Computational aspects of sample surveys

Thèse présentée à la Faculté des sciences
Université de Neuchâtel

Pour l'obtention du grade de docteur ès science

Par

Alina MATEI

Acceptée sur proposition du jury :
Professeur Yves Tillé, directeur de thèse
Professeur Pascal Felber, co-directeur de thèse
M. Jean-Claude Deville, rapporteur
Professeur Sthephan Morgenthaler, rapporteur
Professeur Chris Skinner, rapporteur

Soutenue le 24 novembre 2005

Université de Neuchâtel
2005

Université de Neuchâtel unine

IMPRIMATUR POUR LA THESE

Computational aspects of sample surveys

Alina MATEI

UNIVERSITE DE NEUCHATEL

FACULTE DES SCIENCES

La Faculté des sciences de l'Université de Neuchâtel,
sur le rapport des membres du jury

MM. J. Tillé (directeur de thèse),
P. Felber (co-directeur de thèse),
J.-C. Deville (Rennes F),
S. Morgenthaler (EPF Lausanne)
et C. Skinner (Southampton)

autorise l'impression de la présente thèse.

Neuchâtel, le 25 novembre 2005

Le doyen :

J.-P. Derendinger

Faculté des Sciences
■ Rue Emile-Argand 11 ■ CP 2 ■ CH-2007 Neuchâtel
■ Téléphone : +41 32 718 21 00 ■ Fax : +41 32 718 21 03 ■ E-mail : secretariat.sciences@unine.ch ■ www.unine.ch

Mots clés : Coordination d'échantillons dans le temps, nombres aléatoires permanents, plan ordonné, variance, plan proportionnel à la taille, problème de transport, algorithme, statistiques d'ordre, contrôle de l'arrondissement, imputation, calage, données qualitatives, simulations de Monte Carlo.

Keywords: sample co-ordination, permanent random number, order sampling, variance, πps sampling design, transportation problem, algorithm, order statistics, controlled rounding, imputation, calibration, qualitative data, Monte Carlo simulation.

Résumé : Cette thèse est consacrée à quatre problèmes d'échantillonnage. Après un chapitre de présentation générale (chapitre 1), on s'intéresse dans le chapitre 2 au calcul des probabilités d'inclusion d'ordre un dans des plans ordonnés et proportionnels à la taille. Les chapitres 2 et 3 traitent de la coordination d'échantillons dans le temps. Le chapitre 4 est consacré à l'approximation et à l'estimation de la variance des échantillons de taille fixe à entropie maximale et à probabilités d'inclusion inégales. Les chapitres 5 et 6 traitent d'une modification de l'algorithme de Cox pour l'équilibrage en nombres entiers de tableaux rectangulaires à marges contraintes, avec application à l'imputation d'une variable qualitative. Une place importante est accordée à l'algorithmique et aux simulations de Monte-Carlo.

Art is made to disturb. Science reassures.
Georges Braque

To my parents and Buratino

Acknowledgements

First of all, I would like to thank my supervisor, Professor Yves Tillé, for his will to take me as PhD student, to convert me to the sample survey domain, for our collaboration and for the mutual trust. For all, thank you, Yves.

In addition, I would like to thank all the members of the thesis committee: Jean-Claude Deville (General inspector of the INSEE and Head of the Survey Statistics Laboratory, CREST/ENSAI, France), Professor Pascal Felber (University of Neuchâtel, Switzerland), Professor Stephan Morgenthaler (Swiss Federal Institute of Technology Lausanne, Switzerland), and Professor Chris Skinner (University of Southampton, United Kingdom).

Special thanks to Anne-Catherine Favre, who is the co-author of two presented papers and who marks the beginning of my stay in Neuchâtel.

I thank Thanos Kondylis for his patience to read and to improve each English version of the presented papers, for his criticism, and for his sense of humor.

I would also like to express my gratitude to the University of Neuchâtel for its support. Many thanks to the friends Inès Pasini, Giuseppe Melfi, Desi Nedyalkova, to the colleagues at the Institute of Statistics (or Statistics Group), and to the Office for the equality of women and men of the university.

I thank Monique Graf for our collaboration at the Swiss Federal Statistical Office.

I would also like to thank Professor Lennart Bondesson from University of Umeå, Sweden for his very useful remarks concerning the thesis.

Last but not least, I wish to thank my parents who have followed each step of my development, and to Buratino for his love and support. The thesis is dedicated to these three persons.

Alina Matei,
Neuchâtel, November 2005

Abstract

Sample survey research has attracted for a long time the interest of statistics and applied probability. The high complexity of this research field together with the development of computer science over the last decade have made survey sampling both a statistical and a computational challenge. The present thesis puts together the research results on several computational aspects in sampling methods. These mainly concern the computation of inclusion probabilities in order πps sampling, the conditions to reach optimal sample coordination, the variance estimation in unequal probability sampling designs with fixed sample size, and the imputation for qualitative data. All the presented algorithms and the Monte Carlo simulations are implemented in C++ in order to guarantee a fast, flexible and effective computation. The thesis is based on the following papers:

1) Paper 1: A. Matei and Y. Tillé (2005), Computational aspects in order sampling designs, *submitted*;

2) Paper 2: A. Matei and Y. Tillé (2005), Maximal and minimal sample co-ordination, *accepted in Sankhyā*;

3) Paper 3: A. Matei and Y.Tillé (2005), Evaluation of variance approximations and estimators in maximum entropy sampling with unequal probability and fixed sample size, *accepted in Journal of Official Statistics*;

4) Paper 4[1]: A.-C. Favre, A. Matei and Y. Tillé (2004), A variant of the Cox algorithm for the imputation of non-response of qualitative data, *Computational Statistics & Data Analysis,* 45(4):709-719;

[1]Reprinted from Computational Statistics & Data Analysis, Vol. 45(4), A.-C. Favre, A. Matei, Y. Tillé, A variant of the Cox algorithm for the imputation of non-response of qualitative data, pages 709-719, Copyright (2003), with permission from Elsevier.

5) Paper 5[2]: A.-C. Favre, A. Matei and Y. Tillé (2005), Calibrated random imputation for qualitative data, *Journal of Statistical Planning and Inference,* 128(2): 411-425.

Keywords and phrases: sample co-ordination, permanent random number, order sampling, variance, πps sampling design, transportation problem, algorithm, order statistics, controlled rounding, imputation, calibration, qualitative data, Monte Carlo simulation.

2000 Mathematics Subject Classifications: 62D05.

[2]Reprinted from Journal of Statistical Planning and Inference, Vol. 128(2), A.-C. Favre, A. Matei, Y. Tillé, Calibrated random imputation for qualitative data, pages 411-425, Copyright (2003), with permission from Elsevier.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Some notations and basic concepts

Sample survey gives the modality to make inference about a characteristic of a finite population by using only a part of this population. Let $U = \{1, \ldots, k, \ldots, N\}$ be a finite population. The unit k is the reference unit, and N denotes the population size. A sample is a subset of U. Let $\mathcal{S}$ be the sample support, which is the set of all possible samples drawn from U. Thus $\mathcal{S}$ is the set of 2^N subsets of U. A couple $(\mathcal{S}, p)$ is denoted as a *sampling design*, where p is a probability distribution on $\mathcal{S}$. For a given $p(.)$, any $s \in \mathcal{S}$ is viewed as a realization of a random variable S, such that

$$Pr(S = s) = p(s).$$

Suppose we have $k \in S$. Thus the random event "$S \ni k$" is the event "a sample containing k is realized" (Särndal et al., 1992, p.31). The cardinality of the set s is the *sample size* of s, and we shall denote it by n. We consider only sampling without replacement. Given $p(.)$, the inclusion probability of a unit k is the probability that unit k will be in a sample. It is defined by

$$\pi_k = Pr(k \in S) = \sum_{\substack{s \ni k \\ s \in \mathcal{S}}} p(s).$$

The quantities π_k are denoted as the *first-order inclusion probabilities*, $\forall\, k \in U$. Similarly, the *second-order inclusion probabilities* or the *joint inclusion proba-*

bilities are defined as

$$\pi_{k\ell} = Pr(k \in S, \ell \in S) = \sum_{\substack{s \ni k,\ell \\ s \in \mathcal{S}}} p(s).$$

When the sample size is fixed to n, the inclusion probabilities satisfy the conditions

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

$$\sum_{\substack{\ell \in U \\ \ell \neq k}} \pi_{k\ell} = (n-1)\pi_k.$$

The inclusion probabilities are useful for variance estimation.

The Horvitz-Thompson estimator (or the π-estimator) of the population total $t_y = \sum_{k \in U} y_k$ is defined as

$$\widehat{t}_\pi = \sum_{k \in s} \frac{y_k}{\pi_k}, \tag{1.1}$$

where $\mathbf{y} = (y_1, \ldots, y_k, \ldots, y_N)$ is the variable of interest. This estimator is unbiased for any sampling design that gives $\pi_k > 0, \ \forall \ k \in U$ (Horvitz and Thompson, 1952). The variance of the π-estimator is

$$\text{var}(\widehat{t}_\pi) = \sum_{k \in U} \sum_{\ell \in U} (\pi_{k\ell} - \pi_k \pi_\ell) \frac{y_k y_\ell}{\pi_k \pi_\ell}.$$

If $y_k \propto \pi_k$, $\text{var}(\widehat{t}_\pi)$ is zero. For a sampling with fixed sample size the variance of $\widehat{t}_\pi$ is given by (Yates and Grundy, 1953; Sen, 1953)

$$\text{var}(\widehat{t}_\pi) = -\frac{1}{2} \sum_{k \in U} \sum_{\substack{\ell \in U \\ \ell \neq k}} \left(\frac{y_k}{\pi_k} - \frac{y_\ell}{\pi_\ell} \right)^2 (\pi_{k\ell} - \pi_k \pi_\ell).$$

When $\pi_{k\ell} > 0$ for all pairs of units and the sample size is fixed, the following two variance estimators are unbiased:

- the Horvitz-Thompson estimator (Horvitz and Thompson, 1952)

$$\widehat{\text{var}}_{HT}(\widehat{t}_\pi) = \sum_{k \in s} \sum_{\ell \in s} \frac{\pi_{k\ell} - \pi_k \pi_\ell}{\pi_{k\ell}} \frac{y_k}{\pi_k} \frac{y_\ell}{\pi_\ell};$$

- the Sen-Yates-Grundy estimator (Yates and Grundy, 1953; Sen, 1953)

$$\widehat{\text{var}}_{SYG}(\widehat{t}_\pi) = -\frac{1}{2}\sum_{k\in s}\sum_{\substack{\ell\in s\\ \ell\neq k}}\left(\frac{y_k}{\pi_k}-\frac{y_\ell}{\pi_\ell}\right)^2\frac{\pi_{k\ell}-\pi_k\pi_\ell}{\pi_{k\ell}}. \tag{1.2}$$

When an auxiliary variable $\mathbf{z} = (z_1, \ldots, z_k, \ldots, z_N)$, with $z_k > 0,\ \forall\ k \in U$ related to the variable of interest $\mathbf{y}$ is available for all units in the population, unequal probability sampling is frequently used in order to increase the efficiency of the estimation (generally using the π-estimator). In this case, for a sampling design with fixed sample size n, the inclusion probabilities are computed as

$$\pi_k = \frac{nz_k}{\sum_{\ell\in U} z_\ell}, \tag{1.3}$$

and the sampling design is denoted as πps sampling design.

The sections below present several problems issued from sample co-ordination, order πps sampling design, maximum entropy sampling design, controlled rounding, calibrated imputation, and their relation with the thesis.

1.2 Sample co-ordination

Sample co-ordination (or optimal integration of surveys or overlapping maps as comments Ernst, 1999) is a commonly faced problem in official statistics. It consists in creating a dependence between two or more samples, in order to minimize or maximize their overlap (the number of common units). Populations are sampled on two or more occasions, with the purpose to estimate some characteristic of the population on each occasion (the cross-sectional aspect) as well as changes in it between occasions (the longitudinal aspect). The major aim of the co-ordination is to control the overlap between samples selected from the populations. Populations usually change with time, due to births (the addition of new units), deaths (the deletion of units) or changes in activity in the case of business surveys.

We are interested in the case where samples are selected in two distinct time periods. The time periods are indicated by the exponents 1 and 2 in our notation. Our notation for a sample is s^1 for time period 1 and s^2 for time period 2. Thus, π_k^1 denotes the inclusion probability of unit $k \in U$ for time period 1 in sample s^1. Similarly, π_k^2 denotes the inclusion probability of unit

$k \in U$ for time period 2 in sample s^2. Let also $\pi_k^{1,2}$ be the joint inclusion probability of unit k in both samples. When two samples are drawn independently, without co-ordination, we have

$$\pi_k^1 \pi_k^2 = \pi_k^{1,2}, \text{ for all } k \in U.$$

Due to the Fréchet bounds, $\pi_k^{1,2}$ satisfies

$$\max(0, \pi_k^1 + \pi_k^2 - 1) \leq \pi_k^{1,2} \leq \min(\pi_k^1, \pi_k^2).$$

There are two kinds of sample co-ordination: positive and negative. In the positive case, the overlap is maximized; in the negative case, it is minimized. When the same sample is measured repeatedly in time, we talk about a *panel*. It is possible to rotate some units in the panel (*rotated panel*), that is these units are effectively missing by design, in order to avoid their fatigue or to respect the population changes over time. In the glossary of Eurostat Concepts and Definitions Database [1], the co-ordination of samples is defined as below, by using some ideas from Lessler and Kalsbeek (1992, p.265):

> "Increasing the sample overlap for some surveys rather than drawing the samples independently is known as positive coordination. A positive coordination is often searched in repeated surveys over time (panels) in order to obtain a better accuracy of statistics depending on correlated variables from two surveys. Reducing the overlap between samples for different surveys is known as negative coordination. A negative coordination is used in order to share more equally the response burden among responding units when statistics from surveys are not used together or are not correlated."

In positive co-ordination, the joint inclusion probability satisfies the conditions

$$\pi_k^1 \pi_k^2 < \pi_k^{1,2} \leq \min(\pi_k^1, \pi_k^2), \text{ for all } k \in U,$$

while in the case of negative co-ordination

$$\max(0, \pi_k^1 + \pi_k^2 - 1) \leq \pi_k^{1,2} < \pi_k^1 \pi_k^2, \text{ for all } k \in U.$$

The joint inclusion probability can be zero only if $\pi_k^1 + \pi_k^2 \leq 1$. Cotton and Hesse (1992b, p.27) give the definition of the sample co-ordination on the unit

[1] $http://forum.europa.eu.int/irc/dsis/coded/info/data/coded/en/Theme1.htm\#C$

level. Thus, for unit k the co-ordination is positive if $\pi_k^{1,2} > \pi_k^1 \pi_k^2$, and negative if $\pi_k^{1,2} < \pi_k^1 \pi_k^2$. They comment that "theoretically, we can have a positive co-ordination for some units and negative one for the others".

We are interested here in sample co-ordination procedures. The quality of a co-ordination procedure can be measured using four possible criteria as given in Ohlsson (1996):

1. the procedure provides a maximum/minimum overlap;
2. the sample design is respected in each selection;
3. independent sampling in new strata (for stratified designs) is assured;
4. the procedure can be easily applied.

The expected overlap n_{12} of two samples is given by

$$E(n_{12}) = \sum_{k \in U} \pi_k^{1,2}.$$

We have possible overlap in the case of:

- sampling on two occasions;
- two-phase sampling.
- independent samples.

Figure 1.1 illustrates possible overlapping samples.

In order to fix the frame of Chapters 2 and 3, we briefly present some methods to co-ordinate samples, which are linked to the thesis. We focus on positive sample co-ordination for two successive surveys, since positive and negative co-ordination can be seen as two aspects of the same problem. Our classification criterium is based on the use or not of the Permanent Random Numbers (PRNs). Thus, we classify the existent procedures as *PRN procedures* and *non-PRN procedures*. Other possible classification criteria are given in Ernst (1999):

- the procedure is sequential or simultaneous;
- the procedure is constructed for one selected unit, for small or large number of selected units;

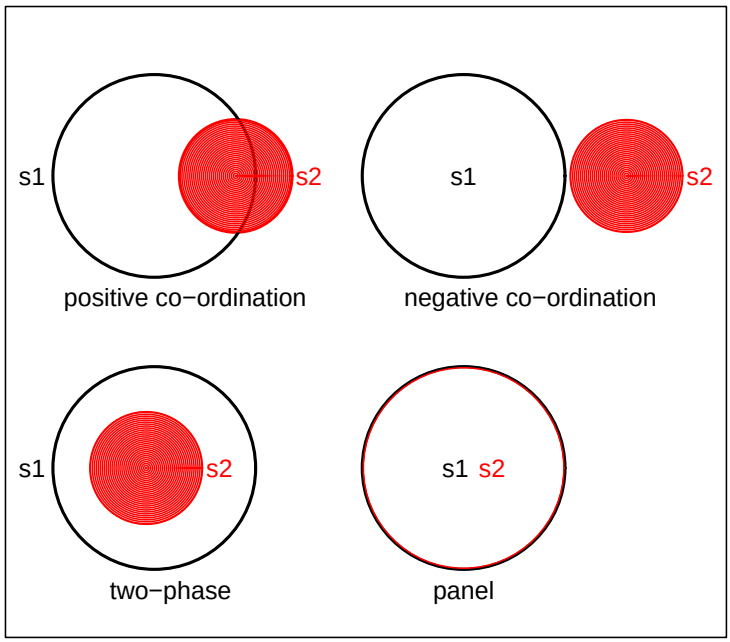


Figure 1.1: Possible overlapping samples $s_1,\ s_2$.

- the procedure maximizes or minimizes the overlap;
- the procedure can take into account different stratifications in the overlapped designs;
- the absolute upper bound defined as

$$\min(\pi_k^1, \pi_k^2) \tag{1.4}$$

  is reached or not;
- the procedure uses linear programming or not;

- the procedure assures the independence of sampling from stratum to stratum;

- how many surveys can be overlapped using the procedure?

Chapter 2 supposes that two sequential surveys are used. Chapter 3 presents a procedure which is sequential, it can be used for a large number of selected units, maximizes or minimizes the sample overlap, can take into account different stratifications, the absolute upper bound can be reached under some conditions, it does not use linear programming, assures the independence of sampling from stratum to stratum, and it is constructed for two surveys. However, the drawback of our procedure consists in the fact that the probability for each possible sample in the times 1 and 2 must be known, as in the linear programming procedures listed below. Chapter 3 studies also the conditions to reach the absolute upper bound as defined in expression (1.4).

Some papers summarize the research on the problem of maximizing / minimizing the sample overlap:

- Ernst (1999) for on overview on the non-PRN procedures;

- Ohlsson (1995a) for PRN procedures;

- Hesse (1998) for a review by country.

1.2.1 Some sample co-ordination procedures

Non-PRN procedures

a. Classical procedures

This category includes the methods based on the change of the inclusion probabilities, in the context of a stratified population. It is possible moreover that for each time period the stratification changes.

One of the "classical" procedures is the method of Keyfitz (1951). If the stratification remains the same in time, the method of Keyfitz (1951) is the optimal solution in the sense that the absolute upper bound defined in (1.4) is reached. The method considers the case of two designs with identical stratification. It selects one unit per stratum. A practical application of this method is described in Brick et al. (1987). The drawback of the Keyfitz method lies in the fact that only one unit per

stratum is selected. More information about Keyfitz method is given in Chapter 3. Two generalizations of the Keyfitz method for the case where the stratification is different at the two occasions were done by Kish and Scott (1971). Their methods select more than one unit per stratum, but in certain examples do not give an optimal solution.

b. Mathematical programming (more precisely linear programming) was used to solve the co-ordination problem of two samples. The application of the transportation problem (a particular form of the linear programming) in sample co-ordination is given by Raj (1968), Arthnari and Dodge (1981), Causey et al. (1985), Ernst and Ikeda (1995), Ernst (1996), Ernst (1998), Ernst and Paben (2002), Reiss et al. (2003). Ernst (1986, 1999) methods use linear programming. There are three principal forms of the transportation problem which are used in the context of two designs. One considers that the sample probabilities are known for all possible samples at each occasion. The first two forms are presented in Chapter 3. The third form uses the controlled rounding principle. Cox and Ernst (1982) are the first to state the controlled rounding problem in the form of a transportation problem. Then, the controlled rounding problem was used by Pruhs (1989); Ernst (1996, 1998); Ernst and Paben (2002). Cox and Ernst's method rounds the elements of a probability matrix to 0 or 1 and uses the simplex algorithm to solve this problem. Afterwards, the selection of the units in the current sample is achieved using another algorithm.

The drawback of using mathematical programming is its huge computational aspect, because for a population size N we have to enumerate up to 2^N distinct samples.

PRN procedures

Usually, the PRN methods are based on the following principle: for each unit $k \in U$, a uniform random number between 0 and 1 is associated from its birth until its death and kept for each survey. For this reason, these numbers are called *Permanent Random Numbers* (PRNs). The PRN of unit k is denoted by ω_k. If a unit disappears, its PRN disappears too; if a new unit appears in the population, a new PRN is assigned to this unit. There are also some techniques which permute the uniform random numbers, see Cotton and Hesse

(1992b); Rivière (2001).

Since Poisson sampling is the pioneer PRN technique, a special reference to this sampling and its extensions is given below.

a. Poisson sampling and its extensions

The PRN technique was introduced by Brewer (1972); Brewer et al. (1972) (see also Brewer et al., 1984) for the case of *Poisson sampling*. The principle of sample co-ordination in this case is the following: if the PRN ω_k is smaller than the inclusion probability, then unit k is included in the sample; otherwise it is not. The next Poisson sample is drawn from the updated population, using the same PRNs as before. Despite the fact that the inclusion probabilities vary from time to time due to the population and design changes, an optimal sample overlap can be obtained using PRNs. Indeed, the Poisson sampling technique guarantees the best possible co-ordination, i.e. $\pi_k^{1,2} = \min(\pi_k^1, \pi_k^2)$.

For negative co-ordination, in order to select a second non-overlapping sample, unit k is selected at time 2, if

$$\pi_k^1 < \omega_k < \pi_k^1 + \pi_k^2.$$

The procedure can be also used in the case of equal inclusion probabilities, where *Bernoulli sampling* is finally obtained (with $\omega_k \leq n/N$ as selection criterium).

A drawback of Poisson sampling is the random size of the sample, which has some undesirable consequences such as a possible small number of selected units in the sample (even 0), or high variance of the Horvitz-Thompson estimator in comparison with simple random sampling without replacement (see Hesse, 1998).

Extensions of Poisson sampling have been defined:

- *Modified Poisson sampling* suggested by Ogus and Clark (1971) (see also Brewer et al., 1972, 1984) was introduced in order to avoid the problem of null sample size. This is achieved by drawing more than one Poisson sample if zero size appears.
- *Collocated sampling* was proposed by Brewer (1972), Brewer et al. (1972) (see also Sunter, 1977b; Brewer et al., 1984; Cotton and Hesse, 1992b; Ernst et al., 2000). This technique was proposed in

order to reduce the variability of sample size for Poisson sampling and to avoid sample size equal to zero. The PRNs ω_k are sorted in ascending order and the rank R_k is addressed to each one of them. A single uniform [0,1] random number ϵ is generated. For each unit k, the quantity

$$u_k = \frac{R_k - \epsilon}{N}$$

is defined. If u_k is smaller then the inclusion probability of k, then unit k is selected. Collocated sampling is almost the same method as Poisson sampling, except that the u_k's replace the PRN. This adjustment has the effect of spreading uniformly the population units and removing any potential cluster of PRNs. The sample size becomes almost non-random. When a stratification is used, the numbers R_k depend on the size of the strata. Hesse (1998) pointed out that "it is difficult to use this technique for several samples at the same time, unless they adopt the same stratification and the same definition of scope."

- *Synchronized sampling* was introduced by Hinde and Young (1984). The method relies on assigning a PRN to each unit and selecting units whose random numbers lie in an interval. Briefly, in the case of positive co-ordination, the method is as follows. The first sample of size n is selected using sequential srswor with PRNs (see section b, fixed ordered procedures). For the next sampling, a selection interval $[s, e)$ is determined as follows. The starting point s is moved in the position of the first PRN. The end point e is positioned at the PRN of unit number $n+1$ to the right of the starting point. If the births or deaths occur in this interval or sample size changes, the interval is adjusted: it is extended to right in order to include more units or it is decreased to left in order to exclude units, until the desired sample size is obtained.
- *Conditional Poisson sampling* or rejective sampling was introduced by Hájek (1964). It is the sampling which maximizes the criterion of entropy given the fixed sample size and the first-order inclusion probabilities. Practically, in order to obtain a sample of fixed size n, a Poisson sampling is drawn; if the number of units in this trial is not equal to n, an other trial is realized and so on, until a sample of size n is finally drawn. This method is known as *rejective*

algorithm. The opportunity to put the rejective sampling in the extensions of Poisson sampling list is determined by the point of view of Brewer (2002), but an effective method to coordinate two or more samples has not yet been studied. This sampling design is revisited in subsection 1.3.2.

- *PoMix sampling* was introduced by Kröger et al. (1999). It is a mixture of Bernoulli and Poisson sampling. Let $z_k > 0$ be a size measure for unit k. The measure $\mathbf{z}$ is known for all units in the population. Let

$$\lambda_k = \frac{nz_k}{\sum_{\ell=1}^{N} z_\ell}, \tag{1.5}$$

for all $k \in U$ be the inclusion probabilities for a Poisson $\pi ps(x)$ sampling, where n is the expected sample size. Let $a \in [0, n/N]$ be a starting point. Compute the PoMix inclusion probabilities defined as

$$\widetilde{\pi}_k = a + \left(1 - a\frac{N}{n}\right)\lambda_k. \tag{1.6}$$

The unit k is included in the sample s if one of the following rules is satisfied:

a. $0 < \omega_k \leq a$;
b. $a < \omega_k \leq 1$ and $\widetilde{\pi}_k \geq \frac{\omega_k - a}{1-a}$.

The form of $\widetilde{\pi}_k$ permits us to have inclusion probability greater than a, in order to avoid the small $\widetilde{\pi}_k$ values (see Brewer, 2002, p.248). The next sample is drawn as in Poisson sampling, from the updated population, using the same PRNs and the same value of a, but possibly different z_k and n. If $a = 0$, we obtain Poisson πps sampling. If $a = n/N$, we obtain Bernoulli sampling. For all other values of a a Poisson-Bernoulli sampling mixture is determined, which is in fact Poisson sampling obtained by using the probabilities $\widetilde{\pi}_k$ instead of λ_k. The authors conducted Monte Carlo simulation studies and showed that such mixture provides estimates with smaller variance than Poisson πps sampling. As in the case of Poisson/Bernoulli sampling, the drawback of this method is, however, the random sample size.

b. Fixed ordered procedures refer to some procedures which sort quantities depending on PRNs, and draw samples with fixed size. They are:

srswor with PRN, order sampling, fixed size PoMix sampling and Ohlsson exponential sampling.

- *Random sort* or *sequential simple random sampling* or *simple random sample without replacement* (srswor) *with PRN* provides a fixed sample size n. The list of units is sorted in ascending or descending order of the ω_k. The sample is composed by the first n units (or the last n units) in the ordered list. This method is described by Fan et al. (1962). A proof that this technique generates a simple random sample without replacement is given for instance in Sunter (1977a) and in Ohlsson (1992). Simulations to compare the Kish and Scott (1971) method (the second one) and srswor with PRN in the case of stratified designs are given in Péa (2004).
- *Order sampling* is a class of sampling designs introduced by Rosén (1997a,b). To each unit $k \in U$ a random variable X_k with the cumulative distribution function (cdf) $F_k(t)$ and the probability density function $f_k(t),\ 0 \leq t < \infty$ is associated. We describe the procedure for the first survey. The second sample is drawn in the same way. A sample of fixed size n is obtained as follows: realizations of independent variables $X_1, X_2, \ldots, X_N$ are given; the first n units in increased order of X-values are selected. The random variables $X_1, \ldots, X_k, \ldots, X_N$ follow the same type of distribution, but are not always identically distributed. If X_k are identically distributed, we obtain *simple random sampling without replacement* and the inclusion probabilities are equal. If not, the inclusion probabilities are unequal. Various types of sampling rise for F_k following different distribution. In particular, we have:
  - *Uniform order sampling* or *sequential Poisson sampling* (Ohlsson, 1990, 1995b,a);
  - *Exponential order sampling* or *successive sampling* (Hájek, 1964);
  - *Pareto order sampling* (Rosén, 1997a,b; Saavedra, 1995).

  More details about the order πps sampling designs are given in Chapter 2.
- *Fixed size PoMix sampling* is a mixture between order sampling and PoMix sampling (see Kröger et al., 2003). It was introduced in order to obtain a fixed sample size equal to n, and to preserve the good

properties of the Pomix sampling concerning the variance in skewed populations. Let us consider the quantities λ_k as in expression (1.5). The fixed PoMix inclusion probabilities are defined as in (1.6). The mechanism to obtain the sample is the same as in order sampling. In the PoMix method, different types of sampling can be obtained by changing the parameter a and the sorted values. While for ranking values $\omega_k/\widetilde{\pi}_k$ and $a = 0$, we obtain sequential Poisson sampling, for the same ranking values and $a = n/N$, we get simple random sampling without replacement. Finally, for ranking values

$$\frac{\omega_k/(1-\omega_k)}{\widetilde{\pi}_k/(1-\widetilde{\pi}_k)},$$

and for all possible a, we obtain Pareto sampling. However, the quantities $\widetilde{\pi}_k$ are not equal to the inclusion probabilities π_k. The second sample is drawn in the same way, using the same PRNs, but possibly different n and z_k.

- *Ohlsson's exponential sampling* (Ohlsson, 1996) is a procedure which follows the same idea to sort values depending on PRNs for one unit selected by stratum. Let λ_k be the selection probability for unit k. It is assumed that $\sum_{k\in U} \lambda_k = 1$. The quantities $\xi_k = -\log(1-\omega_k)/\lambda_k$ are computed. The first unit in the sorted order of ξ_k is selected. The name of the procedure is given by the distribution of $\xi_k \sim \exp(1/\lambda_k)$. We prefer to denote this procedure "Ohlsson exponential sampling" in order to avoid the confusion with the exponential order sampling. The opportunity to put this method in our list is determined by the recent studies of Ernst (2001) and Ernst et al. (2004).

Other used methods in the official statistic centers, but not taken into account in the thesis, are enumerated below: the Jales method (the SAMU system) (Atmer et al., 1975), the EDS system (De Ree, 1983), the Ocean method (Cotton and Hesse, 1992a), the microstrata method (Rivière, 2001).

1.3 Order πps sampling design and maximum entropy sampling design with unequal probability and fixed sample size

Since Chapters 2 and 4 of the thesis are devoted to the sampling designs mentioned in the section title, some references to the already existing works about them is added below. For an interesting presentation of the two designs one can see Brewer (2002, p.259-264).

1.3.1 Order πps sampling design

The order sampling designs are developed by Rosén (1997a,b), based on an idea introduced by Ohlsson (1990). The advantage of the order sampling designs is their easy implementation. The order sampling implementation is linked to the notion of order statistics: to sort the N values of $X_1, .., X_k, ..X_N$ we need about $O(N \log N)$ operations. When N is large, this method is expensive. The implementation can be modified by computing only the first n order statistics $X_{(1)}, \ldots, X_{(n)}$ (see Gentle, 1998, p.125).

In order πps sampling, when an auxiliary information z_k is known for the whole population, the target inclusion probability λ_k is computed as in (1.5). The quantities λ_k are different from the true inclusion probability π_k. The order πps sampling designs are asymptotically πps since λ_k and π_k are very close to one another for large values of n (Rosén, 1997a, 2000). The first algorithm for computing the true inclusion probabilities was given by Aires (1999). She applied her algorithm for the Pareto case. Chapter 2 gives another algorithm to compute π_k in all three cases of order πps sampling. Recently, Ng and Donadio (2005) proposed another method to compute the same thing.

As we have already mentioned, the order samplings are used in sample co-ordination with PRN. The PRNs are used in the expression of the ranking variables. Our simulation results (not mentioned here) show that the three kinds of order πps sampling designs (uniform, exponential and Pareto) have the same performance concerning the sample overlap in the case of positive co-ordination. The same conclusion is given in Ohlsson (1999) for very small sample sizes (for the uniform and Pareto cases). The sample overlap results motivated us to compute the true inclusion probabilities for all three order πps sampling design in Chapter 2 (not only for the Pareto case, which seems

to be preferable in another papers for the good variance properties).

1.3.2 Maximum entropy sampling design with unequal probability and fixed sample size

Given the fixed sample size n and the first-order inclusion probabilities, the maximum entropy sampling design maximizes the entropy criterium defined as

$$-\sum_{s\in\mathcal{S}_n} p(s)\ln(p(s)),$$

subject to the constraints

$$\sum_{s\ni k, s\in\mathcal{S}_n} p(s) = \pi_k, \text{ and}$$

$$\sum_{s\in\mathcal{S}_n} p(s) = 1,$$

where $\mathcal{S}_n = \{s \subseteq U, |s| = n\}$. This sampling design was introduced by Hájek (1964) under the name of rejective sampling. The first implemented algorithm to draw a sample was the rejective algorithm (see previous subsection, PRN procedures). The name of *Conditional Poisson sampling* (CPS) was also used due to the method of conditioning a Poisson sample to a fixed sample size. Nevertheless, CPS can be implemented by at least six algorithms (different from the rejective) as given in Chen et al. (1994); Chen and Liu (1997), and Traat et al. (2004). Some algorithm performances are studied in Grafström (2005).

Other studies devoted to CPS are: Dupacová (1979); Milbrodt (1987); Jonasson and Nerman (1996). Methods to compute the inclusion probabilities for this sampling design have been studied by Chen et al. (1994); Chen and Liu (1997); Aires (1999); Deville (2000b) and more recently by Traat et al. (2004). Chapter 4 uses the method of Chen et al. (1994) to compute the first-order inclusion probabilities and the method of Deville (2000b) to compute the second-order inclusion probabilities . For a large discussion on the maximum entropy sampling one can see Tillé (2005).

Recently, Bondesson et al. (2004); Bondesson and Traat (2005) observed that the vector of the first-order inclusion probabilities $\boldsymbol{\pi}$ for a CPS design is

the eigenvector corresponding to the eigenvalue $n-1$ of a special matrix as given below

$$\mathbf{A}\boldsymbol{\pi} = (n-1)\boldsymbol{\pi},$$

where

$$\mathbf{A} = diag(\mathbf{1}^T\mathbf{C}) + \mathbf{C},$$

$$c_{k\ell} = \frac{\lambda_k(1-\lambda_k)}{\lambda_k - \lambda_\ell}, \quad \lambda_k \neq \lambda_\ell, \quad c_{kk} = 0,$$

and $\mathbf{1}$ is a column vector of 1's.

An interesting approach of multidimensional maximum entropy sampling design has been presented in Qualité (2004). Qualité (2004) comments also that the new French census system is based on a balanced sample drawn by using the cube method of Deville and Tillé (2004), and which is approximatively a maximum entropy sample under some balancing constraints.

Chen et al. (1994) provided two schemes for sample rotation based on the Metropolis-Hastings algorithm. According to our knowledge, sample coordination with maximum entropy sampling was studied by J.-C. Deville, but we can not give a concrete reference.

1.3.3 Relation between Pareto πps order sampling and maximum entropy sampling designs

Order πps sampling and maximum entropy sampling designs belong to the class of πps sampling designs, since the selection probabilities can be computed as in expression (1.5). Moreover, the order sampling is asymptotically πps: when $n \to \infty$, the true inclusion probabilities π_k converge to the prescribed ones λ_k (Rosén, 2000), under general conditions. For the maximum entropy sampling design the following result is available (Hájek, 1964)

$$\frac{\pi_k}{\lambda_k} \to 1, \quad 1 \leq k \leq N, \tag{1.7}$$

uniformly in k if $\sum_{k\in U} \lambda_k(1-\lambda_k) \to \infty$, when $n \to \infty, N-n \to \infty$.

The relation between Pareto πps order sampling and CPS designs was studied by Aires (2000) and Bondesson et al. (2004). Based on the asymptotic results obtained by Rosén (1997a,b) for Pareto πps sampling and Hájek (1964, 1981) for CPS, Aires (2000) compared the two designs by using an asymptotic

variance formula. She concluded that the "resulting variances and second-order inclusion probabilities are very similar" for the two sampling designs (for the same λ_k and n).

The two sampling design probabilities have been compared by Bondesson et al. (2004). They have similar expressions as written below

- for Pareto πps order sampling (Traat et al., 2004)

$$p_{Par}(s) = \left(\sum_{k\in s} c_k\right) \prod_{k\in s} \lambda_k \prod_{k\notin s} (1-\lambda_k),$$

where

$$c_k = \int_0^\infty \frac{x^{n-1}}{1+\tau_k x} \prod_{i\in U} \frac{1+\tau_i}{1+\tau_i x} dx, \tag{1.8}$$

with $\tau_k = \lambda_k/(1-\lambda_k)$.

- for CPS (Hájek, 1981)

$$p_{CPS}(s) = C_{CPS} \prod_{k\in s} \lambda_k \prod_{k\notin s} (1-\lambda_k),$$

where $C_{CPS} = 1/\sum_{s\in\mathcal{S}_n} \prod_{k\in s} \lambda_k \prod_{k\notin s}(1-\lambda_k)$.

The idea of Bondesson et al. (2004) is to generate a conditional Poisson sample from the Pareto πps order probability, by using the acceptance-rejection algorithm, since the sampling probabilities are close to one another. Their method is motivated by the large time execution of the other methods. However, Chapter 4 uses the rejective algorithm, since our goal is not restricted on the performance of the algorithms. In the same paper, Bondesson et al. (2004) provide a formula that allows passing from the inclusion probabilities for CPS (denoted below as π_k^{CPS}) to the Pareto ones (π_k^{Par}). This is given by

$$\pi_k^{Par} = \frac{\sum_{\ell\in U} c_\ell \pi_{k\ell}^{CPS}}{\sum_{\ell\in U} c_\ell \pi_\ell^{CPS}},$$

and

$$\pi_{k\ell}^{Par} = \frac{\sum_{i\in U} c_i \pi_{k\ell i}^{CPS}}{\sum_{i\in U} c_i \pi_i^{CPS}},$$

with c_k as given in expression (1.8).

1.4 Estimation in unequal probability sampling designs

Consider a πps sampling design with fixed sample size n. Suppose that an auxiliary variable x_k is available for all $k \in U$. In the prediction approach, the values $y_1, \ldots, y_N$ are realizations of the random variables $Y_1, \ldots, Y_N$. The Horvitz-Thompson estimator $\widehat{t}_\pi$ defined in expression (1.1) is often studied under a superpopulation model ξ

$$E(Y_k) = x_k\beta, \quad \text{var}(Y_k) = \sigma^2 x_k^2, \quad \text{cov}(Y_k, Y_\ell) = 0, \quad k \neq \ell,$$

or

$$y_k = \beta x_k + \varepsilon_k, \tag{1.9}$$

with

$$E_\xi(\varepsilon_k) = 0, \quad \text{var}_\xi(\varepsilon_k) = \sigma_k^2, \quad E_\xi(\varepsilon_k\varepsilon_\ell) = 0, \quad k \neq \ell, \quad \sigma_k^2 = x_k^2\sigma^2.$$

The anticipated variance (see Isaki and Fuller, 1982) is the variance of $\widehat{t}-t_y$, under the sampling design $p(.)$ and the model ξ. Let $\widehat{t}$ be an estimator of t_y. The anticipated variance is defined as

$$\text{var}(\widehat{t} - t_y) = E_\xi E_p[(\widehat{t} - t_y)^2] - \{E_\xi E_p[(\widehat{t} - t_y)]\}^2. \tag{1.10}$$

For an unbiased estimator, the expression (1.10) simplifies to $E_\xi E_p[(\widehat{t} - t_y)^2]$. Consider the regression estimator of t_y

$$\widehat{t}_{reg} = \sum_{k\in U} \widehat{y}_k + \sum_{k\in s} \frac{y_k - \widehat{y}_k}{\pi_k},$$

where $\widehat{y}_k = \widehat{\beta}x_k$. Särndal et al. (1992, p.451-452) showed that an approximation of the anticipated variance $ANV(\widehat{t}_{reg})$ is equal to the lower bound introduced by Godambe and Joshi (1965) for an unbiased estimator

$$\sum_{k\in U} \sigma_k^2 \left(\frac{1}{\pi_k} - 1\right).$$

Result 12.2.1 of Särndal et al. (1992, p.452) gives the justification of using a πps sampling design, when the model ξ is appropriate. Let n_s be the sample

size. That is for a sampling design with $E(n_s) = n$, when t is estimated by $\widehat{t}_{reg}$, an optimal design, in the sense that $ANV(\widehat{t}_{reg})$ is minimized, has the first-order inclusion probabilities equal to

$$\pi_k = \frac{n\sigma_k}{\sum_{\ell \in U} \sigma_\ell}, \text{ for all } k \in U.$$

The minimum value of the approximate variance is then

$$ANV(\widehat{t}_{reg}) = \frac{1}{n}\left(\sum_{k \in U} \sigma_k\right)^2 - \sum_{k \in U} \sigma_k^2.$$

Under the model ξ and for a fixed size πps sampling design, $\widehat{t}_{reg} = \widehat{\beta}\sum_{k \in U} x_k + \widehat{t}_\pi + \widehat{\beta}\sum_{k \in s} x_k/\pi_k$ and $\widehat{t}_\pi$ are equal. Consequently, in this case, $\widehat{t}_\pi$ reaches also the lower bound of Godambe and Joshi, and is the best in this sense.

In the literature, there are many schemes for selecting πps samples (50 schemes are listed in Brewer and Hanif, 1983). It is also important that the variance of the Horvitz-Thompson estimator always be smaller than the variance obtained by sampling with replacement. Some proofs are given in the literature: for Midzuno (1952) procedure by Chaudhuri (1974), for Sampford (1967) procedure by Gabler (1981), for Chao (1982) procedure by Sengupta (1989). Recently, Qualité (2005) gives a proof for the maximum entropy sampling or conditional Poisson sampling (Hájek, 1964).

Chapter 4 studies the behavior of some estimators of the variance of $\widehat{t}_\pi$. The topic of Chapter 4 is the maximum entropy sampling design with unequal probability and fixed sample size. This sampling design belongs to the class of πps schemes, since Hájek (1964) has provided the expression (1.7). Consequently, one of the used models in the simulation section is equivalent to model (1.9), that is $y_k = 5x_k(1 + \varepsilon_k), \varepsilon_k \sim N(0, 1/3)$. The inferential approach taken in Chapter 4 is basically design-based. However, some model approach ideas are taken into account in the possible explication of the Horvitz-Thompson variance estimator behavior under a model similar to (1.9).

What kind of variance estimator of $\widehat{t}_\pi$ must be used? It is necessary to employ an estimator which uses the second-order inclusion probability $\pi_{k\ell}$? In Brewer (2002, p.143), Ray Chambers in his "response from supervisor" comments:

> "It is also strange that the $\pi_{k\ell}$ should be seen as a necessary input into the variance estimation process. We know that certain unequal

> probability sample designs yield very similar Horvitz-Thompson variances though they have quite different $\pi_{k\ell}$ (Hartley and Rao, 1962; Asok and Sukhatme, 1976). Is it not possible to estimate the variance of the Horvitz-Thompson estimator in such circumstances without recourse to the $\pi_{k\ell}$?"

The criterium of *simplicity* is used in the recommendation given at the end of Chapter 4. Wolter (1985, p.3) emphasized that the choice of an appropriate variance estimator in complex surveys (such an unequal probability sampling survey) typically is a difficult one, "involving the accuracy (usually measured by mean square error), timeliness, cost, *simplicity*, and other administrative considerations". However, Wolter added that "compromises will have to be made because different analysis of the same data may suggest different variance estimators".

1.5 Calibrated imputation and Cox rounding algorithm

Chapter 6 is devoted to the development of an imputation method for qualitative data, which uses the calibration technique (see Deville and Särndal, 1992). We review below some notions about calibration, non-response and imputation.

Let $\mathbf{y}$ be the variable of interest. Suppose to be known the inclusion probability π_k for each unit $k \in s$, where s is the current sample. We can estimate the population total $t_y = \sum_{k\in U} y_k$ with the Horvitz-Thompson estimator $\widehat{t}_\pi = \sum_{k\in s} d_k y_k$, where $d_k = 1/\pi_k$. Deville and Särndal (1992) developed the *calibration estimator* defined as

$$\widehat{t}_{CAL} = \sum_{k\in s} w_k y_k,$$

where

$$\sum_{k\in s} w_k \mathbf{x}_k = \sum_{k\in U} \mathbf{x}_k = \mathbf{t}_x, \tag{1.11}$$

for a row vector of auxiliary variables $\mathbf{x}_k = (x_{1k}, \ldots, x_{Jk})$, for which $\mathbf{t}_x$ is known. The equation (1.11) is called the *calibration equation*. Deville and Särndal required that the difference between the set of sampling design weights

d_k and $w_k, k \in s$, satisfying equation (1.11), minimizes some function. The function to minimize is

$$\sum_{k\in s} d_k q_k G_k(w_k/d_k) - \boldsymbol{\lambda}(\sum_{k\in s} w_k \mathbf{x}_k - \mathbf{t}_x),$$

where $\boldsymbol{\lambda}$ is the vector of the Lagrange multipliers. Minimization leads to the calibrated weights $w_k = d_k F_k(\mathbf{x}_k'\boldsymbol{\lambda}/q_k)$, where q_k is a weight associated with unit k, unrelated to d_k, that accounts for heteroscedastic residuals from fitting $\mathbf{y}$ on $\mathbf{x}$, and F_k is the inverse of the $dG_k(u)/du$ function with the property that $F_k(0) = 1, \quad F_k'(0) = q_k > 0$.

Frequently, the variable $\mathbf{y}$ is observed for only a part of the sample s. Thus a random set of respondents $r \subseteq s$ is observed according to a certain non-response mechanism. Särndal et al. (1992, p.558) suggested the use of two-phase sampling, where the response mechanism is considered as the second phase. Thus t_y is estimated by

$$\widehat{t}_{\pi^*} = \sum_{k\in r} \frac{y_k}{\pi_k p_k}, \tag{1.12}$$

where $p_k = Pr[k \in r \mid s]$ is the probability that unit k responds, given that the sample s was selected. Sometimes p_k can be estimated by $\widehat{p}_k$, and it is replaced in the expression (1.12).

The literature distinguishes between unit non-response and item non-response. In the former case the sample unit does not respond to any item. In the latter, the sample unit responds to some of the requested items, yet not to all. Imputation is a method to treat item non-response, and consists in replacing a missing value by a proxy one. A large number of imputation methods can be approximately described by the general model

$$y_k = f(\mathbf{x}_k'\boldsymbol{\beta}) + \varepsilon_k, \quad k \in r.$$

When $\mathbf{y}$ is a qualitative variable (with values 0,1), $Pr[y_k = 1]$ can be estimated by a logistic model

$$f(u) = \frac{\exp(u)}{1 + \exp(u)},$$

and $f(\mathbf{x}_k'\widehat{\boldsymbol{\beta}}) = \widehat{Pr}[y_k = 1]$. The use of the Poisson sampling in order to impute y_k gives the solution

$$\dot{y}_k = \begin{cases} 1 & \text{with probability } f(\mathbf{x}_k'\widehat{\boldsymbol{\beta}}), \\ 0 & \text{with probability } 1 - f(\mathbf{x}_k'\widehat{\boldsymbol{\beta}}). \end{cases}$$

The term of "Poisson sampling" in this context refers to a very simple application of the inverse CDF algorithm (see Gentle, 1998, p.42,47).

The calibrated imputation consists in computing final imputed values which are close to the initial imputed ones and are calibrated to satisfy the calibration equation(s). The initial imputed values are given by using an imputation model. It is possible to estimate p_k by using the calibration technique, when we desire to have

$$\sum_{k \in r} \frac{\mathbf{x}_k}{\pi_k p_k} = \sum_{k \in s} \frac{\mathbf{x}_k}{\pi_k}.$$

Now the set of calibration weights $\{w_k \mid k \in r\}$ minimize a distance function between them and the set $\{a_k = d_k/p_k \mid k \in r\}$, subject to satisfying the calibration equation. The calibration weights are $w_k = d_k F_k(\mathbf{x}_k' \boldsymbol{\gamma})$ (without taking into account the weights q_k). In this case, the estimator of the response probability for unit $k \in r$ is given by

$$\widehat{p}_k = \frac{1}{F_k(\mathbf{x}_k' \boldsymbol{\gamma})}.$$

$F_k(.)$ equals to $\exp(.)$ or $1 + \exp(.)$ are good choices. The latter form corresponds to a response probability fitted by a logistic function.

Lundström and Särndal (1999) defined two information levels called Info-S ($\mathbf{x}_k$ is known for all $k \in s$) and Info-U ($\mathbf{t}_x$ is known and $\mathbf{x}_k$ is known for all $k \in s$). Chapter 6 uses the Info-S information level, and takes also into account the application of the Poisson sampling in non-response theory, as given above.

Which is the relation between sample co-ordination and imputation for qualitative data? Apparently, none. However, we can use the same algorithm to do both sample co-ordination and qualitative data imputation. It is the case of the Cox (1987) algorithm.

The Cox's algorithm generates unbiased rounding of two-way tables. Motivated by this algorithm, Pruhs (1989) gives an algorithm for sample co-ordination by using the graph theory. Ernst (1999) emphasized that his result from 1996 (see Ernst, 1996) is the same as Pruhs (1989). Deville and Tillé (2000) used the Cox's algorithm to partition a studied population in non-overlapping subsets. They also gave an application of their method in the case of negative sample co-ordination, in order to draw simultaneously two samples. Details on the Cox algorithm are given in Chapter 5.

In our case, a transformation of the Cox's algorithm (denoted as *Cox weighted algorithm*) is used in imputation for qualitative data. Chapter 5

gives details about the implementation of the Cox weighted algorithm, and Chapter 6 presents the frame of its application.

1.6 Summary of chapters

Each chapter of the thesis is self-contained. Chapter 2 is devoted to the computation of the first-order inclusion probabilities in order πps sampling designs, and to an empirical approximation of the joint inclusion probability of a unit in two positively co-ordinated ordered samples. It is based on the paper Matei and Tillé (2005a). Chapter 3 is based on the paper Matei and Tillé (2005c) and focuses on the development of a new strategy in sample co-ordination. Chapter 4 is based on the paper Matei and Tillé (2005b) and it is an empirical study of the variance approximations and estimators in maximum entropy sampling design with unequal probabilities and fixed sample size, using Monte Carlo simulation. Chapters 5 and 6 are based on the papers Favre et al. (2004) and Favre et al. (2005) and provide an imputation method for qualitative data using the Cox's algorithm (Cox, 1987). The notation is not always the same, because it depends on the presented papers.

Chapter 2

In this chapter two algorithms applied in order πps sampling designs are given. The first algorithm computes exactly the first-order inclusion probability for unit k in the case of uniform, exponential and Pareto order sampling. Let $X_1, \ldots, X_N$ be the ranking variables with the cdfs $F_1, \ldots, F_N$. Our algorithm is based on the method given by Cao and West (1997) to determine the cumulative distribution function of the rth order statistic (denoted below as $F_{(r)}$) in the case of independent, but not identically distributed random variables. Their formula is as follows

$$F_{(r)}(t) = F_{(r-1)}(t) - J_r(t)[1 - F_{(1)}(t)], \text{ for } r = 2, \ldots, N, \tag{1.13}$$

where

$$F_{(1)}(t) = 1 - \prod_{j=1}^{N} \bar{F}_j(t),$$

$$J_r(t) = \frac{1}{r-1} \sum_{i=1}^{r-1} (-1)^{i+1} L_i(t) J_{r-i}(t),$$

and

$$L_i(t) = \sum_{j=1}^{N} \left[\frac{F_j(t)}{\bar{F}_j(t)} \right]^i ,$$

with $J_1(t) = 1$, and $\bar{F}_j(t) = 1 - F_j(t)$. Let $X_{(n),k}^{N-1}$ be the n^{th} order statistic out of $N-1$ random variables $X_1 \ldots, X_{k-1}, X_{k+1}, \ldots, X_N$ (computed without X_k) and $F_{(n),k}^{N-1}$ its cdf. To compute effectively π_k we use the formula given in Aires (1999)

$$\pi_k = Pr(k \in s) = Pr(X_k < X_{(n),k}^{N-1}) = \int_0^\infty [1 - F_{(n),k}^{N-1}(t)] f_k(t) dt. \qquad (1.14)$$

The random variables X_k and $X_{(n),k}^{N-1}$ are independent. Equation (1.14) is the convolution of their distributions. The second algorithm is used to approximate the joint inclusion probability of unit k in two positive co-ordinated ordered samples s_1, s_2, with $|s_1| = n1, |s_2| = n_2$. Our approximation uses two steps:

1) the event $C_k = \left(X_k < X_{(n_1),k}^{N-1}, Y_k < Y_{(n_2),k}^{N-1} \right)$ is approximated by

$$\left(\max(X_k, Y_k) < \min(X_{(n_1),k}^{N-1}, Y_{(n_2),k}^{N-1}) \right);$$

2) the cdf

$$F_{\min\ X_{(n_1),k}^{N-1}, Y_{(n_2),k}^{N-1}}(t)$$

is approximated by

$$\max(F_{X_{(n_1),k}^{N-1}}(t), F_{Y_{(n_2),k}^{N-1}}(t)).$$

The result of the two steps above is the formula

$$\pi_k^{1,2} \approx 1 - \int_0^1 \max(F_{X_{(n_1),k}^{N-1}}, F_{Y_{(n_2),k}^{N-1}}) \left(F^{-1}_{\max(X_k, Y_k)}(t) \right) dt.$$

This approximation gives results very close to the simulated ones.

Chapter 3

This chapter concerns sample co-ordination of two designs, using a non-PRN approach. In positive co-ordination, the absolute upper bound given by

$$\sum_{k \in U} \min(\pi_k^1, \pi_k^2)$$

is not always reached. Yet, there are a few cases where this bound is reached. These include: the Keyfitz (1951) method for one unit selection when both designs are identically stratified, the bidimensional simple random sampling without replacement (Cotton and Hesse, 1992b) and the bidimensional Poisson sampling (Cotton and Hesse, 1992b). Sufficient and necessary conditions to reach the absolute upper bound are given in this chapter. Similar conditions are given for negative co-ordination, in order to reach the absolute lower bound defined by

$$\sum_{k \in U} \max(\pi_k^1 + \pi_k^2 - 1, 0).$$

We develop an algorithm based on Iterative Proportional Fitting procedure (Deming and Stephan, 1940), in order to give an optimal solution of co-ordination (when the absolute upper or lower bound are reached). The input of this algorithm consists of the probabilities of all possible samples on the first and second time occasions. The output is the matrix of the joint sample probability $\mathbf{P} = (p_{ij})_{m \times q}$. For a fixed sample s_i^1 on the first occasion, the matrix $\mathbf{P}$ enables us to choose the sample s_j^2 on the second time occasion via $p(s_j^2 | s_i^1)$ computation. In the case where the absolute upper or the absolute lower bound cannot be reached, a message is given.

Chapter 4

Another subject of this thesis treats variance approximation and estimation in unequal probability sampling designs with fixed sample size. The maximum entropy sampling design with unequal probabilities and fixed sample size was chosen in this study since methods to compute the first-order and the second-order inclusion probabilities are available. Chapter 4 uses the method of Chen et al. (1994) for computing the first-order inclusion probabilities from the prescribed ones, and the method of Deville (2000b) for computing the second-order inclusion probabilities. The prescribed inclusion probabilities are proportional to a given size measure. Seven variance approximations and

twenty variance estimators have been compared using simulations. In the approximation class, the fixed-point approximation (Deville and Tillé, 2005) has the best performances. We distinguish three classes of estimators. The first class includes the Horvitz-Thompson estimator (Horvitz and Thompson, 1952) and the Sen-Yates-Grundy estimator (Yates and Grundy, 1953; Sen, 1953). These use the first-order and the joint inclusion probabilities. The second class uses only the first-order inclusion probabilities for all $k \in s$. The variance estimators in the third class use first-order inclusion probabilities, but for all $k \in U$. In this chapter, a large Monte Carlo study is provided. Three population are used in our simulations. These populations are regarded in terms of correlation coefficient between the variable of interest $\mathbf{y}$ and the auxiliary variable $\mathbf{x}$. The size of the correlation between these two variables affect the behavior of the Horvitz-Thompson estimator. According to our simulations, the estimators which use only the first-order inclusion probabilities have similar performances, regardless the correlation between $\mathbf{y}$ and $\mathbf{x}$. The use of the first-order inclusion probabilities over the whole population and the joint inclusion probabilities does not lead to more accurate variance estimators in the case of a maximum entropy sampling design with unequal probability and fixed sample size. The use of a simple estimator such as the one given in Deville (1993) seems to be a good choice.

Chapter 5

In this chapter, a transformation of the Cox's algorithm (Cox, 1987) is given. The original algorithm is a procedure for unbiased controlled rounding (a two-way table $\mathbf{A}$ with elements between 0 and 1 is transformed into a new one $R(\mathbf{A})$, by randomly rounding the elements to 0 or 1, in order to preserve the margins; additionally, the rounding procedure is unbiased i.e. $E\left(R(\mathbf{A})\right) = \mathbf{A}$).

We use a variant of the Cox's algorithm to make imputation. A direct application of the algorithm is not feasible here due to the weighting system which must be taken into account. Cox used in this algorithm a supplementary row and column, in order to have integer total rows and columns. The weighting system makes the total columns to be non-integers. The new variant of the Cox's algorithm (denoted the Cox weighted algorithm) is developed, in order to avoid this problem. However, at the end of the weighted algorithm some elements are not rounded. The solution is to apply the Poisson sampling scheme, in order to set all the elements to 0 or 1. In order to measure the

gain by using the Cox weighted algorithm instead of the Poisson sampling, the variance of imputation is measured via simulations. The results are in favor of the Cox weighted algorithm.

Chapter 6

Handling non-response is a common problem in survey sampling. This problem can be solved by using either a re-weighting technique to adjust for unit non-response, or an imputation technique in the case of item non-response. Chapter 6 proposes a solution to the non-response problem using a new imputation technique. A qualitative variable $\mathbf{y}$ with missing values for a subset of units (item non-response) is considered. The variable $\mathbf{y} = (y_{kj})$ takes v possible exclusive values, for all $k \in U$, and $j = 1, \ldots, v$. That is $y_{kj} = 1$ if the unit k takes the value j, and 0 otherwise, with $\sum_{j=1}^{v} y_{kj} = 1$, for all $k \in U$. The set of respondents is denoted by $R \subseteq S$, where S is the current sample. A set of auxiliary information $x_1, \ldots, x_J$ is available for whole S. Thus, each unit k has attached a row vector $\mathbf{x}_k = (x_{k1}, \ldots, x_{kJ})$. Each unit $k \in S$ is assumed to have a response probability $Pr[k \in R_j]$, where R_j is the column j of the 'matrix' R (see Fig. 6.1). However, in practice the response probabilities are unknown and they have to be estimated.

The proposed method is based on the following five steps: editing (when logical rules are used to identify some possible values for non-response items), estimation of totals $\widehat{Y}_j$ (based on the calibration technique), individual estimation of $p_{kj} = Pr[y_{kj} = 1|\mathbf{x}_k]$, for all $k \in S \backslash R$, calibration on the marginal totals by using IPF procedure (since after the phases of logical rules and the estimation of p_{kj}, the totals for each category j are not respected), and imputation (by using the Cox weighted algorithm, see the section above).

Chapter 2

Computational aspects in order πps sampling designs

Abstract
Order sampling is based on the following principle: the population units are ordered by a ranking variable; a sample is drawn by taking the first n units in this order. For order πps sampling design a new method to compute the inclusion probabilities is given. This method has the advantage to reduce the time execution. We use this algorithm to compute the inclusion probabilities in the case of uniform, exponential and Pareto order πps sampling designs.

The order sampling can be used in sample co-ordination with permanent random numbers. For the case of two positive co-ordinated ordered samples a method to approximate the joint inclusion probability of a unit in both samples is given. All the presented methods use numerical integration.
Key words: survey sampling, πps, fixed sample size, permanent random numbers, co-ordination over time, order statistics, numerical integration.

2.1 Introduction

The order sampling designs is a class of sampling schemes developed by Rosén (1997a,b). The basic idea of ordered sampling is the following. Let $U = \{1, \ldots, k, \ldots, N\}$ be a finite population. The unit k is considered as reference unit. Suppose we have N independent random variables $X_1, \ldots, X_k, \ldots X_N$ (one variable is associated to each unit) usually called 'ordering variables' or 'ranking variables'. Each X_k has a continuous cumulative distribution function (cdf) F_k defined on $[0, \infty)$, and a density function $f_k, k = 1, \ldots, N$. Order sampling with fixed sample size n and order distributions

$$\mathbf{F} = (F_1, \ldots, F_k, \ldots, F_N)$$

is obtained by setting the values X_k in increasing order of magnitude and taking from U the first n units in this order. The sample obtained in this way is a random sample without replacement and has fixed size n. $X_1, \ldots, X_k, \ldots, X_N$ follow the same type of distribution, but are not always identically distributed. In the case where X_k's are identically distributed, a simple random sampling without replacement is obtained, and the inclusion probabilities are equal. Otherwise the inclusion probabilities are unequal.

We focus on the $\pi ps(x)$ sampling, where the quantities λ_k are called 'target inclusion probabilities' and are calculated according to

$$\lambda_k = \frac{n z_k}{\sum_{i=1}^{N} z_i}, \ k = 1, \ldots, N.$$

The quantity $z_k > 0$ denotes an auxiliary information associated to unit k and known for all units in the population. We assume that $0 < \lambda_k < 1, \ \forall \ k = 1, \ldots, N$. Rosén (1997a) defined the order πps sampling design with fixed distribution shape H and target inclusion probabilities $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_k, \ldots, \lambda_N)$ by getting

$$F_k(t) = H\left[tH^{-1}(\lambda_k)\right],$$

and

$$X_k = \frac{H^{-1}(\omega_k)}{H^{-1}(\lambda_k)},$$

where H is a distribution function defined on $[0, \infty)$, and $\boldsymbol{\omega} = (\omega_k)_N$ is a vector of independent and identically distributed (iid) $U[0, 1]$ random variables. This type of sampling is known to be asymptotically a πps sampling (Rosén,

1997a). Different distributions F_k result in various types of order sampling. In particular we have:

1. Uniform order sampling or sequential Poisson sampling (Ohlsson, 1990, 1998), which uses uniform ordering distributions. In this case

$$X_k = \frac{\omega_k}{\lambda_k},\ F_k(t) = \min(t\lambda_k, 1), \forall\ k \in U.$$

2. Exponential order sampling or successive sampling (Hájek, 1964), which uses exponential ordering distributions. In this case

$$X_k = \frac{\ln(1-\omega_k)}{\ln(1-\lambda_k)},\ F_k(t) = 1-(1-\lambda_k)^t, \forall\ k \in U.$$

3. Pareto order sampling (Rosén, 1997a,b; Saavedra, 1995), that uses Pareto ordering distributions. In this case

$$X_k = \frac{\omega_k(1-\lambda_k)}{\lambda_k(1-\omega_k)},\ F_k(t) = \frac{t\lambda_k}{1-\lambda_k+t\lambda_k}, \forall\ k \in U.$$

   The Pareto πps sampling minimizes estimator variances in the class of order sampling schemes with fixed shape (see Rosén, 1997b). For Pareto πps order sampling, see also Aires (1999, 2000) and Holmberg and Swensson (2001). For a formula for the probability function of the design, see also Traat et al. (2004).

Generally, we write the distribution shape H as a generalized Pareto distribution (GPD) function

$$GPD(t,a,b) = \begin{cases} 1-(1-\frac{bt}{a})^{1/b}, & b \neq 0, \\ 1-\exp(-\frac{t}{a}), & b = 0, \end{cases} \tag{2.1}$$

where for $b = 1, a = 1$ we obtain the uniform order sampling, for $b = 0, a = 1$ we have the exponential order sampling, and for $b = -1, a = 1$ we have the Pareto order sampling.

Order sampling design can be used in sample co-ordination over time with Permanent Random Numbers (PRNs). PRN $\omega_k \sim U[0,1]$ associated to each unit $k \in U$ are used. There are two kinds of sample co-ordination: positive co-ordination, when the goal is to maximize the number of common units in

two or more samples, and negative co-ordination, when the goal is to minimize the same number. The positive co-ordination of two or more ordered samples drawn at different time periods is possible by using the same ω_k over time, $\forall\ k \in U$. For a negative co-ordination, the quantities $1 - \omega_k$ are taken into account instead of ω_k in the expression of X_k. The sequential Poisson sampling was used in sample co-ordination in the Swedish Consumer Price Index survey (see Ohlsson, 1990). We are interested here in sample co-ordination since an approximation of the inclusion probability of unit k in two samples is given in Section 2.3.

Let s be an ordered sample, and let $\pi_k = Pr(k \in s)$ be the first-order inclusion probability of the unit k in this sample, $k = 1, \ldots, N$. The values λ_k do not coincide with π_k because the latter depends on the ranks. Rosén (2000) has showed that

$$\frac{\pi_k}{\lambda_k} \to 1, \text{ when } n \to \infty.$$

It is interesting to compute exactly the true inclusion probability or the "factual inclusion probability" π_k for the cases where n is quite small. "Factual inclusion probability" is the expression used by Rosén (2000) to denote the quantity π_k. Since no analytic solution is readily available, numerical integration is implemented to compute the true inclusion probabilities. In the case of Pareto order πps sampling Aires (1999) provided an algorithm to compute numerically the first-order and second-order inclusion probabilities. She uses a double recursion in her algorithm in order to compute a cumulative distribution function for order statistics. Due to the formula provided by Cao and West (1997), a new method to compute the inclusion probabilities for an order sampling design is given. This method uses a simple recursion in order to compute the cumulative distribution function (cdf) of the kth order statistic in the case of independent, but not identically distributed random variables. The algorithm complexity is $O(N^2 n^3 p)$, where p is the number of splits in the applied numerical method. We use this algorithm to compute the inclusion probabilities in the case of uniform, exponential and Pareto order πps sampling design. The results show that the true inclusion probabilities are close to the target inclusion probabilities. In order to check the accuracy of the numerical results the total control is used ($\sum_{k \in U} \pi_k = n$).

The algorithms are implemented in C++ language and use only iterative methods also for the recurrent relations in order to minimize the time execu-

tion. The tests are made on a Pentium 4, 2.8 Ghz computer processor. The numerical integrations are realized using the Simpson method.

The article is organized as follows. Section 2.2 is dedicated to the computation of π_k given λ_k. In the same section, the recurrence formula for cdf of order statistics derived from independent and not identically distributed random variables (Cao and West, 1997) is recalled. The inverse method, to compute λ_k from π_k is also given. Section 2.3 presents an algorithm to compute an approximation of the joint inclusion probability of unit k in two co-ordinated ordered samples. Simulation results are presented in order to compare the proposed approximation with the simulated values. Finally, Section 2.4 presents concluding remarks.

2.2 First-order inclusion probabilities computation

Let $X_{(1)}, , \ldots, X_{(k)}, \ldots, X_{(N)}$ denote the order statistics of $X_1, \ldots, X_k, \ldots X_N$. Let $F_{(k)}$ $(k = 1, \ldots, N)$ denote the cumulative distribution function (cdf) of the k^{th} order statistic $X_{(k)}$, and let $f_{(k)}$ be its probability density function. We make the following definitions: denote the n^{th} order statistic out of N random variables by $X^N_{(n)}$, and its cdf by $F^N_{(n)}$; denote the n^{th} order statistic out of $N-1$ random variables $X_1 \ldots, X_{k-1}, X_{k+1}, \ldots, X_N$ (computed without X_k) by $X^{N-1}_{(n),k}$, and its cdf by $F^{N-1}_{(n),k}$.

Let s be an ordered sample with fixed size n. In order to compute the first-order inclusion probabilities $\pi_k = Pr(k \in s)$ we use the relation given in Aires (1999)

$$\pi_k = Pr(k \in s) = Pr(X_k < X^{N-1}_{(n),k}) = \int_0^\infty [1 - F^{N-1}_{(n),k}(t)] f_k(t) dt. \qquad (2.2)$$

The random variables X_k and $X^{N-1}_{(n),k}$ are independent. Equation (2.2) is the convolution of their distributions.

2.2.1 Recurrence relations on cdf of order statistics

In equation (2.2) it is necessary to compute $F^{N-1}_{(n),k}(t)$ by using an efficient algorithm. Cao and West (1997) provide a recurrence formula for the cdf of

the order statistics for independent, but not identically distributed random variables. Their formula is as follows

$$F_{(r)}(t) = F_{(r-1)}(t) - J_r(t)[1 - F_{(1)}(t)], \text{ for } r = 2, \ldots, N, \tag{2.3}$$

where

$$F_{(1)}(t) = 1 - \prod_{j=1}^{N} \bar{F}_j(t),$$

$$J_r(t) = \frac{1}{r-1} \sum_{i=1}^{r-1} (-1)^{i+1} L_i(t) J_{r-i}(t),$$

and

$$L_i(t) = \sum_{j=1}^{N} \left[\frac{F_j(t)}{\bar{F}_j(t)} \right]^i,$$

with $J_1(t) = 1$, and $\bar{F}_j(t) = 1 - F_j(t)$.

2.2.2 The proposed algorithm and some implementation details

Since no analytic solution is readily available, numerical integration is used. Formula (2.2) can be rewritten as follows

$$\pi_k = 1 - \int_0^1 F_{(n),k}^{N-1} \left[F_k^{-1}(t) \right] dt \tag{2.4}$$

$$= 1 - \int_0^1 F_{(n),k}^{N-1} \left[\frac{g(t)}{g(\lambda_k)} \right] dt, \tag{2.5}$$

where $g = H^{-1}$ depends on the design:

- in the uniform case
$$g(t) = t,$$
- in the exponential case
$$g(t) = -\ln(1-t),$$
- in the Pareto case
$$g(t) = \frac{t}{1-t}.$$

In order to compute numerically π_k using equation (2.5) the Simpson method is an available method. Algorithm 2.2.1 gives the general frame to compute π_k. The algorithm complexity is $O(N^2n^3p)$.

Algorithm 2.2.1 First-order inclusion probabilities

1: **for** $k = 1, \ldots, N$ **do**

2: Evaluate by numerical approximations

$$\pi_k = 1 - \int_0^1 F_{(n),k}^{N-1}\left[\frac{g(t)}{g(\lambda_k)}\right] dt$$

with the corresponding g, and using the next computations in each point $y_j,\ j = 1, \ldots, p$, of the applied numerical method:

3: Compute $t_j = \frac{g(y_j)}{g(\lambda_k)}$;

4: Compute $F_1(t_j), \ldots, F_{k-1}(t_j), F_{k+1}(t_j), \ldots, F_N(t_j)$;

5: Compute

$$F_{(1),k}^{N-1}(t_j) = 1 - \prod_{i=1, i\neq k}^{N} [1 - F_i(t_j)];$$

6:

$$J_{r,k}(t_j) = \frac{1}{r-1}\sum_{i=1}^{r-1}(-1)^{i+1}L_{i,k}(t_j)J_{r-i,k}(t_j),$$

where

$$L_{i,k}(t_j) = \sum_{j=1, j\neq k}^{N}\left[\frac{F_j(t_j)}{1 - F_j(t_j)}\right]^i, \text{ and } J_{1,k}(t_j) = 1;$$

7:

$$F_{(n),k}^{N-1}(t_j) = F_{(1),k}^{N-1}(t_j) - [1 - F_{(1),k}^{N-1}(t_j)]\sum_{r=2}^{n} J_{r,k}(t_j);$$

8: **end for**

Remark 1. The computation of the inclusion probabilities is useful in the case where the sample size n is quite small, since $\frac{\pi_k}{\lambda_k} \to 1$ when $n \to \infty$ under very general conditions (Rosén, 2000). Rosén (2000) emphasizes that the following holds: (a.) "For uniform, exponential and Pareto order πps, inclusion probabilities approach target values fast as the sample size increases, fastest for Pareto πps;" (b.) "For Pareto πps, target and factual inclusion probabilities differ only negligibly at least if $\min(n, N - n) \geq 5$." Our results

agree with the points (a) and (b) above. Example 2.2.1 gives an application of Algorithm 2.2.1 in all three cases.

Remark 2. We use the Simpson method to compute numerically the integral in (2.5). To avoid the problem concerning the domain of definition, the interval $[0, 1]$ was translated to $[1e-11, 0.999999999]$. In practice the interval $[0, 1]$ is split into p very small equal intervals. The question is: how large must be p in order to have a good precision? The answer depends on the function to integrate. In the examples below the number of intervals used is $p = 400$.

Example 2.2.1 *Let $N = 10, n = 3$. Table 2.1 gives the values of λ_k and π_k in the case of uniform, exponential and Pareto order πps sampling computed by using Algorithm 2.2.1. The time execution is also given. Algorithm 2.2.1 was implemented in an iterative way, in order to minimize the time execution. For another example (not shown in the article) with $N = 1000, n = 25$, for the uniform case the time execution is 179.07 seconds, for the exponential case 223.83 seconds, and for the Pareto case 13.98 seconds for $p = 400$.*

Table 2.1: Values of π_k

k	Uniform	Exponential	Pareto	λ_k
1	0.363541	0.367337	0.366244	0.366877
2	0.054914	0.057309	0.060776	0.061399
3	0.350309	0.354898	0.354357	0.355222
4	0.736618	0.714529	0.708367	0.703486
5	0.365803	0.369450	0.368263	0.368855
6	0.070146	0.073164	0.077215	0.078058
7	0.427157	0.425167	0.421471	0.420857
8	0.424307	0.422649	0.419064	0.418509
9	0.148869	0.154624	0.159761	0.161583
10	0.058335	0.060872	0.064483	0.065155
Total	3	3	3	3
Execution time in sec.	0.015	0.015	≈ 0	

2.2.3 How to obtain $\boldsymbol{\lambda}$ from $\boldsymbol{\pi}$

In practice the inclusion probabilities $\boldsymbol{\pi}$ are generally fixed, and it is desired to compute $\boldsymbol{\lambda}$ from $\boldsymbol{\pi}$. The knowledge of $\boldsymbol{\lambda}$ enables us to draw an ordered sample when the auxiliary information $\mathbf{z}$ is not available. Let $\boldsymbol{\phi}(\boldsymbol{\lambda}) = \boldsymbol{\pi}$. Thus $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_k, \ldots, \phi_N)$ where

$$\phi_k(\boldsymbol{\lambda}) = \pi_k = 1 - \int_0^1 F_{(n),k}^{N-1} \left[\frac{g(t)}{g(\lambda_k)} \right] dt.$$

Using the Newton method an iterative solution for $\boldsymbol{\lambda}$ is constructed

$$\boldsymbol{\lambda}^{(i)} = \boldsymbol{\lambda}^{(i-1)} - \mathbf{D}^{-1}\boldsymbol{\phi}(\boldsymbol{\lambda}^{(i-1)})[\boldsymbol{\phi}(\boldsymbol{\lambda}^{(i-1)}) - \boldsymbol{\pi}],$$

with $\boldsymbol{\lambda}^{(0)} = \boldsymbol{\pi}$. $\mathbf{D^{-1}}\boldsymbol{\phi}$ is the inverse of the Jacobian of $\boldsymbol{\phi}$. Since $\boldsymbol{\lambda}$ and $\boldsymbol{\pi}$ are very close to one another, it is possible to approximate $\boldsymbol{\phi}(\boldsymbol{\lambda}) \simeq \boldsymbol{\lambda}$, in order to approximate $\mathbf{D^{-1}}\boldsymbol{\phi}$ by the identity matrix. Thus we can consider the simplest equation

$$\boldsymbol{\lambda}^{(i)} = \boldsymbol{\lambda}^{(i-1)} - \boldsymbol{\phi}(\boldsymbol{\lambda}^{(i-1)}) + \boldsymbol{\pi}. \tag{2.6}$$

The iterative process above is applied until the convergence is attained. The complexity of this method is $O(ni \times N^2 n^3 p)$, where ni denotes the number of iterations in Newton method.

Example 2.2.2 *Table 2.2 gives the values of λ_k computed by using the equation (2.6) in all tree cases. The values of π_k are given in the last column of Table 2.2. The time execution is also specified. The value of the convergence tolerance in the Newton-Raphson method is fixed at 10^{-6}, and $p = 400$.*

2.3 Approximation of the joint inclusion probability in two positive co-ordinated ordered samples

Suppose that the finite population U is surveyed a number of times. For two occasions, which we call respectively the previous (1) and the current (2) occasions, let us denote the ranking variables as X_k and Y_k, $k = 1, \ldots, N$. The cdf are denoted as F_k for X_k and G_k for Y_k. On the previous occasion a sample s_1 is drawn from U with fixed size n_1. On the current occasion a sample s_2 is

Table 2.2: Values of λ_k

k	Uniform	Exponential	Pareto	π_k
1	0.369608	0.365766	0.366903	0.366244
2	0.067966	0.065173	0.061389	0.060776
3	0.359368	0.354747	0.355068	0.354357
4	0.661070	0.694923	0.701995	0.708367
5	0.371332	0.367634	0.373582	0.368263
6	0.085946	0.082469	0.078790	0.077215
7	0.414971	0.416649	0.424313	0.421471
8	0.413072	0.414437	0.421054	0.419064
9	0.173496	0.167236	0.162546	0.159761
10	0.072036	0.0690805	0.065185	0.064483
Total	2.988865	2.998120	3.010825	3
Execution time in sec.	0.031	0.078	0.093	

drawn from U with fixed size n_2. Both s_1 and s_2 are ordered samples of the same family (both are uniform, exponential or Pareto ordered samples). The target inclusion probabilities are denoted as $\boldsymbol{\lambda}^1$ and $\boldsymbol{\lambda}^2$, respectively. X_k and Y_k are dependent random variables since

$$X_k = \frac{g(\omega_k)}{g(\lambda_k^1)},\ Y_k = \frac{g(\omega_k)}{g(\lambda_k^2)},$$

with $\omega_k \sim U[0,1]$. The co-ordination of s_1 and s_2 is possible since the same PRN ω_k are used at each occasion, for all $k \in U$.

Let $\pi_k^{1,2} = Pr(k \in s_1, k \in s_2)$ be the joint inclusion probability of unit k in both samples. Let $\pi_k^1 = Pr(k \in s_1)$ and let $\pi_k^2 = Pr(k \in s_2)$. We are interested in positive co-ordination, where the goal is to maximize the number of common units of both samples. Let n_{12} be this number. The expectation of n_{12} is

$$E(n_{12}) = \sum_{k \in U} \pi_k^{1,2}.$$

We call the quantity $\sum_{k \in U} \min(\pi_k^1, \pi_k^2)$ the *absolute upper bound*. Due to the

Fréchet upper bound we have

$$\sum_{k\in U}\pi_k^{1,2} \leq \sum_{k\in U}\min(\pi_k^1,\pi_k^2). \tag{2.7}$$

The empirical results on order sampling designs show that $\sum_{k\in U}\pi_k^{1,2}$ approaches the absolute upper bound, but does not necessarily reach this quantity. Necessary and sufficient conditions to reach the absolute upper bound for two sampling designs are given in Matei and Tillé (2005c). In the case of equality in relation (2.7) we use the expression *the absolute upper bound is reached.*

2.3.1 Approximation of $\pi_k^{1,2}$

Our goal is to give an approximation of $\pi_k^{1,2}$ for two positive co-ordinated order sampling designs. Generally, for a random variable Z we denote by F_Z its cdf, and by f_Z its probability density function. We have

$$\begin{aligned}
\pi_k^{1,2} &= P(k\in s_1, k\in s_2)\\
&= P(X_k < X_{(n_1),k}^{N-1}, Y_k < Y_{(n_2),k}^{N-1}) \qquad (2.8)\\
&= P\left[g(\omega_k) < \min\left(g(\lambda_k^1)X_{(n_1),k}^{N-1}, g(\lambda_k^2)Y_{(n_2),k}^{N-1}\right)\right]\\
&= \int_0^\infty \left[1 - F_{\min\ g(\lambda_k^1)X_{(n_1),k}^{N-1}, g(\lambda_k^2)Y_{(n_2),k}^{N-1}}\ (t)\right] f_{g(\omega_k)}(t)dt. \qquad (2.9)
\end{aligned}$$

As in Section 2.2, $Y_{(n_2),k}^{N-1}$ denotes the n_2^{th} order statistic out of $N-1$ random variables $Y_1,\ldots,Y_{k-1},Y_{k+1},\ldots,Y_N$ (without Y_k). Expression (2.9) is rewritten as

$$\pi_k^{1,2} = 1 - \int_0^1 F_{\min\ g(\lambda_k^1)X_{(n_1),k}^{N-1}, g(\lambda_k^2)Y_{(n_2),k}^{N-1}}\ \left[F_{g(\omega_k)}^{-1}(t)\right] dt, \tag{2.10}$$

where $F_{g(\omega_k)}^{-1}$ denotes the inverse of $F_{g(\omega_k)}$. The random variable

$$\min(g(\lambda_k^1)X_{(n_1),k}^{N-1}, g(\lambda_k^2)Y_{(n_2),k}^{N-1})$$

is the first order statistic of the dependent variables $g(\lambda_k^1)X_{(n_1),k}^{N-1}$ and $g(\lambda_k^2)Y_{(n_2),k}^{N-1}$. For simplicity, let $F_{\min(g(\lambda_k^1)X_{(n_1),k}^{N-1}, g(\lambda_k^2)Y_{(n_2),k}^{N-1})}$ be denoted by $F_{(1)}^*$, let F_1^* be the cdf of $g(\lambda_k^1)X_{(n_1),k}^{N-1}$ and let F_2^* be the cdf of $g(\lambda_k^2)Y_{(n_2),k}^{N-1}$. The computation of

$F^*_{(1)}(t)$ is difficult due to the fact that $g(\lambda^1_k)X^{N-1}_{(n_1),k}$ and $g(\lambda^2_k)Y^{N-1}_{(n_2),k}$ are dependent and non-identically distributed random variables. Let us consider the events

$$A^t_k = (g(\lambda^1_k)X^{N-1}_{(n_1),k} \leq t),\ B^t_k = (g(\lambda^2_k)Y^{N-1}_{(n_2),k} \leq t).$$

When the joint probability of A^t_k and B^t_k is equal to its Fréchet upper bound

$$P(A^t_k, B^t_k) = \min(F^*_1(t), F^*_2(t)), \tag{2.11}$$

the absolute upper bound is reached, and the equality holds in expression (2.7). We have

$$P(A_t, B_t) = P(\max(g(\lambda^1_k)X^{N-1}_{(n_1),k}, g(\lambda^2_k)Y^{N-1}_{(n_2),k}) \leq t) = F^*_{(2)}(t),$$

which is the cdf of $\max(g(\lambda^1_k)X^{N-1}_{(n_1),k}, g(\lambda^2_k)Y^{N-1}_{(n_2),k})$. When the relation (2.11) fulfills, since

$$F^*_{(1)}(t) = F^*_1(t) + F^*_2(t) - F^*_{(2)}(t),$$

we have

$$F^*_{(1)}(t) = \max(F^*_1(t), F^*_2(t)). \tag{2.12}$$

Let us consider now our approximation. We come back to the relation (2.8). Let $C_k = \left(X_k < X^{N-1}_{(n_1),k}, Y_k < Y^{N-1}_{(n_2),k}\right)$. We approximate C_k by

$$\left(\max(X_k, Y_k) < \min(X^{N-1}_{(n_1),k} < Y^{N-1}_{(n_2),k})\right).$$

Thus an approximation of $\pi^{1,2}_k$ is

$$\pi^{1,2}_k \approx 1 - \int_0^1 F_{\min\ X^{N-1}_{(n_1),k}, Y^{N-1}_{(n_2),k}}\ \left(F^{-1}_{\max(X_k,Y_k)}(t)\right) dt. \tag{2.13}$$

Since $X^{N-1}_{(n_1),k}, Y^{N-1}_{(n_2),k}$ are dependent and non-identically distributed random variables, a method to compute $F_{\min\ X^{N-1}_{(n_1),k}, Y^{N-1}_{(n_2),k}}(t)$ is not trivial. In this context we take

$$F_{\min\ X^{N-1}_{(n_1),k}, Y^{N-1}_{(n_2),k}} = \max(F^*_{X,k}(t), F^*_{Y,k}(t)),$$

as in (2.12), in order to approach the absolute upper bound. $F^*_{X,k}$ and $F^*_{Y,k}$ denote the cdf of $X^{N-1}_{(n_1),k}$ and $Y^{N-1}_{(n_2),k}$, respectively. $F^*_{X,k}$ and $F^*_{Y,k}$ are computed

by using the same method as in Algorithm 2.2.1 (see Section 2.2.1). The relation (2.13) reduces to

$$\pi_k^{1,2} \approx 1 - \int_0^1 \max(F_{X,k}^*, F_{Y,k}^*) \left(F_{\max(X_k,Y_k)}^{-1}(t) \right) dt. \tag{2.14}$$

For the computation of $F_{\max(X_k,Y_k)}^{-1}$ (which is the inverse of the cdf of $\max(X_k, Y_k)$), the relationship between X_k and Y_k via ω_k is used. Thus, we have

- for the case of uniform πps sampling:

$$F_{\max(X_k,Y_k)}^{-1}(t) = \frac{t}{\min(\lambda_k^1, \lambda_k^2)};$$

- for the case of exponential πps sampling:

$$F_{\max(X_k,Y_k)}^{-1}(t) = \frac{\ln(1-t)}{\max(\ln(1-\lambda_k^1), \ln(1-\lambda_k^2))};$$

- for the case of Pareto πps sampling:

$$F_{\max(X_k,Y_k)}^{-1}(t) = \frac{t}{1-t} \left(\frac{1 - \min(\lambda_k^1, \lambda_k^2)}{\min(\lambda_k^1, \lambda_k^2)} \right).$$

Algorithm 2.3.2 gives the general frame of $\pi_k^{1,2}$ approximation. Its complexity is $O(\max(n_1^3, n_2^3)N^2p)$. The approximation (2.14) gives results very close to the values obtained by simulation for any type of population (see Section 2.3.2).
Remark 3: In the case where n_{12} is fixed, the approximation of $\pi_k^{1,2}$ can take into account a normalization in order to have $\sum_{k \in U} \pi_k^{1,2} = n_{12}$ (see steps 6,7 in Algorithm 2.3.2). In this case we have

$$\pi_k^{1,2} \approx P(k \in s_1, k \in s_2 | n_{12}).$$

2.3.2 Examples and simulations

In order to test the performance of Algorithm 2.3.2, a set of simulations was used. Two data sets have been used in Monte-Carlo simulations: one artificial population, and mu284 population from Särndal et al. (1992). A set of 100′000

Algorithm 2.3.2 Approximation of the joint inclusion probabilities

1: **for** $k = 1, \ldots, N$ **do**
2: Evaluate by numerical approximations the integral in (2.14) by using the next computations in each point $y_j, j = 1, \ldots, p$ of the applied numerical method:
3: Compute
$$t_j = F^{-1}_{\max(X_k, Y_k)}(y_j);$$
4: Compute with the same method as in Algorithm 2.2.1
$$F^*_{X,k}(t_j) = F^{N-1}_{(n_1),k}(t_j);$$
$$F^*_{Y,k}(t_j) = G^{N-1}_{(n_2),k}(t_j);$$
and compute
$$\max(F^*_X(t_j), F^*_Y(t_j));$$
5: **end for**
6: **if** n_{12} is fixed **then**
7: Normalize the joint probabilities $\pi^{1,2}_k$ in order to have $\sum_{k \in U} \pi^{1,2}_k = n_{12}$.
8: **end if**

independent ordered πps samples of sizes n_1 and n_2, respectively have been selected from each population. The joint inclusion probabilities were computed by taking into account the sampled units in both samples. The values ω_k have been randomly generated using $U[0, 1]$ distribution. The populations are regarded in terms of correlation coefficient between $\mathbf{z^1}, \mathbf{z^2}$ (or $\boldsymbol{\lambda}^1, \boldsymbol{\lambda}^2$).

The artificial population was obtained by randomly generating $\boldsymbol{\lambda}^1, \boldsymbol{\lambda}^2 \sim U[0, 1]$, and normalizing the target inclusion probabilities in order to have their sums equal to $n_1 = 3$ and $n_2 = 2$, respectively. The size population is $N = 20$. Table 2.3 gives the values of $\boldsymbol{\lambda}^1$ and $\boldsymbol{\lambda}^2$. The correlation coefficient between $\boldsymbol{\lambda}^1$ and $\boldsymbol{\lambda}^2$ is 0.08. The values of $\pi^{1,2}_k$ obtained by using Algorithm 2.3.2 are tabulated in Table 2.4 (for the uniform case), Table 2.5 (for the exponential case) and Table 2.6 (for the Pareto case). In each table, the second column represents the approximated values of $\pi^{1,2}_k$. The third column gives the values of $\pi^{1,2}_k$ obtained by simulations. The column 4 is the $\min(\pi^1_k, \pi^2_k)$ computed using Algorithm 2.2.1. Each table reports the expected overlap and the absolute upper bound $\sum_{k \in U} \min(\pi^1_k, \pi^2_k)$. In all three tables the approximated $\pi^{1,2}_k$ are smaller than $\min(\pi^1_k, \pi^2_k),\ k = 1, \ldots, N$. The approximated values are

Table 2.3: Values of λ_k^1, λ_k^2

k	λ_k^1	λ_k^2	k	λ_k^1	λ_k^2
1	0.172192	0.025246	11	0.122143	0.014400
2	0.199673	0.130083	12	0.226835	0.100582
3	0.167972	0.222296	13	0.241416	0.106913
4	0.083301	0.062166	14	0.191487	0.096061
5	0.052490	0.231067	15	0.114779	0.044966
6	0.018503	0.161529	16	0.248597	0.104245
7	0.043962	0.086747	17	0.182376	0.230550
8	0.225073	0.052452	18	0.099840	0.027754
9	0.255619	0.079211	19	0.019807	0.033286
10	0.084072	0.044691	20	0.249863	0.145756

close to the simulated ones in all three cases.

Tables 2.7 and 2.8 give the approximated and the simulated values for $\pi_k^{1,2}$, when n_{12} is fixed to 1 and 2, respectively, for the values of $\boldsymbol{\lambda}^1, \boldsymbol{\lambda}^2$ given in Table 2.3 (see Remark 3). Our approximations lie close to the simulated values of $\pi_k^{1,2}$.

For the first set of simulations, two data items (P75, 1975 population and P85, 1985 population) have been taken as $\mathbf{z^1}, \mathbf{z^2}$, in mu284 population. This is done in order to compute $\lambda_k^1 = n_1 z_k^1 / \sum_{\ell \in U} z_\ell^1$ and $\lambda_k^2 = n_2 z_k^2 / \sum_{\ell \in U} z_\ell^2$. These two items are highly correlated (the correlation coefficient is 0.9948275). In fact, the items P75 and P85 represent the same variable measured at two different time occasions. This case is refereed below as "mu284 I". For the second set of simulations, we have retained the variables P75 for $\mathbf{z^1}$, but we have taken the item CS82 (number of Conservative seats in municipal council) instead of P85 to compute $\mathbf{z^2}$. Now the correlation coefficient is 0.6230069. This case is refereed below as "mu284 II".

For mu284 population the units are stratified according to region (geographic region indicator; there are 8 regions). Some units were removed, in order to ensure $\lambda_k < 1$: one unit in strata 1, 4 and 5, and three units in stratum 7. In each stratum, two ordered samples were drawn, with sizes $n_1 = 4$ and $n_2 = 6$. The strata are the same in the first and second design. In order to save space in our tabulation only the expected overlaps are reported in the case

Table 2.4: Results for the uniform case

k	$\pi_k^{1,2}$ approx.	$\pi_k^{1,2}$ simulated	$\min(\pi_k^1, \pi_k^2)$
1	0.023150	0.023410	0.023753
2	0.123215	0.123150	0.128719
3	0.160935	0.147440	0.166424
4	0.056760	0.053990	0.059486
5	0.048963	0.049250	0.050085
6	0.016982	0.017800	0.017474
7	0.040065	0.038730	0.041838
8	0.048847	0.050500	0.049963
9	0.074724	0.076460	0.076413
10	0.040649	0.042270	0.042419
11	0.013074	0.013650	0.013483
12	0.095023	0.096930	0.098047
13	0.101446	0.102880	0.104548
14	0.090079	0.094000	0.09343
15	0.041100	0.042480	0.042685
16	0.098964	0.100400	0.101803
17	0.175591	0.161040	0.181604
18	0.025203	0.027000	0.026141
19	0.017787	0.017760	0.018712
20	0.140006	0.141030	0.145424
Total	1.432562	1.420170	1.482451

of mu284 population. Tables 2.9 and 2.10 summarize the results by indicating the expected overlap by stratum for the approximated (denoted as Tapp) and simulated cases (denoted as Tsim). The number of units in each stratum is also given. To compare the expected overlap in the approximated and simulated case we use a measure named Absolute Relative Difference (ARD), equal to

$$ARD = \frac{|Tapp - Tsim|}{Tapp}.$$

In the artificial population, our approximation performs better in the uniform case with ARD=0.006. For the exponential case we have ARD=0.01, and for

Table 2.5: Results for the exponential case

k	$\pi_k^{1,2}$ approx.	$\pi_k^{1,2}$ simulated	$\min(\pi_k^1, \pi_k^2)$
1	0.023948	0.024750	0.024464
2	0.125185	0.121580	0.129666
3	0.162475	0.149830	0.16739
4	0.058253	0.055610	0.060853
5	0.050267	0.050880	0.051213
6	0.017500	0.018270	0.017933
7	0.041182	0.040000	0.042822
8	0.050306	0.051240	0.051208
9	0.076553	0.078720	0.077898
10	0.041837	0.042200	0.043539
11	0.013536	0.013460	0.013912
12	0.097031	0.098870	0.099484
13	0.103441	0.106290	0.105924
14	0.092083	0.093490	0.094897
15	0.042338	0.043220	0.04381
16	0.100943	0.100350	0.103207
17	0.176797	0.160950	0.182184
18	0.026022	0.026340	0.026912
19	0.018333	0.016760	0.019202
20	0.141689	0.143070	0.145878
Total	1.459720	1.435880	1.502396

the Pareto case ARD=0.02. In mu284 I, our approximation performs better in the exponential and the Pareto cases. For the uniform case ARD takes values between 0.02 (in stratum 4) and 0.04 (in stratum 8). For the exponential case ARD takes values between 0.002 (in stratum 7) and 0.02 (in stratum 5). For the Pareto case we have ARD between 0.002 (in stratum 7) and 0.02 (in stratum 5). Similar results are given by mu284 II example. The ARD values for the uniform case lie between 0.002 (in stratum 7) and 0.03 (in stratum 8). For the exponential case we have ARD between 0.002 (in stratum 7) and 0.01 (in stratum 8), and for the Pareto case, ARD takes values between 0.002 (in stratum 3) and 0.01 (in stratum 1).

Table 2.6: Results for the Pareto case

k	$\pi_k^{1,2}$ approx.	$\pi_k^{1,2}$ simulated	$\min(\pi_k^1, \pi_k^2)$
1	0.024842	0.025038	0.02519
2	0.126658	0.123098	0.129907
3	0.163493	0.150420	0.167841
4	0.059766	0.055545	0.062003
5	0.051627	0.052038	0.052406
6	0.018108	0.018363	0.018483
7	0.042390	0.040971	0.043896
8	0.051750	0.052394	0.052315
9	0.078173	0.078939	0.079012
10	0.043116	0.043352	0.044577
11	0.014084	0.014304	0.014372
12	0.098692	0.099403	0.100364
13	0.105036	0.106258	0.106696
14	0.093807	0.094380	0.095843
15	0.043659	0.044494	0.044852
16	0.102536	0.103667	0.104026
17	0.177518	0.161355	0.182287
18	0.026956	0.027863	0.02769
19	0.018973	0.017917	0.019785
20	0.142803	0.141712	0.145644
Total	1.483987	1.451511	1.517189

The following conclusions are made from the results of the empirical study:

a) the approximated and the simulated values of $\pi_k^{1,2}$ are close to one another, no matter the correlation coefficient between $\boldsymbol{\lambda}^1$ and $\boldsymbol{\lambda}^2$;

b) we draw the same conclusion for the case where n_{12} is fixed;

c) the values of ARD lie between 0.002 and 0.04, and show no special behavior from one sampling design to another (uniform, exponential or Pareto).

Table 2.7: Results for the case $n_{12} = 1$

	Uniform		Exponential		Pareto	
k	$\pi_k^{1,2}$ approx.	$\pi_k^{1,2}$ simulated	$\pi_k^{1,2}$ approx.	$\pi_k^{1,2}$ simulated	$\pi_k^{1,2}$ approx.	$\pi_k^{1,2}$ simulated
1	0.016160	0.016333	0.016406	0.016674	0.016740	0.016493
2	0.086010	0.089964	0.085760	0.088803	0.085350	0.088048
3	0.112341	0.104706	0.111306	0.105005	0.110171	0.103987
4	0.039621	0.039287	0.039907	0.038837	0.040274	0.039385
5	0.034179	0.026981	0.034436	0.027020	0.034789	0.027430
6	0.011854	0.009841	0.011989	0.009795	0.012202	0.010421
7	0.027967	0.027074	0.028212	0.027261	0.028565	0.027181
8	0.034098	0.033956	0.034463	0.034446	0.034872	0.034816
9	0.052161	0.052729	0.052444	0.052998	0.052678	0.053360
10	0.028375	0.029062	0.028661	0.030024	0.029054	0.030400
11	0.009126	0.008978	0.009273	0.009506	0.009491	0.009813
12	0.066331	0.069651	0.066472	0.069127	0.066505	0.069469
13	0.070814	0.073941	0.070864	0.073607	0.070780	0.074381
14	0.062880	0.066244	0.063083	0.066700	0.063213	0.066261
15	0.028690	0.029652	0.029004	0.030282	0.029420	0.030875
16	0.069082	0.071658	0.069152	0.072138	0.069095	0.071492
17	0.122571	0.115312	0.121117	0.113793	0.119622	0.113208
18	0.017593	0.018032	0.017827	0.018488	0.018165	0.018893
19	0.012416	0.011897	0.012559	0.012250	0.012785	0.012391
20	0.097731	0.104700	0.097066	0.103247	0.096229	0.101695
Total	1	1	1	1	1	1

2.4 Conclusions

Improved numerical algorithms render possible various types of computation in order πps sampling design with fixed order distribution shape. It is possible to compute the first-order inclusion probabilities in the case of uniform, exponential and Pareto order πps sampling designs in a reasonable time execution. An approximation of the joint inclusion probability of a unit in two co-ordinated ordered samples is also given. The results show that this approximation give values which are close to the simulated ones. These algorithms should facilitate the study and the use of order sampling designs.

Table 2.8: Results for the case $n_{12} = 2$

	Uniform		Exponential		Pareto	
k	$\pi_k^{1,2}$ approx.	$\pi_k^{1,2}$ simulated	$\pi_k^{1,2}$ approx.	$\pi_k^{1,2}$ simulated	$\pi_k^{1,2}$ approx.	$\pi_k^{1,2}$ simulated
1	0.032320	0.034170	0.032812	0.035028	0.033480	0.035026
2	0.172020	0.167568	0.171519	0.167234	0.170700	0.166566
3	0.224681	0.208648	0.222611	0.207268	0.220343	0.205986
4	0.079243	0.075637	0.079814	0.074805	0.080548	0.074713
5	0.068357	0.077243	0.068872	0.077862	0.069579	0.078861
6	0.023709	0.027276	0.023977	0.027443	0.024405	0.028270
7	0.055935	0.057893	0.056425	0.056979	0.057130	0.057051
8	0.068195	0.072502	0.068926	0.071919	0.069745	0.073492
9	0.104322	0.108761	0.104887	0.109982	0.105355	0.108961
10	0.056750	0.059034	0.057322	0.059345	0.058108	0.059600
11	0.018253	0.019416	0.018546	0.020110	0.018981	0.020171
12	0.132662	0.136581	0.132945	0.136930	0.133009	0.136557
13	0.141629	0.146029	0.141727	0.146156	0.141559	0.145410
14	0.125759	0.128096	0.126165	0.128268	0.126426	0.130203
15	0.057380	0.060284	0.058008	0.061023	0.058840	0.061380
16	0.138164	0.142685	0.138305	0.142559	0.138190	0.142919
17	0.245142	0.223198	0.242234	0.221961	0.239245	0.220746
18	0.035186	0.036851	0.035653	0.037810	0.036329	0.038334
19	0.024832	0.024941	0.025119	0.024890	0.025570	0.024524
20	0.195462	0.193187	0.194132	0.192429	0.192459	0.191231
Total	2	2	2	2	2	2

Table 2.9: Results for mu284 I

		uniform		exponential		Pareto	
Stratum	N	Tapp	Tsim	Tapp	Tsim	Tapp	Tsim
1	24	3.84	3.98	3.90	3.98	3.95	3.98
2	48	3.87	3.99	3.91	3.99	3.94	3.98
3	32	3.88	3.99	3.92	3.99	3.95	3.99
4	37	3.88	3.99	3.90	3.99	3.93	3.99
5	55	3.86	3.99	3.88	3.99	3.90	3.99
6	41	3.87	3.99	3.89	3.99	3.93	3.98
7	12	3.89	3.99	3.98	3.99	3.99	3.98
8	29	3.82	3.99	3.92	3.97	3.99	3.95

Table 2.10: Results for mu284 II

		uniform		exponential		Pareto	
Stratum	N	Tapp	Tsim	Tapp	Tsim	Tapp	Tsim
1	24	3.65	3.64	3.70	3.66	3.73	3.67
2	48	3.58	3.50	3.57	3.53	3.56	3.55
3	32	3.64	3.59	3.64	3.62	3.65	3.64
4	37	3.17	3.14	3.18	3.14	3.17	3.15
5	55	3.53	3.49	3.54	3.51	3.54	3.52
6	41	3.59	3.54	3.60	3.56	3.62	3.57
7	12	3.58	3.57	3.57	3.58	3.60	3.57
8	29	3.52	3.41	3.50	3.45	3.50	3.47

Chapter 3

Maximal and minimal sample co-ordination

Abstract
In sample design over time we are interested in maximizing/minimizing the expected overlap between two or more samples drawn on different time points. For this it is necessary to compute the joint inclusion probability of two samples drawn on different time periods. A solution of this computation is given by using linear programming and more precisely by solving a transportation problem. This solution is not computationally fast. We are interested to identify the conditions under which the objective function associated with an optimal solution of the transportation problem is equal to the bound given by maximizing/minimizing the expected overlap. Using these conditions we propose a new algorithm to optimize the co-ordination between two samples without using linear programming. Our algorithm is based on Iterative Proportional Fitting (IPF) procedure. Theoretical complexity is substantially lower than for transportation problem approach, because more than five iterations of IPF procedure are not required in practice.
Keywords and phrases: sample survey, sample co-ordination, IPF procedure, transportation problem.

3.1 Introduction

It is usual to sample populations on two or more occasions in order to obtain current estimates of a character. Sample co-ordination problem consists in managing the overlap of two or more samples drawn in different time occasions. It is either positive or negative. While in the former the expected overlap of two or more samples is maximized, in the latter it is minimized. Positive and negative co-ordination can be formulated as a dual problem. Thus, solving positive co-ordination problem can lead us to the solution of negative sample co-ordination and vice versa.

Various methods have been proposed in order to solve sample co-ordination problem. The co-ordination problem has been the main topic of interest for more than fifty years. The first papers on this subject are due to Patterson (1950) and Keyfitz (1951). Other papers dated from the same period are: Kish and Hess (1959), Fellegi (1963), Kish (1963), Fellegi (1966), Gray and Platek (1963). These first works present methods which are in general restricted to two successive samples or to small sample sizes. A generalization of the problem in the context of a larger sample size has been done by Kish and Scott (1971). Mathematical programming met the domain of the sample co-ordination with the books of Raj (1968) and Arthnari and Dodge (1981) and the paper of Causey et al. (1985). Brewer (1972) and Brewer et al. (1972) introduced the concept of co-ordination based on *Permanent Random Numbers* (PRNs). Furthermore, Rosén (1997a,b) developed *order sampling*, which is another approach that takes into account the concept of PRN.

Let $U = \{1, \ldots, k, \ldots, N\}$ be the population under study. Samples without replacement are selected on two distinct time periods. The time periods are indicated by the exponents 1 and 2 in our notation. Thus, π_k^1 denotes the inclusion probability of unit $k \in U$ for time period 1 in the first sample. Similarly, π_k^2 denotes the inclusion probability of unit $k \in U$ for time period 2 in the second sample. Let $\mathcal{S}_1, \mathcal{S}_2$ be the sets of all samples in the first occasion and the second occasion, respectively. Our notation for a sample is $s_i^1 \in \mathcal{S}_1$ and $s_j^2 \in \mathcal{S}_2$. Let also $\pi_k^{1,2}$ be the joint inclusion probability of unit k in both samples. Thus

$$\max(\pi_k^1 + \pi_k^2 - 1, 0) \leq \pi_k^{1,2} \leq \min(\pi_k^1, \pi_k^2).$$

Let p_i^1, p_j^2 denote the probability distributions on $\mathcal{S}_1, \mathcal{S}_2$, respectively. Let $|s_i^1 \cap s_j^2|$ be the number of common units of both samples, let $I = \{k \in U | \pi_k^1 \leq \pi_k^2\}$ be the set of "increasing" units, and let $D = \{k \in U | \pi_k^1 > \pi_k^2\}$ be the set of "decreasing" units.

Definition 3.1.1 *The quantity* $\sum_{k\in U}\min(\pi_k^1,\pi_k^2)$ *is called the absolute upper bound; the quantity* $\sum_{k\in U}\max(\pi_k^1+\pi_k^2-1,0)$ *is called the absolute lower bound.*

Note that $\sum_{k\in U}\pi_k^{1,2}$ is the expected overlap. The expected overlap is equal to the absolute upper bound when $\pi_k^{1,2}=\min(\pi_k^1,\pi_k^2)$, for all $k\in U$. We use in this case the terminology "the absolute upper bound is reached". Similarly, the absolute lower bound is reached when $\pi_k^{1,2}=\max(\pi_k^1+\pi_k^2-1,0)$, for all $k\in U$. Only a few of the already developed methods can reach the absolute upper/lower bound.

As we have already mentioned, one point of view to solve sample co-ordination problem is to use mathematical programming and more exactly to solve a transportation problem. The form of the sample co-ordination problem in the frame of a transportation problem enables us to compute the joint inclusion probability of two samples drawn on two different occasions, s_i^1 and s_j^2, and then the conditional probability $p(s_j^2|s_i^1)$. This allows to choose the sample s_j^2 drawn in the second occasion given that the sample s_i^1 was drawn in the first. The solution given by using mathematical programming is not computationally fast.

We call a *bi-design* a couple of two sampling designs given on two different occasions. Let $\mathcal{S}=\{s=(s_i^1,s_j^2)|s_i^1\in\mathcal{S}_1,s_j^2\in\mathcal{S}_2\}$. Let $p(s)$ be a probability distribution on $\mathcal{S}$. In our notation $p(s)$ is p_{ij}. We are interested in finding the conditions when the absolute upper/lower bound is reached. We pose this problem because the value of the objective function in the case of an optimal solution given by the linear programming (denoted as *relative upper bound*) is not necessarily equal to the absolute upper/lower bound. In the equality case, for positive co-ordination, we use the terminology "maximal sample co-ordination" instead of "optimal sample co-ordination" to avoid the confusion with the optimal solution given by the linear programming. Similarly, for the case of negative co-ordination, we talk about the "minimal sample co-ordination" when the absolute lower bound is reached.

In this article, we extend the method presented in Matei and Tillé (2004). Two procedures to decide whether the absolute upper bound, respectively the absolute lower bound can be reached or not are developed. In the affirmative case, we propose an algorithm to compute the probability distribution $p(.)$ of a bi-design, without using mathematical programming. The proposed algorithm is based on Iterative Proportional Fitting (IPF) procedure (Deming and Stephan, 1940) and it has lower complexity compared to linear programming. The proposed methods can be applied for any type of sampling design when it is possible to compute the probability distributions for both samples.

The article is organized as follows: Section 3.2 presents the transportation problem in the case of sample co-ordination; Section 3.3 presents some cases where the probability distribution of a bi-design can be computed directly, and gives some conditions to reach the maximal co-ordination; Section 3.4 presents the proposed algorithm and gives two examples of its application for the positive co-ordination. In Section 3.5 the method is applied in the case of negative co-ordination. Finally, in Section 3.6 the conclusions are given.

3.2 Transportation problem in sample co-ordination

3.2.1 Transportation problem

The transportation problem is an application of linear programming. In principle, it consists in finding a flow of least cost that ships from supply sources to consumer destinations. The model is a bipartite graph $G = (A \cup B, E)$, where A is the set of vertex in source, B is the set of vertex in destination, and E is the set of edges from A to B. Each edge $(i, j) \in E$ has an associated cost c_{ij}. The problem is a linear program defined by

$$\min \sum_{i \in A, j \in B} c_{ij} x_{ij}, \tag{3.1}$$

subject to the constraints

$$\left| \begin{array}{l} \sum_{j \in B} x_{ij} = a_i, \text{ for all } i \in A, \\ \sum_{i \in A} x_{ij} = b_j, \text{ for all } j \in B, \\ x_{ij} \geq 0, i \in A, j \in B, \end{array} \right.$$

where a_i is the supply at i-th source, and b_j is the demand at j-th destination. Table 3.1 gives a representation of this problem, with $m = |A|, q = |B|$. In order to obtain the consistency, we must have:

$$\sum_{i \in A} \sum_{j \in B} x_{ij} = \sum_{j \in B} \sum_{i \in A} x_{ij} = \sum_{i \in A} a_i = \sum_{j \in B} b_j.$$

A transportation schedule (x_{ij}) that satisfies the constraints above is said to be feasible with respect to the supply vector $\mathbf{a}$ and the demand vector $\mathbf{b}$.

3.2.2 Some forms of transportation problem

Linear programming was used to solve the co-ordination problem of two samples as a transportation problem. The application of the transportation problem in the sample co-ordination is given by Raj (1968), Arthnari and Dodge (1981), Causey et al. (1985), Ernst and Ikeda (1995), Ernst (1996), Ernst (1998), Ernst and Paben (2002), Reiss et al. (2003).

For a positive co-ordination, we use the following form of transportation problem presented in Causey et al. (1985)

$$\max \sum_{i=1}^{m} \sum_{j=1}^{q} c_{ij} p_{ij}, \tag{3.2}$$

subject to the constraints

$$\left| \begin{array}{l} \sum_{j=1}^{q} p_{ij} = p_i^1, i = 1, \ldots, m, \\ \sum_{i=1}^{m} p_{ij} = p_j^2, j = 1, \ldots, q, \\ p_{ij} \geq 0, i = 1, \ldots, m, j = 1, \ldots, q, \end{array} \right.$$

where

$$c_{ij} = |s_i^1 \cap s_j^2|, p_i^1 = \Pr(s_i^1), p_j^2 = \Pr(s_j^2), p_{ij} = \Pr(s_i^1, s_j^2),$$

$s_i^1 \in \mathcal{S}_1$ and $s_j^2 \in \mathcal{S}_2$ denote all the possible samples on the first and second occasion, respectively, with $m = |\mathcal{S}_1|$ and $q = |\mathcal{S}_2|$. We suppose that $p_i^1 > 0, p_j^2 > 0$ in order to compute the conditional probabilities. A modification of this problem has been done by Ernst (1986). In the case of two selected units per stratum, Ernst and Ikeda (1995) have simplified the computational aspect of the problem (3.2).

Table 3.1: Transportation problem

	1	2	...	q	Σ
1	x_{11}	x_{12}	...	x_{1q}	a_1
2	x_{21}	x_{22}	...	x_{2q}	a_2
...	...	...	...	...	...
m	x_{m1}	x_{m2}	...	x_{mq}	a_m
Σ	b_1	b_2	...	b_q	$\sum_{i=1}^{m} a_i = \sum_{j=1}^{q} b_j$

When only one unit is selected in each design, we obtain a particular case of problem (3.2) (with $c_{kk} = 1$ and $c_{k\ell} = 0,$ for all $k \neq \ell$), that was presented by Raj (1968) as follows

$$\max \sum_{k=1}^{N} \pi_k^{1,2}, \tag{3.3}$$

subject to the constraints

$$\left| \begin{array}{l} \sum_{\ell=1}^{N} \pi_{k\ell}^{1,2} = \pi_k^1, \\ \sum_{k=1}^{N} \pi_{k\ell}^{1,2} = \pi_\ell^2, \\ \pi_{k\ell}^{1,2} \geq 0, k, \ell = 1, \ldots, N, \end{array} \right.$$

where $\pi_{k\ell}^{1,2}$ is the probability to select the units k and ℓ in both samples. Arthnari and Dodge (1981) showed that any feasible solution of problem (3.3), with $\pi_k^{1,2} = \min(\pi_k^1, \pi_k^2)$ for all $k \in U$, is an optimal solution. Keyfitz (1951) gives an optimal solution to the problem (3.3), without application of the linear programming (see 3.1.1). For a negative co-ordination, in problems (3.2) and (3.3) we use min instead of max in expression of the objective function and we keep the same constraints.

3.3 Maximal sample co-ordination

In what follows, we focus the attention on the problem (3.2). Our goal is to define a method that gives an optimal solution for the problem (3.2), without using mathematical programming. We consider the problem (3.2) as a two-dimensional distribution where only the two marginal distributions (the sums along the rows and columns) are given. Information about the inner distribution is available by using the propositions below. It is required to compute the internal values. The technique is based on IPF procedure (Deming and Stephan, 1940).

A measure of positive co-ordination is the number of common sampled units on these two occasions. Let n_{12} be this number. The goal is to maximize the expectation of n_{12}. We have

$$E(n_{12}) = \sum_{k \in U} \pi_k^{1,2} = \sum_{k \in U} \sum_{s_i^1 \ni k} \sum_{s_j^2 \ni k} p_{ij} = \sum_{s_i^1 \in \mathcal{S}_1} \sum_{s_j^2 \in \mathcal{S}_2} |s_i^1 \cap s_j^2| p_{ij},$$

which is the objective function of problem (3.2). To maximize $E(n_{12})$ it amounts to maximize this objective function.

Similarly, the objective function of problem (3.3) is

$$\sum_{k=1}^{N} |\{k\} \cap \{k\}| \Pr(\{k\}, \{k\}) = \sum_{k=1}^{N} \pi_k^{1,2}.$$

3.3.1 Some cases of maximal sample co-ordination

There are three cases when the absolute upper bound equal to $\sum_{k \in U} \min(\pi_k^1, \pi_k^2)$ can be reached, without solving the associated transportation problem. These cases are presented below.

One unit drawn by stratum

Keyfitz (1951) gives an optimal solution to the problem (3.3). This method selects one unit per stratum, when the two designs have the same stratification. The conditional probability to select the unit ℓ in the second sample given that the unit k was selected in the first sample is $\pi_{k\ell}^{1,2}/\pi_k^1$, for all $k, \ell \in U$. Algorithm 3.3.3 computes the values of $\pi_{k\ell}^{1,2}$.

Algorithm 3.3.3 Keyfitz algorithm

1: **for all** $k \in U$ **do**
2: $\quad \pi_k^{1,2} = \min(\pi_k^1, \pi_k^2)$,
3: **end for**
4: **if** $k \in D, \ell \in I, k \neq \ell$ **then**
5: $\quad \pi_{k\ell}^{1,2} = (\pi_k^1 - \pi_k^2)(\pi_\ell^2 - \pi_\ell^1) / \sum_{\ell_1 \in I} (\pi_{\ell_1}^2 - \pi_{\ell_1}^1)$,
6: **else**
7: $\quad \pi_{k\ell}^{1,2} = 0$.
8: **end if**

Example 3.3.3 *Let* $U = \{1, 2, 3, 4\}$, $\pi_1^1 = 0.15$, $\pi_2^1 = 0.25$, $\pi_3^1 = 0.20$, $\pi_4^1 = 0.40$, $\pi_1^2 = 0.10, \pi_2^2 = 0.30, \pi_3^2 = 0.20, \pi_4^2 = 0.40$. *The absolute upper bound equal to* $\sum_{k \in U} \min(\pi_k^1, \pi_k^2) = 0.95$ *is reached. Table 3.2 gives the the values of* $\pi_{k\ell}^{1,2}$ *computed by means of Algorithm 3.3.3.*

Simple random sample without replacement (srswor)

The multidimensional srswor was defined by Cotton and Hesse (1992b). They showed that if the joint sample is srswor, the marginal samples are also srswor.

Table 3.2: Keyfitz method

	$\{1\}$	$\{2\}$	$\{3\}$	$\{4\}$	Σ
$\{1\}$	0.10	0.05	0	0	0.15
$\{2\}$	0	0.25	0	0	0.25
$\{3\}$	0	0	0.20	0	0.20
$\{4\}$	0	0	0	0.40	0.40
Σ	0.10	0.30	0.20	0.40	1

Under the srswor bi-design every sample $s = (s_i^1, s_j^2) \in \mathcal{S}$ of the fixed size $n_s = (n_1^*, n_{12}, n_2^*)$ receives the same probability of being selected, where $n_1^* = |s_i^1 \backslash s_j^2|, n_{12} = |s_i^1 \cap s_j^2|, n_2^* = |s_j^2 \backslash s_i^1|$. That is (see Goga, 2003, p.112)

$$p(s) = \begin{cases} \frac{n_1^*! n_{12}! n_2^*! (N-(n_1^*+n_{12}+n_2^*))!}{N!} & \text{if } s \text{ is of size } n_s, \\ 0 & \text{otherwise.} \end{cases}$$

Let $|s_i^1| = n_1, |s_j^2| = n_2$. In the case of maximal sample co-ordination, this definition reduces to

a. if $k \in I$ which is equivalent with $n_1 \leq n_2$:

$$p(s) = \begin{cases} \frac{n_1!(n_2-n_1)!(N-n_2)!}{N!} & \text{if } n_{12} = n_1, \\ 0 & \text{otherwise.} \end{cases}$$

b. if $k \in D$ which is equivalent with $n_2 < n_1$:

$$p(s) = \begin{cases} \frac{n_2!(n_1-n_2)!(N-n_1)!}{N!} & \text{if } n_{12} = n_2, \\ 0 & \text{otherwise.} \end{cases}$$

Example 3.3.4 *Let $N = 4$. Consider two srswor sampling designs with $n_1 = 2, n_2 = 3, \pi_k^1 = 1/2, \pi_k^2 = 3/4$, for all $k \in U$. The probability distributions are $p_1(s_i^1) = 1/6$ in the first time period and $p_2(s_j^2) = 1/4$ in the second time period. We have $\sum_{k \in U} \min(\pi_k^1, \pi_k^2) = 2$. Table 3.3 gives the values of $p(s)$ in the case of maximal sample co-ordination. Matrix $\mathbf{C} = (c_{ij})_{m \times q}$ is given in Table 3.4. This solution has the property that $\sum_{i=1}^{m} \sum_{j=1}^{q} c_{ij} p_{ij}$ is equal to the absolute upper bound.*

Table 3.3: Srswor bi-design

	{1,2,3}	{1,2,4}	{1,3,4}	{2,3,4}	Σ
{1,2}	1/12	1/12	0	0	1/6
{1,3}	1/12	0	1/12	0	1/6
{1,4}	0	1/12	1/12	0	1/6
{2,3}	1/12	0	0	1/12	1/6
{2,4}	0	1/12	0	1/12	1/6
{3,4}	0	0	1/12	1/12	1/6
Σ	1/4	1/4	1/4	1/4	1

Table 3.4: Values of c_{ij} in the case of srswor

	{1,2,3}	{1,2,4}	{1,3,4}	{2,3,4}
{1,2}	2	2	1	1
{1,3}	2	1	2	1
{1,4}	1	2	2	1
{2,3}	2	1	1	2
{2,4}	1	2	1	2
{3,4}	1	1	2	2

Poisson sampling

A generalization of Poisson sampling in the multidimensional case was given by Cotton and Hesse (1992b). They showed that if the joint sample is Poisson sample, the marginal samples are also Poisson samples. In the bi-dimensional case $s = (s_i^1, s_j^2) \in \mathcal{S}$, we have (see Goga, 2003, p.114)

$$\begin{aligned} p(s) &= p(s_i^1, s_j^2) \\ &= \prod_{k \in s_i^1 \backslash s_j^2} \pi_k^{1*} \prod_{k \in s_j^2 \backslash s_i^1} \pi_k^{2*} \prod_{k \in s_i^1 \cap s_j^2} \pi_k^{1,2} \prod_{k \in U \backslash (s_i^1 \cup s_j^2)} (1 - \pi_k^{1*} - \pi_k^{2*} - \pi_k^{1,2}), \end{aligned}$$

where $\pi_k^{1*} = \pi_k^1 - \pi_k^{1,2}, \pi_k^{2*} = \pi_k^2 - \pi_k^{1,2}$ are the inclusion probabilities for $k \in s_i^1 \backslash s_j^2, s_j^2 \backslash s_i^1$, respectively. In the case of maximal sample co-ordination,

this definition reduces to

$$\begin{aligned} p(s_i^1, s_j^2) \;=\; & \prod_{k \in s_i^1 \setminus s_j^2} (\pi_k^1 - \min(\pi_k^1, \pi_k^2)) \\ & \prod_{k \in s_j^2 \setminus s_i^1} (\pi_k^2 - \min(\pi_k^1, \pi_k^2)) \\ & \prod_{k \in s_i^1 \cap s_j^2} \min(\pi_k^1, \pi_k^2) \prod_{k \in U \setminus (s_i^1 \cup s_j^2)} (1 - \max(\pi_k^1, \pi_k^2)). \end{aligned}$$

An optimal solution for the problem (3.2) can be obtained directly by using the definition above in the case of maximal sample co-ordination. This solution has the property that its optimal objective function is equal to the absolute upper bound.

When the inclusion probabilities are equal in each occasion, Poisson bi-sampling reduces to Bernoulli bi-sampling.

Example 3.3.5 *Let $U = \{1, 2\}$. Consider two Poisson sampling designs with $\pi_1^1 = 1/2, \pi_2^1 = 1/4, \pi_1^2 = 1/3, \pi_2^2 = 2/3$. We have $\sum_{k \in U} \min(\pi_k^1, \pi_k^2) = 0.583$. Table 3.5 gives the values of $p(s)$ in the case of maximal sample co-ordination. Matrix $\mathbf{C} = (c_{ij})_{m \times q}$ is given in Table 3.6. The absolute upper bound is reached.*

Table 3.5: Poisson sampling

	{}	{1}	{2}	{1,2}	Σ
{}	0.167	0	0.208	0	0.375
{1}	0.056	0.111	0.069	0.139	0.375
{2}	0	0	0.125	0	0.125
{1,2}	0	0	0.042	0.083	0.125
Σ	0.223	0.111	0.444	0.222	1

3.3.2 Example where the absolute upper bound cannot be reached

In stratification, when some units change from a stratum to another, the absolute upper bound cannot always be reached. Consider the following simple example:

Table 3.6: Values of c_{ij} in the case of Poisson sampling

	{}	{1}	{2}	{1,2}
{}	0	0	0	0
{1}	0	1	0	1
{2}	0	0	1	1
{1,2}	0	1	1	2

Example 3.3.6 *Let* $U = \{1, 2, 3, 4\}$ *and let the marginal probabilities be*

$$p^1(\{1,3\}) = p^1(\{2,3\}) = p^1(\{1,4\}) = p^1(\{2,3\}) = 1/4,$$

$$p^2(\{1,2\}) = p^2(\{1,4\}) = p^2(\{2,3\}) = p^2(\{3,4\}) = 1/4.$$

In those sampling designs, all the inclusion probabilities are equal to 0.5. *Both designs are stratified and only one unit is selected in each stratum. The definition of the strata is not the same for both designs. Table 3.7 gives the values of* c_{ij}. *The set of optimal solutions is given in Table 3.8. The constant* d *can*

Table 3.7: Values of c_{ij} for stratified sampling designs

	{1,2}	{1,4}	{2,3}	{3,4}
{1,3}	1	1	1	1
{2,3}	1	0	2	1
{1,4}	1	2	0	1
{2,4}	1	1	1	1

be chosen freely in $[-1/8, 1/8]$. *We have*

$$\sum_{i=1}^{m}\sum_{j=1}^{q} c_{ij} = 2 \times \left(\frac{1}{8} + d\right) \times 1 + 2 \times \left(\frac{1}{8} - d\right) \times 1 + 2 \times \frac{1}{4} \times 2 = 1.5.$$

Nevertheless, the absolute upper bound is $\sum_{k \in U} \min(\pi_k^1, \pi_k^2) = 2$. *In this case, the absolute upper bound cannot be reached.*

Table 3.8: Optimal solutions for stratified sampling designs

	{1,2}	{1,4}	{2,3}	{3,4}	Σ
{1,3}	$1/8+d$	0	0	$1/8-d$	$1/4$
{2,3}	0	0	$1/4$	0	$1/4$
{1,4}	0	$1/4$	0	0	$1/4$
{2,4}	$1/8-d$	0	0	$1/8+d$	$1/4$
Σ	$1/4$	$1/4$	$1/4$	$1/4$	1

3.3.3 Conditions for maximal sample co-ordination

Definition 3.3.2 *The relative upper bound is the value of the optimal objective function of the problem (3.2).*

The relative upper bound is smaller or equal to the absolute upper bound, i.e.

$$\max \sum_{i=1}^{m} \sum_{j=1}^{q} c_{ij} p_{ij} \leq \sum_{k \in U} \min(\pi_k^1, \pi_k^2).$$

We have

$$\sum_{i=1}^{m} \sum_{j=1}^{q} c_{ij} p_{ij} = \sum_{k \in U} \pi_k^{1,2}. \tag{3.4}$$

The relative upper bound is equal to the absolute upper bound when $\pi_k^{1,2} = \min(\pi_k^1, \pi_k^2)$, for all $k \in U$. In this case, the sample co-ordination is maximal.

Proposition 1 *The absolute upper bound is reached iff the following two relations are fulfilled:*

a. if $k \in (s_i^1 \backslash s_j^2) \cap I$ *then* $p_{ij} = 0$,

b. if $k \in (s_j^2 \backslash s_i^1) \cap D$ *then* $p_{ij} = 0$,

for all $k \in U$.

Proof 1 *Necessity: Suppose that* $\pi_k^{1,2} = \min(\pi_k^1, \pi_k^2)$ *for all* $k \in U$. *For the case where* $k \in I$

$$\begin{aligned}
\pi_k^1 &= \sum_{s_i^1 \ni k} p^1(s_i^1) \\
&= \sum_{s_i^1 \ni k} \sum_{s_j^2 \in \mathcal{S}_2} p(s_i^1, s_j^2) \\
&= \sum_{s_i^1 \ni k} \sum_{s_j^2 \ni k} p(s_i^1, s_j^2) + \sum_{s_i^1 \ni k} \sum_{s_j^2 \not\ni k} p(s_i^1, s_j^2) \\
&= \pi_k^{1,2} + \sum_{s_i^1 \ni k} \sum_{s_j^2 \not\ni k} p(s_i^1, s_j^2).
\end{aligned}$$

The assumption $\pi_k^1 = \pi_k^{12}$, *for all* $k \in U$ *implies* $\sum_{s_i^1 \ni k} \sum_{s_j^2 \not\ni k} p(s_i^1, s_j^2) = p_{ij} = 0$, *i.e.* $p_{ij} = 0$. *A similar development can be done for the case where* $k \in D$ *in the condition b.*

Sufficiency: Suppose that the relations a and b are fulfilled. We show that the absolute upper bound is reached.

$$\begin{aligned}
\sum_{i=1}^{m} \sum_{j=1}^{q} c_{ij} p_{ij} &= \sum_{k \in U} \pi_k^{1,2} = \sum_{k \in U} \sum_{s_i^1 \ni k} \sum_{s_j^2 \ni k} p_{ij} \\
&= \sum_{k \in U} \sum_{s_i^1 \ni k} \sum_{s_j^2 \ni k} p_{ij} + \sum_{\substack{k \in U, \\ \min(\pi_k^1, \pi_k^2) = \pi_k^1}} \sum_{s_i^1 \ni k} \sum_{s_j^2 \not\ni k} p_{ij} + \sum_{\substack{k \in U, \\ \min(\pi_k^1, \pi_k^2) = \pi_k^2}} \sum_{s_i^1 \not\ni k} \sum_{s_j^2 \ni k} p_{ij} \\
&= \sum_{\substack{k \in U, \\ \min(\pi_k^1, \pi_k^2) = \pi_k^1}} \Big(\sum_{s_i^1 \ni k} \sum_{s_j^2 \ni k} p_{ij} + \sum_{s_i^1 \ni k} \sum_{s_j^2 \not\ni k} p_{ij} \Big) \\
&\quad + \sum_{\substack{k \in U, \\ \min(\pi_k^1, \pi_k^2) = \pi_k^2}} \Big(\sum_{s_j^2 \ni k} \sum_{s_i^1 \ni k} p_{ij} + \sum_{s_j^2 \ni k} \sum_{s_i^1 \not\ni k} p_{ij} \Big) \\
&= \sum_{\substack{k \in U, \\ \min(\pi_k^1, \pi_k^2) = \pi_k^1}} \sum_{s_i^1 \ni k} \sum_{s_j^2 \in \mathcal{S}_2} p_{ij} + \sum_{\substack{k \in U, \\ \min(\pi_k^1, \pi_k^2) = \pi_k^2}} \sum_{s_j^2 \ni k} \sum_{s_i^1 \in \mathcal{S}_1} p_{ij}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{\substack{k\in U,\\ \min(\pi_k^1,\pi_k^2)=\pi_k^1}} \sum_{s_i^1\ni k} p^1(s_i^1) + \sum_{\substack{k\in U,\\ \min(\pi_k^1,\pi_k^2)=\pi_k^2}} \sum_{s_j^2\ni k} p^2(s_j^2) \\
&= \sum_{\substack{k\in U,\\ \min(\pi_k^1,\pi_k^2)=\pi_k^1}} \pi_k^1 + \sum_{\substack{k\in U,\\ \min(\pi_k^1,\pi_k^2)=\pi_k^2}} \pi_k^2 \\
&= \sum_{k\in U} \min(\pi_k^1,\pi_k^2).
\end{aligned}
$$

Proposition 1 shows that any feasible solution for the problem (3.2), which satisfies the conditions a and b, has the property that its objective function is equal to the absolute upper bound. Proposition 1 also gives a method to put zeros in the matrix $\mathbf{P} = (p_{ij})_{m\times q}$ associated to an optimal solution. Note that the necessary and sufficient condition is obviously satisfied in Examples 3.3.3, 3.3.4 and 3.3.5, and is not satisfied in Example 3.3.6.

Proposition 2 *Suppose that all samples have the corresponding probabilities strictly positive, and the relations a and b of Proposition 1 are satisfied. Let be $s_i^1 \in S_1$. If at least one of the following conditions is fulfilled for all $s_j^2 \in \mathcal{S}_2$:*

1) $(s_i^1 \backslash s_j^2) \cap I \neq \emptyset$,

2) $(s_j^2 \backslash s_i^1) \cap D \neq \emptyset$,

the two designs cannot be maximally co-ordinated. This proposition holds in the symmetric sense, too (if s_j^2 is fixed and at least one of the conditions 1 and 2 is fulfilled, for all $s_i^1 \in \mathcal{S}_1$).

Proof 2 *Suppose that two designs can be maximally co-ordinated. Since $(s_i^1 \backslash s_j^2) \cap I \neq \emptyset$, from condition a of Proposition 1 it follows that $p(s_i^1, s_j^2) = 0$. The second relation is fulfilled similarly from condition b of Proposition 1. We have $p(s_i^1, s_j^2) = 0$, for all $s_j^2 \in \mathcal{S}_2$. So $p^1(s_i^1) = 0$. We obtain a contradiction with $p^1(s_i^1) > 0$. The proof is analogous for the symmetric sense of affirmation.*

Example 3.3.7 *Let $U = \{1,2,3,4\}, I = \{3,4\}$ and $D = \{1,2\}$. Two designs with fixed sample size 2 and 3 respectively are considered. In Table 3.9 the zero values are presented. By x is denoted a non-zero value. The sample $\{3,4\}$ in the first occasion has on its row only zero values. The two designs cannot be maximally co-ordinated since $p^1(\{3,4\}) \neq 0$.*

Table 3.9: Impossible maximal co-ordination

	{1,2,3}	{1,2,4}	{1,3,4}	{2,3,4}
{1,2}	x	x	x	x
{1,3}	0	0	x	0
{1,4}	0	0	x	0
{2,3}	0	0	0	x
{2,4}	0	0	0	x
{3,4}	0	0	0	0

Example 3.3.8 *Suppose* $U = \{1, 2, 3, 4, 5\}$ *and the unit 5 is coming in population in the second wave. So* $\pi_5^1 = 0$. *Two sampling designs with fixed sample size 3 are considered. Let* $I = \{1, 3, 4, 5\}, D = \{2\}$. *Table 3.10 gives the zero-values and the non-zero values denoted by* x. *The sample* $\{2, 3, 5\}$ *in the second occasion has on its column only zero values. The two designs cannot be maximally co-ordinated since* $p^2(\{2, 3, 5\}) \neq 0$.

Table 3.10: Impossible maximal co-ordination

	{1,2,3}	{1,2,4}	{1,2,5}	{1,3,4}	{1,3,5}	{1,4,5}	{2,3,4}	{2,3,5}	{2,4,5}	{3,4,5}
{1,2,3}	x	x	x	x	x	x	0	0	0	x
{1,2,4}	0	x	0	0	0	x	0	0	0	0
{1,3,4}	0	0	0	x	0	x	0	0	0	x
{2,3,4}	0	x	0	x	0	x	x	0	x	x

3.4 An algorithm for maximal co-ordination

The following algorithm is based on the Propositions 1 and 2. Let $\mathbf{P} = (p_{ij})_{m\times q}$ be the matrix which corresponds to a feasible solution for problem (3.2). Using Proposition 1, matrix $\mathbf{P}$ is modified by setting zero values to p_{ij}. Now, the total rows and columns of $\mathbf{P}$ are different from the initial values and the constraints

of problem 3.2 are not respected. In order to have the same totals, the non-zero internal values are modified by using the IPF procedure. The algorithm gives an optimal solution in the case where the absolute upper bound can be reached. Otherwise, a message is given. Algorithm 3.4.4 is the proposed algorithm.

Algorithm 3.4.4 The proposed algorithm

1: Let $\mathbf{P} = (p_{ij})_{m\times q}$ be the matrix given by the independence between both designs: $p_{ij} = p^1(s_i^1)p^2(s_j^2)$, for all $i = 1,\ldots,m, j = 1,\ldots,q$;
2: Put the zeros in $\mathbf{P}$ by using Proposition 1.
3: **if** the conditions of Proposition 2 are satisfied **then**
4: Stop the algorithm and give the message "the absolute upper bound can not be reached";
5: **else**
6: Apply the IPF procedure to modify the non-zero internal values and to restore the margins.
7: **end if**

Concerning the IPF procedure, in a first step indicated by the exponent (1) calculate for all rows $i = 1,\ldots,m$

$$p_{ij}^{(1)} = p_{ij}^{(0)} \frac{p^1(s_i^1)}{p^{1,(0)}(s_i^1)}, \text{ for all } j = 1,\ldots,q, \tag{3.5}$$

where $p_{ij}^{(0)} = p^1(s_i^1)p^2(s_j^2)$ and $p^{1,(0)}(s_i^1) = \sum_{j=1}^{q} p_{ij}^{(0)}$. Now the totals $p^1(s_i^1)$ are satisfied. Calculate in a second step for all columns $j = 1,\ldots,q$

$$p_{ij}^{(2)} = p_{ij}^{(1)} \frac{p^2(s_j^2)}{p^{2,(1)}(s_j^2)}, \text{ for all } i = 1,\ldots,m, \tag{3.6}$$

where $p^{2,(1)}(s_j^2) = \sum_{i=1}^{m} p_{ij}^{(1)}$. Now the totals $p^2(s_j^2)$ are satisfied. In a third step, the resulting $p_{ij}^{(2)}$ are used in recursion (3.5) for obtaining $p_{ij}^{(3)}$, and so on until convergence is attained.

In the 1^{st} step of Algorithm 3.4.4 one can use any value for p_{ij}. We start with the values under independence between both designs for a fast convergence of the IPF procedure. The correctness of the algorithm is assured by Proposition 1.

3.4.1 Algorithm applications

Example 3.4.9 *We take this example from Causey et al. (1985). The two designs are one PSU per stratum. The population has size 5. The inclusion probabilities are* $0.5, 0.06, 0.04, 0.6, 0.1$ *for the first design and* $0.4, 0.15, 0.05, 0.3, 0.1$ *for the second design. In the first design, the first three PSU's were in one initial stratum and the other two in a second initial stratum. There are* $m = 12$ *possible samples given in Table 3.12 with the corresponding probabilities:*

$$0.15, 0.018, 0.012, 0.24, 0.04, 0.3, 0.05, 0.036, 0.006, 0.024, 0.004, 0.12.$$

The second design consists of five PSU's ($q = 5$)*. Causey et al. (1985) solve the linear program associated to this problem and give the value 0.88 as the optimal value for the objective function. Yet,* $\sum_{k \in U} \min(\pi_k^1, \pi_k^2) = 0.9$*. We have* $I = \{2, 3, 5\}, D = \{1, 4\}$*. From Proposition 2, the samples* $\{2, 5\}$ *and* $\{3, 5\}$ *have in theirs rows only zero values. Consequently the two designs cannot be maximally co-ordinated. We modify the example by letting* $\pi_5^1 = 0.2$*. Now,* $I = \{2, 3\}, D = \{1, 4, 5\}$*. The samples in the first design have the corresponding probabilities:*

$$0.1, 0.012, 0.008, 0.24, 0.08, 0.3, 0.1, 0.036, 0.012, 0.024, 0.008, 0.08.$$

We apply the proposed algorithm on matrix $\mathbf{P}$*. The absolute upper bound is now reached. Table 3.11 gives the values of* p_{ij} *after the application of steps 1 and 2 of Algorithm 3.4.4. The resulting matrix* $\mathbf{P}$ *is presented in Table 3.12. Table 3.13 gives the values of* c_{ij}*.*

Example 3.4.10 *This example considers unequal probability designs. Two maximum entropy designs or conditional Poisson sampling designs (see Hájek, 1981) are used, with fixed sample size 3 and 4, respectively. The population size is equal to 6. The first occasion sampling design is presented in Table 3.14. The second occasion sampling design is presented in Table 3.15. The first-order inclusion probabilities are presented in Table 3.16. The absolute upper bound is reached and is equal to* 2.468*. Table 3.17 gives the values of* c_{ij}*. Table 3.18 gives the values of* p_{ij} *after steps 1 and 2 of Algorithm 3.4.4. The resulting matrix* $\mathbf{P}$ *is presented in Table 3.19.*

3.5 Minimal sample co-ordination

A similar algorithm can be constructed in the case of negative co-ordination, when the expected overlap is minimized. In an analogous way, the quantity

Table 3.11: Values of p_{ij} after steps 1 and 2 in Example 3.4.9

	{1}	{2}	{3}	{4}	{5}	Σ
{1}	0.0400	0.015	0.005	0	0	0.0600
{2}	0	0.0018	0	0	0	0.0018
{3}	0	0	0.0004	0	0	0.0004
{4}	0	0.0360	0.0120	0.0720	0	0.1200
{5}	0	0.0120	0.0040	0	0.0080	0.0240
{1,4}	0.1200	0.045	0.015	0.0900	0	0.2700
{1,5}	0.0400	0.015	0.005	0	0.0100	0.0700
{2,4}	0	0.0054	0	0	0	0.0054
{2,5}	0	0.0018	0	0	0	0.0018
{3,4}	0	0	0.0012	0	0	0.0012
{3,5}	0	0	0.0004	0	0	0.0004
∅	0	0.0120	0.0040	0	0	0.0160
Σ	0.2000	0.144	0.047	0.1620	0.0180	1

Table 3.12: Values of p_{ij} after step 3 in Example 3.4.9

	{1}	{2}	{3}	{4}	{5}	Σ
{1}	0.098570	0.001287	0.000143	0	0	0.100
{2}	0	0.012	0	0	0	0.012
{3}	0	0	0.008	0	0	0.008
{4}	0	0.009583	0.001065	0.229352	0	0.240
{5}	0	0.003194	0.000355	0	0.076451	0.080
{1,4}	0.226073	0.002952	0.000328	0.070648	0	0.300
{1,5}	0.075358	0.000984	0.000109	0	0.023549	0.100
{2,4}	0	0.036	0	0	0	0.036
{2,5}	0	0.012	0	0	0	0.012
{3,4}	0	0	0.024	0	0	0.024
{3,5}	0	0	0.008	0	0	0.008
∅	0	0.072	0.008	0	0	0.080
Σ	0.400	0.150	0.050	0.300	0.100	1

Table 3.13: Values of c_{ij} in Example 3.4.9

	{1}	{2}	{3}	{4}	{5}
{1}	1	0	0	0	0
{2}	0	1	0	0	0
{3}	0	0	1	0	0
{4}	0	0	0	1	0
{5}	0	0	0	0	1
{1,4}	1	0	0	1	0
{1,5}	1	0	0	0	1
{2,4}	0	1	0	1	0
{2,5}	0	1	0	0	1
{3,4}	0	0	1	1	0
{3,5}	0	0	1	0	1
∅	0	0	0	0	0

Table 3.14: First occasion sampling design in Example 3.4.10

i	s_i^1	$p^1(s_i^1)$	i	s_i^1	$p^1(s_i^1)$
1	{1,2,3}	0.023719	11	{2,3,4}	0.033355
2	{1,2,4}	0.293520	12	{2,3,5}	0.006589
3	{1,2,5}	0.057979	13	{2,3,6}	0.012707
4	{1,2,6}	0.111817	14	{2,4,5}	0.081533
5	{1,3,4}	0.016010	15	{2,4,6}	0.157243
6	{1,3,5}	0.0031626	16	{2,5,6}	0.031060
7	{1,3,6}	0.006099	17	{3,4,5}	0.004447
8	{1,4,5}	0.039137	18	{3,4,6}	0.008577
9	{1,4,6}	0.07548	19	{3,5,6}	0.001694
10	{1,5,6}	0.014909	20	{4,5,6}	0.020966.

$\sum_{k\in U} \max(0, \pi_k^1 + \pi_k^2 - 1)$ is called the absolute lower bound. Retaining the same constraints, we now seek to minimize the objective function of the prob-

Table 3.15: Second occasion sampling design in Example 3.4.10

j	s_j^2	$p^2(s_j^2)$	j	s_j^2	$p^2(s_j^2)$
1	{1,2,3,4}	0.008117	9	{1,3,4,6}	0.002175
2	{1,2,3,5}	0.210778	10	{1,3,5,6}	0.056474
3	{1,2,3,6}	0.045342	11	{1,4,5,6}	0.014264
4	{1,2,4,5}	0.053239	12	{2,3,4,5}	0.034269
5	{1,2,4,6}	0.011453	13	{2,3,4,6}	0.007372
6	{1,2,5,6}	0.297428	14	{2,3,5,6}	0.191446
7	{1,3,4,5}	0.010109	15	{2,4,5,6}	0.048356
8	{3,4,5,6}	0.009182			

Table 3.16: Inclusion probabilities in Example 3.4.10

unit k	1	2	3	4	5	6
π_k^1	0.641830	0.809522	0.116359	0.730264	0.261477	0.440549
π_k^2	0.709377	0.907798	0.575260	0.198533	0.925542	0.683490

lem (3.2). In general,

$$\sum_{i=1}^{m}\sum_{j=1}^{q} c_{ij}p_{ij} \geq \sum_{k\in U} \max(0, \pi_k^1 + \pi_k^2 - 1).$$

By setting $\max(0, \pi_k^1 + \pi_k^2 - 1) = \pi_k^{1,2}$, for all $k \in U$ a proposition similarly to Proposition 1 is given below.

Proposition 3 *The absolute lower bound is reached iff the following conditions are fulfilled:*

a. if $(k \in s_i^1 \cap s_j^2$ *and* $\pi_k^{1,2} = 0)$, *then* $p_{ij} = 0$,

b. if $(k \notin s_i^1 \cup s_j^2$ *and* $\pi_k^{1,2} = \pi_k^1 + \pi_k^2 - 1)$, *then* $p_{ij} = 0$,

for all $k \in U$.

The proof is similar to Proof 1.

Table 3.17: Values of c_{ij} in Example 3.4.10

	s_1^2	s_2^2	s_3^2	s_4^2	s_5^2	s_6^2	s_7^2	s_8^2	s_9^2	s_{10}^2	s_{11}^2	s_{12}^2	s_{13}^2	s_{14}^2	s_{15}^2
s_1^1	3	3	3	2	2	2	2	2	2	1	2	2	2	1	1
s_2^1	3	2	2	3	3	2	2	2	1	2	2	2	1	2	1
s_3^1	2	3	2	3	2	3	2	1	2	2	2	1	2	2	1
s_4^1	2	2	3	2	3	3	1	2	2	2	1	2	2	2	1
s_5^1	3	2	2	2	2	1	3	3	2	2	2	2	1	1	2
s_6^1	2	3	2	2	1	2	3	2	3	2	2	1	2	1	2
s_7^1	2	2	3	1	2	2	2	3	3	2	1	2	2	1	2
s_8^1	2	2	1	3	2	2	3	2	2	3	2	1	1	2	2
s_9^1	2	1	2	2	3	2	2	3	2	3	1	2	1	2	2
s_{10}^1	1	2	2	2	2	3	2	2	3	3	1	1	2	2	2
s_{11}^1	3	2	2	2	2	1	2	2	1	1	3	3	2	2	2
s_{12}^1	2	3	2	2	1	2	2	1	2	1	3	2	3	2	2
s_{13}^1	2	2	3	1	2	2	1	2	2	1	2	3	3	2	2
s_{14}^1	2	2	1	3	2	2	2	1	1	2	3	2	2	3	2
s_{15}^1	2	1	2	2	3	2	1	2	1	2	2	3	2	3	2
s_{16}^1	1	2	2	2	2	3	1	1	2	2	2	2	3	3	2
s_{17}^1	2	2	1	2	1	1	3	2	2	2	3	2	2	2	3
s_{18}^1	2	1	2	1	2	1	2	3	2	2	2	3	2	2	3
s_{19}^1	1	2	2	1	1	2	2	2	3	2	2	2	3	2	3
s_{20}^1	1	1	1	2	2	2	2	2	2	3	2	2	2	3	3

Table 3.18: Values of p_{ij} after steps 1 and 2 in Example 3.4.10

	s_1^2	s_2^2	s_3^2	s_4^2	s_5^2	s_6^2	s_7^2	s_8^2	s_9^2	s_{10}^2	s_{11}^2	s_{12}^2	s_{13}^2	s_{14}^2	s_{15}^2
s_1^1	0	0.004999	0.001075	0	0	0	0	0	0	0	0	0	0	0	0
s_2^1	0.002382	0.061868	0.013309	0.015627	0.003362	0.087301	0	0	0	0	0	0	0	0	0
s_3^1	0	0.012221	0	0	0	0.017245	0	0	0	0	0	0	0	0	0
s_4^1	0	0	0.005070	0	0	0.033258	0	0	0	0	0	0	0	0	0
s_5^1	0.000130	0.003375	0.000726	0	0	0	0.000162	0.000035	0.000904	0	0	0	0	0	0
s_6^1	0	0.000667	0	0	0	0	0	0	0.000179	0	0	0	0	0	0
s_7^1	0	0	0.000277	0	0	0	0	0	0.000344	0	0	0	0	0	0
s_8^1	0	0.008249	0	0.002084	0	0.011640	0.000396	0	0.002210	0.000558	0	0	0	0	0
s_9^1	0	0	0.003422	0	0.000864	0.022449	0	0.000164	0.004262	0.001077	0	0	0	0	0
s_{10}^1	0	0	0	0	0	0.004434	0	0	0.000842	0	0	0	0	0	0
s_{11}^1	0.000271	0.007030	0.001512	0	0	0	0	0	0	0	0.001143	0.000246	0.006386	0	0
s_{12}^1	0	0.001389	0	0	0	0	0	0	0	0	0	0	0.001261	0	0
s_{13}^1	0	0	0.000576	0	0	0	0	0	0	0	0	0	0.002433	0	0
s_{14}^1	0	0.017185	0	0.004341	0	0.024250	0	0	0	0	0.002794	0	0.015609	0.003943	0
s_{15}^1	0	0	0.007130	0	0.001801	0.046768	0	0	0	0	0	0.001159	0.030103	0.007604	0
s_{16}^1	0	0	0	0	0	0.009238	0	0	0	0	0	0	0.005946	0	0
s_{17}^1	0	0.000937	0	0	0	0	0.000045	0	0.000251	0	0.000152	0	0.000851	0	0.000041
s_{18}^1	0	0	0.000389	0	0	0	0	0.000019	0.000484	0	0	0.000063	0.001642	0	0.000079
s_{19}^1	0	0	0	0	0	0	0	0	0.000096	0	0	0	0.000324	0	0
s_{20}^1	0	0	0	0	0	0.006236	0	0	0.001184	0.000299	0	0	0.004014	0.001014	0.000192

Table 3.19: Values of p_{ij} after step 3 in Example 3.4.10

	s_1^2	s_2^2	s_3^2	s_4^2	s_5^2	s_6^2	s_7^2	s_8^2	s_9^2	s_{10}^2	s_{11}^2	s_{12}^2	s_{13}^2	s_{14}^2	s_{15}^2
s_1^1	0	0.021045	0.002674	0	0	0	0	0	0	0	0	0	0	0	0
s_2^1	0.007444	0.122524	0.015565	0.045903	0.007458	0.094626	0	0	0	0	0	0	0	0	0
s_3^1	0	0.032714	0	0	0	0.025265	0	0	0	0	0	0	0	0	0
s_4^1	0	0	0.015795	0	0	0.096023	0	0	0	0	0	0	0	0	0
s_5^1	0.000381	0.006277	0.000797	0	0	0	0.003695	0.000303	0.004556	0	0	0	0	0	0
s_6^1	0	0.001832	0	0	0	0	0	0	0.001330	0	0	0	0	0	0
s_7^1	0	0	0.000909	0	0	0	0	0	0.005191	0	0	0	0	0	0
s_8^1	0	0.010452	0	0.003916	0	0.008072	0.006152	0	0.007586	0.002959	0	0	0	0	0
s_9^1	0	0	0.004753	0	0.002277	0.028894	0	0.001807	0.027154	0.010592	0	0	0	0	0
s_{10}^1	0	0	0	0	0	0.007686	0	0	0.007223	0	0	0	0	0	0
s_{11}^1	0.000291	0.004784	0.000608	0	0	0	0	0	0	0	0.011417	0.001027	0.015229	0	0
s_{12}^1	0	0.001575	0	0	0	0	0	0	0	0	0	0	0.005014	0	0
s_{13}^1	0	0	0.000488	0	0	0	0	0	0	0	0	0	0.012219	0	0
s_{14}^1	0	0.009130	0	0.003421	0	0.007051	0	0	0	0	0.021792	0	0.029067	0.011072	0
s_{15}^1	0	0	0.003586	0	0.001718	0.021798	0	0	0	0	0	0.006059	0.089856	0.034226	0
s_{16}^1	0	0	0	0	0	0.006064	0	0	0	0	0	0	0.024996	0	0
s_{17}^1	0	0.000444	0	0	0	0	0.000261	0	0.000322	0	0.001059	0	0.001413	0	0.000948
s_{18}^1	0	0	0.000169	0	0	0	0	0.000064	0.000967	0	0	0.000286	0.004243	0	0.002847
s_{19}^1	0	0	0	0	0	0	0	0	0.000315	0	0	0	0.001380	0	0
s_{20}^1	0	0	0	0	0	0.001948	0	0	0.001830	0.000714	0	0	0.008029	0.003058	0.005387

Algorithm 3.4.4 can be applied in the case of minimal sample co-ordination by using Proposition 3 instead of Proposition 1, and the absolute lower bound instead of the absolute upper bound.

3.6 Conclusions

The drawback of using linear programming in sample co-ordination is its huge computational aspect. However, it is possible to construct an algorithm to compute the joint probability of two samples drawn on two different occasions, without solving a linear programming problem. The proposed algorithm is based on the Proposition 1 (3), which identifies the conditions when the absolute upper bound (absolute lower bound) is reached and gives a modality to determine the joint sample probabilities equal to zero. The algorithm uses the IPF procedure, which assures a fast convergence. The algorithm has the complexity $O(m \times q \times$ number of iterations in IPF procedure), which is low compared to linear programming, and it is very easy to implement.

Chapter 4

Evaluation of variance approximations and estimators in maximum entropy sampling with unequal probability and fixed sample size

Abstract
Recent developments in survey sampling allow to quickly draw samples with unequal probability, maximum entropy and fixed sample size. The joint inclusion probabilities can be computed exactly. For this sampling design, 7 approximations and 20 estimators of variance have been computed. A large set of simulations shows that the knowledge of the joint inclusion probabilities is not necessary in order to obtain an accurate variance estimator.
Key words: variance, unequal probabilities, maximum entropy, fixed sample size, simulations.

4.1 Introduction

Two of the most commonly used variance estimators in unequal probabilities sampling design are the Horvitz-Thompson estimator (Horvitz and Thompson, 1952) and the Sen-Yates-Grundy estimator (Yates and Grundy, 1953; Sen, 1953). Both estimators use joint inclusion probabilities. It is often a hard task to evaluate the joint inclusion probabilities. Maximum entropy sampling design with fixed sample size allows the fast and exact computation of these probabilities.

The maximum entropy sampling design with fixed sample size is one of the principal topics of the post-mortem book of Hájek (1981). The principal problem of the implementation of this design was the combinatory explosion of the set of all possible samples of fixed size. In order to implement an algorithm for drawing a maximum entropy sample, a very important result has been given by Chen et al. (1994). They have shown that the maximum entropy sampling design can be presented as a parametric exponential family, and they have proposed an algorithm that allows to pass from its parameter to the first-order and the joint inclusion probabilities and vice versa. In a manuscript paper, Deville (2000b) has improved this algorithm. Chen et al. (1994) and Deville (2000b) pointed out that a fast computation of the parameter allows to build several methods: rejective sampling, sequential sampling, draw by draw sampling. Deville (2000b) has shown that the joint inclusion probabilities can be computed exactly by means of a recursive method, without enumerating the possible samples. Using this method, the variance for the Horvitz-Thompson estimator of the total population can be computed exactly. The joint inclusion probabilities are also used to compute the Horvitz-Thompson and Sen-Yates-Grundy variance estimators. Aires (1999) has provided another method to find the exact expression of the inclusion probabilities in the case of the rejective sampling.

The maximum entropy sampling with fixed sample size design is relatively recent, and therefore, it is not yet sufficiently used in practice. The interesting points of this sampling design are however numerous:

a. Generally, the maximization of the entropy consists of defining a sampling design as random as possible. It is a high entropy situation according to Brewer (2002, p.146) definition when "the resulting relationship between the population and the sample follows no particular pattern" and we expect the variance estimators to perform well. In particular, the simple random sampling without replacement and the Poisson sampling

are maximum entropy sampling designs.

b. In the case of fixed sample size, all the samples have strictly positive probabilities of being selected, and, therefore, the joint inclusion probabilities are strictly positive.

c. The joint inclusion probabilities do not depend on the order of the units, and can be easily computed.

d. The algorithm to compute the inclusion probabilities is fast, and particularly convenient to make simulations.

e. Finally, a simple asymptotic argument allows constructing a family of variance approximations and a large set of variance estimators.

Our aim is to review and evaluate a large set of variance approximations and variance estimators. These are generally applicable to unequal probability designs. We test 7 approximations and 20 estimators of variance in several cases of maximum entropy sampling by means of a set of simulations. The ratio of bias and the mean square error under the simulations are derived. Coverage rates of interval estimates for 95% level are reported. The simulations indicate that the knowledge of the joint inclusion probabilities are not necessary to construct a reasonable estimator of variance in the case of the maximum entropy sampling design with fixed sample size and unequal probabilities.

The paper is organized as follows : in Section 4.2, the notation is defined and the maximum entropy sampling design is reviewed. Interest is then focused on the algorithm which allows the transition from the parameter of the exponential family to the first and second-order inclusion probabilities and vice versa. In Sections 3.4 and 4.4, several approximation and estimator expressions for the variance are reviewed. In Sections 4.5 and 4.6, the empirical results are presented in order to compare the different methods of approximation or estimation to the true value of the variance. Section 4.7 presents the concluding remarks.

4.2 The maximum entropy sampling design

4.2.1 Definition and notation

Let $U = \{1, \ldots, k, \ldots, N\}$ be a finite population of size N. A sample s is a subset of U. A support $\mathcal{R}(U)$ is a set of samples of U. Let $\mathcal{S}(U) = \{s \subset U\}$ be

the full support on U with $\#\mathcal{S}(U) = 2^N$, and let $\mathcal{S}_n(U) = \{s \subset U | \#s = n\}$ be the sample support with fixed sample size equal to n. A sampling design $p(s) > 0, s \in \mathcal{R}(U)$ is a probability distribution on $\mathcal{R}(U)$ such that $\sum_{s\in\mathcal{R}(U)} p(s) = 1$. Let S be a random sample such that $Pr[S = s] = p(s)$. The first-order inclusion probability is defined by

$$\pi_k = Pr[k \in S] = \sum_{s\in\mathcal{R}(U)|s\ni k} p(s), k \in U,$$

and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k, \ldots, \pi_N)'$ is the vector of inclusion probabilities.

The entropy of a sampling design $p(.)$ on $\mathcal{R}(U)$ is given by

$$I(p) = -\sum_{s\in\mathcal{R}(U)} p(s) \log p(s).$$

If we calculate the sampling design on $\mathcal{R}(U)$ which maximizes the entropy under the restrictions given by fixed inclusion probabilities, we get

$$p(s, \mathcal{R}(U), \boldsymbol{\lambda}) = \frac{\exp \boldsymbol{\lambda}'\mathbf{s}}{\sum_{z\in\mathcal{R}(U)} \exp \boldsymbol{\lambda}'\mathbf{z}}, \tag{4.1}$$

where $\boldsymbol{\lambda} \in \mathbb{R}^N$ is the vector of Lagrange multipliers, and $\mathbf{s}$ is a vector of $\mathbb{R}^N$ such that

$$s_k = \begin{cases} 1 & \text{if } k \in s, \\ 0 & \text{if } k \notin s. \end{cases}$$

Chen et al. (1994) pointed out that (4.1) belongs to the exponential family and $\boldsymbol{\lambda}$ is its parameter. One of the characteristics of the exponential family is that there exists a one to one correspondence between the parameter and the expectation (on this topic, see for instance Brown, 1986, p. 74). The expectation is the inclusion probability vector

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

Remark 1 *The sampling design which maximizes the entropy on the full support $\mathcal{S}(U)$, when the inclusion probabilities π_k for all $k \in U$ are fixed, is the Poisson sampling design (see Hájek, 1981, p.30). The interest of the Poisson sampling is the independence between the selection of the units, which allows a very simple sequential implementation. The disadvantage of Poisson sampling is its random sample size. For this reason, fixed sample size methods are often used.*

4.2.2 Maximum entropy sampling design with fixed sample size

When the support is $\mathcal{S}_n(U)$, the problem becomes more intricate, because the denominator of

$$p(s, \mathcal{S}_n(U), \boldsymbol{\lambda}) \quad = \quad \frac{\exp \boldsymbol{\lambda}'\mathbf{s}}{\sum_{z \in \mathcal{S}_n(U)} \exp \boldsymbol{\lambda}'\mathbf{z}}, s \in \mathcal{S}_n(U),$$

cannot be simplified. For this reason, one might believe (before the paper of Chen et al., 1994) that it is not possible to select a sample with this design without enumerating all the samples of $\mathcal{S}_n(U)$.

Let $\boldsymbol{\pi}(\boldsymbol{\lambda}, n)$ be the vector of inclusion probabilities for the maximum entropy sampling design with fixed sample size equal to n. The first problem is the derivation of $\boldsymbol{\pi}(\boldsymbol{\lambda}, n)$ from $\boldsymbol{\lambda}$, which is theoretically given by

$$\boldsymbol{\pi}(\boldsymbol{\lambda}, n) \quad = \quad \frac{\sum_{s \in \mathcal{S}_n(U)} \mathbf{s} \exp \boldsymbol{\lambda}'\mathbf{s}}{\sum_{s \in \mathcal{S}_n(U)} \exp \boldsymbol{\lambda}'\mathbf{s}}. \tag{4.2}$$

Unfortunately, expression (4.2) becomes not feasible to compute when U is large, because it becomes impossible to enumerate all the samples. Nevertheless, Chen et al. (1994) have shown a recursive relation between $\boldsymbol{\pi}(\boldsymbol{\lambda}, n-1)$ and $\boldsymbol{\pi}(\boldsymbol{\lambda}, n)$, which allows to pass from $\boldsymbol{\lambda}$ to $\boldsymbol{\pi}(\boldsymbol{\lambda}, n)$, without enumerating all the possible samples of $\mathcal{S}(U)$.

Result 1 *(Chen et al., 1994) For the first-order inclusion probabilities of the maximum entropy fixed sample size (size equal with n)*

$$\pi_k(\boldsymbol{\lambda}, n) = n \frac{\exp \lambda_k [1 - \pi_k(\boldsymbol{\lambda}, n-1)]}{\sum_{\ell \in U} \exp \lambda_\ell [1 - \pi_\ell(\boldsymbol{\lambda}, n-1)]}. \tag{4.3}$$

A proof of Result 1 is given in Appendix 1. Since $\pi_k(\boldsymbol{\lambda}, 0) = 0$, for all $k \in U$, this recursive relation allows computing quickly the inclusion probability vector.

Another recursive relation (Deville, 2000b) allows to compute the joint inclusion probabilities.

Result 2 *(Deville, 2000b) For the joint inclusion probabilities of the maximum entropy fixed sample size (size equal to n)*

$$\pi_{k\ell}(\boldsymbol{\lambda}, n)$$
$$= \frac{n(n-1) \exp \lambda_k \exp \lambda_\ell [1 - \pi_k(\boldsymbol{\lambda}, n-2) - \pi_\ell(\boldsymbol{\lambda}, n-2) + \pi_{kl}(\boldsymbol{\lambda}, n-2)]}{\sum_{i \in U} \sum_{\substack{j \in U \\ j \neq i}} \exp \lambda_i \exp \lambda_j [1 - \pi_i(\boldsymbol{\lambda}, n-2) - \pi_j(\boldsymbol{\lambda}, n-2) + \pi_{ij}(\boldsymbol{\lambda}, n-2)]},$$

with

$$\pi_{k\ell}(\boldsymbol{\lambda},0)=\pi_{k\ell}(\boldsymbol{\lambda},1)=0, \pi_{k\ell}(\boldsymbol{\lambda},2)=\frac{2\exp\lambda_k\exp\lambda_\ell}{\sum_{i\in U}\sum_{\substack{j\in U\\ j\neq i}}\exp\lambda_i\exp\lambda_j}, k,\ell\in U, k\neq\ell.$$

A proof of Result 2 is given in Appendix 2.

In practice, the inclusion probabilities are generally fixed, and the main problem is to compute $\boldsymbol{\lambda}$ from a given inclusion probability vector $\boldsymbol{\pi}$. The knowledge of $\boldsymbol{\lambda}$ permits to calculate the inclusion probabilities and the joint inclusion probabilities for the maximum entropy with fixed sample size design using Results 1 and 2. It is important to point out that the first-order inclusion probabilities of the Poisson design (which maximizes the entropy, but does not have a fixed sample size), denoted by $\widetilde{\boldsymbol{\pi}}$, are not the same as the inclusion probabilities of fixed sample size design, denoted by $\boldsymbol{\pi}$. Deville (2000b) has shown that $\widetilde{\boldsymbol{\pi}}$ can be obtained by means of Algorithm 4.2.5, which is an application of the Newton method. It is straightforward $\lambda_k=\log[\widetilde{\pi}_k/(1-\widetilde{\pi}_k)]$.

Algorithm 4.2.5 Computation of $\widetilde{\boldsymbol{\pi}}$

- Define

$$\boldsymbol{\phi}(\widetilde{\boldsymbol{\pi}},n)=n\frac{\frac{\widetilde{\pi}_k}{1-\widetilde{\pi}_k}\{1-\phi_k(\widetilde{\boldsymbol{\pi}},n-1)\}}{\sum_{\ell\in U}\frac{\widetilde{\pi}_\ell}{1-\widetilde{\pi}_\ell}\{1-\phi_\ell(\widetilde{\boldsymbol{\pi}},n-1)\}}, \text{ with } \boldsymbol{\phi}(\widetilde{\boldsymbol{\pi}},0)=0.$$

- Set $\widetilde{\boldsymbol{\pi}}^{(0)}=\boldsymbol{\pi}$ and, for $i=1,2,\ldots,$ until convergence

$$\widetilde{\boldsymbol{\pi}}^{(i)}=\widetilde{\boldsymbol{\pi}}^{(i-1)}+\boldsymbol{\pi}-\boldsymbol{\phi}(\widetilde{\boldsymbol{\pi}}^{(i-1)},n). \tag{4.4}$$

A justification of the Algorithm 4.2.5 is given in Appendix 3.

4.2.3 The rejective algorithm

Let y_k be the variable of interest associated with the kth individual in the population, and let $x_k>0$ be an auxiliary variable, which is known for all $k\in U$. The first-order inclusion probabilities are computed using the relation

$$\pi_k=\frac{nx_k}{\sum_{\ell\in U}x_\ell}, \tag{4.5}$$

for all $k \in U$, where n is the sample size. If some $\pi_k > 1$, the value 1 is allocated to these units, and the inclusion probabilities are recalculated using (4.5) on the remaining units.

The rejective procedure follows from Result 3.

Result 3 *For all constant* $c \in \mathbb{R}$

$$\begin{aligned} p(s, \mathcal{S}_n(U), \boldsymbol{\lambda}) &= p(s, \mathcal{S}(U), \boldsymbol{\lambda} + c\mathbf{1} | \#S = n) \\ &= \frac{p(s, \mathcal{S}(U), \boldsymbol{\lambda} + c\mathbf{1})}{\sum_{s \in \mathcal{S}_n(U)} p(s, \mathcal{S}(U), \boldsymbol{\lambda} + c\mathbf{1})}, \end{aligned}$$

for all $s \in \mathcal{S}_n(U)$, *where* $\mathbf{1}$ *is a vector* N *ones.*

The proof is obvious. The rejective method can thus be defined in Algorithm 4.2.6.

Algorithm 4.2.6 Rejective Poisson sampling

---

1. Given $\boldsymbol{\pi}$, compute $\widetilde{\boldsymbol{\pi}}$ with Algorithm 4.2.5; next compute (if needed) $\boldsymbol{\lambda}$ by

$$\lambda_k = \log \frac{\widetilde{\pi}_k}{1 - \widetilde{\pi}_k}.$$

   Eventually, vector λ_k can be normalized such that

$$\sum_{k \in U} \lambda_k = 0.$$

2. Select a random sample $\widetilde{S}$, using Poisson design $p(\widetilde{s}, \mathcal{S}(U), \boldsymbol{\lambda} + c\mathbf{1})$. If the sample size is not equal to n, repeat the selection until the sample size is equal to n.

Since the constant c can be any real number, it should be chosen in order to maximize 1/Pr(#S=n), which can be achieved by using the Newton algorithm. A simpler way to fix the value of c consists in using a constant such that

$$\sum_{k \in U} \widetilde{\pi}_k = \sum_{k \in U} \frac{\exp(\lambda_k + c)}{1 + \exp(\lambda_k + c)} = n. \tag{4.6}$$

Note that Algorithm 4.2.5 provides $\widetilde{\pi}_k$'s that have directly such properties.

4.3 Variance approximations for unequal probability sampling

A review of some variance approximations and variance estimators is presented below. Our aim is to compare different variance approximations as well as different variance estimators for the Horvitz-Thompson estimator

$$\widehat{Y}_\pi = \sum_{k \in S} \frac{y_k}{\pi_k}$$

of the total population

$$Y = \sum_{k \in U} y_k.$$

The variance of the Horvitz-Thompson estimator $\widehat{Y}_\pi$ for a fixed sample size is (see Yates and Grundy, 1953; Sen, 1953)

$$\text{var}[\widehat{Y}_\pi] = -\frac{1}{2} \sum_{k \in U} \sum_{\substack{\ell \in U \\ \ell \neq k}} \left(\frac{y_k}{\pi_k} - \frac{y_\ell}{\pi_\ell} \right)^2 (\pi_{k\ell} - \pi_k \pi_\ell). \tag{4.7}$$

Seven variance approximations and twenty variance estimators have been compared using simulations. The notation for each approximation and each estimator is given in the parenthesis in the corresponding paragraph (e.g. $\text{var}_{\text{Hajek}_1}$ for the approximation Hájek 1). For simplicity, in the next formulae, the first-order inclusion probabilities $\pi_k(\lambda, n)$ are denoted by π_k, and the joint inclusion probabilities $\pi_{k\ell}(\lambda, n)$ are denoted by $\pi_{k\ell}$.

Result 3 shows that a sampling design $p(s)$ which maximizes the entropy and has the inclusion probabilities π_k can be viewed as a conditional Poisson sampling design $\widetilde{p}(s)$ given that its sample size $\widetilde{n}_S$ is fixed. If $\text{var}_{poiss}(.)$ denotes the variance and $\text{cov}_{poiss}(.)$ the covariance under the Poisson sampling $\widetilde{p}(s)$ and var(.) the variance under the design $p(.)$, we can write

$$\text{var}\left(\widehat{Y}_\pi\right) = \text{var}_{poiss}\left(\widehat{Y}_\pi | \widetilde{n}_S = n\right).$$

If we suppose that the couple $(\widehat{Y}_\pi, \widetilde{n}_S)$ has a bivariate normal distribution (on this topic see Hájek, 1964; Berger, 1998a), we obtain

$$\text{var}_{poiss}\left(\widehat{Y}_\pi | \widetilde{n}_S = n\right) = \text{var}_{poiss}\left(\widehat{Y}_\pi + (n - \widetilde{n}_S)\beta\right),$$

where

$$\beta = \frac{\text{cov}_{poiss}\left(\widetilde{n}_S, \widehat{Y}_\pi\right)}{\text{var}_{poiss}\left(\widetilde{n}_S\right)},$$

$$\text{var}_{poiss}\left(\widetilde{n}_S\right) = \sum_{k \in U} \widetilde{\pi}_k (1 - \widetilde{\pi}_k).$$

and

$$\text{cov}_{poiss}\left(\widetilde{n}_S, \widehat{Y}_\pi\right) = \sum_{k \in U} \widetilde{\pi}_k (1 - \widetilde{\pi}_k) \frac{y_k}{\pi_k}.$$

Defining $b_k = \widetilde{\pi}_k(1 - \widetilde{\pi}_k)$, we get the following general approximation of the variance for a sampling design with maximum entropy (see Deville and Tillé, 2005; Tillé, 2001, p.117)

$$\text{var}_{approx}\left[\widehat{Y}_\pi\right] = \sum_{k \in U} \frac{b_k}{\pi_k^2} \left(y_k - y_k^*\right)^2, \tag{4.8}$$

where

$$y_k^* = \pi_k \beta = \pi_k \frac{\sum_{\ell \in U} b_\ell y_\ell / \pi_\ell}{\sum_{\ell \in U} b_\ell}.$$

According to the values given to b_k, some variants of this approximation are obtained and presented below.

Hájek approximation 1 ($\text{var}_{\text{Hajek}_1}$)

The most common value for b_k has been proposed by Hájek (1981)

$$b_k = \frac{\pi_k (1 - \pi_k) N}{N - 1}, \tag{4.9}$$

(on this topic see also Rosén, 1997b; Tillé, 2001).

Approximation under sampling with replacement (var_{repl})

A simpler value for b_k could be

$$b_k = \pi_k \frac{N}{N - 1}, \tag{4.10}$$

which leads to the variance under sampling with replacement.

Naive approximation ($\mathbf{var}_{\text{naive}}$)

A finite population correction can be added to (4.10) and thus

$$b_k = \pi_k \frac{N-n}{N} \frac{N}{N-1} = \pi_k \frac{N-n}{N-1},$$

in order to obtain the variance of simple random sampling without replacement in case of equal inclusion probabilities.

Fixed-point approximation ($\mathbf{var}_{\text{Fix}}$)

Deville and Tillé (2005) have proposed to solve the following equation system to find another approximation of b_k

$$b_k - \frac{b_k^2}{\sum_{\ell \in U} b_\ell} = \pi_k(1-\pi_k). \tag{4.11}$$

Since the equation system (4.11) is not linear, the coefficients b_k can be obtained by the fixed-point technique, using the following recurrence equation, until convergence

$$b_k^{(i)} = \frac{\left(b_k^{(i-1)}\right)^2}{\sum_{l \in U} b_l^{(i-1)}} + \pi_k(1-\pi_k), \tag{4.12}$$

for $i = 0, 1, 2, 3, \ldots$, and using the initialization:

$$b_k^{(0)} = \pi_k(1-\pi_k)\frac{N}{N-1}, k \in U.$$

A necessary condition in order that a solution exists in equation (4.11) is

$$\frac{\pi_k(1-\pi_k)}{\sum_{\ell \in U} \pi_\ell(1-\pi_\ell)} < \frac{1}{2}, \text{ for all } k \text{ in } U.$$

If the method (4.12) is not convergent, we can consider one iteration

$$b_k^{(1)} = \pi_k(1-\pi_k)\left(\frac{N\pi_k(1-\pi_k)}{(N-1)\sum_{\ell \in U} \pi_\ell(1-\pi_\ell)} + 1\right).$$

Hartley-Rao approximation 1 ($\mathbf{var}_{\mathrm{H-Rao_1}}$)

An approximation of variance for the randomized systematic sampling was presented by Hartley and Rao (1962) (see also Brewer and Hanif, 1983)

$$\begin{aligned}\mathrm{var}_{\mathrm{H-Rao_1}}(Y) &= \sum_{k\in U}\pi_k\left(1-\frac{n-1}{n}\pi_k\right)\left(\frac{y_k}{\pi_k}-\frac{Y}{n}\right)^2 \\ &- \frac{n-1}{n^2}\sum_{k\in U}\left(2\pi_k^3-\frac{\pi_k^2}{2}\sum_{\ell\in U}\pi_\ell^2\right)\left(\frac{y_k}{\pi_k}-\frac{Y}{n}\right)^2 \\ &+ \frac{2(n-1)}{n^3}\left(\sum_{k\in U}\pi_k y_k-\frac{Y}{n}\sum_{\ell\in U}\pi_\ell^2\right)^2.\end{aligned}$$

Hartley-Rao approximation 2 ($\mathbf{var}_{\mathrm{H-Rao_2}}$)

In the same paper, Hartley and Rao (1962) have also suggested a simpler expression of variance (see also Brewer and Hanif, 1983)

$$\mathrm{var}_{\mathrm{H-Rao_2}}(Y)=\sum_{k\in U}\pi_k\left(1-\frac{n-1}{n}\pi_k\right)\left(\frac{y_k}{\pi_k}-\frac{Y}{n}\right)^2. \tag{4.13}$$

Hájek approximation 2 ($\mathbf{var}_{\mathrm{Hajek_2}}$)

Brewer (2002, p.153) has used the following estimator, starting from Hájek (1964)

$$\mathrm{var}_{\mathrm{Hajek_2}}(Y)=\sum_{k\in U}\pi_k(1-\pi_k)\left(\frac{y_k}{\pi_k}-\frac{\widetilde{Y}}{n}\right)^2, \tag{4.14}$$

where $\widetilde{Y}=\sum_{k\in U}a_k y_k$ and $a_k=n(1-\pi_k)/\sum_{\ell\in U}\pi_\ell(1-\pi_\ell)$.

4.4 Variance estimators

There are three classes of variance estimators. The first class is composed by the Horvitz-Thompson estimator (Horvitz and Thompson, 1952) and the Sen-Yates-Grundy estimator (Yates and Grundy, 1953; Sen, 1953), which use the first-order and the joint inclusion probabilities. The second class uses only first-order inclusion probabilities for all $k \in S$, while, in the third class,

the variance estimators use only first-order inclusion probabilities, but for all $k \in U$.

4.4.1 First class of variance estimators

Horvitz-Thompson estimator ($\widehat{\text{var}}_{\text{HT}}$)

The expression of this estimator is (see Horvitz and Thompson, 1952)

$$\widehat{\text{var}}_{\text{HT}}[\widehat{Y}_\pi] = \sum_{k \in S} \frac{y_k^2}{\pi_k^2}(1 - \pi_k) + \sum_{k \in S} \sum_{\substack{\ell \in S \\ \ell \neq k}} \frac{y_k y_\ell}{\pi_k \pi_\ell \pi_{k\ell}}(\pi_{k\ell} - \pi_k \pi_\ell). \tag{4.15}$$

This estimator has several important drawbacks. In general, when the variable of interest $y_k \propto \pi_k$, $\text{var}[\widehat{Y}_\pi] = 0$, but $\widehat{\text{var}}_{\text{HT}}$ is not necessary equal to 0 in such a case. The Horvitz-Thompson estimator can also take negative values (on this topic see Cumberland and Royall, 1981). For example, if $y_k = \pi_k$, for all $k \in U$, then $\text{var}[\widehat{Y}_\pi] = 0$, and

$$\widehat{\text{var}}_{\text{HT}}[\widehat{Y}_\pi] = n^2 - \sum_{k \in S} \pi_k - \sum_{k \in S} \pi_k \sum_{\substack{\ell \in S \\ \ell \neq k}} \frac{\pi_\ell}{\pi_{k\ell}},$$

which is generally not null, but has a null expectation. Thus, negative values occur.

Sen-Yates-Grundy estimator ($\widehat{\text{var}}_{\text{SYG}}$)

The expression of this estimator is (see Sen, 1953; Yates and Grundy, 1953)

$$\widehat{\text{var}}_{\text{SYG}}[\widehat{Y}_\pi] = \frac{1}{2} \sum_{k \in S} \sum_{\substack{\ell \in S \\ \ell \neq k}} \left(\frac{y_k}{\pi_k} - \frac{y_\ell}{\pi_\ell} \right)^2 \frac{\pi_k \pi_\ell - \pi_{k\ell}}{\pi_{k\ell}}. \tag{4.16}$$

The Horvitz-Thompson and Sen-Yates-Grundy estimators are unbiased.

4.4.2 Second class of variance estimators

From the expression (4.8), a general variance estimator can be derived (see Deville and Tillé, 2005; Tillé, 2001, p.117)

$$\widehat{\text{var}}[\widehat{Y}_\pi] = \sum_{k \in S} \frac{c_k}{\pi_k^2}(y_k - \widehat{y}_k^*)^2, \tag{4.17}$$

where

$$\widehat{y}_k^* = \pi_k \frac{\sum_{\ell \in S} c_\ell y_\ell / \pi_\ell}{\sum_{\ell \in S} c_\ell}.$$

According to the choice of c_k in (4.17), various estimators have been proposed.

Deville estimator 1 ($\widehat{\mathbf{var}}_{\text{Dev1}}$)

Deville (1993) has proposed a simple value for c_k

$$c_k = (1 - \pi_k) \frac{n}{n-1}.$$

Deville estimator 2 ($\widehat{\mathbf{var}}_{\text{Dev2}}$)

In the same manuscript, Deville (1993) has suggested a more complex value (see also Deville, 1999)

$$c_k = (1 - \pi_k) \left[1 - \sum_{k \in S} \left\{ \frac{1 - \pi_k}{\sum_{\ell \in S} (1 - \pi_\ell)} \right\}^2 \right]^{-1}.$$

Variance under sampling with replacement ($\widehat{\mathbf{var}}_{\text{repl}}$)

A simple value for c_k could be

$$c_k = \frac{n}{n-1}, \tag{4.18}$$

which leads to the variance under sampling with replacement (see Särndal et al., 1992, expression 2.9.9, p.53).

Naive estimator ($\widehat{\mathbf{var}}_{\text{naive}}$)

A finite population correction can be added to (4.18) resulting to

$$c_k = \frac{N - n}{N} \frac{n}{n-1},$$

in order to obtain the variance estimator of simple random sampling without replacement, in the case of equal inclusion probabilities.

Fixed-point estimator ($\widehat{\mathbf{var}}_{\mathrm{Fix}}$)

Deville and Tillé (2005) have proposed to use the following development in order to derive a value for c_k. The estimator defined in expression (4.17) can be written as

$$\widehat{\mathrm{var}}[\widehat{Y}_\pi] = \sum_{k\in S} \frac{y_k^2}{\pi_k^2}\left(c_k - \frac{c_k^2}{\sum_{\ell\in S} c_\ell}\right) - \frac{1}{\sum_{\ell\in S} c_\ell}\sum_{k\in S}\sum_{\substack{\ell\in S\\ \ell\neq k}} \frac{y_k y_\ell c_k c_\ell}{\pi_k \pi_\ell}. \tag{4.19}$$

Using the formula (4.15) of $\widehat{\mathrm{var}}_{\mathrm{HT}}$, we can look for c_k which satisfies the equation

$$c_k - \frac{c_k^2}{\sum_{\ell\in S} c_\ell} = (1-\pi_k). \tag{4.20}$$

These coefficients can be obtained by the fixed-point technique, using the following recurrence equation, until the convergence is fulfilled

$$c_k^{(i)} = \frac{\left(c_k^{(i-1)}\right)^2}{\sum_{\ell\in S} c_\ell^{(i-1)}} + (1-\pi_k), \tag{4.21}$$

for $i = 0, 1, 2, 3, \ldots$ and using the initialization

$$c_k^{(0)} = (1-\pi_k)\frac{n}{n-1}, k \in S.$$

A necessary condition in order that a solution exists in equation (4.20) is

$$\frac{1-\pi_k}{\sum_{\ell\in S}(1-\pi_\ell)} < \frac{1}{2}, \text{ for all } k \text{ in } S.$$

If the method (4.21) is not convergent, we can consider one iteration

$$c_k^{(1)} = (1-\pi_k)\left(\frac{n(1-\pi_k)}{(n-1)\sum_{\ell\in S}(1-\pi_\ell)} + 1\right).$$

Rosen estimator ($\widehat{\mathbf{var}}_{\mathrm{R}}$)

Rosén (1991) suggested the following estimator (see also Ardilly, 1994, p.338)

$$\widehat{\mathrm{var}}_{\mathrm{R}}[\widehat{Y}_\pi] = \frac{n}{n-1}\sum_{k\in S}(1-\pi_k)\left(\frac{y_k}{\pi_k} - A\right)^2,$$

where

$$A = \frac{\sum_{k\in S} y_k \frac{1-\pi_k}{\pi_k^2}\log(1-\pi_k)}{\sum_{k\in S}\frac{1-\pi_k}{\pi_k}\log(1-\pi_k)}. \tag{4.22}$$

Deville estimator 3 ($\widehat{\mathbf{var}}_{\mathrm{Dev_3}}$)

Another proposal of Deville (1993) (see also Ardilly, 1994, p.338) is

$$\widehat{\mathrm{var}}_{\mathrm{Dev3}}[\widehat{Y}_\pi] = \frac{1}{1-\sum_{k\in S} a_k^2} \sum_{k\in S}(1-\pi_k)\left(\frac{y_k}{\pi_k} - \frac{\widehat{Y}_\pi}{n}\right)^2, \tag{4.23}$$

where

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

Estimator 1 ($\widehat{\mathbf{var}}_1$)

We propose a new estimator which is defined as below

$$\widehat{\mathrm{var}}_1[\widehat{Y}_\pi] = \frac{n(N-1)}{N(n-1)} \sum_{k\in S} \frac{b_k}{\pi_k^3}(y_k - \widehat{y}_k^*)^2, \tag{4.24}$$

where

$$\widehat{y}_k^* = \pi_k \frac{\sum_{\ell\in S} b_\ell y_\ell/\pi_\ell^2}{\sum_{\ell\in S} b_\ell/\pi_\ell},$$

and the coefficients b_k are defined in the same way as in expression (4.12).

4.4.3 Third class of variance estimators

Berger estimator ($\widehat{\mathbf{var}}_{\mathrm{Ber}}$)

Berger (1998b) has proposed to use

$$c_k = (1-\pi_k)\frac{n}{n-1}\frac{\sum_{k\in S}(1-\pi_k)}{\sum_{k\in U}\pi_k(1-\pi_k)},$$

in the expression (4.17).

Tillé estimator ($\widehat{\mathbf{var}}_{\mathrm{T}}$)

An approximation of the joint inclusion probabilities by means of adjustment to marginal totals was described by Tillé (1996). Using this approximation and the Sen-Yates-Grundy estimator, the following estimator was developed

$$\widehat{\mathrm{var}}_{\mathrm{T}}[\widehat{Y}_\pi] = \sum_{k\in S}\frac{y_k^2}{\pi_k\beta_k}\sum_{\ell\in S}\frac{\pi_\ell}{\beta_l} - \left(\sum_{k\in S}\frac{y_k}{\beta_k}\right)^2 - n\sum_{k\in S}\frac{y_k^2}{\pi_k^2} + \left(\sum_{k\in S}\frac{y_k}{\pi_k}\right)^2. \quad (4.25)$$

The coefficients β_k are calculated using the following algorithm

$$\beta_k^{(0)} = \pi_k, \text{ for all } k,$$

$$\beta_k^{(2i-1)} = \frac{(n-1)\pi_k}{\beta^{(2i-2)} - \beta_k^{(2i-2)}},$$

$$\beta_k^{(2i)} = \beta_k^{(2i-1)}\left(\frac{n(n-1)}{(\beta^{(2i-1)})^2 - \sum_{k\in U}(\beta_k^{(2i-1)})^2}\right)^{1/2},$$

where

$$\beta^{(i)} = \sum_{k\in U}\beta_k^{(i)}, i = 1, 2, 3, ...$$

The coefficients β_k are used to approximate the joint inclusion probabilities such that $\pi_{k\ell} \approx \beta_k\beta_\ell$. The convergence criterion is ensured by the marginal totals

$$\sum_{\substack{k\in U\\ k\neq\ell}}\pi_{k\ell} = \pi_\ell(n-1), \ell \in U.$$

Some new estimators

Four new estimators (named Estimators 2, 3, 4, 5) of variance can be constructed as follows.

Estimator 2 ($\widehat{\mathrm{var}}_2$)

$$\widehat{\mathrm{var}}_2[\widehat{Y}_\pi] = \frac{1}{1-\sum_{k\in U}\frac{d_k^2}{\pi_k}}\sum_{k\in S}(1-\pi_k)\left(\frac{y_k}{\pi_k} - \frac{\widehat{Y}_\pi}{n}\right)^2, \quad (4.26)$$

where

$$d_k = \frac{\pi_k(1-\pi_k)}{\sum_{\ell\in U}\pi_\ell(1-\pi_\ell)}. \quad (4.27)$$

Estimator 3 ($\widehat{\mathbf{var}}_3$)

$$\widehat{\mathrm{var}}_3[\widehat{Y}_\pi] = \frac{1}{1-\sum_{k\in U}\frac{d_k^2}{\pi_k}}\sum_{k\in S}(1-\pi_k)\left(\frac{y_k}{\pi_k}-\frac{\sum_{\ell\in S}(1-\pi_\ell)\frac{y_\ell}{\pi_\ell}}{\sum_{\ell\in S}(1-\pi_\ell)}\right)^2, \tag{4.28}$$

where d_k is defined as in (4.27).

Estimator 4 ($\widehat{\mathbf{var}}_4$)

$$\widehat{\mathrm{var}}_4[\widehat{Y}_\pi] = \frac{1}{1-\sum_{\ell\in U} b_\ell/n^2}\sum_{k\in S}\frac{b_k}{\pi_k^3}(y_k-\widehat{y}_k^*)^2, \tag{4.29}$$

where

$$\widehat{y}_k^* = \pi_k\frac{\sum_{\ell\in S} b_\ell y_\ell/\pi_\ell^2}{\sum_{\ell\in S} b_\ell/\pi_\ell}, \tag{4.30}$$

and the coefficients b_k are defined in the same way as in expression (4.9).

Estimator 5 ($\widehat{\mathbf{var}}_5$)

$$\widehat{\mathrm{var}}_5[\widehat{Y}_\pi] = \frac{1}{1-\sum_{\ell\in U} b_\ell/n^2}\sum_{k\in S}\frac{b_k}{\pi_k^3}(y_k-\widehat{y}_k^*)^2, \tag{4.31}$$

where $\widehat{y}_k^*$ is defined as in (4.30) and the coefficients b_k are defined in the same way as in expression (4.12).

The Brewer family

A set of high-entropy estimators was presented by Brewer (2002); Brewer and Donadio (2003). According to Brewer, "a plausible sample estimator of the approximate design variance of the Horvitz-Thompson estimator, and one that can be constructed so as to be exactly design-unbiased under simple random sampling without replacement, is"

$$\widehat{\mathrm{var}}_{\mathrm{Br}}[\widehat{Y}_\pi] = \sum_{k\in S}(e_k^{-1}-\pi_k)\left(\frac{y_k}{\pi_k}-\sum_{\ell\in S}\frac{y_\ell}{n\pi_\ell}\right)^2. \tag{4.32}$$

Four particular values for e_k were proposed (see Brewer, 2002, p.152,153,158):

Brewer estimator 1 ($\widehat{\mathbf{var}}_{\mathrm{Br}_1}$)

$$e_k = \frac{n-1}{n - \frac{\sum_{\ell \in U} \pi_\ell^2}{n}}.$$

Brewer estimator 2 ($\widehat{\mathbf{var}}_{\mathrm{Br}_2}$)

$$e_k = \frac{n-1}{n-\pi_k}.$$

In this case, the estimator defined in (4.32) could have been placed in the second category, since it uses the inclusion probabilities only for the sample. In order, however, to keep the Brewer estimators in a single category, we place it here.

Brewer estimator 3 ($\widehat{\mathbf{var}}_{\mathrm{Br}_3}$)

$$e_k = \frac{(n-1)/n}{1 - 2\pi_k/n + \frac{\sum_{\ell \in U} \pi_\ell^2}{n^2}}.$$

Brewer estimator 4 ($\widehat{\mathbf{var}}_{\mathrm{Br}_4}$)

$$e_k = \frac{(n-1)/n}{1 - \frac{(2n-1)\pi_k}{n(n-1)} + \frac{\sum_{\ell \in U} \pi_\ell^2}{n(n-1)}}.$$

4.5 Simulations

Three data sets have been used for Monte-Carlo simulations: mu284 population from Särndal et al. (1992), and two artificial populations. A set of 10000 independent samples without replacement have been selected for each different sample size, $n = 10, 20$ and 40, using the rejective sampling. Table 4.3 gives the expected number of the rejected samples (which have sample size different from the fixed size n) under the simulations. From mu284 population, two data items have been taken: the "revenues from 1985 municipal taxation" for the principal characteristic, and the "1985 population" for the auxiliary variable. Three observations (numbers 16, 114, 137) with large x_k were deleted from this population. Thus, $N = 281, Y = 53151 \times 10^6$. The first artificial population was generated using the model $N = 100, x_k = k, y_k = 5x_k(1 + \epsilon_k)$ (see Figure 4.1), where $\epsilon_k \sim N(0, 1/3), k = 1, \cdots, N$. In this case, $Y = 25482.917$.

For the second artificial population the model used is $N = 100, x_k = k, y_k = 1/\pi_k, \pi_k = nx_k/\sum_{k\in U} x_k, k = 1, \cdots, N$ (see Figure 4.2). In this case, for $n = 10, Y = 2619.625$, for $n = 20, Y = 1309.812$, and for $n = 40, Y = 654.906$.

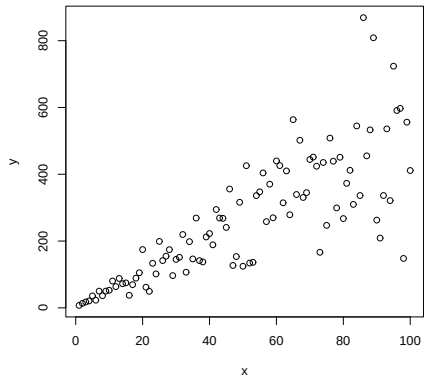


Figure 4.1: Scatter plot for the first artificial population (x versus y).

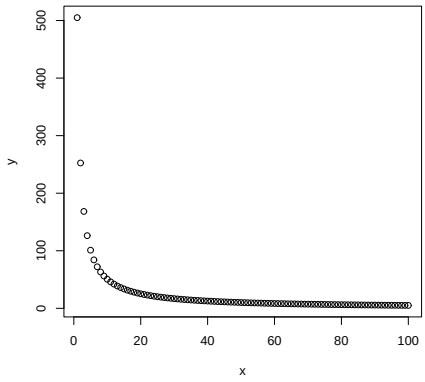


Figure 4.2: Scatter plot for the second artificial population, $n = 10$ (x versus y).

Three measures are used to compare the variance estimators:

- the ratio of bias

$$\mathrm{RB}(\widehat{\mathrm{var}}) = \frac{\mathrm{E}_{sim}(\widehat{\mathrm{var}}) - \mathrm{var}}{\sqrt{\mathrm{var}_{sim}(\widehat{\mathrm{var}})}},$$

  where $\mathrm{E}_{sim}()$ is the average calculated under simulations, $\mathrm{var}_{sim}()$ is the variance calculated under simulations, $\widehat{\mathrm{var}}$ is a variance estimator, and var is the true variance computed from expression (4.7).

- the mean square error

$$\mathrm{MSE}(\widehat{\mathrm{var}}) = \mathrm{var}_{sim}(\widehat{\mathrm{var}}) + (\mathrm{E}_{sim}(\widehat{\mathrm{var}}) - \mathrm{var})^2.$$

- the coverage rate (CR) of an interval estimates for 95% level.

The 95% confidence intervals for the value Y are computed using the t distribution with $n-1$ degrees of freedom $[\widehat{Y_\pi} \pm t_{n-1,0.975}\sqrt{\widehat{\mathrm{var}}}]$. We use the 97.5 quantile of the $t-$distribution with $n-1$ degrees of freedom instead of 1.96 even for $n = 40$ to improve the coverage rate (see Särndal et al., 1992, p.281).

Table 4.1 (for the mu284 population), Table 4.2 (for the first artificial population), and Table 4.4 (for the second artificial population) summarize the performance of the approximations and estimators via simulations. The upper sections of Tables 4.1, 4.2, and 4.4 give the values of the variance approximations presented in Section 3, and the true value, $\text{var}[\widehat{Y}_\pi]$, computed from (4.7). The bottom sections of these tables give the values of RB, MSE and CR for the variance estimators presented in Section 4.4. The ratio of bias and the coverage rates are expressed in percentages. For clarity, a row of exponents is added (for example the Hájek approximation 1 is 3.808×10^{18} in the case of mu284 population, $n = 10$).

4.6 Discussion of the empirical results

The reliable comparison between the different variance approximations is ensured by the fact that the true variance $\text{var}[\widehat{Y}_\pi]$ can be calculated using the formula (4.7). Without any doubt, the fixed-point approximation is the best. The approximations $\text{var}_{\text{Hajek}_1}$, $\text{var}_{\text{H-Rao}_1}$, $\text{var}_{\text{H-Rao}_2}$ and $\text{var}_{\text{Hajek}_2}$ are less precise. The worst results are given by var_{repl} (particulary in the case of the first two populations) and $\text{var}_{\text{naive}}$ (for all populations).

In the case of the variance estimators, the Horvitz-Thompson estimator has the biggest MSE in Tables 4.1 and 4.2. In both populations, the variable of interest y_k and the auxiliary variable x_k are strongly correlated (for the mu284 population the coefficient of correlation is 0.99, and for the first artificial population is 0.86). For the third population (which is badly adapted to the design), the correlation coefficient is approximatively -0.40. In this case, $\widehat{\text{var}}_{\text{HT}}$ performed nearly the same as the other estimators studied. We are led to the same conclusion using an additional simulation study (results not shown in tables), where the variable of interest y_k and the auxiliary variable x_k are not correlated. As seen from the examples above, $\widehat{\text{var}}_{\text{HT}}$ has a big MSE in the cases where y_k and x_k are strongly correlated, which is the usual case in practice. An analytic study of the Horvitz-Thompson variance estimator is given in Stehman and Overton (1994).

Population mu284 as well as the first artificial population arise from a structural model of the form $E(\mathbf{y}) = \beta\mathbf{x}, \text{var}(\mathbf{y}) = \sigma^2\mathbf{x}^2$. In such populations, for sufficiently small sample mean $\mathbf{x}$ and $\beta^2/\sigma^2 > 1$, Cumberland and Royall (1981) showed that $\widehat{\text{var}}_{\text{HT}}$ may take negative values. In the artificial population 1, $\beta^2/\sigma^2 = 9$. We included in Table 4.5 the number of times that $\widehat{\text{var}}_{\text{HT}} < 0$ among the 10000 simulated samples. This could partially explain the large

Table 4.1: Results of simulations for the mu284 population

	n=10			n=20			n=40		
	(10^{18})			(10^{18})			(10^{18})		
$\text{var}_{\text{Hajek}_1}$	3.808			1.778			1.005		
var_{Fix}	3.816			1.782			1.007		
$\text{var}_{\text{H}-\text{Rao}_1}$	3.818			1.788			1.059		
$\text{var}_{\text{H}-\text{Rao}_2}$	3.821			1.789			1.043		
$\text{var}_{\text{Hajek}_2}$	3.794			1.772			1.002		
var_{repl}	4.056			2.031			1.455		
$\text{var}_{\text{naive}}$	3.912			1.887			1.248		
True value	3.817			1.782			1.007		
	RB(%)	MSE (10^{36})	CR(%)	RB(%)	MSE (10^{35})	CR(%)	RB(%)	MSE (10^{34})	CR(%)
$\widehat{\text{var}}_{\text{HT}}$	-1.285	24.239	72.69	0.122	83.829	68.75	-0.657	225.587	71.87
$\widehat{\text{var}}_{\text{SYG}}$	-0.666	6.145	95.19	-0.013	6.041	95.20	-0.638	5.735	94.81
$\widehat{\text{var}}_{\text{Dev1}}$	-0.862	6.097	95.16	-0.343	5.939	95.27	-0.871	5.603	94.83
$\widehat{\text{var}}_{\text{Dev2}}$	-0.814	6.101	95.17	-0.173	5.947	95.27	-0.192	5.621	94.85
$\widehat{\text{var}}_{\text{repl}}$	-7.339	5.655	94.89	18.168	7.221	96.05	134.734	2.679	97.97
$\widehat{\text{var}}_{\text{naive}}$	3.414	6.466	95.45	13.307	6.800	95.85	90.026	1.333	97.04
$\widehat{\text{var}}_{\text{Fix}}$	-0.698	6.104	95.15	-0.054	5.948	95.20	-0.824	5.612	94.84
$\widehat{\text{var}}_{\text{R}}$	-0.835	6.098	95.17	-0.183	5.943	95.27	1.478	5.644	94.89
$\widehat{\text{var}}_{\text{Dev3}}$	-0.699	6.104	95.18	0.539	5.963	95.30	12.945	5.919	95.18
$\widehat{\text{var}}_{1}$	-0.494	6.129	95.22	0.139	5.969	95.29	-0.146	5.626	94.85
$\widehat{\text{var}}_{\text{Ber}}$	-0.697	6.141	95.19	-0.118	6.031	95.23	-0.762	5.719	94.81
$\widehat{\text{var}}_{\text{T}}$	-0.593	6.148	95.19	0.565	6.053	95.28	11.509	6.500	95.15
$\widehat{\text{var}}_{2}$	-0.694	6.105	95.18	0.546	5.964	95.30	12.949	5.918	95.18
$\widehat{\text{var}}_{3}$	-0.808	6.102	95.17	-0.166	5.948	95.28	-0.187	5.620	94.85
$\widehat{\text{var}}_{4}$	-1.429	6.104	95.14	-1.146	5.979	95.24	-2.112	5.755	94.77
$\widehat{\text{var}}_{5}$	-1.019	6.089	95.16	-0.645	5.929	95.25	-1.409	5.594	94.80
$\widehat{\text{var}}_{\text{Br1}}$	-0.869	6.093	95.20	0.177	5.944	95.29	12.746	5.885	95.17
$\widehat{\text{var}}_{\text{Br2}}$	-0.748	6.101	95.17	0.369	5.955	95.29	12.278	5.889	95.18
$\widehat{\text{var}}_{\text{Br3}}$	-0.627	6.109	95.17	0.561	5.966	95.30	11.809	5.894	95.14
$\widehat{\text{var}}_{\text{Br4}}$	-0.614	6.111	95.17	0.571	5.966	95.29	11.797	5.894	95.14

Table 4.2: Results of simulations for the first artificial population

	n=10			n=20			n=40		
	(10^6)			(10^6)			(10^6)		
$\mathrm{var}_{\mathrm{Hajek}_1}$	5.429			2.306			0.745		
$\mathrm{var}_{\mathrm{Fix}}$	5.444			2.312			0.746		
$\mathrm{var}_{\mathrm{H-Rao}_1}$	5.441			2.331			0.858		
$\mathrm{var}_{\mathrm{H-Rao}_2}$	5.455			2.324			0.758		
$\mathrm{var}_{\mathrm{Hajek}_2}$	5.374			2.283			0.737		
$\mathrm{var}_{\mathrm{repl}}$	6.245			3.123			1.563		
$\mathrm{var}_{\mathrm{naive}}$	5.621			2.499			0.938		
True value	5.444			2.312			0.746		
	RB(%)	MSE (10^{12})	CR(%)	RB(%)	MSE (10^{11})	CR(%)	RB(%)	MSE (10^{10})	CR(%)
$\widehat{\mathrm{var}}_{\mathrm{HT}}$	-0.758	7.504	93.71	0.358	8.528	93.09	-1.800	12.931	89.76
$\widehat{\mathrm{var}}_{\mathrm{SYG}}$	-0.827	6.979	94.80	0.669	5.663	94.89	-0.147	2.949	95.31
$\widehat{\mathrm{var}}_{\mathrm{Dev1}}$	-0.949	6.914	94.78	0.299	5.503	94.88	-2.338	2.582	95.19
$\widehat{\mathrm{var}}_{\mathrm{Dev2}}$	-0.887	6.917	94.78	0.557	5.512	94.88	-0.428	2.601	95.24
$\widehat{\mathrm{var}}_{\mathrm{repl}}$	6.444	7.254	95.11	73.064	12.878	97.08	285.872	68.302	99.33
$\widehat{\mathrm{var}}_{\mathrm{naive}}$	6.444	7.254	95.11	25.752	6.349	95.65	114.931	6.545	97.41
$\widehat{\mathrm{var}}_{\mathrm{Fix}}$	-0.895	6.936	94.78	0.581	5.534	94.89	-0.524	2.622	95.23
$\widehat{\mathrm{var}}_{\mathrm{R}}$	-0.936	6.914	94.78	0.346	5.504	94.88	-2.074	2.583	95.19
$\widehat{\mathrm{var}}_{\mathrm{Dev3}}$	-0.833	6.919	94.78	0.774	5.518	94.90	1.115	2.609	95.29
$\widehat{\mathrm{var}}_1$	-0.350	6.939	94.79	1.047	5.519	94.89	-1.576	2.587	95.19
$\widehat{\mathrm{var}}_{\mathrm{Ber}}$	-0.848	6.973	94.79	0.565	5.648	94.89	-1.165	2.917	95.28
$\widehat{\mathrm{var}}_{\mathrm{T}}$	-0.826	6.978	94.80	0.663	5.661	94.89	-0.258	2.921	95.29
$\widehat{\mathrm{var}}_2$	-0.826	6.920	94.78	0.787	5.518	94.90	1.152	2.610	95.29
$\widehat{\mathrm{var}}_3$	-0.880	6.918	94.78	0.569	5.513	94.88	-0.392	2.602	95.24
$\widehat{\mathrm{var}}_4$	-1.699	6.962	94.75	-0.781	5.601	94.84	-3.848	2.659	95.16
$\widehat{\mathrm{var}}_5$	-1.052	6.893	94.77	-0.010	5.481	94.80	-3.147	2.572	95.17
$\widehat{\mathrm{var}}_{\mathrm{Br1}}$	-0.935	6.893	94.77	0.426	5.480	94.87	-1.003	2.559	95.23
$\widehat{\mathrm{var}}_{\mathrm{Br2}}$	-0.895	6.916	94.78	0.514	5.508	94.89	-0.791	2.588	95.22
$\widehat{\mathrm{var}}_{\mathrm{Br3}}$	-0.855	6.939	94.77	0.601	5.537	94.88	-0.581	2.617	95.24
$\widehat{\mathrm{var}}_{\mathrm{Br4}}$	-0.850	6.942	94.77	0.605	5.539	94.88	-0.576	2.618	95.25

Table 4.3: Expected number of the rejected samples under the simulations

	$n = 10$	$n = 20$	$n = 40$
mu284	6.643	9.457	12.372
artificial pop. 1	6.535	8.715	9.841
artificial pop. 2	6.391	8.634	10.013

Table 4.4: Results of simulations for the second artificial population

		n=10			n=20			n=40	
		(10^8)			(10^7)			(10^6)	
$\text{var}_{\text{Hajek}_1}$		1.575			1.966			2.458	
var_{Fix}		1.559			1.948			2.434	
$\text{var}_{\text{H}-\text{Rao}_1}$		1.559			1.948			2.438	
$\text{var}_{\text{H}-\text{Rao}_2}$		1.559			1.948			2.437	
$\text{var}_{\text{Hajek}_2}$		1.559			1.947			2.433	
var_{repl}		1.579			1.978			2.490	
$\text{var}_{\text{naive}}$		1.421			1.583			1.494	
True value		1.559			1.948			2.434	
	RB(%)	MSE (10^{19})	CR(%)	RB(%)	MSE (10^{16})	CR(%)	RB(%)	MSE (10^{14})	CR (%)
$\widehat{\text{var}}_{\text{HT}}$	0.900	1.075	36.13	0.105	6.844	36.98	0.299	5.623	38.73
$\widehat{\text{var}}_{\text{SYG}}$	0.995	1.078	36.12	0.098	6.829	36.87	0.260	5.584	38.60
$\widehat{\text{var}}_{\text{Dev1}}$	0.919	1.043	36.06	-0.028	6.603	36.84	0.007	5.314	38.47
$\widehat{\text{var}}_{\text{Dev2}}$	0.921	1.044	36.06	-0.020	6.618	36.84	0.055	5.364	38.52
$\widehat{\text{var}}_{\text{repl}}$	0.476	0.874	35.65	-0.241	6.229	37.35	0.167	5.443	40.34
$\widehat{\text{var}}_{\text{naive}}$	0.476	0.874	35.65	-1.705	4.418	35.97	-6.356	2.070	35.46
$\widehat{\text{var}}_{\text{Fix}}$	0.989	1.074	36.12	0.105	6.844	36.97	0.300	5.624	38.65
$\widehat{\text{var}}_{\text{R}}$	0.921	1.044	36.06	-0.019	6.619	36.85	0.052	5.359	38.56
$\widehat{\text{var}}_{\text{Dev3}}$	0.930	1.048	36.08	0.015	6.679	36.89	0.221	5.532	38.91
$\widehat{\text{var}}_1$	0.878	1.026	36.04	-0.096	6.486	36.77	-0.088	5.219	38.43
$\widehat{\text{var}}_{\text{Ber}}$	0.993	1.077	36.12	0.091	6.817	36.86	0.215	5.537	38.53
$\widehat{\text{var}}_{\text{T}}$	0.994	1.078	36.12	0.100	6.831	36.91	0.282	5.598	38.91
$\widehat{\text{var}}_2$	0.929	1.048	36.07	0.014	6.676	36.89	0.216	5.527	38.90
$\widehat{\text{var}}_3$	0.920	1.044	36.06	-0.022	6.615	36.84	0.051	5.359	38.52
$\widehat{\text{var}}_4$	0.905	1.027	36.03	-0.055	6.498	36.78	-0.029	5.230	38.44
$\widehat{\text{var}}_5$	0.862	1.019	35.97	-0.122	6.442	36.75	-0.124	5.184	38.43
$\widehat{\text{var}}_{\text{Br1}}$	0.869	1.022	35.99	-0.089	6.497	36.81	0.036	5.339	38.70
$\widehat{\text{var}}_{\text{Br2}}$	0.927	1.047	36.06	0.007	6.665	36.89	0.173	5.481	38.84
$\widehat{\text{var}}_{\text{Br3}}$	0.984	1.072	36.13	0.101	6.835	36.99	0.306	5.625	39.01
$\widehat{\text{var}}_{\text{Br4}}$	0.990	1.075	36.13	0.106	6.844	36.99	0.310	5.629	39.01

Table 4.5: Number of times that $\widehat{\text{var}}_{\text{HT}} < 0$ among 10000 simulated samples

	n=10	n=20	n=40
mu284 population	2312	2708	2450
artificial population 1	61	65	278
artificial population 2	0	0	0

MSE for $\widehat{\text{var}}_{\text{HT}}$ in population mu284 and artificial population 1.

In what concerns the Sen-Yates-Grundy estimator compared to the rest of the estimators (without taking into account $\widehat{\text{var}}_{\text{HT}}$, $\widehat{\text{var}}_{\text{repl}}$ and $\widehat{\text{var}}_{\text{naive}}$), which use only the first-order inclusion probabilities, we see that no big differences in the variance estimation are revealed from the simulations. However, for the second case and in Table 4.2, $\widehat{\text{var}}_{\text{SYG}}$ does not perform better than the other estimators. From the above and if we take seriously into account the fact that Sen-Yates-Grundy estimator uses both first and second-order inclusion probabilities (which makes it harder to compute), we find no reason why it should be preferred to the other estimators.

Concerning the bias, the unbiased Horvitz-Thompson and Sen-Yates-Grundy estimators show non-zero bias due to the measurement error contingent on the finite size of simulations.

The estimator with replacement and the naive estimator are highly biased in the first two populations and overestimate the variance. Therefore, the coverage rates are very good in the case of these populations. In the third population, $\widehat{\text{var}}_{\text{repl}}$ and $\widehat{\text{var}}_{\text{naive}}$ perform better than all the other estimators concerning the RB and MSE, but we must take into account that this population is badly adapted to a real case.

The estimators (different from $\widehat{\text{var}}_{\text{HT}}$, $\widehat{\text{var}}_{\text{SYG}}$, $\widehat{\text{var}}_{\text{repl}}$, $\widehat{\text{var}}_{\text{naive}}$) which use only the first-order inclusion probabilities have similar performances and deserve consideration as practical alternatives. However, in the first population study, which is a real case, for $n = 40$, the estimators $\widehat{\text{var}}_{Dev3}$, $\widehat{\text{var}}_{T}$,$\widehat{\text{var}}_{2}$, $\widehat{\text{var}}_{Br1}$, $\widehat{\text{var}}_{Br2}$, $\widehat{\text{var}}_{Br3}$ and $\widehat{\text{var}}_{Br4}$ get highly biased.

Concerning the coverage rate, $\widehat{\text{var}}_{\text{HT}}$ gives poor coverage rates, compared to all the other estimators in the first two populations. Its coverage percentages range from 68.75% to 72.69% in mu284 population and from 89.76% to 93.71% in artificial population 1. The estimator $\widehat{\text{var}}_{\text{SYG}}$ gives better coverage percentages, and lies closer to the other estimators (without taking into account $\widehat{\text{var}}_{\text{repl}}$ and $\widehat{\text{var}}_{\text{naive}}$). In the first two populations the coverage rate is close to the nominal 95% for all the presented estimators (without $\widehat{\text{var}}_{\text{HT}}$, $\widehat{\text{var}}_{\text{repl}}$, $\widehat{\text{var}}_{\text{naive}}$). In the same populations, the estimators $\widehat{\text{var}}_{\text{repl}}$ and $\widehat{\text{var}}_{\text{naive}}$ give very rich coverage rates with coverage percentages ranging more than the nominal 95%.

The artificial population 2 is a special case: all the presented estimators give very poor coverage rates from 35% to 40%. This is due to the fact that "the exactness of the normal approximation used in computation of the 95% confidence interval depends significantly on the shape of the finite population" and

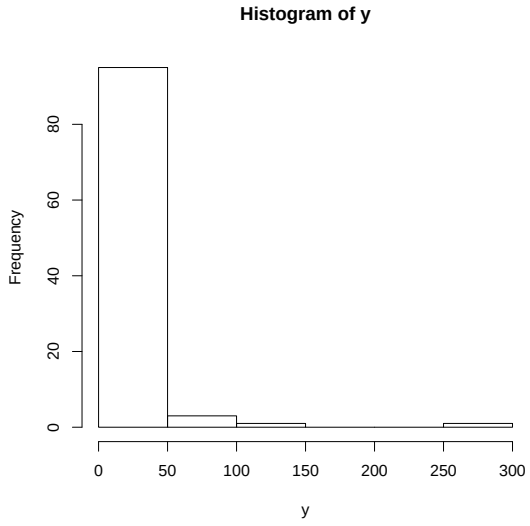


Figure 4.3: The second artificial population, n=10.

"we can expect the approach to normality of the variable $(\widehat{Y}_\pi - Y)/\sqrt{\widehat{var}(\widehat{Y}_\pi)}$ to be slower" in the case of highly skewed population, or with outlying values or other abnormal features (see Remark 2.11.2 Särndal et al., 1992, p.57). Figure 4.3 gives the histogram for $\mathbf{y}$. The artificial population 2 is highly skewed (for example for $n = 10$, $\gamma_1 = 6.019$) and has 12 outlying observations. We have deleted these 12 observations and we have rerun the simulations. Even if the nominal 95% was not reached, the CR were highly improved for all the presented estimators (for $n = 10, CR \approx 72\%$, for $n = 20, CR \approx 79\%$, and for $n = 40, CR \approx 83\%$).

4.7 Conclusions

Using the method of Chen et al. (1994) and Deville (2000b), the joint inclusion probabilities can be computed exactly for a maximum entropy sampling design with fixed sample size and unequal probabilities. The joint inclusion probabilities are used in the formulae of two variance estimators, the Horvitz-Thompson and the Sen-Yates-Grundy. An empirical study demonstrates that inferiority of $\widehat{\text{var}}_{\text{HT}}$ is restricted to populations having high correlation between

the variable of interest y_k and the auxiliary variable x_k, and where $\widehat{\text{var}}_{\text{HT}} < 0$. Apart from these populations, $\widehat{\text{var}}_{\text{HT}}$ performs nearly the same as $\widehat{\text{var}}_{\text{SYG}}$. In the same case, these two estimators have a similar comportment as the estimators which use only the first-order inclusion probabilities (except $\widehat{\text{var}}_{\text{repl}}$ and $\widehat{\text{var}}_{\text{naive}}$). Under simulations, the estimators which use only the first-order inclusion probabilities (different from $\widehat{\text{var}}_{\text{repl}}$, $\widehat{\text{var}}_{\text{naive}}$ which overestimate the variance) have similar performances, regardless of correlation between $\mathbf{y}$ and $\mathbf{x}$. The use of first-order inclusion probabilities over the whole population and joint inclusion probabilities does not lead to more accurate variance estimators in the case of a maximum entropy sampling design with unequal probability and fixed sample size. So, we recommend the use of a simple estimator such as Deville estimator 1, and, in the approximation class, the fixed-point approximation.

Appendix 1: Proof of the Result 1

If we note $C(\boldsymbol{\lambda}, \mathcal{S}_n(U)) = \sum_{s \in \mathcal{S}_n(U)} \exp \boldsymbol{\lambda}'\mathbf{s}$, then

$$\begin{aligned}
\pi_k(\boldsymbol{\lambda}, n) &= \frac{\sum_{s \in \mathcal{S}_n(U)} s_k \exp \boldsymbol{\lambda}'\mathbf{s}}{C(\boldsymbol{\lambda}, \mathcal{S}_n(U))} \\
&= \frac{\exp \lambda_k}{C(\boldsymbol{\lambda}, \mathcal{S}_n(U))} \sum_{s \in \mathcal{S}_{n-1}(U \setminus \{k\})} \exp \boldsymbol{\lambda}'\mathbf{s} \\
&= \frac{\exp \lambda_k}{C(\boldsymbol{\lambda}, \mathcal{S}_n(U))} \left(\sum_{s \in \mathcal{S}_{n-1}(U)} \exp \boldsymbol{\lambda}'\mathbf{s} - \sum_{s \in \mathcal{S}_{n-1}(U)} s_k \exp \boldsymbol{\lambda}'\mathbf{s} \right) \\
&= \frac{\exp \lambda_k C(\boldsymbol{\lambda}, \mathcal{S}_{n-1}(U))}{C(\boldsymbol{\lambda}, \mathcal{S}_n(U))} (1 - \pi_k(\boldsymbol{\lambda}, n-1)).
\end{aligned}$$

Since $\sum_{k \in U} \pi_k(\boldsymbol{\lambda}, n) = n$, we get finally

$$\pi_k(\boldsymbol{\lambda}, n) = n \frac{\exp \lambda_k \{1 - \pi_k(\boldsymbol{\lambda}, n-1)\}}{\sum_{\ell \in U} \exp \lambda_\ell \{1 - \pi_\ell(\boldsymbol{\lambda}, n-1)\}}.$$

Appendix 2: Proof of the Result 2

If we note $C(\boldsymbol{\lambda}, \mathcal{S}_n(U)) = \sum_{s\in\mathcal{S}_n(U)} \exp \boldsymbol{\lambda}'\mathbf{s}$, then

$$
\begin{aligned}
&\pi_{k\ell}(\boldsymbol{\lambda}, n) \\
&= \sum_{s\in\mathcal{S}_n(U)} s_k s_\ell p(s) \\
&= \frac{\sum_{s\in\mathcal{S}_n(U)} s_k s_\ell \exp \boldsymbol{\lambda}'\mathbf{s}}{C(\boldsymbol{\lambda}, \mathcal{S}_n(U))} \\
&= \sum_{\substack{s\in\mathcal{S}_n(U) \\ k,\ell\in s}} \frac{\prod_{j\in S} \exp \lambda_j}{C(\boldsymbol{\lambda}, \mathcal{S}_n(U))} \\
&= \frac{\exp\lambda_k \exp\lambda_\ell \sum_{\substack{s\in\mathcal{S}_{n-2}(U) \\ k,\ell\notin s}} \prod_{j\in s} \exp\lambda_j}{C(\boldsymbol{\lambda}, \mathcal{S}_n(U)} \\
&= \exp\lambda_k \exp\lambda_\ell \mathrm{Pr}\{k,\ell\notin s | s\in\mathcal{S}_{n-2}\} \frac{C(\boldsymbol{\lambda}, \mathcal{S}_{n-2}(U))}{C(\boldsymbol{\lambda}, \mathcal{S}_n(U)} \\
&= \exp\lambda_k \exp\lambda_\ell (1 - \pi_k(\boldsymbol{\lambda}, n-2) - \pi_\ell(\boldsymbol{\lambda}, n-2) + \pi_{k\ell}(\boldsymbol{\lambda}, n-2)) \frac{C(\boldsymbol{\lambda}, \mathcal{S}_{n-2}(U))}{C(\boldsymbol{\lambda}, \mathcal{S}_n(U)}.
\end{aligned}
$$

Since $\sum_{k\in U} \sum_{\substack{k,\ell\in U \\ k\neq\ell}} \pi_{k\ell}(\boldsymbol{\lambda}, n) = n(n-1)$, we get finally

$$
\pi_{k\ell}(\boldsymbol{\lambda}, n) = \frac{n(n-1)\exp\lambda_k \exp\lambda_\ell (1 - \pi_k(\boldsymbol{\lambda}, n-2) - \pi_\ell(\boldsymbol{\lambda}, n-2) + \pi_{k\ell}(\boldsymbol{\lambda}, n-2))}{\sum_{i\in U} \sum_{\substack{j\in U \\ j\neq i}} \exp\lambda_i \exp\lambda_j (1 - \pi_i(\boldsymbol{\lambda}, n-2) - \pi_j(\boldsymbol{\lambda}, n-2) + \pi_{ij}(\boldsymbol{\lambda}, n-2))},
$$

$k, \ell \in U, k \neq \ell$.

Appendix 3: Justification of the Algorithm 4.2.5

Suppose that $\sum_{k\in U} \lambda_k = 0$, in order to have a unique definition of $\boldsymbol{\lambda}$. Indeed,

$$
p(s, \mathcal{S}_n(U), \boldsymbol{\lambda}) = p(s, \mathcal{S}_n(U), \boldsymbol{\lambda}^*), \text{ for all } s \in \mathcal{S}_n(U)
$$

when $\lambda_k^* = \lambda_k + c$ for any $c \in \mathbb{R}$. The inclusion probability vector can be written as a function of $\boldsymbol{\lambda}$ and n

$$
\boldsymbol{\pi}(\boldsymbol{\lambda}, n) = \sum_{s\in\mathcal{S}_n(U)} p(s, \mathcal{S}_n(U), \boldsymbol{\lambda}).
$$

Since $\sum_{k\in U}\pi_k = n$, $\boldsymbol{\pi}(\boldsymbol{\lambda}, n)$ is a one to one application from

$$\left\{\boldsymbol{\lambda}\in\mathbb{R}^N \left| \sum_{k\in U}\lambda_k = 0\right.\right\}$$

to

$$\left\{\boldsymbol{\pi}\in]0,1[^N \left| \sum_{k\in U}\pi_k = n\right.\right\}.$$

Define $\boldsymbol{\pi}(\boldsymbol{\lambda}, n)$ as a function of $\widetilde{\boldsymbol{\pi}}$, that will be denoted $\boldsymbol{\phi}(\widetilde{\boldsymbol{\pi}}, n)$, and

$$\boldsymbol{\phi}(\widetilde{\boldsymbol{\pi}}, n) = \boldsymbol{\pi}(\boldsymbol{\lambda}, n) = \frac{\sum_{s\in\mathcal{S}_n(U)}\mathbf{s}\exp\boldsymbol{\lambda}'\mathbf{s}}{\sum_{s\in\mathcal{S}_n(U)}\exp\boldsymbol{\lambda}'\mathbf{s}} = \frac{\sum_{s\in\mathcal{S}_n(U)}\mathbf{s}\prod_{k\in s}\frac{\widetilde{\pi}_k}{1-\widetilde{\pi}_k}}{\sum_{s\in\mathcal{S}_n(U)}\prod_{k\in s}\frac{\widetilde{\pi}_k}{1-\widetilde{\pi}_k}}.$$

Since $\widetilde{\boldsymbol{\pi}}$ can be derived from $\boldsymbol{\lambda}$ and and vice versa (see Result 1), $\boldsymbol{\phi}(\widetilde{\boldsymbol{\pi}}, n)$ can be computed recursively by means of expression (4.3)

$$\boldsymbol{\phi}(\widetilde{\boldsymbol{\pi}}, n) = n\frac{\frac{\widetilde{\pi}_k}{1-\widetilde{\pi}_k}\left\{1-\phi_k(\widetilde{\boldsymbol{\pi}}, n-1)\right\}}{\sum_{\ell\in U}\frac{\widetilde{\pi}_\ell}{1-\widetilde{\pi}_\ell}\left\{1-\phi_\ell(\widetilde{\boldsymbol{\pi}}, n-1)\right\}}.$$

If the vector of inclusion probabilities $\boldsymbol{\pi}$ (such that $\sum_{k\in U}\pi_k = n$) is given, Chen et al. (1994) have proposed to solve the equation

$$\boldsymbol{\phi}(\widetilde{\boldsymbol{\pi}}, n) = \boldsymbol{\pi},$$

in $\widetilde{\boldsymbol{\pi}}$ by the Newton method, which gives the algorithm

$$\widetilde{\boldsymbol{\pi}}^{(i)} = \widetilde{\boldsymbol{\pi}}^{(i-1)} + \left|\frac{\partial\boldsymbol{\phi}(\widetilde{\boldsymbol{\pi}}, n)}{\partial\widetilde{\boldsymbol{\pi}}}\right|^{-1}_{\widetilde{\boldsymbol{\pi}}=\widetilde{\boldsymbol{\pi}}^{(i-1)}}\left(\boldsymbol{\pi}-\boldsymbol{\phi}(\widetilde{\boldsymbol{\pi}}^{(i-1)}, n)\right),$$

where $i = 1, 2, \ldots$ and with $\widetilde{\boldsymbol{\pi}}^{(0)} = \boldsymbol{\pi}$. Unfortunately, the matrix

$$\left|\frac{\partial\boldsymbol{\phi}(\widetilde{\boldsymbol{\pi}}, n)}{\partial\widetilde{\boldsymbol{\pi}}}\right|_{\widetilde{\boldsymbol{\pi}}=\widetilde{\boldsymbol{\pi}}^{(i-1)}} \tag{4.33}$$

is not easy to compute. However, Deville (2000b) pointed out that the matrix (4.33) is very close to the identity matrix, which allows simplifying significantly the algorithm. Finally we can use

$$\widetilde{\boldsymbol{\pi}}^{(i)} = \widetilde{\boldsymbol{\pi}}^{(i-1)} + \boldsymbol{\pi} - \boldsymbol{\phi}(\widetilde{\boldsymbol{\pi}}^{(i-1)}, n), \tag{4.34}$$

which allows to pass quite quickly from $\boldsymbol{\pi}$ to $\widetilde{\boldsymbol{\pi}}$ and thus to $\boldsymbol{\lambda}$. The number of operations needed to compute $\widetilde{\boldsymbol{\pi}}$ is $O(N^2\times n\times$ number of iterations i).

Chapter 5

A variant of the Cox algorithm for the imputation of non-response of qualitative data

Abstract
The Cox algorithm allows to round randomly and unbiasedly a table of real numbers without modifying the marginal totals. One possible use of this method is the random imputation of a qualitative variable in survey sampling. A modification of the Cox algorithm is proposed in order to take into account a weighting system, which is commonly used in survey sampling. The use of this new method allows to construct a controlled imputation method that reduces the imputation variance.
Keywords: Cox algorithm, non-response, imputation, qualitative data, controlled imputation, survey.

5.1 Introduction

In surveys practice, a certain level of non-response frequently occurs. Essentially, two types of non-response can be distinguished: total or unit non-response and partial or item non-response. The use of weighting adjustment is recommended to compensate the unit non-response, while imputation is used to deal with item non-response. Imputation is a process that allows to replace a missing value by an artificial one. There are two commonly used types of imputation methods: deterministic and random techniques. In the first case, the distribution of the imputed variable is distorted, and the variance is attenuated, while in stochastic imputation procedures, the addition of an estimated residual avoids the distortion of the distribution and the attenuation of variance. For this reason, stochastic imputation procedures are generally preferred. However, these procedures due to their random nature increase the imputation variance, and consequently inflate the variance of survey estimates.

The goal is to develop a new imputation method for item non-response of qualitative data. This method is based on the Cox algorithm (see Cox, 1987), which allows to round randomly the cells of a contingency table without modifying the marginal totals. Deville and Tillé (2000) have shown that the Cox algorithm can be used to select several unequal probability samples in the same population. In this paper, a variant of the Cox method is presented in order to provide a new imputation procedure. This new method merits the advantages of both deterministic and random procedures.

The usual Poisson sampling, and the proposed method described above are tested and compared by simulations. Results, presented in Section 5.5, show that the new method reduces the variance of the imputed totals in comparison to usual Poisson sampling. The paper mainly deals with the modified Cox algorithm and is conducted as follows: Section 5.2 outlines the imputation problem of qualitative variables. Section 5.3 describes the Cox algorithm. Section 5.4 presents the new algorithm (named “Cox weighted algorithm”). In Section 5.5, the variance of this new method is compared to the variance arising from Poisson sampling. Section 5.6 discusses the results of the paper.

5.2 The problem

Consider a finite population U of size N. The interest variable y is qualitative and has v possible exclusive values. The codification is the following

$$y_{kj} = \begin{cases} 1 & \text{if unit } k \text{ takes the value } j, \\ 0 & \text{if not,} \end{cases}$$

for all $k \in U$ and $j = 1, \ldots, v$, with

$$\sum_{j=1}^{v} y_{kj} = 1,$$

for all k in U. The aim is to estimate the totals for each category

$$Y_j = \sum_{k \in U} y_{kj},$$

for $j = 1, \ldots, v$, that have the property

$$\sum_{j=1}^{v} Y_j = N.$$

Suppose that a random sample (or subset) S has been selected in U with inclusion probabilities $\pi_k = Pr(k \in S), k \in U$. Without non-response, the total can be estimated by

$$\widehat{Y}_j = \sum_{k \in S} w_k y_{kj},$$

where the w_k's are weights which are equal to $w_k = 1/\pi_k$ (Horvitz and Thompson, 1952) or can be more complex expressions of a calibrated estimator (see Deville and Särndal, 1992).

Suppose that some of the values of the variables y are missing for a subset $S \backslash R$ of non-respondent units of S, but that J auxiliary variables $x_1, ..., x_J$ are available on the whole set S. Let $\mathbf{x}_k = (x_{k1}, x_{k2}, \ldots, x_{kJ})$ be the vector of the values taken by the x-variables on unit k. Suppose however that a model can be constructed to predict the values of variable y. This model can be written as

$$p_{kj} = Pr(y_{kj} = 1 | \mathbf{x}_k), \text{ for all } k \in S \backslash R.$$

For instance, a multivariate logistic model can be used:

$$Pr(y_{kj} = 1|\mathbf{x}_k) = \frac{\exp(\boldsymbol{\beta}_j'\mathbf{x}_k)}{\sum_{\ell=1}^{J} \exp(\boldsymbol{\beta}_\ell'\mathbf{x}_k)},$$

for all $j = 1 \ldots, v$. The aim here is not to discuss the model. We only suppose that the p_{kj} can be estimated and respect the following property:

$$\sum_{j=1}^{v} p_{kj} = 1, \text{ for all } k \in U.$$

An interesting estimator of Y_j could be

$$\widehat{Y}_j^* = \sum_{k \in R} w_k y_{kj} + \sum_{k \in S \setminus R} w_k p_{kj}.$$

It is noticeable that estimator $\widehat{Y}_j^*$ has no error arising from random imputation. Nevertheless, it could be interesting to provide realistic imputed values for more complex estimation problems. The new developed method, which is based on Cox algorithm, imputes the missing values, and the weights w_k's are taken into account.

5.3 The Cox algorithm

The Cox algorithm provides an unbiased controlled rounding of a matrix. It is based on alternating paths in a matrix; a path is a circular sequence of cells with non-integer values. Cox (1987) uses the following definition of a path:

Definition 5.3.3 *An alternating row-column path in a matrix $\mathbf{A}_{m \times v}$ is a sequence of distinct indexes $(i_1, j_1), (i_2, j_2), ..., (i_l, j_l)$ satisfying $1 \leq i_s \leq m, 1 \leq j_s \leq v, (s = 1, ..., l); j_{s+1} = j_s$ and $i_{s+1} \neq i_s$ if s is even; and $i_{s+1} = i_s$ and $j_{s+1} \neq j_s$ is s is odd. The path begins in row i_1 and alternates horizontal and vertical steps. Reversing the roles of "even" and "odd", an alternating column-row path is obtained, which begins in column j_1 and alternates vertical and horizontal steps. The alternating paths are cycles, that is, $j_l = j_1$ for a row-column path or $i_l = i_1$ for a column-row path.*

The algorithm is as follows :

- *Step 1 :* If the elements of the matrix are all equal to 0 or 1, terminate the procedure.

- *Step 2 :* Choose any non-integer value $a_{i_1 j_1}$ in the matrix. At (i_1, j_1) begin an alternating row-column (or column-row) path of non-integer values :

$$(i_1, j_1), (i_1, j_2), (i_2, j_2), (i_2, j_3), ..., (i_l, j_{l+1}) = (i_l, j_1).$$

  Let :

$$d_- = \min_{1 \leq q \leq l} (a_{i_q j_q}, 1 - a_{i_q j_{q+1}}),$$
$$d_+ = \min_{1 \leq q \leq l} (1 - a_{i_q j_q}, a_{i_q j_{q+1}}).$$

  Both d_- and d_+ consist of values strictly between 0 and 1.

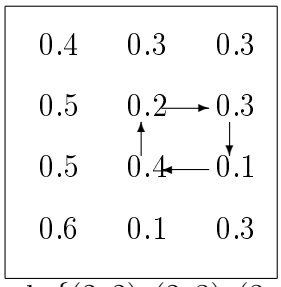
- *Step 3 :* Select either d_- or d_+ randomly with a given probability (p_- or p_+, respectively).

- *Step 4 :* If d_- is selected :
  - transform $a_{i_q j_q}$ to $a_{i_q j_q} - d_-$ along the path.
  - transform $a_{i_q j_{q+1}}$ to $a_{i_q j_{q+1}} + d_-$ along the path.

  If d_+ is selected :
  - transform $a_{i_q j_q}$ to $a_{i_q j_q} + d_+$ along the path.
  - transform $a_{i_q j_{q+1}}$ to $a_{i_q j_{q+1}} - d_+$ along the path.

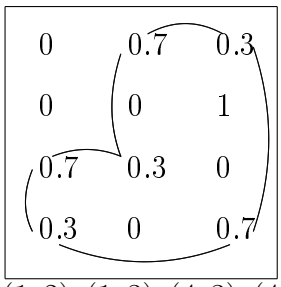
  Return to *Step 1.*
  End.

Figures 5.1 and 5.2 show examples of paths. In Figure 5.1 the path is simple and rectangular, while in Figure 5.2 the path is more complex. A possible random modification of the cells of the path in Figure 5.1 is illustrated in Figure 5.3.



Path $\{(2,2),(2,3),(3,3),(3,2)\}$

Figure 5.1: Example of a simple path.



Path $\{(1,2),(1,3),(4,3),(4,1),(3,1),(3,2)\}$

Figure 5.2: Example of a complex path.

At each iteration, at least one non-integer value is transformed to an integer, whereas all current integers remain fixed. Every non-integer value must appear in one or more iterations. If d_- is selected at step 3, then the value of d_- is subtracted from $a_{i_1 j_1}$; if d_+ is selected at step 3, then the value of d_+ is added to $a_{i_1 j_1}$. The probabilities in order to obtain an unbiased method of selection are respectively:

$$p_- = \frac{d_+}{d_- + d_+} \;\; ; \;\; p_+ = \frac{d_-}{d_- + d_+}.$$

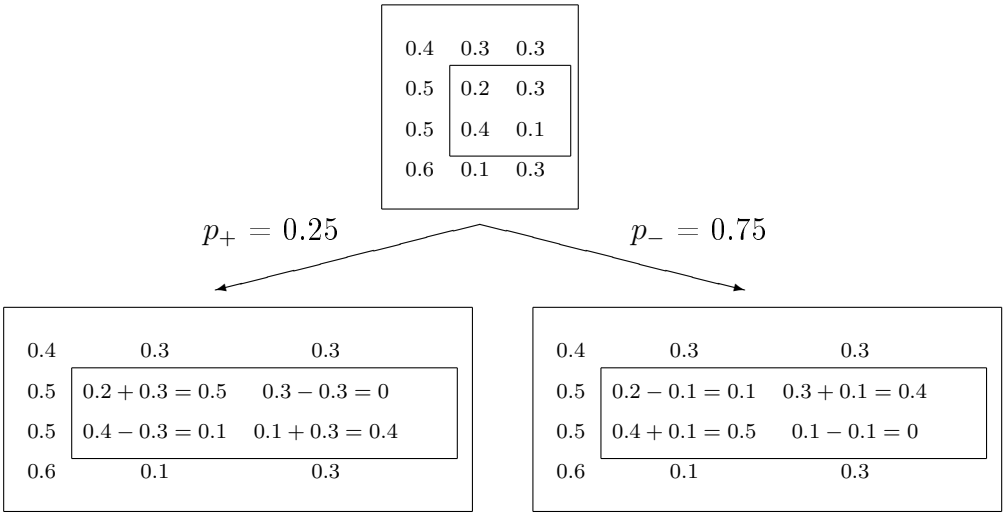


Figure 5.3: Example of an iteration modifying the cells in Figure 5.1.

5.4 The Cox weighted algorithm

The Cox algorithm cannot be used directly to realize the imputation, because of the weights w_k. Those weights are such that the values

$$\sum_{k \in S \backslash R} w_k p_{kj}$$

are not integer. Let $\mathbf{P} = (p_{ij})_{m \times v}$ be the matrix of probabilities given in Section 5.2 and let $\mathbf{w} = (w_k)$ be the vector of weights. The important property that the sum of each column must be integer is no more satisfied. It is thus impossible to find imputed values $\dot{y}_{kj}$ such that

$$\sum_{k \in S \backslash R} w_k p_{kj} = \sum_{k \in S \backslash R} w_k \dot{y}_{kj}. \tag{5.1}$$

For this reason, a variant of the Cox method must be developed. First define matrix $\mathbf{A} = (a_{ij})_{(m \times v)}$ by $a_{ij} = p_{ij} \times w_i$, for all $i \in \{1, \ldots, m\}$ and for all $j \in \{1, \ldots, v\}$ where m is the cardinal of $S \backslash R$.

The new algorithm must thus take into account the weighting system and the impossibility to satisfy exactly expression (5.1). The new procedure is decomposed into two phases. In the first phase, the new algorithm is applied on matrix $\mathbf{A}$ until no more path exists in the matrix. The obtained matrix is denoted $\mathbf{A}^*$. In the second phase, a simple Poisson imputation is applied. At the end of the first phase, each row i of matrix $\mathbf{A}^*$ is divided by the weights w_i in order to obtain a matrix denoted $\mathbf{P}^* = (p^*_{ij})_{(m\times v)}$ with elements $\{0, 1\}$ and some non-integer values that cannot be linked by a path. In matrix $\mathbf{P}^*$, there exists no more path, and the following conditions are fulfilled:

$$\sum_{j=1}^{v} p^*_{ij} = 1, \text{ for all } i \in \{1, \ldots, m\},$$

and

$$\sum_{i=1}^{m} p^*_{ij} w_i = \sum_{i=1}^{m} p_{ij} w_i, \text{ for all } j \in \{1, \ldots, v\}.$$

In each step of the first phase, the existence of a path is firstly controlled. Indeed, since the marginal totals of matrix $\mathbf{A}$ are no more integer, a path can cross a cell only if it exists at least an other non-integer value on its line and on its column. In order to manage this problem, a control matrix is used $\mathbf{C} = (c_{ij})_{(m\times v)}$, indicating the stop criterion.
At the beginning, $\mathbf{C}$ is defined as follows:

$$c_{ij} = \begin{cases} 0 & \text{if } a_{ij} = 0 \text{ or } a_{ij} = w_i, \\ 1 & \text{if not.} \end{cases}$$

Next, when there is only one $c_{ij} = 1$ in a line i or in a column j, it is shifted to 0. The paths must cross only cells such that $c_{ij} = 1$. At each iteration, at least one or more non-integer values a_{ij} in $\mathbf{A}$ is transformed to 0 or w_i, where the position (i, j) is a position of the current path. Every non-integer value must appear in one or more iterations. When a non-integer value a_{ij} is transformed in 0 or w_i, we make also the transformation $c_{ij} = 0$, to indicate the invalidity of a_{ij} for a new path.

When all the elements of the matrix $\mathbf{C}$ are equal to 0, the algorithm stops, indicating that there is no more path in matrix $\mathbf{A}$. Nevertheless, some cells of matrix $\mathbf{P}^*$ can be non-integer. If there is a non-integer value on a line i, there is automatically another non-integer value on the same line, because $\sum_{j=1}^{v} p^*_{ij} = 1$. The maximum number of possible lines with no rounded values

is $v-1$, otherwise a path can be found. Consequently, the maximum possible total number of non-rounded values in the final matrix is $2 \times (v-1)$.

The new procedure is defined as follows:

- *Step 1*:
$$c_{ij} = \begin{cases} 0 & \text{if } a_{ij} = 0 \text{ or } a_{ij} = w_i, \\ 1 & \text{if not.} \end{cases}$$

- *Step 2*: If all the elements of the matrix $\mathbf{C}$ are equal to 0, go to *Step 5*. Else make the rounding of the non-integer values of matrix $\mathbf{A}$ on the rows, determining the first non-integer value on the rows at the position (i_1, j_1), which gives the first position in our path, an alternating row-column or column-row path of non-integer values. Choose randomly at each iteration the type of transformation (plus or minus):
$$d_- = \min_{1 \le q \le l} (a_{i_q j_q}, w_{i_q} - a_{i_q j_{q+1}}), \quad d_+ = \min_{1 \le q \le l} (w_{i_q} - a_{i_q j_q}, a_{i_q j_{q+1}}),$$
where the current path is : $(i_1, j_1), (i_1, j_2), ..., (i_l, j_{l+1}) = (i_l, j_1)$.
If d_- is selected :
- transform $a_{i_q j_q}$ to $a_{i_q j_q} - d_-$ along the path,
- transform $a_{i_q j_{q+1}}$ to $a_{i_q j_{q+1}} + d_-$ along the path.
If d_+ is selected :
- transform $a_{i_q j_q}$ to $a_{i_q j_q} + d_+$ along the path,
- transform $a_{i_q j_{q+1}}$ to $a_{i_q j_{q+1}} - d_+$ along the path.
Transform also the matrix $\mathbf{C}$ into:
$$c_{ij} = \begin{cases} 0 & \text{if } a_{ij} = 0 \text{ or } a_{ij} = w_i, \\ 1 & \text{if not.} \end{cases}$$
Moreover, if there is only one $c_{ij} = 1$ in a line i or in a column j, it is shifted to 0.

- *Step 3*: Return to *Step 2*.

- *Step 4*: Make a division of each row i of matrix $\mathbf{A}$ with w_i.

- *Step 5*: If some cells cannot be rounded, apply a Poisson sampling equalizing all the elements of the matrix to 0 or 1 .
End.

In the previous algorithm, the choice of the paths is deterministic. However, we advocate for a random choice of the paths, which can be done easily in sorting randomly the rows and the columns of the matrix, before applying the procedure.

5.5 Variance

In this section, the variance of column's total ($\widehat{T}_j = \sum_{i \in S \setminus R} w_i p_{ij}$) are compared for usual Poisson sampling and Cox weighted algorithm. In the second case, the variance is only due to the last step of the procedure, when a Poisson procedure is applied on $\mathbf{P}^*$ to round the remaining cells.

The variance of the estimator can be decomposed into 4 components:

1. sampling variance,
2. variance due to non-response,
3. variance due to estimation of the p_{ij},
4. variance due to imputation.

In simulations, we are only interested in the evaluation of the reduction of the 4th component of the variance (due to imputation).

As stated in Section 5.4, the Cox weighted algorithm can be applied until, it exists no more possible paths. At this stage, a maximum of $2(v-1)$ values are not rounded, i.e. it means that on each row, $2(v-1)/v$ cells are on average not rounded. The variance of $\widehat{T}_j$ depends only on these $2(v-1)/v$ not rounded values of matrix $\mathbf{P}^*$, that will be rounded by a Poisson sampling and will form the final matrix noted as $\mathbf{P}_{(end)}$. Thus the variance of an imputed value using Cox algorithm is derived from the Bernoulli distribution:

$$p_{(\text{end})ij}(1 - p_{(\text{end})ij}) \leq \frac{1}{2}\Big(1 - \frac{1}{2}\Big) = \frac{1}{4}.$$

In the worst case (probability = 1/2), a rough approximation of the variance of the imputation of $\widehat{T}_j$ could thus be

$$\text{var}_{\text{cox}}(\widehat{T}_j) \approx \frac{2(v-1)}{v}\frac{1}{4}\sum_i \frac{w_i^2}{m}.$$

If the Poisson algorithm instead of the Cox algorithm is applied on $\mathbf{P}$, the variance of an imputed value is a sum of Bernoulli trials. If we take the approximation $p_{ij} = 1/v$, the variance is roughly equal to

$$\sum_i w_i^2 p_{ij}(1 - p_{ij}) \approx \sum_i w_i^2 \frac{1}{v}\Big(1 - \frac{1}{v}\Big).$$

Thus the variance could be roughly approximated by

$$\mathrm{var}_{\mathrm{poiss}}(\widehat{T}_j) \approx \frac{v-1}{v^2} \sum_i w_i^2.$$

The reduction variance factor should therefore be less than

$$\alpha_j = \frac{\mathrm{var}_{\mathrm{cox}}(\widehat{T}_j)}{\mathrm{var}_{\mathrm{poiss}}(\widehat{T}_j)} \approx \frac{\frac{1}{m}\frac{2(v-1)}{v}\frac{1}{4}\sum_i w_i^2}{\frac{1}{v}(1-\frac{1}{v})\sum_i w_i^2} = \frac{v}{2m}.$$

A set of simulations has been realized to confirm this rough approximation. For different matrix sizes $(5 \times 4, 10 \times 4, 20 \times 4, 40 \times 6)$ we have applied the weighted Cox method and the Poisson algorithm. For each table 10'000 simulations were implemented. Three different weight systems have been used:

- constant weights equal to one;
- uniformly distributed weights: $w \sim \mathcal{U}[1,2]$;
- random weights, function of a uniform random variable $w = \frac{1}{\beta U + (1-\beta)}$ where $U \sim \mathcal{U}[0,1]$ and $\beta = 0.2$.

The probabilities p_{ij} are generated using uniform distribution and are rescaled on 1.

The variance is estimated under the simulations by

$$\widehat{\mathrm{var}}(\widehat{T}_j) = \frac{1}{\mathrm{nbsim}} \sum_{s=1}^{\mathrm{nbsim}} \left(\widehat{T}_j - T_j\right)^2,$$

where *nbsim* is the number of simulations. Results are summarized in Tables 5.1 to 5.3.

Results show that a significant gain of variance is obtained with the Cox modified algorithm compared to the Poisson algorithm. The reduction factor α_j is a good approximation of the reduction variance factor when the weights are not too overspread. Moreover, simulations show that $\widehat{\alpha}_j$ gets closer to α_j, as m increases.

Table 5.1: Simulation results for the variance reduction factor; all the weights are equal to 1.

m	v	j	$\widehat{\text{var}}_{\text{cox}}(\widehat{T}_j)$	$\widehat{\text{var}}_{\text{poiss}}(\widehat{T}_j)$	α_j	$\hat{\alpha}_j$
5	4	1	0.269	0.922	0.4	0.292
		2	0.203	0.728	0.4	0.278
		3	0.260	0.958	0.4	0.271
		4	0.256	0.754	0.4	0.339
10	4	1	0.227	1.383	0.2	0.164
		2	0.245	1.638	0.2	0.150
		3	0.238	2.051	0.2	0.116
		4	0.215	1.802	0.2	0.119
20	4	1	0.289	2.797	0.1	0.103
		2	0.277	3.220	0.1	0.086
		3	0.290	4.043	0.1	0.072
		4	0.290	3.635	0.1	0.080
40	6	1	0.319	4.638	0.075	0.069
		2	0.296	5.424	0.075	0.055
		3	0.311	6.003	0.075	0.052
		4	0.306	4.930	0.075	0.062
		5	0.266	5.140	0.075	0.052
		6	0.331	5.108	0.075	0.065

5.6 Conclusion

It is thus possible to build an imputation technique, which preserves the properties of both deterministic and random methods. A synthesis of deterministic and random process succeeded in remarkably removing imputation's variance. Thus, the new method leads to gain in precision. This gain is in the order of $2v/m$ as compared to usual Poisson sampling. The method allows a calibrated random imputation, which reduces the total variance, but holds the virtue of random imputation.

Table 5.2: Simulation results for the variance reduction factor; $w \sim \mathcal{U}[1,2]$.

m	v	j	$\widehat{\mathrm{var}}_{\mathrm{cox}}(\widehat{T}_j)$	$\widehat{\mathrm{var}}_{\mathrm{poiss}}(\widehat{T}_j)$	α_j	$\hat{\alpha}_j$
5	4	1	0.969	3.027	0.4	0.320
		2	0.615	2.122	0.4	0.290
		3	0.786	2.702	0.4	0.291
		4	0.853	2.028	0.4	0.420
10	4	1	0.682	3.839	0.2	0.178
		2	0.676	4.619	0.2	0.146
		3	0.700	5.537	0.2	0.126
		4	0.656	4.911	0.2	0.134
20	4	1	0.757	6.536	0.1	0.116
		2	0.690	7.932	0.1	0.087
		3	0.760	9.537	0.1	0.080
		4	0.635	8.266	0.1	0.077
40	6	1	1.029	11.565	0.075	0.089
		2	0.750	13.066	0.075	0.057
		3	0.867	14.716	0.075	0.059
		4	0.882	12.498	0.075	0.071
		5	0.736	11.722	0.075	0.063
		6	0.761	12.434	0.075	0.061

Table 5.3: Simulation results for the variance reduction factor; $w = \frac{1}{U\beta+(1-\beta)}$, where $U \sim \mathcal{U}[0,1]$ and $\beta = 0.2$.

m	v	j	$\widehat{\text{var}}_{\text{cox}}(\widehat{T}_j)$	$\widehat{\text{var}}_{\text{poiss}}(\widehat{T}_j)$	α_j	$\hat{\alpha}_j$
5	4	1	3.154	5.538	0.4	0.569
		2	2.327	3.921	0.4	0.594
		3	2.406	5.231	0.4	0.460
		4	1.417	3.564	0.4	0.398
10	4	1	3.568	8.221	0.2	0.434
		2	4.831	10.651	0.2	0.454
		3	4.631	10.775	0.2	0.430
		4	5.062	9.964	0.2	0.508
20	4	1	2.901	17.818	0.1	0.163
		2	3.224	23.184	0.1	0.139
		3	3.194	26.425	0.1	0.121
		4	2.279	21.572	0.1	0.106
40	6	1	3.261	27.609	0.075	0.118
		2	3.887	34.316	0.075	0.113
		3	3.635	38.088	0.075	0.095
		4	3.230	29.058	0.075	0.111
		5	3.456	32.430	0.075	0.107
		6	4.391	34.265	0.075	0.128

Chapter 6

Calibrated Random Imputation for Qualitative Data

Abstract
In official statistics, when a file of microdata must be delivered to external users, it is very difficult to propose them a file where missing values has been treated by multiple imputations. In order to overcome this difficulty, we propose a method of single imputation for qualitative data that respect numerous constraints. The imputation is balanced on totals previously estimated; editing rules can be respected; the imputation is random, but the totals are not affected by an imputation variance.
Keywords: survey, qualitative variables, item non-response, imputation, calibration, Cox algorithm.

6.1 Introduction

In sample surveys two kinds of non-response essentially occur: unit non-response when the entire questionnaire is missing, and item non-response, when one or several items are missing. While weighting methods are usually used to deal with unit non-response, item non-response is handled with imputation-based procedures, where the missing values are filled in and the resultant completed data are analyzed by standard methods.

Two types of imputation methods are commonly used: deterministic and random methods. Several problems remain with imputation techniques. Indeed, random imputation implies an increase of variance, while deterministic imputation distorts the distribution of the imputed variable and attenuates the variance. Stochastic imputation procedures are generally preferred. A way to profit from the advantage of both random and deterministic imputation is to carry out multiple imputations. The distribution is then not distorted, and since the final estimator is the average of the estimators obtained with each imputation, it is not much contaminated by the imputation variance.

In official statistics, the objective is to estimate finite population quantities, such as the total frequencies per category i.e. total of single population. Additionally, the official statistical agencies are confronted with a growing demand by researchers and policymakers for access to microdata, unit record data and low level aggregates, for use in-depth studies and evidence based policy.

In order to use complete data in these analysis, it is important to develop a method which permits to fill in the missing values and to give an accurate estimation for the total frequencies per category. Yet, it is very difficult to provide the final user with multiple imputations.

We propose a new imputation method for qualitative variables when some items non-response occur. The aim of this method is to provide a unique imputation that is concordant with the total frequencies that have been estimated by another method before. The total frequencies per category can be estimated using calibration or multiple imputations. The imputation is not an estimation technique in this context. The total frequencies per category are first estimated, the missing values are then filled in order to preserve coherent principles. Moreover, in the proposed method, we take into account editing constraints i.e. the imputations preserve logical editing rules, for example not widowed with age$<$ 16 years.

The proposed method has several advantages : the variance of the estimates of the total frequencies is significantly reduced in comparison to usually

imputation techniques (see Favre et al., 2004); the final user has only a single imputation; logical editing rules are taken into account; the imputation is coherent with the totals; the distribution of the variable is not distorted by the imputation.

This paper is organized as follows. Section 6.2 outlines the problem of imputation for qualitative variable. The next sections are devoted to the different steps used in the method: editing (Section 6.3), estimation of totals (Section 6.4), individual estimation of category probabilities (Section 6.5), calibration on the marginal totals (Section 6.6), realization of the imputation based on a variant of the Cox algorithm (Cox, 1987) (Section 6.7). The above method is illustrated in an example given in Section 6.8. The last section discusses the advantages of the proposed method.

6.2 The problem

Consider a qualitative variable with some missing values for a subset of units (items non-response). A classical example is a household survey with the variable marital status made up of four categories: single, married, divorced and widowed. We assume that the non-response is missing completely at random, and that each unit has a strictly positive unknown probability of response. Moreover, we assume that the probability of response can be predicted by means of auxiliary variables. The formalization of the problem is as follows.

Consider a finite population U of size N. The interest variable y is qualitative and has v possible exclusive values. The codification is the following

$$y_{kj} = \begin{cases} 1 & \text{if unit } k \text{ takes the value } j, \\ 0 & \text{if not,} \end{cases}$$

for all $k \in U$ and $j = 1, \ldots, v$, with

$$\sum_{j=1}^{v} y_{kj} = 1,$$

for all k in U. The objective of the survey is to estimate the totals for each category

$$Y_j = \sum_{k \in U} y_{kj},$$

that have the property that

$$\sum_{j=1}^{v} Y_j = N.$$

Suppose that a random sample or subset S has been selected in U with inclusion probabilities $\pi_k = \Pr(k \in S), k \in U$. Without non-response, the total could be estimated by

$$\widehat{Y}_j = \sum_{k \in S} w_k y_{kj},$$

where the w_k's are weights that can be equal to $w_k = 1/\pi_k$ (Horvitz and Thompson, 1952) or can be more complex weights of a ratio estimator, regression estimator, raking ratio estimator or a calibrated estimator (see Deville and Särndal, 1992).

Suppose that some of the values of the variables y are missing for a subset $\bar{R}$ of non-respondent units of S. Let R be the subset of respondents of S. Suppose also that J auxiliary variables $x_1, ..., x_J$ are available on the whole set S. Let $\mathbf{x}_k = (x_{k1}, x_{k2}, \ldots, x_{kJ})^t$ be the vector of the values taken by the x-variables on unit k. The variable y can be represented with a frame, where the columns represent the categories and the rows the units. Using a reorder of the frame rows, we obtain two contiguous subsets, corresponding to the respondents R in the top and to the non-respondents units $\bar{R}$ in the bottom. Figure 6.1 shows this frame after reorder. This frame is used to illustrate the different steps of the proposed method, with bold character indicating the changes in the frame from one step to another. The proposed method can be divided into two main parts:

1. Correction by re-weighting

$$\widehat{Y}_j^{calibrate} = \sum_{k \in R_j} \frac{y_{kj} w_k}{\widehat{\Pr}[k \in R_j]} = \sum_{k \in R_j} y_{kj} \nu_k^{(j)},$$

   where R_j is the j^{th} column of 'matrix' R.

2. Imputation taking into account the weights $\nu_k^{(j)}$ obtained in step 1.

In the next paragraphs, we note $\widehat{Y}_j^{calibrate}$ by $\widehat{Y}_j$ for a good visualization.

If we consider further details, the method works in five steps: editing, estimation of totals, individual estimation of category probabilities, calibration on the marginal totals and finally realization of imputation. Figure 6.2 shows an overview of the five steps with the implied methods.

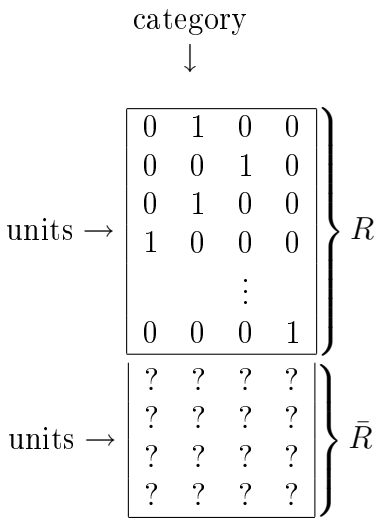


Figure 6.1: Frame representing variable y

6.3 Editing

Fellegi and Holt (1976) proposed a systematic approach of automatic editing and imputation. Edits of qualitative data expressed the judgment of some experts that certain combinations of values or code values in different fields are unacceptable. Logical rules lead to exclusion of some given categories and in some cases automatic imputation.

Example

Consider a very simple record containing two fields each with its possible set of codes given in Table 6.1. We want to impute the marital status for a person

Table 6.1: Example of possible codes

Age	Marital status
0-16	single
16+	married
	divorced
	widowed

younger than 16 years. The four possible editing rules are presented in Table

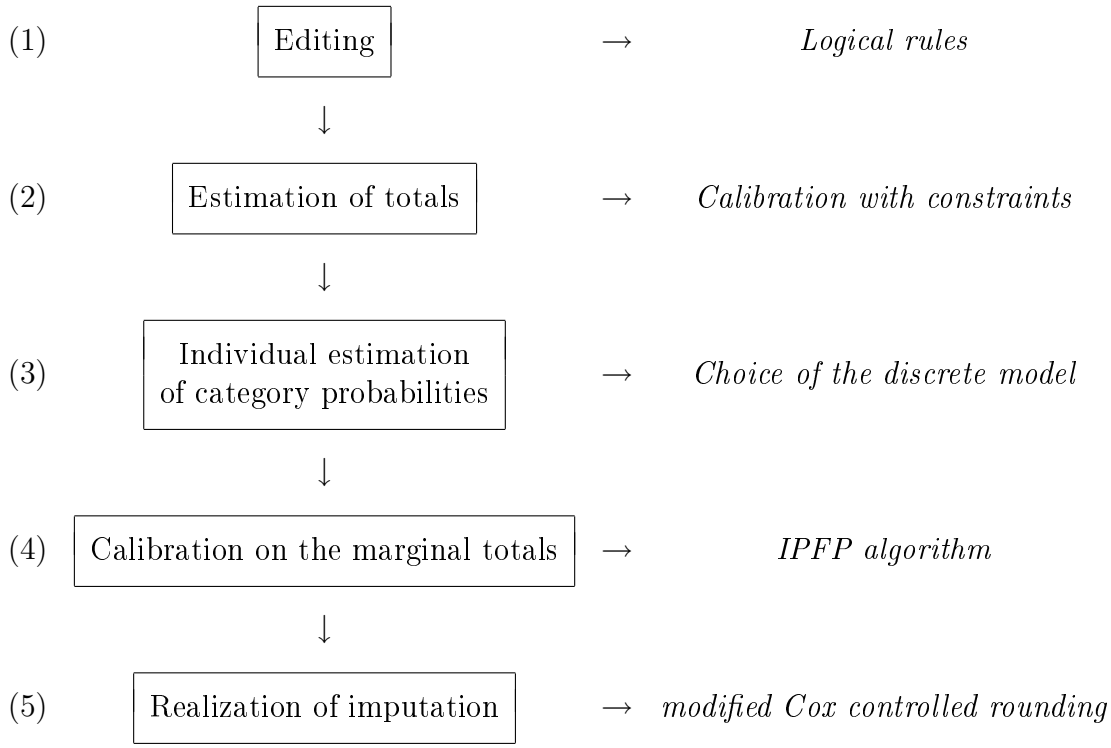


Figure 6.2: Overview of the imputation method with five steps

6.2. Clearly in this very simple case, the only feasible imputation is Marital

Table 6.2: Edit rules for age and marital status

e_1: (Marital status=widowed) & (Age=0-16)	=	False
e_2: (Marital status=divorced) & (Age=0-16)	=	False
e_3: (Marital status=married) & (Age=0-16)	=	False
e_4: (Marital status=single) & (Age=0-16)	=	True

status=single.

In some cases, the editing rules allow to identify the only possible value for the imputed data. Nevertheless, in most of the cases, Fellegi and Holt (1976) pointed out that the logical rules can always be viewed as incompatibilities between categories. For instance, the rule:

(Age = [0-25]) & (Professional status=retired) = False

excludes the category 'retired', but does not allow to derive the professional status. In the frame, it means that in some rare cases the value can be predicted exactly, but in most of the cases categories are excluded. Figure 6.3 shows an example of frame after editing.

$$
\begin{array}{l}
R \left\{ \begin{array}{|cccc|} \hline 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ & & \vdots & \\ 0 & 0 & 0 & 1 \\ \hline \end{array} \right. \\
\bar{R} \left\{ \begin{array}{|cccc|} \hline ? & \mathbf{0} & \mathbf{0} & ? \\ \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ ? & ? & ? & ? \\ \mathbf{0} & ? & ? & ? \\ \hline \end{array} \right.
\end{array}
$$

Figure 6.3: Frame after editing

6.4 Estimation of totals

We estimate the totals Y_j of each category accurately. The non-response is corrected by a calibration of the respondents in the sample. A separate calibration is done for each category with the transversal constraint

$$\sum_{j=1}^{v} \widehat{Y}_j = \widehat{N},$$

where $\widehat{N} = \sum_S w_k$. The problem turns to a minimization in $\nu_k^{(j)}$ of

$$\sum_{k \in R_j} G(\nu_k^{(j)}, w_k)$$

subject to the following system

$$\begin{cases} \sum\limits_{\substack{k\in R_j \\ v}} \nu_k^{(j)}\mathbf{x}_k & = & \sum\limits_S w_k\mathbf{x}_k, \quad \text{for all } j, \\ \sum\limits_{j=1}\sum\limits_{k\in R_j} \nu_k^{(j)} y_{kj} & = & \sum\limits_S w_k, \end{cases} \tag{6.1}$$

where G is a calibration-distance function defined in Deville and Särndal (1992) i.e. $G(\cdot, w_k)$ is positive, strictly convex, differentiable and such that $G(w_k, w_k) = 0$. It is more judicious to choose a function G so that $\nu_k^{(j)} \geq w_k$ for all k and j. With this particular choice, the ratio $w_k/\nu_k^{(j)}$ can be viewed as an estimator of the non-response probabilities (on this topic, see Deville, 2000a) of variable j for unit k, $w_k/\nu_k^{(j)} = \widehat{\Pr}[k \in R_j|S]$, where R_j is the respondent subset of S for j^{th} category. The problem of calibration becomes difficult when the values of the variable of interest are not defined on the same subset of the population for the various categories.

The solution of this optimization problem gives the following new weights for each category:

$$\nu_k^{(j)} = w_k F(\boldsymbol{\lambda}_j\mathbf{x}_k + \gamma y_{kj}),$$

where $w_k F(\cdot) = g^{-1}(\cdot)$ and $g(\cdot) = \partial G(.)/\partial w_k$.

The Lagrange multipliers $\boldsymbol{\lambda}_j \in \mathbb{R}^J$ and $\gamma \in \mathbb{R}$ are derived using the above constraints. We determine the $\boldsymbol{\lambda}_j$ and γ using the calibration equations (6.1).

As a special case we can consider $F(\cdot) = 1 + \exp(\cdot)$, then the system (6.1) becomes

$$\begin{cases} \sum\limits_{\substack{R_j \\ v}} w_k\big[1 + \exp(\boldsymbol{\lambda}_j\mathbf{x}_k + \gamma y_{kj})\big]\mathbf{x}_k = \sum\limits_S w_k\mathbf{x}_k, \\ \sum\limits_{j=1}\sum\limits_{k\in R_j} w_k\big[1 + \exp(\boldsymbol{\lambda}_j\mathbf{x}_k + \gamma y_{kj})\big] y_{kj} = \sum\limits_S w_k. \end{cases}$$

After the determination of $\boldsymbol{\lambda}_j$ and γ, the non-response probability is estimated by

$$\widehat{\Pr}[k \in R_j|S] = \frac{w_k}{\nu_k^{(j)}} = \frac{1}{1 + \exp(\boldsymbol{\lambda}_j\mathbf{x}_k + \gamma y_{kj})}.$$

The non-response behavior is thus modelled by a logistic regression which parameters are estimated with the calibration equations (see Deville, 2000a).

The final result of this optimization problem is an estimator of totals

$$\widehat{Y}_j = \sum_{k\in R_j} \nu_k^{(j)} y_{kj} = \sum_{k\in R_j} \frac{w_k}{\widehat{\Pr}[k \in R_j|S]} y_{kj},$$

which is consistent since

$$\sum_{j=1}^{v} \hat{Y}_j = \sum_{S} w_k = \widehat{N}.$$

The second step of the method provides the Y_j. Figure 6.4 illustrates the frame after estimation of totals. Now, the aim is to fill matrix $\bar{R}$ with values consistent with the Y_j.

$$\begin{array}{c} \\ R \\ \\ \\ \\ \\ \\ \bar{R} \\ \\ \\ \\ Y \end{array} \left\{ \begin{array}{|cccc|} \hline 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ & & \vdots & \\ 0 & 0 & 0 & 1 \\ \hline ? & 0 & 0 & ? \\ 1 & 0 & 0 & 0 \\ ? & ? & ? & ? \\ 0 & ? & ? & ? \\ \hline \widehat{Y}_1 & \widehat{Y}_2 & \cdots & \widehat{Y}_v \\ \hline \end{array} \right.$$

Figure 6.4: Frame after estimation of totals

6.5 Individual estimation of category probabilities

The next step in the procedure is the individual estimation of category probabilities. These probabilities can be estimated by a multinomial logistic model that can be written:

$$p_{kj} = \Pr[y_{kj} = 1 | \mathbf{x}_k] = \frac{\exp(\mathbf{x}_k' \boldsymbol{\beta}_j)}{\sum_{i=1}^{v} \exp(\mathbf{x}_k' \boldsymbol{\beta}_i)}.$$

However, since the values for p_{kj} are given for some cases according to the editing rules applied in phase 1, the estimated category probabilities should

$$
\begin{array}{cc}
R & \left\{\begin{array}{|cccc|}
\hline
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 \\
 & & \vdots & \\
0 & 0 & 0 & 1 \\
\hline
\end{array}\right. \\
\bar{R} & \left\{\begin{array}{|cccc|}
\hline
\mathbf{p_{11}} & 0 & 0 & \mathbf{p_{14}} \\
1 & 0 & 0 & 0 \\
\mathbf{p_{31}} & \mathbf{p_{32}} & \mathbf{p_{33}} & \mathbf{p_{34}} \\
0 & \mathbf{p_{42}} & \mathbf{p_{43}} & \mathbf{p_{44}} \\
\hline
\end{array}\right. \\
Y & \left\{\begin{array}{|cccc|}
\hline
\widehat{Y}_1 & \widehat{Y}_2 & \widehat{Y}_3 & \widehat{Y}_4 \\
\hline
\end{array}\right.
\end{array}
$$

Figure 6.5: Frame after estimation of the category probabilities

take in account this information. The final estimation $p_{kj} = \Pr[y_{kj} = 1|\mathbf{x}_k]$ for these cases is given by:

$$\Pr[y_{kj} = 1|\mathbf{x}_k, \text{ for all } \ell = 1, \ldots, v, \text{ with } y_{k\ell} = 0] = \frac{p_{kj}}{1 - \sum_{\ell=1, y_{k\ell}=0}^{v} p_{k\ell}}.$$

Let $\mathbf{P} = (p_{kj})$ the matrix obtained in this step. The following frame give the result after the individual estimation of category probabilities for the non-respondents:

$$
\bar{R} \left\{\begin{array}{|cccc|}
\hline
\mathbf{p_{11}} & 0 & 0 & \mathbf{p_{14}} \\
1 & 0 & 0 & 0 \\
\mathbf{p_{31}} & \mathbf{p_{32}} & \mathbf{p_{33}} & \mathbf{p_{34}} \\
0 & \mathbf{p_{42}} & \mathbf{p_{43}} & \mathbf{p_{44}} \\
\hline
\end{array}\right. = \mathbf{P}.
$$

The obtained total frame at this step is represented in Figure 6.5. A remaining problem at this stage is that the totals of the columns are not equal to the $\widehat{Y}_j$.

6.6 Calibration on the marginal totals

To fix the problem, the frame $\mathbf{P}$ is calibrated according the following constraints:

$$\left\{\begin{array}{lcl} \displaystyle\sum_{j=1}^{v} p_{kj} & = & 1, \text{ for all } k \in \bar{R}, \\ \displaystyle\sum_{k\in\bar{R}} w_k p_{kj} & = & T_j = \widehat{Y}_j - \displaystyle\sum_{k\in R} w_k y_{kj}, \text{ for all } j = 1, \ldots, v. \end{array}\right. \tag{6.2}$$

In order to simplify, let $\mathbf{Q} = (q_{kj}), q_{kj} = w_k p_{kj}$. The raking ratio procedure, also called algorithm IPFP, Iterative Proportional Fitting Procedure, (see Deming and Stephan, 1940), allows to realize such an adjustment on $\mathbf{Q}$. We search a matrix $\tilde{\mathbf{Q}} = (\tilde{q}_{kj})$ close to $\mathbf{Q}$, in such a way that:

$$\left\{\begin{array}{lcl} \displaystyle\sum_{j=1}^{v} \tilde{q}_{kj} & = & w_k, \text{ for all } k \in \bar{R}, \\ \displaystyle\sum_{k\in\bar{R}} \tilde{q}_{kj} & = & T_j = \widehat{Y}_j - \displaystyle\sum_{k\in R} w_k y_{kj}, \text{ for all } j = 1, \ldots, v. \end{array}\right.$$

At each step of the raking ratio procedure, the rows and the columns are successively adjusted in order to obtain $\tilde{q}_{ij}$ subject to (6.2). The following matrix is obtained with the sum of each row k equal to w_k and the sum of each column equal to T_j :

$$\tilde{\mathbf{Q}} = \begin{array}{|cccc|c|} \hline \tilde{q}_{11} & 0 & 0 & \tilde{q}_{14} & w_1 \\ \tilde{q}_{21} & 0 & 0 & 0 & w_2 \\ \tilde{q}_{31} & \tilde{q}_{32} & \tilde{q}_{33} & \tilde{q}_{34} & w_3 \\ 0 & \tilde{q}_{42} & \tilde{q}_{43} & \tilde{q}_{44} & w_4 \\ \hline T_1 & T_2 & T_3 & T_4 & \end{array}$$

Let $\tilde{\mathbf{P}} = (\tilde{p}_{kj}), \tilde{p}_{kj} = \tilde{q}_{kj}/w_k$. The matrix $\tilde{\mathbf{P}}$ has the properties

$$\left\{\begin{array}{lcl} \displaystyle\sum_{j=1}^{v} \tilde{p}_{kj} & = & 1, \text{ for all } k \in \bar{R}, \\ \displaystyle\sum_{k\in\bar{R}} w_k \tilde{p}_{kj} & = & T_j = \widehat{Y}_j - \displaystyle\sum_{k\in R} w_k y_{kj}, \text{ for all } j = 1, \ldots, v. \end{array}\right.$$

At the end of this step, the frame can be represented as in Figure 6.6. The totals of the frame columns are now exactly equal to the $\widehat{Y}_j$, for all $j = 1, \ldots, v$.

$$
\begin{array}{cc}
R & \left\{\begin{array}{|cccc|}\hline
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 \\
 & & \vdots & \\
0 & 0 & 0 & 1 \\ \hline
\end{array}\right. \\
\bar{R} & \left\{\begin{array}{|cccc|}\hline
\widetilde{\mathbf{p}}_{\mathbf{11}} & 0 & 0 & \widetilde{\mathbf{p}}_{\mathbf{14}} \\
1 & 0 & 0 & 0 \\
\widetilde{\mathbf{p}}_{\mathbf{31}} & \widetilde{\mathbf{p}}_{\mathbf{32}} & \widetilde{\mathbf{p}}_{\mathbf{33}} & \widetilde{\mathbf{p}}_{\mathbf{34}} \\
0 & \widetilde{\mathbf{p}}_{\mathbf{42}} & \widetilde{\mathbf{p}}_{\mathbf{43}} & \widetilde{\mathbf{p}}_{\mathbf{44}} \\ \hline
\end{array}\right. \\
\widehat{Y} & \left\{\begin{array}{|cccc|}\hline
\widehat{Y}_1 & \widehat{Y}_2 & \widehat{Y}_3 & \widehat{Y}_4 \\ \hline
\end{array}\right.
\end{array}
$$

Figure 6.6: Frame after calibration on the marginal totals

6.7 Realization of imputation

The last step consists in the realization of the imputation using the idea developed by Deville and Tillé (2000). This method is based on the Cox algorithm (Cox, 1987), which allows to produce an unbiased rounding of the elements of the matrix, without modifying the marginal totals. As regards frame $\tilde{\mathbf{P}}$, the Cox algorithm cannot be used directly because the weights w_k must be taken into account for the calibration on the totals.

This algorithm has been modified by Favre et al. (2004) in order to take into account a weighting system. This new algorithm named 'Cox weighted algorithm' is applied on matrix $\mathbf{A} = (a_{kj})_{(m \times v)}$ defined by $a_{kj} = \tilde{p}_{kj} \times w_k$, for all $k \in \{1, \ldots, m\}$ and for all $j \in \{1, \ldots, v\}$ where m is the cardinal of $\bar{R}$. Note that matrix $\mathbf{A}$ does not have integer marginal totals.

The algorithm is based on a circular path $(i_1, j_1), (i_1, j_2), .., (i_\ell, j_{\ell+1}) = (i_\ell, j_1)$ or $(i_1, j_1), (i_2, j_1), .., (i_{\ell+1}, j_\ell) = (i_1, j_\ell)$, where (i_h, j_t) represents the position given in the row h and column t and which contains a non-integer value. A path can be simple as $(1,1), (1,2), (3,2), (3,1)$ or more complicate as $(2,2), (2,3), (3,3), (3,1), (4,1), (4,2)$. A detailed description and a discussion of the weighted Cox algorithm are given in Favre et al. (2004). However, the main steps of the procedure are defined as follows:

- *Step 1*: Define the control matrix $\mathbf{C} = (c_{kj})_{(m \times v)}$ as

$$c_{ij} = \begin{cases} 0 & \text{if } a_{ij} = 0 \text{ or } a_{ij} = w_i, \\ 1 & \text{if not.} \end{cases}$$

- *Step 2*: If all the elements of the matrix $\mathbf{C}$ are equal to 0, go to *Step 5*. Else make the rounding of the non-integer values of matrix $\mathbf{A}$ on the rows, determining the first non-integer value on the rows at the position (i_1, j_1), which gives the first position in our path, an alternating row-column or column-row path of non-integer values. Choose randomly at each iteration the type of transformation (plus or minus):

$$d_- = \min_{1 \leq q \leq l}(a_{i_q j_q}, w_{i_q} - a_{i_q j_{q+1}}), \quad d_+ = \min_{1 \leq q \leq l}(w_{i_q} - a_{i_q j_q}, a_{i_q j_{q+1}}),$$

where the current path is : $(i_1, j_1), (i_1, j_2), ..., (i_l, j_{l+1}) = (i_l, j_1)$.
If d_- is selected :
- transform $a_{i_q j_q}$ to $a_{i_q j_q} - d_-$ along the path,
- transform $a_{i_q j_{q+1}}$ to $a_{i_q j_{q+1}} + d_-$ along the path.
If d_+ is selected :
- transform $a_{i_q j_q}$ to $a_{i_q j_q} + d_+$ along the path,
- transform $a_{i_q j_{q+1}}$ to $a_{i_q j_{q+1}} - d_+$ along the path.
Transform also the matrix $\mathbf{C}$ into:

$$c_{ij} = \begin{cases} 0 & \text{if } a_{ij} = 0 \text{ or } a_{ij} = w_i, \\ 1 & \text{if not.} \end{cases}$$

Moreover, if there is only one $c_{ij} = 1$ in a line i or in a column j, it is shifted to 0.

- *Step 3*: Return to *Step 2*.

- *Step 4*: Make a division of each row i of matrix $\mathbf{A}$ with w_i.

- *Step 5*: If some cells cannot be rounded, apply a Poisson sampling equalizing all the elements of the matrix to 0 or 1 .
End.

Concerning the variance of the total frequencies estimators per category due to the imputation, Favre et al. (2004) showed that the accuracy gain of the Cox weighted algorithm is in the order of $2v/m$ as compared with the Poisson sampling. In the previous algorithm, the choice of the paths is deterministic.

However, we advocate for a random choice of the paths, which can be done easily in sorting randomly the rows and the columns of the matrix, before applying the procedure. The final result with the imputed values is illustrated in Figure 6.7.

$$
\begin{array}{cc}
R & \left\{\begin{array}{|cccc|}\hline
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 \\
 & & \vdots & \\
0 & 0 & 0 & 1 \\ \hline
\end{array}\right. \\
\bar{R} & \left\{\begin{array}{|cccc|}\hline
\mathbf{0} & 0 & 0 & \mathbf{1} \\
1 & 0 & 0 & 0 \\
\mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{0} \\
0 & \mathbf{1} & \mathbf{0} & \mathbf{0} \\ \hline
\end{array}\right. \\
Y & \left\{\begin{array}{|cccc|}\hline
\widehat{Y}_1 & \widehat{Y}_2 & \cdots & \widehat{Y}_v \\ \hline
\end{array}\right.
\end{array}
$$

Figure 6.7: Final frame

6.8 An example

In a population U of size $N = 100$ a random sample S of size $n = 15$ is selected. The inclusion probabilities π_k are generated using uniform distribution and are rescaled such that $\sum_{k\in U} \pi_k = n$. The questionnaire is composed by 4 items (single, married, divorced, widowed), and we know the age (variable x) for all persons in the sample. After reorder, the sample is given in Figure 6.8.

The inclusion probabilities π_k for the subset R are:

$$0.175,\ 0.138,\ 0.069,\ 0.178,\ 0.048,\ 0.101,\ 0.264,\ 0.282,\ 0.220,\ 0.069,$$

and for the subset $\bar{R}$: $0.147,\ 0.125,\ 0.207,\ 0.239,\ 0.154$. The initial weights are $w_k = 1/\pi_k$. The calibration function used is $F(\cdot) = 1 + \exp(\cdot)$. The system (6.1) has been solved by using the function FindRoot() of Mathematica. The solution (6.1) is:

$$\lambda_1 = -0.018, \lambda_2 = -0.017, \lambda_3 = -0.018, \lambda_4 = -0.018, \gamma = -0.094.$$

	single	married	divorced	widowed	age
R	0	1	0	0	30
	1	0	0	0	42
	0	0	0	1	35
	1	0	0	0	27
	0	1	0	0	48
	0	0	1	0	33
	0	0	1	0	25
	0	1	0	0	65
	0	0	0	1	52
	0	1	0	0	23
$\bar{R}$	?	?	?	?	60
	?	?	?	?	34
	?	?	?	0	28
	?	?	?	?	25
	0	?	?	?	31

Figure 6.8: An example after reorder

Using the solution of system (6.1), the totals $\widehat{Y}_j$ and T_j are calculated and are given below:

$$\widehat{Y}_1 = 8.924, \widehat{Y}_2 = 57.120, \widehat{Y}_3 = 20.726, \widehat{Y}_4 = 6.217,$$

$$T_1 = 3.170, T_2 = 18.285, T_3 = 7.116, T_4 = 1.618.$$

The predicted probabilities p_{kj} for the non-respondent subset $\bar{R}$ using a multinomial logistic model and the re-estimation for the zero values coming from editing phase are following:

$$\mathbf{P} = \begin{pmatrix} 0.095 & 0.560 & 0.010 & 0.335 \\ 0.236 & 0.394 & 0.192 & 0.178 \\ 0.282 & 0.352 & 0.366 & 0 \\ 0.241 & 0.259 & 0.395 & 0.105 \\ 0 & 0.465 & 0.332 & 0.203 \end{pmatrix}.$$

We apply the IPFP algorithm in the following table $\mathbf{Q}$ and we provide w_k and T_j as marginal totals:

0.638	3.806	0.070	2.276	6.790
1.875	3.130	1.526	1.406	7.939
1.361	1.693	1.759	0	4.814
1.007	1.081	1.639	0.440	4.169
0	3.017	2.147	1.311	6.477
3.170	18.285	7.116	1.618	30.189

The result of the IPFP algorithm is the following:

$$\tilde{\mathbf{P}} = \begin{pmatrix} 0.061 & 0.826 & 0.010 & 0.101 \\ 0.157 & 0.587 & 0.201 & 0.054 \\ 0.171 & 0.479 & 0.349 & 0 \\ 0.162 & 0.390 & 0.415 & 0.032 \\ 0 & 0.629 & 0.314 & 0.056 \end{pmatrix}.$$

The application of the weighted Cox algorithm on the matrix $\mathbf{A} = (\tilde{p}_{kj} w_k)$ is given below (step 2 and 3 of the algorithm are applied 10 times; we note with $\mathbf{A}_{iteration-number}$ the matrix obtained after each iteration):

- *iteration 1:* The path is $(1,1),(1,2),(2,2),(2,1)$. The transformation minus with $d_- = \min(0.420, 6.790 - 5.610, 4.665, 7.939 - 1.247)$ is randomly chosen. The result of the first iteration is:

$$\mathbf{A}_1 = \begin{pmatrix} 0 & 6.029 & 0.072 & 0.688 \\ 1.667 & 4.245 & 1.596 & 0.430 \\ 0.827 & 2.306 & 1.681 & 0 \\ 0.676 & 1.626 & 1.730 & 0.135 \\ 0 & 4.077 & 2.036 & 0.363 \end{pmatrix}.$$

- *iteration 2:* The path is $(1,2),(1,3),(2,3),(2,2)$. The transformation minus with $d_- = \min(6.029, 6.790 - 6.029, 1.596, 7.939 - 4.245)$ is randomly chosen. The result of the second iteration is:

$$\mathbf{A}_2 = \begin{pmatrix} 0 & 4.432 & 1.669 & 0.688 \\ 1.667 & 5.841 & 0 & 0.430 \\ 0.827 & 2.306 & 1.681 & 0 \\ 0.676 & 1.626 & 1.730 & 0.135 \\ 0 & 4.077 & 2.036 & 0.363 \end{pmatrix}.$$

We skip the following 7 iterations. The last iteration is:

- *iteration 10:* The path is: $(2,2),(2,4),(5,4),(5,2)$. The result of the last iteration is:

$$\mathbf{A}_{10} = \begin{pmatrix} 0 & 0 & 6.790 & 0 \\ 3.170 & 3.149 & 0 & 1.618 \\ 0 & 4.814 & 0 & 0 \\ 0 & 4.169 & 0 & 0 \\ 0 & 6.153 & 0.326 & 0 \end{pmatrix}.$$

After the application of the step 4 of the weighted Cox algorithm, the result is the following:

$$\mathbf{P}_{end} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0.399 & 0.397 & 0 & 0.204 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0.949 & 0.051 & 0 \end{pmatrix}.$$

The matrix $\mathbf{P}_{end} = (p_{kj}^{end})$ has the properties that $\sum_{j=1}^{p} p_{kj}^{end} = 1$ and $\sum_k p_{kj}^{end} w_k = T_j$. So, the total frequencies are reproduced exactly and many values are rounded to 0 or 1.

We apply the last step, the Poisson sampling, because some values of the matrix $\mathbf{P}_{end}$ are not rounded. At this stage, the number of the not rounded values is maximum $2(v-1)$, and there are $2(v-1)/v$ cells not rounded in average. The result of the Poisson sampling (which is also the result of the imputation) is:

$$\bar{\mathbf{R}} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}.$$

We observe that the values obtained in the editing phase remain unchanged after the imputation. In our example, the final total frequencies are $T_1^{end} = 1, T_2^{end} = 3, T_3^{end} = 1, T_4^{end} = 0$. The difference between the initial values T_j and the final values T_j^{end} of the total frequencies is due to the frame size and the computational rounding. The total frequencies obtained after the application of the Poisson sampling are close to their initial values, in the context of a big database, and are the best estimations.

6.9 Discussion

It is thus possible to build an imputation technique that preserves the properties of both deterministic and random imputation methods. Moreover, the proposed procedure respects both the editing rules and the estimated calibrated totals. The totals are thus estimated according to a robust method, which does not produce any increase of the variance.

The variance of the imputed totals can be computed easily, because the imputed totals are equal, or quasi equal to the calibrated totals. The variance can thus be derived from a two-phase sampling design theory (see Särndal and Swensson, 1987), where the second phase of sampling is the result of the non-response mechanism.

The proposed method has thus many advantages, but the most important is the simplicity of the processing once the imputation is realized, which can be particularly appreciated when the methodologist who has realized the imputation does not make himself the data processing. The imputation reduces the data handling and estimators computation. The imputation is also random and therefore the variance estimator will not be inflate. Concerning the variance of the total frequencies estimators due to the imputation, Favre et al. (2004) have been showed that the gain of the proposed method is $2v/m$ as comparing with the usual Poisson sampling.

Bibliography

Aires, N. (1999). Algorithms to find exact inclusion probabilities for conditional Poisson sampling and Pareto πps sampling designs. *Methodology and Computing in Applied Probability*, 4:457–469.

Aires, N. (2000). Comparisons between conditional Poisson sampling and Pareto πps sampling designs. *Journal of Statistical Planning and Inference*, 82:1–15.

Ardilly, P. (1994). *Les Techniques de Sondage*. Technip, Paris.

Arthnari, T. and Dodge, Y. (1981). *Mathematical Programming in Statistics*. Wiley, Inc., New York.

Asok, C. and Sukhatme, B. (1976). On Sampford's procedure of unequal probability sampling without replacement. *Journal of American Statistical Association*, 71:912–918.

Atmer, J., Thulin, G., and Baecklund, S. (1975). Co-ordination of samples with the Jales technique (Swedish). *Statistisk Tidskrift (Statistical Review)*, 13:443–450.

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

Berger, Y. (1998b). Variance estimation using list sequential scheme for unequal probability sampling. *Journal of Official Statistics*, 14:315–323.

Bondesson, L. and Traat, I. (2005). On a matrix with integer eigenvalues and its relation to conditional Poisson sampling. *Res. Lett. Inf. Math. Sci.*, 8:155–163.

Bondesson, L., Traat, I., and Lundqvist, A. (2004). Pareto sampling versus Sampford and conditional Poisson sampling. Research Repport 6, University of Umeå, Sweden.

Brewer, K. (1972). Selecting several samples from a single population. *Australian Journal of Statistics*, 14:231–239.

Brewer, K. (2002). *Combined Survey Sampling Inference, Weighing Basu's Elephants*. Arnold, London.

Brewer, K. and Donadio, M. (2003). The high entropy variance of the Horvitz-Thompson estimator. *Survey Methodology*, 29:189–196.

Brewer, K., Early, L., and Hanif, M. (1984). Poisson, modified Poisson and Collocated sampling. *Journal of Statistical Planning and Inference*, 10:15–30.

Brewer, K., Early, L., and Joyce, S. (1972). Selecting several samples from a single population. *Australian Journal of Statistics*, 3:231–239.

Brewer, K. R. W. and Hanif, M. (1983). *Sampling with Unequal Probabilities*, volume 15 of *Lecure Notes in Statistics*. Springer-Verlag, New York.

Brick, J., Morganstein, D., and Wolter, C. (1987). Additional uses for Keyfitz selection. *Proceedings of the Survey Research Methods Section, The American Statistical Association*, pages 787–791.

Brown, L. D. (1986). *Fundamentals of Statistical Exponential Families: With Applications in Statistical Decision Theory*, volume 9 of *Institute of Mathematical Statistics Lecture Notes–Monograph Series*. Hayward, California.

Cao, G. and West, M. (1997). Computing distributions of order statistics. *Comm. Statistics. Theory Methods*, 26(3):755–764.

Causey, B. D., Cox, L. H., and Ernst, L. R. (1985). Application of transportation theory to statistical problems. *Journal of the American Statistical Association*, 80:903–909.

Chao, M. (1982). A general purpose unequal probability sampling plan. *Biometrika*, 69:653–656.

Chaudhuri, A. (1974). On some properties of the sampling scheme due to Midzuno. *Calcutta Statistical Association Bulletin*, 23:1–19.

Chen, S., Dempster, A., and Liu, J. (1994). Weighted finite population sampling to maximize entropy. *Biometrika*, 81:457–469.

Chen, S. and Liu, J. (1997). Statistical applications of the Poisson-binomial and conditional Bernoulli distributions. *Statistica Sinica*, 7:875–892.

Cotton, F. and Hesse, C. (1992a). Co-ordinated selection of stratified samples. *Proceedings of Statistics Canada Symposium 92*, 92:47–54.

Cotton, F. and Hesse, C. (1992b). Tirages coordonnés d'échantillons. Document de travail de la Direction des Statistiques Économiques E9206. Technical report, INSEE, Paris.

Cox, L. H. (1987). A constructive procedure for unbiased controlled rounding. *Journal of the American Statistical Association*, 82:520–524.

Cox, L. H. and Ernst, L. R. (1982). Controlled rounding. *INFOR*, 20:423–432.

Cumberland, W. and Royall, R. (1981). Prediction models in unequal probability sampling. *Journal of the Royal Statistical Society*, B 43:353–367.

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*, pages 673–676.

Deming, W. and Stephan, F. (1940). On a least square adjustment of sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 11:427–444.

Deville, J.-C. (1993). Estimation de la variance pour les enquêtes en deux phases. Manuscript, INSEE, Paris.

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

Deville, J.-C. (2000a). Generalized calibration and application to weighting for non-response. In Bethlehem, J. and van der Heijden, P., editors, *Compstat 2000 - Proceedings in Computational Statistics, 14th Symposium*, pages 65–76, New York. Springer-Verlag.

Deville, J.-C. (2000b). Note sur l'algorithme de Chen, Dempster et Liu. Technical report, CREST-ENSAI, Rennes.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87:376–382.

Deville, J.-C. and Tillé, Y. (2000). Selection of several unequal probability samples from the same population. *Journal of Statistical Planning and Inference*, 86:215–227.

Deville, J.-C. and Tillé, Y. (2004). Efficient balanced sampling: the cube method. *Biometrika*, 91:893–912.

Deville, J.-C. and Tillé, Y. (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference*, 128:411–425.

Dupacová, J. (1979). A note on rejective sampling. In Jurecková, J., editor, *Contributions to Statistics*, pages 71–78. Jaroslav Hájek memorial volume, Reidel, Holland and Academia, Prague.

Ernst, L. R. (1986). Maximizing the overlap between surveys when information is incomplete. *European Journal of Operational Research*, 27:192–200.

Ernst, L. R. (1996). Maximizing the overlap of sample units for two designs with simultaneous selection. *Journal of Official Statistics*, 12:33–45.

Ernst, L. R. (1998). Maximizing and minimizing overlap when selecting a large number of units per stratum simultaneously for two designs. *Journal of Official Statistics*, 14:297–314.

Ernst, L. R. (1999). The maximization and minimization of sample overlap problems: a half century of results. In *Proceedings of the International Statistical Institute, 52nd Session*, pages 168–182, Finland.

Ernst, L. R. (2001). Retrospective assignment of permanent random numbers for Ohlsson's exponential sampling overlap maximization procedure for designs with more than one sample unit per stratum. Technical report, U.S. Department of Labor, Bureau of Labor Statistics.

Ernst, L. R. and Ikeda, M. M. (1995). A reduced-size transportation algorithm for maximizing the overlap between surveys. *Survey Methodology*, 21:147–157.

Ernst, L. R., Izsak, Y., and Paben, S. (2004). Use of overlap maximization in the redesign of the national compensation survey. Technical report, U.S. Department of Labor, Bureau of Labor Statistics.

Ernst, L. R. and Paben, S. P. (2002). Maximizing and minimizing overlap when selecting any number of units per stratum simultaneously for two designs with different stratifications. *Journal of Official Statistics*, 18:185–202.

Ernst, L. R., Vaillant, R., and Casady, R. J. (2000). Permanent and collocated random number sampling and the coverage of births and deaths. *Journal of Official Statistics*, 16:211–228.

Fan, C., Muller, M., and Rezucha, I. (1962). Development of sampling plans by using sequential (item by item) selection techniques and digital computer. *Journal of the American Statistical Association*, 57:387–402.

Favre, A.-C., Matei, A., and Tillé, Y. (2004). A variant of the Cox algorithm for the imputation of non-response of qualitative data. *Computational Statistics & Data Analysis*, 45(4):709–719.

Favre, A.-C., Matei, A., and Tillé, Y. (2005). Calibrated random imputation for qualitative data. *Journal of Statistical Planning and Inference*, 128(2):411–425.

Fellegi, I. (1963). Sampling with varying probabilities without replacement: rotation and non-rotating samples. *Journal of American Statistical Association*, 58:183–201.

Fellegi, I. (1966). Changing the probabilities of selection when two units are selected with PPS without replacement. In *Proceeding of the Social Statistics Section, American Statistical Association*, pages 434–442, Washington.

Fellegi, P. and Holt, D. (1976). A systematic approach to automatic edit and imputation. *Journal of the American Statistical Association*, 71:17–35.

Gabler, S. (1981). A comparison of Sampford's sampling procedure versus unequal probability sampling with replacement. *Biometrika*, 68:725–727.

Gentle, J. E. (1998). *Random Number Generation and Monte Carlo Methods*. Verlag-Springer, New York.

Godambe, V. and Joshi, V. (1965). Admissibility and bayes estimation in sampling finite populations I. *Annals of Mathematical Statistics*, 36:1707–1722.

Goga, C. (2003). *Estimation de la variance dans les sondages à plusieurs échantillons et prise en compte de l'information auxiliaire par des modèles nonparamétriques.* PhD thesis, Université de Rennes II, Haute Bretagne, France.

Grafström, A. (2005). Comparisons of methods for generating conditional Poisson samples and Sampford samples. Master's thesis, UmeåUniversity, Sweden. Preprint.

Gray, G. and Platek, R. (1963). Several methods of re-designing area samples utilizing probabilities proportional to size when the sizes change significantly. *Journal of the American Statistical Association*, 63:1280–1297.

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

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

Hartley, H. and Rao, J. (1962). Sampling with unequal probabilities and without replacement. *Annals of Mathematical Statistics*, 33:350–374.

Hesse, C. (1998). Sampling co-ordination: A review by country. Technical Report E9908, Direction des Statistique d'Entreprises, INSEE, Paris.

Hinde, R. and Young, D. (1984). Synchronised sampling and overlap control manual. Methodology report, Australian Bureau os Statistics, Canberra, Australia.

Holmberg, A. and Swensson, B. (2001). On Pareto πps sampling: reflections on unequal probability sampling strategies. *Theory of Stochastic Processes*, 7(23),no.(1–2):142–155.

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

Isaki, C. and Fuller, W. (1982). Survey design under a regression population model. *Journal of the American Statistical Association*, 77:89–96.

Jonasson, J. and Nerman, O. (1996). On maximum entropy πps-sampling with fixed sample size. Technical Report 13, Dept. of Mathematics, Göteborg University and Charles University of Technology, Sweden.

Keyfitz, N. (1951). Sampling with probabilities proportional to size: adjustmentfor changes in the probabilities. *Journal of American Statistics Association*, 46:105–109.

Kish, L. (1963). Changing strata and selection probabilities. In *Proceeding of the Social Statistics Section, American Statistical Association*, pages 139–143, Washington.

Kish, L. and Hess, I. (1959). Some sampling techniques for continuing surveys operations. In *Proceeding of the Social Statistics Section, American Statistical Association*, pages 139–143, Washington.

Kish, L. and Scott, A. (1971). Retaining units after changing strata and probabilites. *Journal of the American Statistical Association*, 66:461–470.

Kröger, H., Sarndal, C., and Teikari, I. (1999). Poisson mixture sampling: a family of designs for coordinated selection using permanent random numbers. *Survey Methodology*, 25:3–11.

Kröger, H., Särndal, C., and Teikari, I. (2003). Poisson mixture sampling combined with ordered sampling. *Journal of Official Statistics*, 19:59–70.

Lessler, J. and Kalsbeek, W. (1992). *Non Sampling Error in Surveys*. John Wiley, New York.

Lundström, S. and Särndal, C.-E. (1999). Calibration as a standard method for treatment of nonresponse. *Journal of Official Statistics*, 15(2):305–327.

Matei, A. and Tillé, Y. (2004). On the maximal sample coordination. In Antoch, J., editor, *Proceedings in Computational Statistics, COMPSTAT'04*, pages 1471–1480. Physica-Verlag/Springer.

Matei, A. and Tillé, Y. (2005a). Computational aspects in order sampling designs. *submitted*.

Matei, A. and Tillé, Y. (2005b). Evaluation of variance approximations and estimators in maximum entropy sampling with unequal probability and fixed sample size. *Accepted in Journal of Official Statistics*.

Matei, A. and Tillé, Y. (2005c). Maximal and minimal sample co-ordination. *Accepted in Sankhyā*.

Midzuno, H. (1952). On the sampling system with probability proportional to sum of size. *Annals of the Institute of Statistical Mathematics*, 3:99–107.

Milbrodt, H. (1987). A note on Hájek's theory of rejective sampling. *Metrika*, 34:275–281.

Ng, M. and Donadio, M. (2005). Computing inclusion probabilities for order sampling. *to appear in Journal of Statistical Planning and Inference*.

Ogus, J. L. and Clark, D. F. (1971). The annual survey of manufactures: a report on methodology. Technical paper no. 24, Bureau of the Census, Washington D.C., USA.

Ohlsson, E. (1990). Sequential Poisson sampling from a business register and its application to the Swedish Consumer Price Index. R&D Report 1990:6, Statistics Sweden.

Ohlsson, E. (1992). The system for co-ordination of samples from the business register at Statistics Sweden. R&D report 1992:18, Statistics Sweden.

Ohlsson, E. (1995a). Coordination of samples using permanent random numbers. In Cox, B. G., Binder, D. A., Chinnapa, B. N., Christianson, A., Colledge, M. J., and Kott, P. S., editors, *Business Survey Methods*, chapter 9, pages 153–169. Wiley. inc., New York, USA.

Ohlsson, E. (1995b). Sequential Poisson sampling. Research report 182, Stockholm University, Sweden.

Ohlsson, E. (1996). Methods for PPS size: One sample coordination. Research report 194, Stockholm University, Sweden.

Ohlsson, E. (1998). Sequential Poisson sampling. *Journal of Official Statistics*, 14:149–162.

Ohlsson, E. (1999). Comparison of PRN techniques for small sample size pps sample coordination. Researh Report 210, Institute of Actuarial Mathematics and Mathematical Statistics, Stockholm University, Sweden.

Péa, J. (2004). Tirages coordonnés d'échantillons: Simulation des méthodes des substitutions de Kish & Scott et de la méthode PRN. Master's thesis, Université de Neuchâtel, Switzerland.

Patterson, H. (1950). Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society, B*, 12:241–255.

Pruhs, K. (1989). The computational complexity of some rounding and survey overlap problems. In *Proceedings of the Survey Research Methods Section, American Statistical Association*, pages 747–752.

Qualité, L. (2004). Echantillonnage à entropie maximale. Master's thesis, Laboratoire de Statistique d'enquête CREST, Rennes, France. Preprint.

Qualité, L. (2005). A comparison of conditional Poisson sampling versus unequal probability sampling with replacement. *submitted*.

Raj, D. (1968). *Sampling Theory*. McGraw-Hill, New York.

Reiss, P., Șchiopu-Kratina, I., and Mach, L. (2003). The use of the transportation problem in coordinating the selection of samples for business surveys. In *Proceedings of the Survey Methods Section, Statistical Society of Canada Annual Meeting, June 2003*.

Rivière, P. (2001). Coordinating sampling using the microstrata methodology. *Proceedings of Statistics Canada Symposium 2001*.

Rosén, B. (1991). Variance estimation for systematic pps-sampling. Technical Report 1991:15, Statistics Sweden.

Rosén, B. (1997a). Asymptotic theory for order sampling. *Journal of Statistical Planning and Inference*, 62:135–158.

Rosén, B. (1997b). On sampling with probability proportional to size. *Journal of Statistical Planning and Inference*, 62:159–191.

Rosén, B. (2000). On inclusion probabilties for order πps sampling. *Journal of Statistical Planning and Inference*, 90:117–143.

Saavedra, P. (1995). Fixed sample size PPS approximations with a permanent random number. In *Proceedings of the Section on Survey Research Methods*, pages 697–700. American Statistical Association.

Sampford, M. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika*, 54:499–513.

Sen, A. (1953). On the estimate of the variance in sampling with varying probabilities. *Journal of Indian Society for Agricultural Statistics*, 5:119–127.

Sengupta, S. (1989). On Chao's unequal probability sampling plan. *Biometrika*, 76:192–196.

Särndal, C.-E. and Swensson, B. (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse. *International Statistical Review*, 55:279–294.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer Verlag, New York.

Stehman, V. and Overton, W. S. (1994). Comparison of variance estimators of the Horvitz-Thompson estimator for randomized variable probability systematic sampling. *Journal of the American Statistical Association*, 89:30–43.

Sunter, A. (1977a). List sequential sampling with equal or unequal probabilities without replacement. *Applied Statistics*, 26:261–268.

Sunter, A. (1977b). Response burden, sample rotation, and classification renewal in economic surveys. *International Statistics Review*, 45:209–222.

Tillé, Y. (1996). Some remarks on unequal probability sampling designs without replacement. *Annales d'Economie et de Statistique*, 44:177–189.

Tillé, Y. (2001). *Théorie des sondages: échantillonnage et estimation en populations finies*. Dunod, Paris.

Tillé, Y. (2005). Sampling algorithms to equal and unequal probabilities. Technical report, University of Neuchâtel, Suisse.

Traat, I., Bondesson, L., and Meister, K. (2004). Sampling design and sample selection through distribution theory. *Journal of Statistical Planning and Inference*, 123:395–413.

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

Yates, F. and Grundy, P. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society, B*, 15:235–261.