04` that have a very strong correlation (corr=0.99), and `Women03` and `DiffWom` that are less correlated (corr=0.33). For each set of parameters $n_1$, $n_2$ and $n_{12}$, we selected 1'000 samples and computed

the value of estimators 5.7, 5.8, 5.10 and 5.11 for both sets of variable. In Tables 5.1, 5.2, 5.3 and 5.4, we give the empirical relative bias (RB) and Ratio Root Mean Square Error (RRMSE),

$$RB(\widehat{\text{cov}}) = \frac{\text{E}(\widehat{\text{cov}})}{\text{cov}} - 1, \quad RRMSE(\widehat{\text{cov}}) = \left\{ \frac{\text{E}\left[(\widehat{\text{cov}} - \text{cov})^2\right]}{\text{cov}^2} \right\}^{\frac{1}{2}},$$

of these estimators for both sampling designs and different sets of size parameters. We also included, as a reference, values for the Sen-Yates-Grundy (see Sen, 1953; Yates & Grundy, 1953) estimator of covariance based on the first phase sample $s$:

$$\widehat{\text{cov}}_{SYG}(\widehat{X}, \widehat{Y}) = \frac{1}{2} \sum_{k \in s} \sum_{\ell \in s} \frac{\pi_k^s \pi_\ell^s - \pi_{k,\ell}^s}{\pi_{k,\ell}^s} \left( \frac{x_k}{\pi_k^s} - \frac{x_\ell}{\pi_\ell^s} \right) \left( \frac{y_k}{\pi_k^s} - \frac{y_\ell}{\pi_\ell^s} \right).$$

We can read in Tables 5.1, 5.2, 5.3, and 5.4 that the relative bias of the ratio type estimators is not larger than the relative bias of direct estimators on $s_{12}$. We also see that, if the overlap is large enough, estimators 5.10 and 5.11 perform almost as well as the Sen-Yates-Grundy estimator in terms of mean square error, even if they are estimated on a smaller sample. Finally, ratio type estimators seem to have a smaller mean square error than direct estimators when the samples have a small overlap, and when the observed variables are strongly correlated.

| Size | Estimators | | | | |
|---|---|---|---|---|---|
| $a-b-c$ | 5.7 | 5.8 | 5.10 | 5.11 | SYG |
| 10-130-10 | -0.37(11.5) | -0.46(10.4) | -0.15(12.9) | -0.24(11.9) | -0.28(10.8) |
| 25-100-25 | -0.18(14.4) | -0.14(11.2) | 0.02(15.9) | 0.15(12.8) | 0.09(11.2) |
| 60-30-60 | 0.57(30.2) | -0.42(11.7) | -0.07(31.5) | -1.18(13.5) | -0.17(10.5) |
| 10-100-40 | 0.05(14.3) | -0.12(11.1) | 0.55(15.8) | 0.36(12.7) | 0.03(10.9) |

Table 5.1 – $RB(RRMSE) \times 10^2$ *for variables* `Women03` *and* `Women04`, *CP-sampling design*

| Size | Estimators | | | | |
|---|---|---|---|---|---|
| $a-b-c$ | 5.7 | 5.8 | 5.10 | 5.11 | SYG |
| 10-130-10 | 0.02(21.2) | -0.04(20.9) | 0.19(26.0) | 0.15(25.5) | -0.24(22.1) |
| 25-100-25 | 0.47(25.1) | 0.56(24.0) | 0.14(29.7) | 0.25(28.4) | -0.69(21.1) |
| 60-30-60 | -0.73(50.2) | -1.01(44.0) | -1.13(57.2) | -1.64(48.8) | -0.06(21.3) |
| 10-100-40 | -1.08(26.0) | -1.4(24.4) | -1.27(31.5) | -1.79(29.3) | -1.84(22.8) |

Table 5.2 – $RB(RRMSE) \times 10^2$ *for variables* `Women03` *and* `DiffWom`, *CP-sampling design*

Results for actual covariance estimators of $\text{cov}(\widehat{X}_1, \widehat{Y}_2)$, derived from Equation 5.5 and these estimators, are given in Tables 5.5 and 5.6. There again we observe that ratio type estimators may be a great improvement on direct estimators when the variables of interest are strongly correlated and samples $s_1$ and $s_2$ have a small overlap. This performance however

| Size | Estimators | | | | |
|---|---|---|---|---|---|
| $a-b-c$ | 5.7 | 5.8 | 5.10 | 5.11 | SYG |
| 10-130-10 | -0.10(11.5) | -0.22(10.5) | 1.34(13.4) | 1.21(12.4) | -0.28(10.7) |
| 25-100-25 | 0.28(14.4) | 0.10(11.3) | 2.20(16.6) | 1.96(13.4) | 0.06(11.1) |
| 60-30-60 | 1.97(31.9) | -0.03(12.4) | 3.20(33.2) | 1.50(14.1) | -0.11(10.5) |
| 10-100-40 | 0.32(14.5) | 0.09(11.3) | 1.73(16.3) | 1.67(13.3) | 0.06(10.8) |

Table 5.3 – $RB(RRMSE) \times 10^2$ *for variables* `Women03` *and* `Women04`, *Tillé sampling design*

| Size | Estimators | | | | |
|---|---|---|---|---|---|
| $a-b-c$ | 5.7 | 5.8 | 5.10 | 5.11 | SYG |
| 10-130-10 | -0.14(20.6) | -0.04(20.3) | 1.22(25.3) | 1.35(25.0) | -0.19(22.3) |
| 25-100-25 | 0.29(24.5) | 0.48(23.2) | 1.14(29.7) | 1.38(28.2) | -0.59(21.2) |
| 60-30-60 | -1.82(49.1) | -0.76(43.1) | -0.62(57.4) | 0.41(48.7) | 0.02(21.4) |
| 10-100-40 | -0.63(25.3) | -0.99(24.1) | 0.13(30.8) | -0.28(29.3) | -1.76(23.0) |

Table 5.4 – $RB(RRMSE) \times 10^2$ *for variables* `Women03` *and* `DiffWom`, *Tillé sampling design*

is dependent on the correlation of $X$ and $Y$: when they are just mildly correlated, as is the case of `Women03` and `DiffWom`, the improvement is modest as the correlation estimation is less precise.

| Size | CP-design | | Tillé's design | |
|---|---|---|---|---|
| $a-b-c$ | on $s_{12}$ | ratio-type | on $s_{12}$ | ratio-type |
| 10-130-10 | -0.19(12.4) | -0.28(11.4) | 1.05(12.8) | 0.93(11.8) |
| 25-100-25 | -0.01(15.5) | 0.11(12.5) | 1.93(16.1) | 1.69(13.0) |
| 60-30-60 | 1.73(70.9) | 0.99(27.2) | 7.18(72.9) | 6.43(29.3) |
| 10-100-40 | 0.46(15.4) | 0.28(12.3) | 1.49(15.8) | 1.41(12.8) |

Table 5.5 – $RB(RRMSE) \times 10^2$ *of covariance estimator for* `Women03` *and* `Women04`

| Size | CP-design | | Tillé's design | |
|---|---|---|---|---|
| $a-b-c$ | on $s_{12}$ | ratio-type | on $s_{12}$ | ratio-type |
| 10-130-10 | 0.16(25.1) | 0.12(24.6) | 0.99(24.4) | 1.11(24.1) |
| 25-100-25 | 0.18(29.1) | 0.29(27.8) | 1.03(29.0) | 1.27(27.4) |
| 60-30-60 | -1.76(76.7) | -2.66(64.7) | 1.36(79.9) | 2.35(66.9) |
| 10-100-40 | -1.25(30.6) | 0.29(27.8) | 1.03(29.0) | 1.27(27.4) |

Table 5.6 – $RB(RRMSE) \times 10^2$ *of covariance estimator for* `Women03` *and* `DiffWom`

## 5.6 Discussion

We give exact expressions for the variance and covariance of Horvitz-Thompson estimators in a special type of bidimensional sampling designs with fixed size and unequal inclusion probabilities. From these expres-

sions, we derive estimators that can be used in a large family of repeated surveys. The ratio-type estimators that we proposed gives non-negative variance estimates under mild conditions. While these conditions should be satisfied in most cases, it may be difficult to prove that it is actually the case. This work can easily be extended to take into account deaths in the population that occur between the sampling occasions, in a simple case. Assume that these deaths result from an additional Bernoulli sampling phase, thus uniform in the population or strata, and that this phase is independent from the selection process. In this case, the proposed estimators will give results conditional on the number of living units observed in $s_2$ and $s_{12}$. Births in the population can also be accounted for in a simple case. Assume that the population of newly born units $U_b$ is known and that the survey is updated by selecting a fixed size sample $s_{2b}$ in $U_b$, independently from the selection process of $(s_1, s_2)$. The sample $s_2$ is then replaced with $\tilde{s}_2 = s_2 \cup s_{2b}$. We can write that the resulting Horvitz-Thompson estimator on $\tilde{s}_2$, $\widehat{Y}_{\tilde{2}}$ is the sum of $\widehat{Y}_2$ and of the Horvitz-Thomson estimator $\widehat{Y}_{2b}$, defined on $s_{2b}$, of the total of $Y$ in the population of newly born units. Then, the covariance between $\widehat{Y}_{2b}$ and $\widehat{Y}_2$ or $\widehat{X}_1$ is null. The estimators we proposed can still be used to estimate the covariance between $\widehat{Y}_2$ and $\widehat{X}_1$. Hence, in this simple case, we can also estimate the variance of any linear combination of $\widehat{X}_1$ and $\widehat{Y}_{\tilde{2}}$.

## Appendix

*Proof.* Proof of Proposition 5.1
The equations (5.3) and (5.4) are implied by (5.5). The inclusion probabilities of unit $k$ in $s_1$ and $s_2$ are respectively equal to $\pi_k^1 = n_1\pi_k^s/n$ and $\pi_k^2 = n_2\pi_k^s/n$. The Horvitz-Thompson estimators on $s_1$ and $s_2$ are defined by

$$\widehat{X}_1 = \sum_{k\in U} \frac{x_k}{\pi_k^1} 1_{k\in s_1} = \frac{n}{n_1} \sum_{k\in U} \frac{x_k}{\pi_k^s} 1_{k\in s_1},$$
$$\widehat{Y}_2 = \sum_{k\in U} \frac{y_k}{\pi_k^2} 1_{k\in s_2} = \frac{n}{n_2} \sum_{k\in U} \frac{y_k}{\pi_k^s} 1_{k\in s_2}.$$

We use the identity

$$\mathrm{cov}(\widehat{X}_1, \widehat{Y}_2) = \mathrm{E}_s\left[\mathrm{cov}(\widehat{X}_1, \widehat{Y}_2|s)\right] + \mathrm{cov}_s\left[\mathrm{E}(\widehat{X}_1|s), \mathrm{E}(\widehat{Y}_2|s)\right].$$

The second phase being a simple random sampling, we have :

$$\begin{aligned}
\mathrm{E}(\widehat{X}_1|s) &= \sum_{k\in U} \frac{x_k}{\pi_k^s} 1_{k\in s} = \widehat{X}_s, \\
\mathrm{E}(\widehat{Y}_2|s) &= \sum_{k\in U} \frac{y_k}{\pi_k^s} 1_{k\in s} = \widehat{Y}_s, \\
\mathrm{cov}(\widehat{X}_1, \widehat{Y}_2|s) &= n^2 \left(\frac{n_{12}}{n_1 n_2} - \frac{1}{n}\right) s_{\frac{x}{\pi}\frac{y}{\pi}},
\end{aligned}$$

where

$$s_{\frac{x}{\pi}\frac{y}{\pi}} = \frac{1}{n-1}\sum_{k\in s}\left(\frac{x_k}{\pi_k^s} - \frac{\widehat{X}_s}{n}\right)\left(\frac{y_k}{\pi_k^s} - \frac{\widehat{Y}_s}{n}\right).$$

Hence we have

$$\operatorname{cov}(\widehat{X}_1, \widehat{Y}_2) = n^2\left(\frac{n_{12}}{n_1 n_2} - \frac{1}{n}\right)\mathrm{E}_s(s_{\frac{x}{\pi}\frac{y}{\pi}}) + \operatorname{cov}_s(\widehat{X}_s, \widehat{Y}_s),$$

and it is sufficient to remark that

$$\mathrm{E}_s(s_{\frac{x}{\pi}\frac{y}{\pi}}) = \frac{1}{n-1}\sum_{k\in U}\pi_k^s\left(\frac{x_k}{\pi_k^s} - \frac{X}{n}\right)\left(\frac{y_k}{\pi_k^s} - \frac{Y}{n}\right) - \frac{1}{n(n-1)}\operatorname{cov}_s(\widehat{X}_s, \widehat{Y}_s).$$

□

*Proof.* Proof of Proposition 5.2
Let $\sigma_{s_1}^x$ and $\sigma_{s_2}^y$ be estimated standard errors of $\widehat{X}_s$ and $\widehat{Y}_s$, $\rho_{s_{12}}^{xy} \in [-1, 1]$ be an estimation of the design-correlation between $\widehat{X}_s$ and $\widehat{Y}_s$ as in Section 5.4.2. Let also $\sigma_{w_1}^x$, $\sigma_{w_2}^y$ and $\rho_{w_{12}}^{xy}$ be standard error and correlation estimators of the with-replacement variance of $\widehat{X}_s$ and $\widehat{Y}_s$ as in Section 5.4.1. With $\alpha,\ \beta \in \mathbb{R}$, and using Estimator 5.14, we get the estimation of variance:

$$\widehat{\operatorname{var}}_b\left(\alpha\widehat{X}_1 + \beta\widehat{Y}_2\right) = \frac{n}{n-1}\mathbf{u}'\mathbf{M}\mathbf{u},$$

where $\mathbf{u} = (\alpha\sigma_{w_1}^x, \beta\sigma_{w_2}^y, \alpha\sigma_{s_1}^x, \beta\sigma_{s_2}^y)'$ and $\mathbf{M}$ is the symmetric matrix

$$\mathbf{M} = \begin{pmatrix} \frac{n}{n_1} - 1 & \left(\frac{nn_{12}}{n_1 n_2} - 1\right)\rho_{w_{12}}^{xy} & 0 & 0 \\ & \frac{n}{n_2} - 1 & 0 & 0 \\ & & 1 - \frac{1}{n_1} & \left(1 - \frac{n_{12}}{n_1 n_2}\right)\rho_{s_{12}}^{xy} \\ & & & 1 - \frac{1}{n_2} \end{pmatrix}.$$

Unfortunately $\mathbf{M}$ is not necessarily non-negative. Let us note

$$\mathbf{M}_w = \begin{pmatrix} \frac{n}{n_1} - 1 & \left(\frac{nn_{12}}{n_1 n_2} - 1\right)\rho_{w_{12}}^{xy} \\ & \frac{n}{n_2} - 1 \end{pmatrix},$$

and

$$\mathbf{M}_s = \begin{pmatrix} 1 - \frac{1}{n_1} & \left(1 - \frac{n_{12}}{n_1 n_2}\right)\rho_{s_{12}}^{xy} \\ & 1 - \frac{1}{n_2} \end{pmatrix},$$

so that

$$\mathbf{M} = \begin{pmatrix} \mathbf{M}_w & 0 \\ 0 & \mathbf{M}_s \end{pmatrix}.$$

The matrix $\mathbf{M}_w$ is non-negative for all values of $\rho_{w_{12}}^{xy}$, but $\mathbf{M}_s$ can have a negative eigenvalue. However, $\mathbf{M}_s$ is non-negative exactly when

$$\left|\rho_{s_{12}}^{xy}\right| \leq \left(1 - \frac{1}{n_1}\right)^{\frac{1}{2}}\left(1 - \frac{1}{n_2}\right)^{\frac{1}{2}}\left(1 - \frac{n_{12}}{n_1 n_2}\right)^{-1}.$$

This proves that Condition 1 of Proposition 5.2 is sufficient to have that $\widehat{\text{var}}_b\left(\alpha\widehat{X}_1+\beta\widehat{Y}_2\right) \geq 0$. If $\mathbf{u}$ could take any value, this condition would also be necessary. But, with additional restrictions on $\sigma^x_{s_1}$, $\sigma^y_{s_2}$, $\sigma^x_{w_1}$ and $\sigma^y_{w_2}$ (and thus on design-variance estimators $\widehat{\text{var}}_{\lambda,s_i}(\cdot)$, $i=1,2$), we can still have that $\widehat{\text{var}}_b\left(\alpha\widehat{X}_1+\beta\widehat{Y}_2\right) \geq 0$ for all $\alpha,\ \beta \in \mathbb{R}$. First we note that, with $\mathbf{u}_w = (\alpha\sigma^x_{w_1}, \beta\sigma^y_{w_2})'$ and $\mathbf{u}_s = (\alpha\sigma^x_{s_1}, \beta\sigma^y_{s_2})'$, and if $\lambda_w$ (resp. $\lambda_s$) is the smallest eigenvalue of $\mathbf{M}_w$ (resp. $\mathbf{M}_s$), then:

$$\begin{aligned}
\widehat{\text{var}}_b\left(\alpha\widehat{X}_1+\beta\widehat{Y}_2\right) &= \frac{n}{n-1}\left(\mathbf{u}_w'\mathbf{M}_w\mathbf{u}_w + \mathbf{u}_s'\mathbf{M}_s\mathbf{u}_s\right),\\
\mathbf{u}_w'\mathbf{M}_w\mathbf{u}_w &\geq \lambda_w\mathbf{u}_w'\mathbf{u}_w,\\
\mathbf{u}_s'\mathbf{M}_s\mathbf{u}_s &\geq \lambda_s\mathbf{u}_s'\mathbf{u}_s.
\end{aligned}$$

Thus we have the inequality:

$$\widehat{\text{var}}_b\left(\alpha\widehat{X}_1+\beta\widehat{Y}_2\right) \geq \frac{n}{n-1}\left(\lambda_w\mathbf{u}_w'\mathbf{u}_w + \lambda_s\mathbf{u}_s'\mathbf{u}_s\right).$$

The worst case is when $\lambda_w$ and $\lambda_s$ are minimal, the latter being negative. This happens for example when $|\rho^{xy}_{w_{12}}| = |\rho^{xy}_{s_{12}}| = 1$ (generally when $x_k = y_k$ or $x_k = -y_k$ for all $k \in s_{12}$). Suppose for example that $n_{12} \leq n_1 \leq n_2$ and that $\lambda_s$ is negative. If we have:

$$\mathbf{u}_s'\mathbf{u}_s \leq \frac{\lambda_w}{-\lambda_s}\mathbf{u}_w'\mathbf{u}_w,$$

then the variance estimator is non-negative. If on the contrary $\mathbf{u}_s'\mathbf{u}_s/\mathbf{u}_w'\mathbf{u}_w$ is not bounded, or can take large values, then the overall variance estimator can be negative. Gershgorin's circle theorem gives that

$$\begin{aligned}
\lambda_s &\geq 1-\frac{1}{n_1}-\left(1-\frac{n_{12}}{n_1n_2}\right) = \frac{n_{12}-n_2}{n_1n_2}\\
\lambda_w &\geq \frac{n_1(n-n_2)-|n_{12}n-n_1n_2|}{n_1n_2}.
\end{aligned}$$

If $n_{12}n \geq n_1n_2$, then $\lambda_w \geq n\cdot\frac{n_1-n_{12}}{n_1n_2}$ and

$$\frac{\lambda_w}{-\lambda_s} \geq n\cdot\frac{n_1-n_{12}}{n_2-n_{12}},$$

which proves the first case of Condition 2. If $n_{12}n < n_1n_2$, we also get that

$$\frac{\lambda_w}{-\lambda_s} \geq n_{12}+\frac{n_1^2}{n_2-n_{12}},$$

but $n_{12}$ can be small and $n_2$ can be large so this is not very useful. If $n=n_2$, $\mathbf{M}_w$ has a null eigenvalue and we need to be more careful.

Going back to the problem, after having multiplied everything by $n_1n_2$ and replaced $X$ with $sign(\alpha)\sqrt{|\alpha|}X$ and $Y$ with $sign(\beta)\sqrt{|\beta|}Y$, we want conditions under which

$$\begin{aligned}
0 \leq\ & n_2(n-n_1)\left(\sigma^x_{w_1}\right)^2 + n_1(n-n_2)\left(\sigma^y_{w_2}\right)^2 + 2(n_{12}n-n_1n_2)\rho^{xy}_{w_{12}}\sigma^x_{w_1}\sigma^y_{w_2}\\
+\ & n_2(n_1-1)\left(\sigma^x_{s_1}\right)^2 + n_1(n_2-1)\left(\sigma^y_{s_2}\right)^2 + 2(n_1n_2-n_{12})\rho^{xy}_{s_{12}}\sigma^x_{s_1}\sigma^y_{s_2}.
\end{aligned}$$

The worst possible case is when $\rho_{s_{12}}^{xy} = -1$ and $\rho_{w_{12}}^{xy} = \pm 1$, it is thus sufficient to prove that

$$\begin{aligned} 0 \quad &\leq \quad n_2(n-n_1)\left(\sigma_{w_1}^x\right)^2 + n_1(n-n_2)\left(\sigma_{w_2}^y\right)^2 - 2|n_{12}n - n_1n_2|\sigma_{w_1}^x\sigma_{w_2}^y \\ &+ \quad n_2(n_1-1)\left(\sigma_{s_1}^x\right)^2 + n_1(n_2-1)\left(\sigma_{s_2}^y\right)^2 - 2(n_1n_2 - n_{12})\sigma_{s_1}^x\sigma_{s_2}^y. \end{aligned}$$

If $n = n_2$, then $n_1 = n_{12}$ and this inequality becomes

$$2n_1(n_2-1)\sigma_{s_1}^x\sigma_{s_2}^y \leq n_2(n_2-n_1)\left(\sigma_{w_1}^x\right)^2 + n_2(n_1-1)\left(\sigma_{s_1}^x\right)^2 + n_1(n_2-1)\left(\sigma_{s_2}^y\right)^2.$$

It is thus sufficient to have that

$$\begin{aligned} n_1(n_2-1)\left(\sigma_{s_1}^x\right)^2 &\leq n_2(n_2-n_1)\left(\sigma_{w_1}^x\right)^2 + n_2(n_1-1)\left(\sigma_{s_1}^x\right)^2 \\ (n_2-n_1)\left(\sigma_{s_1}^x\right)^2 &\leq n_2(n_2-n_1)\left(\sigma_{w_1}^x\right)^2 \ \hookleftarrow \\ \left(\sigma_{s_1}^x\right)^2 &\leq n_2\left(\sigma_{w_1}^x\right)^2 \hookleftarrow, \end{aligned}$$

which proves the second case of Condition 2. In the general case, we want to have

$$\begin{aligned} 0 \quad &\leq \quad \left\{[n_2(n-n_1)]^{\frac{1}{2}}\sigma_{w_1}^x - [n_1(n-n_2)]^{\frac{1}{2}}\sigma_{w_2}^y\right\}^2 \\ &+ \quad 2\left\{[n_1(n-n_2)n_2(n-n_1)]^{\frac{1}{2}} - |n_{12}n - n_1n_2|\right\}\sigma_{w_1}^x\sigma_{w_2}^y \\ &+ \quad \left\{[n_2(n_1-1)]^{\frac{1}{2}}\sigma_{s_1}^x - [n_1(n_2-1)]^{\frac{1}{2}}\sigma_{s_2}^y\right\}^2 \\ &+ \quad 2\left\{[n_2(n_1-1)n_1(n_2-1)]^{\frac{1}{2}} - n_1n_2 + n_{12}\right\}\sigma_{s_1}^x\sigma_{s_2}^y. \end{aligned}$$

It is sufficient to have that

$$\sigma_{s_1}^x\sigma_{s_2}^y \leq \frac{[n_1(n-n_2)n_2(n-n_1)]^{\frac{1}{2}} - |n_{12}n - n_1n_2|}{n_1n_2 - n_{12} - [n_2(n_1-1)n_1(n_2-1)]^{\frac{1}{2}}}\sigma_{w_1}^x\sigma_{w_2}^y,$$

which proves the last assertion of Condition 2. □

## 5.7 Addendum: Different estimators of covariance on overlapping samples

In Chapter 5, we presented two kind of covariance estimators when faced with two samples $s_1$ and $s_2$ with overlap $s_{12} = s_1 \cap s_2$. One family consists of direct estimators on the data available for $s_{12}$, like estimator 5.7, and the second family consists of ratio type estimators like estimator 5.8. In 5.8, only a correlation coefficient is estimated on $s_{12}$, and all the information on $s_1$ and $s_2$ is used to obtain an estimation of covariance. Berger (2004b) notes that correlations estimated in a matched sample may overestimate the correlation in the total population. Indeed, if the observed samples result from a non-response mechanism in a panel, units that have an average evolution may respond more frequently than units that have atypical evolutions. He also notes that if two random variables $\widehat{X}_1$ and $\widehat{Y}_2$ are such that $\text{var}(\widehat{X}_1) = \text{var}(\widehat{Y}_2) = v^2$, that $\widehat{\rho}$ is an estimator of $\rho$ with a relative bias $RB(\widehat{\rho})$ and that we can make the approximation

$$\widehat{\text{var}}(\widehat{X}_1 - \widehat{Y}_2) \approx 2v^2(1 - \widehat{\rho}),$$

whereas the true variance is

$$\text{var}(\widehat{X}_1 - \widehat{Y}_2) = 2v^2(1 - \rho),$$

we get that $\widehat{\text{var}}(\widehat{X}_1 - \widehat{Y}_2)$ has relative bias:

$$RB\left[\widehat{\text{var}}(\widehat{X}_1 - \widehat{Y}_2)\right] = RB(\widehat{\rho})\frac{\rho}{1-\rho}. \tag{5.16}$$

So, even a small relative bias $RB(\widehat{\rho})$ could result in a large relative bias for $\widehat{\text{var}}(\widehat{X}_1 - \widehat{Y}_2)$ if the true correlation $\rho$ is close to 1. There are a number of reasons for which we should not feel concerned by this potential problem in repeated survey sampling.

1. When both samples are not equal, for example when there is a defined rotation rate, and finite population correction terms are not negligible, the correlation between Horvitz-Thompson estimators $\widehat{X}_1$ and $\widehat{Y}_2$ can never be close to 1. In the case of simple random sampling, for example, it is just not possible.

2. When there is no rotation, the correlation can take values close to 1. But when there is no or very few rotation, estimating a correlation coefficient on $s_{12}$ or estimating $S_{xy}$ directly on $s_{12}$ makes no difference. It is to be noted also that usual estimators of variance perform pretty poorly and that it is always difficult to estimate a variance close to zero using variables that can take large values.

### Notations and description of the problem

Let $s_1$ and $s_2$ be two samples selected in a population $U$ with a joint distribution $p(s_1, s_2)$. Define the marginal sampling designs $p_1(\cdot)$ and $p_2(\cdot)$,

and note $s_{12}$ the intersection of these samples. Resulting sampling design $p_{12}(\cdot)$ for $s_{12}$ is known in the cases we will consider. Let also $\widehat{X}_1$ and $\widehat{Y}_2$ be estimators of the total of a variable $X$, using sample $s_1$ and of a variable $Y$ using sample $s_2$, with variances $\text{var}_1(\widehat{X}_1)$, $\text{var}_2(\widehat{Y}_2)$, and adapted estimators $\widehat{\text{var}}_1(\widehat{X}_1)$, $\widehat{\text{var}}_2(\widehat{Y}_2)$. Finally, let $\widehat{X}_{12}$ and $\widehat{Y}_{12}$ denote estimators based on $s_{12}$, with variances and covariance $\text{var}_{12}(\widehat{X}_{12})$, $\text{var}_{12}(\widehat{Y}_{12})$, $\text{cov}_{12}(\widehat{X}_{12}, \widehat{Y}_{12})$ given by $p_{12}(\cdot)$, and $\text{cov}(\cdot,\cdot)$ denote the covariance operator given by $p(\cdot,\cdot)$.

Berger (2004b) attributes to Kish (see Kish, 1965, p.457, but I could not find where) the idea of using

$$\widehat{\text{cov}}_{wr}(\widehat{X}_1, \widehat{Y}_2) = \widehat{\rho}_{12}\left[\widehat{\text{var}}_{wr,1}(\widehat{X}_1)\widehat{\text{var}}_{wr,2}(\widehat{Y}_2)\right]^{\frac{1}{2}}, \tag{5.17}$$

as an estimator of $\text{cov}(\widehat{X}_1, \widehat{Y}_2)$, where

$$\widehat{\rho}_{12} = \frac{\widehat{\text{cov}}_{wr,12}(\widehat{X}_{12}, \widehat{Y}_{12})}{\left[\widehat{\text{var}}_{wr,12}(\widehat{X}_{12})\widehat{\text{var}}_{wr,12}(\widehat{Y}_{12})\right]^{\frac{1}{2}}}$$

is a correlation coefficient estimated only on $s_{12}$. One advantage of estimator 5.17 is that, if estimators $\widehat{\text{var}}_{wr,1}$ and $\widehat{\text{var}}_{wr,2}$ are used for the variances of $\widehat{X}_1$ and $\widehat{Y}_2$, it leads to non-negative variance estimators for linear combinations of $\widehat{X}_1$ and $\widehat{Y}_2$. Simulations in Berger (2004b) show that this estimator does not perform well, and that it appears to lead to a bias consistent with 5.16.

### Discussion

One remark we should make is that an estimator $C(\widehat{X}_1, \widehat{Y}_2)$ defined as

$$C(\widehat{X}_1, \widehat{Y}_2) = \widehat{\rho}_{12}\left[\widehat{\text{var}}_1(\widehat{X}_1)\widehat{\text{var}}_2(\widehat{Y}_2)\right]^{\frac{1}{2}}, \tag{5.18}$$

where

$$\widehat{\rho}_{12} = \frac{\widehat{\text{cov}}_{12}(\widehat{X}_{12}, \widehat{Y}_{12})}{\left[\widehat{\text{var}}_{12}(\widehat{X}_{12})\widehat{\text{var}}_{12}(\widehat{Y}_{12})\right]^{\frac{1}{2}}}$$

is a correlation coefficient estimated only on $s_{12}$, is clearly a very bad estimator of $\text{cov}(\widehat{X}_1, \widehat{Y}_2)$. Indeed, we can see in example 5.1 that, even in the simplest case, estimator 5.1 makes a total mess of finite population corrections and of sample sizes.

**Example 5.1** *With simple random sampling, as we have seen in Chapter 5, the covariance between Horvitz-Thompson estimators is equal to*

$$\text{cov}(\widehat{X}_1, \widehat{Y}_2) = N^2\left(\frac{n_{12}}{n_1 n_2} - \frac{1}{N}\right)\rho_{xy}S_x S_y,$$

*where $S_x$, $S_y$ and $\rho_{xy} = S_{xy}/S_x S_y$ are standard errors and correlation of variables $X$ and $Y$ in the population, $n_{12} = |s_{12}|$, $n_1 = |s_1|$ and $n_2 = |s_2|$. Consequently,*

*the correlation between* $\widehat{X}_1$ *and* $\widehat{Y}_2$ *is equal to*

$$\operatorname{corr}(\widehat{X}_1, \widehat{Y}_2) = \rho_{xy} \frac{\left(\frac{n_{12}}{n_1 n_2} - \frac{1}{N}\right)}{\left[\left(\frac{1}{n_1} - \frac{1}{N}\right)\left(\frac{1}{n_2} - \frac{1}{N}\right)\right]^{\frac{1}{2}}},$$

*and, since* $n_{12} \leq \min(n_1, n_2)$*, it cannot be close to* 1 *if* $n_1$*,* $n_2$ *and* $n_1 n_2 / n_{12}$ *are not equal. It would be a bad idea to estimate directly* $\operatorname{cov}(\widehat{X}_1, \widehat{Y}_2)$ *with estimator 5.18. Indeed, in that case 5.18 gives*

$$C(\widehat{X}_1, \widehat{Y}_2) = N^2 \left[\left(\frac{1}{n_1} - \frac{1}{N}\right)\left(\frac{1}{n_2} - \frac{1}{N}\right)\right]^{\frac{1}{2}} \rho_{xy,s_{12}} s_{x,s_1} s_{y,s_2},$$

*where*

$$\rho_{xy,s_{12}} = \frac{s_{xy,s_{12}}}{s_{x,s_{12}} s_{y,s_{12}}}.$$

*Finite population correction as well as sample size terms are clearly wrong in this expression of* $C(\widehat{X}_1, \widehat{Y}_2)$*. It is however legitimate to estimate* $S_{xy}$ *with*

$$\widehat{S}_{xy} = \rho_{xy,s_{12}} s_{x,s_1} s_{y,s_2},$$

*and to use it in an estimator of* $\operatorname{cov}(\widehat{X}_1, \widehat{Y}_2)$*:*

$$\widehat{\operatorname{cov}}(\widehat{X}_1, \widehat{Y}_2) = N^2 \left(\frac{n_{12}}{n_1 n_2} - \frac{1}{N}\right) \rho_{xy,s_{12}} s_{x,s_1} s_{y,s_2}.$$

Estimator 5.1 is asymptotically unbiased and simulations let us believe that it performs honorably, or at least as well as a direct estimator on $s_{12}$. For some other bidimensional sampling designs, such as those discussed in Chapter 5, we can also use ratio type estimators of the covariance. A correlation is estimated on the samples overlap and multiplied by standard errors estimated on samples $s_1$ and $s_2$, but corrective terms that are functions of the sample size are used. This is particularly easy with viable approximations of $\operatorname{cov}_s(\widehat{X}_s, \widehat{Y}_s)$ that depend only on the first order inclusion probabilities (on this subject, see Hájek, 1964; Deville, 1993, 1999; Brewer, 2002; Brewer & Donadio, 2003). These approximations have in common that they can be seen as estimators of correlations of reweighted variables $\widetilde{X}$ and $\widetilde{Y}$ in the population. The correlation and standard errors involved here are not directly correlation on $s_{12}$ of $\widehat{X}_{12}$ and $\widehat{Y}_{12}$ and standard errors of $\widehat{X}_1$ and $\widehat{Y}_2$, but correlation and standard errors of $\widetilde{X}$ and $\widetilde{Y}$ in the population. Simulations in Chapter 5, Section 5.5 did not reveal dramatic flaws with this kind of estimator either.

# Coordinated Poisson sampling

6

**Abstract**

The Swiss Federal Statistical Office intends to manage all its business surveys with a single system of sample coordination. This system should provide Poisson transversal sampling designs. It should enable us to coordinate positively or negatively one time surveys, but also panels and rotating panels, with an optimal coordination if possible. Finally, it must be adequate for a dynamic population in which births and deaths of units take place, and in which units can also split or merge as is quite common in business surveys. We generalized Brewer's method of coordination and detailed how the system can be used to answer these needs in the case of a dynamic population.

**Keywords:** Rotating panels, Dynamic population, Brewer's method

## Introduction

Several methods of sampling coordination have already been developed in national institutes of statistics (see for example in France Cotton & Hesse, 1992; Rivière, 2001). There is a review of some of them by Hesse (1999) and a more detailed description of their properties in Nedyalkova et al. (2009). Each of these methods performs perfectly in the ideal case of a static population and of coordinations that are all positive or all negative. And for each one of them, solutions have been developed to adapt to dynamic populations. There remain however some problems that make these methods unsuitable for a global system of coordination of all surveys conducted by an institute. In particular, it seems difficult with these methods to mix positive and negative coordination as is required when one wants to have two separate rotating panels for example. Moreover, most of these methods provide simple random with fixed size or stratified transversal sampling designs, at least approximately. They require difficult adaptations to allow for deaths and births in the population while still giving such transversal designs. Methods that provide stratified samples often necessitate to create the strata once and for all such as in De Ree (1983), or

to use the intersection of all strata, as in Rivière (2001). This last problem also makes these methods unpractical for an institute that wants to have a consistent and durable system of sample coordination.

Most of the technical problems raised by existing coordination methods come from their stratified or fixed size transversal sampling designs. Using Poisson sampling for transversal designs as in Brewer et al. (1972) allows to create a much more flexible system. The birth (respectively death) of a unit simply translates to its inclusion probability becoming positive (respectively null) without requiring an intervention on the inclusion probabilities or selection process of other units. Strata are replaced by domains in which inclusion probabilities may be equal, or in which the sum of inclusion probabilities may be chosen so as to obtain a sufficient size of sample with a high probability. The sample size is random, but it is also the case when there is non-response and, even with stratified sample, one has to chose a sufficient size of sample within strata or domains of interest to compensate for non-response. The main drawback remains that variance of the estimators of variables proportional to inclusion probabilities does not benefit as much as with fixed size sampling. That can be mostly compensated by the use of a calibrated estimator, and variance-covariance estimation between samples is largely simplified. Most importantly, there just does not exist yet a coordination system that satisfies every requirements asked of a general system for all business surveys and allows to use fixed size or modern transversal sampling designs that take better advantage of auxiliary information.

## 6.1 Survey burden

Each year, the Federal Statistical Office (FSO) conducts several business surveys. Some units are selected on multiple occasions. Sometimes it cannot be avoided, for example for large companies that receive an inclusion probability equal to one in every survey. It is nevertheless desirable to limit as much as possible the survey burden of smaller companies, and to guarantee that an unit is not selected more often than necessary. There are two aspects to survey burden: the number of occasions on which an unit is selected and the time between two selections. The average number of selection depends only on the first order inclusion probabilities and it is usually not possible to adjust it in order to diminish the survey burden. Indeed, inclusion probabilities are computed on each occasion so as to obtain a good precision for every transversal survey. A unit that receives inclusion probabilities $\pi^1, \ldots, \pi^r$ for surveys over a time period will be selected on average $\pi^1 + \cdots + \pi^r$ over this period, independently from the joint sampling design of the $r$ surveys. The only way to reduce this figure is to conduct fewer surveys, or to sacrifice the precision of some of those and this is usually not acceptable.

However, for surveys that are conducted on different occasions, one can try to ascertain that a unit that has just been selected will be left alone during a certain period. The regularity with which an unit is selected can be controlled with an adequate coordinated sampling method. For

example, if five surveys are to be conducted and an unit receives inclusion probability 0.2 at each of these surveys, it will be selected on average one time. The naive method would be to draw each sample independently, leading this unit to be selected 0 time, 1 time, et cœtera, and up to five times. By using a coordinated sampling design, we can make sure that this unit will be selected exactly one time, and that once it has been selected, it will not be selected again. In the general case, when the sum of inclusion probabilities of a given unit is between two natural numbers $j$ and $j+1$, negatively coordinated sampling leads to selecting this unit on either $j$ or $j+1$ occasions, while independent sampling would have resulted in any possible number of selections. We studied some of the existing coordination methods in Nedyalkova et al. (2009) and found out that they all lead approximately to the same time out of sample for a unit that has just been selected. It can be argued that the best longitudinal design for repeated surveys with negative coordination is systematic sampling, which is the only design that will give a perfectly regular time out of sample in an ideal case.

Coordinated sampling is especially interesting for units that have a sum of inclusion probabilities over time not greater than one since in that case one can guarantee them that they will be selected on at most one occasion. The coordinated sampling system developed in this paper makes it possible to organize any type of survey (on one occasion, on many occasions with a panel or with a rotating panel), in a dynamic population (with births, deaths, merges and splits), and to guarantee that each unit, on an individual level, will be selected as sparsely as possible under the constraints imposed by their inclusion probabilities. The resulting longitudinal design for each unit is the systematic design. In some cases, there is but a slight difference between independent and coordinated samples. For example, when inclusion probabilities are very small independent samples will naturally have no overlap. In other cases, the effect of sampling coordination will be obvious. For example, when there is a survey with a high sampling rate and another one with a small sampling rate (one can think of sampling rates of 0.9 and 0.1), independent sampling will lead to samples that have a very big overlap while coordinated sampling could have produced non-overlapping samples.

## 6.2 Method

We describe here a method that is a generalization of the method in Brewer et al. (1972) and that allows to select samples with a negative or positive coordination with previous surveys for every unit in the population. It is based on the use of permanent random numbers, allows for births and deaths in the population, gives Poisson cross-sectional designs and systematic longitudinal designs. Units are treated independently one from the other, and they may receive different coordination rules, though we do not see any use for different units having different coordinations. Finally, a first implementation using the SAS® software shows that this method is practical and requires only a few minutes of computation to coordinate

thirty surveys in a population of 400′000 units (the number of companies in Switzerland).

Let us recall shortly some useful definitions for sampling on many occasions: we are interested in drawing samples from a population $U^t$ at times $t = 1, 2, \ldots, T$. At time $t$, a sample without replacement is a subset of the population $U^t$. Without loss of generality, we can define $U = \{1, \ldots, N\}$ as the union of populations $U^t$, where we identify all occurrences of a same unit and we assign an index to any unit that appears in $U^t$, $t = 1, 2, \ldots, T$ on at least one occasion. It is often useful to define "existence indicators", but in this case it is not necessary as a non living unit will naturally receive a null inclusion probability and everything will work out nicely.

**Definition 6.1** *A cross-sectional sample is denoted by a vector*

$$\mathbf{s}^t = (s_1^t, \ldots, s_k^t, \ldots, s_N^t)' \in \{0,1\}^N,$$

*for all $t \in \{1, 2, \ldots, T\}$, and the longitudinal sample by a vector*

$$\mathbf{s}_k = (s_k^1, \ldots, s_k^t, \ldots, s_k^T)' \in \{0,1\}^T,$$

*where*

$$s_k^t = \begin{cases} 1 & \text{if, at time } t, \text{unit } k \text{ is in the sample } \mathbf{s}^t \\ 0 & \text{if, at time } t, \text{unit } k \text{ is not in the sample } \mathbf{s}^t, \end{cases}$$

*for all $k \in U$.*

**Definition 6.2** *A sampling design $p(\mathbf{s}^t)$, $t = 1, 2, \ldots, T$, will be called a cross-sectional sampling design.*

**Definition 6.3** *A sampling design $p(\mathbf{s}_k)$, $k = 1, 2, \ldots, N$, will be called a longitudinal sampling design.*

The joint (or complete) sampling design $p(\mathbf{s})$ is given by

$$p(\mathbf{s}) = p(\mathbf{s}^1, \ldots, \mathbf{s}^t, \ldots, \mathbf{s}^T).$$

From this joint sampling design, we can derive the marginal cross-sectional design for a time $t$

$$p_t(\mathbf{s}^t) = \sum_{\mathbf{s}^1, \ldots, \mathbf{s}^{t-1}, \mathbf{s}^{t+1}, \ldots, \mathbf{s}^T} p(\mathbf{s}^1, \ldots, \mathbf{s}^t, \ldots, \mathbf{s}^T),$$

and the marginal longitudinal design for a unit $k$,

$$p_k(\mathbf{s}_k) = \sum_{\mathbf{s}_1, \ldots, \mathbf{s}_{k-1}, \mathbf{s}_{k+1}, \ldots, \mathbf{s}_N} p(\mathbf{s}_1, \ldots, \mathbf{s}_k, \ldots, \mathbf{s}_N).$$

Let $S_k^t$ be the random variable that takes the value 1 if unit $k$ is selected at time $t$ and 0 otherwise. The first-order inclusion probabilities of the cross-sectional design at time $t$ are given by:

$$\pi_k^t = \mathrm{E}(S_k^t),$$

where $\mathrm{E}(\cdot)$ is the expectation under the probability distribution $p(\cdot)$, $k \in U, t = 1, \ldots, T$. The longitudinal joint inclusion probabilities for times $t$ and $i$ are given by:

$$\pi_k^{t,i} = \mathrm{E}(S_k^t S_k^i),\ k \in U, t, i = 1, \ldots, T.$$

Naturally, if a unit $k$ does not belong to $U^t$, then $S_k^t = 0$ and the inclusion probabilities $\pi_k^t$ and $\pi_k^{t,i}$ are also null.

When selection of unit $k$ at times $t$ and $i$ are uncorrelated, as is the case when cross-sectional sampling designs $p_t(\mathbf{s}^t)$ and $p_i(\mathbf{s}^i)$ are independent, we have that $\pi_k^{t,i} = \pi_k^t \pi_k^i$. We will say that there is a positive coordination between surveys $i$ and $t$ for unit $k$ when $\pi_k^{t,i} > \pi_k^t \pi_k^i$ and negative coordination when $\pi_k^{t,i} < \pi_k^t \pi_k^i$. Remark that the following inequality always holds:

$$\max(0, \pi_k^t + \pi_k^i - 1) \leq \pi_k^{t,i} \leq \min(\pi_k^t, \pi_k^i). \tag{6.1}$$

We will say that there is an optimal negative or positive coordination when one of the bounds of 6.1 is reached. This definition of correlation is not the only one possible. For sampling designs with simple random or stratified cross-sectional designs it is not uncommon to define negative (resp. positive) coordination as the property that the intersection of samples holds less (resp. more) units than would have been the case with independent sampling.

We will also give definitions of weakly and strictly sequential (longitudinal) sampling algorithms. A strictly sequential procedure may be necessary for the longitudinal design when we are sampling over time. This is the case when the inclusion probabilities for the future occasions are not known (e.g. they are proportional to a variable that is not available in advance), or when the total number of occasions is not known. The usual algorithm for systematic sampling is strictly sequential.

**Definition 6.4** *A longitudinal sampling algorithm, for a unit k, is said to be weakly sequential if at step $t = 1, \ldots, T$ of the procedure, the decision concerning whether the unit k is in the sample $\mathbf{s}^t$ is definitively taken.*

**Definition 6.5** *A longitudinal sampling algorithm is said to be strictly sequential if it is weakly sequential and if the decision concerning the unit k at time t does not depend on the inclusion probabilities of the unit k at times $t+1, \ldots, T$ and on the number T of sampling occasions.*

Our system requires that an order of priority among the coordinations be defined, before each new sample is selected. One can for example chose to give priority to the coordination with the preceding survey, then with the one before, and so on. The sign of coordination is chosen freely by the user and can be positive with some past surveys and negative with others. For a repeated survey, the updated sample can be positively coordinated with previous samples of the same survey and negatively coordinated with other surveys.

### 6.2.1 Description of the method

Each unit is treated independently. We will describe our coordination method for one unit. Assume for example that one wants to coordinate three surveys and that a unit $k$ receives inclusion probabilities $\pi_k^1$, $\pi_k^2$, $\pi_k^3$.

- On the first sampling occasion, the unit has an inclusion probability $\pi_k^1$ and receives a permanent random number $u_k$ uniformly generated in $[0,1]$. The first sample is selected using the usual procedure for Poisson sampling: unit $k$ is selected when $u_k \leq \pi_k^1$ and is not selected otherwise. Line segment $[0,1]$ is thus divided in two subsets one of which can be trivial if $\pi_k^1$ equals 0 or 1, and $[0, \pi_k^1]$ is the selection zone for unit $k$ in sample $\mathbf{s}^1$ (see Figure 6.1).

Figure 6.1 – *First sampling occasion*

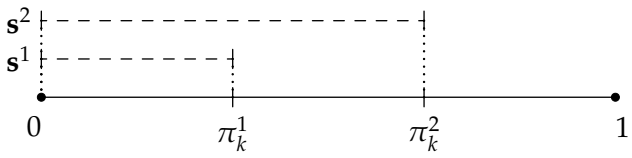

- On the second sampling occasion, coordination is obtained as in Brewer et al. (1972): a selection zone is defined for unit $k$ in sample $\mathbf{s}^2$. This zone consists of one or two intervals with total length $\pi_k^2$. If we want to obtain an optimal positive coordination for unit $k$ between these two sampling occasions, we will chose a selection zone that will have a maximum overlap with the selection zone of the first sampling occasion. If on the contrary we want to have an optimal negative coordination, we will chose a selection zone that has no overlap, if possible (if $\pi_k^1 + \pi_k^2 \leq 1$), with the selection zone of the first sampling occasion, or that will have the smallest possible overlap. For example, if $\pi_k^2 \leq \pi_k^1$ and we want to have a positive coordination, the selection zone at the second occasion is included in the selection zone at the first occasion (see Figure 6.2). If $\pi_k^2$ is

Figure 6.2 – *Positive coordination when* $\pi_k^2 \leq \pi_k^1$


greater than $\pi_k^1$, we have the situation of Figure 6.3.

The selection of unit $k$ is determined by the inclusion of $u_k$ to selection zones and each intersection of selection zones corresponds to a

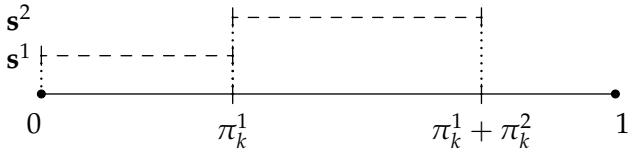
Figure 6.3 – *Positive coordination when* $\pi_k^2 \geq \pi_k^1$




longitudinal sample for unit $k$. After the second sampling occasion, $[0,1]$ is split into three intervals that correspond to longitudinal samples $(1,1),(1,0),(0,0)$ in the case of Figure 6.2 and $(1,1),(0,1),(0,0)$ in the case of Figure 6.3.

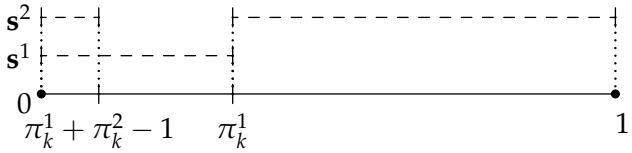
If we want to obtain a negative coordination for unit $k$ between sampling occasions, we need to chose a selection zone for the second occasion that has the smallest possible overlap with the selection zone of the first sampling occasion and thus that is preferably included in $[\pi_k^1, 1]$, if $\pi_k^1 + \pi_k^2 \leq 1$ (see Figure 6.4).

Figure 6.4 – *Negative coordination if* $\pi_k^1 + \pi_k^2 \leq 1$




If $\pi_k^1 + \pi_k^2 \geq 1$ the selection zone on the second sampling occasion will be the union of $[\pi_k^1, 1]$ and of $[0, \pi_k^2 + \pi_k^1 - 1]$ (see Figure 6.5).

Figure 6.5 – *Negative coordination if* $\pi_k^1 + \pi_k^2 \geq 1$




Here again $[0,1]$ is split into three intervals, each of these corresponds to a longitudinal sample for unit $k$: $(1,0),(0,1),(0,0)$ in the case of Figure 6.4 and $(1,1),(1,0),(0,1)$ in the case of Figure 6.5.

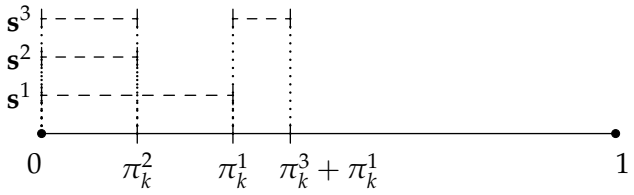
- Assume that after $t$ sampling occasions $[0,1]$ is split into $t+1$ intervals, each one corresponding to a longitudinal sample. In order to define a selection zone for unit $k$ in sample $\mathbf{s}^{t+1}$, we assign a score to these $t+1$ intervals in a way that is determined by the coordination signs and priorities with past surveys. If sample $\mathbf{s}^{t+1}$ is to be coordinated positively with $\mathbf{s}^i$, with maximum priority, every interval

| Interval | $\mathbf{s}_k^1$ | $\mathbf{s}_k^2$ | ... | $\mathbf{s}_k^t$ |
|---|---|---|---|---|
| $a_k^1$ | 1 | 0 | ... | 0 |
| $a_k^2$ | 1 | 1 | ... | 0 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| $a_k^{t+1}$ | 0 | 0 | ... | 1 |

Table 6.1 – *Sampling design for unit k*

that corresponds to a longitudinal sample where unit $k$ is selected in $\mathbf{s}^i$ will receive a higher score than all intervals where unit $k$ is not selected in $\mathbf{s}^i$. Then, inside these two groups, intervals will receive scores sorted according to the coordination desired with the second sampling occasion by order of priority, and so on. The selection zone for unit $k$ in sample $\mathbf{s}^{t+1}$ will be the union of intervals with the highest scores up to a total length no greater than $\pi_k^{t+1}$ and of a subinterval of the next interval by decreasing score, up to a total length of $\pi_k^{t+1}$. For example, if the first two sampling designs are positively coordinated, as in Figure 6.2 and sample $\mathbf{s}^3$ is to be, in priority, positively coordinated with sample $\mathbf{s}^2$ then negatively coordinated with sample $\mathbf{s}^1$, we obtain Figure 6.6.

Figure 6.6 – *Coordination of a third sample*




For each unit $k$ a list of intervals $a_k^1, \ldots, a_k^{t+1}$ that form a partition of $[0, 1]$ must be created and updated, and to each interval $a_k^i$ is associated a longitudinal sample $\mathbf{s}_{k,a_k^i}$. These data allow to effectively select a sample, that will be determined by the permanent random number $u_k$. They are necessary in order to compute the joint sampling design $p(\mathbf{s}^1, \ldots, \mathbf{s}^{t+1})$ on the next sampling occasion. Table 6.1 gives an idea of the data that must be kept for each unit. In that table $\sum_{i=1}^{t+1} a_k^i = 1$ ($a_k^i$ is used indifferently to represent an interval or its length), and a random number $u_k$ must also be stored. Note that longitudinal sample $\mathbf{s}_{k,a_k^i}$ can be read in Table 6.1 as the $t$ last columns of row $i$.

### 6.2.2 Births and deaths in the population

The method we described can easily be adapted to a dynamic population where births and deaths occur. Indeed, units are treated independently and the birth or death of a unit does not modify the selection process of

other units (other than maybe in the computation of inclusion probabilities) as would be the case, for example, with fixed size simple random sampling. Moreover, the longitudinal sampling design is strictly sequential, hence the selection of a unit at a given time does not depend on its future characteristics or on its remaining lifespan. In order to add a new unit to the sampling frame, it is sufficient to assign to this unit null inclusion probabilities for past surveys, and fictitious intervals of length 0, corresponding to fictitious longitudinal samples, that may be defined to be equal to $(0, \ldots, 0)$. This unit will receive a first non trivial interval on the first sampling occasion when it has a non null inclusion probability. In order to take into account the death of a unit, it is sufficient to assign to it a null inclusion probability at all future sampling occasions.

Algorithm 4 describes the main procedure of our coordination method. Parameters $o_k^j$, $j = 1, \ldots, t$ give the priority of coordination of survey $j$

**Algorithm 4**: Coordination of Poisson sampling designs

$t = 1$: *Initialization*
**for** each unit $k$ in the population $U^1$ **do**
  Define $a_k^1 = \pi_k^1$, $a_k^2 = 1 - \pi_k^1$;
  Define $\mathbf{s}_k^{a_1} = 1$, $\mathbf{s}_k^{a_2} = 0$;
  Draw a random number $u_k$ uniformly in $[0, 1]$;
**end for**
$t \to t + 1$: *Addition of a sampling occasion*
**Require:** Coordination rules $o_k^j$, $c_k^j$, $j = 1, \ldots, t$;
Define $U = \bigcup_{r=1}^{t+1} U^r$;
Every unit in $U \backslash U^{t+1}$ receives inclusion probability 0;
**for** every new unit $k$, in $U^{t+1} \backslash \bigcup_{r=1}^{t} U^r$, **do**
  Define $a_k^1 = \ldots = a_k^t = 0$, $a_k^{t+1} = 1$;
  Define $\mathbf{s}_k^{a_1} = \ldots = \mathbf{s}_k^{a_{t+1}} = (0, \ldots, 0) \in \mathbb{R}^t$;
  Draw a random number $u_k$ uniformly in $[0, 1]$;
**end for**
**for** each unit $k$ in $U$ **do**
  **for** $i = 1$ to $t + 1$ **do**
    Compute a score $\sigma_k^i$ for $a_k^i$ as $\sigma_k^i = \sum_{j=1}^{t} 2^{o_k^j} c_k^j(\mathbf{s}_{k,a_k^i})$;
  **end for**
  Define selection zone for unit $k$ in $\mathbf{s}^{t+1}$ as the union of intervals $a_k^i$ that have highest scores until their total length exceeds $\pi_k^{t+1}$;
  Split the last added interval into two parts and remove a part from the selection zone so that its total length is equal to $\pi_k^{t+1}$;
  Renumber the list of intervals $a_k^1, \ldots, a_k^{t+2}$ and update the corresponding longitudinal samples $\mathbf{s}_k^1, \ldots, \mathbf{s}_k^{t+2}$.
**end for**
*Selection of a longitudinal sample*
**for** each unit $k$ in $U$ **do**
  Select the longitudinal sample corresponding to the interval $a_k^i$ in which permanent random number $u_k$ lies.
**end for**

with survey $t+1$ ($o_k^j = t$ means highest priority and $o_k^j = 1$ means lowest priority), and $c_k^j(\mathbf{s}_{k,a_k^i})$ is defined as

$$c_k^j(\mathbf{s}_{k,a_k^i}) = \begin{cases} \mathbf{s}^j_{k,a_k^i} & \text{if the desired coordination of survey } j \\ & \text{with survey } t+1 \text{ is positive,} \\ 1 - \mathbf{s}^j_{k,a_k^i} & \text{if the desired coordination of survey } j \\ & \text{with survey } t+1 \text{ is negative,} \end{cases}$$

where $\mathbf{s}^j_{k,a_k^i}$ is the $j^{th}$ line of $\mathbf{s}_{k,a_k^i}$. Function $c_k^j(\mathbf{s}_{k,a_k^i})$ is the indicator variable that, in the longitudinal sample corresponding to interval $a_k^i$, unit $k$ is selected in $\mathbf{s}^j$ if the desired coordination is positive, or on the contrary that unit $k$ is not selected in $\mathbf{s}^j$ if the desired coordination is negative.

### 6.2.3 Merging and splitting units

In dynamic populations, and particularly in dynamic populations of companies, it is not uncommon for two (or more) units to split or merge. Having units that split into two (or more) units, does not create new problems for coordinated sampling. One just has to decide if the new units inherit the characteristics of the parent unit or if they are considered as newly born units. In any case, the units receive a consistent 'past', i.e. intervals, longitudinal samples and selection zones consistent with the inclusion probabilities of their parent and with the observed data (which may or may not be split between the new units).

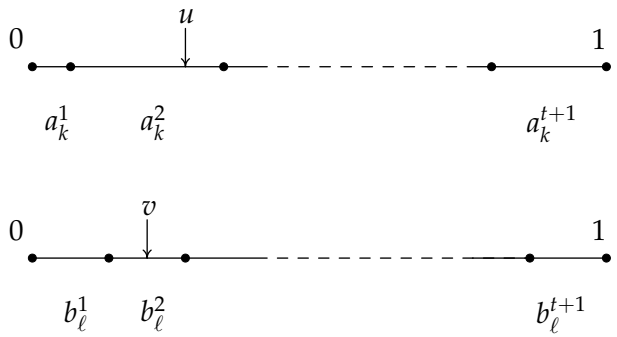
The case of merging units is much more difficult as two units that merge have their own different pasts and may belong to different panels, so we need to chose a fictitious past for the merger unit, depending on the way we want this unit to be sampled in the future. Here are some possible choices:

- the merger unit is considered as a newly born unit. The past of merging units is discarded and everything goes as if they were deceased.
- One of the merging units is considered to be dominant (e.g. it is much bigger than the other units). It is then natural to affect the past of this unit to the merger.
- The merger inherits characteristics of two or more merging units. For example if units $k$ and $\ell$ with selection indicators $s_k^t$ and $s_\ell^t$ merge in a new unit $m$, we need to decide how we will use information $(s_k^t, s_\ell^t)$ to recreate a past for $m$.

The last case is the only one that requires further developments for our problem of sample selection and coordination. Two units have to be merged with their longitudinal sampling designs and their permanent random numbers. Let us consider unit $k$ with intervals $a_k^1, \ldots, a_k^{t+1}$, random number $u$ and unit $\ell$ with intervals $b_\ell^1, \ldots, b_\ell^{t+1}$ and random number $v$. The longitudinal marginal sampling designs of units $k$ and $\ell$ can be

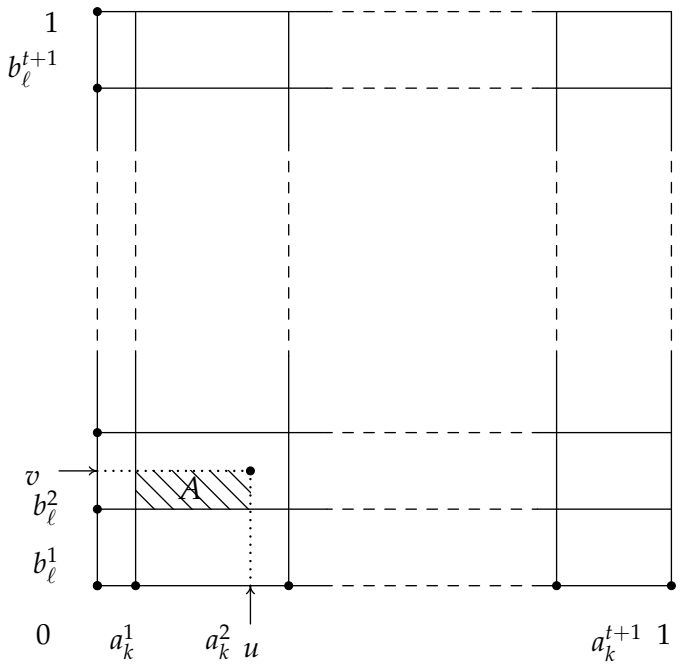
represented as in Figure 6.7. To each interval $a_k^i$ corresponds a longitudinal

Figure 6.7 – *Merging two units: marginal sampling designs*




sample $\mathbf{s}_{k,a_k^i}$ for unit $k$ and to each interval $b_\ell^i$ corresponds a longitudinal sample $\mathbf{s}_{\ell,b_\ell^i}$ for unit $\ell$. Since these sampling designs are independent, couples $(\mathbf{s}_{k,a_k^i}, \mathbf{s}_{\ell,b_\ell^i})$ are in bijection with rectangles $a_k^i \times b_\ell^j$ of Figure 6.8. The probability of selecting a couple of longitudinal samples $(\mathbf{s}_{k,a_k^i}, \mathbf{s}_{\ell,b_\ell^i})$ is equal to the area of the corresponding rectangle.

Figure 6.8 – *Merging two units: joint sampling design*




Our aim is to suppress differences between merger units and 'standard' units of the population: we want them to have at most $t+1$ possible longitudinal samples at time $t$, coupled with intervals $(c_j)_{j=1,\dots,t+1}$ and a permanent random number.

A first and painless step is to transform the square $[0,1]^2$ and its subdivision in rectangles of Figure 6.8 into line segment $[0,1]$ and an adequate subdivision. That can be done by reporting areas of rectangles $a_k^i \times b_\ell^j$ on $[0,1]$ in any chosen order, and the couple of random numbers $(u,v)$ can be placed in the corresponding interval. If it falls into interval $[a,b]$, we can chose the permanent random number to be $w = a + A$ where $A$ is the area drawn in Figure 6.8. That way, we have a partition of $[0,1]$ into $(t+1)^2$ intervals, each of them corresponding to couples of longitudinal samples $(\mathbf{s}_k, \mathbf{s}_\ell)$, and a random number $w$ that falls into the interval that corresponds to the couple of longitudinal samples that where effectively selected.

A second and necessary step in order to insert the merger unit $m$ in the coordination system and compute scores as in Algorithm 4 is to map couples $(\mathbf{s}_k, \mathbf{s}_\ell)$ to vectors $\mathbf{s}_m \in \{0,1\}^t$. This transformation implies a loss of information but is standard practice when units merge in business surveys. We define a merging function

$$f_{k,\ell} = \begin{pmatrix} \{0,1\}^t \times \{0,1\}^t & \longrightarrow & \{0,1\}^t \\ (\mathbf{s}_k, \mathbf{s}_\ell) & \longmapsto & \mathbf{s}_m \end{pmatrix}.$$

Usual merging functions are:

- $f_{k,\ell}(\mathbf{s}_k, \mathbf{s}_\ell) = \mathbf{s}_k$,
- $f_{k,\ell}(\mathbf{s}_k, \mathbf{s}_\ell) = \mathbf{s}_\ell$,
- $f_{k,\ell}(\mathbf{s}_k, \mathbf{s}_\ell) = (\max(s_k^1, s_\ell^1), \ldots, \max(s_k^t, s_\ell^t))'$,
- $f_{k,\ell}(\mathbf{s}_k, \mathbf{s}_\ell) = (\min(s_k^1, s_\ell^1), \ldots, \min(s_k^t, s_\ell^t))'$.

A posteriori inclusion probabilities can be computed for unit $m$, that depend on the choice of merging function $f_{k,\ell}$. With the usual functions given above, we get respectively: $\pi_m^i = \pi_k^i$, $\pi_m^i = \pi_\ell^i$, $\pi_m^i = \pi_k^i + \pi_\ell^i - \pi_k^i \cdot \pi_\ell^i$, and $\pi_m^i = \pi_k^i \cdot \pi_\ell^i$.

At this point, we could coordinate the selection of unit $m$ in future surveys with past surveys using Algorithm 4 and all $(t+1)^2$ longitudinal samples for unit $m$. However, that would make the system unstable since, if several units merge, the amount of computation and of data to be stored grows rapidly.

The third and optional step is to map the $(t+1)^2$ longitudinal samples to at most $t+1$ samples. Once this is done, the merger unit will fit in the system exactly as a 'standard' unit does. Note that for some units there is in fact nothing to do. Big companies, for example, are always selected, and if such a unit merges with another unit, it will naturally have at most $t+1$ possible longitudinal samples. For other units we need to define a compression function $g_m$:

$$g_m = \begin{pmatrix} \{0,1\}^t & \longrightarrow & \{0,1\}^t \\ \mathbf{s}_m & \longmapsto & \tilde{\mathbf{s}}_m \end{pmatrix}$$

that takes at most $t+1$ values when $\mathbf{s}_m$ is taken in the set of $(t+1)^2$ possible longitudinal samples given by $f_{k,\ell}$. This implies an additional

loss of information that must be carefully controlled. If all coordinations were negative and the only objective was to maximize the time out of sample, we could keep for only information the last occasion on which unit $m$ has been selected:

$$g_m : \mathbf{s}_m = (s_m^1, \ldots, s_m^{i-1}, 1^i, 0, \ldots, 0) \mapsto \tilde{\mathbf{s}}_m = (0, \ldots, 0, 1^i, 0 \ldots, 0).$$

Unfortunately, this simple solution is not adequate when we need to coordinate panels or rotating panels, for which it is not sufficient to keep only the last time of selection. Hence we need to make a better choice for the compression function $g_m$. Some of the important aspects that we should to consider, are:

1. keep the longitudinal sample $\mathbf{s}_m(w)$ that was actually selected. In that way, the available data for unit $m$ is consistent with the history of selections of $m$, modulo the function $f_{k,\ell}$.

2. Preserve the a posteriori inclusion probabilities $\pi_m^1, \ldots, \pi_m^t$ that are given by $f_{k,\ell}$, so that the resulting coordination is coherent with the choice expressed in $f_{k,\ell}$.

3. Select a function $g_m$ that can be implemented: enumeration problems of vertices of polytopes in $[0,1]^t$ can be complex and lead to heavy computation burden, when they are at all feasible.

4. Once the above points are respected, chose a method that keeps the most information, such as

   - choose if possible the length of the interval to which $w$ belongs, to be for example equal to $p(\mathbf{s}_m(w))$,
   - choose a function $g_m$ which takes exactly $t+1$ values,
   - use an information criterion.

It is theoretically possible to preserve the inclusion probabilities $\pi_m^1, \ldots, \pi_m^t$, the longitudinal sample $\mathbf{s}_m(w)$ and to use $t$ other longitudinal samples $\mathbf{s}_m^{i_1}, \ldots, \mathbf{s}_m^{i_t}$ among those that were given by $f_{k,\ell}$ (see Section A.5 in Appendix). Unfortunately, it is not possible in general to define a sampling design $\tilde{p}(\cdot)$ on $\mathbf{s}_m(w), \mathbf{s}_m^{i_1}, \ldots, \mathbf{s}_m^{i_t}$ with inclusion probabilities $\pi_m^1, \ldots, \pi_m^t$ and such that $\tilde{p}(\mathbf{s}_m(w)) = p(\mathbf{s}_m(w))$. There are however several choices available of samples $\mathbf{s}_m^{i_1}, \ldots, \mathbf{s}_m^{i_t}$, sampling designs $\tilde{p}(\cdot)$, and values of $\tilde{p}(\mathbf{s}_m(w))$. It is not clear that choosing $\pi_m^1, \ldots, \pi_m^t$ among the possible samples is the best strategy. And it would be much easier to keep only $\mathbf{s}_m(w)$, $p(\mathbf{s}_m(w))$ and to chose $t$ other samples among all samples and not only among those that are in the support of $p(\cdot)$.

### 6.2.4 Properties of the joint sampling design

- In the case of negative coordinations, the longitudinal sampling design is systematic, and, according to Nedyalkova et al. (2009), it has good properties for this kind of use.

- The whole selection process of different units are independent, hence we have that $\pi^t_{k,\ell} = \pi^t_k \pi^t_\ell$ and $\pi^{t,i}_{k,\ell} = \pi^t_k \pi^i_\ell$. Second order inclusion probabilities $\pi^{t,i}_k$ are not so simple, but can be rapidly computed as the number of sampling occasions is usually small, or at least it is for business surveys. It is thus easy to have a variance formula and to compute the second order inclusion probabilities involved. The main problem of covariance computation remains that negatively coordinated samples are usually and preferably non-overlapping, making the correlations impossible to estimate without a model and strong assumptions.

- The coordination obtained with this method is optimal in the sense that, for each unit, the coordination between a survey $\mathbf{S}^t$ and the survey with which it was to be coordinated with the highest priority is indeed optimal: the bound in 6.1 is reached. Then, among sampling designs that reach this bound, this method gives maximum coordination between $\mathbf{S}^t$ and the survey which was second by order of priority, and then it gives maximum coordination with the third, etc.

## 6.3 Application

### 6.3.1 Rotating panels

The system of coordination we described allows to select rotating panels coordinated with other surveys. Assume that $t-1$ sampling occasions have passed and that we want to organize a new rotating panel, with, for example, a rotation rate of one fifth. Inclusion probabilities of the panel are noted $\pi^p_k$, $k \in U$. We will start by selecting five subsamples $\mathbf{s}^t, \cdots, \mathbf{s}^{t+1}$ that constitute the initial sample $\mathbf{s}^p$ in the panel.

First we define coordination rules and priority for the panel. Then we select, using our coordination system, a first subsample $\mathbf{s}^t$ with inclusion probabilities $\pi^p_k/5$ and the coordination rules defined above. Subsample $\mathbf{s}^{t+1}$ is also selected with inclusion probabilities $\pi^p_k/5$ but we add to the coordination rules that its first priority is to be negatively coordinated with $\mathbf{s}^t$, ensuring that these samples are non-overlapping. Subsample $\mathbf{s}^{t+2}$ is selected with the augmented coordination rules that it must be negatively coordinated with $\mathbf{s}^{t+1}$ and $\mathbf{s}^t$ in priority, and we proceed similarly for $\mathbf{s}^{t+3}$ and $\mathbf{s}^{t+4}$. Once this is done, we define the first sample of the rotating panel as

$$\mathbf{s}^p = \mathbf{s}^t + \cdots + \mathbf{s}^{t+4}.$$

The five subsamples have inclusion probabilities $\pi^p_k/5$ that are at most equal to 0.2. Since these subsamples are negatively coordinated with the highest priority, our method ensures that they do not overlap and that a unit $k$ such that $\pi^p_k = 1$ is indeed selected in one of them. The first subsample to be drawn $\mathbf{s}^t$ is the one that has best coordination with previous surveys, while the last one $\mathbf{s}^{t+4}$ has a deteriorated coordination. It is thus

preferable to discard $\mathbf{s}^{t+4}$ when the panel is updated and to retain $\mathbf{s}^t$ as long as possible.

When the panel must be updated, after $u-1$ sampling occasions, we start by selecting a new subsample $\mathbf{s}^u$, negatively coordinated in priority with $\mathbf{s}^t, \ldots, \mathbf{s}^{t+4}$, then coordinated according to chosen rules with other surveys. Then we update subsamples $\mathbf{s}^t, \ldots, \mathbf{s}^{t+3}$ : first we select a sample $\mathbf{s}^{u+1}$ positively coordinated with $\mathbf{s}^t$ (with highest priority) then negatively coordinated with $\mathbf{s}^u, \mathbf{s}^{t+1}, \mathbf{s}^{t+2}, \mathbf{s}^{t+3}$, then coordinated as chosen with other surveys, then we select $\mathbf{s}^{u+2}, \mathbf{s}^{u+3}, \mathbf{s}^{u+4}$ to update $\mathbf{s}^{t+1}, \mathbf{s}^{t+2}, \mathbf{s}^{t+3}$ along the same lines. Subsample $\mathbf{s}^{t+4}$ is discarded from the panel and we define the updated sample of the rotating panel as

$$\mathbf{s}^{p+1} = \mathbf{s}^u + \cdots + \mathbf{s}^{u+4}.$$

All these operations are described in Algorithm 5.

**Algorithm 5**: Rotating panel

*First selection of the panel*

**Require:** Coordination rules: $o_k^j$, $c_k^j$, $j = 1, \ldots, t-1$;

Define inclusion probabilities $\pi_k^t = \pi_k^p / n$ for a rotating panel with a rotation rate of $1/n$;

Select sample $\mathbf{s}^t$ with inclusion probabilities $\pi_k^t$ using algorithm 4;

**for** $i = t+1$ to $t+n-1$ **do**

  *Update coordination rules to obtain non overlapping samples:*

  Add to coordination rules $o_k^{i-1} = i-1$, $c_k^{i-1}(\mathbf{s}_{k,a}) = 1 - s_{k,a}^{i-1}$;

  Select sample $\mathbf{s}^i$ with inclusion probabilities $\pi_k^t$ using algorithm 4;

**end for**

Define first panel sample $\mathbf{s}^p = \sum_{i=t}^{t+n-1} \mathbf{s}^i$

*Update of the panel*

**Require:** Coordination rules: $o_k^j$, $c_k^j$, $j \in \{1, \ldots, t-1, t+n, \ldots, u-1\}$;

*Select the new subsample* $\mathbf{s}^u$*:*

Add to coordination rules $o_k^{t+r} = u-r-1$, $c_k^{t+r}(\mathbf{s}_{k,a}) = 1 - s_{k,a}^{t+r}$, $r = 0, \ldots, n-1$;

Select sample $\mathbf{s}^u$ with inclusion probabilities $\pi_k^t$ using algorithm 4;

*Update the old subsamples, except* $\mathbf{s}^{t+n-1}$*:*

**for** $i = 0$ to $n-2$ **do**

  *Coordinate in priority with the subsample to be updated:*

  Update coordination rule $o_k^{t+i} = u+i$, $c_k^{t+i}(\mathbf{s}_{k,a}) = s_{k,a}^{t+i}$;

  **for** $r = t+i+1$ to $u+i$ **do**

    $o_k^r \leftarrow o_k^r - 1$;

  **end for**

  *Coordinate negatively with the other subsamples:*

  Add coordination rule $o_k^{u+i} = u+i-1$, $c_k^{u+i}(\mathbf{s}_{k,a}) = 1 - s_{k,a}^{u+i}$;

  Select sample $\mathbf{s}^{u+i+1}$ with inclusion probabilities $\pi_k^t$ using algorithm 4;

**end for**

Define new panel sample $\mathbf{s}^p = \sum_{i=u}^{u+n-1} \mathbf{s}^i$

Coordination rules to select the new subsample $\mathbf{s}^u$ are subject to discussion. Indeed, coordinating $\mathbf{s}^u$ negatively with $\mathbf{s}^{t+4}$ means that, if possible, we absolutely do not want a unit to stay in the panel on more than five consecutive occasions and break a moral contract with the sampled population. It is however not clear that this is the best strategy. Indeed, a unit that drops out of the panel may be selected in another survey, even another panel for which it will have to be trained, so it may be better to coordinate $\mathbf{s}^u$ negatively with $\mathbf{s}^t, \ldots, \mathbf{s}^{t+3}$, then to insert some coordination rules with other surveys before coordinating with $\mathbf{s}^{t+4}$.

### 6.3.2 "Erasing" previous surveys

When a survey is considered as ancient and we do not want to actively coordinate with it any longer, it is possible to remove it from the coordination system. In order to remove survey $i$, it is sufficient to remove column $i$ from Table 6.1, and merge intervals that correspond to the same longitudinal sample obtained after this deletion. It may be necessary to update random number $u_k$ if the interval in which it lies is modified by this merge. Once survey $i$ is removed from the system, it is not possible anymore to coordinate actively a new survey with survey $i$. However future samples will, in general, still be coordinated, in an uncontrolled manner, with survey $i$. Indeed, they will be coordinated with other surveys that were themselves coordinated with survey $i$.

## 6.4 Conclusion

We propose a method of coordinated sampling that is a generalization of the method in Brewer et al. (1972). It is adapted to dynamic populations and has good properties. The independence between selection processes of units in the population is a key factor that allows to design a system capable of dealing with complex problems. These problems are much harder to solve with other methods of coordination. We do not have a perfect solution for merging units: there are simple solutions that have obvious flaws, but a good solution almost certainly requires a large amount of computations. Finally, this method gives Poisson transversal designs with the usual drawback that sample size is random. This is not critical in business surveys, where very large units usually receive an inclusion probability equal to 1, and where sampling rates of other large units are high. Moreover, most of the added variance compared to a fixed size design can be compensated with a calibration estimator. If this system was to be used for household or general population surveys, it would be interesting to look into the possibility of a modified rejective method.

# General conclusion

Instead of a conclusion, I will put down a list of follow-up tasks and open questions motivated by the work exposed in preceding pages.

Over dispersion properties and extremal entropy sampling, in Chapter 3 we gave several results that are necessary conditions for a sampling design with given inclusion probabilities to give extremal dispersion of the eigenvalues of its variance operator, be this extremum a minimum or a maximum. These results depend on the chosen metric on the parameter space of variables of the population. Minimum support designs appear to always be candidate to have an extremal dispersion, as they appeared in Section 3.1.2 to be candidate to having a minimum entropy. Out of curiosity, we still have to determine if they all have indeed locally maximal dispersion and minimal entropy in the convex polytope of sampling designs with given first order inclusion probabilities.

Over the variance of evolution estimators in repeated surveys, in Chapter 4, we estimated the variance of the difference of ratio estimators by substituting the variable of interest $y_k$ with $y_k - \widehat{r}x_k$ where $x_k$ is the auxiliary variable used for the ratio and $\widehat{r}$ is the estimated ratio (in domains). That is to say, we used the estimator $v_C$ of Royall & Cumberland (1981). There has been extensive work published on the properties of this estimator and whether it should be replaced by other estimators, such as $v_2 = (X/\widehat{X})^2 v_C$ (on this subject, see for example Royall & Cumberland, 1981; Särndal et al., 1989; Binder, 1996; Deville, 1999). Other estimators were proposed that use corrective terms in order to address issues raised in Royall & Cumberland (1981) and make a better use of the available information. In Wu (1982), it is proposed to use $(X/\widehat{X})^g$, $g \in \mathbb{R}$, with a particular interest for the case $g = 1$ and in Isaki (1983), it is proposed to use $S_x^2/s_x^2$. The Swiss Federal Statistical Office initially used estimator $v_2$ (inside strata) for its transversal variance estimations. Results where very close to those obtained with estimator $v_C$, at least on an aggregate level. Nevertheless, we should probably look into the opportunity of using a modified estimator of the covariance $S_{xy}$, such as the ratio estimator

$$\widehat{s}_{xy} = \frac{X^2}{\widehat{X}_1 \widehat{X}_2} \frac{s_{xy,s_{12}}}{s_{x,s_{12}} s_{y,s_{12}}} s_{x,s_1} s_{y,s_2}.$$

This work on the evolution of the value added was carried out on a 'true panel', that is to say the selected sample was the same on the three considered sampling occasions. The observed samples differed only due to non-response, be it caused by death or not. The value-added survey sample is however updated about once every three years. Further work includes the adaptation to these years of sample update.

Results of Chapter 5 on the covariance of Horvitz-Thompson estimators in repeated surveys with unequal probability need to be improved. It is particularly true of the lower bound that guarantees that the ratio-type variance estimator is non-negative. We should also try to find out which of the available variance approximations satisfy the required inequality and under which sampling designs they can be used. An important development would be to describe extensively how this work can be adapted to repeated sampling in dynamic populations.

Finally, the coordination method of Chapter 6 needs to be completed. The problem of merging units is complex, both on a theoretical level and on a practical level. Choosing which information must be retained and which can be discarded is delicate. Even more delicate is to find an appropriate method that does not require an enormous amount of computation. This coordination system also presents the much despised flaw of giving random sample sizes. In business surveys, that may not be very important. Weights are strongly dispersed, and big units are always selected. In household surveys, it is more problematic, even if just for cultural reasons. An exploration should be made as to the remote possibility of making a rejective or conditional-on-size adaptation of this coordination method. Permanent Random Numbers need not really be permanent. They can be moved inside the selected interval.

# Appendix

A

## A.1 Proof of proposition 3.1

1. *Minimum support designs are the extremal points of $\mathcal{C}_\pi$, or of $\mathcal{C}_\pi^n$ if we consider fixed size designs only.*

2. *Entropy reaches its local minima in $\mathcal{C}_\pi$ or $\mathcal{C}_\pi^n$ at minimum support designs, and a design that is a global minimum of entropy is also a minimum support design.*

*Proof.* 1. We start by proving that minimum support designs are vertices of $\mathcal{C}_\pi$. Indeed, let $\mathbf{p} \in \mathcal{C}_\pi$ be a minimum support design with support $\mathcal{Q}_\mathbf{p}$ Let also $\mathbf{p}_1$ and $\mathbf{p}_2$ be two sampling designs with the same vector of inclusion probabilities $\boldsymbol{\pi}$, and note their respective supports $\mathcal{Q}_{\mathbf{p}_1}$ and $\mathcal{Q}_{\mathbf{p}_2}$. Now assume that there exists $\alpha \in [0,1]$ such that $\mathbf{p} = \alpha\mathbf{p}_1 + (1-\alpha)\mathbf{p}_2$. If $\alpha$ is not equal to 0 or 1, it follows that $\mathcal{Q}_{\mathbf{p}_1} \subset \mathcal{Q}_\mathbf{p}$ and that $\mathcal{Q}_{\mathbf{p}_2} \subset \mathcal{Q}_\mathbf{p}$. Since $\mathbf{p}$ is a minimum support design, these supports are equal. But, according to Lemma 2.1, there is only one probability distribution on support $\mathcal{Q}_\mathbf{p}$ that gives inclusion probabilities $\boldsymbol{\pi}$. Thus $\mathbf{p}_1 = \mathbf{p}_2 = \mathbf{p}$.

Now we prove that all vertices of $\mathcal{C}_\pi$ are minimal support designs. Indeed, by contradiction, if $\mathbf{p} \in \mathcal{C}_\pi$ is not a minimum support design, using Lemma 2.1, we know that there is a non null vector $\boldsymbol{\lambda}$ in the kernel of its support matrix $\mathbf{S}_\mathbf{p}$ and with coefficients that sum to 0. As we did in the proof of this lemma, we can construct a non trivial line segment $[\mu^a, \mu^b]$, $\mu^a < 0 < \mu^b$ such that, for all $\mu \in [\mu^a, \mu^b]$, $\mathbf{p} + \mu\boldsymbol{\lambda}$ is a sampling design with inclusion probabilities $\boldsymbol{\pi}$. In order to conclude, it is sufficient to remark that $\mathbf{p}$ cannot be an extremal point of $\mathcal{C}_\pi$ as it is a non trivial convex combination of any two sampling designs $\mathbf{p} + \mu_1\boldsymbol{\lambda}$, $\mathbf{p} + \mu_2\boldsymbol{\lambda}$ with $\mu^a \leq \mu_1 < 0$ and $0 < \mu_2 \leq \mu^b$.

2. In this context, entropy is defined as the function

$$\mathcal{H} = \begin{pmatrix} \mathcal{C}_\pi & \longrightarrow & \mathbb{R} \\ \mathbf{p} = (p_1, \ldots, p_{2^N})' & \mapsto & -\sum_i p_i \log p_i \end{pmatrix},$$

where $0 \log 0 = 0$. This function is strictly concave and thus its local minima are necessarily vertices of the convex polytope $\mathcal{C}_\pi$. Indeed,

if $\mathbf{p}$ is not an extremal point of $\mathcal{C}_\pi$, there exist $\mathbf{p}_1$ and $\mathbf{p}_2$ in $\mathcal{C}_\pi$ arbitrarily close to $\mathbf{p}$ and $\alpha \in (0,1)$ such that $\mathbf{p} = \alpha\mathbf{p}_1 + (1-\alpha)\mathbf{p}_2$. Using the strict concavity of $\mathcal{H}$, we get

$$\mathcal{H}(\mathbf{p}) > \alpha\mathcal{H}(\mathbf{p}_1) + (1-\alpha)\mathcal{H}(\mathbf{p}_2) \geq \min\left[\mathcal{H}(\mathbf{p}_1), \mathcal{H}(\mathbf{p}_2)\right],$$

and thus $\mathbf{p}$ cannot be a local minimum of $\mathcal{H}$. Hence, all local minima and the global minimum of $\mathcal{H}$ on $\mathcal{C}_\pi$ are minimum support designs. This proof is still valid with $\mathcal{C}_\pi^n$ instead of $\mathcal{C}_\pi$.

□

## A.2 Proof of proposition 3.2

It is sufficient to prove the following proposition.

*For a given vector of inclusion probabilities $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)$, $d(p, \mathbf{I})$, the dispersion $\delta(p, \mathbf{I})$ and $r(p, \mathbf{I})$ are minimal for any sampling design $p$ that has a diagonal variance matrix $\boldsymbol{V}_p$, that is to say for any sampling design such that $\pi_{k,\ell} = \pi_k\pi_\ell$, for all $1 \leq k \neq \ell \leq N$. Poisson sampling is one such design. Moreover, if the inclusion probabilities are not all equal, this minimal dispersion $\delta(p, \mathbf{I})$ is positive.*

*Proof.* Let $\boldsymbol{\pi}$ be a vector of numbers in $(0,1]$, $p(\cdot)$ be a sampling design with inclusion probabilities $\boldsymbol{\pi}$, and $\lambda_1, \ldots, \lambda_N$ be the eigenvalues of $\boldsymbol{V}_p$. Then

$$Tr(\boldsymbol{V}_p) = \sum_{k=1}^{N} \frac{1-\pi_k}{\pi_k} = \lambda_1 + \ldots + \lambda_N,$$

and

$$\begin{aligned} Tr(\boldsymbol{V}_p^2) &= \lambda_1^2 + \ldots + \lambda_N^2, \\ &= \sum_{k=1}^{N} \left(\frac{1-\pi_k}{\pi_k}\right)^2 + \sum_{i=1}^{N}\sum_{j\neq i} \left(\frac{\pi_{ij}}{\pi_i\pi_j} - 1\right)^2. \end{aligned}$$

Since $\delta(p, \mathbf{I}) = (\lambda_1^2 + \ldots + \lambda_N^2)/N - (\lambda_1 + \ldots + \lambda_N)^2/N^2$ and the $\pi_k$ are given, $\delta(p, \mathbf{I})$ is minimal when $\sum_{i=1}^{N}\sum_{j\neq i}\left(\frac{\pi_{ij}}{\pi_i\pi_j} - 1\right)^2$ is minimal. This last quantity is null for any sampling design where the selection variables of different units are not correlated, for example with Poisson sampling. When it is null, $\delta(p, \mathbf{I})$ is just the dispersion of $\frac{1-\pi_1}{\pi_1}, \ldots, \frac{1-\pi_N}{\pi_N}$, and this dispersion is positive when the $\pi_k$ are not all equal. The last statement of 3.2 also holds. Indeed, note $\mathbf{V}_\mathcal{P}$ the (diagonal) variance matrix of Poisson sampling design with inclusion probabilities $\pi_1, \ldots, \pi_N$ and $\mathbf{V}_{\tilde{p}}$ the variance matrix of another sampling design with the same first order inclusion probabilities. We can assume, for example, that $\pi_1 \leq \ldots \leq \pi_N$. It follows that the maximum eigenvalue of $\mathbf{V}_\mathcal{P}$ is equal to $(1-\pi_1)/\pi_1$, for the normed eigenvector $\mathbf{e}_1 = (1, 0, \ldots, 0)'$, and we have that $\mathbf{e}_1'\mathbf{V}_{\tilde{p}}\mathbf{e}_1$ is also equal to $(1-\pi_1)/\pi_1$. Hence $r(\tilde{p}, \mathbf{I}) \geq r(\mathcal{P}, \mathbf{I})$. These arguments can easily be adapted to the case where $\mathbf{I}$ is replaced with a diagonal positive matrix $\mathbf{D}$. □

## A.3 Proof of proposition 3.4

### A.3.1 Proof of statement 2

*There is no sampling design with fixed size and positive unequal inclusion probabilities $\pi_1, \dots, \pi_N$, $N \geq 3$, such that all the non null eigenvalues of its variance operator $\mathbf{V}_p$ for the Horvitz-Thompson estimator are equal.*

*Proof.* Let us note $\boldsymbol{\pi}$ a vector of inclusion probabilities, and

$$\mathbf{V}_p = \begin{pmatrix} \frac{1-\pi_1}{\pi_1} & \frac{\pi_{1,2}-\pi_1\pi_2}{\pi_1\pi_2} & \dots & \frac{\pi_{1,N}-\pi_1\pi_N}{\pi_1\pi_N} \\ \frac{\pi_{1,2}-\pi_1\pi_2}{\pi_1\pi_2} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ \frac{\pi_{1,N}-\pi_1\pi_N}{\pi_1\pi_N} & \dots & \dots & \frac{1-\pi_N}{\pi_N} \end{pmatrix},$$

so that the variance of the Horvitz-Thompson estimator of the total of a variable $Y$ is equal to $Y'\mathbf{V}_pY$. For a sampling design with fixed size, the variable $\boldsymbol{\pi}$ is always an eigenvector of $\mathbf{V}_p$ associated with the eigenvalue 0. Indeed, the Horvitz-Thompson estimator of the total of $\boldsymbol{\pi}$ is constant and equal to $n$. The sum of the other eigenvalues is equal to $Tr(\mathbf{V}_p) = \sum_{k\in U} \frac{1-\pi_k}{\pi_k}$. Let us note $\boldsymbol{\pi}^\perp = \{\mathbf{x} \in \mathbb{R}^N \ : \ \mathbf{x}'\boldsymbol{\pi} = 0\}$ the orthogonal of $\boldsymbol{\pi}$ in $\mathbb{R}^N$. The other eigenvectors of the symmetric matrix $\mathbf{V}_p$ are all in $\boldsymbol{\pi}^\perp$. If all the eigenvalues of $\mathbf{V}_p$ corresponding to these eigenvectors were equal to $\lambda$, $\mathbf{V}_p$ would be proportional to the orthogonal projector on $\boldsymbol{\pi}^\perp$. Indeed, in that case, for any variable $X$,

$$\begin{aligned} \mathbf{V}_pX = & \quad \mathbf{V}_p \left\{ (\boldsymbol{\pi}'\boldsymbol{\pi})^{-1}\boldsymbol{\pi}\boldsymbol{\pi}'X + \left[I_N - (\boldsymbol{\pi}'\boldsymbol{\pi})^{-1}\boldsymbol{\pi}\boldsymbol{\pi}'\right] X \right\} \\ = & \quad 0 + \mathbf{V}_p \left[I_N - (\boldsymbol{\pi}'\boldsymbol{\pi})^{-1}\boldsymbol{\pi}\boldsymbol{\pi}'\right] X, \\ = & \quad \lambda \left[I_N - (\boldsymbol{\pi}'\boldsymbol{\pi})^{-1}\boldsymbol{\pi}\boldsymbol{\pi}'\right] X, \end{aligned}$$

where $(\boldsymbol{\pi}'\boldsymbol{\pi})^{-1}\boldsymbol{\pi}\boldsymbol{\pi}'$ is the orthogonal projector on $\mathbb{R}^N\boldsymbol{\pi}$, $I_N$ is the identity matrix, and the orthogonal projector on $\boldsymbol{\pi}^\perp$ is equal to $I_N - (\boldsymbol{\pi}'\boldsymbol{\pi})^{-1}\boldsymbol{\pi}\boldsymbol{\pi}'$.

Since the sum of the eigenvalues is given, $\lambda$ would be equal to $\frac{1}{N-1}\sum_{k\in U} \frac{1-\pi_k}{\pi_k}$, and we would have the identity:

$$\mathbf{V}_p = \frac{1}{N-1} \sum_{k\in U} \frac{1-\pi_k}{\pi_k} \left[I_N - (\boldsymbol{\pi}'\boldsymbol{\pi})^{-1}\boldsymbol{\pi}\boldsymbol{\pi}'\right],$$

that is to say

$$\mathbf{V}_p = \frac{1}{N-1} \sum_{k\in U} \frac{1-\pi_k}{\pi_k} \begin{pmatrix} 1 - \frac{\pi_1^2}{\sum_{k\in U}\pi_k^2} & \frac{-\pi_1\pi_2}{\sum_{k\in U}\pi_k^2} & \dots & \frac{-\pi_1\pi_N}{\sum_{k\in U}\pi_k^2} \\ \frac{-\pi_1\pi_2}{\sum_{k\in U}\pi_k^2} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ \frac{-\pi_1\pi_N}{\sum_{k\in U}\pi_k^2} & \dots & \dots & 1 - \frac{\pi_N^2}{\sum_{k\in U}\pi_k^2} \end{pmatrix}. \quad \text{(A.1)}$$

We are going to prove that this matrix cannot be the variance operator of a fixed size sampling design with Horvitz-Thompson estimator strategy

if the $\pi_k$ are not all equal. By contradiction, suppose that there is fixed size sampling design that gives this variance matrix for the Horvitz-Thompson estimator and that the $\pi_k$ are positive and not all equal.

- *First step: there are at most two different values of $\pi_k$ in the population.* Looking at the diagonal of the matrix in A.1, and recalling the expression of $\mathbf{V}_p$, we obtain that

$$\frac{1}{N-1}\sum_{j\in U}\frac{1-\pi_j}{\pi_j}\left(1-\frac{\pi_k^2}{\sum_{j\in U}\pi_j^2}\right)=\frac{1-\pi_k}{\pi_k},\ \text{for all } k.$$

  Now, if we note $B=\sum_{j\in U}\pi_j^2$ and $C=\frac{1}{N-1}\sum_{j\in U}\frac{1-\pi_j}{\pi_j}$, all the $\pi_k$ satisfy the same equation

$$\pi_k^3-\frac{B}{C}(1+C)\pi_k+\frac{B}{C}=0. \tag{A.2}$$

  Hence $\pi_k$ can take no more than three different values in the population, or else we would have found a non-null polynomial of the third degree with more than three roots. Let $\alpha$, $\beta$ and $\gamma$ be the roots of this polynomial. Looking more closely at A.2, we observe that the constant $B/C$ is positive, yielding that this polynomial has a negative root, say $\alpha$. The roots satisfy the following equations:

$$\alpha\beta\gamma = \frac{-B}{C}, \tag{A.3}$$
$$\alpha+\beta+\gamma = 0,\ \text{and} \tag{A.4}$$
$$\alpha\beta+\beta\gamma+\gamma\alpha = -\frac{B}{C}(1+C). \tag{A.5}$$

  It results that the two other roots, $\beta$ and $\gamma$, are either complex numbers with a positive real part, or positive real numbers. Hence $\pi_k$ can take at most two different values. This proves that not all unequal inclusion probability vectors can be obtained with a sampling design with fixed size that enjoys the same property as simple random sampling. In order to finish the proof, we need to prove that all the $\pi_k$ are in fact equal.

  The rest of the proof, which is hard on the eyes, eliminates the possibility for a sampling design with only two different inclusion probabilities to have a variance matrix $\mathbf{V}_p$ as in A.1. Unfortunately, the equation A.2 can have two solutions in $(0,1)$. For example, when the $\pi_k$ are all equal to $n/N$, the other positive root of A.2 is equal to $\frac{n}{N}\left(\sqrt{1+4\frac{N(N-1)}{N-n}}-1\right)$, that can be in $(0,1)$ so it is not sufficient to look at this polynomial. We must also use, at least, the fact that the sum of the inclusion probabilities is equal to $n$.

- *Second step: at least $N-2$ units have the same inclusion probabilities.* Suppose that the population $U$ is split into two parts, one part consists of $N_1$ units with an inclusion probability equal to $\beta$ and the

other part consists of $N_2$ units with an inclusion probability equal to $\gamma$. Let us also suppose that $\gamma > \beta > 0$. We have

$$\begin{aligned} N_1 + N_2 &= N, && \text{(A.6)} \\ N_1\beta + N_2\gamma &= n, && \text{(A.7)} \\ N_1\beta^2 + N_2\gamma^2 &= B, && \text{(A.8)} \\ \frac{1}{N-1}\left(\frac{N_1}{\beta} + \frac{N_2}{\gamma} - N\right) &= C. && \text{(A.9)} \end{aligned}$$

Remark that $0 < \beta < \frac{n}{N} < \gamma \leq 1$.

If $\beta$ and $\gamma$ are solutions of A.2, we also have, using A.3, A.4 and A.5 that

$$\begin{aligned} \beta\gamma(\beta+\gamma) &= \frac{B}{C}, \text{ and} && \text{(A.10)} \\ (\beta+\gamma)^2 - \beta\gamma &= \frac{B}{C}(1+C). && \text{(A.11)} \end{aligned}$$

Since $\beta$ and $\gamma$ are roots of A.2, we also have that

$$\begin{aligned} \beta^3 - B\beta &= \frac{B}{C}(\beta - 1), \text{ and} && \text{(A.12)} \\ \gamma^3 - B\gamma &= \frac{B}{C}(\gamma - 1). && \text{(A.13)} \end{aligned}$$

From A.12 and A.13 we get

$$\frac{\beta(B-\beta^2)}{1-\beta} = \frac{\gamma(B-\gamma^2)}{1-\gamma}\left(= \frac{B}{C}\right).$$

Using Rolle's theorem, it follows that the derivative of $x \mapsto \frac{x(B-x^2)}{1-x}$ has a zero $x_0 \in (\beta, \gamma)$. Hence we have that:

$$B = x_0^2(3 - 2x_0).$$

It follows that $B < 3x_0^2 < 3\gamma^2$ and, using A.8 we have that $N_2 \leq 2$.

Combining equations A.10 and A.11, we have the following equation that will help discard these possibilities:

$$(\beta+\gamma)^2 - \beta\gamma = \beta\gamma(\beta+\gamma) + B,$$

and thus

$$(\beta+\gamma)^2 - \beta\gamma(1+\beta+\gamma) = B. \quad \text{(A.14)}$$

- *Third step:* $N_2 \neq 1$. Suppose that $N_2 = 1$, then A.7 becomes $n = (N-1)\beta + \gamma$ and A.8 becomes $B = (N-1)\beta^2 + \gamma^2$. Using A.14, we have

$$\begin{aligned} (\beta+\gamma)^2 - \beta\gamma(1+\beta+\gamma) &= (N-1)\beta^2 + \gamma^2 \\ \beta^2 + \gamma^2 + 2\beta\gamma - \beta\gamma - \beta^2\gamma - \beta\gamma^2 &= (N-1)\beta^2 + \gamma^2 \quad \hookleftarrow \\ \beta^2 + \beta\gamma - \beta^2\gamma - \beta\gamma^2 &= (N-1)\beta^2 \quad \hookleftarrow \\ \beta(1-\gamma) + \gamma(1-\gamma) &= (N-1)\beta \quad \hookleftarrow \\ (\beta+\gamma)(1-\gamma) &= n - \gamma. \quad \hookleftarrow \end{aligned}$$

Now, since $\frac{n}{N} < \gamma \leq 1$ and $\frac{n-1}{N-1} \leq \beta < \frac{n}{N}$ (using A.7), we also have that

$$n - \gamma = (\beta + \gamma)(1 - \gamma) < 1 - \frac{n^2}{N^2}.$$

Hence $n - \gamma < 1$, and $n = 1$. However, if $n = 1$, the second order inclusion probabilities are null, and thus the terms that are not on the diagonal of the matrix $\mathbf{V}_p$, that must be equal to $(\pi_{k\ell} - \pi_k\pi_\ell)/(\pi_k\pi_\ell)$ are all equal (to $-1$). Going back to the expression of $\mathbf{V}_p$ in A.1, it follows that all the products $\pi_k\pi_\ell$ are equal and thus, if $N \geq 3$, that all the inclusion probabilities are equal, and we have a first contradiction.

- *Fourth and last step:* $N_2 \neq 2$. If $N_2 = 2$, then, using A.7 and A.8, we get that $n = (N-2)\beta + 2\gamma$ and $B = (N-2)\beta^2 + 2\gamma^2$. It follows from A.14 that

$$\begin{aligned} \beta^2 + \gamma^2 + 2\beta\gamma - \beta\gamma - \beta^2\gamma - \beta\gamma^2 &= (N-2)\beta^2 + 2\gamma^2 \\ (\beta + \gamma)(1 - \gamma) - \frac{\gamma^2}{\beta} &= (N-2)\beta \\ &= n - 2\gamma. \end{aligned}$$

However, $(\beta + \gamma)(1 - \gamma) \leq 1 - \gamma^2$ and thus, reporting this inequality in the previous equation, $n - 2\gamma < 1$. It follows that $n \leq 2$. As in the preceding case, $n = 1$ is not possible. If $n = 2$, $(\beta + \gamma)(1 - \gamma) - \frac{\gamma^2}{\beta} = 2(1 - \gamma)$, and:

$$(\beta + \gamma - 2)(1 - \gamma) = \frac{\gamma^2}{\beta}.$$

Necessarily, we have that $\beta + \gamma > 2$. However, if $n = 2$, $\frac{1}{N-2} \leq \beta < \frac{2}{N}$ and $\frac{2}{N} < \gamma \leq 1$, thus:

$$\beta + \gamma \leq 1 + \frac{2}{N} \leq 2.$$

This last contradiction puts an end to the proof.

□

### A.3.2 Proof of statement 3

*$\tilde{\delta}(p, \boldsymbol{\Psi})$ can only be null in degenerate cases where $\pi_k$ takes only two different values that sum to one, and $\widetilde{\mathbf{V}}_p$ is the same as in simple random sampling.*

*Proof.* With the same arguments as in the preceding proof, we observe that $\tilde{\delta}(p, \boldsymbol{\Psi})$ is null only if $\widetilde{\mathbf{V}}_p$ is proportional to the orthogonal projector on the orthogonal of $\boldsymbol{\tau}$. The trace of $\widetilde{\mathbf{V}}_p$ is equal to $N$, and thus we should have

that:

$$\widetilde{V}_p = \frac{N}{N-1}\left(I_N - (\boldsymbol{\tau}'\boldsymbol{\tau})^{-1}\boldsymbol{\tau}\boldsymbol{\tau}'\right),$$

$$= \frac{N}{N-1}\begin{pmatrix} 1-\frac{\pi_1(1-\pi_1)}{\sum_{k\in U}\pi_k(1-\pi_k)} & \dots & -\frac{\left[\pi_1(1-\pi_1)\pi_j(1-\pi_j)\right]^{\frac{1}{2}}}{\sum_{k\in U}\pi_k(1-\pi_k)} & \dots \\ \vdots & \ddots & \ddots & \vdots \\ -\frac{\left[\pi_1(1-\pi_1)\pi_j(1-\pi_j)\right]^{\frac{1}{2}}}{\sum_{k\in U}\pi_k(1-\pi_k)} & \ddots & \ddots & \vdots \\ \vdots & \dots & \dots & 1-\frac{\pi_N(1-\pi_N)}{\sum_{k\in U}\pi_k(1-\pi_k)} \end{pmatrix}.$$

Since all the terms on the diagonal of $\widetilde{V}_p$ must be equal to one, it follows that all the $\pi_k(1-\pi_k)$ are equal. Hence, the inclusion probabilities can only take two values, say $\alpha$ and $1-\alpha$. And in that case, we have that:

$$\widetilde{V}_p = \begin{pmatrix} 1 & -\frac{1}{N-1} & \dots & -\frac{1}{N-1} \\ -\frac{1}{N-1} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ -\frac{1}{N-1} & \dots & \dots & 1 \end{pmatrix}.$$

This situation may or may not to be possible. I must admit that I do not see any obvious reason for it not to be possible, but that I was unable to come up with a fixed size sampling design that has these properties.

□

## A.4 Proof of propositions 3.5 and 3.2

### A.4.1 Proof of proposition 3.5

*For a sampling design to be a local extremum of $d(p,\mathbf{D})$, it must satisfy the necessary condition that there exists a $(N+1)-$vector $(\lambda_0,\dots,\lambda_N)'$ solution of the linear system of P equations:*

$$\sum_{k\in s_i}\sum_{\ell\in s_i}\frac{a_{k,\ell}}{\pi_k\pi_\ell} = \lambda_0 + \sum_{k\in s_i}\lambda_k, \ (3.1),$$

*for all $s_i$, $i=1,\dots,P$, where $a_{k,\ell}$ are coefficients of $\mathbf{D}^{-\frac{1}{2}}\mathbf{M}\mathbf{D}^{-\frac{1}{2}}$.*

*Proof.* $d(p,\mathbf{D})$ is equal to the sum of squares of eigenvalues of $\mathbf{M} = \mathbf{D}^{-\frac{1}{2}}\mathbf{V}_p\mathbf{D}^{-\frac{1}{2}}$. The problem is thus to find extrema of

$$\mathrm{Tr}\left[\left(\mathbf{D}^{-\frac{1}{2}}\mathbf{V}_p\mathbf{D}^{-\frac{1}{2}}\right)^2\right] \quad \text{s.t.} \quad \sum_{i=1}^{P} p(s_i) = 1, \text{ and}$$
$$\sum_{i=1}^{P} I_k(s_i)p(s_i) = \pi_k, \ k=1,\dots,N.$$

Here $\mathbf{V}_p$ depends on $p(\cdot)$ through the relation:

$$\mathbf{V}_p = \boldsymbol{\Phi}^{-1}\left(\mathbf{S}_p\mathbf{P}\mathbf{S}_p' - \boldsymbol{\pi}\boldsymbol{\pi}'\right)\boldsymbol{\Phi}^{-1},$$

where $\mathbf{S}_p$ is the support matrix of $p(\cdot)$ and $\mathbf{P}$ is the diagonal matrix defined by

$$\mathbf{P} = \begin{pmatrix} p(s_1) & & \\ & \ddots & \\ & & p(s_P) \end{pmatrix}.$$

Let $p_i$ denote the $i^{th}$ term on the diagonal of $\mathbf{P}$ (abusively: $p_i = p(s_i)$), and observe that

$$\frac{\partial \mathbf{V}_p}{\partial p_i} = \mathbf{\Phi}^{-1}\mathbf{s}_i\mathbf{s}_i'\mathbf{\Phi}^{-1}.$$

Hence,

$$\begin{aligned} \frac{\partial}{\partial p_i}\mathrm{Tr}\left[\left(\mathbf{D}^{-\frac{1}{2}}\mathbf{V}_p\mathbf{D}^{-\frac{1}{2}}\right)^2\right] &= 2\mathrm{Tr}\left(\mathbf{D}^{-\frac{1}{2}}\mathbf{V}_p\mathbf{D}^{-1}\mathbf{\Phi}^{-1}\mathbf{s}_i\mathbf{s}_i'\mathbf{\Phi}^{-1}\mathbf{D}^{-\frac{1}{2}}\right) \\ &= 2\mathrm{Tr}\left(\mathbf{s}_i'\mathbf{\Phi}^{-1}\mathbf{D}^{-1}\mathbf{V}_p\mathbf{D}^{-1}\mathbf{\Phi}^{-1}\mathbf{s}_i\right) \\ &= 2\sum_{k\in s_i}\sum_{\ell\in s_i}\frac{a_{k,\ell}}{\pi_k\pi_\ell}, \end{aligned}$$

where $a_{k,\ell}$ are coefficients of $\mathbf{D}^{-1}\mathbf{V}_p\mathbf{D}^{-1}$. Remark also that

$$\frac{\partial \pi_k}{\partial p_i} = I_k(s_i),\ k = 1,\ldots,N.$$

A sampling design that is a local extremum of $d(p,\mathbf{D})$ will thus be such that

$$\sum_{k\in s_i}\sum_{\ell\in s_i}\frac{a_{k,\ell}}{\pi_k\pi_\ell} = \lambda_0 + \sum_{k\in s_i}\lambda_k,$$

for all $i = 1,\ldots,P$, where $\lambda_0,\ldots,\lambda_N$ are Lagrange multipliers. □

### A.4.2 Proof of proposition 3.2

1 *Maximum entropy sampling with fixed size usually does not satisfy the preceding conditions and thus is not always a minimum of* $\tilde{\delta}(p,\mathbf{D})$, $\mathbf{D} = \mathbf{I},\ \mathbf{\Psi},\ \boldsymbol{\phi}^{-1}$.

2 *Maximum entropy sampling with fixed size is also not always a minimum of* $r(p,\mathbf{D})$, $\mathbf{D} = \mathbf{I},\ \mathbf{\Psi},\ \boldsymbol{\phi}^{-1}$.

3 *A minimum support design, such as systematic sampling, always satisfies the preceding conditions and thus may be a local extremum of* $d(p,\mathbf{D})$, $\delta(p,\mathbf{D})$, *or* $\tilde{\delta}(p,\mathbf{D})$ *for any metric* $\mathbf{D}$.

*Proof.*

of statement 1: Taking a simple example with $N = 4$, $n = 2$, $\pi_1 = 1/3$, $\pi_2 = 2/3$, $\pi_3 = 2/5$, $\pi_4 = 3/5$, we get that the maximum entropy sampling design does not satisfy the conditions of Corollary 3.1. Here is the output obtained using the package 'sampling' of the 'R' language:

```
> library(sampling)
>
> ##Compute matrices V, tilda(S)
> pik=c(1/3,2/3,2/5,3/5)
> pikl=UPmaxentropypi2(pik)
> phi=diag(1/pik)
> V=phi%*%(pikl-pik%*%t(pik))%*%phi
> S=t(writesample(sum(pik),length(pik)))
> St=rbind(rep(1,times=ncol(S)),S)
>
> ##Check rank of tilda(S)
> qr(St)$rank
[1] 4
>
> ##Compute diagonal of omega, case D=I and
> ##check rank of augmented matrix
> O=diag(t(S)%*%phi%*%V%*%phi%*%S)
> qr(rbind(St,O))$rank
[1] 5
>
> ##Compute diagonal of omega, case D=Psi and
> ##check rank of augmented matrix
> Psi=diag(1/(1-pik))
> U=diag(t(S)%*%phi%*%Psi%*%V%*%Psi%*%phi%*%S)
> qr(rbind(St,U))$rank
[1] 5
>
> ##Compute diagonal of omega, case D=Phi^-1
> ##and check rank of augmented matrix
> U=diag(t(S)%*%(pikl-pik%*%t(pik))%*%S)
> qr(rbind(St,U))$rank
[1] 5
>
```

of statement 2: Also with an example, and having faith in the numeric precision of the procedures involved. With $n = 2$, $N = 4$, the code

```
##write in S all samples of size 2 out of 4
S=t(writesample(2,4))
##assign probabilities to each sample
p=c(1,2,3,1,1,2)
p<-p/sum(p)
pik=S%*%p
##compute variance matrix corresponding to p
##and variance matrix of conditional poisson
##with same inclusion probabilities
Vp=(S%*%diag(p)%*%t(S)-pik%*%t(pik))/(pik%*%t(pik))
Vcp=(UPmaxentropypi2(pik)-pik%*%t(pik))/(pik%*%t(pik))
max(eigen(Vp)$values)-max(eigen(Vcp)$values)

##compute matrix tilda(Vp) corresponding to p
##and matrix tilda(Vp) of conditional poisson
##with same inclusion probabilities
Psi2=(diag(as.vector(pik)/(1-as.vector(pik))))^(.5)
Vcpt=Psi2%*%Vcp%*%Psi2
```

```
Vpt=Psi2%*%Vp%*%Psi2
max(eigen(Vpt)$values)-max(eigen(Vcpt)$values)

##compute matrix corresponding to p and phig^-1
##and matrix tilda(Vp) of conditional poisson
##with same inclusion probabilities
Vcpf=Vcp*((pik%*%t(pik))^(.5))
Vpf=Vp*((pik%*%t(pik))^(.5))
max(eigen(Vpf)$values)-max(eigen(Vcpf)$values)
```

yields result:

```
>max(eigen(Vp)$values)-max(eigen(Vcp)$values)
[1] -0.1165994
```

for the first part. Using

```
p=c(3.1,1.8,3.2,.2,1.2,.5)
```

the result of the second part is:

```
>max(eigen(Vpt)$values)-max(eigen(Vcpt)$values)
[1] -0.03381239
```

With

```
p=c(3.5,1.2,1.7,1.2,1.5,0.8)
```

we also get a design for which $r(p, \mathbf{\Phi}^{-1})$ is smaller than $r(\mathcal{CP}, \mathbf{\Phi}^{-1})$, where $\mathcal{CP}$ is the conditional Poisson sampling design with the same inclusion probabilities as $p$. The numeric result is:

```
> max(eigen(Vpf)$values)-max(eigen(Vcpf)$values)
[1] -0.03238187
```

of statement 3: If $p(\cdot)$ is a minimum support design, then $P \leq N+1$, and, if $p(\cdot)$ is a fixed size design, $P \leq N$ (see Wynn, 1977). Moreover, lemma 2.1 states that $\mathrm{Ker}(\mathbf{S}_p) \cap \mathbf{1}_P^{\perp} = \{\mathbf{0}\}$, hence the matrix $\widetilde{\mathbf{S}}$ is a full rank matrix and there are always solutions $(\lambda_0, \lambda_1, \ldots, \lambda_N)'$ to the considered linear system.

□

## A.5 Geometry for merging units

There is nothing new in the following proposition, but it is easier to prove it here than to find a reference for it.

**Proposition A.1** *Let $\mathcal{K}$ be a convex compact polyhedron of $\mathbb{R}^P$, $P \geq 2$ and note $\mathbf{s}_1, \ldots, \mathbf{s}_r$ the vertices of $\mathcal{K}$, so that $\mathcal{K} = \mathrm{Conv}(\mathbf{s}_1, \ldots, \mathbf{s}_r)$, where $\mathrm{Conv}(\cdot)$ stands for "convex hull". Then, if $\mathcal{A}$ denotes the set of vectors that are convex combinations of $\mathbf{s}_1$ and of $P$ other vertices of $\mathcal{K}$:*

$$\mathcal{A} = \bigcup_{\{i_1,\ldots,i_P\} \subset \{2,\ldots,r\}} \mathrm{Conv}\left(\mathbf{s}_1, \mathbf{s}_{i_1}, \ldots, \mathbf{s}_{i_P}\right),$$

*we have that $\mathcal{K} = \mathcal{A}$.*

*Proof.* Obviously $\mathcal{A} \subset \mathcal{K}$, and it is thus sufficient to prove that $\mathcal{K} \subset \mathcal{A}$. Let $\boldsymbol{\pi}$ be in $\mathcal{K}$. If $\boldsymbol{\pi} = \mathbf{s}_1$, we have that $\boldsymbol{\pi} \in \mathcal{A}$. If $\boldsymbol{\pi} \neq \mathbf{s}_1$, consider

$$\Lambda = \{\lambda \in \mathbb{R} \; s.t. \; \mathbf{s}_1 + \lambda \cdot (\boldsymbol{\pi} - \mathbf{s}_1) \in \mathcal{K}\}.$$

Since $(\boldsymbol{\pi} - \mathbf{s}_1) \neq 0$, and $\mathcal{K}$ is a convex compact set of $\mathbb{R}^P$ that holds $\boldsymbol{\pi}$ and $\mathbf{s}_1$, $\Lambda$ is a convex compact set in $\mathbb{R}$. $\Lambda$ is the preimage of $\mathcal{K}$ by the bijective bicontinuous map $f$,

$$f = \begin{pmatrix} \mathbb{R} & \rightarrow & \mathbf{s}_1 + \mathbb{R} \cdot (\boldsymbol{\pi} - \mathbf{s}_1) \\ \lambda & \mapsto & \mathbf{s}_1 + \lambda \cdot (\boldsymbol{\pi} - \mathbf{s}_1) \end{pmatrix}.$$

Moreover, we have that $[0, 1] \subset \Lambda$.

Note $\lambda_m = \max(\Lambda)$ and $\mathbf{x}_m = \mathbf{s}_1 + \lambda_m \cdot (\boldsymbol{\pi} - \mathbf{s}_1)$. It is sufficient to prove that $\mathbf{x}_m$ lies in the convex hull of $P$ extremal vectors of $\mathcal{K}$ in order to conclude that $\boldsymbol{\pi}$ is in $\mathcal{A}$. Since, by construction, $\mathbf{x}_m$ is in the boundary of $\mathcal{K}$, it lies on a face $\mathcal{F}$ of $\mathcal{K}$. $\mathcal{F} \cap \mathcal{K}$ is a convex compact polyhedron of $\mathbb{R}^{P-1}$, and its vertices are also vertices of $\mathcal{K}$. Using Carathéodory's theorem on convex sets in a finite dimensional space, we get that $\mathbf{x}_m$ is a convex combination of at most $P$ vertices of $\mathcal{F} \cap \mathcal{K}$ and hence of vertices of $\mathcal{K}$. □

**Remark A.1**

- *The vector $\boldsymbol{\pi}$, is thus a convex combination,*

$$\boldsymbol{\pi} = p_1 \mathbf{s}_1 + \sum_{i=2}^{P+1} p_i \mathbf{s}_{a_i}$$

  *of sample $\mathbf{s}_1 (= \mathbf{s}_m(w))$ and of $P$ other samples on the exit face of $[\mathbf{s}_1, \boldsymbol{\pi})$, if $\boldsymbol{\pi} \neq \mathbf{s}_1$. The weight $p_1$ of $\mathbf{s}_1$ is equal to $1 - 1/\lambda_m$, where $\lambda_m$ is defined in the proof of proposition A.1.*

- *There are usually many choices for the $P$ other vertices, and using again proposition A.1, we can force any vertex on the exit face of $[\mathbf{s}_1, \boldsymbol{\pi})$ to be in the convex combination. All these combinations use the same weight $p_1$ for $\mathbf{s}_1$, which is the maximal weight that can be attributed to $\mathbf{s}_1$ when writing $\boldsymbol{\pi}$ as a convex combination of $\mathbf{s}_1$ and of other vertices of $\mathcal{K}$.*

- *For each intersection of $[\mathbf{s}_1, \boldsymbol{\pi})$ with the convex hull of a subset of $\mathbf{s}_2, \ldots, \mathbf{s}_r$, there is a matching weight $p_1$ and several convex decompositions of $\boldsymbol{\pi}$ on $\mathbf{s}_1$ and other samples in $\mathcal{K}$.*

- *Finding adequate weights $p_i$ and samples $\mathbf{s}_{a_i}$, and a fortiori enumerating all possible solutions, involves a large amount of computations, and thus may not be practical.*

- *In the case of merging units, it is not clear that we should use the value $\lambda_m$. Maybe it would be best to find a value for $p_1$ that is close to the original probability of selecting sample $\mathbf{s}_1$.*

# Bibliography


Ardilly, P. & Tillé, Y. (2003). *Exercices corrigés de méthodes de sondage*. Paris: Ellipses. (Cited page 64.)

Berger, Y. (1998). Rate of convergence for asymptotic variance for the Horvitz-Thompson estimator. *Journal of Statistical Planning and Inference* **74**, 149–168. (Cited page 42.)

Berger, Y. (2004a). A simple variance estimator for unequal probability sampling without replacement. *Journal of Applied Statistics* **31**, 305–315. (Cited page 84.)

Berger, Y. (2004b). Variance estimation for measures of change in probability sampling. *The Canadian Journal of Statistics* **32**, 451–467. (Cited pages 18, 19, 61, 77, 80, 94, and 95.)

Berger, Y. (2005). Variance estimation with chao's sampling scheme. *Journal of Statistical Planning and Inference* **127**, 253–277. (Cited page 84.)

Binder, D. (1996). Linearization methods for single phase and two-phase samples: a cookbook approach. *Survey Methodology* **22**, 17–22. (Cited page 113.)

Binder, D. & Patak, Z. (1994). Use of estimating functions for estimation from complex surveys. *Journal of the American Statistical Association* **89**, 1035–1043. (Cited page 68.)

Brewer, K. (1963). A model of systematic sampling with unequal probabilites. *Australian Journal of Statistics* **5**, 5–13. (Cited page 30.)

Brewer, K. (2002). *Combined Survey Sampling Inference, Weighing Basu's Elephants*. London: Arnold. (Cited pages 84 and 96.)

Brewer, K. & Donadio, M. (2003). The high entropy variance of the Horvitz-Thompson estimator. *Survey Methodology* **29**, 189–196. (Cited pages 42, 84, and 96.)

Brewer, K., Early, L. & Joyce, S. (1972). Selecting several samples from a single population. *Australian Journal of Statistics* **3**, 231–239. (Cited pages 19, 98, 99, 102, and 112.)

Brewer, K. & Hanif, M. (1983). *Sampling with Unequal Probabilities*. New York: Springer-Verlag. (Cited pages 25, 27, 29, 30, 38, 40, and 57.)

Caron, N. & Ravalet, P. (2000). Estimation dans les enquêtes répétées: Application à l'enquête emploi en continu. Tech. Rep. 0005, Méthodologie Statistique, INSEE, Paris. (Cited pages 61 and 67.)

Cassel, C., Särndal, C. & Wretman, J. (1977). *Foundations of Inference in Survey Sampling*. New York: Wiley. (Cited page 26.)

Chao, M. (1982). A general purpose unequal probability sampling plan. *Biometrika* **69**, 653–656. (Cited pages 24 and 81.)

Chen, S., Dempster, A. & Liu, J. (1994). Weighted finite population sampling to maximize entropy. *Biometrika* **81**, 457–469. (Cited pages 17, 23, 24, 25, and 26.)

Cheng, C. & Li, K. (1983). A minimax approach to sample surveys. *Annals of Statistics* **11**, 552–563. (Cited page 53.)

Connor, W. (1966). An exact formula for the probability that specified sampling units will occur in a sample drawn with unequal probabilities and without replacement. *Journal of the American Statistical Association* **61**, 384–490. (Cited page 30.)

Cotton, F. & Hesse, C. (1992). Tirages coordonnes d'échantillons. Document de travail de la Direction des Statistiques Économiques E9206. Tech. rep., INSEE, Paris. (Cited page 97.)

De Ree, S. (1983). A system of co-ordinated sampling to spread response burden of enterprises. In *Contributed paper, 44th Session of the ISI Madrid*. (Cited page 97.)

Deville, J.-C. (1993). Estimation de la variance pour les enquêtes en deux phases. Manuscript, INSEE, Paris. (Cited pages 84 and 96.)

Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: linearization and residual techniques. *Survey Methodology* **25**, 193–204. (Cited pages 68, 84, 96, and 113.)

Deville, J.-C. & Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association* **87**, 376–382. (Cited page 68.)

Deville, J.-C. & Tillé, Y. (1998). Unequal probability sampling without replacement through a splitting method. *Biometrika* **85**, 89–101. (Cited pages 24, 29, 31, 40, and 81.)

Fuller, W. & Rao, J. (2001). A regression composite estimator with application to the canadian labour force survey. *Survey Methodology* **27**, 45–51. (Cited page 61.)

Gabler, S. (1981). A comparison of Sampford's sampling procedure versus unequal probability sampling with replacement. *Biometrika* **68**, 725–727. (Cited pages 24 and 81.)

Gabler, S. (1984). On unequal probability sampling: sufficient conditions for the superiority of sampling without replacement. *Biometrika* **71**, 171–175. (Cited pages 17 and 24.)

Gabler, S. (1990). *Minimax Solutions in Sampling from Finite Populations*, vol. 64. Berlin: Springer-Verlag. (Cited pages 49 and 54.)

Gabler, S. & Schweigkoffer, R. (1990). The existence of sampling designs with preassigned inclusion probabilities. *Metrika* **37**, 87–96. (Cited page 50.)

Goga, C. (2003). *Estimation de la variance dans les sondages à plusieurs échantillons et prise en compte de l'information auxiliaire par des modèles non-paramétriques*. Ph.D. thesis, Université de Rennes II, Haute Bretagne, France. (Cited pages 62, 63, and 79.)

Gray, G. (1971). Joint probabilities of selection of units in systematic samples. In *Proceedings of the American Statistical Association, Survey Research Methods Section*. (Cited page 30.)

Hájek, J. (1959). Optimum strategy and other problems in probability sampling. *Casopis pro Pestování Matematiky* **84**, 387–423. (Cited page 47.)

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics* **35**, 1491–1523. (Cited pages 17, 18, 23, 77, 84, and 96.)

Hansen, M. & Hurwitz, W. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics* **14**, 333–362. (Cited pages 25 and 81.)

Hardy, G., Littlewood, J. & Pólya, G. (1956). *Inequalities*. Cambridge Univ. Press. (Cited page 26.)

Hedayat, A., Bing-Ying, L. & Stufken, J. (1989). The construction of $\pi ps$ sampling designs through a method of emptying boxes. *Annals of Statistics* **4**, 1886–1905. (Cited page 40.)

Hesse, C. (1999). Sampling co-ordination: a review by country. Tech. Rep. E9908, Direction des Statistiques d'Entreprises, INSEE, Paris. (Cited page 97.)

Hidiroglou, M. & Gray, G. (1980). Construction of joint probability of selection for systematic P.P.S. sampling. *Applied Statistics* **29**, 107–112. (Cited page 30.)

Hidiroglou, M., Särndal, C.-E. & Binder, D. (1995). Weighting and estimation in business surveys. In *Business Survey Methods*, B. Cox, D. Binder, B. Chinnappa, A. Christianson, M. Colledge & P. Kott, eds. New York: Wiley, pp. 477–502. (Cited page 61.)

Hidiroglou, M. A. & Srinath, K. P. (1981). Some estimators of a population total from simple random samples containing large units. *Journal of the American Statistical Association* **76**, 690–695. (Cited pages 18 and 71.)

Holmes, D. & Skinner, C. (2000). Variance estimation for labour force survey estimates of level and change. Tech. rep., Government Statistical Service Methodology Series., 21, London, England. (Cited page 61.)

Horvitz, D. & Thompson, D. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47**, 663–685. (Cited pages 24 and 78.)

Hulliger, B. (1999). Simple and robust estimators for sampling. In *Proceedings of the Section on Survey Research Methods*. American Statistical Association. (Cited pages 69 and 71.)

Iachan, R. (1982). Systematic sampling: A critical review. *International Statistical Review* **50**, 293–303. (Cited page 42.)

Isaki, C. (1983). Variance estimation using auxiliary information. *Journal of the American Statistical Association* **78**, 117–123. (Cited page 113.)

Jessen, R. (1969). Some methods of probability non-replacement sampling. *Journal of American Statistics Association* **64**, 175–193. (Cited pages 29 and 40.)

Kish, L. (1965). *Survey Sampling*. New York: Wiley. (Cited pages 19, 61, and 95.)

Laniel, N. (1988). Variances for a rotating sample from a changing population. In *Proceedings of the Business and Economic Statistics Section*. American Statistical Association. (Cited page 61.)

Madow, W. (1949). On the theory of systematic sampling, II. *Annals of Mathematical Statistics* **20**, 333–354. (Cited page 30.)

Matei, A. & Tillé, Y. (2005). Evaluation of variance approximations and estimators in maximum entropy sampling with unequal probability and fixed sample size. *Journal of Official Statistics* **21**, 543–570. (Cited page 84.)

Midzuno, H. (1950). An outline of the theory of sampling systems. *Annual Institute of Statistical of Mathematics* **1**, 149–156. (Cited page 57.)

Narain, R. (1951). On sampling without replacement with varying probabilities. *Journal of Indian Society for Agricultural Statistics* **3**, 169–174. (Cited page 78.)

Nedyalkova, D., Qualité, L. & Tillé, Y. (2009). General framework for the rotation of units in repeated survey sampling. *Accepted in Statistica Neerlandica* . (Cited pages 97, 99, and 109.)

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. (Cited page 61.)

Pea, J., Qualité, L. & Tillé, Y. (2007). Systematic sampling is a minimum support design. *Computational Statistics and Data Analysis* **51**, 5591–5602. (Cited page 17.)

Peters, R., Renfer, J.-P. & Hulliger, B. (2001). Statistique de la valeur ajoutée : procédure d'extrapolation des données. Tech. rep., Office fédéral de la Statistique. (Cited page 71.)

Pinciaro, S. J. (1978). An algorithm for calculating joint inclusion probabilities under PPS systematic sampling. In *ASA Proceedings of the Section on Survey Research Methods*. American Statistical Association. (Cited page 30.)

Qualité, L. (2008). A comparison of conditional poisson sampling versus unequal probability sampling with replacement. *Journal of Statistical Planning and Inference* **138**, 1428–1432. (Cited pages 17 and 81.)

Qualité, L. & Tillé, Y. (2008). Variance estimation of changes in repeated surveys and its application to the swiss survey of value added. *Survey Methodology* **34**, 173–181. (Cited pages 18, 77, and 79.)

Renfer, J.-P. (2000). Enquête sur la production et la valeur ajoutée : échantillonnage complémentaire. Tech. rep., Office fédéral de la Statistique. (Cited pages 70 and 71.)

Rivière, P. (2001). Coordinating samples using the microstrata methodology. *Proceedings of Statistics Canada Symposium 2001* . (Cited pages 97 and 98.)

Royall, R. & Cumberland, W. (1981). An empirical study of the ratio estimator and its variance. *Journal of the American Statistical Association* **76**, 66–77. (Cited page 113.)

Sampford, M. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika* **54**, 499–513. (Cited page 81.)

Sen, A. (1953). On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics* **5**, 119–127. (Cited pages 24 and 88.)

Sen, A. (1973). Theory and application of sampling on repeated occasions with several auxiliary variables. *Biometrics* **29**, 381–385. (Cited page 61.)

Sengupta, S. (1989). On Chao's unequal probability sampling plan. *Biometrika* **76**, 192–196. (Cited pages 24 and 81.)

Sinha, B. (1973). On sampling schemes to realize preassigned sets of inclusion probabilities of first two orders. *Bulletin of the Calcutta Statistical Association* **22**, 89–110. (Cited page 50.)

Särndal, C.-E., Swensson, B. & Wretman, J. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika* **76**, 527–537. (Cited page 113.)

Stenger, H. (1979). A minimax approach to randomization and estimation in survey sampling. *Annals of Statistics* **7**, 395–399. (Cited page 49.)

Stenger, H. & Gabler, S. (1996). A minimax property of Lahiri-Midzuno-Sen's sampling scheme. *Metrika* **43**, 213–220. (Cited pages 49 and 57.)

Tam, S. (1984). On covariances from overlapping samples. *The American Statistician* **38**, 288–289. (Cited pages 61, 63, 64, 75, 77, and 79.)

Tillé, Y. (1996). An elimination procedure of unequal probability sampling without replacement. *Biometrika* **83**, 238–241. (Cited pages 24, 81, and 87.)

Tillé, Y. (2006). *Sampling algorithms*. New York: Springer-Verlag. (Cited pages 24 and 40.)

Tillé, Y. & Matei, A. (2005). *The R package Sampling*. The Comprehensive R Archive Network, http://cran.r-project.org/, Manual of the Contributed Packages. (Cited page 38.)

Traat, I., Bondesson, L. & Meister, K. (2004). Sampling design and sample selection through distribution theory. *Journal of Statistical Planning and Inference* **123**, 395–413. (Cited page 24.)

Wolter, K. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag. (Cited page 61.)

Woodruff, R. (1971). A simple method for approximating de variance of a complicated estimate. *Journal of the American Statistical Association* **66**, 411–414. (Cited page 68.)

Wu, C.-F. (1982). Estimation of variance of the ratio estimator. *Biometrika* **69**, 183–189. (Cited page 113.)

Wynn, H. (1976). Connected finite population sampling plans. *Biometrika* **63**, 208–210. (Cited page 49.)

Wynn, H. (1977). Convex sets of finite population plans. *Annals of Statistics* **5**, 414–418. (Cited pages 29, 31, 52, and 124.)

Yates, F. & Grundy, P. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society* **B15**, 235–261. (Cited pages 24 and 88.)

# Notations

| | |
|---|---|
| $U = \{1, \ldots, N\}$ | Population; |
| $s \subset U$ | Sample (without replacement); |
| $\mathbf{s} \in \{0,1\}^N$ | Sample (vector notation); |
| $I_k(s)$ | Indicator variable of unit $k$ in sample $s$; |
| $\mathcal{S} = \{0,1\}^N$ | Set of all possible samples of $U$; |
| $\mathbf{S}$ | Matrix of all possible samples; |
| $\mathcal{S}_n$ | Set of all samples of size $n$; |
| $p(\cdot)$ or $P(\cdot)$ | Sampling design (without replacement), probability distribution on $\mathcal{S}$; |
| $\mathbf{p} \in [0,1]^{2^N}$ | Sampling design (vector notation); |
| $\mathcal{Q}_{\mathbf{p}} \subset \mathcal{S}$ | Support of a sampling design; |
| $\mathbf{S}_{\mathbf{p}}$ | Support matrix of a sampling design; |
| $\mathcal{H}(\mathbf{p})$, $I(p)$ | Entropy of a sampling design; |
| $\pi_k = \mathrm{E}(I_k)$ | First order inclusion probabilities, $\mathrm{E}(\cdot)$ is the expectation under $p(\cdot)$; |
| $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)'$ | Vector of inclusion probabilities; |
| $\mathcal{C}_{\boldsymbol{\pi}}$ | Set of sampling designs on $\mathcal{S}$ with inclusion probabilities $\boldsymbol{\pi}$; |
| $\mathcal{C}_{\boldsymbol{\pi}}^n$ | Set of sampling designs on $\mathcal{S}_n$ with inclusion probabilities $\boldsymbol{\pi}$; |
| $\pi_{k,\ell} = \mathrm{E}(I_k I_\ell)$ | Second order inclusion probabilities; |
| $\Delta_{k,\ell} = \pi_{k,\ell} - \pi_k \pi_\ell$ | Covariance of $I_k$ and $I_\ell$; |
| $U^t$ | Population at time $t$ (in Chapter 6); |
| $\mathbf{s}^t$ | Sample at time $t$; |
| $S_k^t$ | Indicator variable of unit $k$ in sample $\mathbf{s}^t$; |
| $\pi_k^t$ | Inclusion probability of unit $k$ in sample $\mathbf{s}^t$; |
| $\pi_{k,\ell}^{t,i} = \mathrm{E}(S_k^t S_\ell^i)$ | Second order inclusion probabilities over different sampling occasions; |
| $\widehat{Y}_{HT}$, $\widehat{Y}$ | Horvitz-Thompson estimator; |
| $\widehat{Y}_{HH}$ | Hansen-Hurwitz estimator; |
| $s_{y,s_a}$ | Standard error of variable $Y$ in sample $s_a$; |
| $s_{xy,s_a}$ | Covariance of variables $X$ and $Y$ in sample $s_a$. |

Ce document a été préparé à l'aide de l'éditeur de texte TeXnicCenter et du logiciel de composition typographique LaTeX $2_\varepsilon$.

**Titre** Sondage à probabilités inégales et enquêtes répétées

**Résumé** Ce document est constitué de deux parties. Dans la première partie, nous nous intéressons à certains plans de sondage à probabilités inégales, et dans la deuxième partie nous étudions le problème des enquêtes répétées. Bien que les sujets développés dans ces deux parties semblent entièrement différents, ils sont en fait reliés. La première partie est principalement consacrée à l'étude des propriétés de deux plans de sondage de taille fixe. Dans un premier chapitre, il est démontré que le plan de sondage à entropie maximale et de taille fixe est plus efficace que le sondage avec remise. Dans le second chapitre, nous montrons que le sondage systématique est un plan à support minimal. Nous donnons aussi quelques résultats sur la variance de l'estimateur de Horvitz-Thompson pour les plans à entropie maximale et pour les plans à support minimal. La deuxième partie débute par une étude de cas sur l'estimation de précision des évolutions dans le panel suisse sur la valeur ajoutée. Dans le chapitre suivant, nous proposons un estimateur de covariance pour les panels rotatifs à probabilités inégales. Enfin, nous présentons un système de coordination d'échantillons poissoniens développé pour l'Office Fédéral de la Statistique Suisse.

**Mots-clés** Entropie maximale, Coordination d'échantillons, Variance, Plan systématique

**Title** Unequal probability sampling and repeated surveys

**Abstract** This document is divided into two parts. The first part revolves around the properties of some unequal probability survey sampling designs, and the second part deals with repeated surveys. While the topics developed in these two parts appear to be largely different, they are in fact related. The first part is devoted to the study of properties of two sampling designs with fixed size. In a first chapter we show that maximum entropy sampling with fixed size is more efficient than sampling with replacement. In a second chapter we prove that systematic sampling is a minimum support design. We also give some results on the variance of the Horvitz-Thompson estimator for maximum entropy and for minimum support designs. The second part begins with a case study of the estimation of variance of evolutions in the Swiss panel on value added. In a second chapter, we give covariance estimators for rotating panels with unequal inclusion probabilities. Finally, we describe a coordination method of maximum entropy samples that was developed for the Swiss Federal Statistical Office.

**Keywords** Maximal entropy, Sample coordination, Variance, Systematic sampling design