Université de Neuchâtel
Institut de Statistique

Thèse

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

par

Erika Antal

Variance estimation via resampling methods for complex sampling designs

Acceptée sur proposition du jury composé de :

Pr.	Catalin	STARICA	Univ. de Neuchâtel	Président du jury
Pr.	Yves	TILLE	Univ. de Neuchâtel	Directeur de thèse
Pr.	Hervé	CARDOT	Univ. de Bourgogne	Rapporteur
Pr.	Isabel	MOLINA PERALTA	Univ. Carlos III de Madrid	Rapporteur

Thèse soutenue le 14.12.2012.

UNIVERSITÉ DE
NEUCHÂTEL

IMPRIMATUR POUR LA THÈSE

Variance estimation via resampling methods for complex sampling designs

Erika ANTAL

UNIVERSITÉ DE NEUCHÂTEL

FACULTÉ DES SCIENCES ÉCONOMIQUES

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

Prof. Yves Tillé (directeur de thèse, Université de Neuchâtel)
Prof. Catalin Starica (président du jury, Université de Neuchâtel)
Prof. Hervé Cardot (Université de Bourgogne)
Prof. Isabel Molina (Université Carlos 3, Madrid)

Autorise l'impression de la présente thèse.

Neuchâtel, le 21 février 2013

Le doyen

Gerald Reiner

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

...

Acknowledgements

First, I would like to thank my supervisor Prof. Yves Tillé for his help and for sharing his knowledge and experience with me. I am especially grateful for his constant support throughout my Ph.D studies. Even when I was about to lose hope he always managed to re-motivate me, which was a real encouragement.

I extend my sincere thanks to the members of the thesis committee: the president of the jury Prof. Catalin Starica and the two referees Prof. Hervé Cardot and Prof. Isabel Molina Peralta for accepting to be a part of it.

I would also like to thank the Office of Equal Opportunities at the University of Neuchâtel, which supported me financially for a semester through the 'Subside Tremplin' grant.

I am really grateful to all my colleagues at the Institute of Statistics, in particular, Anthea Monod, Desislava Nedyalkova and Matti Langel for their help and friendship.

Last, but not least, many thanks to my family and my friends for their support and encouragement.

Neuchâtel, 26 October, 2012.

Abstract and keywords

Variance estimation via resampling methods for complex sampling designs

Abstract This thesis is devoted to the variance estimation via re-sampling methods for complex sampling designs. These techniques are often called Bootstrap methods. Although the first Bootstrap methods were proposed in infinite population context, now there are several versions of the basic method dedicated for finite populations which is the context of survey sampling. The presentation and definition of the basic concept of the theory of survey sampling, with its usual notation, followed by the representation of divers existing resampling methods are the subjects of the first chapter (1). The following chapters present four papers submitted or published in international journals. The second chapter (2) is devoted to a new sampling method which will be very useful for the other new resampling methods for variance estimation, presented later. Chapter 3 proposes new algorithms for simple random sampling without replacement design, for Poisson design and also for a maximum entropy-type design with unequal inclusion probabilities. The next chapter (4) contains a simpler version of a common part of these methods presented previously, then the last chapter (5) can be seen as an extension to the treatment of the non-response often occurred in practice.

Keywords survey sampling, variance estimation, bootstrap, Poisson sampling design, Maximum entropy sampling design, two-phase sampling design, non-responses.

Estimation de variance par ré-échantillonage pour des plans de sondage complexes

Résumé Cette thèse est consacrée à l'estimation de la variance en utilisant des techniques de Bootstrap pour des plans de sondage complexes. Bien que les premières méthodes de Bootstrap ont été proposées dans un contexte de population infinie, il existe aujourd'hui plusieurs versions de la méthode de base adaptés aux populations finies, ce qui est le contexte de l'échantillonnage. La présentation et la définition du concept de base avec la notation habituelle de la théorie de l'échantillonnage, suivi par la présentation des divers méthods de Bootstrap existantes sont les sujets du premier chapitre (1) de cette thèse. Les 4 chapitres suivants présentent des articles soumis ou publiés dans des revues internationales. Notamment le deuxième chapitre (2) est dédié à un nouveau plan de sondage qui aura une très grande utilité pour les nouveaux plans de bootstrap proposés ultérieurement pour des problèmes d'estimation de variance, et qui sont, eux présentés dans les chapitres suivants. Notamment, dans le chapitre 3 des algorithmes de bootstrap pour un plan simple sans remise, pour un plan de Poisson puis pour un plans de type maximum entropy avec des probabilités

d'inclusion inégales sont proposés. Le chapitre suivant (4) contient une version plus simple d'une partie commune de ces algorithmes présentés précédemment, puis le thème du dernier chapitre (5) peut être vu comme une extension du problème au traitement de la non-réponse souvent présent dans la pratique.

Mots-clés sondage, estimation de variance, bootstrap methods, plan de Poisson, maximum entropy, plan de sondage en deux-phase, non-réponse.

Contents

List of Tables

Introduction

The field of survey methodology is one of the most widespread domains of applied statistics. Its aim is to draw a statistical inference concerning a real population, by samples selected according to defined rules. In general, it is impossible to observe the whole population, since it would be very expensive and would take a lot of time. For this reason, most of the time, only a part of the population is observed. There are several methods for drawing a sample. Different goals and different situations require different sampling designs.

Since the main objective is to estimate a given property of the population as reliably as possible, the most important issues are the bias of an estimator, and its variance. For most estimators, the property of unbiasedness has already been explored, proven theoretically and empirically. Thus the question of the variance estimation of an estimator is a more current field of research. In fact, in practice we only have one sample, thus we can calculate only one realisation of an estimator. Even if this estimator is unbiased, however the value obtained with this single sample is not necessarily the same as the parameter to be estimated. It is logical that one might wonder how likely this single value closely matches the parameter, or in other words, what is the variability of this estimator. The usual measure for this accuracy is the variance of the estimator. Of course, the more stable and reliable the variance, the better the estimator. Only the point estimation is not sufficient to judge wether an estimator is adequate or not. This is the reason why in general when examining the performance of an estimator, information about its variance is also needed. This information could be obtained by estimating this variance or by constructing a confidence interval directly.

Usually, in an infinite population, the variance formulae are derived using an approximation based on a Taylor series expansion of the estimator, or via resampling methods. In the environment of a finite population, the applicability of these methods is not so apparent; these methods require modifications. Moreover, these modified methods could be very

time-consuming, their efficiency depends on the estimators and also on the sampling designs. The goal of this thesis is to develop such resampling methods that could efficiently estimate the variance of an estimator.

This essay is structured as follows: the first chapter (1) contains a brief introduction of the two main topics. First we present the basic concepts and notation used in survey sampling. Then, we give an overview of the most commonly used resampling methods. Chapters 2 to 5 are self-contained papers published or submitted in peer-reviewed international journals. In Chapter 2 a new and useful tool for variance estimation is introduced. Based on this tool, which is a sampling design, called simple random sampling with over-replacement, a family of new methods to estimate the variance of an estimator is presented in Chapter 3. Chapter 4 is dedicated to a simplification of a common part of these methods and the method proposed in Chapter 5 may be considered as an extension of the resampling methods presented in Chapter 2 for special two-phase designs. Since a two-phase sampling design can always be seen as a non-response problem from a structural standpoint; a brief overview of this topic will be presented at the end of the first chapter (1). The document ends with a discussion and concluding remarks.

1 An introduction to survey sampling and variance estimation methods

Abstract

In this chapter first the concepts and notations in survey sampling methods are introduced. Then a short overview of existing methods for variance estimation is described. Section 1.1 presents the basic notions and well-known results concerning the most used estimators and their variance. In Section 1.2 we present some important and relevant variance estimation methods.

Keywords: complex sampling, resampling methods, bootstrap

1.1 Survey Sampling

1.1.1 The approach, the population and interest functions

In statistics, a population is in most cases a set of the elements that one would like to examine. Generally it is denoted by U. Its elements are the individuals and can be identified by their label. Let us use the notation:

$$U = \{u_1, \ldots, u_N\} \equiv \{1, \ldots, N\}.$$

This finite population can be viewed as a sample drawn from an infinite super-population; this is the case when a statistical model is supposed for this super-population. This approach is referred to as a model-based approach. Nevertheless, model misspecification can often occur. In order to avoid this problem, the design-based approach incorporates the whole stochastic structure in the sampling design. Moreover, large sample

asymptotic results under the design are generally preferable to estimators under model assumptions. Anyhow, most statisticians adopt both a model and design-based approach depending on the context. Here, through the entire document the design-based approach will be considered. Note that between the two main philosophies a good consensus could be the model-assisted approach (Särndal et al., 1992), especially for problems where the design-based one does not work, such as small area estimation.

The variable of interest is often denoted by y. It can be measured on the whole population, but its values are not known for each unit, since we observe only a subset of the population. Let y_k be the value of y taken on the k^{th} individual. The objective of the survey is to estimate a certain function (parameter) of the variable of interest and evaluate the precision of the estimate. The general notation of the function to be estimated is:

$$\theta = \theta(y_k, k \in U).$$

Regarding the functions commonly used to be estimated, one can find for example:

the total:

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

the mean:

$$\bar{Y} = \frac{1}{N} \sum_{k \in U} y_k = \frac{Y}{N},$$

the variance:

$$\sigma_y^2 = \frac{1}{N} \sum_{k \in U} (y_k - \bar{y})^2 = \frac{1}{2N^2} \sum_{k \in U} \sum_{l \in U} (y_k - y_l)^2 \text{ where } k \neq l,$$

or the corrected variance:

$$S_y^2 = \frac{1}{N-1} \sum_{k \in U} (y_k - \bar{y})^2 = \frac{N}{N-1} \sigma_y^2,$$

or some more complex parameters, such as the ratio of two totals of different variables:

$$R = \frac{Y}{X}.$$

Because of its important role in inequality theory, the Gini index also frequently needs to be estimated. Its formal definition for an ordered finite population is:

$$G = \frac{2}{NY} \sum_{k \in U} k y_k - \frac{N+1}{N}.$$

Another group of parameters contains non-smooth functions, such as

the median and the other quantiles and functions based on them. Among these functions, the median is the one most commonly used in estimation.

1.1.2 Sample and its design

As mentioned earlier, it is generally not possible to observe the whole population. Even if it were possible, it would require a lot of time and effort. Usually only a part of the population is examined. This observed subset is called a sample.

Let s denote a sample without replacement; a subset of the population. This sample is used to extrapolate results on the whole population. The set of all possible samples without replacement is denoted by $\mathcal{S}$;

$$\mathcal{S} = \{s | s \subset U\}.$$

A sampling design $p(.)$ is a probability distribution on the set $\mathcal{S}$ such that

$$\text{for all } s \subset U, \quad p(s) \geq 0 \text{ with } \sum_{s \in \mathcal{S}} p(s) = 1.$$

The random sample S takes a value s with probability

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

Let S denote the support of a sampling design which contains all subsets having a non-null probability of being selected as a sample:

$$\mathsf{S} = \{s \in U \text{ where } p(s) > 0\}.$$

There are countless ways to obtain a sample. Different situations and different goals require different criteria to be satisfied by the sampling design. Several basic properties may, however, be used to group the designs. For example, the sampling procedure can be conducted with or without replacement, the design can guarantee or not the same sample size for the possible samples, or the units can have the same or different probabilities of being selected in the sample.

Designs without replacement are preferable in practice and are generally more accurate in the point of view of estimation than design with replacement. The size of the sample, denoted n_S, is the number of selected individual observations. If the support of a sample design is included in the set of *n-size* samples (i.e. only these samples have a probability greater than zero, the size of the sample is not random and var $(n_S) = 0$) the sampling design is called a fixed size sampling design.

Whenever it is possible, it is better to use a fixed size sampling design to avoid additional variance caused by the design, which would make it incompressible. There are many designs (e.g. Poisson sampling or cluster sampling) that are important and useful, but lack this property.

When one performs a survey, the probabilities of selecting any subset of the population could be calculated. In most cases, however, finding the whole probability distribution is quite laborious. Fortunately, in general, the estimators may be calculated via the inclusion probabilities.

With the notation of Särndal et al. (1992) let $I_k = \delta_{k\in S}$ denote an indicator variable for unit k, which is defined by:

$$I_k = \begin{cases} 1 & \text{if unit } k \in S \\ 0 & \text{if unit } k \notin S. \end{cases}$$

For every sampling design, the probability of including the k^{th} observation in the sample can be defined. This probability is known as the inclusion probability, which can be derived from the sampling design:

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

Furthermore the joint inclusion probability of two different units k and ℓ is:

$$\pi_{k\ell} = \Pr(k \in S \text{ and } \ell \in S) = \sum_{\substack{s\subset U \\ k,\ell\in s}} p(s).$$

For every sampling design, the indicator variable has the following properties:

$$\mathrm{E}(I_k) = \pi_k,$$

$$\mathrm{var}(I_k) = \pi_k(1-\pi_k),$$

and

$$\mathrm{cov}(I_k, I_\ell) = \pi_{k\ell} - \pi_k\pi_\ell = \Delta_{k\ell}.$$

For the proof, see Tillé (2001, page 31.).

In addition, if the sampling design has a fixed size n, the following properties can be proven (Tillé, 2001):

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

$$\text{for all } \ell \in U \quad \sum_{\substack{k\in U \\ k\neq l}} \pi_{k\ell} = \pi_\ell(n-1),$$

and

$$\text{for all } \ell \in U \quad \sum_{k \in U} (\pi_{k\ell} - \pi_k \pi_\ell) = 0.$$

1.1.3 Estimators and their properties

Let Z denote the statistic as a function of the observations:

$$Z = u(D),$$

where u is a function and $D = \{(y_k; k), k \in S\}$ is the set of the observed data. The expectation E, is defined by the sampling design:

$$\mathrm{E}(Z) = \sum_{s \in \mathsf{S}} \Pr(S = s) Z(s),$$

The probability distribution of Z can be derived from the sampling design:

$$\Pr(Z = z) = \sum_{s | Z(s) = z} p(s),$$

thus

$$\mathrm{E}(Z) = \sum_{z \in \mathsf{Z}} z \Pr(Z = z),$$

where Z is the set of all possible values of Z. The variance (var) of a statistic can be defined via the expectation:

$$\mathrm{var}(Z) = \mathrm{E}(Z - \mathrm{E}(Z))^2.$$

Finally, an estimator $\widehat{\theta}$ is a statistic used to estimate a function of interest θ of y_N.

Since many estimators can be created to estimate the same function of interest, we would like to know which of these estimators is the best. This requires to consider the various properties that determine the quality of the estimation. In this aspect, one important property of an estimator is its bias. An estimator $\widehat{\theta}$ is unbiased if and only if:

$$\mathrm{E}(\widehat{\theta}) = \theta, \quad \text{for all } \mathbf{y}_N \in \mathbb{R}^{\mathbf{N}}$$

where

$$y_N = (y_1 \dots y_N)^T.$$

Its bias is defined by:

$$B(\widehat{\theta}) = \mathrm{E}(\widehat{\theta}) - \theta,$$

and its mean square error by:

$$MSE(\widehat{\theta}) = \mathrm{E}(\widehat{\theta} - \theta)^2 = \mathrm{var}(\widehat{\theta}) + B^2(\widehat{\theta}),$$

where the $\mathrm{var}(\widehat{\theta})$ is the variance of $\widehat{\theta}$:

$$\mathrm{var}(\widehat{\theta}) = \mathrm{E}\big[\widehat{\theta} - \mathrm{E}(\widehat{\theta})\big]^2.$$

The basic estimator of the total is the Horvitz-Thompson estimator (Horvitz & Thompson, 1952). It is defined as:

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

if for all $k \in U$, $\pi_k > 0$, and

$$\mathrm{E}(\widehat{Y}_\pi) = \mathrm{E}\left(\sum_{k \in U} \frac{y_k}{\pi_k} I_k \right) = \sum_{k \in U} \frac{y_k}{\pi_k} \mathrm{E}(I_k) = Y. \tag{1.1}$$

Generally if one or some inclusion probabilities are equal to zero, it indicates a coverage problem.

The Horvitz-Thompson estimator is also called the π-estimator or the expansion estimator, as it is a weighted estimator giving weight $w_k = 1/\pi_k$ to individual k in the sample. It means that this observation k represents $1/\pi_k$ observations of the population. It can be proven that if the first order inclusion probabilities are all greater than zero, the π-estimator is an unbiased estimator for every linear function of totals, independently of the sampling design.

The π-estimator of the mean is:

$$\widehat{\bar{Y}}_\pi = \frac{1}{N} \sum_{k \in S} \frac{y_k}{\pi_k}, \tag{1.2}$$

and of the size of the population is:

$$\widehat{N}_\pi = \sum_{k \in S} \frac{1}{\pi_k}.$$

From the expression (1.1), these estimators are unbiased. For their mean square errors, calculation (or estimation) of their variances is needed. The variance of the π-estimator of the total is:

$$\mathrm{var}(\widehat{Y}_\pi) = \sum_{k \in U} \sum_{\ell \in U} \frac{y_k}{\pi_k} \frac{y_\ell}{\pi_\ell} \Delta_{k\ell}, \tag{1.3}$$

if $\pi_k > 0$ and $k \in U$.

If $\pi_{k\ell} > 0$ for all $k, \ell \in U$ this estimator can be estimated unbiasedly by:

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

Since the variance of a linear function of the total can be derived from the variance of the total, only the variance estimation techniques for the total estimator will be considered. Of course when the estimator is not a linear function of the total estimator, we need to use other techniques.

In cases where the sample size is fixed n, this variance formula can be written in a particular form (Sen, 1953; Yates & Grundy, 1953).

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

and if $\pi_{k\ell} > 0$ for all $k, \ell \in U$, it may be estimated unbiasedly by:

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

which can also be written under the quadratic form

$$\widehat{\text{var}}_D(\widehat{Y_\pi}) = \sum_{k \in S} \sum_{\ell \in S} \frac{y_k y_\ell}{\pi_k \pi_\ell} D_{k\ell},$$

with

$$D_{k\ell} = \begin{cases} -\sum\limits_{\substack{j \in S \\ j \neq k}} \dfrac{\Delta_{kj}}{\pi_{kj}} & \text{if } k = \ell \\ \dfrac{\Delta_{k\ell}}{\pi_{k\ell}} & \text{if } k \neq \ell. \end{cases}$$

However, this expression can be less then zero, which is problematic, considering that the variance can never be negative. Thus, this estimator needs a condition to be positive. A sufficient condition is (Sen-Yates-Grundy condition):

$$\pi_{k\ell} \leq \pi_k \pi_\ell \quad \text{for all } k \neq \ell \in S,$$

and to guarantee that it is satisfied for each possible sample:

$$\pi_{k\ell} \leq \pi_k \pi_\ell \quad \text{for all } k \neq \ell \in U.$$

An interesting problem with the Horvitz-Thompson estimator can be illustrated by the following situation. In cases where $y_k = C$ for all $k \in U$

but the variance of the sum of the weights in the sample is not zero:

$$\text{var}\left(\sum_{k\in S} w_k\right) = \text{var}\left(\sum_{k\in S} \frac{1}{\pi_k}\right) \neq 0,$$

for example, when the sample size is random, or the inclusion probabilities are not equal for each unit, and one would like to estimate the mean of y, the Horvitz-Thompson estimator is written as

$$\widehat{\widehat{Y}}_\pi = \frac{C}{N}\sum_{k\in S}\frac{1}{\pi_k},$$

thus it is not equal to C, but it is a variable whose mean is C. For this problem, another estimator, the Hájek estimator gives the solution (Hájek, 1964):

$$\widehat{\widehat{Y}}_H = \left(\sum_{\ell\in S}\frac{1}{\pi_\ell}\right)^{-1}\sum_{k\in S}\frac{y_k}{\pi_k}.$$

Thus even if N is known in the formula (1.2), it is replaced by $\widehat{N} = \sum_{\ell\in S}\frac{1}{\pi_\ell}$ in order to compensate for the variance of $\sum_{k\in S}\frac{y_k}{\pi_k}$. The sum of the weights in the populations is equal to 1. Generally this estimator is biased, but in most cases this bias is negligible.

The Hájek estimator for the total is $N\widehat{\widehat{Y}}_H$, where the size of the population is used as auxiliary information to improve the estimation.

Most of the main sampling designs are described in detail in the following chapters. Only the two-phase sampling design is discussed here, because of its connection to the topic of missing data.

1.1.4 Missing data problems and its link with 2-phase sampling designs

The missing data problem, in its simplest version, is when one or more values of variables of the data set are missing. This lack of information could cause problems for estimation. However, the extent of the problem depends on its structure. Basically there are two levels of non-response:

- unit non-response and
- item non-response.

The unit non-response occurs when values are missing for each of the variables for a given unit or observation. Item non-response occurs when some but not all variables are affected by missing information. Generally, the main reasons for the non-response are really different for these two types. Consequently, they should also be handled differently.

In case of unit non-response keeping only the complete data set for analysis is quite common. It is simple, does not use artificial values but the sampling size is reduced. Thus the estimation is not efficient. The sampling weights can not be used for estimation and there is considerable risk of bias. Another method used is reweighting. It consists of increasing the sample weight of the respondent units using the inverse of the response-probability or calibration.

For item non-response, the main treatment applied is imputation. In its simplest version, it consists of replacing the missing value by an artificial or existing value. With multiple imputation, instead of one single value several are calculated. The main advantage of imputation is to provide a complete data set with single sample weights. Its principal disadvantage is that the inference made using imputed data is valid only if the hypothesis concerning the method of imputation is valid. In addition, variance can be considerably under-estimated.

From the standpoint of estimation, item non-response is a much more important subject than unit non-response. This is the case when some information is available and the question is how it could be used to increase the quality of the estimation. Whatever the reason and approach used to handle missing information, the untreated data set contains observed and non-observed values. Thus, in both types of non-response, the sample may be considered as if it were divided into two parts. A part with observed values and the other part with unobserved ones. Note that non-response, and hence a missing data problem, also exists when the value of the variable of interest is present but the value/s of other variable/s are missing. Even if it is less problematic than the case where it is the variable of interest which is affected by missing information, it is important, especially when the estimator uses the value of an auxiliary variable for estimation.

Anyhow, the most crucial case of non-response is when it is the value of the variable of interest that is missing. Thus, the sample is divided into two parts, respondents and non-respondents with respect to the variable of interest. Structurally, this situation is equivalent to a two-phase sampling design. In the case of a two-phase sampling design, an initial sample is selected by means of any sampling design then in the second phase, another random sample is selected from the first sample, using another sampling design. This second sample is the final sample.

In the context of missing data, the first sample can be considered as the initial sample and the second one as the sample of respondents. The sampling design of the first phase corresponds to the initial sampling design

and is not subject to the non-response problem. Concerning the second phase, a corresponding design can be found depending on the type of the non-response mechanism. Generally the willingness to respond is supposed to be independent between the units. Regarding the non-response mechanism, three types can be distinguished.

- uniform,
- ignorable and
- non-ignorable.

The main difference between these three types is whether the probability of response depends on a variable or not. If so, we need to know, which variable. With a uniform mechanism, the probability of response does not depend on any variable. In such cases, we say that the data set is Missing Completely At Random (MCAR). The sampling design of the second phase can be seen as a simple random sampling without replacement design.

If this probability depends on an auxiliary variable, but not on the variable of interest, then the non-response mechanism is ignorable. Consequently, the non-response mechanism is conditionally independent of the variable of interest. In such cases, we say that the data set is Missing At Random (MAR). The corresponding second-phase sampling design can be considered as a design where the first order inclusion probabilities (π_k) depend on the values of this auxiliary variable.

The third and most important case is when the non-response mechanism is non-ignorable because the probability of response depends on the variable of interest. In such cases, we say that the data set is Not Missing At Random (NMAR). The bias of the estimators due to non-response could be major. In this situation, the second-phase sampling design could be seen as a design where the (π_k) depend on the values of the variable of interest. However, these values are not observed for each unit. Thus, in order to determine the sampling design of the second phase, other relevant information is needed.

This connection pointed out above between the two topics is the reason why estimation methods developed for two-phase designs could also be useful in handling the non-response problem. The last chapter (5) of this thesis proposes a variance estimation method for a special type of two-phase design.

1.1.5 Precision of the estimator, inference

As mentioned earlier, a confidence interval around $\widehat{\theta}$ is generally used to determine the precision of the estimator. It requires the probability distribution of $\widehat{\theta}$ which is, in practice, not known. Usually it is assumed to follow a Normal distribution, at least approximately. The problem with this approach is that it supposes that units are selected independently (central limit theorem), which is never the case in practice. Firstly, because in the survey environment, the population is always finite, secondly because the selection is in general without replacement.

In the case of a finite population, the central limit theorem was proven for most of sampling designs. For example, Hájek (1960) proved the theorem for a simple random sampling design without replacement. The basic idea of his proof is to consider the population, its size and the size of the sample as the realizations of three increasing series. For more details, see Hájek (1960).

Regarding other sampling designs, this theorem was studied by:

- Hájek (1964) for the π-estimator of the total for a conditional Poisson sampling design (or rejective method).
- Rosén (1972a,b) for the π-estimator of the total for a successive unit selection.
- Bickel & Freedman (1984) and Krewski & Rao (1981) for the estimator of the average, in the case of a stratified simple random sampling design.

Thus in most cases, the probability distribution of $\widehat{\theta}$ is considered as a Normal distribution, of course approximately. Hence the confidence interval (with a level of confidence $1-\alpha$) around the estimator is defined as:

$$CI(\theta,\alpha) = \left[\widehat{\theta} - z_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\widehat{\theta})}, \widehat{\theta} + z_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\widehat{\theta})}\right].$$

However, in order to have θ in this interval with a probability $1-\alpha$, it is also necessary to have a variance estimator with negligible bias. This is one of the reasons why estimating the variance of an estimator is a crucial question and consequently a subject of research.

1.2 Variance estimation methods

For a more complex function of the interest variable or in the case where the sampling design is more sophisticated, the variance estimators described in the previous section could become very complicated. For this

reason, a variance estimation method is applied. The most commonly used methods include the following:

1. The linearization method, (1.2.1)
2. The Jackknife method, (1.2.2)
3. Balanced repeated replications, (1.2.3)
4. Bootstrap methods (1.2.4).

The last three methods are particular cases of resampling methods.

1.2.1 The linearization method

Most of the estimators can be expressed as a differentiable function (f) of a vector of linear estimators.

$$\theta = f(Y_1, \ldots, Y_p)$$

and they can be estimated by their π-estimators:

$$\widehat{\theta} = f(\widehat{Y}_{1\pi}, \ldots, \widehat{Y}_{p\pi}).$$

The key idea of the method of linearization is to approach the estimators with a linear estimator $\widehat{\theta}_0$ obtained by the Taylor-linearization of the function f, at the point $(Y_1, \ldots, Y_p)$.

$$\widehat{\theta} \simeq \widehat{\theta}_0 = \theta + \sum_{i=1}^{p} c_i(\widehat{Y}_{i\pi} - Y)$$

where

$$c_i = \frac{\partial f(\widehat{Y}_{1\pi}, \ldots, \widehat{Y}_{p\pi})}{\partial \widehat{Y}_{i\pi}}|_{(Y_1,\ldots,Y_p)},$$

and the variance of $\widehat{\theta}$ is approximated by the variance of $\widehat{\theta}_0$.

According to Särndal et al. (1992), using the variance formula done by Horvitz and Thompson (1.3), the approximated variance of $\widehat{\theta}$ is:

$$\text{var}(\widehat{\theta}) \simeq \text{var}(\widehat{\theta}_0) = \sum_{k\in U}\sum_{\ell\in U} \frac{u_k}{\pi_k}\frac{u_\ell}{\pi_\ell}\Delta_{k\ell},$$

if $\pi_k > 0$ and $k \in U$. The $u_k = \sum_{i=1}^{p} c_i y_{ki}$, where y_{ki} is the value of the variable y_i taken on the k^{th} observation.

A consistent estimator of this variance could be written as:

$$\widehat{\text{var}}(\widehat{\theta}) = \widehat{\text{var}}(\widehat{\theta}_0) = \sum_{k\in S}\sum_{\ell\in S} \frac{\widehat{u}_k}{\pi_k}\frac{\widehat{u}_\ell}{\pi_\ell}\frac{\Delta_{k\ell}}{\pi_{k\ell}},$$

where

$$\widehat{u}_k = \sum_{i=1}^{p} \widehat{c}_i y_{ki}$$

and $\widehat{c}_i$-s come from the c_i-s, where each total is replaced by the corresponding π-estimator.

Even if there are several methods to determine the confidence intervals around some non-linear parameters, in the form mentioned above, linearization is used only to estimate the variance of explicit functions of totals. For this reason, a more general approach based on the influence-function is proposed by Deville (1999).

The influence function of θ at point k is defined as (if M exists):

$$I\theta_k(M) = \lim_{\epsilon \to 0} \frac{\theta(M + \epsilon\delta_k) - \theta(M)}{\epsilon},$$

where M is a measure that gives a unit weight to each unit k in U and δ_k is a Dirac measure. This influence function can be calculated in a general sampling case, where a sample S is selected from the population U, by a sampling design p, applying the inclusion probabilities π_k for the unit $k \in U$. This definition is different from the classical definition of the influence function because the total weight of M is often unknown. In this case, it can be estimated by its π-estimator $\widehat{M}$, which gives a weight of $1/\pi_k$ for the unit k in the sample. Thus $\theta(M)$ is estimated by $\theta(\widehat{M}) = \widehat{\theta}$.

Under certain hypothesis (see in Deville (1999)), the approximated variance of $\widehat{\theta}$ can be written as:

$$\mathrm{var}(\widehat{\theta}) = \sum_{k \in U} \sum_{\ell \in U} \frac{u_k}{\pi_k} \frac{u_\ell}{\pi_\ell} \Delta_{k\ell},$$

if $\pi_k > 0$ and $k \in U$, where $u_k = I\theta_k(M)$.

Its linearized estimator is:

$$\widehat{\mathrm{var}}(\widehat{\theta}) = \sum_{k \in S} \sum_{\ell \in S} \frac{\widehat{u}_k}{\pi_k} \frac{\widehat{u}_\ell}{\pi_\ell} \frac{\Delta_{k\ell}}{\pi_{k\ell}},$$

where the $\widehat{u}_k$-s are estimated from the $I\theta_k(M)$ obtained by replacing the u_k-s by their π-estimators.

1.2.2 The Jackknife method

This method was originally introduced to estimate the bias of a statistic (Quenouille, 1949) and then proposed for variance estimation in an infinite population by Tukey (1958). The basic idea is to systematically delete a group of units and recalculate the statistic on the remaining observa-

tions. The variance of the parameter is estimated by the variance of these Jackknife statistics. The simplest techniques consist of deleting only one observation at a time, recalculating the statistics without this observation and then comparing them with the initial value. The bias and the variance estimator can be calculated in this manner.

If the parameter to estimate is denoted by θ and its estimator coming from the initial sample by $\widehat{\theta}$, the statistics obtained from the n new samples are $\widehat{\theta}_{-1}, \ldots, \widehat{\theta}_{-n}$. Thus the bias of $\widehat{\theta}$ is estimated by:

$$\widehat{B}(\widehat{\theta}) = \frac{n-1}{n} \sum_{i=1}^{n} (\widehat{\theta}_{-i} - \widehat{\theta}) = (n-1)(\bar{\widehat{\theta}} - \widehat{\theta})$$

where

$$\bar{\widehat{\theta}} = \frac{1}{n} \sum_{i=1}^{n} \widehat{\theta}_{-i},$$

and the estimated variance is:

$$\widehat{\mathrm{var}}(\widehat{\theta}) = \frac{n-1}{n} \sum_{i=1}^{n} (\widehat{\theta}_{-i} - \bar{\widehat{\theta}})^2.$$

This simple Jackknife method provides a consistent estimation of the variance only for smooth statistics. For all parameters based on the quantiles, for example, it is inconsistent. The solution to this problem is proposed by Shao & Wu (1989). They generalized the method by deleting not only one observation from the sample in each turn, but a group of d units. They showed that if d is large enough ($d \to \infty$ while also $n \to \infty$), the inconsistency can be repaired. For a less smooth statistic, a larger d should be chosen.

In a survey environment, the application of the Jackknife method is quite complex. In many cases, it needs to be adjusted, because it can not capture the correction of a finite population $(1 - f)$ where $f = \frac{n}{N}$ the sampling fraction. When the initial sample is too large, the recalculations could became laborious. Even the delete-d jackknife method would require $\binom{n}{d}$ recalculations. In order to reduce this number, it is common to first divide the population into groups of d units and delete one group at a time. In this way, the number of recalculations is reduced, but a rounding problem could appear.

Rao & Wu (1985) showed that in the most cases, the linearization techniques and Jackknife method are asymptotically equivalent at the first order. They produce exactly the same estimation for the variance. In gen-

eral, the Jackknife estimator has a smaller bias, but a larger variance than the estimator obtained by linearization.

1.2.3 Balanced repeated replications

Among the various balanced repeated replications designs, the simplest is the balanced half-sampling method. It was originally introduced by Mac Carthy (1969). He proposed the method for stratified multistage designs, for cases where two primary sampling units for the first stage are selected with replacement in each strata. One of the two primary units is selected for the half-sample in each strata. Hence, there are 2^H different half-samples on which the statistic is recalculated. Thus, the number of recalculations can be quite high, when H, the number of strata, is large.

The generalization of this basic form is to apply the balanced repeated replications method when the size of the strata n_h is greater than 2. This can be achieved by grouping the primary sampling units into two groups in each stratum.

If $\alpha_{hr} = 1$ denotes that the first primary unit of the h^{th} stratum is selected in the r^{th} half-sample, and $\alpha_{hr} = -1$ denotes when it is not, the balanced equation can be written as:

$$\sum_{r=1}^{R} \alpha_{hr}\alpha_{kr} = 0 \qquad \text{for all } h,k = 1,\dots H, \quad h \neq k$$

and the estimator of the variance of $\widehat{\theta}$ as:

$$\widehat{\text{var}}_{BRR}(\widehat{\theta}) = R^{-1} \sum_{r=1}^{R} (\widehat{\theta}_r - \widehat{\theta})^2.$$

In practice, balanced repeated sampling could be created using a Hadamard matrix, and if necessary, $\widehat{\theta}$ could be replaced by $R^{-1}\sum_{r=1}^{R}\widehat{\theta}_r$.

The technique of balanced repeated replications produces a much smaller number of replications, which is a good argument in favour of using this method. However, it does not solve the problem of possible sensitivity to weight-perturbation. There are several parameters (e.g. ratio) which are very sensitive to this. In some cases, it would became impossible to calculate all replicated estimations, because the denominator could be cancelled for some re-weightings. The solution was proposed by Dippo et al. (1984). They propose to use a smoother re-weighting method, applying a factor ϵ, where $0 < \epsilon \leq 1$. In this manner, the estimator gives less weight to the unit instead of deleting it.

For many parameters, the

$$\widehat{\text{var}}_{BRR} = \widehat{\text{Var}}_{Jack} = \widehat{\text{Var}}_{Lin}.$$

Furthermore, if there are no missing values, the BRR variance estimator is also consistent for non-linear functions and non-smooth functions - as the median and other estimators based on the quantiles. This consistency was proven by Shao & Wu (1992).

Like Jackknife variance estimator, the BRR variance estimator is unable to capture the correction of a finite population in the case of the random sampling without replacement design, and therefore needs to be adjusted. A possible correction is proposed by Wu (1991).

1.2.4 Bootstrap methods

At present, the bootstrap method is probably the most largely used resampling method for variance estimation. It was originally proposed by Efron (1979) for an infinite population, where the observations are independent and identically distributed and where the sample is taken with replacement. This is the environment in which the method was profoundly studied at first. However, the conditions mentioned earlier are never satisfied in a survey context, thus the adaptation of the method is not direct. This failure of the basic bootstrap prompted the development of several modified bootstrap methods, such as the bootstrap without-replacement method, introduced by Booth et al. (1994), the bootstrap with-replacement method (Mac Carthy & Snowden, 1985), the rescaling bootstrap (Rao & Wu, 1988) or the mirror match bootstrap (Sitter, 1992a). It is worth mentioning the recent work of Beaumont & Patak (2012). Their bootstrap algorithm is very interesting, especially for Poisson sampling designs, which are often used to select samples for the purpose of studying economic indicators.

In a finite population, Presnell & Booth (1994) divide these methods into two groups: plug-in type methods and ad-hoc methods. The plug-in type methods, as their common name implies, apply a principle of plug-in while ad-hoc resampling designs are constructed so that the unbiased estimators of the first moments are reproduced by the method. Among the bootstrap variants described below, the basic bootstrap method and the bootstrap without replacement method are plug-in type methods, the four other (the bootstrap with replacement, the rescaling bootstrap, the mirror match bootstrap and the method of Beaumont and Patak) are part of the various ad-hoc methods.

For other relevant works in the context of an infinite population see: Hall (1992), Efron & Tibshirani (1993) and in a finite population see: Presnell & Booth (1994), Shao & Tu (1995) or Davison & Hinkley (1997).

Basic bootstrap method

The key idea of the basic bootstrap method is to estimate the population from where one can select a bootstrap sample and recalculate the required statistic. In order to obtain R bootstrap replicates, this procedure is repeated R times independently. As in most cases, the parameter θ, and the variance of its estimator, thus $\text{var}(\widehat{\theta})$ need to be estimated.

Let $\widehat{U}$ denote the estimated population, the r^{th} bootstrap sample S_r^B is obtained by a sampling design with replacement from this estimated population, and the statistic is $\widehat{\theta}_r^B = \theta(S_r^B)$. Thus, the bootstrap estimation for the variance of the parameter θ can be written as:

$$\widehat{\text{var}}_B(\widehat{\theta}) = \frac{1}{R-1}\sum_{r=1}^{R}(\widehat{\theta}_r^B - \overline{\widehat{\theta}^B})^2,$$

where

$$\overline{\widehat{\theta}^B} = \frac{1}{R}\sum_{r=1}^{R}\widehat{\theta}_r^B.$$

In the case of a stratified sample, this resampling process is estimated independently in each stratum.

As mentioned earlier, because of the departure from the real sampling design (which is almost never with-replacement) and the basic-bootstrap design (which is with replacement), the estimator does not include the finite population correction. It is thus biased and could be inconsistent, for example, when in a stratified sampling design the number of strata is large, but the sampling-ratios ($f_h = n_h/N_h$) have a distance limited from zero. If the mean square error of the estimator is small, disregarding the asymptotic problem could be a solution. However, this is not always a viable option. Modified bootstrap methods could provide better solutions.

Bootstrap without replacement

This bootstrap method is the most commonly used plug-in type method. In the case of non-stratified sampling designs, the simplest approach is to create a pseudo-population by replicating N/n times the original sample, and select the bootstrap samples without replacement from it. In the case of stratified sample design, the h^{th} stratum is replicated N_h/n_h times. However, in this case, the probability of having a rounding problem (at

least one of the N_h/n_h is not integer) increases considerably. This rounding problem may be resolved by randomization between $[N_h/n_h - 0.5]$ and $[N_h/n_h + 0.5]$ but there are some situations where this method does not provide a solution (Mac Carthy & Snowden, 1985).

An unbiased and consistent estimator for the linear statistics was given by Sitter (1992b). He proposes to select a bootstrap sample of size m_h without replacement, in each pseudo-stratum with size $M_h = k_h n_h$, where:

$$m_h = n_h - (1 - f_h) \quad \text{and} \quad k_h = \left(1 - \frac{1 - f_h}{n_h}\right) / f_h.$$

In this way, the sampling fraction in each pseudo-stratum M_h will be the same then in the real stratum N_h, thus $\frac{m_h}{M_h} = \frac{n_h}{N_h} = f_h$. As mentioned earlier, if N_h/n_h is not an integer, this method requires randomization. The details for such randomization are very important, and therefore require careful programming. In the case of huge surveys its performance is quite weak.

Bootstrap with replacement

Without creating pseudo-populations, Mac Carthy & Snowden (1985) suggest imitating the sampling design by selecting samples of size m_h, with replacement in each stratum, where

$$m_h = \frac{n_h - 1}{1 - f_h}.$$

They show that such an estimator is unbiased and consistent for linear statistics. Of course, when m_h is not an integer, this method also, requires randomization, thus thoughtful programming.

Rescaling bootstrap

Rao & Wu (1988) propose another type of modification of the basic bootstrap method. They also suggest a sampling design with replacement, but for the sample size, they propose:

$$m_h = \frac{1 - f_h}{(1 - 2f_h)^2} \frac{(n_h - 2)^2}{n_h - 1},$$

in each stratum, and they obtain the variance by rescaling the bootstrap sample as:

$$y^*_{hi} = \bar{y}_h + \left[\frac{m_h(1 - f_h)}{n_h - 1}\right]^2 (y_{hi} - \bar{y}_h)$$

in each stratum.

They show that the variance estimator obtained in this way is unbiased and consistent for smooth functions of the average. This method has the advantage of not creating pseudo-populations. In the case where the expression for the sample size does not give an integer number, it requires randomization, which affects the stability of the estimator. Moreover in the case where $m_h > n_h$, the method can provide an impossible value for the bootstrap statistic.

Mirror match bootstrap

The mirror-match bootstrap method, introduced by Sitter (1992a) is the crossing of the bootstrap with and without replacement methods. The main idea is to repeatedly select sub-samples of size n^* from the original sample, then put them together to obtain a resample S^{mb} of size $n^{mb} = k^* n^*$, where k^* is the number of repetitions carried out in the original sample S.

The proposed algorithm is the following:

suppose that $n^* \in (1, n)$, an integer number, where n is the size of the sample S, $f^* = n^*/n$ and

$$k^* = \frac{n(1-f^*)}{n^*(1-f)}$$

where k^* is also an integer number. The k^* sub-samples $s_1^*, \ldots, s_{k^*}^*$ of size n^* are selected independently from S. The mirror match bootstrap sample S^{mb} is the union of these subsamples.

Sitter suggests using $n^* = fn$ and proves that such a variance estimator is consistent for linear statistics. This method may also require randomization, because fn is rarely an integer (Presnell & Booth, 1994).

Wu (1986, 1990) mentions that the mirror match bootstrap could be considered as an adaptation of the delete-d Jackknife method. Shao & Tu (1995) shows that the use of $n^* = fn$ allows the finite population correction to be captured.

The method of Beaumont and Patak

As mentioned earlier, the bootstrap algorithm proposed by Beaumont & Patak (2012) is worth mentioning, especially for a Poisson sampling design. They show that a plug-in type bootstrap method with creation of pseudo-populations can be avoided by using an appropriate adjustment to the sampling weights w_k. Solutions are provided even if the multiply-

ing factors (w_k) used to create the artificial populations are not all integers. The bootstrap estimator of the total is written as:

$$\widehat{Y}^* = \sum_{k \in S} y_k w_k^*$$

where $w_k^* = w_k a_k$ and a_k are the bootstrap adjustments. They propose generating a_k, independently, for $k \in S$, from the binomial distribution $Bin(w_k, \pi_k)$ for the case where w_k are all integers. If this is not the case, instead of generating a_k, a_k^r is generated independently from the binomial distribution $Bin(w_k^r, \pi_k)$, where w_k^r are the randomly rounded sampling weights:

$$w_k^r = \begin{cases} \lfloor w_k \rfloor, & \text{with probability } \lfloor w_k \rfloor + 1 - w_k \\ \lfloor w_k \rfloor + 1 & \text{with probability } w_k - \lfloor w_k \rfloor, \end{cases}$$

where $\lfloor w_k \rfloor$ are the largest integer smaller than or equal to w_k. The final bootstrap adjustments are $a_k = a_k^r + (1 - \pi_k w_k^r)$. Unfortunately, these bootstrap adjustments are not always positive. However, with rescaling techniques, negative bootstrap weights can be avoided.

Confidence intervals

Using bootstrap methods, the whole probability distribution function of the estimator can be estimated. Thus, a confidence interval can be created around the point estimation. The most common approaches are the method of percentiles intervals (Efron, 1981) and the method of t-Bootstrap intervals (Efron, 1982). The main difference between them is that the first method directly applies the estimated probability function while the t-Bootstrap first estimates a studentized statistic of the parameter.

The full algorithms of these methods will not be discussed here, only their advantages and disadvantages in terms of the length of the provided intervals (i.e. the coverage) and their behaviour for the change of scale will be mentioned. Typically, when an interval is very short, the probability that it contains the parameter is lower, so the risk that the real value of the parameters is not in this interval (under-coverage) is higher. However, a very "careful" interval is too large, which makes no sense. The property of being invariant to the scale on which the parameter is calculated, means that if we calculate confidence interval for the transformed parameter and than back-transform it to the initial scale, we get the same intervals.

With regards to these two aspects, the method of percentiles intervals

is scale-invariant but frequently provides a too short interval leading to possible under-coverage. It is intuitive, easy to implement and does not require too many calculations. For small sample sizes, it is typically less precise than the t-Bootstrap, which is generally scale-invariant and provides good coverage (Davison & Hinkley, 1997).

In Chapters 3, 4 and 5, new resampling methods are proposed and their performances are compared to the other existing methods. Simulation studies are also carried out for numerical comparison purposes. In these simulation studies, performance is measured by, inter alia, the lower and the upper error rate, which requires to create confidence intervals. These confidence intervals are created using the t-Bootstrap method.

1.2.5 Summary of the behaviour of the mentioned methods in a survey environment

The last subsection summarizes the above-mentioned variance estimation methods in a survey environment, their advantages, disadvantages and difficulties.

- Linearization
  - In its basic form, the method is used only to estimate the variance of explicit smooth functions of totals.
  - Its version based on the influence function provides a consistent estimator of variance for non-explicit functions.
- Jackknife
  - The simple Jackknife method provides a consistent estimation of variance only for the smooth statistics.
  - In the case of non-smooth function, inconsistency can be repaired using its generalized version.
  - The jackknife cannot capture the correction of a finite population, so it needs to be adjusted.
  - When the initial sample is too large, recalculations could became laborious.
  - The delete-d jackknife method would require numerous recalculations and even if the population is first divided into groups of d units to reduce the number of recalculations, a rounding problem could appear.
  - In most of the cases, the linearization method and the Jackknife method produce exactly the same estimation of variance.

  - In general, the Jackknife estimator has a smaller bias, but a larger variance than the estimator obtained by linearization.

- Balanced repeated replications
  - The technique of balanced repeated replications produces a much smaller number of replications.
  - It is sensitive to weight-perturbations, which can be solved by using a smoother re-weighting method. This usually gives less weight to the unit, instead of deleting it.
  - For many parameters, it holds that

$$\widehat{\mathrm{Var}}_{BRR} = \widehat{\mathrm{Var}}_{Jack} = \widehat{\mathrm{Var}}_{Lin}.$$

  - Furthermore in absence of missing values, the BRR variance estimator is consistent not only for linear functions but also for non-linear, non-smooth functions.
  - It cannot capture the correction of a finite population in the case of a random sampling without replacement design. Thus it needs to be adjusted.

- Bootstrap
  - This is probably the most widely used resampling method for variance estimation.
  - Originally, it was developed for independent and identically distributed observations with a sampling with-replacement design, which is never the case in a survey environment.
  - The finite population correction is not considered by the estimator. Thus it is biased and in some cases could be inconsistent.
  - The modified bootstrap methods try to be more adaptable to the survey environment:
  - The bootstrap without replacement version creates pseudo-populations, which could be very inefficient.
  - A rounding problem usually occurs. Even if it could be solved by randomization, there are some situations where it does not provide a solution.
  - An unbiased and consistent estimator for a linear statistic can be found.
  - The bootstrap with replacement method does not require pseudo-populations to be created.

- It provides an unbiased and consistent estimator for a linear statistic.
- The rounding problem is also quite usual.
- With the rescaling version, an unbiased and consistent variance estimator can be obtained for the smooth functions of the total.
- It does not create pseudo-populations.
- In some cases, it requires randomization, which affects the stability of the estimator.
- It could provide an impossible value for the bootstrap statistic.
- The variance estimator obtained by the mirror match bootstrap method is consistent for linear statistics.
- It could also require randomization.
- The Baumont-Patak method calculates bootstrap weight adjustments for each unit in the sample.
- rounding problem often occurred, but solved by randomization.
- Possible negative weights are avoided by rescaling.

The objective of this thesis is to develop such a variance estimation method that can provide an unbiased and consistent estimator for many sampling designs. Besides these criteria, it should be easy to implement, not requiring adjustment or correction factor and it should have negligible difficulties concerning imputation, calibration, or weighting for non-response.

2 Simple Random Sampling With Over-replacement

Abstract

There are several ways for selecting units with replacement and an equal inclusion expectation. We present a new sampling design called simple random sampling with over-replacement. Its interest lies in the high variance produced for the Horvitz-Thompson estimator. This characteristic could be useful for resampling methods. [1]

Keywords: survey sampling, simple random sampling with replacement, discrete probability distribution, resampling method

2.1 Introduction

There are several methods for drawing a sample, different goals, and different situations that require different sampling designs. The most basic sampling procedures are simple random sampling with and without replacement. In this paper we show that there are several ways to select units with replacement with an equal inclusion expectation. A new method is proposed where the repetition of the units in the sample is more important than with usual simple random sampling with replacement. This sampling design called simple random sampling with over-replacement provides a larger variance. This property could be interesting for resampling methods. We show how to implement this design and we compare it to simple random sampling with and without replacement.

[1] This chapter is a reprint of: Antal, E. and Tillé, Y. (2011). Simple random sampling with over-replacement *Journal of Statistical Planning and Inference* **141**, 597–601.

2.2 Main concept and notation

A sampling design on a population $U = \{1, \ldots, k, \ldots, N\}$ is a procedure that allows us to randomly select statistical units. Some statistical units can be selected several times in the sample. In survey sampling theory, it is usual to define a sample as a subset of the population U. However, this definition is rather restrictive because it is limited to samples for which the units are selected only once, i.e. when sampling is done without replacement.

A more flexible notation consists in defining a sampling design by a positive, discrete random vector $\mathbf{S} = (S_1, \ldots, S_k, \ldots, S_N)'$, where S_k is the number of times unit k is selected in the sample. The same notation can thus be used to define sampling designs with or without replacement. If the sample is selected without replacement, then S_k can only take the values 0 and 1. If the sample has a fixed sample size n, then $\sum_{k \in U} S_k = n$.

The inclusion expectation of unit k is $\pi_k = \mathrm{E}(S_k)$. Since a unit can be selected several times in the sample, π_k can take any nonnegative value. The joint inclusion expectation of two units k and ℓ is the expectation of the product of S_k and S_ℓ, i.e. $\pi_{k\ell} = \mathrm{E}(S_k S_\ell)$. Moreover, $\Delta_{k\ell} = \mathrm{cov}[S_k, S_\ell] = \pi_{k\ell} - \pi_k \pi_\ell$. If the sample is selected without replacement, then the inclusion expectation is called inclusion probability.

Let $y_1, ..., y_N$ denote the values taken on the units of the population by an interest variable y. Suppose now that we want to estimate the total of these values $Y = \sum_{k \in U} y_k$. If all the $\pi_k > 0$, this total can be estimated without bias by $\widehat{Y} = \sum_{k \in U} S_k y_k / \pi_k$. This estimator is called the Horvitz-Thompson estimator if the sample is selected without replacement and the Hansen-Hurwitz estimator if the sample is selected with replacement (see Hansen & Hurwitz, 1949; Horvitz & Thompson, 1952).

The variance of $\widehat{Y}$ is

$$\mathrm{var}(\widehat{Y}) = \sum_{k \in U} \sum_{\ell \in U} \frac{y_k y_\ell}{\pi_k \pi_\ell} \Delta_{k\ell}.$$

If all the $\pi_{k\ell} > 0$, this variance can be estimated without bias by means of the following formula

$$\widehat{\mathrm{var}}(\widehat{Y}) = \sum_{k \in U} \sum_{\ell \in U} \frac{S_k S_\ell y_k y_\ell}{\pi_k \pi_\ell} \frac{\Delta_{k\ell}}{\pi_{k\ell}}. \tag{2.1}$$

Nevertheless, this variance estimator is often very unstable. It can take negative values (see, for instance Tillé, 2006, p. 26-29). When the sampling

design has a fixed sample size, the variance can be written as

$$\text{var}(\widehat{Y}) = \frac{-1}{2} \sum_{k \in U} \sum_{\ell \in U} \left(\frac{y_k}{\pi_k} - \frac{y_\ell}{\pi_\ell} \right)^2 \Delta_{k\ell},$$

and can be estimated by

$$\widehat{\text{var}}(\widehat{Y}) = \frac{-1}{2} \sum_{k \in U} \sum_{\ell \in U} S_k S_\ell \left(\frac{y_k}{\pi_k} - \frac{y_\ell}{\pi_\ell} \right)^2 \frac{\Delta_{k\ell}}{\pi_{k\ell}},$$

which can also be written under the quadratic form

$$\widehat{\text{var}}_D(\widehat{Y}) = \sum_{k \in U} \sum_{\ell \in U} \frac{S_k S_\ell y_k y_\ell}{\pi_k \pi_\ell} D_{k\ell}, \tag{2.2}$$

with

$$D_{k\ell} = \begin{cases} -\sum\limits_{\substack{j \in U \\ j \neq k}} S_j \dfrac{\Delta_{kj}}{\pi_{kj}} & \text{if } k = \ell \\ \dfrac{\Delta_{k\ell}}{\pi_{k\ell}} & \text{if } k \neq \ell. \end{cases}$$

2.3 Simple random sampling without replacement

A sampling design is said to be simple and without replacement if $\Pr(\mathbf{S} = \mathbf{s}) = n!(N-n)!/N!$, for all $\mathbf{s} \in \mathcal{S}_n^N$, where $\mathcal{S}_n^N = \left\{ \mathbf{s} \in \{0,1\}^N \left| \sum_{k=1}^N s_k = n \right. \right\}$. In simple random sampling without replacement, $\Delta_{k\ell} = -n(N-n)/\{N^2(N-1)\}$, if $k \neq \ell \in U$ and $\Delta_{kk} = n(N-n)/N^2, k \in U$, which gives the variance of the estimator of the total $\text{var}(\widehat{Y}) = N^2(N-n)\sigma^2/\{(N-1)n\}$, where $\sigma^2 = N^{-1}\sum_{k \in U}(y_k - \overline{Y})^2$, and $\overline{Y} = N^{-1}\sum_{k \in U} y_k$. Moreover, we have $\Delta_{k\ell}/\pi_{k\ell} = -(N-n)/\{N(n-1)\}$ when $k \neq \ell \in U$ and $\Delta_{kk}/\pi_{kk} = (N-n)/N, k \in U$, which gives $\widehat{\text{var}}(\widehat{Y}) = N^2(N-n)\widehat{\sigma}^2/(Nn)$, where $\widehat{\sigma}^2 = (n-1)^{-1}\sum_{k \in U} S_k(y_k - \widehat{\overline{Y}})^2$, and $\widehat{\overline{Y}} = n^{-1}\sum_{k \in U} S_k y_k$.

2.4 Simple random sampling with replacement

A sampling design is said to be simple and with replacement if

$$\Pr(\mathbf{S} = \mathbf{s}) = \frac{1}{N^n} \binom{n}{s_1 \cdots s_k \cdots s_N}^{-1}, \text{ for all } \mathbf{s} \in \mathcal{R}_n^N,$$

where $\mathcal{R}_n^N = \left\{ \mathbf{s} \in \mathbb{N} \left| \sum_{k=1}^N s_k = n \right. \right\}$. Vector $\mathbf{S}$ therefore has a multinomial distribution. A well-known result is that a multinomial distribution can be derived from a sequence of Poisson independent random variables

given their sum. More formally, consider N random Poisson variables $X_1, \ldots, X_N$ with the same parameter λ, i.e $\Pr(X_k = x_k) = e^{-\lambda}\lambda^{x_k}/x_k!, x_k = 0, 1, 2, 3, \ldots$ Then, one can prove that

$$\Pr\left(X_1 = x_1, \ldots, X_N = x_N \middle| \sum_{i=1}^{N} X_i = n\right) = \frac{1}{N^n}\binom{n}{x_1 \cdots x_k \cdots x_N}^{-1},$$

for all $(x_1, \ldots x_N) \in \mathcal{R}_n^N$. The conditional distribution no longer depends on λ anymore (see Bol'shev, 1965; Johnson et al., 1997, p. 65).

Two ways of implementing simple random sampling with replacement are given in Tillé (2006, pp. 60-61). In simple random sampling with replacement, $\pi_k = n/N$ for all $k \in U$, and $\Delta_{k\ell} = -n(N-1)/\{N^2(N-1)\}$, when $k \neq \ell \in U$ and $\Delta_{kk} = n(N-1)/N^2, k \in U$, which gives the variance of the Hanssen-Hurwitz estimator of the total $\text{var}(\widehat{Y}) = N^2\sigma^2/n$. Moreover, we have

$$\frac{\Delta_{k\ell}}{\pi_{k\ell}} = \begin{cases} \dfrac{N-1}{N-1+n} & \text{if } k = \ell \\ -\dfrac{1}{n-1} & \text{if } k \neq \ell. \end{cases}$$

Although it is possible to construct an unbiased estimator of the variance by using expression (2.1), the result obtained is very strange and should not be used (see Tillé, 2006, p. 58). It is nevertheless possible to construct an unbiased estimator by using the quadratic form based on the $D_{k\ell}$ given in expression (2.2)

$$D_{k\ell} = \begin{cases} 1 & \text{if } k = \ell \\ -\dfrac{1}{n-1} & \text{if } k \neq \ell, \end{cases}$$

which gives $\widehat{\text{var}}(\widehat{Y}) = N^2\widehat{\sigma}^2/n$.

2.5 Simple random sampling with over-replacement

Simple random sampling with replacement can be viewed as a conditional distribution of independent Poisson variables. What happens if instead of using the Poisson distribution, we use another discrete distribution? If we use a sequence of geometric random variables given their sum, we obtain another sampling design with replacement and with a fixed sample size. We have called this design simple random sampling with over-replacement because the repetitions of the units are more frequent than in a usual simple random sampling with replacement.

First, consider a sequence of N independent geometric random variables X_k: $\Pr(X_k = x_k) = (1-p)p^{x_k}, x_k = 0,1,2,3,\ldots$ with parameters $\pi_k \in (0,1)$. The sample size $n_s = \sum_{k=1}^{N} X_k$ is random. Let us now calculate the conditional geometric sample design. If S_k denotes the random variable that gives the number of times that unit k is selected in the sample, we have:

$$\begin{aligned}\Pr(S_1 = x_1, \ldots S_N = x_N) &= \Pr\left(X_1 = x_1, \ldots X_N = x_N \left| \sum_{k=1}^{N} X_k = n \right.\right) \\ &= \frac{\prod_{k=1}^{N}(1-p)p^{x_k}}{\sum_{\mathcal{R}_n^N}\prod_{k=1}^{N}(1-p)p^{x_k}} = \frac{q^N p^n}{\sum_{\mathcal{R}_n^N} q^N p^n} = \frac{1}{\mathrm{card}\mathcal{R}_n^N} = \frac{1}{\binom{N+n-1}{n}}.\end{aligned}$$

All the samples have exactly the same probability of being selected. By noting that

$$\#\mathcal{R}_n^N = \binom{N+n-1}{n} \text{ and } \#\mathcal{R}_{n-j}^{N-1} = \binom{N-1+n-j-1}{n-j},$$

we can derive the marginal distribution of S_k:

$$\Pr(S_k = j) = \frac{\binom{N-1+n-j-1}{n-j}}{\binom{N+n-1}{n}}, j = 0, \ldots, n,$$

which is an inverse (or negative) hypergeometric distribution (see Johnson et al., 1992, p.239, 264). We thus have $\mathrm{E}(S_k) = n/N$ and

$$\mathrm{var}(S_k) = \frac{n(N-1)(N+n)}{N^2(N+1)}.$$

This sampling design has a fixed sample size, which implies that $\sum_{k\in U}\mathrm{cov}(S_k, S_\ell) = \mathrm{cov}(n, S_\ell) = 0$. Moreover, since all the units are treated symmetrically, $\mathrm{cov}(S_k, S_\ell) = -\mathrm{var}(S_k)/(N-1)$. The matrix of $\Delta_{k\ell}$ is thus given by

$$\Delta_{k\ell} = \frac{(N-1)(N+n)n}{N^2(N+1)} \times \begin{cases} 1 & \text{if } k = \ell \\ -\frac{1}{N-1} & \text{if } k \neq \ell, \end{cases}$$

which allows us to compute the variance of the Hansen-Hurwitz estimator:

$$\mathrm{var}(\widehat{Y}) = \frac{(N+n)N^2\sigma^2}{(N+1)n}.$$

This variance is much larger than the variance obtained under simple random sampling with replacement.

Simple random sampling with over-replacement can be implemented by a rejective procedure that consists in selecting geometric samples until a sample size n is obtained. Tillé (2006, p. 34) also proposed a general sequential algorithm in order to quickly generate multivariate random variables. This algorithm is based on the computation at each step of the conditional distribution probabilities of the S_k, that is

$$\Pr(S_k = j | S_{k-1}, \ldots, S_1) = \frac{\binom{N-k-1+n_k-j}{n_k-j}}{\binom{N-k+n_k}{n_k}}, j = 0, 1, 2, 3, \ldots, n_k$$

where $n_1 = n$ and

$$n_k = n - \sum_{j=1}^{k-1} S_j, k = 2, \ldots, N.$$

Algorithm 1 is the application of the general algorithm presented in Tillé (2006, p. 34) to sampling with over-replacement. It provides an efficient implementation of sampling with over-replacement.

Algorithm 1 Algorithm for simple random sampling with over-replacement

- For $k = 1, \ldots, N$ unit k is selected S_k times, where

$$\Pr(S_k = j) = \frac{\binom{N-k-1+n_k-j}{n_k-j}}{\binom{N-k+n_k}{n_k}}, j = 0, 1, 2, 3, \ldots, \ldots, n_k.$$

2.6 Discussion

Table 2.1 shows the variances of the three sampling designs. Compared to simple random sampling with replacement, we find that simple random sampling without replacement and simple random sampling with over-replacement have a symmetric position. Indeed, for random sampling without replacement, the finite population correction factor is $(N-n)/(N-1)$ and for simple random sampling with over-replacement, the over-replacement correction factor is $(N+n)/(N+1)$.

Simple random sampling with over-replacement is interesting because it shows that there are several methods of sampling with replacement that have an equal inclusion expectation in the sample. It is also possible to

Table 2.1 – *Comparison of the variance of the three simple designs*

Sampling design	variance of the estimator of the total
Simple without replacement	$\frac{(N-n)N^2\sigma^2}{(N-1)n}$
Simple with replacement	$\frac{N^2\sigma^2}{n}$
Simple with over-replacement	$\frac{(N+n)N^2\sigma^2}{(N+1)n}$

define a large range of simple random samplings by combining several simple random sampling designs. For instance, one can select a subset of observations by simple random sampling with replacement and a second subset by simple random sampling with over-replacement. So, a large range of sampling designs with replacement can be defined with different variances of the estimator of the total. Antal & Tillé (2011a) have used simple random sampling with over-replacement to construct new bootstrap methods for complex sampling designs. The main idea consists of mixing simple random sampling with over-replacement with other sampling designs in order to construct ad hoc resampling designs to reproduce the correct estimator of the variance in a complex sampling design. Sampling with over-replacement is thus not only a simple mathematical curiosity but can be used in practical applications.

3

A Direct Bootstrap Method for Complex Sampling Designs from a Finite Population

Abstract

In complex designs, classical bootstrap methods result in a biased variance estimator when the sampling design is not taken into account. Resampled units are usually rescaled or weighted in order to achieve unbiasedness in the linear case. In the present article, we propose novel resampling methods that may be directly applied to variance estimation. These methods consist of selecting subsamples under a completely different sampling scheme from that one used to generate the original sample composed of several sampling designs. In particular, a portion of the subsampled units is selected without replacement, while another is selected with replacement, thereby adjusting for the finite population setting. We show that these bootstrap estimators directly - and precisely - reproduce unbiased estimators of the variance in the linear case in a time-efficient manner, and eliminate the need for classical adjustment methods such as rescaling, correction factors, or artificial populations. Moreover, we show via simulation studies that our method is at least as efficient as those currently existing, which call for additional adjustment. This methodology can be applied to classical sampling designs, including simple random sampling with and without replacement, Poisson sampling, and unequal probability sampling with and without replacement. [1]

Keywords: survey sampling, one-one design, poisson sampling, resampling method

[1]This chapter is a reprint of: Antal, E. and Tillé, Y. (2011). A Direct Bootstrap Method for Complex Sampling Designs from a Finite Population *Journal of the American Statistical Association* **106**, 534–543.

3.1 Introduction

Resampling methods such as the bootstrap and jackknife are largely used to estimate variances across a broad spectrum of statistical contexts. In survey sampling, the variances of even simple estimators depend on the sampling design, and can take very complex forms, particularly when the sampling design is elaborate. The classical bootstrap method, developed by Efron (1979) cannot be directly applied to cases of sampling from a finite population because the identical and independent distribution assumption fails under sampling without replacement. Gross (1980) and Chao & Lo (1985) have proposed a method for variance estimation based on reconstructing artificial populations from the sample. Bootstrap samples are then selected from this artificial population using the original sampling scheme. Another important class of methods arises from the rescaled bootstrap (Rao & Wu, 1988) which consists of modifying the sample values of the variable of interest to construct an unbiased estimator of the variance in the linear case. Other methods have also been proposed by Mac Carthy & Snowden (1985); Kuk (1989); Rao et al. (1992); Shao & Tu (1995); Sitter (1992a,b); Booth et al. (1994); Holmberg (1998).

In this article, we propose a new methodology that can be applied to classical sampling designs both with and without replacement, as well as both equal and unequal inclusion probabilities. Our methodology consists of selecting bootstrap samples from the original sample in such a way that it eliminates the need for scaling, weighting of the sample, and using artificial populations. We argue that if the aim is variance estimation, the resampling design must be radically different from that which generates the original data. We then proceed to construct an *ad hoc* resampling design by mixing several designs, such that the bootstrap variance is equal to the estimator of the variance in the linear case, and such that the bootstrap sample has the same expected sample size as that of the actual sample size of the data, and can thus be treated as the original sample. This feature is particularly attractive because imputation, weighting for nonresponse and calibration can thus be carried out without the need for any additional considerations or corrective techniques. In sampling without replacement, the main idea consists in selecting bootstrap samples by mixing sampling with and without replacement in order to reproduce a variance estimator that comprises the finite population correction.

The remainder of the article is structured as follows. We will first review basic notions of the theory of survey sampling, and provide an overview of the most frequently used sampling designs. We will then introduce two new sampling designs, simple random sampling with over-

replacement and one-one resampling, that are used exclusively in resampling. We will then establish sufficient conditions for a direct unbiased estimator for the variance of the total in resampling designs, and provide the construction of the algorithms used to draw such samples for several basic sampling designs. Finally, we supplement the theoretical proofs with results of simulation studies performed on several functions of interest, including the total, the median, the Gini index, and the ratio of totals. These results are compared to those obtained under resampling methods currently used, such as the classical bootstrap with and without replacement. We conclude with comparative remarks on our proposed methodology, and propose additional development and future research on the topic.

3.2 Sampling design and estimation

Consider the finite population $U = \{1, \ldots, k, \ldots, N\}$ and the variable of interest y that takes the value y_k on unit k, for all k in U. A first aim is to estimate the total of the interest variable: $Y = \sum_{k \in U} y_k$. A random sample is a random vector $\mathbf{S} = (S_1, \ldots, S_k, \ldots, S_N)'$, where S_k is the number of times unit k is selected in the sample. If the sample is selected without replacement, then S_k can only take the values 0 and 1. If the sample has a fixed sample size n, then $\sum_{k \in U} S_k = n$.

Let π_k be the expectation of S_k, that is, $\pi_k = \mathrm{E}(S_k)$. The joint expectation of two units k and ℓ is $\pi_{k\ell} = \mathrm{E}(S_k S_\ell)$. Moreover, $\Delta_{k\ell} = \mathrm{cov}(S_k, S_\ell) = \pi_{k\ell} - \pi_k \pi_\ell$. If the sample is selected without replacement, π_k is the inclusion probability of unit k and $\pi_{k\ell}$ is the joint inclusion probability of unit k and ℓ.

If $\pi_k > 0$, for all $k \in U$, then the total Y can be estimated in an unbiased manner by using the Horvitz-Thompson estimator $\widehat{Y} = \sum_{k \in U} S_k y_k / \pi_k$. The variance of $\widehat{Y}$ is

$$\mathrm{var}(\widehat{Y}) = \sum_{k \in U} \sum_{\ell \in U} \frac{y_k y_\ell}{\pi_k \pi_\ell} \Delta_{k\ell}. \tag{3.1}$$

Theoretically, if $\pi_{k\ell} > 0$, for all $k \neq \ell \in U$, this variance can be estimated in an unbiased manner by

$$\widehat{\mathrm{var}}(\widehat{Y}) = \sum_{k \in U} \sum_{\ell \in U} \frac{S_k S_\ell y_k y_\ell}{\pi_k \pi_\ell} \frac{\Delta_{k\ell}}{\pi_{k\ell}}. \tag{3.2}$$

Nevertheless, this variance estimator is often very unstable. It can even take negative values. When the sampling design has a fixed sample size,

then the variance can be written

$$\text{var}(\widehat{Y}) = \frac{-1}{2} \sum_{k \in U} \sum_{\ell \in U} \left(\frac{y_k}{\pi_k} - \frac{y_\ell}{\pi_\ell} \right)^2 \Delta_{k\ell},$$

and, if $\pi_{k\ell} > 0$, for all $k \neq \ell \in U$, can be estimated by the Yates-Grundy estimator of variance:

$$\widehat{\text{var}}(\widehat{Y}) = \frac{-1}{2} \sum_{k \in U} \sum_{\ell \in U} S_k S_\ell \left(\frac{y_k}{\pi_k} - \frac{y_\ell}{\pi_\ell} \right)^2 \frac{\Delta_{k\ell}}{\pi_{k\ell}}. \tag{3.3}$$

Expression (3.3) holds for sampling with or without replacement and can also be written in the quadratic form

$$\widehat{\text{var}}_D(\widehat{Y}) = \sum_{k \in U} \sum_{\ell \in U} \frac{S_k S_\ell y_k y_\ell}{\pi_k \pi_\ell} D_{k\ell}, \tag{3.4}$$

with

$$D_{k\ell} = \begin{cases} -\sum\limits_{\substack{j \in U \\ j \neq k}} S_j \dfrac{\Delta_{kj}}{\pi_{kj}} & \text{if } k = \ell \\ \dfrac{\Delta_{k\ell}}{\pi_{k\ell}} & \text{if } k \neq \ell. \end{cases} \tag{3.5}$$

When the sampling design with or without replacement has a fixed sample size, estimator (3.3) must be preferred to estimator (3.2). We shall show below in Result 3.2 that the presentation of estimator (3.3) in a quadratic form is needed to construct a resampling method that produces an unbiased estimator.

3.3 Basic sampling designs

In a *Poisson sampling design* with inclusion probabilities π_k, the S_k are N independent Bernoulli random variables with parameter π_k. Thus, $\Delta_{k\ell} = \pi_k(1 - \pi_k)$ if $k = \ell$ and 0 otherwise. So $\Delta_{k\ell}/\pi_{k\ell} = 1 - \pi_k$ if $k = \ell$ and 0 otherwise.

Simple random sampling with replacement is very common. The sampling design is given by $\Pr(\mathbf{S} = \mathbf{s}) = N^{-n} \binom{n}{s_1 \cdots s_k \cdots s_N}^{-1}$, for all $\mathbf{s} \in \mathcal{R}_n$, where $\mathcal{R}_n = \left\{ \mathbf{s} \in \mathbb{N}^n \,\middle|\, \sum_{k=1}^N s_k = n \right\}$. It follows that $\Delta_{k\ell} = -n(N-1)/\{N^2(N-1)\}$ when $k \neq \ell \in U$ and $\Delta_{kk} = n(N-1)/N^2$ when $k \in U$. Since the sample size is fixed, we can construct an unbiased estimator by using the quadratic form based on the $D_{k\ell}$, defined in Expression (3.5), $D_{k\ell} =$

$-1/(n-1)$ when $k \neq \ell \in U$, and $D_{kk} = 1$ when $k \in U$, which gives

$$\widehat{\text{var}}(\widehat{Y}) = \frac{N^2}{n}\frac{1}{n-1}\sum_{k\in U} S_k(y_k - \widehat{\overline{Y}})^2, \tag{3.6}$$

where $\widehat{\overline{Y}} = n^{-1}\sum_{k\in U} S_k y_k$.

Unequal probability with replacement with fixed sample size, is a generalization of simple random sampling with replacement to unequal probabilities of selection. The distribution of this sampling design is multinomial:

$$\Pr(\mathbf{S} = \mathbf{s}) = \binom{n}{s_1 \cdots s_k \cdots s_N}^{-1} \prod_{k\in U}\left(\frac{\pi_k}{n}\right)^{s_k}, \text{ for all } \mathbf{s} \in \mathcal{R}_n.$$

In unequal probability sampling with replacement,

$$\Delta_{k\ell} = \frac{n(N-1)}{N^2} \times \begin{cases} \pi_k\left(1 - \dfrac{\pi_k}{n}\right) & \text{if } k = \ell \\ -\dfrac{\pi_k\pi_\ell}{n} & \text{if } k \neq \ell. \end{cases}$$

In order to construct an unbiased estimator of the variance, we can use the $D_{k\ell}$ defined in Expression (3.5), and we get $D_{k\ell} = -1/(n-1)$ when $k \neq \ell \in U$, and $D_{kk} = 1$ when $k \in U$. Curiously, $D_{k\ell}$ does not depend on the π_k's of the sampling design and are the same as in simple random sampling with replacement. The unbiased variance estimator (3.3) becomes

$$\widehat{\text{var}}(\widehat{Y}) = \frac{n}{n-1}\sum_{k\in U} S_k\left(\frac{y_k}{\pi_k} - \frac{\widehat{Y}}{n}\right)^2. \tag{3.7}$$

Simple random sampling without replacement is defined by the following sampling design $\Pr(\mathbf{S} = \mathbf{s}) = \binom{N}{n}^{-1}$, for all $\mathbf{s} \in \mathcal{S}_n$, where

$$\mathcal{S}_n = \left\{ \mathbf{s} \in \{0,1\}^N \,\middle|\, \sum_{k=1}^N s_k = n \right\}.$$

We thus have $\Delta_{k\ell} = -n(N-n)/\{N^2(N-1)\}$ when $k \neq \ell \in U$, $\Delta_{kk} = n(N-n)/N^2$ when $k \in U$, $\Delta_{k\ell}/\pi_{k\ell} = -(N-n)/\{N(n-1)\}$ when $k \neq \ell \in U$, and $\Delta_{kk}/\pi_{kk} = (N-n)/N$ when $k \in U$.

Unequal probability sampling without replacement and with fixed sample size is much more complex. The first problem is that there are many methods of sampling without replacement and with unequal probabilities. Each method provides a specific matrix of joint inclusion probabilities. These inclusion probabilities are, however, very similar if the sampling has a large entropy (Berger, 1998; Brewer & Donadio, 2003; Henderson,

2006), such as the random systematic design (Madow, 1949) or the Rao-Sampford design (Rao, 1965; Sampford, 1967), the Brewer design (Brewer, 1975), the maximum entropy design or the random pivotal design (Tillé, 2006, p. 79-95 and p.106). The second problem is that these inclusion probabilities can never be simplified. So, a simpler expression of variance than (3.1) and its estimator (3.2) cannot be constructed. Several approximations of variance based on a simple sum have been proposed, however. These approximations are obviously biased, but simulations have shown that they have smaller mean squared errors than estimators (3.2) and (3.4) (Hájek, 1981; Matei & Tillé, 2005). There are thus various ways to estimate the variance. The strictly unbiased estimator consists of computing the $D_{k\ell}$ by expression (3.5). A general biased and simple estimator of variance is given by

$$\widehat{\text{var}}(\widehat{Y}) = \sum_{k\in S} c_k \left(\frac{y_k}{\pi_k} - \frac{\sum_{k\in S} c_k y_k/\pi_k}{\sum_{k\in S} c_k} \right)^2,$$

where the c_k are weights that we discuss later. This expression can be viewed as an approximation of the $D_{k\ell}$ given in expression (3.5), by

$$\widetilde{D}_{k\ell} = \begin{cases} c_k - \dfrac{c_k^2}{\sum_{j\in U} S_j c_j} & \text{if } k = \ell \\ -\dfrac{c_k c_\ell}{\sum_{j\in U} S_j c_j} & \text{if } k \neq \ell. \end{cases}$$

Diverse values have been proposed for the weights c_k.

1. A simple value was given by Hájek (1981), who proposed to use

$$c_{k1} = \frac{n}{n-1}(1-\pi_k). \tag{3.8}$$

2. Deville & Tillé (2005) proposed weights c_{k2} such that

$$c_{k2} - \frac{c_{k2}^2}{\sum_{j\in U} S_j c_{j2}} = 1 - \pi_k. \tag{3.9}$$

   In this case, the diagonal elements $\widetilde{D}_{kk2}$ of the approximated matrix are equal to $1-\pi_k$. A solution does not always exist for this equation, for instance when $n = 2$.

3. One could also take the weights c_{k3} such that

$$c_{k3} - \frac{c_{k3}^2}{\sum_{j\in U} S_j c_{j3}} = -\sum_{\substack{j\in U \\ j\neq k}} S_j \frac{\Delta_{kj}}{\pi_{kj}}, \tag{3.10}$$

but this requires to solve a nonlinear system of equations. In this case, the diagonal elements of the approximated matrix are

$$\widetilde{D}_{kk3} = -\sum_{\substack{j\in U\\ j\neq k}} S_j \frac{\Delta_{kj}}{\pi_{kj}}$$

and correspond to the diagonal of the matrix used for the Yates-Grundy estimator of variance given in (3.5).

Simple random sampling with over-replacement was recently proposed by Antal & Tillé (2011b). The sampling design is defined by $\Pr(S_1 = x_1, \ldots, S_N = x_N) = (\text{card}\mathcal{R}_n)^{-1} = \binom{N+n-1}{n}^{-1}$. The $\binom{N+n-1}{n}$ samples with replacement have exactly the same probability of being selected. The marginal distribution of S_k is given by

$$\Pr(S_k = j) = \binom{N+n-1}{n}^{-1}\binom{N-1+n-j-1}{n-j}, j = 0, \ldots, n,$$

which is an inverse hypergeometric distribution.The expectation is $\mathrm{E}(S_k) = n/N$, and the matrix of $\Delta_{k\ell}$ is given by

$$\Delta_{k\ell} = \frac{(N-1)(N+n)n}{N^2(N+1)} \times \begin{cases} 1 & \text{if } k = \ell \\ -\frac{1}{N-1} & \text{if } k \neq \ell. \end{cases}$$

This design has a larger variance than sampling with replacement and will be used to define a new resampling method.

3.4 Resampling and sufficient conditions

Define the random set S that contains the list of labels for the units selected in the sample $\mathbf{S}$. If a unit is selected several times in the sample, the labels can appear several times in S. For instance, if from population $U = \{1,2,3,4,5,6\}$, we select a sample $\mathbf{S}$ that takes the value $(0,2,1,0,3,1)$, the set S takes the value $\{2,2,3,5,5,5,6\}$. A resampling method is a second stage on sampling from sample S. A subsample $\mathbf{S}^* = (S_k^*, k \in S)$ can thus be presented as a sequence of discrete nonnegative random variables S_k^* that denote the number of times unit k is resampled. For example, if, in the above example, $\mathbf{S}^*$ takes the values $(1,0,3,0,2,0,1)$, then the subsample set S^* will be $\{2,3,3,3,5,5,6\}$. The S_k^* are generally not independent. A correlation is indeed necessary to obtain an unbiased estimation of the variance when the sample size is fixed. The resampling sample size is denoted by n^*.

In fact, a resampling method is a second phase of sampling that can depend on the first phase. Let $\mathrm{E}^*(.) = \mathrm{E}(.|S)$, $\mathrm{var}^*(.) = \mathrm{var}(.|S)$ and $\mathrm{cov}^*(.,.) = \mathrm{cov}(.,.|S)$ denote, respectively, the conditional expectation, variance and covariance under the resampling design with respect to the original design. Moreover, let $\Pr^*(.) = \Pr(.|S)$ denote the probability under the resampling design given the original design. Let $\alpha_k = \mathrm{E}^*(S_k^*)$, $\alpha_{k\ell} = \mathrm{E}^*(S_k^* S_\ell^*)$ and $\mathrm{cov}^*(S_k^*, S_\ell^*) = \Sigma_{k\ell} = \alpha_{k\ell} - \alpha_k \alpha_\ell$. The resampled estimator of the total is defined as $\widehat{Y}^* = \sum_{k\in S} y_k S_k^* / \pi_k$. This estimator is generally biased for $\widehat{Y}$ given S since its conditional expectation is

$$\mathrm{E}^*(\widehat{Y}^*) = \sum_{k\in S} \frac{y_k \mathrm{E}^*(S_k^*)}{\pi_k} = \sum_{k\in S} \frac{y_k \alpha_k}{\pi_k}. \tag{3.11}$$

Note that α_k can depend on S. If $\alpha_k = 1$, then the estimator is unbiased.

The conditional variance of the resampled estimator is

$$\mathrm{var}^*(\widehat{Y}^*) = \sum_{k\in S} \sum_{\ell\in S} \frac{y_k y_\ell}{\pi_k \pi_\ell} \Sigma_{k\ell}. \tag{3.12}$$

This directly leads to two fundamental results:

Result 3.1 *A sufficient condition for* $\mathrm{E}^*(\widehat{Y}^*) = \widehat{Y}$ *is* $\alpha_k = 1$, *for all* $k \in U$.

This result directly comes from the equality between Expression (3.11) and the Horvitz-Thompson estimator.

Result 3.2 *A sufficient condition for* $\mathrm{var}^*(\widehat{Y}^*) = \widehat{\mathrm{var}}(\widehat{Y})$, *is* $\Sigma_{k\ell} = \Delta_{k\ell}/\pi_{k\ell}$, *for all* $k, \ell \in U$ *and a sufficient condition for* $\mathrm{var}^*(\widehat{Y}^*) = \widehat{\mathrm{var}}_D(\widehat{Y})$, *is* $\Sigma_{k\ell} = D_{k\ell}$, *for all* $k, \ell \in U$ *if the sample size is fixed.*

This result directly comes from the equality between Expressions (3.2) and (3.12) or between Expressions (3.4) and (3.12) when the sample size of S is fixed.

In fact, the main idea of this paper is to develop resampling methods that satisfy conditions given in Results 3.1 and 3.2. This idea leads us to choose a sampling design for S^* that is completely different from the sampling design used for S. Indeed, Result 3.2 is generally not satisfied by using the same sampling design for S and S^*, because $\Sigma_{k\ell}$ and $D_{k\ell}$ are of very different natures: the $\Sigma_{k\ell}$ are variances and covariances, but the $D_{k\ell}$ are not.

Let $\widehat{\theta}$ be an estimator of a function of interest θ. Estimator $\widehat{\theta}$ is a function of the observed data $\{(y_k, \pi_k), k \in S\}$. The bootstrap estimator $\widehat{\theta}^*$ is the same function as $\widehat{\theta}$, applied on the bootstrap data $\{(y_k, \pi_k), k \in S^*\}$. In practice, a sequence of bootstrap samples $\mathbf{S}^{*1}, \ldots, \mathbf{S}^{*m}$ is selected with the

bootstrap design. The bootstrap variance given in (3.12) is approximated by

$$\widetilde{\text{var}}^*(\widehat{\theta}^*) = \frac{1}{m-1}\sum_{j=1}^{m}\left(\widehat{\theta}_j^* - \overline{\widehat{\theta}}\right)^2,$$

where $\widehat{\theta}_j^*$ is the bootstrap estimator computed on the jth bootstrap sample and $\overline{\widehat{\theta}} = (1/m)\sum_{j=1}^{m}\widehat{\theta}_j^*$.

3.5 The simplest example: resampling from a Poisson sample

In a Poisson sampling design, $\Delta_{k\ell}/\pi_{k\ell} = 0$ when $k \neq \ell \in U$, and $\Delta_{kk}/\pi_{kk} = 1 - \pi_k$, $k \in U$. The resampling design must be such that $\mathrm{E}^*(S_k^*) = 1$, $\text{var}^*(S_k^*) = 1 - \pi_k$, and $\text{cov}^*(S_k^*, S_\ell^*) = 0$, for all $k \neq \ell$. Algorithm 2 can be used to generate such S_k^*'s. The main idea consists of selecting a part of the units without replacement and the other part with replacement by generating the counts from Poisson random variables, in order to reproduce the finite population correction $1 - \pi_k$.

Algorithm 2 Resampling procedure for Poisson sampling

Define, independently for $k \in S$:

- S_{kA}^* is a Bernoulli random variable with parameter π_k.
- If $S_{kA}^* = 1$ then $S_{kB}^* = 0$;
  otherwise S_{kB}^* is a Poisson random variable with parameter $\lambda = 1$.
- The resampling design is $S_k^* = S_{kA}^* + S_{kB}^*$.

With Algorithm 2, the expectations, variances and covariances of S_k^* can be computed easily $\mathrm{E}^*(S_k^*) = \mathrm{E}^*(S_{kA}^*) + \mathrm{E}^*(S_{kB}^*) = \pi_k + 1 \times (1 - \pi_k) = 1$. Moreover, $\text{var}^*(S_k^*) = \mathrm{E}^*\left[\text{var}^*(S_k^*|S_{kA}^*)\right] + \text{var}^*\left[\mathrm{E}^*(S_k^*|S_{kA}^*)\right] = 1 - \pi_k$. This bootstrap method provides the exact Horvitz-Thompson estimator in the linear case. Indeed, $\text{var}^*(\widehat{Y}^*) = \text{var}^*\left(\sum_{k\in S} y_k S_k^*/\pi_k\right) = \sum_{k\in S} y_k^2(1 - \pi_k)/\pi_k^2 = \widehat{\text{var}}(\widehat{Y})$.

3.6 The one-one resampling design

The one-one design is a sampling design defined only for resampling. It is an *ad hoc* construction used to randomly select n units from a sample of size n in such a way that the expectation and the variance of S_k^* are equal to 1, that is, $\mathrm{E}^*(S_k^*) = 1$ and $\text{var}^*(S_k^*) = 1$. This sampling design

is a mixture between a simple random sampling with replacement and a simple random sampling with over-replacement. Its implementation is given in Algorithm 3.

Algorithm 3 The one-one resampling design

:

- If $n = 2$, then

$$S_1^* = \begin{cases} 0 \text{ with probability } 1/2 \\ 2 \text{ with probability } 1/2 \end{cases}, \text{ and } S_2^* = 2 - S_1^*.$$

- If $n \geq 3$, then
  - Compute:

$$m = \left\lfloor \frac{1}{2}\left(1 + \sqrt{\frac{4n^2 + 5n - 1}{n - 1}}\right)\right\rfloor, \tag{3.13}$$

    where $\lfloor x \rfloor$ is the largest integer less than or equal to x and

$$\alpha = \frac{m(n-1)(m+1) - n(n+1)}{2m(n-1)}. \tag{3.14}$$

  - Define the random variable

$$\tilde{n} = \begin{cases} m & \text{with a probability } \alpha \\ m+1 & \text{with a probability } 1 - \alpha. \end{cases}$$

  - Select a simple random sample with overreplacement with sample size $\tilde{n}$ from S. This sample is denoted by S_{kA}^*.
  - Select a simple random sample with replacement with sample size $n - \tilde{n}$ from S. This sample is denoted by S_{kB}^*. This second sample is independent from the first one.
  - The final sample is $S_k^* = S_{kA}^* + S_{kB}^*, k \in S$.

Result 3.3 *If S_k^* is the number of times unit k is selected by the one-one resampling design described in Algorithm 3, then $\mathrm{E}^*(S_k^*) = 1, \mathrm{var}^*(S_k^*) = 1$ and $\mathrm{cov}^*(S_k^*, S_\ell^*) = -1/(n-1)$, for all $k \neq \ell$.*

Proof

The case where $n = 2$ is obvious. For the case where $n \geq 3$, we have that $\mathrm{E}^*(S_k^*|\tilde{n}) = \mathrm{E}^*(S_{kA}^*|\tilde{n}) + \mathrm{E}^*(S_{kB}^*|\tilde{n}) = \tilde{n}/n + (n - \tilde{n})/n = 1$. Thus

$\mathrm{E}^*(S_k^*) = \mathrm{E}^*\mathrm{E}^*(S_k^*|\tilde{n}) = 1$. Moreover,

$$\begin{aligned}\mathrm{var}^*(S_k^*|\tilde{n}) &= \mathrm{var}^*(S_{kA}^*|\tilde{n}) + \mathrm{var}^*(S_{kB}^*|\tilde{n}) \\ &= \frac{(n-1)(n+\tilde{n})\tilde{n}}{n^2(n+1)} + \frac{(n-\tilde{n})(n-1)}{n^2} = \frac{n-1}{n^2}\left[\frac{(n+\tilde{n})\tilde{n} + (n+1)(n-\tilde{n})}{(n+1)}\right].\end{aligned}$$

Since $\mathrm{E}^*(S_k^*|\tilde{n}) = 1$,

$$\begin{aligned}\mathrm{var}^*(S_k^*) &= \mathrm{E}^*\mathrm{var}^*(S_k^*|\tilde{n}) \\ &= \alpha\frac{n-1}{n^2}\left[\frac{(n+m)m + (n+1)(n-m)}{(n+1)}\right] \\ &\quad +(1-\alpha)\frac{n-1}{n^2}\left[\frac{(n+m+1)(m+1) + (n+1)(n-(m+1))}{(n+1)}\right] \\ &= \frac{(n-1)[n+n^2+m(1-2\alpha+m)]}{n^2(1+n)}. \qquad (3.15)\end{aligned}$$

By plugging the value of α given in (3.14) and the value of m given in (3.13) in Expression (3.15), we get $\mathrm{var}^*(S_k^*) = 1$. This sampling design has a fixed sample size, which implies that $\sum_{k\in S}\mathrm{cov}^*(S_k^*, S_\ell^*) = \mathrm{cov}^*(n, S_\ell^*) = 0$. Moreover, since all the units are treated symmetrically, $\mathrm{cov}^*(S_k^*, S_\ell^*) = -\mathrm{var}^*(S_k^*)/(n-1)$. We thus have $\Sigma_{k\ell} = -1/(n-1)$ when $k \neq \ell \in U$ and $\Sigma_{kk} = 1$ for $k \in U$.

□

3.7 Resampling from a simple random sample with replacement

3.7.1 The usual bootstrap with replacement

If the sample S is selected by means of simple random sampling with replacement, the formula of the estimated variance of the total estimator is already given in Expression (3.6). The usual bootstrap consists of selecting a sample from S with the same sampling design, that is, a simple random sampling design with replacement from S. In this case, the variance of the resampled estimator is $\mathrm{var}^*(\widehat{Y}^*) = (N^2/n^2)\sum_{k\in S}(y_k - \overline{Y})^2$. The bootstrap variance slightly underestimates the unbiased estimator given in Expression (3.6). Indeed, $\widehat{\mathrm{var}}(\widehat{Y}) = n/(n-1) \times \mathrm{var}^*(\widehat{Y}^*)$. Actually, this underestimation is not very important if the sample size is large but can create problems if the samples are selected in strata with small sample sizes. Obviously, a correction factor can be applied in each stratum, but these procedures require a particular treatment of the bootstrap sample in each stratum.

3.7.2 Bootstrap by using the one-one sampling design

The one-one simple random sampling design allows us to avoid the use of correction factors for the variance. Indeed, if the bootstrap sample is selected by a one-one design then the bootstrap variance is $\text{var}^*(\widehat{Y}^*) = \left[N^2/\{n(n-1)\}\right] \sum_{k\in S}(y_k - \overline{Y})^2$. In a one-one simple random sampling, the repetition of the units is slightly larger than with simple random sampling with replacement, which increases the variance by a factor of $n/(n-1)$. It is thus no longer necessary to multiply the bootstrap variance by this factor.

3.8 Resampling from a sample selected with unequal probabilities with replacement

3.8.1 The usual bootstrap with replacement

If the sample is selected with unequal probabilities, with replacement and with fixed sample size, the estimator of variance is given in (3.7). In this case, the matrix of $D_{k\ell}$ given in (3.5) does not depend on the π_k's, which means that the resampling design must be done with equal selection probabilities. A usual design consists of resampling by means of simple random sampling with replacement, which gives the bootstrap variance

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

With simple random sampling with replacement, the bootstrap variance thus suffers from a small underestimation. This problem can be annoying when the sample size is small and can be fixed by using a one-one simple random sampling.

3.8.2 Bootstrap by using the one-one sampling design

If the bootstrap samples are selected with a one-one design, the bootstrap variance becomes

$$\text{var}^*(\widehat{Y}^*) = \frac{n}{n-1}\sum_{k\in S}\left(\frac{y_k}{\pi_k} - \frac{\widehat{Y}}{n}\right)^2,$$

and is exactly equal to the estimator of variance (3.7). The one-one design is thus a convenient design for resampling from a sample selected with

unequal probabilities with replacement, particularly when the sample size is small.

3.9 Resampling from a simple random sample selected without replacement

3.9.1 Resampling using simple random sampling with replacement

In simple random sampling without replacement, the estimator of variance is

$$\widehat{\text{var}}(\widehat{Y}) = \frac{N^2(N-n)}{nN}\frac{1}{n-1}\sum_{k\in S}(y_k - \widehat{\overline{Y}})^2. \tag{3.16}$$

A simple way of resampling consists of using a simple random sampling with replacement as a resampling design. In this case, $\text{var}^*(\widehat{Y}^*) = (N^2/n^2)\sum_{k\in S}(y_k - \overline{Y})^2$. Obviously, the bootstrap variance is not equal to the variance estimator. Indeed, $\widehat{\text{var}}(\widehat{Y}) = \text{var}^*(\widehat{Y}^*)(N-n)n/\{N(n-1)\}$, which means that the resampling variance does not take into account the loss of one degree of freedom and the finite population correction. The bootstrap variance must be corrected by a factor. This correction can become intricate if a large number of samples are selected in strata.

3.9.2 Resampling using a with replacement and a one-one design

In order to avoid the use of a correction factor, one can use a mixture of a simple sampling without replacement and a one-one design as described in Algorithm 4 in order to directly reproduce the unbiased estimator of variance for the totals.

Result 3.4 gives the properties of Algorithm 4.

Result 3.4 *If Algorithm 4 is used for the resampling design, (i)* $\text{E}^*(S_k^*) = 1$, *(ii)* $\text{var}^*(S_k^*) = (N-n)/N$, *(iii)* $\text{cov}^*(S_k^*, S_\ell^*) = -(N-n)/\{N(n-1)\}$.

Proof

The case where $n - n^2/N < 2$ is trivial. For the case where $n - n^2/N \geq 2$, the expectation is given by: $\text{E}^*(S_k^*) = \text{E}^*(S_{kA}^*) + \text{E}^*(S_{kB}^*|S_{kA}^*)\text{Pr}^*(S_{kA} = 0) = n/N + 1 \times (1 - n/N) = 1$. Next, the variance is $\text{var}^*(S_k^*) = \text{E}^*\left[\text{var}^*(S_k^*|S_{kA}^*)\right] + \text{var}^*\left[\text{E}^*(S_k^*|S_{kA}^*)\right] = \text{var}^*(S_{kB}^*|S_{kA}^* = 0)\text{Pr}^*(S_{kA}^* = 0) = 1 - n/N$. Finally, the covariances can be derived from the symmetry of treatment of the units, which implies that $\text{cov}^*(S_k^*|S_{kA}^*, S_{\ell A}^*) = -\text{var}^*(S_k^*)/(n-1) = -(N-n)/\{N(n-1)\}$. □

Algorithm 4 Resampling using a with replacement and a one-one design

- If $n - n^2/N < 2$:
  - With a probability $q = n(N-n)/(2N)$, select randomly without replacement and with equal probabilities two units in S denoted by i and j. Next define $S_i^* = 2, S_j^* = 0, S_k = 1$, for all $k \notin \{i, j\}$.
  - With a probability $1 - q$, $S_k^* = 1$, for all $k \in S$.
- If $n - n^2/N \geq 2$:
  - Define

$$m = \begin{cases} \left\lfloor \frac{n^2}{N} \right\rfloor & \text{with probability } q \\ \left\lfloor \frac{n^2}{N} \right\rfloor + 1 & \text{with probability } 1 - q, \end{cases}$$

    where $q = \lfloor n^2/N \rfloor + 1 - n^2/N$.
  - Select a sample S_{kA}^* from S with simple random sampling design without replacement with a sample size m.
  - From the set of units of S such that $S_{kA}^* = 0$, select a sample S_{kB}^* according to a one-one design, so S_{kB}^* has size $n - m$.
  - The resampling design is $S_k^* = S_{kA}^* + S_{kB}^*$.

If Algorithm 4 is used, the resampling variance is thus

$$\text{var}(\widehat{Y}^*) = N^2 \frac{N-n}{nN} \frac{1}{n-1} \sum_{k \in S} (y_k - \widehat{\overline{Y}})^2,$$

and is exactly equal to the estimator of variance given in Expression (3.16).

3.10 Resampling from a sample selected with unequal probabilities without replacement

Unequal probability without replacement is obviously a more complicated problem. The main reason is that the unbiased estimators given in (3.2) and (3.4) of the variance can never been simplified, which makes it necessary to compute all the joint inclusion probabilities to estimate the variance. When the entropy of the sampling design is large, biased estimators given in (3.8), (3.9) and (3.10) have smaller mean square errors than estimators (3.2) and (3.4) (see Matei & Tillé, 2005). For this reason, we do not propose using a bootstrap method that exactly reproduces the estimator of

variance, but rather one that gives one of the three approximations. These methods are described in Algorithms 5 and 6.

Algorithm 5 Resampling for unequal probability sampling without replacement: Case 1

Case 1: $n - \sum_{k \in S} \phi_k \geq 2$.

- Select a sample S^*_{kA} without replacement with unequal inclusion probabilities ϕ_k (the choice of ϕ_k is discussed below) and fixed sample size. This sampling design is the same as the original design. If $n^* = \sum_{k \in S} \phi_k$ is not an integer, then define

$$m = \begin{cases} m_1 = \lfloor n^* \rfloor & \text{with probability } q \\ m_2 = \lfloor n^* \rfloor + 1 & \text{with probability } 1 - q, \end{cases}$$

  where $q = \lfloor n^* \rfloor + 1 - n^*$. Also define ϕ_{k1} and ϕ_{k2} as the inclusion probabilities such that

$$\sum_{k \in S} \phi_{k1} = m_1, \quad \sum_{k \in S} \phi_{k2} = m_2, \quad q\phi_{k1} + (1-q)\phi_{k2} = \phi_k, \text{ for all } k \in S.$$

  Let $\phi_{k\ell 1}$ and $\phi_{k\ell 2}$ also be the joint inclusion probabilities of the design where sample sizes m_1 or m_2 were selected.

- From the set of units of S such that $S^*_{kA} = 0$, select a sample S^*_{kB} according to a one-one design.

- The resampling design is $S^*_k = S^*_{kA} + S^*_{kB}$.

Result 3.5 gives the properties of Algorithm 5.

Result 3.5 *If Algorithm 5 is used for the resampling design, (i)* $\mathrm{E}^*(S^*_k) = 1$, *(ii)* $\mathrm{var}^*(S^*_k) = 1 - \phi_k$, *(iii)* $\mathrm{cov}^*(S^*_k, S^*_\ell) = -q(1 - \phi_{k1} - \phi_{\ell 1} + \phi_{k\ell 1})/(n - m_1 - 1) - (1 - q)(1 - \phi_{k2} - \phi_{\ell 2} + \phi_{k\ell 2})/(n - m_2 - 1)$.

Proof

First, the conditional expectation is given by $\mathrm{E}^*(S^*_k|m_j) = \mathrm{E}^*(S^*_{kA}|m_j) + \mathrm{E}^*(S^*_{kB}|m_j) = \phi_{kj} + 1 \times (1 - \phi_{kj}) = 1$. Thus, $\mathrm{E}^*(S^*_k) = \mathrm{E}^*\mathrm{E}^*(S^*_k|m) = 1$. Next, the conditional variance is $\mathrm{var}^*(S^*_k|m_j) = \mathrm{E}^*\left[\mathrm{var}^*(S^*_k|S^*_{kA}, m_j)|m_j\right] + \mathrm{var}^*\left[\mathrm{E}^*(S^*_k|S^*_{kA}|m_j), m_j\right] = \mathrm{var}^*(S^*_{kB}|S^*_{kA} = 0, m_j)\Pr^*(S^*_{kA} = 0|m_j) = 1 - \phi_{kj}, j = 1, 2$. Thus, $\mathrm{var}^*(S^*_k) = \mathrm{E}^*\mathrm{var}^*(S^*_k|m) + \mathrm{var}^*\mathrm{E}^*(S^*_k|m) = q(1 - \phi_{k1}) +$

$(1-q)(1-\phi_{k2}) = 1-\phi_k$. Finally, the covariance is given by:

$$\begin{aligned}
&\mathrm{cov}^*(S_k^*, S_\ell^* | m_j) \\
&= \mathrm{cov}^*\left[\mathrm{E}^*(S_k^*|S_{kA}^*, S_{\ell A}^*, m_j), \mathrm{E}^*(S_\ell^*|S_{kA}^*, S_{\ell A}^*, m_j)|m_j\right] \\
&\quad +\mathrm{E}^*\left[\mathrm{cov}^*(S_k^*, S_\ell^*|S_{kA}^*, S_{\ell A}^*, m_j)|m_j\right] \\
&= \mathrm{cov}^*(S_k^*, S_\ell^*|S_{kA}^* = 0, S_{\ell A}^* = 0, m_j)\mathrm{Pr}^*(S_{kA}^* = 0, S_{\ell A}^* = 0|m_j) \\
&= -\frac{1}{n-m_j-1} \times (1-\phi_{kj}-\phi_{\ell j}+\phi_{k\ell j}).
\end{aligned}$$

Thus

$$\begin{aligned}
\mathrm{cov}^*(S_k^*, S_\ell^*) &= \mathrm{E}^*\mathrm{cov}^*(S_k^*, S_\ell^*|m) + \mathrm{cov}^*[\mathrm{E}^*(S_k^*|m), \mathrm{E}^*(S_\ell^*|m)] \\
&= -\frac{q}{n-m_1-1} \times (1-\phi_{k1}-\phi_{\ell 1}+\phi_{k\ell 1}) \\
&\quad -\frac{1-q}{n-m_2-1} \times (1-\phi_{k2}-\phi_{\ell 2}+\phi_{k\ell 2}).
\end{aligned}$$

□

We have seen that according to the definition of the c_k, there are several ways to approximate the matrix of $D_{k\ell}$ by a matrix of $\widetilde{D}_{k\ell}$. The values of ϕ_k that reconstruct as best as possible the three approximations for c_k given in (3.8), (3.9) and (3.10) can be chosen by taking $1-\phi_k = \widetilde{D}_{kk}$. Obviously, these resampling variances are not exactly equal to the estimator of variance, but they take into account the correction for finite population. Moreover, the diagonal terms are exactly the same as usual estimators of variance.

The case where $n - \sum_{k\in s}\phi_k < 2$ must also be treated. Consider the procedure used to compute the inclusion probabilities from a vector of positive values x_k. First, compute the quantities

$$\frac{nx_k}{\sum_{\ell\in U} x_\ell}, \tag{3.17}$$

$k = 1, \ldots, N$. For units for which these quantities are larger than 1, set $\pi_k = 1$. Next, the quantities are recalculated using (3.17) restricted to the remaining units. This procedure is repeated until each π_k is in $]0,1]$. Some π_k are 1 and others are proportional to x_k. Let $H(x_1, \ldots, x_N; n)$ denote the function that allows us to construct these inclusion probabilities from a vector of positive values $(x_1, \ldots, x_N)$. Function $H(.;.)$ allows us to define Algorithm 6 in order to select the bootstrap sample in the case where $n - \sum_{k\in s}\phi_k < 2$.

Algorithm 6 Resampling for unequal probability sampling without replacement: Case 2

Case 2: $n - \sum_{k \in S} \phi_k < 2$.

- Compute $(\psi_k, k \in S) = 1 - H(1 - \phi_k, k \in S; 2)$ and $q = (n - \sum_{k \in S} \phi_k)/2$.
- With a probability q select a sample without replacement denoted by S^*_{kA} of size $n - 2$ from S by using inclusion probabilities ψ_k. Let $\psi_{k\ell}$ denote the joint inclusion probability of this design. From the two remaining units, select a one-one design denoted by S^*_{kB}. The final sample is $S^*_{kA} + S^*_{kB}$.
- With a probability $1 - q$, $S^*_k = 1$, for all k in S.

With Algorithm 6, $\mathrm{E}^*(S^*_k) = 1$ and

$$\mathrm{var}^*(S^*_k) = \frac{(1 - \psi_k)(n - \sum_{k \in S} \phi_k)}{2},$$

and

$$\mathrm{cov}^*(S^*_k, S^*_\ell) = -\frac{(1 - \psi_k - \psi_\ell + \psi_{k\ell})(n - \sum_{k \in S} \phi_k)}{2}.$$

The ϕ_k can be chosen according to the three approximations given above in (3.8), (3.9) and (3.10).

3.11 Monte Carlo simulation study for numerical comparisons

First, we developed simulations for matrix reconstruction in order to confirm the theoretical results obtained in Section 3.10 on the new bootstrap methods for unequal probability sampling. As seen earlier, we distinguished the two cases depending on whether $n - \sum_{k \in S} \phi_k$ is greater than or equal to 2 or less then 2. We generated a population for each of these cases. We computed the matrices of Horvitz-Thompson and of Yates-Grundy variance estimators, as well as their approximations, we then ran sets of simulations to obtain the matrices of the variances using the new bootstrap method. We noticed that these matrices were very close to the respective approximations, so the method should provide estimators of variance that are very similar to the estimators given by the approximations. In order to be concise, we do not include the results of these simulations in this paper.

Secondly, we also ran a set of simulations for the variance estimators under different sampling designs. In each case a population of 150 units

was generated from the model $y_k = (\beta_0 + \beta_1 x_k^{1.2} + \sigma\varepsilon_k)^2 + c$, with $x_k = |i_k|$ and $i_k \sim \mathcal{N}(0,7)$, $\varepsilon_k \sim \mathcal{N}(0,1)$ and $\sigma = 15$. The regression parameters are $\beta_0 = 12.5$, $\beta_1 = 3$ and $c = 4000$. The model and its parameters were chosen intentionally to have a distribution for y similar to a lognormal - as it is often used for income distributions - with a correlated and positive explanatory variable x in the regression model. From this population, 1000 samples were drawn with a sample size $n = 50$. We intentionally used a large sample rate $n/N = 1/3$ and a skewed population in order to better illustrate the performance of the tested bootstrap methods. From each of these samples, we calculated four statistics: the total, the median, the Gini index of variable y and the ratio of total of variable y on the total of variable x.

Three sampling designs were tested: Poisson sampling, simple random sampling without replacement and a maximum entropy design with unequal inclusion probabilities. Concerning the inclusion probabilities, they were calculated proportional to the values of a variable z, which was generated from equation $z = y^{0.2}p$ where $p \sim \ln\mathcal{N}(0, 0.25)$. In this manner the correlation between y and z is about 0.5. In the case where the total was the function of interest, the goal was to reproduce the estimator of variance of the total. In fact, for the estimation of the total, estimators of variance can directly be computed. A resampling method is thus not necessary. However, simulations were also run in this case in order to test the performance of the methods.

From each of the 1000 initial samples, 1000 bootstrap samples were selected by means of five different bootstrap methods. Besides the new bootstrap method, four other resampling methods were tested. The first one is the bootstrap with replacement proposed by Mac Carthy & Snowden (1985) for which a correction factor for the finite population is used. The second one is the bootstrap without replacement, which consists of creating an artificial population from the initial sample and drawing bootstrap samples with the same design as the initial one (Gross, 1980; Chao & Lo, 1985). In the cases of simple random sampling without replacement or unequal inclusion probability sampling design as initial sampling designs, the third method is the rescaled bootstrap of Rao & Wu (1988). For the Poisson sampling design, we used the Beaumont & Patak (2012) method. Nonlinear functions of interest were also tested: the ratio of two totals, the median and the Gini index. For these functions of interest, the variances under the simulations, say the Monte Carlo variances, were considered as the true variances of the estimators. In the case where the total was the function of interest, the results were directly compared with the variance

of the total that can be exactly computed, and not with the Monte Carlo simulation variance. After drawing the bootstrap samples, the estimators, their variances and the means of these variances were computed for each of the initial samples and were then compared with the approximations of the true variances. Note that the median is not a smooth function of the total. Estimating its variance can therefore be difficult, but the simulations show that in this case bootstrap methods perform well.

In order to measure the performance of the new method and compare it with the other ones, the following five indicators were used:

- Lower error rate (L) in %

$$L = \frac{100}{sim}\sum_{i=1}^{sim} I\left[\widehat{\theta} - 1.96 \times \sqrt{\text{var}(\widehat{\theta}^*)} > \theta\right],$$

  where $I[a] = 1$ if a is true and $I[a] = 0$ elsewhere,

- Upper error rate (U) in %

$$U = \frac{100}{sim}\sum_{i=1}^{sim} I\left[\widehat{\theta} + 1.96 \times \sqrt{\text{var}(\widehat{\theta}^*)} < \theta\right],$$

- Total error rate (ER) in %

$$ER = 100 - \frac{100}{sim}\sum_{i=1}^{sim} I\left[\widehat{\theta} - 1.96 \times \sqrt{\text{var}(\widehat{\theta}^*)} \leq \theta \leq \widehat{\theta} + 1.96 \times \sqrt{\text{var}(\widehat{\theta}^*)}\right],$$

- Relative Bias

$$RB = 100 \times \frac{\text{var}(\widehat{\theta}^*) - \text{var}_{sim}(\widehat{\theta})}{\text{var}_{sim}(\widehat{\theta})} = 100 \times \frac{B}{\text{var}_{sim}(\widehat{\theta})},$$

- Relative Root Mean Squared Error

$$RRMSE = 100 \times \frac{\sqrt{B^2 + \text{var}[\text{var}(\widehat{\theta}^*)]}}{\text{var}_{sim}(\widehat{\theta})}.$$

The *RB* gives a measure of the bias of the estimator of variance. The *RRMSE* measures its accuracy. The *Error Rates* allow us to evaluate the capacity of the methods to provide a valid inference. The lower and the upper error rates give us an idea of how skewed the distribution of the estimator $\widehat{\theta}$ is. Tables 1, 2 and 3 present the numerical performances of the estimators of variance for the three sampling designs, the four functions of interest and the four resampling methods.

Table 3.1 presents the outcomes achieved using the Poisson sampling design with inclusion probabilities proportional to variable z. The variance estimator provided by the proposed method is unbiased for the total. For the other considered functions it is nearly unbiased according to the MC simulation. The relative biases are small, even for the Gini index (around -5%). For the total and the ratio, the total error rates are about 5%, and for the two other functions of interest about 10%. The bootstrap with replacement is clearly inefficient for the total. In fact, despite the use of a correction factor, the bootstrap with replacement with fixed sample size cannot catch the variance due to the randomness of the sample size of the Poisson sampling design. The variance estimator can thus largely underestimate the true variance. For the other functions of interest, the bootstrap with replacement provides a relatively high coverage rate, but the estimators themselves are biased. With regard to the bootstrap without replacement, the variance estimators are also strongly biased. For the total, the Gini index and the ratio, the variance estimators underestimate the true variance, and give lower coverage rates. For the median, the coverage rate is 97.5% which is only due to the large overestimation of the variance. In general, the performance of the proposed method and the method of Beaumont & Patak (2012) are equivalent. The estimators are unbiased, or have a slight bias for each function. The RRMSE have the same order and the error rates show a slightly positively skewed distribution, with coverage rates between 90 and 95%. We can conclude that the new method provides essentially the same results as the others, but its application is simpler: it does not require a correction factor, rescaling or artificial population.

Table 3.2 shows the results of the applications of resampling methods for simple random sampling without replacement. Here, the original sampling design has a fixed sample size, which explains why the bootstrap with replacement performs better. Instead of the method of Beaumont & Patak (2012) dedicated to Poisson sampling, we have used the rescaled bootstrap proposed by Rao & Wu (1988). The simulations show that, for the total error rates, the bootstrap with replacement method performs slightly better than the three others, but the coverage rates provided by these others are also between 93% and 94% for each function of interest. The lower and upper error rates for each method and for each function of interest show the same behavior: the distributions are right skewed. There are small biases, positive in the case of the total, the median and the ratio of two totals, except for the rescaled bootstrap method, where the variance of the median is underestimated. For the Gini index, the

Table 3.1 – *Performance of resampling methods in Poisson sampling*

POISSON	L	U	ER	Relative bias	RRMSE
TOTAL					
New method	0.5	4.7	5.2	-0.0278	38.5813
Bootstrap WR	10.1	16.2	26.3	-76.4830	78.6988
Bootstrap WOR	4.9	5.4	10.3	-35.4241	36.1937
Method of Patak-Beaumont	1.0	6.1	7.1	-2.8247	40.0502
MEDIAN					
New method	3.9	6.2	10.1	0.4701	60.1267
Bootstrap WR	2.3	4.3	6.6	-12.9935	50.2141
Bootstrap WOR	1.9	0.6	2.5	66.7149	113.1575
Method of Patak-Beaumont	3.1	4.8	7.9	8.0193	64.6926
GINI					
New method	1.1	9.8	10.9	-5.3805	38.4937
Bootstrap WR	0.0	5.2	5.2	15.3095	44.8152
Bootstrap WOR	3.5	13.9	17.4	-41.5459	48.1382
Method of Patak-Beaumont	0.6	8.8	9.4	8.1915	65.8452
RATIO					
New method	2.3	4.0	6.3	1.6710	59.6199
Bootstrap WR	0.6	2.6	3.2	-4.8825	49.1502
Bootstrap WOR	8.3	6.2	14.5	-45.2226	48.6318
Method of Patak-Beaumont	1.8	4.8	6.6	8.0236	76.6924

first three methods give an estimator that underestimates the true variance, in contrast to the rescaling bootstrap method. In general, for simple random sampling without replacement, there is no crucial difference in performance between the resampling methods. They all provide a slightly biased estimator, with relatively high coverage rates - around 94% - and the variabilities of the variance estimators are also similar.

Table 3.3 shows the performance of resampling methods under a maximum entropy design with inclusion probabilities proportional to variable z. In the proposed bootstrap method, the second approximation (3.9) is used, which gives us $\phi_k = \pi_k$. In the case where the function of interest is the total, the new method gives an unbiased estimator with a coverage rate of 92.2%. The bootstrap with replacement provides a lower error rate and thus a higher coverage rate for the total. However, it is due to a larger estimated confidence interval caused by a slight overestimation of the variance. The bootstrap without replacement with an artificial population and the rescaling bootstrap method strongly underestimate the variance and consequently give a smaller coverage rate. The RRMSE are essentially the same, and again, the distributions of the estimators are right skewed. Concerning the median, the variance estimator of the new method is practically unbiased while the bootstrap with replacement, and the rescaled bootstrap underestimate the variance. The bootstrap without replacement seriously overestimates it. For the Gini index, the new method and the

Table 3.2 – *Performance of resampling methods in simple random sampling without replacement sampling design*

SRSWOR	L	U	ER	Relative bias	RRMSE
TOTAL					
New method	1.3	6.3	7.6	5.9195	35.9356
Bootstrap WR	0.0	4.1	4.1	6.5763	36.3808
Bootstrap WOR	1.2	6.3	7.5	4.6716	35.4567
RW Bootstrap	1.0	6.5	7.5	0.6130	33.0132
MEDIAN					
New method	1.8	6.4	8.2	9.4256	56.6512
Bootstrap WR	0.5	4.4	4.9	3.8235	49.2184
Bootstrap WOR	2.1	6.1	8.2	10.7279	58.4537
RW Bootstrap	1.9	6.1	8.0	-1.5549	49.6286
GINI					
New method	1.7	5.0	6.7	-4.4216	17.6308
Bootstrap WR	0.6	2.5	3.1	-2.5877	18.1130
Bootstrap WOR	1.7	5.6	7.3	-3.4073	19.4626
RW Bootstrap	0.7	7.8	8.5	12.3624	42.5067
RATIO					
New method	1.7	4.2	5.9	1.2438	28.5170
Bootstrap WR	0.2	2.6	2.8	3.0868	29.1121
Bootstrap WOR	1.7	4.3	6.0	1.3686	28.5030
RW Bootstrap	1.8	4.8	6.6	0.0379	27.1146

bootstrap without replacement perform almost identically: the estimators of the variance are slightly biased (1-3% in absolute value) with a coverage rate of around 92-93%. The coverage rate provided by the bootstrap with replacement method is slightly larger, but the variance estimator is biased. The rescaled bootstrap method strongly underestimates the variance and this is the reason why the error rate is higher. Concerning the ratio, the estimator under the new resampling method has a small negative bias. In contrast, the bootstrap with replacement method gives an unbiased estimator and the bootstrap without replacement method gives a variance estimator that is 41% larger than the true variance. To summarize these results: the new method performs at least as well as the other methods considered. At the same time, it is simpler and does not require any additional calculation to estimate the variance of the estimators.

These simulations show that the new bootstrap method works at least as well as the usual bootstrap methods. In Poisson sampling design, the inefficiency of the bootstrap with replacement is clear. It is due to the randomness of the sample size. In general, the new method provides an unbiased or a slightly biased estimator with a coverage rate between 89% and 95% for each of the functions of interest studied here, under each cinsidered sampling design. Besides having at least the same performance as the other methods, the main advantage of the new method is that it does

Table 3.3 – *Performance of the resampling methods in maximum entropy sampling design*

UPWOR	L	U	ER	Relative bias	RRMSE
TOTAL					
New Method	0.4	7.4	7.8	-0.9515	35.8027
Bootstrap WR	0.0	2.8	2.8	6.8616	34.9417
Bootstrap WOR	3.1	8.8	11.9	-22.6490	33.1929
RW Bootstrap	2.8	10.6	13.4	-36.7334	49.8869
MEDIAN					
New Method	3.5	6.8	10.3	0.9405	58.6158
Bootstrap WR	1.3	5.0	6.3	-12.9157	48.5572
Bootstrap WOR	0.2	0.0	0.2	233.5629	280.1593
RW Bootstrap	16.6	19.0	35.6	-71.0074	72.6283
GINI					
New Method	2.1	5.6	7.7	-3.8518	7.1675
Bootstrap WR	1.0	3.1	4.1	-12.9006	5.5983
Bootstrap WOR	1.3	5.2	6.5	-1.2589	3.3370
RW Bootstrap	7.7	16.7	24.4	-64.8759	65.8130
RATIO					
New Method	2.6	3.6	6.2	-2.3080	36.2905
Bootstrap WR	1.2	1.5	2.7	1.0800	30.9940
Bootstrap WOR	2.0	0.7	2.7	41.4054	53.3239
RW Bootstrap	15.3	13.5	28.8	-71.3400	71.9911

not require rescaling, correction factors or an artificial population. Thus, the samples can be directly used to compute the variance of the functions of interest.

3.12 Discussion

The main idea driving the new methodology presented in this article is that if the original sample is drawn with replacement, the one-one sampling design can be directly used in the bootstrap method even if the units are selected with unequal probabilities. If it is drawn without replacement, the true variances are smaller than that of a design with replacement and thus a portion of the resampled units are selected without replacement and another is selected according to a one-one design in order to achieve the correct variance. The implementation of selecting resampled units according to a mixture of sampling designs is straightforward and extremely fast. It consists in computing the sample sizes of the different components of the mixtures, and then proceeds to select the bootstrap samples, which do not need to be rescaled. The Horvitz-Thompson weights remain unchanged from the original sample.

The simulations show that the classical bootstrap with replacement might be seriously biased under unequal probability sampling without re-

placement, or if the sample size is random. For simple random sampling without replacement, the bootstrap with replacement requires a rescaling factor. The class of methods based on the construction of artificial populations has limitations in its time-consuming execution. In addition, inaccuracy may arise due to rounding problems in the multiplication of sample units by the inverse of their inclusion probabilities, which are almost never integer (Holmberg, 1998). This problem is bypassed in the methodology proposed in the present work. Regarding the method of Rao & Wu (1988), the bootstrap values need not be values from the original sample because of the redefinition technique; although this indeed provides unbiased estimators, difficulties may arise in cases of calibration, reweighting and imputation.

The method of Beaumont & Patak (2012) entails noninteger weights that may even be negative, which can lead to counterintuitive bootstrap estimations. This problem is mitigated via a rescaling method, but it requires a rescaling factor for the variance, which also presents difficulties under imputation, calibration, and weighting for total nonresponse. The work proposed here can be seen as a variant of the method of Beaumont & Patak (2012), but we impose weights that are positive and integer.

The use of artificial populations produces the correct variance, but, as shown in simulation studies, it can be cumbersome and time consuming. The present work avoids these difficulties and attains bootstrap samples in a direct manner that have precisely the same weights as in the original sample, and do not present any of the previous limitations when weighting, calibration or imputation is required.

4 New Resampling Method for Sampling Designs Without Replacement: the Doubled Half Bootstrap

Abstract

A new and very fast method of bootstrap for sampling without replacement from a finite population is proposed. This method can be used to estimate the variance in sampling with unequal inclusion probabilities and it does not require generation of artificial populations, calculation of bootstrap weights or rescaling. The bootstrap samples are directly selected from the original sample. The bootstrap procedure contains two steps: In the first step, units are selected once with Poisson sampling using the same inclusion probabilities as the original design. In the second step, amongst the non-selected units, half of the units are randomly selected twice. This procedure enables us to efficiently estimate the variance. A set of simulations show the advantages of this new resampling method. [1]

Keywords: Poisson sampling, simple random sampling, unequal probability sampling, variance estimation

4.1 Introduction

Resampling methods are frequently used to lead inference in survey statistics. The main difficulty, however, is that the variance of an estimator depends on the sampling design. Bootstrap methods must thus be adapted

[1] This chapter is a reprint of: Antal, E. and Tillé, Y. (2012). New Resampling Method for Sampling Designs Without Replacement: the Doubled Half Bootstrap. *Submitted.*

for each sampling design. Moreover, the variance of the Horvitz-Thomson estimator can have a very different form from the variance estimator. The original bootstrap method, developed by Efron (1979) is not directly applicable in sampling from a finite population because the units of the sample are not independent and identically distributed when the sample is selected without replacement. Gross (1980) and Chao & Lo (1985) proposed a method based on the reconstruction of pseudo-populations from the sample. Another important family of methods is the rescaled bootstrap (Rao & Wu, 1988) which consists of modifying the values of the interest variable to reconstruct an unbiased variance estimator for statistics that are linear functions of the observations. Other methods were also proposed by Mac Carthy & Snowden (1985); Kuk (1989); Rao et al. (1992); Shao & Tu (1995); Sitter (1992a,b); Booth et al. (1994); Holmberg (1998). The main approach presented of this paper could be seen as similar to the general proposal of Bertail & Combris (1997), which has also been used in Lahiri (2003).

Beaumont & Patak (2012) propose a bootstrap method that directly reconstructs the variance for linear cases. The drawback of this method is that the observations must be weighed by non-integer values. Antal & Tillé (2011a) propose another method that uses non-integer weights and that is based on mixture of discrete multivariate distributions. In this paper, we propose a new methodology that is less complex than the one developed in Antal & Tillé (2011a) to select a bootstrap sample when the original sample is drawn by sampling without replacement. This method enables one to quickly implement and directly reconstruct the appropriate variance without need of reweighting the statistical units.

The paper is organized as follows: In Section 4.2, the notation for a sampling design, the estimator of the total and its variance estimator are defined. Section 4.3 establishes the conditions needed to obtain unbiased bootstrap estimates of the variances. In Section 4.4, a new method is proposed for Poisson sampling. Next, the doubled half sampling is introduced in Section 4.5. This tool is used to define a new bootstrap method for simple random sampling in Section 4.6 and for unequal probability sampling in Section 4.7. Simulations are presented in Sections 4.8 and conclusions are drawn in Section 4.9.

4.2 Sampling design, total and variance

Let $p(.)$ be a sampling design on a population $U = \{1, \ldots, N\}$ of size N such that

$$p(s) \geq 0, \text{ for all } s \subset U, \text{ and } \sum_{s \subset U} p(s) = 1.$$

Let S be the random sample such that $\Pr(S = s) = p(s)$. The sample size n of S can be random or not. Define also the inclusion probabilities $\pi_k = \Pr(k \in S)$ for $k \in U$, and the joint inclusion probabilities $\pi_{k\ell} = \Pr(k \text{ and } \ell \in S)$ for $k, \ell \in U$. Moreover, define $\Delta_{k\ell} = \pi_{k\ell} - \pi_k \pi_\ell$ for $k, \ell \in U$, and $\check{\Delta}_{k\ell} = \Delta_{k\ell}/\pi_{k\ell}$. When $k = \ell$, we obtain $\Delta_{kk} = \pi_k(1 - \pi_k), k \in U$ and $\check{\Delta}_{kk} = 1 - \pi_k$.

If all the inclusion probabilities are strictly positive, then the total $Y = \sum_{k \in U} y_k$ of the values $y_1, \ldots, y_k, \ldots, y_N$ taken by the interest variable y can be unbiasedly estimated by the Horvitz-Thompson estimator (HT) (Horvitz & Thompson, 1952)

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

whose variance is given by

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

and can be unbiasedly estimated by the Horvitz-Thompson (HT) variance estimator

$$\widehat{\text{var}}_{HT}(\widehat{Y}_\pi) = \sum_{k \in S} \sum_{\ell \in S} \frac{y_k y_\ell}{\pi_k \pi_\ell} \check{\Delta}_{k\ell}.$$

This variance estimator is however very unstable because

$$\sum_{k \in S} \check{\Delta}_{k\ell}$$

is generally different from zero. In the particular case where the sample size is fixed and $y_k = \pi_k, k \in U$, the HT-estimator of the total is equal to the sample size and is thus not random. Nevertheless, the HT-variance estimator is generally not zero. Indeed, in this case,

$$\widehat{\text{var}}_{HT}(\widehat{Y}_\pi) = \sum_{k \in S} \sum_{\ell \in S} \check{\Delta}_{k\ell}.$$

When the sample size is fixed, the Sen-Yates-Grundy (SYG) estimator

is also unbiased (Sen, 1953; Yates & Grundy, 1953):

$$\widehat{\text{var}}_{SYG}(\widehat{Y}_\pi) = \sum_{k\in S}\sum_{\ell\in S}\frac{y_k y_\ell}{\pi_k\pi_\ell}D_{k\ell},$$

where

$$D_{k\ell} = \begin{cases} -\sum\limits_{\substack{j\in S\\ j\neq k}}\frac{\Delta_{kj}}{\pi_{kj}} & \text{if } k=\ell \\ \frac{\Delta_{k\ell}}{\pi_{k\ell}} & \text{if } k\neq\ell. \end{cases}$$

Several other variance estimators exist. They all have the same form as the SYG-estimator with different values for $D_{k\ell}$. Matei & Tillé (2005) discussed the merits of a family of estimators based on another value of $D_{k\ell}$ given by:

$$\widetilde{D}_{k\ell} = \begin{cases} c_k - \frac{c_k^2}{\sum_{j\in U} S_j c_j} & \text{if } k=\ell \\ -\frac{c_k c_\ell}{\sum_{j\in U} S_j c_j} & \text{if } k\neq\ell. \end{cases}$$

Diverse values have been proposed for the c_k. Matei & Tillé (2005) ran a set of simulations that shows that the choice proposed by Hájek (1981):

$$c_k = \frac{n}{n-1}(1-\pi_k). \tag{4.1}$$

produces a very efficient and slightly biased estimator. We refer to this estimator as the H-estimator of the variance.

4.3 Bootstrap

A bootstrap sample is a sample with replacement - not necessarily a simple random sample - selected from S. Let S_k^* be the number of times unit k is repeated in the bootstrap sample. The HT estimator of the total for a single bootstrap sample is given by

$$\widehat{Y}^* = \sum_{k\in S}\frac{y_k}{\pi_k}S_k^*.$$

Let $\text{Pr}^*(.) = \text{Pr}(.|S)$, $\text{E}^*(.) = \text{E}(.|S)$ and $\text{var}^*(.) = \text{var}(.|S)$ respectively denote the probability, the expectation and the variance of the total estimator in the bootstrap sample given the original sample. Then

$$\text{E}^*(\widehat{Y}^*) = \sum_{k\in S}\frac{y_k}{\pi_k}\text{E}(S_k^*),$$

and

$$\text{var}^*(\widehat{Y}^*) = \sum_{k\in S}\sum_{\ell\in S}\frac{y_k y_\ell}{\pi_k\pi_\ell}\text{cov}(S_k^*, S_\ell^*|S).$$

A necessary and sufficient condition for the expected value $\text{E}^*(\widehat{Y}^*)$ to equal the H-T estimator of the total is

$$\text{E}^*(S_k^*) = 1, k \in S. \tag{4.2}$$

Moreover, in order to have an unbiased variance estimator of the total, a first condition is that

$$\text{var}^*(S_k^*) = \check{\Delta}_{kk} = 1 - \pi_k, k \in S. \tag{4.3}$$

The bootstrap estimator of the variance $v_{boot}(\widehat{Y}^*)$ is computed by generating a set of bootstrap samples and computing the variance of the outcomes of $\widehat{Y}^*$. Moreover, if a bootstrap method provides an approximately unbiased estimator for the variance of the HT estimator of the totals, it will also provide approximately unbiased variance estimators for smooth functions of HT estimator of several totals.

Ideally, another condition for a bootstrap method to unbiasedly estimate the variance of the HT-estimator is that

$$\text{cov}(S_k^*, S_\ell^*|S) = \check{\Delta}_{k\ell}, k \neq \ell \in S. \tag{4.4}$$

These conditions on the covariances are however difficult to meet when the sample is selected with fixed sample size and unequal inclusion probabilities. In this particular case, it is difficult to exactly satisfy more than conditions (4.2) and (4.3). Condition (4.4) can however be approximately satisfied.

When the sample size is fixed, another way of constructing an unbiased estimator of the variance is to equate the bootstrap estimator to the SYG-estimator. Doing that, the sufficient conditions become

$$\text{E}^*(S_k^*) = 1, k \in S,$$

$$\text{var}^*(S_k^*) = D_{kk}, k \in S,$$

and

$$\text{cov}(S_k^*, S_\ell^*|S) = D_{k\ell}, k \neq \ell \in S. \tag{4.5}$$

But again, conditions (4.5) on the covariances are difficult to meet when the sample is selected with unequal inclusion probabilities, but it could be approximately satisfied. Below we will see that, for unequal probability

sampling with fixed sample size, there are two bootstrap strategies that may be used to approximate either the HT or the SYG-estimator.

4.4 Bootstrap for Poisson design

Suppose the original sample is obtained by a so-called Poisson design, where the observation k is included with probability π_k and the decision is made for each observation independently. The name reflects the fact that if all π_k are small, then the sample size n has approximately a Possion distribution with mean $\sum_{k=1}^{N} \pi_k$. In a Poisson design with inclusion probabilities π_k,

$$p(s) = \prod_{k=1}^{N} \pi_k^{\mathbb{1}(k\in s)}(1-\pi_k)^{\mathbb{1}(k\notin s)} \text{ for all } s \subset U,$$

where $\mathbb{1}(A)$ is equal to 1 if A is true and 0 otherwise. The inclusion probability is $\Pr(k \in S) = \pi_k$. Moreover, $\pi_{k\ell} = \pi_k\pi_\ell$ when $k \neq \ell \in U$ and $\pi_{kk} = \pi_k$. Thus $\Delta_{k\ell} = 0$, when $k \neq \ell \in U$ and $\Delta_{kk} = \pi_k(1-\pi_k)$. We thus have, $\check{\Delta}_{k\ell} = 0$, when $k \neq \ell \in U$ and $\check{\Delta}_{kk} = 1-\pi_k$. Under Poisson sampling design, the sample size n is random and thus the estimator of variance is calculated by $\widehat{\text{var}}_{HT}(\widehat{Y}_\pi)$.

Beaumont & Patak (2012) propose a bootstrap method for Poisson design that uses normal independent variables with expectation equal to 1 and variances equal to $1-\pi_k$ thus

$$S_k^* \sim N(1, 1-\pi_k).$$

Unfortunately, this method requires the use of non-integer weights that can be negative. Instead we recommend the use of a discrete random variable for S_k^*.

Antal & Tillé (2011a) proposed a simple bootstrap method that uses n independent Bernoulli random variables X_k with parameter π_k and n independent Poisson random variables Z_k with parameter $\lambda = 1$. For this method, the bootstrap sample is given by

$$S_k^* = X_k + (1-X_k)Z_k, k \in S.$$

Thus, the probability mass function of S_k^* is given by:

$$\Pr^*(S_k^* = r) = \pi_k \mathbb{1}[r=1] + \frac{(1-\pi_k)}{e \cdot r!}, r = 0, 1, 2, \ldots$$

where $e \approx 2.71$ is the Euler constant. The bootstrap variable S_k^* satisfies conditions (4.2), (4.3), and (4.4).

An even simpler method is to consider n independent Bernoulli random variables $X_k, k \in S$ with parameter π_k and n independent Bernoulli random variables Y_k with parameter $1/2$. Define the bootstrap sample by

$$S_k^* = X_k + 2(1 - X_k)Y_k, k \in S.$$

The probability distribution of S_k^* is thus

$$S_k^* = \begin{cases} 0 & \text{with a probability } (1 - \pi_k)/2 \\ 1 & \text{with a probability } \pi_k \\ 2 & \text{with a probability } (1 - \pi_k)/2. \end{cases}$$

Again, the bootstrap variable S_k^* meets conditions (4.2), (4.3), and (4.4). Here, the bootstrap sample does not contain the same unit more than twice.

4.5 One-one design and doubled half sampling

When the original sample has a fixed sample size, Antal & Tillé (2011a) proposed an important tool in order to estimate the variance of an estimator via bootstrap method. This tool is the family of one-one designs. The members of this family are discrete probability distributions that have common properties concerning their expectation and their variance, where the name of the family comes from, that is:

$$\mathrm{E}^*(S_k^*) = 1,$$

$$\mathrm{var}^*(S_k^*) = 1.$$

Two other important properties are satisfied by these designs. The first one guarantees the same fixed sample size for the bootstrap sample, the second one concerns the covariance between S_k^* and S_ℓ^*.

$$\sum_{k \in S} S_k^* = n,$$

$$\mathrm{cov}^*(S_k^*, S_\ell^*) = -\frac{1}{n-1}, k \neq \ell \in S.$$

Antal & Tillé (2011a) showed that such a sampling design can be obtained by using a mixture between two samples selected by simple random sampling with replacement and by simple random sampling with over-

replacement (Antal & Tillé, 2011b). One-one designs can next be mixed with other sampling designs in order to reproduce an unbiased estimator of variance for most of the sampling methods with fixed sample size.

Before describing the new bootstrap procedure, we first propose a simpler method for selecting a one-one design that we call 'doubled half sampling'. If the size n of the initial sample is *even*, then a sample from S of size $n/2$ is selected with simple random sampling without replacement. Each selected unit is taken twice. In this case, we obtain

$$\mathrm{E}^*(S_k^*) = 2 \times \frac{1}{2} = 1,$$

$$\mathrm{var}^*(S_k^*) = 4 \times \frac{1}{2}\left(1 - \frac{1}{2}\right) = 1,$$

and

$$\mathrm{cov}(S_k^*, S_\ell^* | S) = 4 \times \frac{1}{2}\left(1 - \frac{1}{2}\right)\frac{-1}{n-1} = \frac{-1}{n-1}.$$

If n is odd, then we can have the same property by means of the following slightly modified procedure:

- Select $(n-1)/2$ units from S and take them twice in the bootstrap sample.
- With a probability $1/4$, select a unit from the set of units selected twice. This unit is selected one more times.
- Otherwise, with a probability $3/4$, select a unit with equal probabilities among the units that are not selected. This unit is selected only once.

This procedure gives the following distribution for S_k^*:

$$\Pr^*(S_k^* = j) = \begin{cases} \frac{n+1}{2n} \times \left(1 - \frac{3}{4} \times \frac{2}{n+1}\right) = \frac{2n-1}{4n} & \text{if } j = 0 \\ \frac{n+1}{2n} \times \frac{3}{4} \times \frac{2}{n+1} = \frac{3}{4n} & \text{if } j = 1 \\ \frac{n-1}{2n} \times \left(1 - \frac{1}{4} \times \frac{2}{n-1}\right) = \frac{2n-3}{4n} & \text{if } j = 2 \\ \frac{n-1}{2n} \times \frac{1}{4} \times \frac{2}{n-1} = \frac{1}{4n} & \text{if } j = 3. \end{cases}$$

After some algebra, it can be shown that this design is one-one. A one-one design can thus be selected for any sample size except when $n = 1$.

4.6 Bootstrap for simple random sampling without replacement

In this chapter, a bootstrap method is proposed for the case where the original sample was selected by simple random sampling without replacement. In simple random sampling without replacement,

$$p(s) = \begin{cases} \frac{n!(N-n)!}{N!} & \text{for all the samples } s \text{ of size } n \\ 0 & \text{otherwise.} \end{cases}$$

The inclusion probability is $\pi_k = n/N$. Moreover,

$$\pi_{k\ell} = \frac{n(n-1)}{N(N-1)}$$

when $k \neq \ell \in U$ and $\pi_{kk} = n/N$. Thus,

$$\Delta_{k\ell} = -\frac{n(N-n)}{N^2(N-1)},$$

when $k \neq \ell \in U$ and

$$\Delta_{kk} = \frac{n(N-n)}{N^2}.$$

We thus have,

$$\check{\Delta}_{k\ell} = -\frac{N-n}{N(n-1)},$$

when $k \neq \ell \in U$ and $\check{\Delta}_{kk} = 1 - n/N$. Note also that in this case, the HT-estimator and the SYG-estimator of the variance are equal, i.e. $\check{\Delta}_{kk} = D_{kk} = 1 - n/N$ for all $k \in U$.

The selection of a bootstrap sample can be done by using the following two-stage procedure. Let S_k^* denotes the number of times unit k is selected in the bootstrap sample.

- Select a sample from S by using independent Bernoulli random variables $X_k, k \in S$, with probabilities $\pi_k = n/N$. The selected units are taken once in the bootstrap sample $S_k^*, k \in S$. Let $m = \sum_{k \in S} X_k$. Thus $\mathrm{E}(m) = n^2/N$.

- – If the number of non-selected units is equal to or larger than 2, then select a doubled half sample design in $S_k^*, k \in S$ amongst the units $k \in S$ such that $X_k = 0$.
  – If the number of units such that $X_k = 0$ is equal to 1 (say unit

ℓ), select unit ℓ with the following distribution

$$S_\ell^* = \begin{cases} 0 & \text{with probability } 1/4 \\ 1 & \text{with probability } 1/2 \\ 2 & \text{with probability } 1/4. \end{cases}$$

Next, randomly select one of the units such that $X_k = 1$ (say z) with equal probability and select it $S_z^* = 2 - S_\ell^*$ times.

– If the number of units such that $X_k = 0$ is null, then the bootstrap sample S_k^* is the same as the original sample.

Note that

$$\Pr^*(S_k^* = j | m = r \text{ and } n - r \text{ is even}) = \begin{cases} (1 - r/n)/2 & \text{if } j = 0 \\ r/n & \text{if } j = 1 \\ (1 - r/n)/2 & \text{if } j = 2, \end{cases}$$

$$\Pr^*(S_k^* = j | m = r, n - r \text{ is odd, and } r < n - 1) = \begin{cases} \left(1 - \frac{r}{n}\right) \frac{2n-1}{4n} & \text{if } j = 0 \\ \frac{r}{n} + \left(1 - \frac{r}{n}\right) \frac{3}{4n} & \text{if } j = 1 \\ \left(1 - \frac{r}{n}\right) \frac{2n-3}{4n} & \text{if } j = 2 \\ \left(1 - \frac{r}{n}\right) \frac{1}{4n} & \text{if } j = 3, \end{cases}$$

and

$$\Pr^*(S_k^* = j | m = n - 1) = \begin{cases} 1/(2n) & \text{if } j = 0 \\ 1 - 1/n & \text{if } j = 1 \\ 1/(2n) & \text{if } j = 2. \end{cases}$$

It can be checked that $\mathrm{E}(S_k^*|m) = 1$ and $\mathrm{var}(S_k^*|m) = 1 - m/n$, for all three cases. We obtain $\mathrm{E}(S_k^*) = \mathrm{E}\mathrm{E}(S_k^*|m) = 1$ and

$$\begin{aligned} \mathrm{var}(S_k^*) &= \mathrm{E}\ \mathrm{var}(S_k^*|m) + \mathrm{var}\ \mathrm{E}(S_k^*|m) = 1 - \mathrm{E}(m)/n \\ &= 1 - \mathrm{E}(m)/n = 1 - (n^2/N)n = 1 - n/N. \end{aligned}$$

We thus have $\mathrm{E}^*(S_k^*) = 1, \mathrm{var}^*(S_k^*) = \check{\Delta}_{kk}, k \in S$. The $\mathrm{cov}^*(S_k^*, S_\ell^*) = \check{\Delta}_{k\ell}$ is equal because units have all been treated symmetrically.

This procedure passably differs from the method proposed in Antal & Tillé (2011a). Indeed, in the first stage a Poisson design is used whereas in Antal & Tillé (2011a) an almost fixed sample size is used. In the second stage, an half-double sampling whereas in Antal & Tillé (2011a) a complex mixture of distributions is applied. The new method is thus much less complex.

4.7 Bootstrap for unequal probability sampling without replacement

In this section we propose a bootstrap strategy for estimating the variance of an estimator when the original sample is selected by means of an unequal probability sampling design without replacement. There is a large set of sampling methods with unequal inclusion probabilities and fixed sample size (see among others Brewer & Hanif, 1983; Tillé, 2006). Each method has a particular matrix of elements $\breve{\Delta}_{k\ell}$ but the diagonal of this matrix is always $\breve{\Delta}_{kk} = 1 - \pi_k, k \in U$. However, the sampling designs with large entropy have very similar joint inclusion probabilities (see Brewer & Donadio, 2003; Matei & Tillé, 2005; Henderson, 2006; Preston & Henderson, 2007).

Consider the procedure used to compute the inclusion probabilities from a vector of positive values x_k. First, compute the quantities

$$\frac{nx_k}{\sum_{\ell \in U} x_\ell}, \tag{4.6}$$

$k = 1, \ldots, N$. For units where the quantities are larger than 1, set $\pi_k = 1$. Next, the quantities are recalculated using (4.6) restricted to the remaining units. This procedure is repeated until each π_k is in $(0, 1]$. Some π_k are 1 and others are proportional to x_k. Let $\pi_k = H_k(x_1, \ldots, x_N; n)$ denote the function that allows us to construct these inclusion probabilities from a vector of positive values $(x_1, \ldots, x_N)$. Function $H(.;.)$ will be used in the new bootstrap we propose for unequal probability sampling without replacement.

A bootstrap method for unequal probability sampling without replacement should satisfy

$$\mathrm{E}^*(S_k^*) = 1,$$

and must have a fixed sample size i.e.

$$\sum_{k \in S} S_k^* = n.$$

Moreover $\mathrm{var}^*(S_k^*)$ should be equal either to the diagonal of matrix $(\breve{\Delta}_{k\ell})$ used for the HT-estimator or to the diagonal of matrix $(D_{k\ell})$ used for the SYG-estimator, i.e.

$$\mathrm{var}^*(S_k^*) = \breve{\Delta}_{k\ell} = 1 - \pi_k, k \in S,$$

or, by posing $\phi_k = 1 - D_{kk}$,

$$\text{var}^*(S_k^*) = D_{kk} = 1 - \phi_k, k \in S.$$

With unequal inclusion probabilities, it is difficult to construct a bootstrap method that meets the properties of the covariances given by

$$\text{cov}^*(S_k^*, S_\ell^*) = \check{\Delta}_{k\ell}, \text{ or } \text{cov}^*(S_k^*, S_\ell^*) = D_{k\ell}, k \neq \ell \in S. \tag{4.7}$$

Nevertheless, since the bootstrap sample has a fixed sample size, the relation

$$\sum_{k \in S} \text{cov}^*(S_k^*, S_\ell^*) = 0$$

ensures that (4.7) will be approximately satisfied when the sampling design has large entropy. Let S_k^* denotes the number of times unit k is selected in the bootstrap sample.

- Select a sample from S by using a Poisson random variable $X_k, k \in S$, with the same inclusion probabilities as the original design π_k. The selected units are taken once in the bootstrap sample S_k^*. Let $m = \sum_{k \in S} X_k$. Thus $\text{E}(m) = \sum_{k \in S} \pi_k$.
-
  - If the number of non-selected units is equal to or larger than 2, then select a doubled half sample design in S_k^* amongst the units such that $X_k = 0$.
  - If the number of units such that $X_k = 0$ is equal to 1 (say unit ℓ),
    - With a probability 1/2, select the same bootstrap sample as the original sample.
    - Otherwise, with a probability 1/2, compute

$$\pi_{k|n-1} = \text{E}(X_k | m = n - 1) = 1 - \frac{\frac{1-\pi_k}{\pi_k}}{\sum_{\ell \in S} \frac{1-\pi_\ell}{\pi_\ell}}$$

(see the Appendix for the proof). With a sampling method with unequal inclusion probabilities with fixed sample size, select $n - 2$ units from S with probability

$$\psi_k = 1 - H_k(1 - \pi_{k|n-1}, k \in S; 2)$$

and take them once in the bootstrap sample S_k^*. Select a doubled half sample from the two units that are not se-

lected. Note that $\psi_k = 2\pi_{k|n-1} - 1$ except when one of the $\pi_{k|n-1}$ is less than $1/2$.

Let $\pi_{k|r} = \mathrm{E}(X_k|m = r)$. These conditional probabilities are not easy to compute. A recursive relation for computation is given for instance in Tillé (2006, p. 81). Fortunately, we do not need to compute this conditional expectation in order to implement the method except for case $r = n - 1$. However we will use it in the following reasoning.

We have

$$\Pr^*(S_k^* = j|m = r \text{ and } n - r \text{ is even}) = \begin{cases} (1 - \pi_{k|r})/2 & \text{if } j = 0 \\ \pi_{k|r} & \text{if } j = 1 \\ (1 - \pi_{k|r})/2 & \text{if } j = 2, \end{cases}$$

$$\Pr^*(S_k^* = j|m = r, n-r \text{ is odd, and } r < n-1) = \begin{cases} (1 - \pi_{k|r})\,\frac{2n-1}{4n} & \text{if } j = 0 \\ \pi_{k|r} + (1 - \pi_{k|r})\,\frac{3}{4n} & \text{if } j = 1 \\ (1 - \pi_{k|r})\,\frac{2n-3}{4n} & \text{if } j = 2 \\ (1 - \pi_{k|r})\,\frac{1}{4n} & \text{if } j = 3, \end{cases}$$

and

$$\Pr^*(S_k^* = j|m = n - 1) = \begin{cases} (1 - \psi_k)/4 & \text{if } j = 0 \\ (1 + \psi_k)/2 & \text{if } j = 1 \\ (1 - \psi_k)/4 & \text{if } j = 2. \end{cases}$$

It can be checked that $\mathrm{E}(S_k^*|m) = 1$,

$$\mathrm{var}(S_k^*|m = r) = 1 - \pi_{k|r}, r = 0, 2, 3, \ldots, n,$$

$$\mathrm{var}(S_k^*|m = 1) = \frac{1 - \psi_k}{2} = 1 - \pi_{k|n-1} + \pi_{k|n-1} - \frac{1 + \psi_k}{2}.$$

We thus have $\mathrm{E}(S_k^*) = \mathrm{EE}(S_k^*|m) = 1$ and

$$\begin{aligned} \mathrm{var}(S_k^*) &= \mathrm{Evar}(S_k^*|m) + \mathrm{varE}(S_k^*|m) \\ &= \mathrm{E}(1 - \pi_{k|r}) + \left(\pi_{k|n-1} - \frac{1 + \psi_k}{2}\right) \Pr^*(m = n - 1) \\ &= 1 - \pi_k + \left(\pi_{k|n-1} - \frac{1 + \psi_k}{2}\right) \Pr^*(m = n - 1). \end{aligned}$$

The variance of the diagonal is very slightly biased. Indeed

$$\mathrm{var}^*(S_k^*) = \check{\Delta}_{kk} + \left(\pi_{k|n-1} - \frac{1 + \psi_k}{2}\right) \Pr^*(m = n - 1), k \in S.$$

The bias is small and often nonexistent. Indeed, $(m = n - 1)$ is a rare

event. Moreover, $\pi_{k|n-1} - (1+\psi_k)/2$ is null except if one of the $\pi_{k|n-1}$ is smaller than $1/2$, which is also rare except in case where the sample size is very small. Moreover, we always have

$$\sum_{k \in S} \left(\pi_{k|n-1} - \frac{1+\psi_k}{2} \right) = 0.$$

Below, the simulations will show that even for very small sample sizes, the bias is negligible.

The same results can be derived by taking ϕ_k in place of π_k. The D_{kk} can sometimes be larger than 1. In this case, we advocate to take $\phi_k = 0$. We can thus define two bootstrap methods according to the fact that the inclusion probabilities are π_k or ϕ_k. The first case will be referred to as π-bootstrap and the second as ϕ-bootstrap. The bootstrap sample size always remains fixed, i.e.

$$\sum_{k \in S} S_k^* = n,$$

which implies that

$$\sum_{k \in S} \mathrm{E}^*(S_k^*) = n \text{ and } \sum_{k \in S} \mathrm{cov}^*(S_k^*, S_\ell^*) = 0.$$

4.8 Simulation studies

4.8.1 Comparison with existing variance estimators for the total

In the first part of the simulation study, we examined the performance of the estimator using the proposed method and then compared this estimator with other variance estimators. We ran simulations on the MU284 population from Särndal et al. (1992) from where we selected samples of size $n = 2$, $n = 10$ and $n = 40$ with inclusion probabilities proportional to variable P75 (population in 1975). We used a maximum entropy design (also called conditional Poisson sampling) because this method maximizes the entropy of the sampling design subject to given inclusion probabilities and fixed sample size and can be implemented very quickly (see Tillé, 2006). The variable of interest was RMT85 (revenues from 1985 municipal taxation). We compared the HT-estimator, the SYG-estimator, the H-estimator, the π-bootstrap and the ϕ-bootstrap of the variance for the total of RMT85. We ran 10,000 simulations and, in each of them, we used 10,000 bootstrap replications. Due to the simplicity of the method, the simulations were

achieved in a few hours. Table 4.1 shows the relative bias given by

$$\mathrm{RB} = \frac{\mathrm{E}_{sim}[v_{boot}(\widehat{Y}^*)] - \mathrm{var}(\widehat{Y})}{\mathrm{var}(\widehat{Y})}$$

and the coefficients of variation given by

$$\mathrm{CV} = \frac{\sqrt{\mathrm{var}_{sim}[v_{boot}(\widehat{Y}^*)]}}{\mathrm{var}(\widehat{Y})}$$

of the HT-estimator, SYG-estimator and the Bootstrap estimators, where $\mathrm{E}_{sim}(.)$ and $\mathrm{var}_{sim}(.)$ respectively denote the expectation and the variance under the bootstrap replications. Although only the HT-estimator and the SYG-estimator are strictly unbiased, the relative bias of the π-bootstrap method given by the simulations is still smaller. All the biases computed by simulation are nevertheless very small and are not significantly different from zero. The simulations also show that the HT-estimator is very unstable and that the bootstrap method performs as well as the SYG-estimator and the H-estimator. These simulations show that the bootstrap leads to an estimation of the variance that is at least as efficient as the SYG-estimator even for a very small sample size (n=2).

Table 4.1 – *Relative bias and coefficients of variation of the HT-estimator, the SYG-estimator, H-estimator, the π-bootstrap and the ϕ-bootstrap*

	Estimator	Relative bias in percentages	Coefficients of variation
$n = 2$	HT-estimator	1.78511	1.98240
	SYG-estimator	1.40424	1.80025
	H-estimator	3.56754	1.84788
	π-Bootstrap	3.77461	1.85225
	ϕ-Bootstrap	1.14898	1.80014
$n = 10$	HT-estimator	2.62246	1.31354
	SYG-estimator	0.69995	0.50915
	H-estimator	2.74196	0.53513
	π-Bootstrap	0.76149	0.52311
	ϕ-Bootstrap	0.30085	0.50914
$n = 40$	HT-estimator	-1.53914	1.38550
	SYG-estimator	-0.11598	0.26809
	H-estimator	-0.35775	0.26119
	π-Bootstrap	-0.19534	0.26211
	ϕ-Bootstrap	-0.15363	0.26830

4.8.2 Performance in the case of variance estimation of other functions of interest

In the second part of the simulation study, we ran simulations in order to examine performance in relation to the variance of nonlinear functions of interest. Besides the total, the ratio of two totals, the median and the Gini index were also used as a function of interest. In the case of nonlinear statistics, the variances under the simulations, say the Monte Carlo variances were considered as the true variances of the estimators. A population of 150 units was generated from the model $y_k = (\beta_0 + \beta_1 x_k^{1.2} + \sigma\varepsilon_k)^2 + c$, with $x_k = |i_k|$ and $i_k \sim \mathcal{N}(0,7)$, $\varepsilon_k \sim \mathcal{N}(0,1)$ and $\sigma = 15$. The regression parameters were $\beta_0 = 12.5$, $\beta_1 = 3$ and $c = 4000$. The model and its parameters were chosen intentionally to have a distribution for y similar to a lognormal - as it is often used for income distributions - with a correlated and positive explanatory variable x in the regression model. From this population, 1000 samples were drawn using - as in the previous section - a maximum entropy sampling design with unequal inclusion probabilities. Concerning the inclusion probabilities, they were calculated proportional to the values of a variable z, which was generated from equation $z = y^{0.2}p$ where $p \sim \ln\mathcal{N}(0, 0.25)$. In this manner, the correlation between y and z is about 0.5. We knowingly used a large sample rate $n/N = 1/3$ and a skewed population in order to better illustrate the performance of the tested bootstrap methods. From each of these samples, we calculated four statistics: the total, the median, the Gini index of variable y and the ratio of total of variable y on the total of variable x.

From each of the 1,000 initial samples, 1,000 bootstrap samples were selected using three different bootstrap methods. Besides the new bootstrap method, two other resampling methods were tested. The first one was the rescaled bootstrap of Rao & Wu (1988) and the second one was the generalization of the bootstrap method without replacement proposed by Booth et al. (1994) for unequal inclusion probabilities. This bootstrap method of Booth et al. (1994) is itself a variant of the initial bootstrap without replacement method that consists of creating an artificial population from the initial sample and then drawing bootstrap samples from it with the same design as the initial one (Gross, 1980; Chao & Lo, 1985). After drawing the bootstrap samples, the estimators and their variances were computed for each of the initial samples and then the averages of these variances were then compared with the Monte Carlo approximations of the true variances. Note that the median is not a smooth function of the total. Estimating its variance can therefore be difficult, but the simulations show that in this case bootstrap methods perform well.

In order to measure the performance of the new method and compare it with the other ones, the following five indicators were used:

- Lower error rate (L) in %

$$L = \frac{100}{sim}\sum_{i=1}^{sim} I\left[\widehat{\theta} - 1.96 \times \sqrt{\text{var}(\widehat{\theta}^*)} > \theta\right],$$

  where $I[a] = 1$ if a is true and $I[a] = 0$ elsewhere,

- Upper error rate (U) in %

$$U = \frac{100}{sim}\sum_{i=1}^{sim} I\left[\widehat{\theta} + 1.96 \times \sqrt{\text{var}(\widehat{\theta}^*)} < \theta\right].$$

- Total error rate (ER) in %

$$ER = 100 - \frac{100}{sim}\sum_{i=1}^{sim} I\left[\widehat{\theta} - 1.96 \times \sqrt{\text{var}(\widehat{\theta}^*)} \leq \theta \leq \widehat{\theta} + 1.96 \times \sqrt{\text{var}(\widehat{\theta}^*)}\right].$$

- Relative Bias

$$RB = 100 \times \frac{\text{var}(\widehat{\theta}^*) - \text{var}_{sim}(\widehat{\theta})}{\text{var}_{sim}(\widehat{\theta})} = 100 \times \frac{B}{\text{var}_{sim}(\widehat{\theta})},$$

  where B is the Bias of the $\text{var}(\widehat{\theta}^*)$.

- Relative Root Mean Squared Error

$$RRMSE = 100 \times \frac{\sqrt{B^2 + \text{var}[\text{var}(\widehat{\theta}^*)]}}{\text{var}_{sim}(\widehat{\theta})}.$$

The *RB* gives a measure of the bias of the estimator of variance. The *RRMSE* measures its accuracy and in the case of unbiasedness of the variance estimator it is equal to the coefficient of variation. The *Error Rates* allow us to evaluate the capacity of the methods to provide a valid inference. The lower and the upper error rates give us an idea of how skewed the distribution of the estimator $\widehat{\theta}$ is.

Table 4.2 presents the results of the application of the resampling methods for a maximum entropy design with inclusion probabilities proportional to variable z. In the proposed bootstrap method, the Hájek approximation given in (4.1) is used, which gives us $\phi_k = \pi_k$. One can notice that the rescaled bootstrap clearly performs worse than the two other methods for the error rates and the bias and the relative root mean square errors.

While the new method and the bootstrap without replacement provide confidence intervals around 94% for the total, the ratio and the Gini index, and $89-90\%$ for the median, the error rates of the rescaled bootstrap are 15% for the total, $26-30\%$ for the Gini index and the ratio and nearly 40% for the median. The column of the relative biases directly shows that, in each case of the four functions of interest, the new method and the bootstrap without replacement perform well and give relative biases of $1-3\%$. On the other hand, the rescaled bootstrap method strongly underestimates the variance in each case and gives relative biases of $40-65-70-75\%$ respectively for the total, the Gini index, the ratio and the median.

Note that a high underestimation of the variance of a function of interest could result in a low coverage rate, and therefore high error rates for the function of interest, which is probably the case here. Regarding the relative root mean square errors, the same trend can be observed. The new method and the bootstrap with replacement perform identically, giving a value of RRMSE around $30-40\%$ for the total, the Gini index and the ratio and 60% for the median. The RRMSE's values of the rescaled bootstrap method are substantially higher for each of the functions of interest, essentially because of the high biases. In general, there is no major difference in performance between the proposed method and the method of Booth et al. (1994). The estimators are unbiased, or have a slight bias for each function. The RRMSE have the same order and the error rates show a slightly positively skewed distribution, with coverage rates between 90 and 95%.

Table 4.2 – *Performance of the resampling methods in maximum entropy sampling design*

UPWOR	L	U	ER	Relative bias	RRMSE
TOTAL					
New Method	1.1	6.4	7.5	0.1121	35.4938
RW Bootstrap	4.4	10.7	15.1	-39.0907	49.7916
Bootstrap WOR	1.1	6.9	8.0	-1.6084	34.7805
MEDIAN					
New Method	4.1	6.9	11.0	-1.0564	58.8889
RW Bootstrap	18.9	20.0	38.9	-74.5727	75.7896
Bootstrap WOR	3.4	7.5	10.9	2.8753	61.9642
GINI					
New Method	1.5	5.1	6.6	3.5753	39.4669
RW Bootstrap	7.8	18.7	26.5	-64.3092	65.3181
Bootstrap WOR	1.6	5.1	6.7	-1.0276	30.9325
RATIO					
New Method	2.0	3.7	5.7	2.0403	41.1975
RW Bootstrap	16.5	13.8	30.3	-71.3889	71.9735
Bootstrap WOR	2.1	4.7	6.8	-2.8664	38.2802

The new method thus provides essentially the same results as the boot-

strap with replacement method, but its application is simpler: it does not require a correction factor, rescaling or artificial population. Besides having at least the same performance as the method of artificial populations, its main advantage is that it is easy to implement and fast. Moreover, the bootstrap sample does not need to be rescaled. Thus, the samples can be directly used to compute the variance of the functions of interest.

4.9 Conclusions

These bootstrap methods compare favorably with the best of the classical variance estimates for linear statistics, and also apply to nonlinear statistics. Its simplicity, its rapidity and its efficiency speak in its favour. The bootstrap sample does not need to be reweighted and the observations do not need to be rescaled. There is no need for artificial populations and extreme samples are also avoided because the units can be repeated twice or rarely three times. The bootstrap samples can directly be used to provide an estimation. It can eventually be calibrated, reweighted for nonresponse and imputed as the original sample.

Appendix

If a sample S is selected by a Poisson sampling design with inclusion probabilities π_k in a population U of size N, if n_s denotes the random sample size, then

$$\pi_{k|N-1} = \Pr(k \in S | n_S = N-1) = 1 - \frac{\frac{1-\pi_k}{\pi_k}}{\sum_{\ell \in U} \frac{1-\pi_\ell}{\pi_\ell}}.$$

Proof. We have

$$\Pr(k \notin S \text{ and } n_S = N-1) = (1-\pi_k) \prod_{\ell \neq k} \pi_\ell = \frac{1-\pi_k}{\pi_k} \prod_{\ell \in U} \pi_\ell.$$

Thus

$$\Pr(n_S = N-1) = \sum_{k \in U} \Pr(k \notin \text{ and } n_S = N-1) = \sum_{k \in U} \frac{1-\pi_k}{\pi_k} \prod_{\ell \in U} \pi_\ell,$$

which gives the complementary of the conditional probability of Lemma 1.

$$\Pr(k \notin S | n_S = N-1) = \frac{\Pr(k \notin S \text{ and } n_S = N-1)}{\Pr(n_S = N-1)} = \frac{\frac{1-\pi_k}{\pi_k}}{\sum_{\ell \in U} \frac{1-\pi_\ell}{\pi_\ell}}.$$

Lemma 1. can also be derived from Expression (5.12) of Result 22 in Tillé (2006). □

5

Bootstrap Methods for Two-Phase Sampling with Poisson Design at the Second Phase

Abstract

In order to provide an unbiased estimator of the variance, the most frequently used sampling design in the existing bootstrap methods is simple random sampling with replacement. Nevertheless, when these methods do not take the sampling design into account, they provide biased variance estimators. Resampled units usually need to be rescaled or weighted to correct this bias. Another set of methods consists of constructing artificial populations and to resample from them. These methods are very often time-consuming and have rounding problems. Furthermore, when the sampling design has several phases, implementation of these bootstrap methods becomes very difficult. We present new sampling methods for two-phase sampling with Poisson design at the second-phase. In this paper, only the cases where the second phase design is Poisson are considered. The reason for this restriction is the connection between this type of two-phase design and lots of missing data problems where the non response mechanism is non ignorable. These methods consist of resampling only a subsample of the units in the second phase. This subsample is selected randomly in such a way that it directly reproduces the appropriate variance, without having to rescale or create artificial population. The main advantage of the method is its simplicity, especially for after treatments, such as calibration or imputation for nonresponse. These techniques can be directly applied to bootstrap samples. That is why the proposed method could be particularly worthwhile in real applications.[1]

[1]This chapter is a reprint of: Antal, E. and Tillé, Y. (2012). Bootstrap Methods for Two-Phase Sampling with Poisson Design at the Second Phase. *Submitted.*

Keywords: poisson sampling, two-phase sampling, variance estimation

5.1 Introduction

Resampling methods are very often used to lead inference in survey statistics. However, as the variance depends on the sampling design, bootstrap methods must be adapted to each of them. Moreover, the variance of the Horvitz-Thompson estimator can be different from the variance estimator, especially for complex designs such as those in the family of multi phase sampling designs.

The first bootstrap methods were developed by Efron (1979) for infinite population situations with identical and independent distribution assumption. When the sample is selected without replacement, the units of the sample are not i.i.d. Consequently, the classical methods are not directly applicable. For a finite population, two other families of bootstrap have been developed. Methods based on reconstruction of pseudo-populations from the sample, proposed by Gross (1980) and Chao & Lo (1985), and the family of the rescaled bootstrap (Rao & Wu, 1988) which consists of modifying the values of the interest variable in order to obtain an unbiased variance estimator. Other methods have also been proposed by Mac Carthy & Snowden (1985); Kuk (1989); Rao et al. (1992); Shao & Tu (1995); Sitter (1992a,b); Booth et al. (1994); Holmberg (1998). The method proposed by Beaumont & Patak (2012) directly reconstructs the variance for the linear cases but uses non-integer values to weight the observations in the sample. In this paper, we propose bootstrap methods for specific two-phase designs. They can be viewed as an extension of the method proposed by Antal & Tillé (2011a) based on mixture of discrete multivariate distributions. These methods are simple, use integer weights and directly reproduce the appropriate variance without having to rescale or use pseudo-populations.

The paper is structured as follows: in Section 5.2, the notation for a general two-phase sampling design, the estimator of the total and its variance estimator are defined. Section 5.3 reviews the conditions needed to obtain unbiased bootstrap variance estimates (see Antal & Tillé, 2011a). In Section 5.4, new methods are proposed for special two-phase designs, with Poisson sampling at the second phase, then a simulations are presented in Section 5.5. Finally, conclusions are drawn in Section 5.6.

5.2 Basic notation for two-phase sampling design

Let $p(.)$ be a sampling design on a population $U = \{1, \ldots, k, \ldots, N\}$ of size N such that

$$p(s) \geq 0, \text{ for all } s \subset U, \text{ and } \sum_{s \subset U} p(s) = 1.$$

Let S be the random sample such that $\Pr(S = s) = p(s)$. The sample size n of S can be random or not. Define also the inclusion probabilities $\pi_k = \Pr(k \in S), k \in U$, and the joint inclusion probabilities $\pi_{k\ell} = \Pr(k \text{ and } \ell \in S), k, \ell \in U$. Moreover, define $\Delta_{k\ell} = \pi_{k\ell} - \pi_k \pi_\ell, k \neq \ell \in U$, and $\check{\Delta}_{k\ell} = \Delta_{k\ell}/\pi_{k\ell}$. When $k = \ell$, we obtain $\Delta_{kk} = \pi_k(1 - \pi_k), k \in U$ and $\check{\Delta}_{kk} = 1 - \pi_k$.

If all the inclusion probabilities are positive, then the total $Y = \sum_{k \in U} y_k$ of the values $y_1, \ldots, y_k, \ldots, y_N$ taken by the interest variable y can be unbiasedly estimated by the Horvitz-Thompson estimator (HT) (Horvitz & Thompson, 1952)

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

whose the variance is given by

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

and can be unbiasedly estimated by the Horvitz-Thompson (HT) variance estimator

$$\widehat{\text{var}}_{HT}(\widehat{Y}_\pi) = \sum_{k \in S} \sum_{\ell \in S} \frac{y_k y_\ell}{\pi_k \pi_\ell} \check{\Delta}_{k\ell}. \tag{5.1}$$

This variance estimator is however very unstable because $\sum_{k \in S} \check{\Delta}_{k\ell}$ is generally different from zero. In the case where the sample size is fixed, the Sen-Yates-Grundy (SYG) estimator is also unbiased for this variance (Sen, 1953; Yates & Grundy, 1953)

$$\widehat{\text{var}}_{SYG}(\widehat{Y}_\pi) = \sum_{k \in S} \sum_{\ell \in S} \frac{y_k y_\ell}{\pi_k \pi_\ell} D_{k\ell},$$

where

$$D_{k\ell} = \begin{cases} -\displaystyle\sum_{\substack{j \in S \\ j \neq k}} \frac{\Delta_{kj}}{\pi_{kj}} & \text{if } k = \ell \\ \dfrac{\Delta_{k\ell}}{\pi_{k\ell}} & \text{if } k \neq \ell. \end{cases}$$

In two-phase designs, at the first phase, a sample S_1 is selected by

means of any sampling design $p_1(S_1)$ of expected size n_1. Then at the second phase, a random sample S_2 is selected in S_1 according to another sampling design $p(s_2|S_1) = \Pr(S_2 = s_2|S_1)$ of expected size n_2. The final sample is $S = S_2$ of expected size $n = n_2$. The different notions mentioned above should therefore be defined separately for the two-phases. Let us use the following notation.

$$\pi_{1k} = \Pr(k \in S_1) \quad k \in U,$$

$$\pi_{1k\ell} = \Pr(k \text{ and } \ell \in S_1) \quad k \neq \ell \in U \text{ with } \pi_{1kk} = \pi_{1k},$$

$$\Delta_{1k\ell} = \begin{cases} \pi_{1k}(1-\pi_{1k}) & k = \ell \\ \pi_{1k\ell} - \pi_{1k}\pi_{1\ell} & k \neq \ell. \end{cases}$$

As the second sampling design depends on the first phase, we define conditional probabilities.

$$\pi_{2k} = \Pr(k \in S_2|S_1) \quad k \in U,$$

$$\pi_{2k\ell} = \Pr(k \text{ and } \ell \in S_2|S_1) \quad k \neq \ell \in U \text{ with } \pi_{2kk} = \pi_{2k},$$

$$\Delta_{2k\ell} = \begin{cases} \pi_{2k}(1-\pi_{2k}) & k = \ell \\ \pi_{2k\ell} - \pi_{2k}\pi_{2\ell} & k \neq \ell. \end{cases}$$

Thus π_{2k}, $\pi_{2k\ell}$ and $\Delta_{2k\ell}$ are random variables that depend on S_1. As $\pi_k = \pi_{1k}\mathrm{E}(\pi_{2k})$, the total $Y = \sum_{k\in U} y_k$ can not be estimated by the Horvitz-Thompson estimator because π_{2k} is random.

The unbiased estimator generally used is the expansion estimator defined as

$$\widehat{Y}_E = \sum_{k\in S_2} \frac{y_k}{\pi_{1k}\pi_{2k}}.$$

This estimator is unbiased (see Särndal & Swensson, 1987) and its variance is given by

$$\mathrm{var}(\widehat{Y}_E) = \sum_{k\in U}\sum_{\ell\in U} \frac{y_k y_\ell}{\pi_{1k}\pi_{1\ell}}\Delta_{1k\ell} + \mathrm{E}\Big(\sum_{k\in S_1}\sum_{\ell\in S_1} \frac{y_k y_\ell}{\pi_{1k}\pi_{2k}\pi_{1\ell}\pi_{2\ell}}\Delta_{2k\ell}\Big),$$

which can be estimated unbiasedly by

$$\widehat{\mathrm{var}}(\widehat{Y}_E) = \sum_{k\in S_2}\sum_{\ell\in S_2} \frac{y_k y_\ell}{\pi_{1k}\pi_{1\ell}}\frac{\Delta_{1k\ell}}{\pi_{1k\ell}\pi_{2k\ell}} + \sum_{k\in S_2}\sum_{\ell\in S_2} \frac{y_k y_\ell}{\pi_{1k}\pi_{2k}\pi_{1\ell}\pi_{2\ell}}\frac{\Delta_{2k\ell}}{\pi_{2k\ell}}.$$

This variance estimator can be written as

$$\widehat{\mathrm{var}}(\widehat{Y}_E) = \sum_{k\in S_2}\sum_{\ell\in S_2} \frac{y_k y_\ell}{\pi_{1k}\pi_{2k}\pi_{1\ell}\pi_{2\ell}}\Big(\frac{\Delta_{1k\ell}}{\pi_{1k\ell}}\frac{\pi_{2k}\pi_{2\ell}}{\pi_{2k\ell}} + \frac{\Delta_{2k\ell}}{\pi_{2k\ell}}\Big),$$

or

$$\widehat{\text{var}}(\widehat{Y}_E) = \sum_{k\in S_2}\sum_{\ell\in S_2}\frac{y_k y_\ell}{\pi_{1k}\pi_{2k}\pi_{1\ell}\pi_{2\ell}}\breve{\Delta}_{GG.k\ell}$$

where

$$\breve{\Delta}_{GG.k\ell} = \frac{\Delta_{1k\ell}}{\pi_{1k\ell}}\frac{\pi_{2k}\pi_{2\ell}}{\pi_{2k\ell}} + \frac{\Delta_{2k\ell}}{\pi_{2k\ell}}$$

or

$$\breve{\Delta}_{GG.k\ell} = \frac{\Delta_{1k\ell}}{\pi_{1k\ell}} + \frac{\Delta_{2k\ell}}{\pi_{2k\ell}}\frac{\pi_{1k}\pi_{1\ell}}{\pi_{1k\ell}}.$$

We would like to reconstruct this variance estimator using bootstrap technics.

5.3 Bootstrap

A bootstrap sample is a random sample with replacement selected from S. Let S_{kB} be the number of times that unit k is repeated in the bootstrap sample. The bootstrap estimator of the total is given by

$$\widehat{Y}_B = \sum_{k\in S}\frac{y_k}{\pi_k}S_{kB}.$$

Antal & Tillé (2011a) gave the necessary and sufficient condition for the unbiasedness of this estimator

$$\text{E}^*(S_{kB}) = 1 \quad k \in S, \tag{5.2}$$

where $\text{E}^*(.) = \text{E}(.|S)$ is the expectation subject to the original sample. The conditional variance of the bootstrap estimator is

$$\text{var}^*(\widehat{Y}_B) = \sum_{k\in S}\sum_{\ell\in S}\frac{y_k y_\ell}{\pi_k\pi_\ell}\Sigma_{k\ell}, \tag{5.3}$$

where

$$\Sigma_{k\ell} = \begin{cases} \text{var}(S_{kB}|S) = \text{var}^*(S_{kB}) & \text{if } k = \ell \\ \text{cov}(S_{kB}, S_{\ell B}|S) = \text{cov}^*(S_{kB}, S_{\ell B}) & \text{if } k \neq \ell. \end{cases}$$

Thus, the condition for the variance given in (5.1) to be equal to the variance estimator given in (5.3) is

$$\Sigma_{k\ell} = \breve{\Delta}_{k\ell}.$$

In fact, condition (5.3) involves two different conditions, one for the variance and the other for the covariance. The most important condition from

these two is the condition applying to the variance, given by

$$\text{var}^*(S_{kB}) = \check{\Delta}_{kk} = 1 - \pi_k \quad k \in S. \tag{5.4}$$

Ideally, the conditions for the covariance

$$\text{cov}^*(S_{kB}, S_{\ell B}) = \check{\Delta}_{k\ell} \quad k \neq \ell \in S, \tag{5.5}$$

should also be satisfied. However, in the case where the sample is selected with a fixed sample size and unequal inclusion probabilities, it is difficult to exactly satisfy more than conditions (5.2) and (5.4). Condition (5.5) can nevertheless be approximately satisfied. Another way to construct an unbiased estimator of the variance is to equate the bootstrap estimator to the SYG-estimators. The conditions become

$$\text{E}^*(S_{kB}) = 1 \quad k \in S,$$

$$\text{var}^*(S_{kB}) = D_{kk} \quad k \in S,$$

and

$$\text{cov}^*(S_{kB}, S_{\ell B}) = D_{k\ell} \quad k \neq \ell \in S.$$

Again, the conditions on the covariances are difficult to meet when the sample is selected with a fixed sample size and unequal inclusion probabilities, but they can however be approximately satisfied. With the notation of the two-phase design, these three conditions are

$$\text{E}^*(S_{kB}) = 1 \quad k \in S, \tag{5.6}$$

$$\text{var}^*(S_{kB}) = \check{\Delta}_{GG.kk} = 1 - \pi_{1k} + \pi_{1k}(1 - \pi_{2k}) \quad k \in S, \tag{5.7}$$

and

$$\text{cov}^*(S_{kB}, S_{\ell B}) = \check{\Delta}_{GG.k\ell} = 1 - \frac{\pi_{1k}\pi_{2k}\pi_{1\ell}\pi_{2\ell}}{\pi_{1k\ell}\pi_{2k\ell}} \quad k \neq \ell \in S. \tag{5.8}$$

The bootstrap estimator of the variance of a statistic of interest, $\widehat{\text{var}}_{boot}(\widehat{\theta})$, is computed by generating a set of bootstrap samples and then computing the $\text{var}(\widehat{\theta}^*)$, the variance of the outcomes of $\widehat{\theta}^*$. Moreover, if a bootstrap method provides an approximately unbiased estimator for the variance of totals, it will also provide approximately unbiased variance estimators for smooth functions of totals.

5.4 Poisson design at the second phase

As mentioned earlier, in this paper only the two-phase sampling with Poisson design at the second phase will be considered. In fact, a general data set - containing missing data - can always be seen as a first phase sample S_1 and its respondent part as the final sample $S_2 = S$. If the probability of non-response is supposed to be proportional to the values of a measured auxiliary variable, the observed part of the sample can be considered as a Poisson sample from the first sample, i.e. a two-phase sampling with Poisson design at the second phase.
As no assumption is made for the first phase design, the notation the first and second order inclusion probabilities and the matrix of elements $\Delta_{1k\ell}$ remain the same as before. However, as in Poisson sampling it holds that $\pi_{k\ell} = \pi_k \pi_\ell$, the matrix $\Delta_{2k\ell}$ can be simplified as follows

$$\Delta_{2k\ell} = \begin{cases} \pi_{2k}(1-\pi_{2k}), & k = \ell; \\ 0, & k \neq \ell. \end{cases}$$

The estimator of the variance of the expansion estimator of the total is also simplified as

$$\widehat{\text{var}}(\widehat{Y}_E) = \sum_{k \in S_2} \sum_{\ell \in S_2} \frac{y_k y_\ell}{\pi_{1k}\pi_{2k}\pi_{1\ell}\pi_{2\ell}} \check{\Delta}_{GP.k\ell}$$

where $\check{\Delta}_{GP.k\ell}$ can be written as

$$\check{\Delta}_{GP.k\ell} = \begin{cases} 1 - \pi_{1k}\pi_{2k} = (1-\pi_{1k}) + \pi_{1k}(1-\pi_{2k}) & \text{if } k = \ell; \\ 1 - \frac{\pi_{1k}\pi_{1\ell}}{\pi_{1k\ell}} & \text{if } k \neq \ell. \end{cases}$$

The conditions (5.6), (5.7) and (5.8), for a bootstrap design that estimates without bias the variance estimator of the total, becomes

$$\text{E}^*(S_{kB}) = 1 \quad k \in S, \tag{5.9}$$

$$\text{var}^*(S_{kB}) = \check{\Delta}_{GP.kk} = (1-\pi_{1k}) + \pi_{1k}(1-\pi_{2k}) \quad k \in S, \tag{5.10}$$

and

$$\text{cov}^*(S_{kB}, S_{\ell B}) = \check{\Delta}_{GP.k\ell} = 1 - \frac{\pi_{1k}\pi_{1\ell}}{\pi_{1k\ell}} \quad k \neq \ell \in S. \tag{5.11}$$

5.4.1 Bootstrap for a two-phase design with Poisson sampling at the second phase

When a Poisson sampling design is applied in the second phase, the condition on the covariance (5.8) does not depend on the second phase. The main idea here is to apply a bootstrap for the corresponding first-phase design. For example, if a simple random sampling without replacement (srswor) design is used, apply a bootstrap method for srswor that reproduces the variance estimator of the total for the first phase (see for example in Antal & Tillé, 2011a), and then complete it with a random variable that adds the missing part of the variance to it. We introduce two different strategies to construct such a random variable.

First proposal

If the bootstrap method for the sampling design used in the first phase is supposed to be known and the variable that reproduce $\Delta_{1k\ell}/\pi_{1k\ell}$ is S_{kB1}, then the variable S_{kB} is

$$S_{kB} = S_{kB1}S_{kB2},$$

where S_{kB2} is independent from S_{kB1}. The conditional expectation, the conditional variance and covariance can be calculated, as follows

$$\mathrm{E}^*(S_{kB}) = \mathrm{E}^*(S_{kB1}S_{kB2}) = \mathrm{E}^*(S_{kB1})\mathrm{E}^*(S_{kB2}) = 1 \times \mathrm{E}^*(S_{kB2}) \quad k \in S,$$

$$\begin{aligned}
\mathrm{var}^*(S_{kB}) &= \mathrm{var}^*(S_{kB1}S_{kB2}) \\
&= \mathrm{var}^*(S_{kB1})\mathrm{var}^*(S_{kB2}) + \mathrm{var}^*(S_{kB1})\mathrm{E}^{*2}(S_{kB2}) \\
&\quad +\mathrm{var}^*(S_{kB2})\mathrm{E}^{*2}(S_{kB1}) \\
&= (2-\pi_{1k})\mathrm{var}^*(S_{kB2}) + 1 - \pi_{1k} \quad k \in S,
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{cov}^*(S_{kB}, S_{\ell B}) &= \mathrm{cov}^*(S_{kB1}S_{kB2}, S_{\ell B1}S_{\ell B2}) \\
&= \mathrm{E}^*(S_{kB1}S_{\ell B1})\mathrm{E}^*(S_{kB2})\mathrm{E}^*(S_{\ell B2}) \\
&\quad -\mathrm{E}^*(S_{kB1})\mathrm{E}^*(S_{kB2})\mathrm{E}^*(S_{\ell B1})\mathrm{E}^*(S_{\ell B2}) \\
&= 1 - \frac{\pi_{1k}\pi_{1\ell}}{\pi_{1k\ell}} \quad k \neq \ell \in S.
\end{aligned}$$

If we compare these results with conditions (5.9), (5.10) and (5.11), we see that S_{kB2} has to satisfy

$$\mathrm{E}^*(S_{kB2}) = 1 \quad k \in S, \tag{5.12}$$

$$\text{var}^*(S_{kB2}) = \frac{\pi_{1k}(1-\pi_{2k})}{2-\pi_{1k}} \quad k \in S, \tag{5.13}$$

and

$$\text{cov}^*(S_{kB2}, S_{\ell B2}) = 0 \quad k \neq \ell \in S. \tag{5.14}$$

Let us define the probability distribution of S_{kB2} (independent from $S_{\ell B2}$) as

$$S_{kB2} = \begin{cases} 0 & \text{with a probability } q/2 \\ 1 & \text{with a probability } 1-q \\ 2 & \text{with a probability } q/2, \end{cases}$$

where $q = \pi_{1k}(1-\pi_{2k})/(2-\pi_{1k})$.

After some algebra, we can verify that the expectation, the variance and the covariances between S_{kB2} and S_{lB2} are exactly (5.12), (5.13) and (5.14). Consequently, the conditions (5.6), (5.7) and (5.8) are satisfied for S_{kB}.

$$\text{E}^*(S_{kB}) = \text{E}^*(S_{kB1})\text{E}^*(S_{kB2}) = 1 \times 1 = 1, \quad k \in S,$$

$$\begin{aligned} \text{var}^*(S_{kB}) &= \text{var}^*(S_{kB1}S_{kB2}) = (1-\pi_{1k}) + \pi_{1k}(1-\pi_{2k}) \\ &= \check{\Delta}_{GP.kk} \quad k \in S, \end{aligned}$$

and

$$\begin{aligned} \text{cov}^*(S_{kB}, S_{\ell B}) &= \text{cov}^*(S_{kB1}S_{kB2}, S_{\ell B1}S_{\ell B2}) = 1 - \frac{\pi_{1k}\pi_{1\ell}}{\pi_{1k\ell}} \\ &= \check{\Delta}_{GP.k\ell}, \quad k \neq \ell \in S. \end{aligned}$$

Second proposal

In the first proposition, we have defined a variable S_{kB} that satisfies the bootstrap conditions as the product of two independent variables S_{kB1} and S_{kB2}. The first one, S_{kB1}, is the variable that reproduces the variance of the total estimator due to the first phase (which is supposed to be known) and S_{kB2} is the variable that adds the missing part of the variance in order to satisfy conditions (5.9), (5.10) and (5.11). Here we consider an S_{kB2} variable, which is not independent of S_{kB1}. To find this variable, we define S_{B1L} as the list of the units selected in the first phase sample of the bootstrap. For example, if the initial sample is S=(1,3,5,6,8) and the variable S_{kB1}=(0,1,2,1,1), the list of the units S_{B1L}=(3,5,5,6,8). The second variable S_{kB2} is defined on this list on the first phase of the bootstrap strategy, therefore the bootstrap estimator will also depend on this set. The

estimator

$$\widehat{Y}_B = \sum_{S_{B1L}} \frac{y_k}{\pi_k} S_{kB2}$$

is the total estimator defined on the bootstrap sample with S_{kB2} showing the times that unit k is in this sample S_{B1L}. The variance of the bootstrap estimator is

$$\begin{aligned} \operatorname{var}(\widehat{Y}_B) &= \mathrm{E}\Big[\operatorname{var}(\widehat{Y}_B|S_{B1L})\Big] + \operatorname{var}\Big[\mathrm{E}(\widehat{Y}_B|S_{B1L})\Big] \\ &= \sum_{S_{B1L}} \frac{y_k}{\pi_k} \operatorname{var}(S_{kB2}|S_{B1L}) + \sum_{S_{B1L}} \frac{y_k}{\pi_k} \operatorname{var}(S_{kB1}) \\ &= \sum_{S_{B1L}} \frac{y_k}{\pi_k} \Big(\operatorname{var}(S_{kB2}|S_{B1L}) + 1 - \pi_{1k}\Big). \end{aligned}$$

In order to satisfy condition (5.10), we should have

$$\operatorname{var}(S_{kB2}|S_{B1L}) + 1 - \pi_{1k} = (1 - \pi_{1k}) + \pi_{1k}(1 - \pi_{2k}).$$

thus

$$\operatorname{var}(S_{kB2}|S_{B1L}) = \pi_{1k}(1 - \pi_{2k}).$$

Finally, the variable S_{kB2} has to satisfy the following properties

$$\mathrm{E}(S_{kB2}|S_{B1L}) = 1$$

and

$$\operatorname{var}(S_{kB2}|S_{B1L}) = \pi_{1k}(1 - \pi_{2k}).$$

Let us define the probability distribution of S_{kB2} as

$$S_{kB2}|S_{B1L} = \begin{cases} 0 & \text{with a probability } q/2 \\ 1 & \text{with a probability } 1 - q \\ 2 & \text{with a probability } q/2, \end{cases}$$

where $q = \pi_{1k}(1 - \pi_{2k})$. The expectation of this variable is equal to 1 and the conditional variance is also the required, $\operatorname{var}(S_{kB2}|S_{B1L}) = \pi_{1k}(1 - \pi_{2k})$.

5.5 Simulation study

In order to examine the performance of the proposed methods, simulations were ran. Besides the total, another function of interest, the Gini

index, was also calculated. In the case of the total, the variance estimator, and in the case of the Gini index, the variances under the simulations, say the Monte Carlo variances, were considered as the true variances of the estimators. A population of $N = 1500$ units was generated from the model $y_k = (\beta_0 + \beta_1 x_k^{1.2} + \sigma\varepsilon_k)^2 + c$, with $x_k = |i_k|$ and $i_k \sim \mathcal{N}(0,7)$, $\varepsilon_k \sim \mathcal{N}(0,1)$ and $\sigma = 15$. The regression parameters are $\beta_0 = 12.5$, $\beta_1 = 3$ and $c = 4000$. The model and its parameters were chosen intentionally to have a distribution for y similar to a lognormal (as it is often used for income distributions) with a correlated and positive explanatory variable x in the regression model.

From this population, 500 samples were drawn using given two-phase sampling with Poisson design at the second phase. Concerning the first phase, two different sampling designs were used. The first one was the simple random sampling design and the second one was the brewer design (Brewer, 1975) with inclusion probabilities proportional to the values of a variable z_1. This design was chosen among the sampling designs with unequal inclusion probabilities thanks to its simplicity and computational rapidity. The z_1 variable was generated from equation $z_1 = y^{0.3}p$, where $p \sim \ln\mathcal{N}(0, 0.25)$. Concerning the unequal inclusion probabilities used for the second phase (Poisson sampling), a second variable z_2, generated from the equation $z_2 = y^2 q$ where $q \sim \ln\mathcal{N}(0, 0.49)$ was used. In this manner, the correlations between y and z_1 and between y and z_2 are about 0.5. We knowingly used a skewed population and a large sample rate $1/3$ in the first phase, in order to better illustrate the performance of the tested bootstrap methods.

The second phase consists of decreasing 20% of the sample size, which can be viewed as a 20% non-response rate. From each of the final samples, the estimator of the total, its variance estimator and the estimator of the Gini index were calculated. At each step, 500 bootstrap samples were selected by means of the two proposed bootstrap methods (Method 1 and 2). For comparison purposes the performance of an existing resampling method was also tested. This method is the generalization of the bootstrap without replacement proposed by Booth et al. (1994) for unequal inclusion probabilities (BWOR). This method is a variant of the initial bootstrap with replacement method (Gross, 1980; Chao & Lo, 1985), which consists of creating an artificial population from the sample and then drawing bootstrap samples from it with the same design as the initial one.

After drawing the bootstrap samples, the bootstrap estimator of the total and the Gini index and their variances were computed for each of the ini-

tial samples. Finally, the means of these variances were compared to the approximations of the true variances.

In order to measure the performance of the new methods, the following four indicators were used:

- Lower error rate (L) in %

$$L = \frac{100}{sim}\sum_{i=1}^{sim} \mathbb{1}\left[\widehat{\theta} - 1.96 \times \sqrt{\widehat{\text{var}}(\widehat{\theta})} > \theta\right],$$

- Upper error rate (U) in %

$$U = \frac{100}{sim}\sum_{i=1}^{sim} \mathbb{1}\left[\widehat{\theta} + 1.96 \times \sqrt{\widehat{\text{var}}(\widehat{\theta})} < \theta\right].$$

- Relative Bias

$$RB = 100 \times \frac{\widehat{\text{var}}(\widehat{\theta}) - \text{var}_{sim}(\widehat{\theta})}{\text{var}_{sim}(\widehat{\theta})} = 100 \times \frac{B}{\text{var}_{sim}(\widehat{\theta})},$$

  where B is the Bias of $\widehat{\text{var}}(\widehat{\theta})$.

- Relative Root Mean Squared Error

$$RRMSE = 100 \times \frac{\sqrt{B^2 + \text{var}[\widehat{\text{var}}(\widehat{\theta})]}}{\text{var}_{sim}(\widehat{\theta})}.$$

The *Relative Bias* gives a measure of the bias of the variance estimator. The *RRMSE* measures its accuracy and in the case of non-biased variance estimator it is equal to the variation coefficients. The sum of the lower L and the upper U error rates gives the total error rates which allows the capacity of the methods to provide a valid inference to be evaluated. In addition, comparing the lower and the upper error rates, enables us to analyze the skewness of the distribution of the estimator $\widehat{\theta}$.

Table 5.1 – *The lower error rate (L), the upper error rate (U), the relative bias and the relative root mean squared error (RRMSE) of the tested resampling methods in two-phase Simple random sampling without replacement-Poisson design*

srswor-Poisson	L	U	Relative bias(%)	RRMSE(%)
TOTAL				
Method 1	1.4	3.2	5.3494	19.4767
Method 2	2.0	3.0	-6.9343	16.6099
BWOR	0.0	0.0	218.5604	219.9920
GINI index				
Method 1	1.6	3.4	6.6895	16.7219
Method 2	1.4	3.4	5.4464	16.0379
BWOR	1.4	2.2	16.8604	25.0477

Table 5.2 – *The lower error rate (L), the upper error rate (U), the relative bias and the relative root mean squared error (RRMSE) of the tested resampling methods in two-phase Brewer-Poisson design*

Brewer-Poisson	L	U	Relative bias(%)	RRMSE(%)
TOTAL				
Method 1	0.2	3.4	15.1237	35.3265
Method 2	0.2	3.4	12.3285	32.8453
BWOR	0.0	0.0	8959.1448	140.1256
GINI index				
Method 1	1.4	3.8	4.1287	16.1113
Method 2	1.2	4.0	2.8025	15.0011
BWOR	0.8	1.6	116.8477	178.7661

5.6 Conclusions

The results above show that the two new methods perform much better than the bootstrap without replacement. At first sight, the error rates of the BWOR method seem to be very low. However, this is due to the extremely high overestimation of the variance. In fact, it is not surprising that with a very large estimated variance, the length of the 95% confidence interval also becomes very large. Consequently the Monte Carlo estimators fall in the confidence interval with a probability of quasi 1, inducing an error rate near 0. In general, the bootstrap without replacement provides quite poor results. It is very likely that the rounding problems that often occur with this methods contribute to the overestimation. Concerning the two new methods, their performance is essentially the same. Maybe the second could have a little advantage with a slightly smaller relative bias and RRMSEs. Anyhow, the differences are not important. In three of the four cases, the two methods provide a slightly biased estimators for the variance of the total and the Gini estimators with RRMSEs around 15-20%. When the estimation is almost unbiased, the provided coverage rate is the most important property enabling valid inference.

Both of the new methods give a coverage rate of around 95%. If the two-phase design is a Brewer-Poisson design and the function of interest is the variance of the total estimator, the relative biases are a little bit higher and the RRMSEs are also around 35%. However, they remain much smaller than that of the BWOR method. The simplicity, the rapidity and efficiency of the proposed method speak in its favor. Moreover, the bootstrap sample does not need to be reweighted, the observations do not need to be rescaled. No artificial populations need to be generated and the bootstrap samples can directly be used to provide an estimation. the bootstrap samples can eventually be calibrated, reweighted for nonresponse and imputed as the original sample. The variances of nonlinear statistics can also be estimated directly.

General conclusion

As mentioned at various points in this thesis, the variance estimation of an estimator is a fundamental question in survey sampling. This topic is important because it enables us to measure the accuracy of the estimation. Even if an estimator is unbiased, a confidence interval around the point estimation is generally needed. In order to provide appropriate confidence intervals, the variance of the estimator must be known, or must be estimated without bias.

Nevertheless, there are many different sampling designs with different characteristics and there are also several estimators with different properties. The way the sample is selected depends on numerous factors. Consequently, the choice of estimator must be based on various criteria. Thus, there are uncountable situations. It is almost impossible to develop a method that provides the best solution for each sampling design, for each estimator and for each point of view.

However, among these sampling designs and estimators there are some that are used much more frequently in practice than others, such as simple random sampling without replacement or Poisson design. While the former uses equal sampling inclusion probabilities, the latter presumes that these probabilities are different and proportional to the values of a given variable. Both situations are frequent. Equal probabilities can be justifiable if no auxiliary information is available. However, the assumption of unequal probabilities is more realistic. The simplest sampling design using unequal inclusion probabilities is the Poisson design. The main disadvantage of this design is its random sampling size. This is the reason why sampling designs with unequal inclusion probabilities providing a fix sample size were developed.

The most usual parameters of interest to estimate are totals and functions of totals. Recently greater attention has been paid to non-smooth functions as quantiles or inequality indexes. In the second part of Chapter 1 some of the numerous variance estimation methods are presented. Except linearization, most methods apply replication procedures, such as the new algorithms presented in this thesis.

The main objective of this thesis was to develop new methods that can provide an unbiased variance estimator for most sampling designs, particularly the ones mentioned above. As there are many different situations, criteria and preferences, it is almost impossible to determine an optimal method. We do not pretend to have found the best variance estimation method. We have only proposed a new one that may be preferable to the others, in some respects.

The basic method presented in Chapter 3 provides a good estimate for the variance of estimators in several sampling designs. Unlike the plug-in type bootstrap algorithms, this method does not require an artificial population to be created, avoiding thereby handling with rounding problems. It takes the sampling design into account through the inclusion probabilities, thus rescaling or reweighting are not needed to create unbiased estimators. In this way, it is directly applicable for the usual after-treatments, such as weighting, calibration or imputation for non-response. Moreover, a particularly worthwhile feature of the proposed method is that it always furnishes positive and integer weights. In Chapter 4 a simpler alternative is presented.

In Chapter 5, we present an extension of the basic method developed in the previous chapters. An extension for a special two-phase design is also explained. Usually, when the sampling design is composed of several phases, the variance formula could be very complicated and implementation of bootstrap methods could become difficult. Due to the connection to missing data problems, the two-phase designs are particularly important among the complex sampling designs. In fact, a sample with a respondent and a non-respondent part can always be seen as a two-phase sample, where the second phase sample is the respondent part. In this thesis, variance estimation methods were developed for a sampling design which has a Poisson design in the second phase. Further investigation could be worthwhile, supposing another sampling design for the second phase, depending on the assumptions made on the non-response mechanism.

The new methods proposed in this thesis may seem complex. However, they all have the interesting advantage that, once the bootstrap samples have been selected, the inference is very simple. Indeed, the bootstrap samples are similar to the original one. They have the same size and the same characteristic as the original sample. The main trick consists in using a bootstrap design that is different from the original design. The bootstrap samples can thus be directly used to conduct the inference. There is no need for rescaling or reweighting of the observations.

There still are several interesting lines of research. For instance, it could be interesting to identify appropriate bootstrap methods for balanced sampling (Deville & Tillé, 2004). Indeed, a bootstrap method could capture part of the variance since a sample can rarely be exactly balanced. Another line of future research could be to develop methods based on the same basic idea, but for repeated surveys.

Bibliography

Antal, E. & Tillé, Y. (2011a). A direct bootstrap method for complex sampling designs from a finite population. *Journal of the American Statistical Association* **106**, 534–543. (Cited pages 47, 74, 78, 79, 82, 94, 97, and 100.)

Antal, E. & Tillé, Y. (2011b). Simple random sampling with over-replacement. *Journal of Statistical Planning and Inference* **141**, 597–601. (Cited pages 55 and 80.)

Beaumont, J.-F. & Patak, Z. (2012). Generalized bootstrap for prices surveys. *International statistical Review* **80**, 127–148. (Cited pages 32, 35, 66, 68, 72, 74, 78, and 94.)

Berger, Y. G. (1998). Variance estimation using list sequential scheme for unequal probability sampling. *Journal of Official Statistics* **14**, 315–323. (Cited page 53.)

Bertail, P. & Combris, P. (1997). Bootstrap généralisé d'un sondage. *Annales d'Economie et de Statistique* **46**, 49–83. (Cited page 74.)

Bickel, P. J. & Freedman, D. A. (1984). Asymptotic normality and the bootstrap in stratified sampling. *Annals of Statistics* **12**, 470–482. (Cited page 27.)

Bol'shev, L. N. (1965). On a characterization of the Poisson distribution. *Teoriya Veroyatnostei i ee Primeneniya* **10**, 64–71. (Cited page 44.)

Booth, J. G., Butler, R. W. & Hall, P. (1994). Bootstrap methods for finite populations. *Journal of the American Statistical Association* **89**, 1282–1289. (Cited pages 32, 50, 74, 88, 90, 94, and 103.)

Brewer, K. R. W. (1975). A simple procedure for πpswor. *Australian Journal of Statistics* **17**, 166–172. (Cited pages 54 and 103.)

Brewer, K. R. W. & Donadio, M. E. (2003). The high entropy variance of the Horvitz-Thompson estimator. *Survey Methodology* **29**, 189–196. (Cited pages 53 and 83.)

Brewer, K. R. W. & Hanif, M. (1983). *Sampling with Unequal Probabilities*. New York: Springer. (Cited page 83.)

Chao, M.-T. & Lo, S.-H. (1985). A bootstrap method for finite population. *Sankhyā* **A47**, 399–405. (Cited pages 50, 66, 74, 88, 94, and 103.)

Davison, A. C. & Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge University Press. (Cited pages 33 and 37.)

Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: linearization and residual techniques. *Survey Methodology* **25**, 193–204. (Cited page 29.)

Deville, J.-C. & Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika* **91**, 893–912. (Cited page 109.)

Deville, J.-C. & Tillé, Y. (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference* **128**, 569–591. (Cited page 54.)

Dippo, C. S., Fay, R. E. & Morganstein, D. H. (1984). Computing variances from complex samples with replicate weights. In *Proceedings of the Section on Survey Research Methods*. Washington DC: American Statistical Association. (Cited page 31.)

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics* **7**, 1–26. (Cited pages 32, 50, 74, and 94.)

Efron, B. (1981). Nonparametric standard errors and confidence intervals (with discussion). *Canadian Journal of Statistics* **9**, 139–172. (Cited page 36.)

Efron, B. (1982). *The jackknife, the Bootstrap and Other Resampling Plans*, vol. 38. ACBMS-N SIAM. (Cited page 36.)

Efron, B. & Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman and Hall Ltd. (Cited page 33.)

Gross, S. T. (1980). Median estimation in sample surveys. In *ASA Proceedings of the Section on Survey Research Methods*. American Statistical Association. (Cited pages 50, 66, 74, 88, 94, and 103.)

Hájek, J. (1960). Limiting distributions in simple random sampling from finite population. *Matematikai Kutatï¿½ Intézeténk közleményei (Publication of the Mathematical Institute of the Hungarian Academy of Sciences)* **A5**, 361–374. (Cited page 27.)

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics* **35**, 1491–1523. (Cited pages 24 and 27.)

Hájek, J. (1981). *Sampling from a Finite Population*. New York: Marcel Dekker. (Cited pages 54 and 76.)

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. New York: Springer-Verlag. (Cited page 33.)

Hansen, M. H. & Hurwitz, W. N. (1949). On the determination of the optimum probabilities in sampling. *Annals of Mathematical Statistics* **20**, 426–432. (Cited page 42.)

Henderson, T. (2006). *Estimating the variance of the Horvitz-Thompson estimator*. Master's thesis, School of Finance and Applied Statistics, The Australian National University. (Cited pages 53 and 83.)

Holmberg, A. (1998). A bootstrap approach to probability proportional-to-size sampling. In *ASA Proceedings of the Section on Survey Research Methods*. American Statistical Association. (Cited pages 50, 72, 74, and 94.)

Horvitz, D. G. & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47**, 663–685. (Cited pages 22, 42, 75, and 95.)

Johnson, N. L., Kotz, S. & Balakrishnan, N. (1997). *Discrete Multivariate Distributions*. New York: Wiley. (Cited page 44.)

Johnson, N. L., Kotz, S. & Kemp, A. W. (1992). *Univariate Discrete Distributions*. New York: Wiley. (Cited page 45.)

Krewski, D. & Rao, J. N. K. (1981). Inference from stratified samples: properties of linearization, jackknife and balanced repeated replication methods. *Annals of Statistics* **9**, 1010–1019. (Cited page 27.)

Kuk, A. Y. C. (1989). Double bootstrap estimation of variance under systematic sampling with probability proportional to size. *Journal of Statistical Computation and Simulation* **31**, 73–82. (Cited pages 50, 74, and 94.)

Lahiri, P. (2003). On the impact of bootstrap in survey sampling and small-area estimation. *Statistical Science* **18**, 199–210. (Cited page 74.)

Mac Carthy, P. J. (1969). Pseudo-replication: Half samples. *Review of the International Statistics Institute* **37**, 239–264. (Cited page 31.)

Mac Carthy, P. J. & Snowden, C. B. (1985). The bootstrap and finite population sampling. Tech. rep., Public Health Service Publication. (Cited pages 32, 34, 50, 66, 74, and 94.)

Madow, W. G. (1949). On the theory of systematic sampling, II. *Annals of Mathematical Statistics* **20**, 333–354. (Cited page 54.)

Matei, A. & Tillé, Y. (2005). Evaluation of variance approximations and estimators in maximum entropy sampling with unequal probability and fixed sample size. *Journal of Official Statistics* **21**, 543–570. (Cited pages 54, 62, 76, and 83.)

Presnell, B. & Booth, J. G. (1994). Resampling methods for sample surveys. Tech. rep., Department of Statistics, University of Florida, Gainesville,2008. 04. 22, Technical Report 470. (Cited pages 32, 33, and 35.)

Preston, J. & Henderson, T. (2007). Replicate variance estimation and high entropy variance approximations. In *Papers presented at the ICES-III, June 18-21, 2007, Montreal, Quebec, Canada*. (Cited page 83.)

Quenouille, M. H. (1949). Approximation tests of correlation in time series. *Journal of the Royal Statistical Society* **B11**, 18–84. (Cited page 29.)

Rao, J. N. K. (1965). On two simple schemas of unequal probability sampling without replacement. *Journal of the Indian Statistical Association* **3**, 173–180. (Cited page 54.)

Rao, J. N. K. & Wu, C. F. J. (1985). Inference from stratified samples: Second-order analysis of three methods for nonlinear statistics. *Journal of the American Statistical Association* **80**, 620–630. (Cited page 30.)

Rao, J. N. K. & Wu, C. F. J. (1988). Resampling inference for complex survey data. *Journal of the American Statistical Association* **83**, 231–241. (Cited pages 32, 34, 50, 66, 68, 72, 74, 88, and 94.)

Rao, J. N. K., Wu, C. F. J. & Yue, K. (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology* **18**, 209–217. (Cited pages 50, 74, and 94.)

Rosén, B. (1972a). Asymptotic theory for successive sampling I. *Annals of Mathematical Statistics* **43**, 373–397. (Cited page 27.)

Rosén, B. (1972b). Asymptotic theory for successive sampling II. *Annals of Mathematical Statistics* **43**, 748–776. (Cited page 27.)

Sampford, M. R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika* **54**, 499–513. (Cited page 54.)

Särndal, C.-E. & Swensson, B. (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and non-response. *International Statistical Review* **55**, 279–294. (Cited page 96.)

Särndal, C.-E., Swensson, B. & Wretman, J. H. (1992). *Model Assisted Survey Sampling*. New York: Springer. (Cited pages 18, 20, 28, and 86.)

Sen, A. R. (1953). On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics* **5**, 119–127. (Cited pages 23, 76, and 95.)

Shao, J. & Tu, D. (1995). *The Jacknife and Bootstrap*. New York: Sprinter-Verlag. (Cited pages 33, 35, 50, 74, and 94.)

Shao, M. T. & Wu, C. F. J. (1989). A general theory for jackknife variance estimation. *Annals of Statistics* **17**, 1176–1197. (Cited page 30.)

Shao, M. T. & Wu, C. F. J. (1992). Asymptotic properties of the balanced repeated replication method for sample quantiles. *Annals of Statistics* **20**, 1571–1593. (Cited page 32.)

Sitter, R. R. (1992a). Comparing three bootstrap methods for survey data. *Canadian Journal of Statistics* **20**, 135–154. (Cited pages 35, 50, 74, and 94.)

Sitter, R. R. (1992b). A resampling procedure for complex survey data. *Journal of the American Statistical Association* **87**, 755–765. (Cited pages 50, 74, and 94.)

Sitter, R. R. (1992a). Comparing three bootstrap methods for survey data. *Canadian Journal of Statistics* **20**, 135–154. (Cited page 32.)

Sitter, R. R. (1992b). A resampling procedure for complex survey data. *Journal of the American Statistical Association* **87**, 755–765. (Cited page 34.)

Tillé, Y. (2001). *Théorie des sondages: échantillonnage et estimation en populations finies*. Paris: Dunod. (Cited page 20.)

Tillé, Y. (2006). *Sampling Algorithms*. New York: Springer. (Cited pages 42, 44, 46, 54, 83, 85, 86, and 92.)

Tukey, J. W. (1958). Bias and confidence in not quiet large samples. *Annals of Mathematical Statistics* **29**, 614. (Cited page 29.)

Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis (with discussions). *Annals of Statistics* **14**, 1261–1350. (Cited page 35.)

Wu, C. F. J. (1990). On the asymptotic properties of the jackknife histogram. *Annals of Statistics* **18**, 1438–1452. (Cited page 35.)

Wu, C. F. J. (1991). Balanced repeated replications based on mixed orthogonal arrays. *Biometrika* **78**, 181–188. (Cited page 32.)

Yates, F. & Grundy, P. M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society. Series B. Methodological* **15**, 235–261. (Cited pages 23, 76, and 95.)