Sapienza University of Rome
Department of Statistical Sciences

Ph.D. in School of Statistical Sciences

Estimation methods for data from nonprobability samples

Candidate:
Simona Rosati

Thesis advisor:
Prof. Pier Luigi Conti

Thesis submitted in 2019

Abstract

The main goal of the present dissertation is to evaluate the asymptotic behaviour of estimators for data from nonprobability samples. In this context some target population units do not have positive inclusion probabilities, which means that estimation is affected by biases associated with under-coverage or self-selection errors. For this purpose, we aim at developing a model for the mechanism which caused self-selection in order to estimate the inclusion probabilities for each unit. In this way, pseudo estimators which mimic classical ones can be constructed. More specifically, pseudo Horvitz-Thompson and Hájek estimators are proposed, where propensity score plays the role of inclusion probability. We show that weighting by the inverse of nonparametric estimate of the propensity score leads to an efficient estimate of the population mean. Resampling techniques are used to study the variance asymptotic behaviour and to address the issue of its estimation. A simulation study is carried out in order to assess the validity of the proposed methodology.

Contents

Introduction

This study aims at investigating inferential potential of data from nonprobability samples. It is well known how the traditional surveys are increasingly replaced by web surveys, since they are less expensive, quicker and get easily access to a large number of respondents. There are, however, two phenomena that can make unreliable the results of web surveys: *under-coverage* and *self-selection*. The quality of web surveys may be seriously affected by these problems, making it difficult, if not impossible to make proper inference with respect to the target population of the survey (Bethlehem, 2010).

Under-coverage means that some units of the target population are excluded from the sample selection mechanism; therefore such units have no chance to be selected in the survey. If data are collected by means of the Internet, only people with Internet can access the questionnaire, while those without Internet are excluded from the survey. Research shows that people who are covered by the Internet technology differ, on the average, from those who are not. As a consequence, web survey results cannot be used to say something about the entire population; web survey results only apply to the sub-population of people having Internet. This is unavoidable, unless a sample of non-Internet units is available.

Self-selection means that individuals are allowed to decide completely for themselves whether or not they want to participate in a survey. In case of web surveys, the questionnaire is put on the web. Respondents are those individuals who visit the website and decide to participate in the survey or, in addition, individuals are invited via e-mail and asked them to complete the questionnaire.

Self-selection may also occur in CAWI (Computer Assisted Web Interviewing) surveys, where sampled units are asked to complete the questionnaire by filling in a form online. As a consequence, people with no internet connection or not familiar with computers or mobile devices cannot be interviewed.

Both under-coverage and self-selection have serious impact on the quality of survey results. The theory of probability sampling cannot be applied and estimates are often biased. Horvitz and Thompson (1952) show that unbiased estimates of population characteristics can be computed only if a real probability sample has been drawn, every element in the population has a non-zero probability of selection, and all these probabilities are known to the researcher. Furthermore, only under these conditions, the accuracy of estimates can be computed.

Many web surveys are not based on probability sampling. The problem is that the survey researcher is not in control of the selection process. Selection probabilities are unknown and, moreover, they are considerably smaller than in traditional probability surveys. Therefore, neither unbiased estimates can be computed nor the accuracy of estimates can be determined (Bethlehem, 2010).

In this work we propose different estimation methods for data from nonprobability samples. The main idea consists of finding a model for the process that is supposed to have caused self-selection. Therefore, on the basis of the specified self-selection model estimate inclusion probabilities. The work is organized as follows. We begin in Chapter 0 with some preliminaries on nonprobability sampling. In Chapter 1 we introduce the theoretical framework and the methodology, including the estimators. In Chapter 2 we investigate the large sample properties of the proposed estimators. Then we present in Chapter 3 various bootstrap approaches to estimate variance and confidence intervals. We conclude in Chapter 4 with a simulation study aimed at evaluating the performance of the proposed estimators. Finally, the last Section contains some final comments and conclusions.

Chapter 0

Preliminaries

0.1 Nonprobability samples

In the last decade, many statistical applications on samples that are not randomly selected from a well-defined finite population have become common. These samples often come from huge data sources, such as customers electronic data, but also administrative data on persons and households, and those for business statistics. Some vendors and survey organizations have also formed large panels of persons who are willing to participate in surveys via the Internet. Many of these databases, despite being large, are not probability samples, but analysts want to project them to full finite populations (Valliant et al., 2018).

Because of declining response rates and ever increasing costs, pressures to find alternatives to expensive probability sampling have been building. A nonprobability sample may do very well on a criterion like timeliness, but evaluating its accuracy may be difficult.

Nonprobability surveys capture participants through various methods. Not all of these are equally dependable for making inferences. According to Baker et al. (2013) these samples can be characterized into three broad categories:

(1) Convenience sampling

(2) Sample matching

(3) Network sampling.

Convenience sampling is a form of nonprobability sampling in which easily locating and recruiting participants is the primary consideration. No formal sample design is used. Some types of convenience samples are shopping mall intercepts, volunteer samples, river samples, observational studies and snowball samples.

In a mall intercept sample, interviewers try to recruit shoppers to take part in some study. Usually, neither the malls nor the people are probability samples. A more modern equivalent to a mall intercept is an online popup survey where visitors to a set of websites are asked to participate in a survey. For example, Google Surveys[1] allow a questionnaire to be constructed and a target audience specified by age group, gender, country, and language. Google then posts the survey across a network of news, reference, and entertainment sites. Even though a target audience can be specified, the set of persons who respond cannot be considered to be a probability sample of that target population.

Volunteer samples are common in social science, medicine and market research. Volunteers may participate in a single study or become part of a panel whose members may be recruited for different studies over the course of time. A recent development is the opt-in web panel in which volunteers are recruited when they visit particular web sites. After becoming part of a panel, the members may participate in many different surveys, often for some type of incentive. River samples are a version of opt-in web sampling in which volunteers are recruited at a number of websites.

In *sample matching*, the members of a nonprobability sample are selected to match a set of important population characteristics. For example, a sample of persons may be constructed so that its distribution by age, race/ethnicity and sex closely matches the distribution of the inference population. Quota sampling is an example of sample matching. The matching is intended to reduce selection biases as long as the covariates that predict survey responses can be used in matching. Rubin (1979) presents the theory for matching in observational studies.

A variation of matching in survey sampling is to match the units in a nonprobability sample with those in a probability sample. Each unit in the nonprobability sample is then assigned the weight of its match in the probability sample. River (2007) describes this type of sample matching in the context of web survey panels. Other techniques developed by Rosenbaum and Rubin (1983) and others for analyzing observational data have also been applied when attempting to develop weights for some volunteer samples.

In *network sampling*, members of some target population are asked to identify other members of the population with whom they are somehow connected. Members

[1] `https://www.google.com/analytics/surveys/`

of the population that are identified in this way are then asked to join the sample. This method of recruitment may proceed for several rounds. Snowball sampling is an example of network sampling in which existing study subjects recruit additional subjects from among their acquaintances. These samples typically do not represent any well-defined target population, although they are a way to potentially accumulate a sizeable collection of units from a rare population. The size of the collection is heavily dependent on locating "seed" (starting points) and their willingness to recruit others from the network.

0.2 Potential problems

According to several authors some different types of problems can arise during a survey process (Baker et al., 2013; Valliant et al., 2018). We mention in particular three major categories:

- Selection bias
- Nonresponse
- Measurement error.

For sake of simplicity we refer to volunteer Internet surveys (also called opt-in surveys).

Selection bias occurs if the observed part of the population (the sample) differs from the unobserved (the nonsample) in such a way that the sample cannot be projected to the full population. Coverage error, for instance, will lead to selection bias. For example, in a volunteer web panel only persons with access to the Internet can join a panel.

To describe three components of coverage survey bias, Valliant and Dever (2011) defined three populations, illustrated in Figure 1: (1) the target population of interest for the study U; (2) the potentially covered population given the way that data are collected, F_{pc}; and (3) the actual covered population, F_c, the portion of the target population that is recruited for the study through the essential survey conditions. The inferential problem is to project the set of sample units s to the universe U, accounting for the facts that part of the population is only potentially covered and part is not covered at all.

In a volunteer web panel, F_{pc} might be the set of all persons who visit websites where recruiting is done, F_c are the people who visit those websites and volunteer for the panel, and s is a sample of persons from the panel selected for a particular survey. The set $U - F_{pc}$ consists of all the people who have Internet access but never visit the sites where recruiting is done plus all people who do not have Internet access at all.

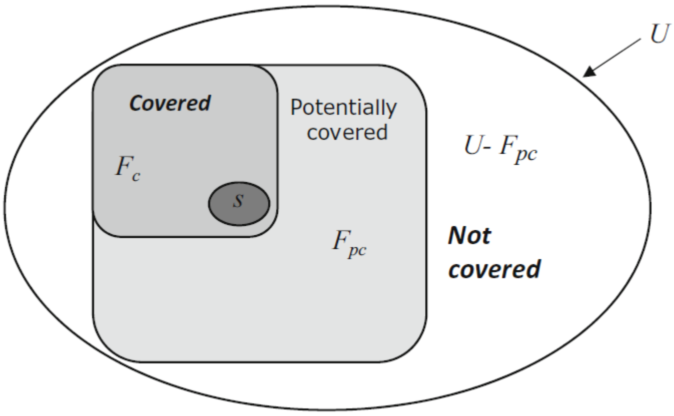

Figure 1: Universe and sample with coverage errors - Source Valliant et al. (2018)

Nonresponse of several kinds affects web surveys. Usually the vendor sends the person an email with a link that must be clicked in order to access the questionnaire. After that the questionnaire need to be filled in to participate the survey. People may also click on a banner ad advertising the survey but never complete the questionnaire.

Measurement error is a common problem in nonprobability samples as it is in probability samples. For a specific item it is often defined as a random error due to the discrepancy between the observed value in the sample and the true value in the population. It occurs when respondent's answer to a question is inaccurate. In traditional surveys interviewers themselves can sometimes be a source of measurement error. For example, if interviewers suggest by their nonverbal (or verbal) behaviour that they want to get the interview over with as quickly as possible. In contrast, in web surveys with self-administered questionnaire respondents themselves may be a potential cause of measurement error.

In general, all surveys may be subject to these problems, but the degree of difficulties, like selection bias, can be worse for nonprobability samples. In order to obtain good quality estimates, these problems have to be corrected.

0.3 Approaches to inference

In finite population sampling more than one approach can be used to make inference about unknown population parameters. It is convenient to distinguish two major approaches: the *design-based approach* and the *model-based approach*.

The principal difference between the two philosophies lies in the element of randomness they utilize in order to give stochastic structure to the inference (Särndal et al., 1978). Classical survey sampling, following in the tradition of Neyman (1934) extremely influential paper, relies on what we call a design base. This means that the primary source of randomness is the probability ascribed by the sampling design to the various subsets of the finite population $1, 2, \ldots, N$ (Särndal et al., 1978).

In the model-based approach the values $y_1, y_2, \ldots, y_N$ associated with the N units of the population are views as the realized outcome of random variables $Y_1, Y_2, \ldots, Y_N$ having an N-dimensional joint distribution ξ, where the *superpopulation* ξ is modeled. In very broad terms, it is a model specified by assumptions about the statistical properties of the study variable values y. In some cases the model can correctly specified to describe the stochastic process that generates the variable values. Generally, it will depend on one or more unknown parameters that are named *superpopulation parameters.*

In the model-based approach the objective of the inference can be twofold:

1. we can either be interested in estimation of the descriptive population parameters, such as the total or the population mean of the study variable. The attention is addressed to the specific model realization $y = (y_1, y_2, \ldots, y_N)$ in the population;

2. or we can be interested in estimation of the density or probability function $f(y; \theta)$ of the random variable Y: in this case the attention is focused on the model assumed to have generated the population, that is the process underlying finite population and the vector of parameters upon which it depends on.

In case 2. it is reasonable to think that the interest is in the process that generates y and in the complex of relationships between the variables Y and the auxiliary variables X, that is the interest is in superpopulation parameters rather than in descriptive population parameters.

In contrast to descriptive population parameters, which could be known exactly in a census not affected by measurement errors and non-responses, superpopulation parameters are hypothetical constructs not directly observable, neither in a census. However, census observations of realizations y will be hardly available. In real applications, observed values of Y are available only for a sample that can also be not random.

To summarize, in a design-based approach the randomness required to make inference comes from the sampling design, and the values, $Y_1, Y_2, \ldots, Y_N$, forming the population are fixed. In a model-based approach, instead, y is considered to be a realization of Y whose joint distribution is specified by the model ξ.

Chapter 1

Methodology proposed

This chapter provides a theoretical framework for estimating population mean in nonprobability samples, such as opt-in sample surveys. After introducing basic notations as well as concepts and exploring some effects of self-selection when the inclusion probabilities are unknown, two estimators of population mean are proposed under the model-based approach.

1.1 Basic setup

Let $\mathcal{U}_N$ be a finite population of N units labeled by integers $1, 2, \ldots, N$. A variable of interest, Y, associated with each unit of the population, is considered. We denote the value of the variable of interest for unit i by y_i, $i = 1, 2, \ldots, N$. The values $y_1, \ldots, y_N$ are not known and the parameter of interest is the population mean:

$$\overline{Y}_N = \frac{1}{N}\sum_{i=1}^{N} y_i.$$

We assume that for each N, y_i, $i = 1, 2, \ldots, N$, are realizations of a *superpopulation* $Y_N = (Y_1, \ldots, Y_N)$, composed by independent and identically distributed (i.i.d.) random variables, $\{Y_1, Y_2, \ldots, Y_N\}$, with common distribution function, F. We also suppose that a vector of covariates denoted by X_i, $i = 1, \ldots, N$ is available. The values $x_1, x_2, \ldots, x_N$ are known for each unit i in the population and can be used in the estimators in order to improves their properties. Such a model enables us to make inferences about population characteristics based on sample measurements and other supplementary information for each unit of the population (auxiliary information). Essentially, a *model-based* approach (Section 0.3) is adopted.

The random variables, $\{Y_1, Y_2, \ldots, Y_N\}$, are assumed to be marginally independent and identically distributed. They are also assumed to be conditionally independent given covariates, that is $Y_i|X_i$ are still independent, but not identically distributed, $i = 1, 2, \ldots, N$. In this way, any possible influence of the auxiliary variables on the variable of interest is accounted for.

Suppose a sample s including n units is observed; they are viewed as a nonprobability sample from a large population.

Let δ_i be the *sample membership indicator*, which indicates whether or not unit i is included in the sample:

$$\begin{cases} \delta_i = 1 & \text{if } i \in s \\ \delta_i = 0 & \text{otherwise.} \end{cases}$$

For each unit in the sample the triple (δ_i, y_i, x_i) is observed, where y_i is the value of the variable of interest. Basically, the probability distribution of (δ, Y, X) refers to the distribution induced by the random sampling from the superpopulation.

Since the variables Y_i, $i = 1, 2, \ldots, N$, may depend on the values of the auxiliary variables, we denote by

$$\begin{aligned} \mu(x) &\equiv \mathbb{E}[Y|X = x] \\ \sigma^2(x) &\equiv \mathrm{Var}[Y|X = x] \end{aligned}$$

the conditional expectation and the conditional variance of the variable of interest with respect to the values of the auxiliary variables, respectively.

Finally, define the *inclusion* (or *selection*) *probability* of unit $i \in \mathcal{U}_N$, given the covariates:

$$\begin{aligned} \pi(x_i) &\equiv Pr(\delta_i = 1|X_i = x_i) \\ &= \mathbb{E}(\delta_i|X_i = x_i), \quad i \in s. \end{aligned}$$

It is essentially the *first-order inclusion probability* of unit i, conditionally on X_i. The first-order inclusion probability $\pi_i \equiv \pi(x_i)$ refers to the probability that unit i is included in the sample, given the values of the covariates.

The inclusion probability $\pi(x_i)$, $i = 1, 2, \ldots, N$, can be interpreted in terms of *propensity score*, a concept first introduced by Rosenbaum and Rubin (1983). They developed a technique to compare two populations, treated units and control units. They attempt to make the two populations comparable by simultaneously controlling for all variables that were thought to explain the differences. From this point of view, the case of self-selected sample essentially parallels causality model in Rosenbaum-Rubin approach, where sample units play the role of "treated units" and the self-selection mechanism is similar to the random assignment of treatment levels to units.

From a formal perspective this context can be associated to a *Poisson design* where the propensity score is equivalent to the first-order inclusion probability of unit i. In symbols:

$$P(s \mid X_1, \ldots, X_n) = \prod_{i=1}^{N} \pi(x_i)^{\delta_i}(1-\pi(x_i))^{1-\delta_i}.$$

This design was introduced by Hájek (1964): it consists of performing N independent Bernoulli trials with probability π_i that unit i is selected in the sample. All the samples have a positive probability of being selected and there is a non-null probability of selecting an empty sample. Since the units are selected independently, the *second-order inclusion probability*, that is the probability that both units i and j are included in the sample, is $\pi_{i,j} = \pi_i \pi_j$, for all $i \neq j$.

Under this sampling design, the variance of the Horvitz and Thompson (1952) estimator of the population mean reduces to

$$\mathrm{Var}(\hat{T}_{HT}) = \frac{1}{N^2} \sum_{i \in U} \frac{1-\pi_i}{\pi_i} y_i^2,$$

which can be unbiasedly estimated by means *shrinkage* techniques without involving joint inclusion probabilities, thus providing a simple formula for variance estimation.

It is worth noting that the Poisson sampling design maximizes the entropy (Hájek, 1981) given by

$$I(p) = -\sum_{s \subset \mathcal{U}} p(s) \log p(s),$$

subject to given inclusion probabilities π_i, $i \in \mathcal{U}$. Since the entropy is a measure of spread of the sampling design $p(\cdot)$, the Poisson sampling design can be viewed as the most random sampling design that satisfies given inclusion probabilities. This means that there is a high amount of uncertainty or randomness in the samples which will be selected, which in turns make the design more robust.

Despite the good properties, Poisson sampling is rarely applied in practice because its sample size $n(s)$ is random implying a nonfixed cost of sampling. This design is, however, often used to model nonresponse.

1.2 The Horvitz-Thompson estimator

Under unequal probability sampling without replacement, the Horvitz-Thompson estimator is an unbiased estimator of the population mean (Horvitz and Thompson, 1952). It is defined as

$$\hat{T}_{HT} = \frac{1}{N} \sum_{i \in s} \frac{y_i}{\pi_i}, \tag{1.1}$$

where $\pi_i = Pr(i \in s)$ is the first order inclusion probability of the ith unit.

The variance of $\hat{T}_{HT}$ is

$$\text{Var}(\hat{T}_{HT}) = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} (\pi_{ij} - \pi_i \pi_j) \frac{y_i y_j}{\pi_i \pi_j},$$

with unbiased estimates

$$\hat{V}(\hat{T}_{HT}) = \frac{1}{N^2} \sum_{i \in s} \sum_{j \in s} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \right) \frac{y_i y_j}{\pi_i \pi_j},$$

where π_{ij} is the joint inclusion probability of the ith and the jth units, with $\pi_{ii} = \pi_i$.

For a sampling design of fixed size, $n(s) = n$, equivalent formulas can be deduced for the variance and variance estimator of $\hat{T}_{HT}$, as obtained by Yates and Grundy (1953) and Sen (1953):

$$\text{Var}_{YG}(\hat{T}_{HT}) = \frac{1}{2N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} (\pi_{ij} - \pi_i \pi_j) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2,$$

$$\hat{V}_{YG}(\hat{T}_{HT}) = \frac{1}{2N^2} \sum_{i \in s} \sum_{j \in s} (\pi_{ij} - \pi_i \pi_j) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2.$$

Note that to calculate the Horvitz-Thompson and Sen-Yates-Grundy variance estimator, knowledge of the second-order inclusion probabilities is required for all possible pairs of the units sampled, that is the probability that any pair of units is included in the sample. These probabilities are usually problematic to calculate for complex sampling designs, such as unequal probability sampling.

To address this issue, Hájek (1964) explored the properties of joint inclusion probabilities and derived a formula based on rejective sampling, a sampling procedure in which a Poisson sample is rejected unless it contains exactly n sample units as required by the sample design (Hájek, 1981). Rejective sampling is also called *conditional Poisson sampling.*

The quality of the Horvitz-Thompson estimator, $\hat{T}_{HT}$, does not depend on any modeling. Information can be incorporated in this estimator only by the first-and second-order sample inclusion probabilities in the design phase of the survey, in which the sampling method is determined. Hence, $\hat{T}_{HT}$ is a pure design-based estimator, meaning that its accuracy depends solely on the applied sampling method, the inclusion probabilities assigned by this method, and the sample size (Quatember, 2015).

1.3 The Hájek estimator

Assume that a sample is taken according to a randomization scheme having unknown inclusion probabilities $\pi_i = Pr(i \in s)$ and a predetermined sample size n. Then assume that values x_i of a positive auxiliary variable are available for all units in the population, $i = 1, 2, \ldots, N$, which can be assumed approximately proportional to the variable of interest Y:

$$\frac{y_i}{x_i} \approx constant, \quad i = 1, 2, \ldots, N. \tag{1.2}$$

If (1.2) holds it seems reasonable to calculate the first-order inclusion probabilities as

$$\pi_i = \frac{nx_i}{\sum_{j=1}^{N} x_j} = \frac{nx_i}{N\mu_x}, \tag{1.3}$$

where μ_x is the population mean of X.

When the first-order inclusion probability is defined according to the criteria (1.3) the sampling design is said to be πpps (*inclusion probabilities proportional to size*).

Under this scheme, a well known and popular estimator attributed to Hájek (1971) is defined by

$$\hat{T}_H = \frac{\displaystyle\sum_{i \in s} \frac{1}{\hat{\pi}(x_i)} Y_i}{\displaystyle\sum_{i \in s} \frac{1}{\hat{\pi}(x_i)}}.$$

He suggested this estimator in response to an observation by Basu (1971) on paradoxical behaviour of the πpps unbiased Horvitz and Thompson (1952) estimator.

Särndal et al. (1992) give several cases for regarding the Hájek as "usually the better estimator" comparing to the Horvitz-Thompson estimator (1.1) when:

(a) the y_i are relatively homogeneous (the difference $y_i - \mu_y$ tend to be small);

(b) sample size is not fixed;

(c) π_i are weakly or negatively correlated with the y_i.

By using Taylor expansion (Section 2.2) it is possible to show that

$$\hat{T}_H = \overline{Y} + \sum_{i \in s} \frac{1}{\pi_i}(y_i - \overline{Y}) + O_p(n^{-1})$$

.

Hence, Hájek variance estimator can be approximated by

$$\hat{V}_H = \text{Var}\left[\sum_{i \in s} \frac{1}{\pi_i}(y_i - \overline{Y})\right] + O_p(\frac{1}{n^2}).$$

1.4 Effect of self-selection

In this section we show that the sample mean is not an unbiased estimator of the population mean when inclusion probabilities are unknown.

Consider the sample mean:

$$\overline{y}_s = \frac{1}{n} \sum_{i \in s} y_i = \frac{\sum_{i=1}^N y_i \delta_i}{\sum_{i=1}^N \delta_i},$$

where the sample size $n = \sum_{i=1}^N \delta_i$ is a random variable.

By using a first Taylor expansion (Section 2.2) of the sample mean and taking into account that

$$\mathbb{E}[n \mid X_1, X_2, \dots, X_N] = \sum_{i=1}^N \mathbb{E}[\delta_i \mid X_i] = \sum_{i=1}^N \pi(x_i),$$

we may write

$$\frac{1}{n} = \frac{1}{\sum_{i=1}^N \delta_i} \simeq \frac{1}{\sum_{i=1}^N \pi(x_i)} - \frac{1}{\left[\sum_{i=1}^N \pi(x_i)\right]^2} \left[\sum_{i=1}^N \delta_i - \sum_{i=1}^N \pi(x_i)\right],$$

where the symbol $\simeq$ means "approximately equal to".

From the above inequality we get

$$\frac{N}{n} \simeq \frac{N}{\sum_{i=1}^N \pi(x_i)} - \left[\frac{N}{\sum_{i=1}^N \pi(x_i)}\right]^2 \left\{\frac{1}{N} \sum_{i=1}^N [\delta_i - \pi(x_i)]\right\}.$$

From the Weak Law of large Numbers we have

$$\frac{1}{N} \sum_{i=1}^N \pi(x_i) \xrightarrow{P} \mathbb{E}[\pi(x)] = \overline{\pi}, \quad \text{as } N \to \infty;$$

$$\frac{1}{N} \sum_{i=1}^N [\delta_i - \pi(x_i)] \xrightarrow{P} \mathbb{E}[\delta_i - \pi(x_i)] = 0, \quad \text{as } N \to \infty,$$

where

$$\begin{aligned}\mathbb{E}[\delta_i - \pi(x_i)] &= \mathbb{E}[\delta_i] - \mathbb{E}[\pi(x_i)] \\ &= \mathbb{E}[\mathbb{E}(\delta_i \mid X_i)] - \overline{\pi} \\ &= \mathbb{E}[\pi(x_i)] - \overline{\pi} \\ &= \overline{\pi} - \overline{\pi} \\ &= 0.\end{aligned}$$

As a consequence, a crude first-order approximation gives the following result:

$$\mathbb{E}[\overline{y}_s] \simeq \frac{\mathbb{E}[\sum_{i=1}^{N} y_i\delta_i]}{\mathbb{E}\big[\sum_{i=1}^{N} \delta_i\big]}.$$

Since

$$\begin{aligned}\mathbb{E}\Big[\sum_{i=1}^{N} \delta_i\Big] &= \sum_{i=1}^{N} \mathbb{E}[\delta_i] \\ &= \sum_{i=1}^{N} \overline{\pi} \\ &= N\overline{\pi} \\ &= N\mathbb{E}[\delta_i]\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^{N} y_i\delta_i\right] &= \sum_{i=1}^{N} \mathbb{E}[y_i\delta_i] \\ &= N\mathbb{E}[y_i\delta_i],\end{aligned}$$

we finally obtain

$$\mathbb{E}[\overline{y}_s] \simeq \frac{N\mathbb{E}[y_i\delta_i]}{N\mathbb{E}[\delta_i]} = \frac{\mathbb{E}[y_i\delta_i]}{\mathbb{E}[\delta_i]} \neq \mathbb{E}[y_i].$$

This show that the expected value of the sample mean is not equal to the population mean. The only situation in which $\overline{y}_s$ is approximately unbiased is that in which y_i and δ_i are independent.

1.5 Aim of the study

Given a sample s including $n(s)$ units, that are selected according to the sampling scheme described in Section 1.1, the estimation process consists of three different steps.

STEP 1:

On the basis of the values of δ_i and x_i finding an estimate of $\pi(x_i)$, $i \in s$.

We adopt two different approaches in order to achieve this aim:

- sieve estimator (Hirano et al., 2003);
- logit model estimator.

We describe these methods in more detail in the next section.

STEP 2:

Construct an estimator for the population mean, $\overline{Y}_N$.

For this purpose, we define the **pseudo Horvitz-Thompson estimator** as follows:

$$\begin{aligned} \hat{T}_{pHT} &= \frac{1}{N} \sum_{i \in \underline{s}} \frac{1}{\hat{\pi}(x_i)} Y_i \\ &= \frac{1}{N} \sum_{i=1}^{N} \frac{\delta_i}{\hat{\pi}(x_i)} Y_i. \end{aligned} \tag{1.4}$$

Similarly, we define the **pseudo Hájek estimator**:

$$\begin{aligned} \hat{T}_{pH} &= \frac{\displaystyle\sum_{i \in \underline{s}} \frac{1}{\hat{\pi}(x_i)} Y_i}{\displaystyle\sum_{i \in \underline{s}} \frac{1}{\hat{\pi}(x_i)}} \\ &= \frac{\displaystyle\frac{1}{N} \sum_{i=1}^{N} \frac{\delta_i}{\hat{\pi}(x_i)} Y_i}{\displaystyle\frac{1}{N} \sum_{i=1}^{N} \frac{\delta_i}{\hat{\pi}(x_i)}}, \end{aligned} \tag{1.5}$$

which is especially useful when the population size N is unknown. When N is unkown we have to remark that the denominator $\sum_{i=1}^{N} \frac{\delta_i}{\hat{\pi}(x_i)}$ can be viewed as the Horvitz-Thompson estimator for N, which is consistent. The effect of this result will be clearer in the next chapter.

STEP 3:

Study the behaviour of the estimators chosen in step 2 by assessing their asymptotic properties. Our aim is to obtain consistent estimator for the population mean.

1.6 Assumptions

The basic assumptions on which our work is based are listed below.

Assumption 1 (Unconfoundedness):

$$Y \perp \delta | X \tag{1.6}$$

This assumption was first introduced by Rosenbaum and Rubin (1983), who refer to it as "ignorable treatment assignment". In our context it seems to be logical to refer to it as the conditional indipendence assumption, that is variables Y and δ are independent conditionally on X.

As a consequence, if the inclusion indicator variable and the variable of interest are independent conditionally on all covariates, they are also independent conditionally on the (conditional) probability of being included given covariates (i.e. propensity score). Formally, as shown by Rosenbaum and Rubin (1983), this assumption implies

$$Y \perp \delta | \pi(X). \tag{1.7}$$

Assumption 2 (Overlap):

$$\epsilon < Pr(\delta_i = 1 | X_i) < 1 - \epsilon, \quad \text{for some positive } \epsilon. \tag{1.8}$$

Given assumption 1, the following equalities hold:

$$\begin{aligned} \mu(x) &= \mathbb{E}[Y | X = x] \\ &= \mathbb{E}[Y | \delta, X = x], \end{aligned}$$

and thus $\mu(x)$ is identified. To make this feasible, one needs to be able to estimate the expectations $\mathbb{E}[Y|\delta, X = x]$ for all values of δ and x in the support of these variables. This is where the second assumption enters.

In addition to the uncounfoundedness assumption, the following assumptions are used to derive the properties of the estimator. First, we restrict the distribution of X and Y.

Assumption 3 (Distribution of X):

(i) the support $\mathcal{X}$ of the r-dimensional covariate X is a Cartesian product of compact intervals, $\mathcal{X} = \prod_{j=1}^{r} [x_{lj}, x_{uj}]$;

(ii) the density of X is bounded, and bounded away from 0, on $\mathcal{X}$.

Assumption 4 (Distribution of Y):

(i) $\mathbb{E}[Y^2] < \infty$;

(ii) $\mu(x)$ is continuously differentiable for all $x \in \mathcal{X}$.

The next assumption requires sufficient smoothness of the propensity score.

Assumption 5 (Selection Probability): The propensity score $\pi(x)$ satisfies the following conditions. For all $x \in \mathcal{X}$:

(i) $\pi(x)$ is continuously differentiable of order $s \geq 7 \cdot r$ where r is the dimension of $\mathcal{X}$;

(ii) $\pi(x)$ is bounded away from zero and one: $0 < \pi(x) < 1$.

Finally, we restrict the rate at which additional terms are added to the series approximation to $\pi(x)$, depending on the dimension of X and the number of derivatives of $\pi(x)$.

Assumption 6 (Series Estimator): The series logit estimator of $\pi(x)$ uses a power series with $L = N^v$ for some $1/(4(s/r - 1)) < v < \frac{1}{9}$.

The restriction on the derivatives (Assumption 5(i)) guarantees the existence of a v that satisfies the conditions in Assumption 6.

1.7 Propensity score methods

1.7.1 Hirano-Imbens-Ridder estimator

This section is devoted to introduce the main features of the estimator suggested by Hirano et al. (2003) in the context of estimation of propensity score for average treatment effects.

Estimating the average effect of a binary treatment or policy on a scalar outcome is a basic goal of many empirical studies in economics. If assignment to the treatment is exogenous or *unconfounded* (i.e., independent of potential outcomes conditional on covariates or pre-treatment variables, an assumption also known as selection observables), the average treatment effect can be estimated by matching (Abadie and Imbens, 2002) or by averaging within-subpopulation differences of treatment and control averages. If there are many covariates, such strategies may not be desirable or even feasible. An alternative approach is based on the *propensity score*, the conditional probability of receiving treatment given covariates.

Rosenbaum and Rubin (1983) show that, under the assumption of unconfoundedness, adjusting solely for differences in the propensity score between treated and control units removes all biases. Although adjusting for differences in the propensity score removes all bias, it need not be as efficient as adjusting for differences in all covariates, as shown by Hahn (1998). However, Rosenbaum (1987), Rubin and Thomas (1992), and Robins et al. (1995) show that using parametric estimates of the propensity score, rather than the true propensity score, can avoid some of these efficiency losses.

Hirano et al. (2003) propose estimators that are based on adjusting for nonparametric estimates of the propensity score, leading to an efficient estimate of the average treatment effect. The proposed estimators weight observations by the inverse of nonparametric estimates of the propensity score, rather than the true propensity score. They also show that for the case in which the propensity score is known, the proposed estimators can be interpreted as empirical likelihood estimators that efficiently incorporate the information about the propensity score.

The authors estimate the propensity score in a sieve approach (e.g., Geman and Hwang, 1982) by the Series Logit Estimator. More precisely, they first specify a sequence of functions of the covariates, such as power series $h_l(x)$, $l = 1, \ldots, \infty$. Next, they choose a number of terms, $L(N)$, as a function of the sample size, and then estimate the L-dimensional vector γ_L in

$$Pr(\delta = 1 \mid X = x) = \frac{exp[(h_1(x), \ldots, h_L(x))\gamma_L]}{1 + exp[((h_1(x), \ldots, h_L(x))\gamma_L]},$$

by maximizing the associated likelihood function. Let $\hat{\gamma_L}$ be the maximum likelihood estimate. In the third step, the estimated propensity score is calculated as

$$Pr(\delta = 1 \mid X = x) = \frac{exp[(h_1(x), \ldots, h_L(x))\hat{\gamma_L}]}{1 + exp[((h_1(x), \ldots, h_L(x))\hat{\gamma_L}]}.$$

Under the Assumptions 1-6 (Section 1.6), where the role of δ is played here by the treatment, the authors show that with a nonparametric estimator for $\pi(x)$ the estimator of the average treatment effect is efficient, whereas with the true propensity score the estimator would not be fully efficient.

To provide some intuition for these results the authors consider the simpler problem of estimating the population average of a variable Y, $\mu_0 = \mathbb{E}[Y]$, given a random sample of size N of the triple $(\delta_i, X_i, \delta_i \cdot Y_i)$. In other words, δ_i and X_i are observed for all units in the sample, but Y_i is only observed if $\delta_i = 1$.

The analog to the unconfoundedness assumption here is the assumption that the Y_i are Missing At Random (MAR; Rubin (1976)), or

$$\delta \perp Y|X.$$

The role of the propensity score is played here by the selection probability $\pi(x) = \mathbb{E}[\delta|X = x] = Pr(\delta = 1|X = x)$. First, the attention is restricted to the case with a single binary covariate. Let N_{tx} denote the number of observations with $\delta_i = t$ and $X_i = x$, for $t, x \in \{0, 1\}$. Furthermore, suppose the true selection probability is constant, $\pi(x) = 1/2$ for all $x \in \{0, 1\}$. The normalized variance bound for μ_0 is

$$V_{bound} = 2 \cdot \mathbb{E}[V(Y|X)] + V[\mathbb{E}(Y|X)]. \tag{1.9}$$

The first estimator, named the "true weights" estimator, weights the complete observations by the inverse of the true selection probability:

$$\hat{\mu}_{tw} = \frac{1}{N}\sum_{i=1}^{N}\frac{Y_i\delta_i}{\pi(X_i)} = \frac{1}{N}\sum_{i=1}^{N}\frac{Y_i\delta_i}{1/2}.$$

Its large sample normalized variance is

$$\begin{aligned} V_{tw} &= 2 \cdot \mathbb{E}[V(Y|X)] + V[\mathbb{E}(Y|X)] + \mathbb{E}[\mathbb{E}(Y|X)^2] \\ &= V_{bound} + \mathbb{E}[\mathbb{E}(Y|X)^2] \end{aligned}$$

strictly larger than the variance bound (1.9) unless $\mathbb{E}(Y|X) = 0$.

The second estimator weights the complete observations by the inverse of a non-parametric estimate of the selection probability. This estimator is the main focus of the paper by Hirano et al. (2003). In the current setting the estimated selection probability is simply the proportion of observed outcomes for a given value of the covariate. For units with $X_i = 0$ the proportion of observed outcomes is $N_{10}/(N_{00} + N_{10})$, and for units with $X_i = 1$ the proportion of observed outcomes is $N_{11}/(N_{01} + N_{11})$. Thus the estimated selection probability is

$$\hat{\pi}(x) = \begin{cases} N_{10}/(N_{00} + N_{10}) & \text{if } x = 0, \\ N_{11}/(N_{01} + N_{11}) & \text{if } x = 1. \end{cases}$$

The proposed "estimated weights" estimator is then

$$\hat{\mu}_{ew} = \frac{1}{N}\sum_{i=1}^{N}\frac{Y_i\delta_i}{\hat{\pi}(X_i)}.$$

The normalized variance of this estimator is equal to the variance bound:

$$V_{ew} = 2 \cdot \mathbb{E}[V(Y|X)] + V[\mathbb{E}(Y|X)] = V_{bound}.$$

Not only does the weighting estimator with nonparametrically estimated weights have a lower variance than the estimator using the “true” weights in this simple case, but it is in fact fully efficient. This will suggest why this efficiency property may carry over to the case with continuous and vector-valued covariates, as well as with general dependence of the selection probability or propensity score on the covariates.

However, these estimators are relevant whether the propensity score is known or not. In randomized experiments, for example, the propensity score is known by design. In that case the proposed estimators can be used to improve efficiency over simply differencing treatment and control averages. With the propensity score known, an attractive choice for the nonparametric series estimator for the propensity score is to use the true propensity score as the leading term in the series.

The estimators proposed by Hirano et al. (2003) require fewer functions to be estimated nonparametrically than other efficient estimators previously proposed in the literature, such as *regression* estimators. One difficulty with these estimators that are based on the estimated propensity score is the problem of choosing the smoothing parameter. Hirano et al. (2003) use series estimators, which requires choosing the number of terms in the series; for regression method it is the bandwidth of the kernel chosen.

1.7.2 Logit model estimator

When the propensity score must be estimated, typically, researchers assume a parametric propensity score model $\pi_\beta(X_i)$,

$$Pr(\delta_i = 1 \mid X_i) = \pi_\beta(X_i),$$

where $\beta \in \Theta$ is an L-dimensional column vector of unknown parameters. For example, a popular choice is the logistic model:

$$\pi_\beta(X_i) = \frac{exp(X_i^T\beta)}{1 + exp(X_i^T\beta)}$$

in which case we have $L = K$. Then the empirical fit of the model is maximized so that the estimated propensity score predicts the selection probability of unit i given covariates as well. This can be done by maximizing the log-likelihood function:

$$\hat{\beta}_{MLE} = \arg\max_{\beta\in\Theta} \sum_{i=1}^{N} \delta_i \log\{\pi_\beta(X_i)\}.$$

Assuming that $\pi_\beta(\cdot)$ is twice continuously differentiable with respect to β, this implies the first-order condition:

$$\frac{1}{N}\sum_{i=1}^{N} S_\beta(\delta_i, X_i) = 0, \qquad S_\beta(\delta_i, X_i) = \frac{\delta_i \pi'_\beta(X_i)}{\pi_\beta(X_i)} \tag{1.10}$$

and $\pi'_\beta(X_i) = \partial\pi(X_i)/\partial\beta^T$.

As several authors noticed, the major difficulty of this standard approach is that the propensity score model may be misspecified, yielding biased estimates of target parameter (e.g., Kang and Schafer, 2007).

Chapter 2

Estimators of the population mean and their large sample properties

In this chapter we aim at deriving the large sample properties for both pseudo Horvitz-Thompson estimator and pseudo Hájek estimator. At first we assume that the true value of propensity score is known. Then the propensity score is assumed to be estimated according to Hirano-Imbens-Ridder method. It is worth noting that similar properties to Hirano-Imbens-Ridder method can be expected for parametric propensity score model using logistic regression model, provided that the model is correctly specified.

2.1 Pseudo Horvitz-Thompson estimator

2.1.1 Properties when the propensity score is known

Theorem 1. *The pseudo Horvitz-Thompson estimator, $\hat{T}_{pHT}$, is an unbiased estimator of the expectation of the population mean, $\mathbb{E}[\overline{Y}_N]$, when the propensity score is known.*

Proof. We have to prove that

$$\begin{aligned}\mathbb{E}\left[\hat{T}_{pHT}\right] &= \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\frac{\delta_i}{\pi^*(x_i)}Y_i\right] \\ &= \mathbb{E}[\overline{Y}_N]\end{aligned}$$

where $\pi^*(x_i)$ is the "true" propensity score and $\mathbb{E}$ denotes the expected value under the superpopulation model ξ as specified in Chapter 1.

Given the Assumptions (Section 1.6) the following chain of equalities holds:

$$\begin{aligned}
\mathbb{E}\big[\hat{T}_{pHT}\big] &= \mathbb{E}\bigg[\frac{1}{N}\sum_{i=1}^{N}\frac{\delta_i}{\pi^*(x_i)}Y_i\bigg] \\
&= \frac{1}{N}\cdot\mathbb{E}\bigg[\mathbb{E}\bigg(\sum_{i=1}^{N}\frac{\delta_i}{\pi^*(x_i)}Y_i\bigg|X_i\bigg)\bigg] \\
&= \frac{1}{N}\cdot\mathbb{E}\bigg[\sum_{i=1}^{N}\mathbb{E}\bigg(\frac{\delta_i}{\pi^*(x_i)}Y_i\bigg|X_i\bigg)\bigg] \\
&= \frac{1}{N}\cdot\mathbb{E}\bigg[\sum_{i=1}^{N}\frac{\mathbb{E}(\delta_i|X_i)\mathbb{E}(Y_i|X_i)}{\pi^*(x_i)}\bigg] \\
&= \frac{1}{N}\cdot\mathbb{E}\bigg[\sum_{i=1}^{N}\frac{\pi^*(x_i)\mu(x_i)}{\pi^*(x_i)}\bigg] \\
&= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\big[\mu(x_i)\big] \\
&= \frac{1}{N}N\mu_y.
\end{aligned}$$

Hence

$$\begin{aligned}
\mathbb{E}\big[\hat{T}_{pHT}\big] &= \mu_y \\
&= \mathbb{E}[\overline{Y}_N]
\end{aligned}$$

which means that the pseudo Horvitz-Thompson estimator is an unbiased estimator of the expectation of the population mean when the true value of the propensity score is known. □

Theorem 2. *The variance of the pseudo Horvitz-Thompson estimator, $\hat{T}_{pHT}$, when $\pi(x_i)$ is known, is given by the sum of two components as follows*

$$V(\hat{T}_{pHT}) = V_1 + V_2$$

where

$$\begin{aligned}
V_1 &= \frac{1}{N^2}\mathrm{Var}\bigg(\sum_{i=1}^{N}\frac{\delta_i}{\pi^*(x_i)}Y_i\bigg) \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\frac{\mathbb{E}\big(\sigma^2(x_i)\big)}{\pi^*(x_i)}
\end{aligned}$$

$$\begin{aligned}
V_2 &= \mathrm{Cov}\bigg[\sum_{i=1}^{N}\sum_{j\neq i}^{N}\bigg(\frac{\delta_i}{\pi^*(x_i)}Y_i\bigg)\bigg(\frac{\delta_j}{\pi^*(x_j)}Y_j\bigg)\bigg] \\
&= 0
\end{aligned}$$

and $\pi^*(x_i)$ is the true propensity score.

Proof.

$$
\begin{aligned}
\mathrm{V}_1 &= \frac{1}{N^2}\mathrm{Var}\left(\sum_{i=1}^{N}\frac{\delta_i}{\pi^*(x_i)}Y_i\right) \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\left\{\mathbb{E}\left(\frac{\delta_i Y_i}{\pi^*(x_i)}\right)^2 - \left[\mathbb{E}\left(\frac{\delta_i Y_i}{\pi^*(x_i)}\right)\right]^2\right\} \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\left\{\mathbb{E}\left[\mathbb{E}\left(\frac{\delta_i^2 Y_i^2}{(\pi^*(x_i))^2}\middle| X_i\right)\right] - \left[\mathbb{E}\left(\mathbb{E}\left(\frac{\delta_i Y_i}{\pi^*(x_i)}\middle| X_i\right)\right)\right]^2\right\} \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\left\{\mathbb{E}\left[\frac{\mathbb{E}(\delta_i^2|X_i)\mathbb{E}(Y_i^2|X_i)}{(\pi^*(x_i))^2}\right] - \left[\mathbb{E}\left(\frac{\mathbb{E}(\delta_i|X_i)\mathbb{E}(Y_i|X_i)}{\pi^*(x_i)}\right)\right]^2\right\} \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\left\{\mathbb{E}\left(\frac{\pi^*(x_i)(\sigma^2(x_i)+\mu_y^2)}{(\pi^*(x_i))^2}\right) - \left[\frac{\pi^*(x_i)\mathbb{E}\big(\mu(x_i)\big)}{\pi^*(x_i)}\right]^2\right\} \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\left\{\frac{\mathbb{E}\big(\sigma^2(x_i)\big)}{\pi^*(x_i)} + \mu_y^2 - \mu_y^2\right\} \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\frac{\mathbb{E}\big(\sigma^2(x_i)\big)}{\pi^*(x_i)}
\end{aligned}
$$

As far as V_2 is concerned, we could observe that δ_i and Y_i are independent and identically distributed conditionally on X_i and therefore the covariance between them is zero. However, a proof of this result is provided.

$$\mathrm{V}_2 = \mathrm{Cov}\Big[\sum_{i=1}^{N}\sum_{j\neq i}^{N}\Big(\frac{\delta_i}{\pi^*(x_i)}Y_i\Big)\Big(\frac{\delta_j}{\pi^*(x_j)}Y_j\Big)\Big]$$

$$= \mathbb{E}\Big[\sum_{i=1}^{N}\sum_{j\neq i}^{N}\Big(\frac{\delta_i}{\pi^*(x_i)}Y_i\Big)\Big(\frac{\delta_j}{\pi^*(x_j)}Y_j\Big)\Big] - \mathbb{E}\Big[\sum_{i=1}^{N}\Big(\frac{\delta_i}{\pi^*(x_i)}Y_i\Big)\Big]\mathbb{E}\Big[\sum_{j\neq i}^{N}\Big(\frac{\delta_j}{\pi^*(x_j)}Y_j\Big)\Big]$$

$$= \mathbb{E}\Big\{\mathbb{E}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\Big[\Big(\frac{\delta_i}{\pi^*(x_i)}Y_i\Big)\Big(\frac{\delta_j}{\pi^*(x_j)}Y_j\Big)\Big|X_i, X_j\Big]\Big\}$$

$$- \mathbb{E}\Big[\mathbb{E}\sum_{i=1}^{N}\Big(\frac{\delta_i}{\pi^*(x_i)}Y_i\Big|X_i\Big)\Big]\mathbb{E}\Big[\mathbb{E}\sum_{j\neq i}^{N}\Big(\frac{\delta_j}{\pi^*(x_j)}Y_j\Big|X_j\Big)\Big]$$

$$= \mathbb{E}\Big[\mathbb{E}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\Big(\frac{\delta_i}{\pi^*(x_i)}Y_i\Big|X_i\Big)\Big(\frac{\delta_j}{\pi^*(x_j)}Y_j\Big|X_j\Big)\Big]$$

$$- \mathbb{E}\Big[\sum_{i=1}^{N}\mathbb{E}\Big(\frac{\delta_i}{\pi^*(x_i)}Y_i\Big|X_i\Big)\Big]\mathbb{E}\Big[\sum_{j\neq i}^{N}\mathbb{E}\Big(\frac{\delta_j}{\pi^*(x_j)}Y_j\Big|X_j\Big)\Big]$$

$$= \mathbb{E}\Big[\sum_{i=1}^{N}\mathbb{E}\Big(\frac{\delta_i}{\pi^*(x_i)}Y_i\Big|X_i\Big)\Big]\mathbb{E}\Big[\sum_{j\neq i}^{N}\mathbb{E}\Big(\frac{\delta_j}{\pi^*(x_j)}Y_j\Big|X_j\Big)\Big]$$

$$- \mathbb{E}\Big[\sum_{i=1}^{N}\mathbb{E}\Big(\frac{\delta_i}{\pi^*(x_i)}Y_i\Big|X_i\Big)\Big]\mathbb{E}\Big[\sum_{j\neq i}^{N}\mathbb{E}\Big(\frac{\delta_j}{\pi^*(x_j)}Y_j\Big|X_j\Big)\Big] = 0$$

□

2.1.2 Estimating the propensity score

When the propensity score is unknown the pseudo Horvitz-Thompson estimator can be represented as asymptotically linear (Hirano et al., 2003):

$$\hat{T}_{pHT} = \mu_y + \frac{1}{N}\sum_{i=1}^{N}\Big\{\psi\big(Y_i, \delta_i, X_i, \mu_y, \pi^*(x_i)\big) + \alpha\big(\delta_i, X_i\big)\Big\} + o_p(1/\sqrt{N})$$

where

$$\psi\big(Y_i, \delta_i, X_i, \mu_y, \pi^*(x_i)\big) = \frac{\delta_i}{\pi^*(x_i)}Y_i - \mu_y$$

and

$$\alpha\big(\delta_i, X_i\big) = -\frac{\mathbb{E}(Y_i|X_i)}{\pi^*(x_i)}\Big(\delta_i - \pi^*(x_i)\Big),$$

being $\pi^*(x_i)$ the true propensity score.

By computing the expectation for $\psi(\cdot)$ and $\alpha(\cdot)$ we have:

$$\begin{aligned}
\mathbb{E}\big[\psi\big(Y_i,\delta_i,X_i,\mu_y,\pi^*(x_i)\big)\big] &= \mathbb{E}\left[\frac{\delta_i}{\pi^*(x_i)}Y_i\right]-\mu_y \\
&= \mathbb{E}\left[\mathbb{E}\left(\frac{\delta_i}{\pi^*(x_i)}Y_i\middle|X_i\right)\right]-\mu_y \\
&= \mathbb{E}\left[\mathbb{E}(\delta_i|X_i)\cdot\mathbb{E}\left(\frac{Y_i}{\pi^*(x_i)}\middle|X_i\right)\right]-\mu_y \\
&= \mathbb{E}\left[\mathbb{E}(\delta_i|X_i)\cdot\frac{\mu(x_i)}{\pi^*(x_i)}\right]-\mu_y \\
&= \mathbb{E}\left[\pi^*(x_i)\cdot\frac{\mu(x_i)}{\pi^*(x_i)}\right]-\mu_y \\
&= \mathbb{E}[\mu(x_i)]-\mu_y=\mu_y-\mu_y=0,
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}\big[\alpha\big(\delta_i,X_i\big)\big] &= -\mathbb{E}\left[\frac{\mathbb{E}(Y_i|X_i)}{\pi^*(x_i)}\Big(\delta_i-\pi^*(x_i)\Big)\right] \\
&= -\mathbb{E}\left[\delta_i\cdot\frac{\mu(x_i)}{\pi^*(x_i)}\right]+\mathbb{E}[\mu(x_i)] \\
&= -\mathbb{E}\left[\mathbb{E}\left(\delta_i\cdot\frac{\mu(x_i)}{\pi^*(x_i)}\middle|X_i\right)\right]+\mu_y \\
&= -\mathbb{E}\left[\frac{\mu(x_i)}{\pi^*(x_i)}\cdot\mathbb{E}(\delta_i|X_i)\right]+\mu_y \\
&= -\mathbb{E}\left[\frac{\mu(x_i)}{\pi^*(x_i)}\cdot\pi^*(x_i)\right]+\mu_y \\
&= -\mathbb{E}[\mu(x_i)]+\mu_y=-\mu_y+\mu_y=0,
\end{aligned}$$

where it is easy to understand the role of the assumptions (Section 1.6).

Hence

$$\mathbb{E}(\hat{T}_{pHT})=\mu_y+o_p(1/\sqrt{N}),$$

which means that the pseudo Horvitz-Thompson estimator is asymptotically unbiased when the propensity score is estimated according to Hirano-Imbens-Ridder method.

The asymptotically linear representation of $\hat{T}_{pHT}$ implies that its asymptotic variance equals

$$\frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\big[\big(\psi\big(Y_i,\delta_i,X_i,\mu_y,\pi^*(x_i)\big)+\alpha\big(\delta_i,X_i\big)\big)^2\big]+o_p(1/N).$$

The three components of this variance are reported below:

$$
\begin{aligned}
\mathrm{V}_1 &= \mathbb{E}\big[\psi\big(Y_i,\delta_i,X_i,\mu_y,\pi^*(x_i)\big)^2\big] = \mathbb{E}\left[\left(\frac{\delta_i}{\pi^*(x_i)}Y_i-\mu_y\right)^2\right] \\
&= \mathbb{E}\left[\frac{\sigma^2(x_i)}{\pi^*(x_i)}+\frac{\mu^2(x_i)}{\pi^*(x_i)}\right]-\mu_y^2 \\
\mathrm{V}_2 &= \mathbb{E}\big[\alpha\big(\delta_i,X_i\big)^2\big] = \mathbb{E}\left[\left(\frac{\mathbb{E}(Y_i|X_i)}{\pi^*(x_i)}\cdot\Big(\delta_i-\pi^*(x_i)\Big)\right)^2\right] = \mathbb{E}\left[\frac{\mu^2(x_i)}{\pi^*(x_i)}\right]-\mathbb{E}\big[\mu(x_i)\big]^2 \\
\mathrm{V}_3 &= -2\mathbb{E}\big[\psi\big(Y_i,\delta_i,X_i,\mu_y,\pi^*(x_i)\big)\cdot\alpha\big(\delta_i,X_i\big)\big] \\
&= -2\mathbb{E}\left[\left(\frac{\delta_i}{\pi^*(x_i)}Y_i-\mu_y\right)\cdot\frac{\mathbb{E}(Y_i|X_i)}{\pi^*(x_i)}\Big(\delta_i-\pi^*(x_i)\Big)\right] \\
&= -2\mathbb{E}\left[\frac{\mu^2(x_i)}{\pi^*(x_i)}\right]+2\mathbb{E}\big[\mu^2(x_i)\big],
\end{aligned}
$$

so that

$$
\begin{aligned}
\mathrm{Var}(\hat{T}_{pHT}) &= \frac{1}{N^2}\sum_{i=1}^{N}\big(\mathrm{V}_1+\mathrm{V}_2+\mathrm{V}_3\big) \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\Big[\Big(\psi\big(Y_i,\delta_i,X_i,\mu_y,\pi^*(x_i)\big)+\alpha\big(\delta_i,X_i\big)\Big)^2\Big] \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\left\{\mathbb{E}\left[\frac{\sigma^2(x_i)}{\pi^*(x_i)}\right]+\mathbb{E}\big[\mu(x_i)\big]^2-(\mu_y)^2\right\}+o_p(1/N)
\end{aligned}
$$

Proof. :

$$\begin{aligned}
\mathrm{V}_1 = \mathbb{E}\big[\psi\big(Y_i, \delta_i, X_i, \mu_y, \pi^*(x_i)\big)^2\big] &= \mathbb{E}\left[\left(\frac{\delta_i}{\pi^*(x_i)}Y_i - \mu_y\right)^2\right] \\
&= \mathbb{E}\left[\left(\frac{\delta_i}{\pi^*(x_i)}Y_i\right)^2 + \mu_y^2 - 2\mu_y \frac{\delta_i}{\pi^*(x_i)}Y_i\right] \\
&= \mu_y^2 + \mathbb{E}\left[\mathbb{E}\left(\frac{\delta_i}{\pi^*(x_i)}Y_i\right)^2 \middle| X_i\right] - 2\mu_y \mathbb{E}\left[\mathbb{E}\left(\frac{\delta_i}{\pi^*(x_i)}Y_i \middle| X_i\right)\right] \\
&= \mu_y^2 + \mathbb{E}\left[\mathbb{E}(\delta_i^2|X_i)\mathbb{E}\left(\frac{Y_i^2}{(\pi^*(x_i))^2}\middle| X_i\right)\right] - 2\mu_y \mathbb{E}\left[\mathbb{E}(\delta_i|X_i)\mathbb{E}\left(\frac{Y_i}{\pi^*(x_i)}\middle| X_i\right)\right] \\
&= \mu_y^2 + \mathbb{E}\left[\pi^*(x_i)\frac{\sigma^2(x_i) + \mu^2(x_i)}{(\pi^*(x_i))^2}\right] - 2\mu_y \mathbb{E}\left[\pi^*(x_i)\frac{\mu(x_i)}{\pi^*(x_i)}\right] \\
&= \mu_y^2 + \mathbb{E}\left[\frac{\sigma^2(x_i)}{\pi^*(x_i)} + \frac{\mu^2(x_i)}{\pi^*(x_i)}\right] - 2\mu_y^2 = \mathbb{E}\left[\frac{\sigma^2(x_i)}{\pi^*(x_i)} + \frac{\mu^2(x_i)}{\pi^*(x_i)}\right] - \mu_y^2 \qquad (2.1)
\end{aligned}$$

$$\begin{aligned}
\mathrm{V}_2 = \mathbb{E}\big[\alpha\big(\delta_i, X_i\big)^2\big] &= \mathbb{E}\left[\left(\frac{\mathbb{E}(Y_i|X_i)}{\pi^*(x_i)}\Big(\delta_i - \pi^*(x_i)\Big)\right)^2\right] \\
&= \mathbb{E}\left[\left(\frac{\mu(x_i)}{\pi^*(x_i)}\right)^2 \delta_i^2 + \mu^2(x_i) - 2\frac{\delta_i}{\pi^*(x_i)}\mu^2(x_i)\right] \\
&= \mathbb{E}\left[\left(\frac{\mu(x_i)}{\pi^*(x_i)}\right)^2 \mathbb{E}\big(\delta_i^2|X_i\big)\right] - 2\mathbb{E}\left[\frac{\mu^2(x_i)}{\pi^*(x_i)}\mathbb{E}\big(\delta_i|X_i\big)\right] + \mathbb{E}\big[\mu(x_i)\big]^2 \\
&= \mathbb{E}\left[\left(\frac{\mu(x_i)}{\pi^*(x_i)}\right)^2 \pi^*(x_i)\right] - 2\mathbb{E}\left[\frac{\mu^2(x_i)}{\pi^*(x_i)}\pi^*(x_i)\right] + \mathbb{E}\big[\mu(x_i)\big]^2 \\
&= \mathbb{E}\left[\frac{\mu^2(x_i)}{\pi^*(x_i)}\right] - \mathbb{E}\big[\mu(x_i)\big]^2 \qquad (2.2)
\end{aligned}$$

$$\mathrm{V}_3 = -2\mathbb{E}\big[\psi\big(Y_i, \delta_i, X_i, \mu_y, \pi^*(x_i)\big) \cdot \alpha\big(\delta_i, X_i\big)\big] =$$

$$= -2\mathbb{E}\left[\left(\frac{\delta_i}{\pi^*(x_i)}Y_i - \mu_y\right) \cdot \frac{\mathbb{E}(Y_i|X_i)}{\pi^*(x_i)}\left(\delta_i - \pi^*(x_i)\right)\right]$$

$$= -2\mathbb{E}\left[\mu(x_i)Y_i\left(\frac{\delta_i}{\pi^*(x_i)}\right)^2 - \mu(x_i)Y_i\frac{\delta_i}{\pi^*(x_i)} - \mu(x_i)\mu_y\frac{\delta_i}{\pi^*(x_i)} + \mu(x_i)\mu_y\right]$$

$$= -2\mathbb{E}\left[\mathbb{E}\left(\mu(x_i)Y_i\left(\frac{\delta_i}{\pi^*(x_i)}\right)^2\right)\Big|X_i\right] + 2\mathbb{E}\left[\mathbb{E}\left(\mu(x_i)Y_i\frac{\delta_i}{\pi^*(x_i)}\right)\Big|X_i\right]$$

$$+ 2\mathbb{E}\left[\mathbb{E}\left(\mu(x_i)\mu_y\frac{\delta_i}{\pi^*(x_i)}\right)\Big|X_i\right] - 2\mathbb{E}\left[\big(\mu(x_i)\mu_y\big)|X_i\right]$$

$$= -2\mathbb{E}\left[\frac{\mu(x_i)}{(\pi^*(x_i))^2}\mathbb{E}\big(Y_i|X_i\big)\mathbb{E}\big(\delta_i^2|X_i\big)\right] + 2\mathbb{E}\left[\frac{\mu(x_i)}{\pi^*(x_i)}\mathbb{E}\big(Y_i|X_i\big)\mathbb{E}\big(\delta_i|X_i\big)\right]$$

$$+ 2\mathbb{E}\left[\frac{\mu(x_i)}{\pi^*(x_i)}\mu_y\mathbb{E}\big(\delta_i|X_i\big)\right] - 2\mu_y^2$$

$$= -2\mathbb{E}\left[\frac{\mu(x_i)}{(\pi^*(x_i))^2}\mu(x_i)\pi^*(x_i)\right] + 2\mathbb{E}\left[\frac{\mu(x_i)}{\pi^*(x_i)}\mu(x_i)\pi^*(x_i)\right]$$

$$+ 2\mathbb{E}\left[\frac{\mu(x_i)}{\pi^*(x_i)}\mu_y\pi^*(x_i)\right] - 2\mu_y^2$$

$$= -2\mathbb{E}\left[\frac{\mu^2(x_i)}{\pi^*(x_i)}\right] + 2\mathbb{E}\left[\mu^2(x_i)\right] + 2\mu_y\mathbb{E}\left[\mu(x_i)\right] - 2\mu_y^2$$

$$= 2\mathbb{E}\left[\frac{\mu^2(x_i)}{\pi^*(x_i)}\right] + 2\mathbb{E}\left[\mu^2(x_i)\right] \quad (2.3)$$

$\square$

2.2 Pseudo Hájek estimator

In this section we study the large sample properties of the pseudo Hájek estimator when the propensity score is estimated by Hirano-Imbens-Ridder method.

Theorem 3. *Let us assume that propensity score is estimated by Hirano-Imbens-Ridder method, then the pseudo Hájek estimator is asymptotically unbiased and its variance is*

$$V(\hat{T}_{pH}) = \frac{1}{N^2}\sum_{i=1}^{N}\left[\mathbb{E}\left(\frac{\sigma^2(x_i)}{\pi^*(x_i)} + \frac{\mu^2(x_i)}{\pi^*(x_i)}\right) - \frac{\mu_y^2}{\pi^*(x_i)}\right] + \mathbb{E}[O_p(1/N^2)].$$

Proof. We use the first order Taylor expansion to get a linear approximation of the estimator under study. This method makes it possible to approximate a general differentiable function to a linear function by which expectation and variance of estimators can be computed. In this regard, consider a regular function of two variables

$$f(x,y) = f(x_0,y_0) + f'_x(x_0,y_0)(x-x_0) + f'_y(x_0,y_0)(y-y_0) + \epsilon$$

where $f'_x(x_0,y_0)$ and $f'_y(x_0,y_0)$ are the first order partial derivatives of the function $f(x,y)$ at the point (x_0,y_0) and ϵ is the rest of the expansion including the higher order partial derivatives that converges to zero faster than the other terms as $(x,y)\to(x_0,y_0)$. Therefore as (x,y) gets closer and closer to (x_0,y_0) the rest ϵ can be considered negligible and the original function can be approximated with the remaining terms, that is

$$f(x,y) = f(x_0,y_0) + f'_x(x_0,y_0)(x-x_0) + f'_y(x_0,y_0)(y-y_0). \tag{2.4}$$

The second term of (2.4) is named linear approximation of the function $f(x,y)$ at the point (x_0,y_0). According to this method the pseudo Hájek estimator can be approximated as follows:

$$\hat{T}_{pH} = \mu_y + \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\pi^*(x_i)}\delta_i(Y_i-\mu_y) + O_p(1/N), \tag{2.5}$$

where $O_p(1/N)$ means that the remainder term is bounded in probability.

By computing the expectation of the expression (2.5) we get

$$
\begin{aligned}
\mathbb{E}(\hat{T}_{pH}) &= \mu_y + \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left(\frac{1}{\pi^*(x_i)}\delta_i Y_i\right) - \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left(\frac{1}{\pi^*(x_i)}\delta_i \mu_y\right) \\
&\quad + \mathbb{E}[O_p(1/N)] \\
&= \mu_y + \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\mathbb{E}\left(\frac{1}{\pi^*(x_i)}\delta_i Y_i\middle| X_i\right)\right] - \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\mathbb{E}\left(\frac{1}{\pi^*(x_i)}\delta_i \mu_y\middle| X_i\right)\right] \\
&\quad + \mathbb{E}[O_p(1/N)] \\
&= \mu_y + \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\frac{1}{\pi^*(x_i)}\mathbb{E}\big(\delta_i\big|X_i\big)\mathbb{E}\big(Y_i\big|X_i\big)\right] - \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\frac{1}{\pi^*(x_i)}\mu_y\mathbb{E}\big(\delta_i\big|X_i\big)\right] \\
&\quad + \mathbb{E}[O_p(1/N)] \\
&= \mu_y + \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\pi^*(x_i)}\pi^*(x_i)\mathbb{E}\big[\mu(x_i)\big] - \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\pi^*(x_i)}\mu_y\pi^*(x_i) + \mathbb{E}[O_p(1/N)] \\
&= \mu_y + \frac{1}{N}\sum_{i=1}^{N}\mu_y - \frac{1}{N}\sum_{i=1}^{N}\mu_y + \mathbb{E}[O_p(1/N)] \\
&= \mu_y + \mathbb{E}[O_p(1/N)],
\end{aligned}
$$

which means that the estimator $\hat{T}_{pH}$ is asymptotically unbiased. □

Now we consider the variance of $\hat{T}_{pH}$. Because of (2.5), an approximate variance of the pseudo Hájek estimator is given by:

$$
\mathrm{V}(\hat{T}_{pH}) = \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\frac{1}{\pi^*(x_i)}\delta_i(Y_i - \mu_y)\right]^2 + \mathbb{E}[O_p(1/N^2)].
$$

Proof. By developing the first term we have:

$$
\begin{aligned}
\mathbb{E}\left[\frac{1}{\pi^*(x_i)}\delta_i(Y_i-\mu_y)\right]^2 &= \mathbb{E}\left[\frac{\delta_i^2}{(\pi^*(x_i))^2}(Y_i-\mu_y)^2\right] \\
&= \mathbb{E}\left[\frac{\delta_i^2}{(\pi^*(x_i))^2}(Y_i^2+\mu_y^2-2\mu_y Y_i)\right] \\
&= \mathbb{E}\left[\frac{\delta_i^2}{(\pi^*(x_i))^2}Y_i^2+\mu_y^2\frac{\delta_i^2}{(\pi^*(x_i))^2}-2\mu_y\frac{\delta_i^2}{(\pi^*(x_i))^2}Y_i\right] \\
&= \mathbb{E}\left[\mathbb{E}\left(\frac{\delta_i^2}{(\pi^*(x_i))^2}Y_i^2\middle|X_i\right)\right]+\mu_y^2\mathbb{E}\left[\mathbb{E}\left(\frac{\delta_i^2}{(\pi^*(x_i))^2}\middle|X_i\right)\right] \\
&\quad -2\mu_y\mathbb{E}\left[\mathbb{E}\left(\frac{\delta_i^2}{(\pi^*(x_i))^2}Y_i\middle|X_i\right)\right] \\
&= \mathbb{E}\left[\mathbb{E}(\delta_i^2|X_i)\frac{\mathbb{E}(Y_i^2|X_i)}{(\pi^*(x_i))^2}\right]+\frac{\mu_y^2}{\pi^*(x_i)} \\
&\quad -2\mu_y\mathbb{E}\left[\mathbb{E}(\delta_i^2|X_i)\frac{\mathbb{E}(Y_i|X_i)}{(\pi^*(x_i))^2}\right] \\
&= \mathbb{E}\left(\frac{\sigma^2(x_i)}{\pi^*(x_i)}+\frac{\mu^2(x_i)}{\pi^*(x_i)}\right)+\frac{\mu_y^2}{\pi^*(x_i)}-2\mu_y\frac{\mathbb{E}(\mu(x_i))}{\pi^*(x_i)} \\
&= \mathbb{E}\left(\frac{\sigma^2(x_i)}{\pi^*(x_i)}+\frac{\mu^2(x_i)}{\pi^*(x_i)}\right)+\frac{\mu_y^2}{\pi^*(x_i)}-\frac{2\mu_y^2}{\pi^*(x_i)} \\
&= \mathbb{E}\left(\frac{\sigma^2(x_i)}{\pi^*(x_i)}+\frac{\mu^2(x_i)}{\pi^*(x_i)}\right)-\frac{\mu_y^2}{\pi^*(x_i)}.
\end{aligned}
$$

Therefore

$$
\mathrm{V}(\hat{T}_{pH})=\frac{1}{N^2}\sum_{i=1}^{N}\left[\mathbb{E}\left(\frac{\sigma^2(x_i)}{\pi^*(x_i)}+\frac{\mu^2(x_i)}{\pi^*(x_i)}\right)-\frac{\mu_y^2}{\pi^*(x_i)}\right]+\mathbb{E}[O_p(1/N^2)].
$$

□

Finally, we provide another way to prove that the pseudo Hájek estimator is asymptotically unbiased as follows.

Proof. At first let us consider the distance between the average of the estimated weights and the average of the true weights:

$$D = \left| \frac{1}{N} \sum_{i=1}^{N} \frac{\delta_i}{\hat{\pi}(x_i)} - \frac{1}{N} \sum_{i=1}^{N} \frac{\delta_i}{\pi^*(x_i)} \right| = \frac{1}{N} \sum_{i=1}^{N} \delta_i \left| \frac{1}{\hat{\pi}(x_i)} - \frac{1}{\pi^*(x_i)} \right|$$

$$\leq \frac{1}{N} \sum_{i=1}^{N} \delta_i \cdot \sup_{x \in X} \left| \frac{1}{\hat{\pi}(x_i)} - \frac{1}{\pi^*(x_i)} \right|. \quad (2.6)$$

Since $\delta_1, \delta_2, \ldots, \delta_N$ are independent and have the same distribution we can apply the Law of Large Numbers:

$$\frac{1}{N} \sum_{i=1}^{N} \delta_i \xrightarrow{P} \mathbb{E}(\delta_i) = \mathbb{E}_x[\pi^*(x)] = \pi^* \in (0,1), \quad \text{as } N \to \infty \quad (2.7)$$

and

$$\frac{1}{N} \sum_{i=1}^{N} \frac{\delta_i}{\pi^*(x_i)} \xrightarrow{P} \mathbb{E}\left[\frac{\delta_i}{\pi^*(x_i)}\right] = \mathbb{E}_x\left[\frac{1}{\pi^*(x)} \mathbb{E}(\delta|X)\right] = \mathbb{E}_x\left[\frac{\pi^*(x)}{\pi^*(x)}\right] = 1. \quad (2.8)$$

We may also write

$$\sup_{x \in \mathcal{X}} \left| \frac{1}{\hat{\pi}(x_i)} - \frac{1}{\pi^*(x_i)} \right| = \sup_{x \in \mathcal{X}} \left| \frac{\pi^*(x_i) - \hat{\pi}(x_i)}{\hat{\pi}(x_i)\pi^*(x_i)} \right| \leq$$

$$\sup_{x \in \mathcal{X}} \left| \pi^*(x_i) - \hat{\pi}(x_i) \right| \xrightarrow{P} 0, \quad \text{as } N \to \infty \quad (2.9)$$

where (2.8) holds if the propensity score is estimated by Hirano-Imbens-Ridder method.

By combining the previous results we obtain:

$$
\begin{aligned}
\sqrt{n}\left(\frac{\frac{1}{N}\sum_{i=1}^{N}\frac{\delta_i}{\hat{\pi}(x_i)}Y_i}{\frac{1}{N}\sum_{i=1}^{N}\frac{\delta_i}{\hat{\pi}(x_i)}} - \mu_y\right) &= \sqrt{n}\left(\frac{\frac{1}{N}\sum_{i=1}^{N}\frac{\delta_i}{\hat{\pi}(x_i)}(Y_i - \mu_y)}{\frac{1}{N}\sum_{i=1}^{N}\frac{\delta_i}{\hat{\pi}(x_i)}}\right) \\
&= \sqrt{n}\left(\frac{\frac{1}{N}\sum_{i=1}^{N}\frac{\delta_i}{\hat{\pi}(x_i)}(Y_i - \mu_y)}{\frac{1}{N}\sum_{i=1}^{N}\frac{\delta_i}{\hat{\pi}(x_i)} + \frac{1}{N}\sum_{i=1}^{N}\frac{\delta_i}{\pi^*(x_i)} - \frac{1}{N}\sum_{i=1}^{N}\frac{\delta_i}{\pi^*(x_i)}}\right) \\
&= \sqrt{n}\left(\frac{\frac{1}{N}\sum_{i=1}^{N}\frac{\delta_i}{\hat{\pi}(x_i)}(Y_i - \mu_y)}{\frac{1}{N}\sum_{i=1}^{N}\frac{\delta_i}{\hat{\pi}(x_i)}}\right)
\end{aligned}
$$

where the denominator converges to 1 in probability, while the numerator follows the same distribution of the pseudo Horvitz-Thompson estimator with estimated weights (Section 2.1.2). Hence the pseudo Hájek estimator is asymptotically unbiased when the propensity score is estimated by means Hirano-Imbens-Ridder method. □

Chapter 3

Estimating variance and confidence intervals

In this chapter we describe the principles to estimate variance and confidence intervals of the proposed estimators. We focus on the bootstrap approach for finite population surveys based on the idea of generating pseudo-populations. Then we concentrate on the bootstrap method by Holmberg (1998) for probability proportional-to-size designs without replacement that is the starting point to develop the simulation study. Some of the discussion is abridged from Quatember (2015).

3.1 The bootstrap method

When no explicit variance formula is available and the calculations for Taylor linearization are too cumbersome, computer-intensive methods that use computer power instead of heavy calculations can be applied. One technique of estimating the theoretical variance of an estimator is the *bootstrap* method. This strategy falls under the family of resampling methods. The basic bootstrap procedure generates resamples of the same size as the original sample, while another strategy, the *jackknife* method, generates resamples from the original sample, which consist all but one or a certain number of elements of the original sample drawn.

Boostrap was originally developed by Efron (1979) for the estimation of the sampling distribution of an estimator $\hat{\theta}$ for the parameter θ on the basis of a random sample and an unknown probability distribution ϕ of a variable Y under study. For this purpose, a sample of *i.i.d.* variables is observed. This procedure can be described as follows:

1. Construct the empirical distribution of the study variable, $\hat{\phi}$ (e.g. for a random sample of size n putting mass $1/n$ at each sample point). The empirical distribution of a random variable Y as observed in the *i.i.d.* sample can be interpreted as the non-parametric Maximum-Likelihood (ML) estimator of the true probability distribution ϕ of Y.

2. Draw *i.i.d.* bootstrap samples of the same size as the original sample from this empirical distribution. Call each of them *bootstrap sample*.

3. Approximate the true sample distribution of $\hat{\theta}$ by the theoretical bootstrap distribution of the estimator calculated in all possible resamples. Call this the *bootstrap distribution.*

This bootstrap distribution equals the sampling distribution of the estimator if the empirical distribution of the variable under study equals its probability distribution. In symbols if $\hat{\phi} = \phi$.

As Efron (1979) stated: "the difficult part of the bootstrap procedure is the actual calculation of the bootstrap distribution". Three methods are possible:

Method 1. The direct theoretical calculation.

Method 2. An approximation by Taylor expansion.

Method 3. A Monte Carlo approximation.

The latter has turned out to be most common. In this case, B bootstrap samples of the same size as that of the original sample are drawn with replacement from the empirical distribution of Y, which can be seen as (non-parametric) the Maximum-Likelihood estimator of the underlying probability distribution ϕ of Y. Within each of the B bootstrap samples, $s_1^*, \ldots, s_B^*$, the estimator $\hat{\theta}_b^*$ is calculated in the same way that the estimator $\hat{\theta}$ was calculated in the original *i.i.d.* sample s ($b = 1, \ldots, B$).

For a large B, the distribution of $\hat{\theta}_b^*$ is interpreted as an estimation of the sample distribution of $\hat{\theta}$. Hence, the theoretical variance $\mathrm{V}(\hat{\theta})$ is estimated by the Monte Carlo variance estimator given by

$$\hat{V}_b(\hat{\theta}) = \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\theta}_b^* - \bar{\hat{\theta}}^*\right)^2 \tag{3.1}$$

with

$$\bar{\hat{\theta}}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^*,$$

being the mean value of estimators $\hat{\theta}_b^*$ from the B bootstrap samples. For approximately normally distributed $\hat{\theta}_b^*$ values, this variance estimator can be used for the calculation of an approximate confidence interval. For a large B, also for nonnormally distributed bootstrap estimators, a confidence interval can be calculated by applying the percentile method (Efron, 1981). This method directly uses the $\alpha/2$ and $(1-\alpha/2)$ quantile of the observed distribution of the estimators $\hat{\theta}_b^*$ as the lower and the upper bound of the confidence interval, respectively.

With increasing computer power, this technique has also become attractive for finite population surveys. However, 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. Consequently, in complex designs, classical bootstrap methods result in a biased variance estimator when the sampling design is not taken into account. Suitable adaptations are needed in order to consider complex sampling designs consisting of complex estimators and sampling schemes drawing the sample units without replacement. For this purpose, two main approaches are available in the literature: *ad hoc* approach and *plug-in* approach (Ranalli and Mecatti, 2012).

Several methods can be included in the first approach. One of them rescales the observations in the resamples drawn with replacement from the original without-replacement sample in a way that the bootstrap variance (3.1) approximates the actual variance for a given sampling design (Rao and Wu, 1988). Another *ad hoc* method is to use the with-replacement bootstrap technique and adjust its bootstrap variance estimator to the parameter by choosing an appropriate size for the bootstrap samples (McCarthy and Snowden, 1985). Sitter (1992) presented the Mirror-Match Method, in which subsamples of the original sample are drawn repeatedly according to the original sampling plan with a subsample size chosen to appropriately match the original variance of the estimator. Antal and Tillé (2011) discuss another method, in which different with and without replacement resampling designs are combined in such a way that the bootstrap estimators 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.

The second major approach (called *plug-in*), to deal with non-*i.i.d.* data, is to generate an artificial population, the “pseudo-population” from the observed sample data. A pseudo-population is built up by using sample data, and assumed to estimate the unknown actual population. According to the mimicking principle (Hall, 1992), bootstrap samples (i.e. the resampling result) are selected from this estimated population with the same sample size as the original sample and by mimicking the original sampling design to the largest extent (Ranalli and Mecatti, 2012).

3.2 Pseudo-population bootstrap methods

For a direct extension of the *i.i.d.* bootstrap to finite population sampling, the population $\mathcal{U}$ of N elements plays the role of the unknown probability distribution in the *i.i.d.* bootstrap. The population elements are characterized by their values y_k of the variable Y under study and X_k of possible auxiliary variables X $(k = 1, \dots, N)$.

Gross (1980) was the first to adapt the original bootstrap method to the specific case of a simple random sample without replacement (SI), but only with the restriction of integer design weights $\frac{N}{n}$ (Figure 3.1). For this purpose, from a SI sample s, a set-valued finite population estimator $\mathcal{U}_G^*$ of size $N_G^* = N$ of the true population $\mathcal{U}$ of size N is generated by replicating each sample value y_k exactly $\frac{N}{n}$ providing a variable Y^* denoting these "clones" of the sample values. Hence, the bootstrap population $\mathcal{U}_G^*$ can be interpreted as the finite population with the Maximum Likelihood regarding the sample drawn (Chao and Lo, 1994).

The idea behind pseudo-populations is simple: as the sample and population sizes increase, the pseudo-population tends to be "similar" to the real finite population. Hence, it would be intuitive to use a pseudo-population that is as similar as possible to the actual population. In a sense, the pseudo-population should be somehow calibrated with respect to the population (Conti et al., 2017).

In practical applications, i.e. for finite n, a crucial aspect that would potentially affect the performance of resampling, is how the pseudo-population is constructed. Recently Conti et al. (2017) raised the question of how different choices for constructing the pseudo-population $\mathcal{U}^*$ (where resampling is actually performed) may affect the accuracy of the resulting inference in practical applications. They showed that the construction process of the pseudo-population is a crucial choice for small to moderate population and sample sizes, under general sampling designs.

In the next subsections more proposals based on different approaches are illustrated, which lead to different pseudo-populations. In particular, a detailed description of the Holmberg's bootstrap algorithm for complex sampling design with inclusion probability proportional to an auxiliary variable X is provided.

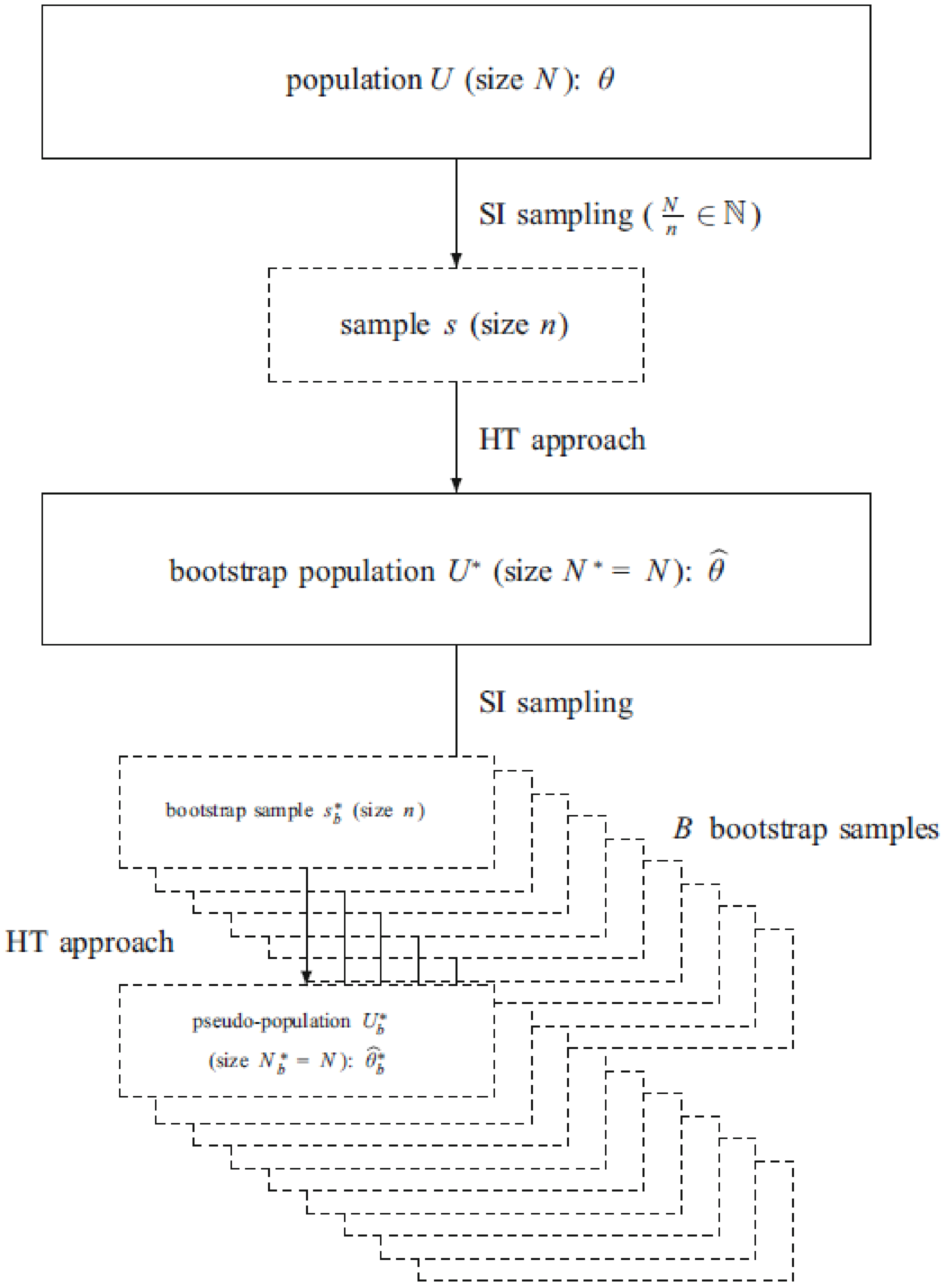

Figure 3.1: Estimating the sampling distribution of an estimator $\hat{\theta}$ applying the bootstrap method in SI sampling with integer design weights according to Gross (1980) - Source Quatember (2015)

3.2.1 Horvitz-Thompson pseudo-population

The rationale behind the Horvitz-Thompson estimation process expressed by

$$\hat{Y}_{HT} = \sum_{i \in s} \frac{1}{\pi_i} y_i$$

can be described by the idea of generating an artificial population estimating appropriately the original population with respect to the parameter under study, i.e. the total $Y_T = \sum_{i=1}^{N} y_i$ of a variable Y.

The generation process starts at population $\mathcal{U}$. Each element i of $\mathcal{U}$ is assigned a certain value y_i of variable Y, but the parameter Y_T is unknown. In the next step, one of all possible samples, which can be drawn according to a given probability sampling scheme, is selected. In this sample s of n elements, variable Y is observed. In the next step, the original population $\mathcal{U}$ of size N is estimated with respect to the parameter Y_T of variable Y by a pseudo-population $\mathcal{U}^*_{HT}$. In the final step, the Horvitz-Thompson estimator of Y_T, i.e. $\hat{Y}^*_{HT}$, is calculated as the total of the replications of y in $\mathcal{U}^*_{HT}$.

For the generation of the pseudo-population $\mathcal{U}^*_{HT}$, the variable value y_i of the unit i in the sample is replicated $\frac{1}{\pi_i}$ times, for each $i \in s$. Hence, the design weights can be seen as the replication factors of this process. Pseudo-population $\mathcal{U}^*_{HT}$ has $N^*_{HT} = \sum_s \frac{1}{\pi_i}$ elements that is in general not equal to N, while the expectation is $\mathbb{E}(N^*_{HT}) = \sum_{\mathcal{U}} \frac{1}{\pi_k} I_k = N$. However, the ratio $\frac{N^*_{HT}}{N}$ tends in probability to 1 as N and n increase (Conti et al., 2017).

Note that the design weights $\frac{1}{\pi_i}$ are not integers as a rule. Hence, the Horvitz-Thompson pseudo-population $\mathcal{U}^*_{HT}$ is special in the sense that it may not only contain $\lfloor \frac{1}{\pi_i} \rfloor$ whole units with the same value y_i of variable Y (where $\lfloor \cdot \rfloor$ denotes the integer part of a real number), but also $\frac{1}{\pi_i} - \lfloor \frac{1}{\pi_i} \rfloor$ piece of unit with that value when $\frac{1}{\pi_i} - \lfloor \frac{1}{\pi_i} \rfloor > 0$, $i \in s$.

Consequently, the efficiency of the unbiased Horvitz-Thompson estimator $\hat{Y}^*_{HT}$ for Y_T depends on the quality of the estimation of $\mathcal{U}$ by $\mathcal{U}^*_{HT}$ with respect to Y or, to be even more precise, with respect to parameter Y_T.

3.2.2 Multinomial pseudo-population

For $k = 1, \ldots, N$ perform independent trials consisting in choosing a unit from the original sample, where each unit i is selected with probability

$$\frac{\dfrac{1}{\pi_i}}{\sum\limits_{j\in s}\dfrac{1}{\pi_j}} = \frac{\dfrac{1}{x_i}}{\sum\limits_{j\in s}\dfrac{1}{x_j}}. \tag{3.2}$$

If at trial k unit i is selected, unit k of the pseudo-population will take values $y_k^* = y_i$ and $x_k^* = x_i$. If N_i^*, $i \in s$, is the number of replications for unit i in the pseudo-population, then N_i^* has a multinomial distribution with expectation

$$\mathbb{E}[N_i^*|\delta_N, Y_N, X_N] = N\frac{\dfrac{\delta_i}{\pi_i}}{\sum\limits_{j=1}^{N}\dfrac{\delta_j}{\pi_j}}, \tag{3.3}$$

variance

$$\mathrm{V}[N_i^*|\delta_N, Y_N, X_N] = N\left(\frac{\dfrac{\delta_i}{\pi_i}}{\sum\limits_{j=1}^{N}\dfrac{\delta_j}{\pi_j}}\right)\left(1 - \frac{\dfrac{\delta_i}{\pi_i}}{\sum\limits_{j=1}^{N}\dfrac{\delta_j}{\pi_j}}\right) \tag{3.4}$$

and covariance

$$\mathrm{Cov}[N_i^*, N_h^*|\delta_N, Y_N, X_N] = -N\frac{\dfrac{\delta_i\delta_h}{\pi_i\pi_h}}{\left(\sum\limits_{j=1}^{N}\dfrac{\delta_j}{\pi_j}\right)^2} \qquad h \neq i. \tag{3.5}$$

This approach goes essentially back to Pfeffermann and Sverchkov (2004) and guarantees by construction a pseudo-population calibrated with respect to the population size (Conti et al., 2017). This means that pseudo-population replications satisfy constraint on population size: they are as close as possible to the initial N.

3.2.3 The Holmberg's bootstrap algorithm

As we have seen in Section 3.1 several proposals to adapt the original Efron's bootstrap to handle with non-*i.i.d.* situations have been introduced, particularly for the without replacement selection. Among the methods based on pseudo-population, Holmberg (1998) proposed a generalization of this approach for a general sampling design without replacement and with inclusion probability proportional to an auxiliary variable X (usually referred as IPPS sampling or πPS sampling), i.e. $\pi_i \propto xi/X_T$, where $X_T = \sum_{i=1}^{N} x_i$ is the population auxiliary total.

A sampling design without replacement and with inclusion probability proportional to an auxiliary variable X paired with the well-known unbiased Horvitz-Thompson estimator $\hat{Y}_{HT} = \sum_{i=1}^{n} y_i/\pi_i$ devises a strategy methodologically appealing, since the estimator variance $V(\hat{Y}_{HT})$ tends to zero as the relationship between X and Y approaches proportionality (Barbiero and Mecatti, 2009).

Let $\pi(x_i) = nx_i/X_T$ be the first-order inclusion probability under the πPS sampling design, let $s \subset \mathcal{U}_N$ be a sample of size n selected according to the design $p(\cdot)$, and let

$$r_i = \frac{1}{\pi(x_i)} - \left\lfloor \frac{1}{\pi(x_i)} \right\rfloor, \qquad 0 \leq r_i < 1, \quad i \in s,$$

where $\lfloor \cdot \rfloor$ denotes the greatest integer equal to or smaller than.

Finally, for $i \in s$, let ϵ_i be independent Bernoulli random variables with parameters r_i, i.e.

$$r_i = Pr(\epsilon_i = 1)$$

$$1 - r_i = Pr(\epsilon_i = 0).$$

The bootstrap approach suggested by Holmberg (1998) can be described as follows:

1. For $i \in s$, let ϵ_i be independent realizations of the Bernoulli random variables, and define
$$N_i^* = \left\lfloor \frac{1}{\pi(x_i)} \right\rfloor + \epsilon_i.$$

2. Create a resampling population $\mathcal{U}^*$ by copying each element $i \in s$ in such a way that element i is copied N_i^* times, i.e.
$$\mathcal{U}^* = \{N_i^*, i \in s\},$$
with $N^* = \sum_{i \in s} N_i^*$. All N_i^* elements that are copies of element $i \in s$ are assigned the value $\{y_i, x_i\}$.

3. Draw a sample s_1^* of size $n^* = n$ from $\mathcal{U}^*$ by applying the same sample selection scheme as for selecting s, which means that pseudo unit i is included in the sample with probability $\pi(\cdot)$ and not included with probability $1 - \pi(\cdot)$. We refer to it as the *bootstrap sample*.

4. Compute a bootstrap replicate $\hat{\theta}_1^* = \hat{\theta}(s_1^*)$.

5. Repeat steps 3 and 4 B times. The Monte Carlo bootstrap variance estimator for $\hat{\theta}$ is then given by

$$\hat{V}_b(\hat{\theta}) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_b^* - \bar{\hat{\theta}}^*\right)^2,$$

where $\bar{\hat{\theta}}^* = B^{-1}\sum_{b=1}^{B}\hat{\theta}_b^*$.

Note that in the Holmberg's method a further step in the bootstrap algorithm is needed for constructing the bootstrap population $\mathcal{U}^*$. Particularly, in step 1 n random variables have to be simulated in order to compute the weights ${N_i}^*$. Then, if ϵ_i does not equal zero for some i, an entire class $\mathcal{U}^* = \{\mathcal{U}_h^*, h = 1, 2, \ldots, 2^n\}$ of 2^n possible bootstrap populations remains defined. The further step is actually performed to select a unique bootstrap population by randomization into $\mathcal{U}^*$. As a consequence, the Holmberg's πPS-bootstrap results computationally heavy and resource consuming (Barbiero and Mecatti, 2009).

Chapter 4

Simulation and empirical studies

This chapter is devoted to the main results of the simulation study carried out to evaluate the methodology proposed. We tested and compared the validity of the proposed estimators by measuring their accuracy in terms of bias, variance and confidence intervals. For this purpose, a specific bootstrap method for complex sampling design was applied that is based on the concept of pseudo-population. We refer to it as the pseudo-population bootstrap method of Chauvet (Chauvet, 2007).

As seen in the previous chapter (Chapter 3), the unknown quantity in the classical *i.i.d.* model of classical statistics is the distribution ϕ of the variable of interest Y. To perform the bootstrap procedure for this model, ϕ is first estimated by the empirical distribution function $\hat{\phi}_n$, and then *i.i.d.* observations from $\hat{\phi}_n$ are generated. In survey sampling, the unknown is the population $\mathcal{U}$ from which the sample is drawn. Therefore, under the pseudo-population bootstrap (PPB) approach, $\mathcal{U}$ is estimated by creating a pseudo-population via repeating the original sample using principles from the original sampling design. Then, the bootstrap sample is drawn from the resulting pseudo-population using the original sampling design. By obeying the original scheme to draw the bootstrap sample from the pseudo-population, the finite population correction factors, e.g. the $1 - f$ in the case of simple random sample without replacement, are naturally captured by the bootstrap variance estimator. This important property has persuaded many researchers to widely study this approach (Mashreghi et al., 2016).

4.1 The bootstrap algorithm for unequal probability sampling

In survey sampling, the unknown is the population $\mathcal{U}$ from which the sample is drawn. Therefore, under the pseudo-population bootstrap approach, $\mathcal{U}$ is estimated by creating a pseudo-population via repeating the original sample using principles from the original sampling design. Then, the bootstrap sample is drawn from the resulting pseudo-population using the original sampling design.

In this section, we focus on bootstrap method for unequal probability sampling design based on the concept of pseudo-population. More specifically, we present the bootstrap algorithm to evaluate the performances of the estimators proposed. This method is inspired by the bootstrap method of Chauvet (2007) for Poisson Sampling and also reported by Mashreghi et al. (2016).

As we have seen in Section 1.1, in Poisson sampling each element of the population is selected independently in the sample with probability π_i and therefore the sample size is random.

The general algorithm for unequal probability sampling can be described as follows:

1. Repeat the pair (y_i, π_i), $\lfloor \frac{1}{\pi_i} \rfloor$ times for all i in s to create, $\mathcal{U}^f$, the fixed part of the pseudo-population.

2. To complete the pseudo-population, $\mathcal{U}^*$, draw $\mathcal{U}^{c*}$ from $\{(y_i, \pi_i)\}_{i \in s}$ using Poisson sampling with inclusion probability $r_i = \frac{1}{\pi_i} - \lfloor \frac{1}{\pi_i} \rfloor$ for the i^{th} pair. Denote the pseudo-population by $\mathcal{U}^* = \mathcal{U}^f \cup \mathcal{U}^{c*} = \{(\breve{y}_i, \breve{\pi}_i)\}_{i \in \mathcal{U}^*}$ where $(\breve{y}_i, \breve{\pi}_i)$ is the i^{th} pair of the pseudo-population and corresponds to one of the values of the variable obtained from the sample and its corresponding probability of selection according to the sample design.

3. Take the bootstrap sample s^* from $\mathcal{U}^*$ using the same sampling design that led to s, but with inclusion probability $\pi_i^{'}$ for the i^{th} unit in $\mathcal{U}^*$, as defined in the sequel.

4. Compute the bootstrap statistic, $\hat{\theta}^*$, on the bootstrap sample s^*.

5. Repeat Steps 3 and 4 a large number of times, B, to get $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$. Let

$$\hat{V}_B^* = \frac{1}{B-1} \sum_{b=1}^{B} \big(\hat{\theta}_b^* - \bar{\hat{\theta}}^*\big)^2,$$

where $\bar{\hat{\theta}}^* = B^{-1} \sum_{b=1}^{B} \hat{\theta}_b^*$.

6. Repeat Steps 2 to 5 a large number of times, D, to get $\hat{V}_{1B}^*, \ldots, \hat{V}_{DB}^*$.

It is worth noting the main similarities and differences between this algorithm and the Holmberg (1998) algorithm as described by Mashreghi et al. (2016). Both are designed for unequal (single-stage) probability sampling design and aim to emulate the original sampling design as was the case with simple random sample without replacement: the method of Chauvet (2007) for Poisson sampling and the method of Holmberg (1998) for probability proportional to size sampling.

We observe that the pseudo-population is constructed the same way as Holmberg method (Section 3.2.3). However, to draw the bootstrap sample, the original sampling mechanism used to draw s from $\mathcal{U}$ is applied, but with inclusion probability $\pi_i^{'}$. Note that $\pi_i^{'}$ may be different from the original inclusion probability, that is the inclusion probability of unit i in the original sample.

Holmberg (1998) proposed his bootstrap method for inclusion probability proportional to size sampling designs; since the size distribution for the pseudo-population is not the same as the original, the first order inclusion probability used in Step 3 of the algorithm is modified to $\pi_i^{'} = n\breve{\pi}_i / \sum_{j \in \mathcal{U}^*} \breve{\pi}_i$. However, to compute the Monte Carlo variance estimator, he ignores the variability induced by creating the pseudo-population.

Chauvet (2007) estimates the variance of the population total for Poisson sampling design. To obtain the bootstrap variance estimator of Chauvet, Poisson sampling with the original inclusion probabilities $\pi_i^{'} = \breve{\pi}_i$ in Step 3 of the algorithm is used. Recall that $\breve{\pi}_i$ is the probability of selection of the value $\breve{y}_i$, one of the pairs making the pseudo-population and therefore one of the pairs (y_j, π_j) of the original sample. Under this method, the bootstrap variance estimator is $\mathbb{E}_{u^*}\left[V_{p*}(\hat{\theta}^* | \mathcal{U}^*)\right]$ which is approximated by

$$\hat{V}^* = \frac{1}{D} \sum_{d=1}^{D} V_{dB}^*.$$

Furthermore, note that the resulting pseudo-population may not have the same size as the original population size, N. But, letting M_i be the number of times unit i appears in $\mathcal{U}^*$, we have $\mathbb{E}_p \mathbb{E}_{u^*}(\sum_{i \in s} M_i)$

While Holmberg (1998) did not address the problem of constructing confidence intervals, Chauvet (2007) computed bootstrap percentile intervals, more specifically percentile intervals constructed from the DB values of $\hat{\theta}_i^*$. Given that the bootstrap parameter θ^* changes with each pseudo-population, the bootstrap percentile intervals should be computed from the quantiles of $\hat{\theta}_i^* - \theta_i^*$, where the pseudo-population changes with each bootstrap sample.

The basic scheme of the simulation process is shown in Figure 4.1.

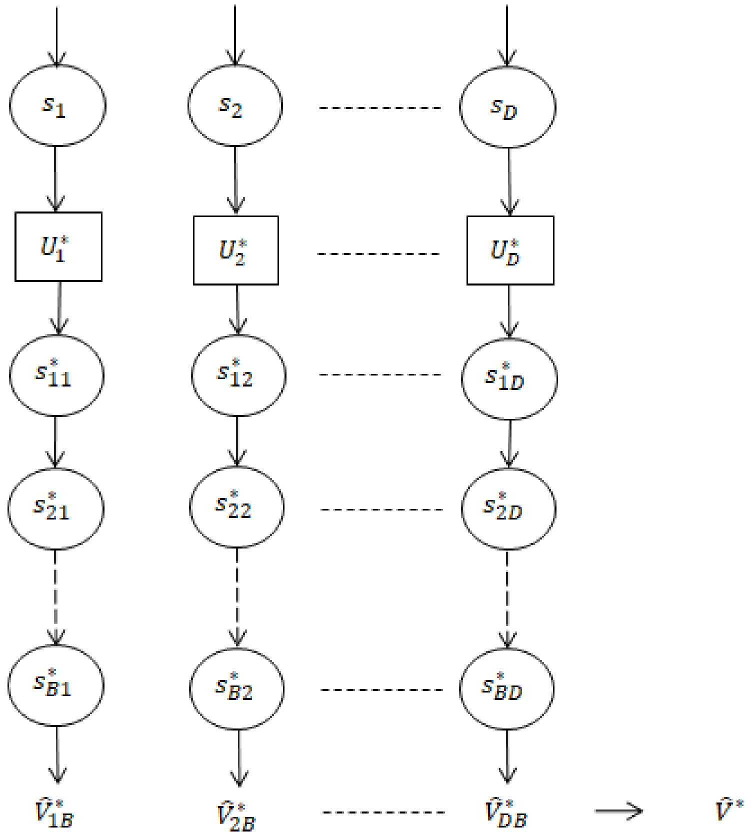

Figure 4.1: Simulation scheme under pseudo-population approach

4.2 Simulation design

Given the population size, N, the variable of interest Y and two auxiliary variables, X_1 and X_2, were generated according to a multinormal distribution with given mean vector $\underline{\mu}$ and covariance matrix Σ. The values of σ_{y,x_1} and σ_{y,x_2} were chosen to guarantee high correlation between Y and X (not less than 0.7), since high correlation is desirable to obtain more efficient estimates given the assumptions. At the same time, the auxiliary variables had low correlation among themselves in order to achieve considerable gain in efficiency. In fact, when the auxiliary variables are highly correlated, there is practically no gain in efficiency by use of an additional variable, and a use of a single auxiliary variable is recommended.

The true inclusion probabilities (propensity scores) were modeled by logistic regression model where coefficients of $x1$ and $x2$ were assumed to be known. More specifically, for each unit i of the population ($i = 1, 2, \ldots, N$), a value was generated from the *logit* model $g(x) = \beta_0 + \beta_1 x1 + \beta_2 x2$ where $\beta_0 = -0.8$, $\beta_1 = 0.2$ and $\beta_2 = 0.3$. Therefore the true inclusion probability, $\pi^*(x)$, was obtained by the inverse of the *logit* function, i.e. $\pi^*(x) = \exp[g(x)]/1 + \exp[g(x)]$.

Given the values of the true inclusion probabilities, $\pi_1^*, \pi_2^*, \ldots, \pi_N^*$, for all units of the population, N *i.i.d.* Bernoulli random variables δ_i, $i = 1, 2, \ldots, N$, were generated, each of them with probability equals to the true propensity score assigned to the corresponding unit of the population. Those units with $\delta_i = 1$ represents the units included in a sample.

After generating original data of the population, the propensity score was estimated for each unit i ($i = 1, 2, \ldots, N$) by using logistic regression model, now representing Hirano-Imbens-Ridder estimator with a finite number of terms, where the response variable was given by the Bernoulli random variable, δ, as above generated and independent variables were the auxiliary variables, X_1 and X_2, whose values are known for all units in the population.

On the basis of the estimated propensity scores one sample was drawn following the original sampling scheme and one pseudo-population was generated from it as described in the previous Section (4.1). Then, B bootstrap samples were drawn from the pseudo-population following the original sampling scheme, and in each bootstrap sample the pseudo Horvitz-Thompson (1.4) and the pseudo Hájek estimators were calculated (1.5). This gives a bootstrap estimation of the sample distribution of the estimators, $\hat{T}_{pHT}$ and $\hat{T}_{pH}$, of the unknown population mean in the original population $\mathcal{U}_N$. This process has been replicated a large number of times, D, in order to take into account the variability of the pseudo-population.

Two simulation trials were carried out: in the first one the size of the original population was set to $N = 500$, while in the second one N was set to 1000. The number of bootstrap samples was set at four times the size of the original population, thus 2000 and 4000, respectively, while the number of iterations was set at twice the size of the original population. For example, if the size of the population is set equal to 500, the number of samples, D, from which pseudo-populations are generated is equal to 1000, while the number of bootstrap samples, B, which are drawn from each generated pseudo-population, is equal to 2000. D Monte Carlo runs, simulating the sample space, have been combined with B resampling runs from each generated sample.

Simulation has been performed in the R environment. The full script for this simulation can be found in Appendix.

4.3 Simulation results

In order to evaluate the performance of the proposed estimators the following Monte Carlo (MC) indicators have been computed:

- Percentage Relative Bias (PRB), concerning the ability of the resampled distribution of an estimator of the population mean to match the (original) sample mean as its empirical first moment

$$PRB = \mathbb{E}_{MC}\left[\frac{\mathbb{E}^*(\hat{\theta}^*) - \hat{\theta}}{\hat{\theta}}\right] \times 100$$

  where $\mathbb{E}^*$ indicates the empirical average over the B resampling runs and by taking $\hat{\theta} = \overline{Y}_N$ (Conti et al., 2017);

- 95% Confidence Interval based on the bootstrap percentile method (bootstrap distribution).
  The percentile method for the construction of a reasonable $(1-\alpha)100\%$ confidence interval for a parameter θ directly uses the $\alpha/2$ and $1-\alpha/2$ quantile of the observed bootstrap distribution of the estimator $\hat{\theta}$ as the lower and the upper bound of the confidence interval, respectively (Efron, 1981). Given that the bootstrap parameter $\hat{\theta}^*$ changes with each pseudo-population, the bootstrap percentile intervals should be computed from the quantiles of $\hat{\theta}_i^* - \theta_i^*$ where the pseudo-population changes with each bootstrap sample.

- 95% Confidence Interval Coverage (or Coverage Probability), i.e. the proportion of intervals which contain the parameter of interest, based on two methods: (i) the bootstrap percentile method; (ii) the boostrap-normal confidence interval method given by

$$[\hat{\theta}^* - z_{1-\alpha/2}\sqrt{\hat{V}^*}, \hat{\theta}^* - z_{\alpha/2}\sqrt{\hat{V}^*}],$$

  where z_β is the β-quantile of the standard normal distribution. This interval is based on the approximation of $(\hat{\theta}^* - \theta)/\sqrt{\hat{V}^*}$ by a standard normal distribution. The intervals were constructed from the $D \times B$ values of θ_i^*.

The PRB gives a measure of the bias of the proposed estimators. The confidence intervals and the coverage probability allows us to evaluate the capacity of the proposed estimators to provide a valid inference.

The simulated scenarios, parameters and estimators are summarized in Table 4.1.

Table 4.1: Simulated scenarios, population parameters, and estimators

Scenarios 1	N=500	$Cor(Y, X_1) = 0.73$	$Cor(Y, X_2) = 0.78$
Scenarios 2	N=1 000	$Cor(Y, X_1) = 0.75$	$Cor(Y, X_2) = 0.78$
Parameters	$\overline{Y}_N = \sum_{i=1}^{N} y_i/N$		
$\pi(x)$ estimator		$\hat{\pi}(x) = \dfrac{exp(\hat{\beta}_0 + \hat{\beta}_1 x1 + \hat{\beta}_2 x2)}{1 + exp(\hat{\beta}_0 + \hat{\beta}_1 x1 + \hat{\beta}_2 x2)}$	
Estimators		$\hat{T}_{pHT} = \dfrac{1}{N}\sum_{i=1}^{N} \dfrac{\delta_i}{\hat{\pi}(x_i)} Y_i$	$\hat{T}_{pH} = \dfrac{\dfrac{1}{N}\sum_{i=1}^{N} \dfrac{\delta_i}{\hat{\pi}(x_i)} Y_i}{\dfrac{1}{N}\sum_{i=1}^{N} \dfrac{\delta_i}{\hat{\pi}(x_i)}}$

Tables 4.2 and 4.3 present the numerical performance of the proposed estimators. Table 4.2 summarizes the simulation results with respect to both pseudo Horvitz-Thompson and pseudo Hájek estimation of population mean for $N = 500$. Similarly, the results for $N = 1\,000$ are presented in Table 4.3.

The results in Tables 4.2 show a slightly better performance of the pseudo Hájek estimator than the pseudo Horvitz-Thompson estimator. PRB is quite small for both of them, meaning that they are slightly biased with respect to the true population parameter. However, the bias of the pseudo Hájek estimator is smaller than that of the pseudo Horvitz-Thompson estimator (1.22 vs 2.77). In terms of variance and confidence intervals, there is no appreciable difference between the two estimators. The variance is approximately zero for both of them, while the length of the estimated confidence intervals is smaller for the pseudo Hájek estimator, indicating more precise estimates. Concerning the coverage probability, it exceeds the nominal level for the pseudo Hájek estimator (0.97), whereas it is lower than the nominal level for the pseudo Horvitz-Thompson estimator (0.86). This occurs when confidence intervals are calculated using the normal approximation. If the percentile method is used, the coverage probability equals 1 for both estimators (as shown in brackets).

The results in Table 4.3 confirm the properties of the proposed estimators. We are now considering a larger population size ($N = 1\,000$), therefore larger samples size. PRB is slightly higher for both the estimators with respect to the previous trial (2.95 and 1.34 for $\hat{T}_{HT}$ and $\hat{T}_H$, respectively). However, also in this case bias is lower for the pseudo Hájek estimator (1.34 vs 2.95). Coverage probability seems to get worse for the pseudo Horvitz-Thompson estimator when normal approximation is used (0.84), whereas it is quite high for the pseudo Hájek estimator (0.97). It still equals 1 for both estimators when percentile method is used. Variance estimation can be considered equals zero for both the pseudo estimators and the confidence intervals are more precise than the previous ones.

Table 4.2: Boostrap results: N=500, B=2 000, D=1 000

	Pseudo Horvitz-Thompson	Pseudo Hájek
Population mean	−0.084 470	−0.084 470
Population mean estimate	−0.086 813	−0.085 499
Variance	0.001 975	0.001 839
Confidence intervals 95%:		
Lower bound	−0.116 648	−0.113 464
Upper bound	0.000 326	0.001 052
Percentage Relative Bias	2.77	1.22
Coverage probability	0.8625(1)	0.9825(1)

Table 4.3: Boostrap results: N=1 000, B=4 000, D=2 000

	Pseudo Horvitz-Thompson	Pseudo Hájek
Population mean	−0.044 075	−0.044 075
Population mean estimate	−0.045 376	−0.044 667
Variance	0.000 973	0.000 939
Confidence intervals 95%:		
Lower bound	−0.068 277	−0.067 030
Upper bound	0.019 392	0.020 492
Percentage Relative Bias	2.95	1.34
Coverage probability	0.839(1)	0.974(1)

Conclusions

Nonprobability samples, such as those from opt-in web surveys, are getting more and more attention, since they are less expensive, quicker and get easily access to a large number of respondents. Nevertheless, they are affected by under-coverage and self-selection, which may lead to unreliable estimates. In addition, inclusion probabilities are unknown, which means that the sample mean is not an unbiased estimator of the population mean. For this reason, our main interest was in overcoming the problem of self-selection. It is worth noting that self-selection also occurs in traditional web surveys.

The first major result of this study was to obtain a model for the process that is supposed to have caused the self-selection. Therefore, on the basis of the specified model obtain an estimator of inclusion probabilities. For this purpose, the estimator of the propensity score by Hirano et al. (2003) was chosen. This choice was made based on the good properties of this estimator. Once defined the theoretical framework and the inclusion probability estimator, two estimators of the population mean were proposed: the pseudo Horvitz-Thompson estimator and the pseudo Hájek estimator, both with estimated inclusion probabilities.

The second major result was to study the large sample properties of the proposed estimators ($\hat{T}_{pHT}$ and $\hat{T}_{pH}$). It was shown that both of them are asymptotically unbiased and the asymptotic variance was derived.

In order to verify the validity of the proposed methodology a simulation study was carried out. The simulation results revealed that both of the proposed estimators can be considered efficient. They also have reasonable bias, but the PRB is lower for the pseudo Hájek estimator. The coverage probability is quite high for $\hat{T}_{pH}$ when the normal approximation is used, whereas it decreases as N increases for the pseudo Horvitz-Thompson estimator. The 95% confidence intervals improve as N and n increase in size, but they are more precise for the pseudo Hájek estimator. As a result, we conclude that the pseudo Hájek estimator is preferable over the pseudo Horvitz-Thompson estimator.

We conclude by discussing further areas of research. There are a few interesting issues that should be addressed. The first concerns the method used to estimate inclusion probabilities. It was stated that Hirano et al. (2003) use series estimators, which requires choosing the number of terms in the series (smoothing parameter). So the question is: How to choose this number in order to achieve the efficiency bound? This is especially true in real situations. The second is to determine the number of simulation runs which are needed to ensure that the properties of the proposed estimator are stable over different sets of simulations. This would enable to arrive at a firm conclusion about the behaviour of the proposed estimators over all possible samples in a population. Given that simulations are the main approach to study the performance of estimators, it is important that a sufficient number of simulations are used to ensure the analysis is reliable. It would also be interesting to investigate the effect of different pseudo-population bootstrap method on the proposed estimators as well as the effect of different resampling designs. Antal and Tillé (2011) argued that if the aim is variance estimation, the resampling design must be radically different from that which generates the original data. Moreover, an application to real data would be needed in order to evaluate the validity of the proposed methodology also in real situations.

Bibliography

Abadie, A. and Imbens, G. (2002). Simple and bias-corrected matching estimators for average treatment effects. NBER Technical Working Paper.

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.

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M. P., Couper, M. P., Dever, J. A., Gile, K., and Tourangeau, R. (2013). Report of the AAPOR task force on non-probability sampling. Technical report, The American Association for Public Opinion Research, Deerfield, IL.

Barbiero, A. and Mecatti, F. (2009). Bootstrap algorithms for variance estimation in complex survey sampling. In *S. Co. 2009 Sixth Conference on Complex Data Modeling and Computationally Intensive Statistical Methods for Estimation and Prediction*, pages 55–60. Maggioli Editore.

Basu, D. (1971). An essay on the logical foundations of survey sampling. Part 1. In *Foundations of Statistical Inference*, pages 203–233. Godambe, V.P., Sprott, D.A. eds., Toronto: Holt, Rinehart & Winston.

Bethlehem, J. (2010). Selection bias in web surveys. *International Statistical Review*, 78(2):161–188.

Chao, M. T. and Lo, S. H. (1994). Maximum likelihood summary and the bootstrap method in structured finite populations. *Statistica Sinica*, 4:389–406.

Chauvet, G. (2007). *Méthodes de bootstrap en population finie*. PhD thesis, Université de Rennes 2.

Conti, P. L., Marella, D., Mecatti, F., and Andrei, F. (2017). A unified principled framework for resampling based on pseudo-populations: asymptotic theory. Also available as `https://arxiv.org/pdf/1705.03827.pdf`.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7(1):1–26.

Efron, B. (1981). Censored data and the bootstrap. *Journal of the American Statistical Association*, 76(374):312–319.

Geman, S. and Hwang, C. R. (1982). Nonparametric Maximum Likelihood Estimation by the Method of Sieves. *Annals of Statistics*, 10(2):401–414.

Gross, S. (1980). Median estimation in sample surveys. In *Proceedings of the Section on Survey Research Methods*, pages 181–184. American Statistical Association.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66(2):315–331.

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

Hájek, J. (1971). Comment on: An essay on the logical foundations of survey sampling by Basu. Part 1. In *Foundations of Statistical Inference*. Godambe, V.P., Sprott, D.A. eds., Toronto: Holt, Rinehart & Winston.

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

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

Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189.

Holmberg, A. (1998). A bootstrap approach to probability proportional-to-size sampling. In *Proceedings of the Section on Survey Research Methods*, pages 378–383. American Statistical Association.

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

Kang, J. D. and Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4):523–539.

Mashreghi, Z., Haziza, D., and Léger, C. (2016). A survey of bootstrap methods in finite population sampling. *Statistics Surveys*, 10:1–52.

McCarthy, P. J. and Snowden, C. B. (1985). The bootstrap and finite population sampling. In *Vital and Public Health Statistics*, Public Health Service Publication, Series 2, number 95. Washington, U.S. Government Printing Office.

Neyman, J. (1934). On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society*, 97(4):558–625.

Pfeffermann, D. and Sverchkov, M. (2004). Prediction of finite population totals based on the sample distribution. *Survey Methodology*, 30:79–92.

Quatember, B. (2015). *Pseudo-Populations. A Basic Concept in Statistical Surveys.* Springer.

Ranalli, M. G. and Mecatti, F. (2012). Comparing recent approaches for bootstrapping sample survey data: a first step towards a unified approach. In *Proceedings of the Section on Survey Research Methods*, pages 4088–4099. American Statistical Association.

Rao, J. N. K. and Wu, C. F. J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83(401):231–241.

River, D. (2007). Sampling for web surveys. In *Proceedings of Joint Statistical Meetings, section on Survey Research Methods*. American Statistical Association.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995). Semiparametric Efficiency in Multivariate Regression Models with Missing Data. *Journal of the American Statistical Association*, 90(429):106–121.

Rosenbaum, P. (1987). The Role of a Second Control Group in an Observational Study: Rejoinder. *Statistical Science*, 2(3):313–316.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rubin, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, 74:318–328.

Rubin, D. B. and Thomas, N. (1992). Affinely Invariant Matching Methods with Ellipsoidal Distributions. *The Annals of Statistics*, 20(2):1079–1093.

Särndal, C. E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York.

Särndal, C. E., Thomsen, I., Hoem, J. M., Lindley, D. V., Barndorff-Nielsen, O., and Dalenius, T. (1978). Design-based and model-based inference in survey sampling (with discussion and reply). *Scandinavian Journal of Statistics*, 5(1):27–52.

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.

Sitter, R. R. (1992). A resampling procedure for complex survey data. *Journal of the American Statistical Association*, 87(419):755–765.

Valliant, R. and Dever, J. A. (2011). Estimating propensity adjustments for volunteer web surveys. *Sociological Methods and Research*, 40:105–137.

Valliant, R., Dever, J. A., and Kreuter, F. (2018). *Practical Tools for Designing and Weighting Survey Samples*. Springer, New York, 2nd edition.

Yates, F. and Grundy, P. M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society*, 15(2):253–261.

R code used for simulation

This annex contains the complete R code for performing the simulation.

```

library(MASS)
set.seed(12345)

# set N=500
N<-500
###################################################
# Generating population values: Y, X1, X2
###############################################

# define the mean vector
mu = c(0,1,3)

# define the variance-covariance matrix
Sigma = matrix(c(1, 0.74, 0.9, 0.74, 1, 0.2, 0.9, 0.2, 1.3), nrow =
    3, ncol = 3)

# generate Y, X1 and X2 from a multivariate normal distribution
   with
# mean vector 'mu' and variance-covariance matrix 'sigma'

df = data.frame(mvrnorm(N, mu, Sigma))

# rename the variables
names(df)<-c("Y", "X1", "X2")
str(df)

# scatterplot of Y, X1 and X2
plot(df, pch = 20, cex = 0.5)

# correlation between variables
mycorr.data=cor(df)
mycorr.data
```

```

#####################################################
# Population parameters
##################################################

# compute the population mean
mean.y<- mean(df$Y)
mean.y

# Create a function for variance
var_pop <- function(x) {
  mean((x - mean(x))^2)
}

#####################################################
# Generating the inclusion probabilities by
# logistic model
##################################################

# set the coefficients of the logit model
z <- -0.8 + 0.2*df$X1 + 0.3*df$X2

# initialize the variable 'pr'
df$pr<-c(1:N)

# compute the inclusion probabilities
df$pr <- exp(z)/(1+exp(z))

# generate N i.i.d. Bernoulli random variables on the basis of 'pr'
df$Be<-rbinom(N, size=1, p=df$pr)

####################################################
# Fitting logistic regression on data produced
##############################################
fit <- glm(Be ~ X1+X2, data = df, family = binomial(logit))
summary(fit) # results
coef(fit)    # estimated coefficients

# initialize a new variable 'prob'
df$prob<-c(1:N)

# compute the estimate of inclusion probabilities
df$prob<-predict(fit, type="response")

# expected sample size
ns <- sum(df$prob)
```

```

###################################
# Setup of bootstrap parameters
##############################

npop<-2*N # number of pseudo-populations
nboot <-4*N #number of bootstrap samples

boot.ht<-matrix(data = NA, nrow = nboot, ncol = npop, byrow = FALSE
    ,
                dimnames = NULL)
boot.hj<-matrix(data = NA, nrow = nboot, ncol = npop, byrow = FALSE
    ,
                dimnames = NULL)
boot.theta.star.ht <-NULL
boot.theta.star.hj <-NULL
boot.var.ht<-NULL
boot.var.hj<-NULL

mean.ht.camp<-NULL
mean.hj.camp<-NULL

# create a function for pseudo Horvitz-Thompson estimator
mean.ht<- function (x, p) {
  sum(x/p)/N
}
# create a function for pseudo Hajek pseudo estimator
mean.hj<- function (x, p) {
  sum(x/p)/sum(1/p)
}

##############################################################
# Draw a sample and from it generate one pseudo population
# Take B bootstrap samples
# Repeat this process D times
#########################################################

for(j in 1:npop){
# generate N i.i.d. Bernoulli random variables
df$ber<-rbinom(N,size=1,df$prob)

# selects units with 'ber=1', i.e. select the sample s
camp<-subset(df, ber == 1)

#####################################################
# Generating the pseudo population from the sample
```

```
##################################################

####################################################
# Construct the fix part of the pseudo population
#################################################
ncamp<-data.frame(camp$Y, camp$prob)
ncamp$int<-floor(1/camp$prob)
ncamp$rest<-(1/camp$prob)-ncamp$int

nfixpop<-ncamp[rep(rownames(ncamp), ncamp$int),]

# renumber the rows
rownames(nfixpop)<-1:NROW(nfixpop)

# draw and rename the first two variables
fixpop <- nfixpop[c(1:2)]
names(fixpop)<-c("ps.Y", "ps.prob")

############################################################
# Completing the remaining part of the pseudo population
# adopting Poisson sampling with probabilities included
# in 'ncamp$rest'
#######################################################

ncamp$b<-rbinom(sb,size=1,ncamp$rest)
# select units with bernoulli variable = 1
nrestpop<-subset(ncamp, ncamp$b==1)

# renumber the rows
rownames(nrestpop)<-1:NROW(nrestpop)

# draw and rename the first two variables
restpop<- nrestpop[c(1:2)]
names(restpop)<-c("ps.Y", "ps.prob")

# obtain the pseudo-population
pspop<-rbind(fixpop, restpop)
# number of units
NR<-nrow(pspop)
NR

######################################################
# Take B bootstrap samples s* from U* ('pspop') by
# using the same sampling design that led to s and
# for each of them computing the estimates
#################################################
```

```
for(i in 1:nboot){
  pspop$ber<-rbinom(NR,size=1,pspop$ps.prob)
  bcamp<-subset(pspop, pspop$ber==1)
  boot.ht[i,j]<-mean.ht(bcamp$ps.Y, bcamp$ps.prob)
  boot.hj[i,j]<-mean.hj(bcamp$ps.Y, bcamp$ps.prob)
} # close loop for i
} # close loop for j

install.packages("matrixStats")
library(matrixStats)

# compute theta(s) star
boot.theta.star.ht<-colMeans2(boot.ht)
boot.theta.star.hj<-colMeans2(boot.hj)

# compute theta star estimation
theta.star.est.ht<-mean(boot.theta.star.ht)
theta.star.est.hj<-mean(boot.theta.star.hj)

# compute variances (V*)
boot.var.ht<-colVars(boot.ht)
boot.var.hj<-colVars(boot.hj)

# standard deviations
boot.sd.ht<-sqrt(boot.var.ht)
boot.sd.hj<-sqrt(boot.var.hj)

####################################################
# Variance estimation for pseudo HT estimator
###############################################
boot.var.est.ht<-mean(boot.var.ht)
boot.var.est.ht

boot.sd.est.ht<-sqrt(boot.var.est.ht)
boot.sd.est.ht

####################################################
# Variance estimation for pseudo Hajek estimator
###############################################
boot.var.est.hj<-mean(boot.var.hj)
boot.var.est.hj

boot.sd.est.hj<-sqrt(boot.var.est.hj)
boot.sd.est.hj
```

```
##################################
#       BIAS pseudo HT
############################

diff.ht<-NULL
diff.ht1<-NULL

diff.ht1<-(boot.theta.star.ht-mean.y)/mean.y
bias.ht1<-100*(sum(diff.ht1))/npop
bias.ht1

##################################
#       BIAS pseudo Hajek
############################

diff.hj<-NULL
diff.hj1<-NULL

diff.hj1<-(boot.theta.star.hj-mean.y)/mean.y
bias.hj1<-100*(sum(diff.hj1))/npop
bias.hj1

####################################
# Confidence Intervals
################################

# normal intervals
cnor.ht<-c(mean(boot.theta.star.ht)-1.96*sd(boot.theta.star.ht)/
    sqrt(npop),
            mean(boot.theta.star.ht)+1.96*sd(boot.theta.star.ht)/
    sqrt(npop))
cnor.hj<-c(mean(boot.theta.star.hj)-1.96*sd(boot.theta.star.hj)/
    sqrt(npop),
           mean(boot.theta.star.hj)+1.96*sd(boot.theta.star.hj)/
    qrt(npop))

# student intervals
cin.ht<-c(mean(boot.theta.star.ht)-qt(0.975,df=npop-1)*sd(boot.
    theta.star.ht)/sqrt(npop),
          +mean(boot.theta.star.ht)-qt(0.025,df=npop-1)*sd(boot.
    theta.star.ht)/sqrt(npop))
cin.hj<-c(mean(boot.theta.star.hj)-qt(0.975,df=npop-1)*sd(boot.
    theta.star.hj)/sqrt(npop),
          +mean(boot.theta.star.hj)-qt(0.025,df=npop-1)*sd(boot.
    theta.star.hj)/sqrt(npop))

```

```
library(stats)

# CI - quantile method
c.star.ht=sort(boot.theta.star.ht)
c.star.hj=sort(boot.theta.star.hj)

cq.ht.star = c(quantile(c.star.ht, probs=0.025), quantile(c.star.ht, probs=0.975))
cq.hj.star = c(quantile(c.star.hj, probs=0.025), quantile(c.star.hj, probs=0.975))

########################################
# COVERAGE PROBABILITY
###################################
nch<-0
ncj<-0

nch.t<-0
ncj.t<-0

nch.n<-0
ncj.n<-0

for (k in 1:nboot){

  # quantile method
  sort(boot.ht[k,])
  cq.ht<-c(quantile(boot.ht[k,], probs=0.025), quantile(boot.ht[k,], probs=0.975))
  if (cq.ht[1]<=mean.y & cq.ht[2]>=mean.y)  {nch=nch+1}

  sort(boot.hj[k,])
  cq.hj<-c(quantile(boot.hj[k,], probs=0.025), quantile(boot.hj[k,], probs=0.975))
  if (cq.hj[1]<=mean.y & cq.hj[2]>=mean.y)  {ncj=ncj+1}

  # student intervals
  cin.ht.t<-c(mean(boot.ht[k,])-qt(0.975,df=npop-1)*sd(boot.ht[k,])/sqrt(npop),
           + mean(boot.ht[k,])-qt(0.025,df=npop-1)*sd(boot.ht[k,])/sqrt(npop))
  cin.hj.t<-c(mean(boot.hj[k,])-qt(0.975,df=npop-1)*sd(boot.hj[k,])/sqrt(npop),
           + mean(boot.hj[k,])-qt(0.025,df=npop-1)*sd(boot.hj[k,])/sqrt(npop))

```

```
  if (cin.ht.t[1]<=mean.y & cin.ht.t[2]>=mean.y)  {nch.t=nch.t+1}
  if (cin.hj.t[1]<=mean.y & cin.hj.t[2]>=mean.y)  {ncj.t=ncj.t+1}

  # normal intervals
  cnor.ht.n<-c(mean(boot.ht[k,])-1.96*sd(boot.ht[k,])/ sqrt(npop),
               mean(boot.ht[k,])+1.96*sd(boot.ht[k,])/ sqrt(npop))
  cnor.hj.n<-c(mean(boot.hj[k,])-1.96*sd(boot.hj[k,])/ sqrt(npop),
               mean(boot.hj[k,])+1.96*sd(boot.hj[k,])/ sqrt(npop))

  if (cnor.ht.n[1]<=mean.y & cnor.ht.n[2]>=mean.y)  {nch.n=nch.n+1}
  if (cnor.hj.n[1]<=mean.y & cnor.hj.n[2]>=mean.y)  {ncj.n=ncj.n+1}

}

# Calculate the proportion of intervals that cover the parameter

# quantile method
CP.HT.n<-100*nch/nboot
CP.HJ.n<-100*ncj/nboot

# student intervals
CP.HT.nt<-100*nch.t/nboot
CP.HJ.nt<-100*ncj.t/nboot

# normal intervals
CP.HT.nn<-100*nch.n/nboot
CP.HJ.nn<-100*ncj.n/nboot
```