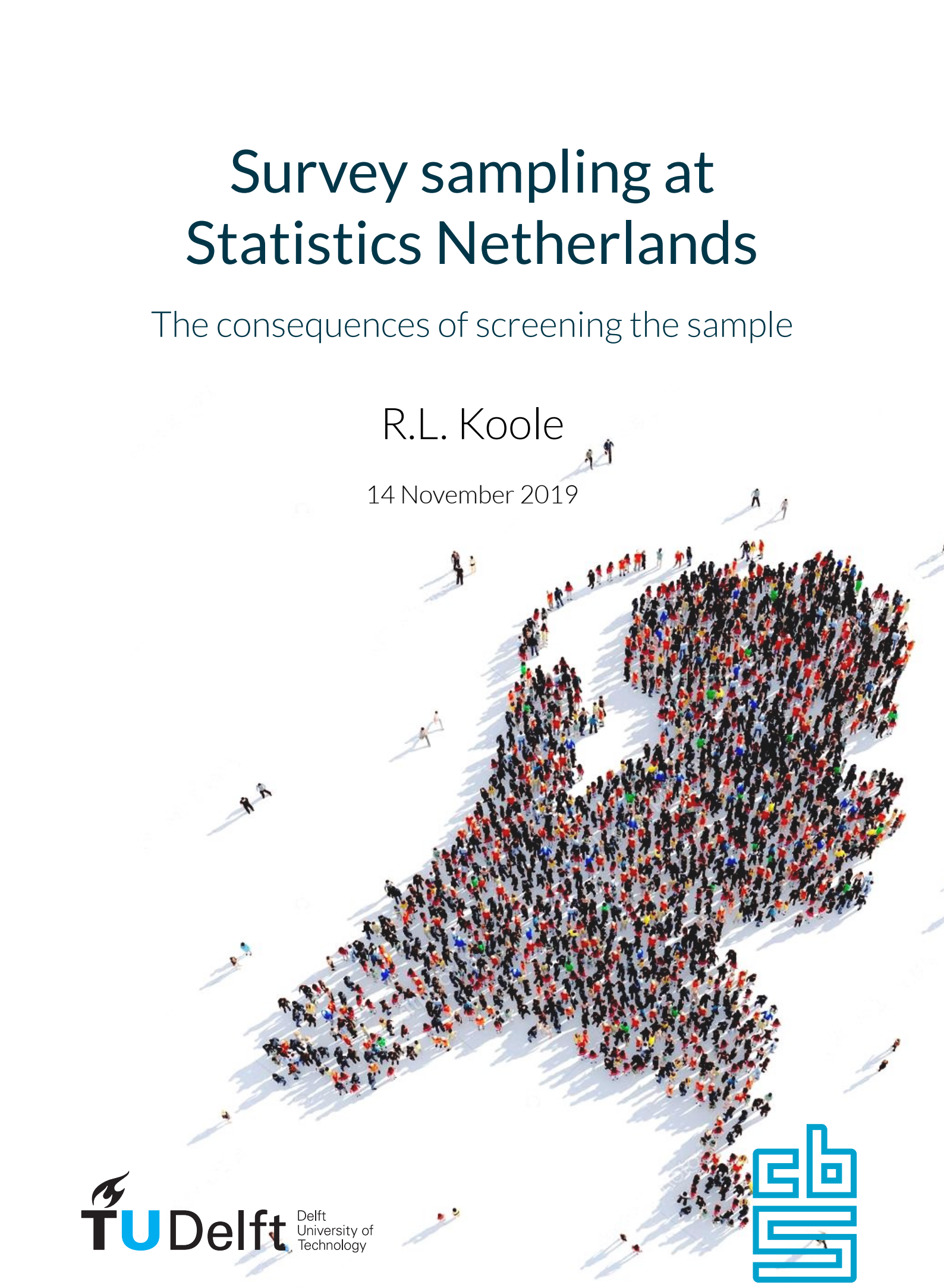
Survey sampling at Statistics Netherlands

The consequences of screening the sample

R.L. Koole

14 November 2019



TU Delft Delft University of Technology

Survey sampling at Statistics Netherlands

The consequences of screening the sample

by

R. L. Koole

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Thursday November 14, 2019 at 14:00.

Student number:	4225392	
Project duration:	February 11, 2019 – November 14, 2019	
Thesis committee:	Dr. H. P. Lopuhaä,	TU Delft, chair & supervisor
	Dr. J. M. Gouweleeuw,	Statistics Netherlands (CBS), supervisor
	Dr. C. Kraaikamp,	TU Delft

An electronic version of this thesis is available at http://repository.tudelft.nl/.

Abstract

Statistics Netherlands performs many different surveys to obtain estimates of unknown characteristics of the Dutch population. To keep the response burden on the Dutch households low, Statistics Netherlands applies a screening procedure to their selected samples. In our research, we investigate the effects of the screening procedure on the survey sampling process. We conclude that the effects of the screening process cannot be considered negligible.

We derive an approximation of the inclusion probability of an element in the sample after screening. This probability is dependent on the number of people on address and the sampling fraction. Consequently, the probability is not equal for all inhabitants and the effects of the screening procedure become larger as sample sizes increase.

Two different statistical tests are developed and applied to existing samples that have recently been selected and screened by Statistics Netherlands, to determine whether the sample after screening is representative for the population (and for the sample before screening) with respect to relevant auxiliary variables.

From a super-population viewpoint, we investigate the properties of the generalised regression estimator. We prove that under modest conditions the generalised regression estimator is consistent and asymptotically unbiased for the self-weighting two-stage sampling design that is used at Statistics Netherlands. When screening is applied, we cannot conclude that the generalised regression estimator is consistent and asymptotically unbiased. We show how the Horvitz-Thompson estimator and the generalised regression estimator can be used to undo the effects of the screening procedure during the estimation of population characteristics.

Keywords: Survey sampling, sampling design, inclusion probability, Horvitz-Thompson estimator, multivariate hypergeometric distribution, parametric bootstrap, generalised regression estimator, super-population model, consistency, asymptotically unbiased.

Preface

Before you lies the thesis *Survey sampling at Statistics Netherlands, the consequences of screening the sample*. In this research we investigated the effects of the screening procedure that is applied to the survey samples at Statistics Netherlands. The report you are reading describes the work that has been conducted over the previous months, from February 2019 until November 2019. It has been written to obtain the degree of Master of Science at Delft University of Technology, within the Applied Mathematics programme.

It has been quite a challenge to figure out all the details of the survey sampling process that are conducted at Statistics Netherlands. After diving into the process and extracting the relevant elements for the screening procedure, I am happy with the result of my research. José, you have been a great help figuring out all the steps in the survey sampling process. I would like to thank my supervisors Rik and José for their guidance, time, critical notes and valuable feedback on my work over the past few months. Furthermore, I want to thank all members of the 'Steekproefgroep' Den Haag and Heerlen for their interest in my research and the enjoyable time as a graduate intern at Statistics Netherlands.

This work also concludes my time as a student at Delft University of Technology. I would like to thank all my friends for the support and the unforgettable time as a student. A special note of gratitude goes to Sven, who has been an incredible support during the previous few months. Last but not least, I want to thank my parents for the endless support you gave me.

I want to thank everyone that is taking the time to read this thesis. I wish you all a pleasant reading.

R. L. Koole
Leidschenveen, November 2019

Contents

1

Introduction

In a society where the amount of information is growing explosively, free access to reliable and integral data is crucial. There is an ever-growing demand for statistical information about the economic, social, political and cultural shape of countries. As the national statistical office, Statistics Netherlands (CBS) provides reliable statistical information and data to produce insight into social issues, thus supporting the public debate, policy development and decision-making while contributing to prosperity, well-being and democracy [1]. Sometimes, this information can be retrieved from existing sources, from for example administrative records, but quite often there is a lack of such sources. A survey is then a powerful instrument to collect new statistical information.

Many surveys are carried by Statistics Netherlands to obtain information about the Dutch population. In order to do this, a sample of inhabitants is selected from all Dutch inhabitants. The inhabitants that are in selected in this sample receive a letter with the request to fill in a questionnaire. The data that is collected from the inhabitants in the sample, is then used to estimate characteristics about the whole Dutch population.

Statistics Netherlands aims to spread their surveys equally among the Dutch inhabitants, such that each inhabitant is selected in a sample of Statistics Netherlands approximately as often as the others. To ensure this, inhabitants have an equal probability to be selected in a sample. Furthermore, it is ensured that inhabitants can only be selected in a sample once per year.

Statistics Netherlands tries to keep the response burden on the Dutch households low. However, it can sometimes occur that two or more inhabitants who live on the same address are approached in a short period of time (or even at the same time). If this occurs, it is assumed that those inhabitants experience this as response burden. This is not desirable and it is expected that this leads to nonresponse.

To prevent this from happening, Statistics Netherlands applies a so-called screening procedure to their samples, which makes some inhabitants in the sample not eligible for participating in the survey. Inhabitants that are not eligible after screening do not receive a letter with the request to participate in the survey. This screening procedure for example ensures that if an inhabitant is selected in the sample, any other inhabitant who lives on the same address cannot be selected in a sample of Statistics

Netherlands in the next twelve months.

A consequence of this screening procedure is that not every inhabitant has the same probability to become non-eligible after screening. For example, an inhabitant who lives on his/her own on his/her address, will never become not eligible from the sample by that reason, whereas an inhabitant who lives with many others on his/her address has a higher probability to become non-eligible after screening. Consequently, not every inhabitant has the same probability to be selected in the sample after screening.

Thus far, it is assumed that the effects of the screening procedure are negligible. In other words, despite applying the screening procedure, it is assumed that if inhabitants have equal probabilities to be selected in the sample, inhabitants also have an equal probability to be selected in the sample after screening.

Over the past few years, sample sizes have been increasing. In the past the amount of inhabitants that become not eligible by the screening procedure was relatively small in comparison to the current situation. Now that sample sizes are increasing, the question arises whether it is still fair to assume that the effects of the screening procedure are negligible.

1.1. Research questions

The aim of this thesis is to investigate the consequences of the screening procedure. We will determine if it is fair to assume that the screening procedure is negligible. To investigate this, we have defined four main research questions:

(**RQ1**) *What is the probability that an inhabitant becomes not eligible after screening the sample? And what is the probability that an inhabitant is selected in the sample after screening?*

(**RQ2**) *Is the sample after screening representative for the population with respect to several auxiliary variables? And is the sample after screening representative for the sample before screening with respect to several auxiliary variables?*

(**RQ3**) *What are the effects of the screening procedure on the procedure of estimation of population characteristics?*

(**RQ4**) *If the screening procedure cannot be considered negligible, how can we undo the effects of the screening procedure?*

To formulate an answer to these questions, we use two approaches throughout this thesis. The first one is by a more theoretical point of view by computing the probabilities that an inhabitant becomes not eligible after screening the sample. Secondly, we can use the samples that have recently been selected and screened by Statistics Netherlands to measure the effects of the screening procedure.

1.2. Data

Throughout this thesis we make use of data that is made available by Statistics Netherlands. Since the Dutch population is continuously changing, we make use of the Dutch population at different specified time stamps (one per month). The data that is used is pseudonymised, such that inhabitants are not traceable. For example, no names, addresses or social security numbers (BSN) are included in the data. However, additional information for these inhabitants such as gender, marital status and age are included in the data.

Furthermore we use data of several surveys samples that are performed by Statistics Netherlands recently. For each sample, it is known which inhabitants are included in the sample. Throughout this thesis we often focus on the mobility survey.

1.2.1. Mobility survey

Statistics Netherlands performs different surveys to collect all types of data about the Dutch population. Since 1978 Statistics Netherlands investigates the mobility of Dutch inhabitants using the mobility survey. In the past this was done by surveys which were called *Onderzoek Verplaatsingsgedrag (OVG)* or *Onderzoek Verplaatsingen Nederland (OViN)* [2]. Since January 2018 the mobility survey continued under the name *Onderweg in Nederland (ODiN)*. The goal of the mobility survey is to deliver useful information about the mobility of Dutch inhabitants on a daily basis to the Ministry of Infrastructure and Water Management and others [2].

The Dutch inhabitants that are participating in the survey are requested to provide for one specific day where he/she traveled to, with what purpose, how long the traveling took, and what means of transportation were used. Additionally, there are questions that relate to the possession of (electrical) bicycles or cars and the average use of different means of transport [2].

We have chosen to focus on this survey because its size is relatively large. The aim is that 45.000 Dutch inhabitants participate in the survey on a yearly basis [2]. The target population for the mobility survey contains all Dutch inhabitants that are six years or older.

1.3. Thesis outline

This thesis is structured as follows. In Chapter 2 we introduce basic (mathematical) concepts on survey sampling, which are used in the further chapters. In Chapter 3 the procedure for selecting a sample at Statistics Netherlands is described. Subsequently, the procedure for screening the sample is described in detail in Chapter 4. In this chapter, an approximation for the probability that an inhabitant becomes not eligible by the screening procedure is derived. In Chapter 5 we discuss two different statistical tests that are applied to a recent sample, which allows us to compare the distributions of auxiliary variables in the population, sample before screening and the sample after screening. To investigate the influence of the sample size on the screening procedure, statistical tests are applied to a hypothetical situation where sample sizes are larger than in the current situation. Chapter 7 is dedicated to estimation of population characteristics from the collected data of inhabitants in the sample. We investigate the properties of the estimator that is used and the influences of the screening procedure on this estimator. In Chapter 8 we discuss the possibilities to correct for the screening procedure in the estimation procedure. We complete this thesis by providing a summarising conclusion and discussion on our work in Chapter 9, and our view on future research on the matter in Chapter 10.

2

Survey sampling

Carrying out a survey is a complex process that requires careful consideration and decision making. The main idea of a survey is to collect data for a part of the population and use that data to obtain an estimate for some characteristic of the total population.

This section gives an overview of the various steps in the process and basic (mathematical) concepts on survey sampling. Figure 2.1 shows the different steps in the survey process. First, a sample of the population is selected (step 1). Then data is collected for that part of the population (step 2), followed by estimating characteristics of the whole population (step 3). Finally, the obtained estimates can be published (step 4).

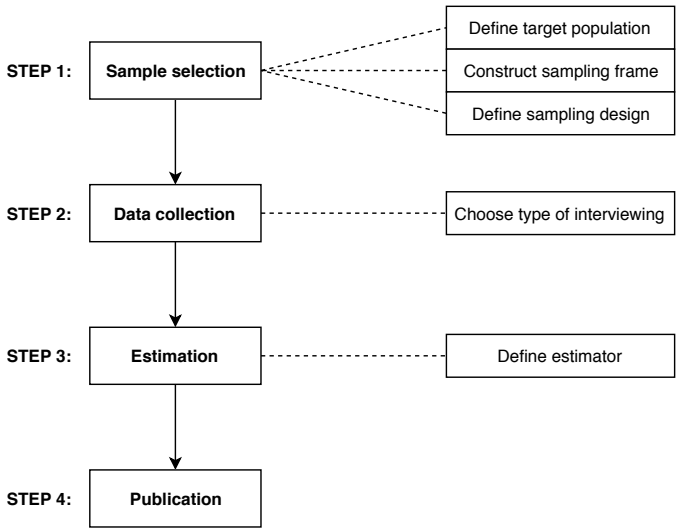


Figure 2.1: Graphical representation of the survey sampling process

2.1. Sample selection

The main goal of survey sampling is to obtain an estimate for a parameter of the population that is not yet known. The first step in the survey process is the step where elements are selected from a

population. Before any elements can be selected, it needs to be clear which population will exactly be investigated. For example, for the mobility survey only the Dutch inhabitants that are six years or older are considered.

Definition 2.1. The *target population* U is a finite set

$$U = \{1, \dots, k, \dots, N\} \tag{2.1}$$

of N elements. Here the quantity N is called the population size, which is usually known. The numbers $1, 2, \dots, N$ denote the sequence numbers of the elements in the target population [3]. The k-th element of the population is represented by its label k and we usually refer to it as element k. Sometimes we may refer to element k as inhabitant or person.

It is important to define the target population properly. For example, for an unemployment survey one should define if inhabitants below or above a certain age should or should not be included in the target population. The next step is to define the target variables that need to be measured by the questions in the survey.

Definition 2.2. A *target variable* y represents an unknown characteristic of the elements in the target population. Let y_k represent the unknown characteristic for the k-th element and let $\mathbf{y} = (y_1, \dots, y_N)$ denote the vector of characteristics of the elements in the population U [4].

Two important population parameters that are usually estimated with the use of a survey are the *population total* of the target variable y

$$t_y = \sum_{k=1}^{N} y_k \tag{2.2}$$

or the *population mean* of y

$$\overline{y}_U = \frac{1}{N} \sum_{k=1}^{N} y_k \tag{2.3}$$

Observing y for all of elements of U would usually prove too expensive or impractical [4], so to estimate the parameters we observe y_k for a subset of U. This limited set of y_k can then be used to calculate estimates of t_y and $\overline{y}_U$.

Definition 2.3. A subset of the population U is called a *sample* of U. A sample is usually denoted by s and the set of possible samples is denoted by $\mathcal{S}$[4].

While discussing a sample, one often makes use of the vector of indicators

$$\mathbf{I} = (I_1, \dots, I_N) \tag{2.4}$$

where the value of the indicator I_k (for $k = 1, 2, \dots, N$) is equal to 1 if element k is included in sample s and 0 otherwise [4]. Note that the indicator $I_k = I_k(s)$ is a function of the sample s. These indicators are sometimes referred to as the *sample membership indicators*.

A sample should be selected such that it allows drawing conclusions for the population as a whole. As Bethlehem [3] describes, a requirement for this is the availability of a *sampling frame*, which is a list of elements in the population. For every element in the target population, there must be information on how to contact that element. The sampling frame should be an accurate representation of the

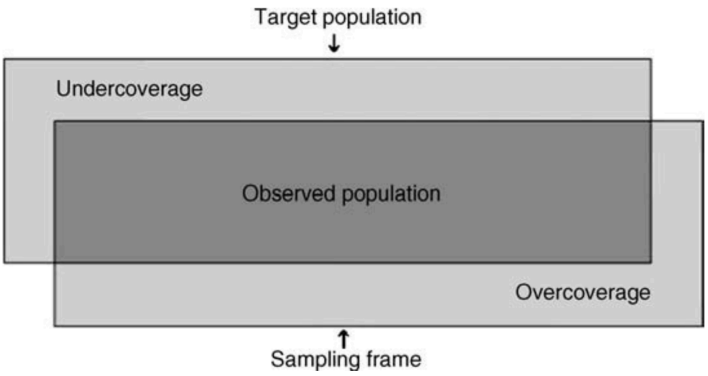


Figure 2.2: Figure from *Applied Survey Methods, a Statistical Perspective* by Bethlehem [3].

population, or else there is a risk of drawing wrong conclusions from the survey [3]. Figure 2.2 shows two situations that can cause problems: *undercoverage* and *overcoverage*.

Undercoverage occurs if the target population contains elements that are not included in the sampling frame, which means that those elements can never be selected in the sample. In the case of overcoverage, the sampling frame contains elements that are not part of the target population.

For selecting a sample from the total population in the Netherlands, a population register is available. Each municipality in the Netherlands maintains its register of their inhabitants and all municipal information is combined in one register that is used by Statistics Netherlands for its surveys. This register is known as the Personal Records Database[1]. Because time passes by between selecting the sample and observing the inhabitants, a little under- and overcoverage occurs, because inhabitants may be born, pass away, emigrate or immigrate in the meantime. However we assume that the under- and overcoverage is negligible and a sampling frame covering the whole population is available for selecting samples.

Another requirement is to select the sample by means of a probability sample, where every element in the target population has a nonzero probability of being selected in the sample [3]. In general, sample selection is carried out by a series of randomised experiments, which will result in a selected sample s.

Given the procedure to select a sample from the population U, it is possible to compute the probability of selecting a specific sample s. We will assume that there exists a function $p(\cdot)$ that assigns a probability to each possible sample.

Definition 2.4. A *sampling design* $p(\cdot)$ assigns a probability $p(s)$ to every possible sample $s \in \mathcal{S}$ from population U, such that [4]

(i) $0 \leq p(s) \leq 1$

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

For a given sampling design, we can regard any sample s as the outcome of a set-valued random variable S, whose probability distribution is given by $p(\cdot)$, i.e.

$$\mathbb{P}(S = s) = p(s) \tag{2.5}$$

[1] In Dutch better known as Basisregistratie Personen (BRP).

Definition 2.5. The *sample size* n_s of a sample s from a population U is defined as the number of elements in the sample s. The sample size can be computed from s by

$$n_s = \sum_{k=1}^{N} I_k(s) \tag{2.6}$$

Note that n_s is not necessarily the same for all samples $s \in \mathcal{S}$. However, in many cases the sampling design defines the sample size to be fixed. The fixed sample size means that every sample s with positive probability $p(s)$ has the same sample size n_s. In this case we write n for the sample size. The quantity n/N is known as the *sampling fraction* and is denoted by the letter f.

Note that for some sampling designs elements can be selected more than once in the sample. For such a sampling design, n_s in Equation (2.13) denotes the number of unique elements in the sample.

If a sampling design has been defined, it is possible to compute the probability π_k that element k is included in the sample.

Definition 2.6. The *first-order inclusion probability* π_k of element k is the probability that element k is included in a sample. It can be obtained by a given sampling design $p(\cdot)$ by [4]

$$\pi_k = \mathbb{P}(k \in S) = \mathbb{P}(I_k = 1) = \sum_{s \ni k} p(s) \tag{2.7}$$

Here $s \ni k$ denotes that the sum is over those samples s that contain the given element k.

In the same way, the probability that both elements k and l are included can be obtained from $p(\cdot)$.

Definition 2.7. The probability that both elements k and l are included in the sample is known as the *second-order inclusion probability*, and it is obtained by

$$\pi_{kl} = \mathbb{P}(k, l \in S) = \mathbb{P}(I_k I_l = 1) = \sum_{s \ni k,l} p(s) \tag{2.8}$$

We have that $\pi_{kl} = \pi_{lk}$ for all $k, l \in U$. Note that in case $k = l$

$$\pi_{kk} = \mathbb{P}\left(I_k^2 = 1\right) = \mathbb{P}(I_k = 1) = \pi_k \tag{2.9}$$

The inclusion probabilities can be used to obtain useful properties of the sample membership indicators of Equation (2.4). For an arbitrary sampling design $p(\cdot)$ it holds:

(i) The expected value of the indicator I_k is equal to the inclusion probability of element k $(k = 1, \ldots, N)$, i.e.

$$\mathbb{E}(I_k) = \mathbb{P}(I_k = 1) = \pi_k \tag{2.10}$$

(ii) The variance of the indicator for an arbitrary sampling design $p(\cdot)$ can be expressed as

$$\begin{aligned} \mathbb{V}(I_k) &= \mathbb{E}\left(I_k^2\right) - \mathbb{E}(I_k)^2 \\ &= \mathbb{E}(I_k) - \mathbb{E}(I_k)^2 \\ &= \pi_k - \pi_k^2 \\ &= \pi_k(1 - \pi_k) \end{aligned} \tag{2.11}$$

(iii) For an arbitrary sampling design and two elements $k, l = 1, \ldots N$ the covariance of the indicators is

$$\mathrm{Cov}(I_k, I_l) = \mathbb{E}(I_k I_l) - \mathbb{E}(I_k)\,\mathbb{E}(I_l) = \pi_{kl} - \pi_k \pi_l \tag{2.12}$$

(iv) The sample size is equal to the sum of the sample membership indicators

$$n_s = \sum_{k=1}^{N} I_k(s) \tag{2.13}$$

and the expected value of the sample size is equal to the sum of the inclusion probabilities for all elements

$$\mathbb{E}(n_s) = \mathbb{E}\left(\sum_{k=1}^{N} I_k(s)\right) = \sum_{k=1}^{N} \mathbb{E}(I_k(s)) = \sum_{k=1}^{N} \pi_k \tag{2.14}$$

If all elements in a population have the same inclusion probability, then such a sample is called a *self-weighting* sample.

Definition 2.8. If all elements in the population have the same probability of being selected in the sample, the sample is called *self-weighting*. A sampling design can be chosen such that the obtained sample becomes self-weighting. In that case, the sampling design is also called self-weighting. [3]

We will now introduce two self-weighting sampling designs: *simple random sampling without replacement* and *systematic sampling*. For both designs, we describe the design and compute the first- and second-order inclusion probabilities. The sampling design that is used by Statistics Netherlands is build on these two designs and is introduced in Section 3.

2.1.1. Simple random sampling without replacement (SRSWR)

Simple random sampling without replacement (SRSWR) is a widely used sampling design. The SRSWR design is a fixed size sampling design, which means that every sample that can be selected contains exactly n distinct elements of the population. Then the number of possible samples is

$$\binom{N}{n} = \frac{N!}{n!(N-n)!} \tag{2.15}$$

Simple random sampling without replacement assigns equal probabilities to each possible sample [5]. Sampling without replacement, means that this is a way of sampling in which each element can appear at most once in a sample.

Sampling design We have already seen that the possible number of samples with size n is $\binom{N}{n}$. Any sample s with size n then has the probability of $1/\binom{N}{n}$ to be selected, so the sampling design for SRSWR is

$$p(s) = \begin{cases} \frac{1}{\binom{N}{n}} & \text{if } s \text{ has } n \text{ elements} \\ 0 & \text{otherwise} \end{cases} \tag{2.16}$$

First-order inclusion probability Let k be an element from population $U = \{1, \ldots, N\}$. There are exactly $\binom{N-1}{n-1}$ samples s that contain element k. We have seen that every possible sample s has probability $1/\binom{N}{n}$ to be selected. Combining these results gives that the first-order inclusion probability

equals

$$\begin{aligned}\pi_k &= \sum_{s \ni k} p(s) \\ &= \binom{N-1}{n-1} \cdot \frac{1}{\binom{N}{n}} \\ &= \frac{(N-1)!}{(n-1)!(N-n)!} \cdot \frac{n!(N-n)!}{N!} \\ &= \frac{(N-1)!}{N!} \cdot \frac{n!}{(n-1)!} \\ &= \frac{n}{N}\end{aligned} \tag{2.17}$$

Second-order inclusion probability Let k and l be two elements from population $U = \{1, \ldots, N\}$, such that $k \neq l$. Then there are $\binom{N-2}{n-2}$ possible samples s that include both elements k and l. The second-order inclusion probability is

$$\begin{aligned}\pi_{kl} &= \sum_{s \ni k,l} p(s) \\ &= \binom{N-2}{n-2} \cdot \frac{1}{\binom{N}{n}} \\ &= \frac{(N-2)!}{(n-2)!(N-n)!} \cdot \frac{n!(N-n)!}{N!} \\ &= \frac{(N-2)!}{N!} \cdot \frac{n!}{(n-2)!} \\ &= \frac{n(n-1)}{N(N-1)}\end{aligned} \tag{2.18}$$

If $k = l$, we have $\pi_{kk} = \pi_k = \frac{n}{N}$.

2.1.2. Systematic sampling

Systematic sampling offers several practical advantages, particularly because of its simplicity of execution [3]. For systematic sampling in its most basic form, assume that the population size N is a multiple of the sample size n. Systematic sampling starts by dividing the population into $F = N/n$ equal parts. F is a positive integer and is known as the *step length* [3] or *sampling interval* [4]. The first element of the sample is determined by randomly drawing element $b \in \{1, F\}$. Then the other elements are determined by systematically taking every F-th element thereafter [4]. The element b is known as the *random start* and can only assume F different values. Consequently, there are only F different samples possible.

Sampling design Let $\mathcal{S}_{sys}$ denote the set of all possible samples that can be selected by systematic sampling. There are only F different and non-overlapping samples in $\mathcal{S}_{sys}$. Consequently, the sampling design is given by

$$p(s) = \begin{cases} \frac{1}{F} & \text{if } s \in \mathcal{S}_{sys} \\ 0 & \text{otherwise} \end{cases} \tag{2.19}$$

First-order inclusion probability We have seen that there are exactly F different and non-overlapping samples possible. Each element k is included in exactly one of those elements. Consequently, the

first-order inclusion probability of element k is

$$\pi_k = \mathbb{P}(k \in s) = \frac{1}{F} = \frac{n}{N} \tag{2.20}$$

Note that the first-order inclusion probabilities of systematic sampling are identical to the first-order inclusion probabilities of simple random sampling without replacement.

Second-order inclusion probability The second-order inclusion probabilities for systematic sampling are not equal to the second-order inclusion probabilities for SRSWR. For systematic sampling, elements can only be selected in the same sample if they are both included in the sample $s \in \mathcal{S}_{sys}$. This means that two inhabitants can only be selected in the sample together, if the distance between them is exactly one step length. The second-order inclusion probability is therefore dependent on the order of the inhabitants. Consequently, there is no explicit formula for the second-order inclusion probability [3].

Systematic sampling is also possible without the assumption that N is a multiple of n. Again let the step length be defined as $F = \frac{N}{n}$. Note that F is not necessarily integer, but it is real-valued. The population U of size N can be represented by the interval $(0, N]$, which can be divided into N intervals of length 1:

$$(0,1], (1,2], \ldots, (N-1, N]$$

A *random start* is then defined by selecting a random value b from the interval $(0, F]$ uniformly. Then the values

$$b, b+F, b+2F, \ldots, b+(n-1)F$$

all correspond to one of the intervals of length one. If a value is contained in the interval $(k-1, k]$ then element k is selected in the sample. The procedure that is used to select a sample by systematic sampling is summarised in Algorithm 2.1.

In the next example, we show that if N is not a multiple of n, the possible samples have equal probability to be selected. The elements have an equal inclusion probability of $\frac{1}{F}$.

Example 2.1. Suppose a sample of size $n = 2$ has to be selected from a population of size $N = 5$ with equal probabilities. Then $F = \frac{N}{n} = \frac{5}{2}$. There are five possible samples, that all have probability $\frac{1}{5}$ to be selected. These samples are $s_1 = \{1,3\}, s_2 = \{1,4\}, s_3 = \{2,4\}, s_4 = \{2,5\}$ and $s_5 = \{3,5\}$. A graphical representation of this example is given in Figure 2.3. We can now easily compute the probability that sample s_2 is selected by

$$\mathbb{P}(S = s_2) = \mathbb{P}\left(b \in \left(\frac{1}{2}, 1\right]\right) = \frac{1}{2} \cdot \frac{1}{F} = \frac{1}{5} \tag{2.21}$$

Algorithm 2.1: Procedure that is used to select a sample according to a systematic sampling design.

Systematic sampling

1. Initialisation:
   - N population size
   - n sample size
2. Compute $F = \frac{N}{n}$
3. Select a random number $b \in (0, F]$ uniformly
4. Determine the sample by $s = \{k \mid b + (j-1)F \in (k-1, k],\ j = 1, \ldots, n\}$

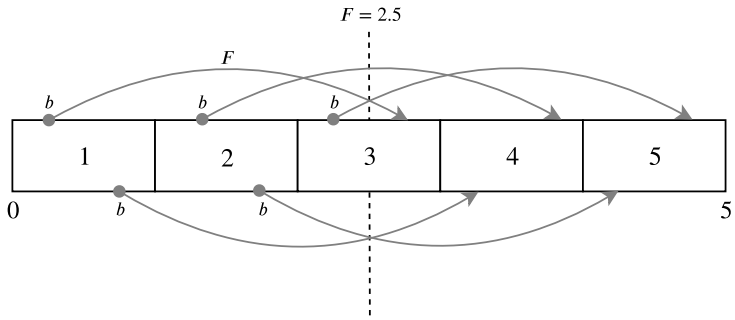


Figure 2.3: Graphical representation of the five different samples that are possible for Example 2.1.

The probabilities for the other samples can be computed similarly. The inclusion probability for the elements can also be computed. For element 3 we have

$$\begin{aligned}\pi_3 &= \mathbb{P}(S = s_1) + \mathbb{P}(S = s_5) = \mathbb{P}\left(b \in \left(2, \frac{5}{2}\right]\right) + \mathbb{P}(b + F \in (2, 3]) \\ &= \frac{1}{2} \cdot \frac{1}{F} + \frac{1}{2} \cdot \frac{1}{F} = \frac{2}{5}\end{aligned} \tag{2.22}$$

2.2. Data collection

After the sample is selected from the target population using a sampling design, the elements in the sample are requested to complete a questionnaire to collect data. It may occur that inhabitants in the sample do not respond to the questionnaire. This phenomenon is known as *nonresponse*. Throughout this thesis, we will assume all elements in the sample will respond to the questionnaire (so nonresponse does not occur) unless we state otherwise.

The collection of data is done with the use of electronic or paper questionnaires, which are used for three types of interviews:

(i) Computer-Assisted Personal Interviewing (CAPI)

(ii) Computer-Assisted Telephone Interviewing (CATI)

(iii) Computer-Assisted Web Interviewing (CAWI)

In the past, most questionnaires were to be completed in face-to-face (or personal) interviews. Interviewers visited the persons selected in the sample and filled in the information on a questionnaire. Although the rate of response was high the quality of the data tended to be good, it required a lot of interviewers and (travel) time which was expensive. In order to reduce costs, the *cluster size* m was introduced, which is defined as the minimum number of persons in a region that should be selected in the sample [3]. It means that in case a person in a region is selected in the sample, we should select $m - 1$ or more elements in the same region, such that the interviewer has to travel a shorter distance, which reduces costs. The introduction of the cluster size introduces a phenomenon called the *cluster effect*. The cluster effect occurs in case inhabitants in a region are more similar to one another than inhabitants in different regions with respect to the target variable of the survey. Collecting data from m 'similar' elements implies that less information is obtained than collecting data from m elements that are randomly chosen throughout the whole population [3]. For example, if aim of the survey is to draw conclusions about the religion of the Dutch population, selecting multiple inhabitants in the same region

may lead to less information than selecting inhabitants randomly. This is because religious involvement may be dependent on the region.

An alternative for face-to-face interviews is telephone interviewing, which requires interviewers who call the inhabitants to obtain information. No more traveling is necessary with telephone interviewing, which makes it a cheaper option. However, telephone interviewing is not always possible because not all inhabitants have a telephone number that is listed and questionnaires cannot be too long or too complicated.

Nowadays, web interviewing is used for most surveys of Statistics Netherlands. Persons that are selected in the sample receive a letter with the request to complete the form on the internet. Although reminders are set, fewer people tend to respond through this way of interviewing compared to the other types of interviewing [6]. Just like for telephone interviewing, web interviewing is not always possible, because not all inhabitants have internet access in the Netherlands. Because web interviewing is used for most surveys, the use of the cluster size became unnecessary. Therefore, the cluster size is usually equal to 1 nowadays.

2.3. Estimation

The data that is collected by completing the questionnaires can be used to obtain estimates for population parameters. Suppose the goal of the survey is to obtain an estimate for the population mean $\overline{y}_U$, which was defined in Equation (2.3). Any estimator $\hat{\overline{y}}_U$ of $\overline{y}_U$ is a statistic that produces values that, for most samples, lie near the unknown population parameter $\overline{y}_U$ [4]. Often the estimator is denoted by $\hat{\overline{y}}_U = \hat{\overline{y}}_U(S)$, which means that for any realisation s of S it is possible to compute $\hat{\overline{y}}_U$ from the target variables y_k for $k \in s$.

The performance of estimators can be determined by four quantities: expected value, bias, variance and mean squared error [3]. These four quantities are defined as:

Definition 2.9. The *expectation* of an estimator $\hat{\overline{y}}_U$ of $\overline{y}_U$ is defined by

$$\mathbb{E}\left(\hat{\overline{y}}_U\right) = \sum_{s \in \mathcal{S}} p(s)\hat{\overline{y}}_U(s) \tag{2.23}$$

where $p(s)$ is the sampling design from Definition 2.4. The expected value is the weighted average of the possible values of $\hat{\overline{y}}_U(s)$ where the weight is the probability the sample s is selected [4].

Definition 2.10. The *bias* of an estimator $\hat{\overline{y}}_U$ of $\overline{y}_U$ is defined as the difference between the expected value of the estimator and the population parameter $\overline{y}_U$ [4], i.e.

$$\mathbb{B}\left(\hat{\overline{y}}_U\right) = \mathbb{E}\left(\hat{\overline{y}}_U\right) - \overline{y}_U \tag{2.24}$$

An estimator $\hat{\overline{y}}_U$ is said to be *unbiased* for $\overline{y}_U$ if $\mathbb{B}\left(\hat{\overline{y}}_U\right) = 0$. Note that in general it is not possible to compute the bias, because $\overline{y}_U$ is unknown.

Definition 2.11. The *variance* of an estimator $\hat{\overline{y}}_U$ of $\overline{y}_U$ is given by [4]

$$\begin{aligned}\mathbb{V}\left(\hat{\overline{y}}_U\right) &= \sum_{s \in \mathcal{S}} p(s)\left(\mathbb{E}\left(\hat{\overline{y}}_U\right) - \hat{\overline{y}}_U(s)\right)^2 \\ &= \mathbb{E}\left(\left(\mathbb{E}\left(\hat{\overline{y}}_U\right) - \hat{\overline{y}}_U(s)\right)^2\right) \\ &= \mathbb{E}\left(\hat{\overline{y}}_U^2\right) - \mathbb{E}\left(\left(\hat{\overline{y}}_U\right)\right)^2\end{aligned} \tag{2.25}$$

Definition 2.12. The *mean square error (MSE)* of $\hat{\bar{y}}_U$ is defined as [4]

$$\text{MSE}\left(\hat{\bar{y}}_U\right) = \mathbb{E}\left(\left(\hat{\bar{y}}_U - \bar{y}_U\right)^2\right) \tag{2.26}$$

Note that the mean squared error is equal to the sum of the variance and the square of the bias, i.e.

$$\begin{aligned}\text{MSE}\left(\hat{\bar{y}}_U\right) &= \mathbb{E}\left(\hat{\bar{y}}_U^2\right) - 2\mathbb{E}\left(\hat{\bar{y}}_U\right)\bar{y}_U + \bar{y}_U^2 + \left(\mathbb{E}\left(\hat{\bar{y}}_U\right)\right)^2 - \left(\mathbb{E}\left(\hat{\bar{y}}_U\right)\right)^2 \\ &= \mathbb{E}\left(\hat{\bar{y}}_U^2\right) - \left(\mathbb{E}\left(\hat{\bar{y}}_U\right)\right)^2 + \left(\mathbb{E}\left(\hat{\bar{y}}_U\right) - \bar{y}_U\right)^2 \\ &= \mathbb{V}\left(\hat{\bar{y}}_U\right) + \left(\mathbb{B}\left(\hat{\bar{y}}_U\right)\right)^2\end{aligned} \tag{2.27}$$

2.3.1. The Horvitz-Thompson estimator

Horvitz and Thompson [7] introduced an estimator for the population mean that uses the inclusion probabilities of the elements in the population to estimate the population total $\bar{y}_U$. This resulted in the *Horvitz-Thompson estimator*, which is also known as the *π-estimator* [4].

Definition 2.13. The *Horvitz-Thompson estimator* $\hat{\bar{y}}_{\text{HT}}$ for $\bar{y}_U$ is defined as

$$\hat{\bar{y}}_{\text{HT}} = \frac{1}{N}\sum_{k\in s}\frac{1}{\pi_k}y_k \tag{2.28}$$

It is often useful to express the Horvitz-Thompson estimator as a linear function of the indicators

$$\hat{\bar{y}}_{\text{HT}} = \frac{1}{N}\sum_{k=1}^{N}\frac{y_k}{\pi_k}I_k \tag{2.29}$$

The operation dividing the value variable for element k by the inclusion probability of element k will be used often in this thesis. So like Särndal et al. [4] do in their book, we introduce the notation

$$\breve{y}_k = \frac{y_k}{\pi_k} \tag{2.30}$$

In the following theorems the most important properties of the Horvitz-Thompson estimator are described and proved.

Theorem 2.1 (Unbiasedness Horvitz-Thompson estimator)**.** The Horvitz-Thompson estimator $\hat{\bar{y}}_{\text{HT}}$ is an unbiased estimator for the population mean $\bar{y}_U$ [4].

Proof. Using the expectation of the sample membership indicator in Equation (2.10), the expectation of $\hat{\bar{y}}_{\text{HT}}$ is

$$\mathbb{E}\left(\hat{\bar{y}}_{\text{HT}}\right) = \mathbb{E}\left(\frac{1}{N}\sum_{k=1}^{N}\frac{y_k}{\pi_k}I_k\right) = \frac{1}{N}\sum_{k=1}^{N}\breve{y}_k\mathbb{E}\left(I_k\right) = \frac{1}{N}\sum_{k=1}^{N}y_k = \bar{y}_U \tag{2.31}$$

□

Theorem 2.2 (Variance Horvitz-Thompson estimator)**.** The variance of the Horvitz-Thompson estimator for $\bar{y}_U$ is [4]

$$\mathbb{V}\left(\hat{\bar{y}}_{\text{HT}}\right) = \frac{1}{N^2}\sum_{k=1}^{N}\sum_{l=1}^{N}\left(\pi_{kl} - \pi_k\pi_l\right)\breve{y}_k\breve{y}_l \tag{2.32}$$

Proof. The proof can be found in Appendix A.1.

Theorem 2.3 (Variance Horvitz-Thompson estimator for fixed sample size designs)**.** If the sample size n of the sampling design $p(\cdot)$ is fixed, the variance of the Horvitz-Thompson estimator in Equation (2.32) can be written alternatively as [4]

$$\mathbb{V}\left(\hat{\bar{y}}_{\mathsf{HT}}\right) = -\frac{1}{2N^2}\sum_{k=1}^{N}\sum_{l=1}^{N}\left(\pi_{kl}-\pi_k\pi_l\right)\left(\breve{y}_k-\breve{y}_l\right)^2 \tag{2.33}$$

Proof. The proof can be found in Appendix A.2

2.3.2. The Horvitz-Thompson estimator for SRSWR

Since simple random sampling without replacement is a widely used sampling design and the Horvitz-Thompson estimator is a widely used estimator, we examine the Horvitz-Thompson estimator for SRSWR.

For simple random sampling without replacement, the Horvitz-Thompson estimator for the population mean $\bar{y}_U$ equals the the sample mean of y, i.e.

$$\hat{\bar{y}}_{\mathsf{HT}} = \frac{1}{N}\sum_{k\in s}\frac{1}{\pi_k}y_k = \frac{1}{N}\sum_{k\in s}\frac{N}{n}y_k = \frac{1}{n}\sum_{k\in s}y_k \tag{2.34}$$

Simple random sampling without replacement is a fixed sample size design, which gives that we can use Theorem 2.3 to find the variance for the Horvitz-Thompson estimator for SRSWR.

First, note that for SRSWR, $k \neq l$ and the sampling fraction $f = \frac{n}{N}$ we have that

$$\begin{aligned}
\pi_{kl}-\pi_k\pi_l &= \frac{n(n-1)}{N(N-1)} - \frac{n}{N}\frac{n}{N} = f\left(\frac{n-1}{N-1}-\frac{n}{N}\right)\\
&= f\left(\frac{(n-1)N-n(N-1)}{N(N-1)}\right)\\
&= f\left(\frac{n-N}{N(N-1)}\right) = -\frac{f}{N-1}\frac{N-n}{N}\\
&= -\frac{f(1-f)}{N-1}
\end{aligned} \tag{2.35}$$

The variance of the Horvitz-Thompson estimator then equals [4]

$$\begin{aligned}
\mathbb{V}\left(\hat{\bar{y}}_{\mathsf{HT}}\right) &= -\frac{1}{2N^2}\sum_{k=1}^{N}\sum_{l=1}^{N}\left(\pi_{kl}-\pi_k\pi_l\right)\left(\breve{y}_k-\breve{y}_l\right)^2\\
&= -\frac{1}{2N^2}\sum_{k=1}^{N}\sum_{k\neq l=1}^{N}\left(\pi_{kl}-\pi_k\pi_l\right)\left(\breve{y}_k-\breve{y}_l\right)^2 - \frac{1}{2N^2}\sum_{k=1}^{N}\left(\pi_{kk}-\pi_k\pi_k\right)\left(\breve{y}_k-\breve{y}_k\right)^2\\
&= -\frac{1}{2N^2}\sum_{k=1}^{N}\sum_{k\neq l=1}^{N}-\frac{f(1-f)}{N-1}\left(\frac{N}{n}\left(y_k-y_l\right)\right)^2 + 0\\
&= \frac{1}{2N^2}\frac{f(1-f)}{N-1}\frac{1}{f^2}\sum_{k=1}^{N}\sum_{k\neq l=1}^{N}\left(y_k-y_l\right)^2
\end{aligned} \tag{2.36}$$

Since $(y_k - y_l)^2$ is zero for $k = l$, we can just write

$$\mathbb{V}\left(\hat{\bar{y}}_{\mathrm{HT}}\right) = \frac{1}{2N^2}\frac{f(1-f)}{N-1}\frac{1}{f^2}\sum_{k=1}^{N}\sum_{l=1}^{N}(y_k - y_l)^2 \tag{2.37}$$

Recall from Equation (2.3) that the population mean of y is denoted by $\bar{y}_U$, which gives

$$\begin{aligned}
\mathbb{V}\left(\hat{\bar{y}}_{\mathrm{HT}}\right) &= \frac{1}{2N^2}\frac{1-f}{N-1}\frac{1}{f}\sum_{k=1}^{N}\sum_{l=1}^{N}\left(\left(y_k - \bar{y}_U\right) - \left(y_l - \bar{y}_U\right)\right)^2 \\
&= \frac{1}{2N^2}\frac{1-f}{N-1}\frac{1}{f}\left(2N\sum_{k=1}^{N}\left(y_k - \bar{y}_U\right)^2 - 2\sum_{k=1}^{N}\sum_{l=1}^{N}\left(y_k - \bar{y}_U\right)\left(y_l - \bar{y}_U\right)\right) \\
&= \frac{1}{2N^2}\frac{1-f}{N-1}\frac{1}{f}\left(2N\sum_{k=1}^{N}\left(y_k - \bar{y}_U\right)^2 - 2\sum_{k=1}^{N}\left(\left(y_k - \bar{y}_U\right)\sum_{l=1}^{N}\left(y_l - \bar{y}_U\right)\right)\right) \\
&= \frac{1}{2N^2}\frac{1-f}{N-1}\frac{1}{f}\left(2N\sum_{k=1}^{N}\left(y_k - \bar{y}_U\right)^2 - 2\sum_{k=1}^{N}\left(y_k - \bar{y}_U\right)\left(N\bar{y}_U - N\bar{y}_U\right)\right) \\
&= \frac{1-f}{n}\frac{1}{N-1}\sum_{k=1}^{N}\left(y_k - \bar{y}_U\right)^2 \\
&= \frac{1-f}{n}S_{yU}^2
\end{aligned} \tag{2.38}$$

where S_{yU}^2 denotes the population variance, i.e.

$$S_{yU}^2 = \frac{1}{N-1}\sum_{k=1}^{N}\left(y_k - \bar{y}_U\right)^2 \tag{2.39}$$

2.4. Publication

After the collected data is used to obtain estimates for population characters, the results of the survey are published in, for example, a report or on the internet. Sometimes the results may also contain some further analysis on the collected data, such as possible patterns/relations in the collected data or testing hypotheses that have been formulated about the population [3]. Results of surveys of Statistics Netherlands are often published on Statline [8].

3

Sampling design

An important part of selecting the sample, is the decision for the sampling design that is used. Throughout this thesis, we only consider sampling designs where elements are selected with equal probability. In survey sampling terminology such a sample is known as *self-weighting*, see Definition 2.8.

In this section, we describe the sampling design that is used for selecting self-weighting samples at Statistics Netherlands. The sampling design was developed at a time when there was no sampling frame available for the total population of the Netherlands, but each municipality had its own population register [3]. Therefore, it was decided to use a two-stage sampling design to select samples from the Dutch population [3]. In practice this means that first municipalities are selected and subsequently inhabitants are selected from the selected municipalities. Another reason to use this sampling design is that costs can be reduced by the introduction of the cluster size, see Section 2.2. In the first stage, municipalities are selected by systematic sampling with probabilities proportional to the population sizes of the municipalities. In the second stage, inhabitants from the selected municipalities are selected using simple random sampling without replacement (SRSWR). The municipalities are sometimes referred to as the *primary sampling units* and the inhabitants as the *secondary sampling units*.

The Netherlands is divided into 40 NUTS-3-regions[1], see Figure 3.1. The NUTS (Nomenclature des Unités Territoriales Statistiques) is a regional classification of the European statistical office Eurostat [9]. Each NUTS-3-region consists of a number of municipalities. To make sure the elements in the sample are proportionally distributed among the different NUTS-3-regions, the sampling design that is described in this section is applied to each region independently. This means that a *stratified sampling* is applied, where the *strata* are the NUTS-3-regions [4]. The sample size in a NUTS-3-region is proportional to the population size in the corresponding NUTS-3-region.

3.1. Selecting municipalities with probabilities proportional to size

Let the target population in a NUTS-3-region be denoted by $U = \{1, 2, \ldots, N\}$. Suppose a sample of size n should be selected from the population U such that each element in the population has an equal probability to be selected in the sample.

[1] In Dutch known as COROP-region, which is an abbreviation for *Coördinatiecommissie Regionaal Onderzoeksprogramma*.

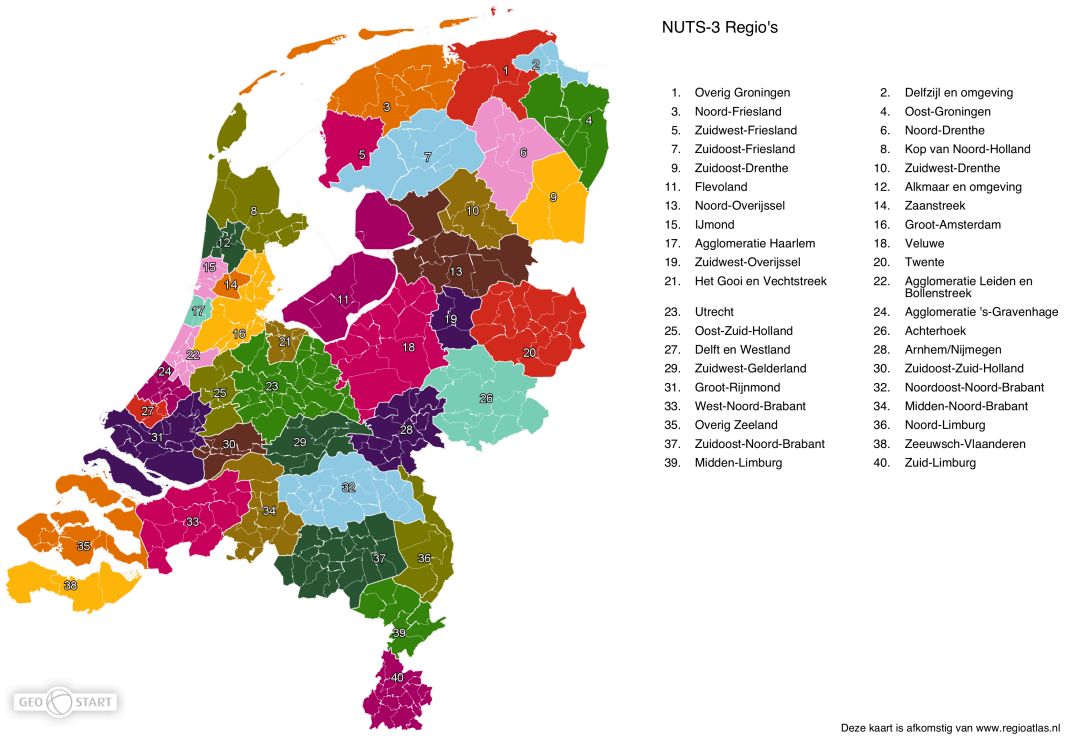


Figure 3.1: Map of the Netherlands divided into 40 NUTS-3-regions. Figure from regioatlas.nl [9].

The population is divided into I municipalities which do not overlap and cover the complete population U. We denote the i-th $(i = 1, \dots, I)$ municipality by U_i such that

$$U = U_1 \cup \cdots \cup U_I \quad \text{and} \quad U_i \cap U_j = \varnothing \; \forall \, i, j = 1, \dots, I \tag{3.1}$$

Let the population size of municipality U_i be denoted by N_i. Then the sum of the population sizes of the municipalities equals the total population size, i.e.

$$\sum_{i=1}^{I} N_i = N \tag{3.2}$$

The municipalities $1, \dots, I$ are now selected by systematic sampling, with probabilities proportional to the sizes $N_1, \dots, N_I$. The idea is similar to systematic sampling with equal probabilities, that was introduced in Section 2.1.2, but now municipalities do not have equal inclusion probabilities. Suppose the municipalities are sorted in random order. Let the total population U be represented by the interval $\mathcal{I} = (0, N]$ and the population of municipality U_i by the interval

$$\mathcal{I}_i = \left(\sum_{j=1}^{i-1} N_j, \sum_{j=1}^{i} N_j \right] \quad \text{for } i = 2, \dots, I \tag{3.3}$$

and $\mathcal{I}_1 = (0, N_1]$. Note that this is defined such that $\mathcal{I}_i$ has length N_i, for $i = 1, \dots, I$ and

$$\mathcal{I} = \bigcup_{i=1}^{I} \mathcal{I}_i \tag{3.4}$$

Denote the number of municipalities to select by n_I. We will sometimes refer to n_I as the sample size of municipalities. As was described in Section 2.2 the *cluster size* was introduced in order to reduce costs. The cluster size is defined as the minimum number of inhabitants to be selected in a municipality, if the municipality is selected. We will denote the cluster size by m. It means that when a municipality U_i is selected, at least m inhabitants from municipality U_i should be selected in the sample. It follows that the number of municipalities to select is at most n/m, which we denote by $n_I^{\max}$, i.e.

$$n_I \leq n_I^{\max} = \frac{n}{m} \tag{3.5}$$

Let the *step length* or *sampling interval* F be computed by

$$F = \frac{N}{n_I^{\max}} = N\frac{m}{n} \tag{3.6}$$

The first municipality is selected by choosing a random number b in the interval $(0, F]$ uniformly. The municipality U_i for which $b \in \mathcal{I}_i$ is selected in the sample of municipalities, that we will denote by s_I. The other municipalities are determined by selecting the municipalities that correspond to the values $b+F, b+2F, \ldots, (n/m-1)F$. So we can denote the sample of municipalities as

$$s_I = \left\{ U_i \;\middle|\; b + (j-1)F \in \mathcal{I}_i,\ j = 1, \ldots, \frac{n}{m} - 1 \right\} \tag{3.7}$$

Note that there may be municipalities that have a population size that is greater than or equal to the step length. This would mean that those municipalities are selected at least once in the sample of municipalities s_I. Municipalities U_i for which $N_i \geq F$ are called *self-selecting* municipalities, and municipalities for which $N_i < F$ are called *non-self-selecting* municipalities.

The next example shows that self-selecting municipalities are always selected and can be selected more than once in the sample of municipalities.

Example 3.1. Suppose a population of size 10 contains 3 municipalities U_1, U_2, U_3 with sizes $N_1 = 6$, $N_2 = 3$ and $N_3 = 1$. Let the cluster size be $m = 2$ and the sample size $n = 4$. Then we should select at most $\frac{4}{2} = 2$ municipalities and the step length is $F = 10 \cdot \frac{2}{4} = 5$. In Figure 3.2 this is illustrated graphically. In the example in Figure 3.2, the random start is $b = 3$. Note that municipality U_1 has a population size that is greayer than the step length. Consequently, municipality U_1 is self-selecting, and hence it is selected in the sample s_I for all possible values of b. Depending on the value of the random start, U_2 or U_3 can be selected in the sample of municipalities. Note that if b is smaller than 1, municipality U_1 is selected in the sample twice.

Under the circumstances that there are self-selecting municipalities, municipalities can be selected more than once in the sample. In Section 3.4, we will describe a simplification of the design that ensures municipalities can only be selected once in the sample s_I. But first, we compute the first-order inclusion probabilities of the municipalities.

3.2. First-order inclusion probabilities of the municipalities

We have seen that there are self-selecting municipalities and non-self-selecting municipalities. For both municipalities, we compute the first-order inclusion probabilities.

Consider a non-self-selecting municipality U_i with population size N_i. Let s_I be the outcome of a set-valued random variable S_I. Then the inclusion probability of the municipality, which we denote by

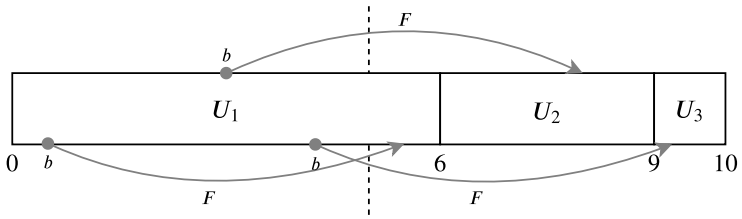


Figure 3.2: Graphical representation of Example 3.1, with a population of size 10 containing 3 municipalities.

π_{U_i} is [3]

$$\begin{aligned}\pi_{U_i} &= \mathbb{P}\left(U_i \in S_I\right) \\ &= \mathbb{P}\left(b \in \mathcal{I}_i\right) + \mathbb{P}\left(b + F \in \mathcal{I}_i\right) + \cdots + \mathbb{P}\left(b + \left(\frac{n}{m} - 1\right) F \in \mathcal{I}_i\right) \\ &= \frac{N_i}{F} = \frac{N_i}{N}\frac{n}{m}\end{aligned} \tag{3.8}$$

Next, consider a self-selecting municipality U_i. Then the population size N_i is greater than or equal to F. We have seen that self-selecting municipalities are selected at least once in the sample of municipalities s_I. Consequently, if U_i is a self-selecting municipality we define its inclusion probability equal to 1.

To sum up, the first-order inclusion probability of a municipality U_i is

$$\pi_{U_i} = \mathbb{P}(U_i \in S_I) = \begin{cases} \frac{N_i}{N}\frac{n}{m} & \text{if } N_i < N\frac{m}{n} \\ \\ 1 & \text{if } N_i \geq N\frac{m}{n} \end{cases} \tag{3.9}$$

3.3. Selecting inhabitants with equal probabilities

The sample of municipalities s_I is determined by applying systematic sampling with probabilities proportional to the population size of the municipalities. The goal is to obtain a sample of inhabitants s of size n such that each inhabitant has the same probability to be selected. This means that each inhabitant should have an inclusion probability of n/N. This can be obtained by choosing the number of inhabitants to be selected from each municipality properly.

Let n_i denote the number of inhabitants to be selected from municipality U_i. The n_i inhabitants are selected from the population of size N_i by simple random sampling without replacement, that was described in Section 2.1.1. Recall that a municipality can either be non-self-selecting or self-selecting. We will consider those cases separately.

Non-self-selecting municipality Consider a non-self-selecting municipality U_i with population size N_i. Then

$$N_i < F \quad \text{and} \quad \pi_{U_i} = \frac{N_i}{N}\frac{n}{m} \tag{3.10}$$

Consider element k in municipality U_i. Then the inclusion probability of element k is

$$\pi_k = \mathbb{P}\left(k \in S\right) = \mathbb{P}\left(k \in S | U_i \in S_I\right) \cdot \mathbb{P}\left(U_i \in S_I\right) = \frac{n_i}{N_i} \cdot \frac{N_i}{N}\frac{n}{m} = \frac{n_i}{N}\frac{n}{m} \tag{3.11}$$

By selecting exactly $n_i = m$ inhabitants from U_i, the inclusion probability of element k is n/N. So, for each non-self-selecting municipality U_i select $n_i = m$ inhabitants by SRSWR.

Self-selecting municipality Consider a self-selecting municipality U_i with population size N_i. Then

$$N_i \geq F \quad \text{and} \quad \pi_{U_i} = 1 \tag{3.12}$$

Again, consider inhabitant k in municipality U_i. Then

$$\pi_k = \mathbb{P}(k \in S) = \mathbb{P}(k \in S | U_i \in S_I) \cdot \mathbb{P}(U_i \in S_I) = \frac{n_i}{N_i} \tag{3.13}$$

This inclusion probability is equal to n/N if and only if $n_i = \frac{n}{N} N_i$. So for each self-selecting municipality U_i, select $n_i = \frac{n}{N} N_i$ inhabitants by SRSWR.

To summarise, the sample size of a municipality U_i is

$$n_i = \begin{cases} m & \text{if } U_i \text{ non-self-selecting} \\ \\ \frac{n}{N} N_i & \text{if } U_i \text{ self-selecting} \end{cases} \tag{3.14}$$

Remark 3.1. Note that this procedure can result in several non-integer sample sizes. In practice that cannot be the case, so sample sizes are rounded. This can cause slight deviations, which are usually ignored.

3.4. Practical simplification of the sampling design

The two-stage sampling design that was described in the previous sections, is applied for selecting self-weighting samples that are selected from the whole Dutch population at Statistics Netherlands. We have seen that the self-selecting municipalities have an inclusion probability of 1, which means that those municipalities are always included in the sample s_I. Consequently, we can start by including the self-selecting municipalities in s_I and excluding them from the population. This leads to a simplification of the implementation of the two-stage sampling design.

Let $\widetilde{N}$ denote the total population size for all non-self-selecting municipalities and let $\widetilde{I}$ denote the number of non-self-selecting municipalities. Then the total population size of the self-selecting municipalities is $N - \widetilde{N}$ and there are $I - \widetilde{I}$ self-selecting municipalities. We have seen that from each self-selecting U_i, we select $n_i = \frac{n}{N} N_i$ inhabitants. Consequently, the sample size of inhabitants in all self-selecting municipalities equals

$$\frac{n}{N} \left(N - \widetilde{N} \right) \tag{3.15}$$

Let $\tilde{n}$ denote the sample size of inhabitants from all non-self-selecting municipalities. Then $\tilde{n}$ is equal to

$$\tilde{n} = n - \frac{n}{N} \left(N - \widetilde{N} \right) \tag{3.16}$$

Recall that the municipalities are sorted in random order. Without loss of generality, we may assume that the first $\widetilde{I}$ municipalities are non self-selecting. Then the population of all non-self-selecting municipalities can be represented by the interval

$$\widetilde{\mathcal{I}} = \left(0, \widetilde{N} \right] \tag{3.17}$$

and each non-self-selecting municipality U_i can be represented by a subinterval

$$\widetilde{\mathcal{I}}_i = \left(\sum_{j=1}^{i-1} N_j, \sum_{j=1}^{i} N_j \right], \quad \text{for } i = 2, \ldots, \widetilde{I} \tag{3.18}$$

and $\widetilde{\mathcal{I}}_1 = (0, N_1]$. From each non-self-selecting municipality at least m inhabitants should be selected, which gives that the number of non-self-selecting municipalities to select is

$$\tilde{n}_I = \frac{\tilde{n}}{m} = \frac{1}{m}\left(n - \frac{n}{N}\left(N - \widetilde{N}\right)\right) = \frac{n}{m} - \frac{n}{m} + \frac{n}{m}\frac{\widetilde{N}}{N} = \frac{n}{m}\frac{\widetilde{N}}{N} \tag{3.19}$$

We can now apply systematic sampling with unequal probabilities again, with a step length $\widetilde{F}$ of size

$$\widetilde{F} = \frac{\widetilde{N}}{\tilde{n}_I} = N\frac{m}{n} \tag{3.20}$$

Note that the step length $\widetilde{F}$ is the same as the step length F that was defined in Equation (3.6). Now the first non-self-selecting municipality is determined by choosing the random start $b \in (0, F]$ uniformly. The municipality U_i for which $b \in \widetilde{\mathcal{I}}_i$ is selected in s_I. The other municipalities in s_I are the municipalities that correspond to the values $b + F, b + 2F, \ldots, b + (\tilde{n}_I - 1)F$. The sample of municipalities s_I can be denoted by

$$s_I = \{U_i \mid N_i \geq F\} \cup \left\{ U_i \;\middle|\; b + (j-1)F \in \widetilde{\mathcal{I}}_i, j = 1, \ldots, \tilde{n}_I \right\} \tag{3.21}$$

The procedure is summarised in Algorithm 3.1.

Remark 3.2. Recall that in Section 2.2 we have explained that since web interviewing is used for most surveys nowadays, the cluster size is usually defined as 1. This means that the step size F is equal to N/n. If sample sizes are large enough, most municipalities are self-weighting. We use data that is available on Statline [8]. In 2018, the Dutch population size was $17,181,084$ [8]. If we use a sample size of $n = 4,000$, then the step size is equal to $4,295$. For this sample size, only four out of 380 municipalities are non-self-selecting. Consequently, if the cluster size is equal to 1 and the sample sizes are large enough, the self-weighting two-stage sampling design is comparable to SRSWR.

3.5. Second-order inclusion probabilities

In Section 7 we will discuss the estimator that is used at Statistics Netherlands. For computing the variance of the estimator, the second-order inclusion probabilities are required. In this section we compute the second-order inclusion probabilities corresponding to the self-weighting two stage design, first for municipalities and then for inhabitants.

3.5.1. Second-order inclusion probabilities of municipalities

Consider two municipalities U_i and U_j in the Netherlands. We will compute the second-order inclusion probability $\pi_{U_iU_j}$, that denotes the probability that both municipalities U_i and U_j are included in the sample of municipalities s_I. Without mentioning it explicitly, we assume that $U_i \neq U_j$, because for $U_i = U_j$ we simply have

$$\pi_{U_iU_i} = \pi_{U_i} \tag{3.22}$$

where π_{U_i} is given by Equation (3.9).

Algorithm 3.1: Procedure to select a sample according to the two-stage self-weighting sampling design that is used by Statistics Netherlands.

Two-stage systematic sampling design with probabilities proportional to size

1. Initialisation:
   - N population size
   - n sample size
   - I number of municipalities
   - N_i population size municipality U_i $(i = 1, \ldots, I)$
   - m cluster size
2. Compute $F = N\frac{m}{n}$
3. **For** each municipalities $U_i\,(i = 1, \ldots, I)$
   If $N_i \geq F$
   - Municipality U_i is self-selecting
   - Add U_i to s_I
   - Select $n_i = \frac{n}{N}N_i$ elements from U_i by SRSWR

   If $N_i < F$
   - Municipality U_i is non-self-selecting
4. Compute:
   - $\widetilde{N}$ population size of all non-self-selecting municipalities
   - $\tilde{n} = n - \frac{n}{N}(N - \widetilde{N})$ number of inhabitants to select of all self-selecting municipalities
   - $\tilde{n}_I = \frac{n}{m}\frac{\widetilde{N}}{N}$ number of non-self-selecting municipalities to select
   - λ_i indicator function that is 1 for non-self-selecting municipalities
5. Compute the subintervals $\widetilde{\mathcal{I}}_i = \left(\sum_{j=1}^{i-1}\lambda_j N_j, \sum_{j=1}^{i}\lambda_j N_j\right]$
6. Choose $b \in (0, F]$ randomly
7. Add the municipalities $\left\{U_i \;\middle|\; b + (j-1)F \in \widetilde{\mathcal{I}}_i, j = 1, \ldots, \tilde{n}_I\right\}$ to s_I
8. **For** each non-self-selecting municipalities $U_i \in s_I$
   Select $n_i = m$ elements from U_i by SRSWR

Note that the two-stage sampling design is applied to each NUTS-3-region independently. Due to the independence, second-order inclusion probabilities for a pair of municipalities in different NUTS-3-regions are easy to calculate. If U_i and U_j are in different NUTS-3-regions, then

$$\pi_{U_iU_j} = \pi_{U_i}\pi_{U_j} \tag{3.23}$$

where π_{U_i} is the first-order inclusion probability as was described in Equation (3.9). From now on, assume that U_i and U_j are municipalities in the same NUTS-3-region, then municipalities U_i and U_j can be self-selecting or non-self-selecting municipalities. We consider each case for the municipalities U_i and U_j separately.

(i) Suppose U_i and U_j are both self-selecting municipalities. Then

$$\pi_{U_iU_j} = \mathbb{P}\left(U_i, U_j \in S_I\right) = 1 \tag{3.24}$$

(ii) Suppose U_i is a self-selecting municipality and U_j is a non-self-selecting municipality, then

$$\pi_{U_iU_j} = \mathbb{P}\left(U_i, U_j \in S_I\right) = \mathbb{P}\left(U_j \in S_I\right) = \pi_{U_j} = \frac{N_i}{N}\frac{n}{m} \tag{3.25}$$

Note that this case is similar to the case when U_i is non-self-selecting and U_j is self-selecting.

(iii) Suppose U_i and U_j are both non-self-selecting municipalities. Then the second-order inclusion probability is more difficult to compute, because the second-order inclusion probabilities are dependent on the order of the municipalities, see Section 2.1.2.

Vondenhoff and van Berkel [10] derive the second-order inclusion probability for two non-self-selecting municipalities. The calculations of these probabilities is computationally intensive, so several approximations for the second-order inclusion probability are derived [10]. The simplest approximation for $\pi_{U_iU_j}$ is bilinear in the pair $\left(N_i, N_j\right)$[10], which gives

$$\pi^B_{U_iU_j} = CN_iN_j \tag{3.26}$$

where $C > 0$ is a constant

$$C = \frac{1}{F^2}\left(1 - \frac{1}{\tilde{n}_I}\right)\left(1 - \frac{1}{\sum_{g=1}^{\tilde{I}} \frac{N_g^2}{\tilde{N}^2}}\right) \tag{3.27}$$

3.5.2. Second-order inclusion probabilities of inhabitants

Let k and l be two elements of the population $U = \{1, 2, \ldots, N\}$, such that k is in municipality U_i and l in municipality U_j. We can now compute the probability that both element k and l ($k \neq l$) are selected in the sample. We will use the first- and second order inclusion probabilities of the municipalities from Sections 3.2 and 3.5.1, and the first- and second-order inclusion probabilities for SRSWR from Section 2.1.1. This gives the conditional first-order inclusion probability

$$\mathbb{P}\left(k \in S \mid U_i \in S_I\right) = \frac{n_i}{N_i} \tag{3.28}$$

and if k and l both in municipality U_i, the conditional second-order inclusion probability

$$\mathbb{P}\left(k, l \in S \mid U_i \in S_I\right) = \frac{n_i(n_i - 1)}{N_i(N_i - 1)} \tag{3.29}$$

To compute the unconditional second-order inclusion probabilities for inhabitants we will distinguish the different cases:

(i) If inhabitant k and l are in the same municipality, so $U_i = U_j$ then

$$\begin{aligned} \pi_{kl} &= \mathbb{P}\left(k, l \in S \mid U_i \in S_I\right) \cdot \mathbb{P}\left(U_i \in S_I\right) \\ &= \frac{n_i(n_i - 1)}{N_i(N_i - 1)}\pi_{U_i} \end{aligned} \tag{3.30}$$

Here the sample size n_i and the inclusion probability π_{U_i} depend on whether U_i is self-selecting or not, see Equations (3.9) and (3.14). If U_i is self-selecting we have

$$\pi_{kl} = \frac{\frac{n}{N}N_i\left(\frac{n}{N}N_i - 1\right)}{N_i(N_i - 1)} = \frac{n}{N}\frac{\frac{n}{N}N_i - 1}{N_i - 1} \tag{3.31}$$

And if U_i is non-self-selecting we have

$$\pi_{kl} = \frac{m(m-1)}{N_i(N_i - 1)} \cdot \frac{N_i}{N}\frac{n}{m} = \frac{n}{N}\frac{m-1}{N_i - 1} \quad \text{for } m \geq 2 \tag{3.32}$$

Note that if the cluster size m is defined as 1, the second-order inclusion probability is zero in this case.

(ii) Suppose inhabitants k and l are in different municipalities that are in different NUTS-3-regions. Recall that the sampling design is applied to each NUTS-3-region independently, which gives that the second-order inclusion probability is rather simple in this case

$$\begin{aligned}\pi_{kl} &= \mathbb{P}\left(k,l \in S \;\middle|\; U_i, U_j \in S_I\right) \cdot \mathbb{P}\left(U_i, U_j \in S_I\right) \\ &= \mathbb{P}\left(k \in S \mid U_i \in S_I\right) \cdot \mathbb{P}\left(l \in S \;\middle|\; U_j \in S_I\right) \cdot \mathbb{P}\left(U_i \in S_I\right) \cdot \mathbb{P}\left(U_j \in S_I\right) \\ &= \frac{n_i}{N_i}\frac{n_j}{N_j}\pi_{U_i}\pi_{U_j}\end{aligned} \tag{3.33}$$

The sample sizes n_i and n_j and the inclusion probabilities of the municipalities π_{U_i} and π_{U_j} depend on whether U_i and U_j are self-selecting or not, see Equations (3.9) and (3.14). Note that in each case we have that $\pi_{kl} = \frac{n}{N}\frac{n}{N}$. Recall that the sampling design is applied to each NUTS-3-region independently. Given the independency, we could have stated immediately that $\pi_{kl} = \frac{n}{N}\frac{n}{N}$.

Besides assuming $k \neq l$, from now on we will assume $U_i \neq U_j$ and that municipalities U_i and U_j are in the same NUTS-3-region.

(iii) Suppose U_i and U_j are both self-selecting municipalities, then

$$\begin{aligned}\pi_{kl} &= \mathbb{P}\left(k,l \in S \;\middle|\; U_i, U_j \in S_I\right) \cdot \mathbb{P}\left(U_i, U_j \in S_I\right) \\ &= \mathbb{P}\left(k \in S \mid U_i \in S_I\right) \cdot \mathbb{P}\left(l \in S \;\middle|\; U_j \in S_I\right) \cdot \mathbb{P}\left(U_i, U_j \in S_I\right) \\ &= \frac{n_i}{N_i} \cdot \frac{n_j}{N_j} \cdot 1 \\ &= \frac{n}{N}\frac{n}{N}\end{aligned} \tag{3.34}$$

(iv) If U_i is a self-selecting municipality and U_j a non-self-selecting municipality, then

$$\begin{aligned}\pi_{kl} &= \mathbb{P}\left(k,l \in S \;\middle|\; U_i, U_j \in S_I\right) \cdot \mathbb{P}\left(U_i, U_j \in S_I\right) \\ &= \mathbb{P}\left(k \in S \mid U_i \in S_I\right) \cdot \mathbb{P}\left(l \in S \;\middle|\; U_j \in S_I\right) \cdot \mathbb{P}\left(U_j \in S_I\right) \\ &= \frac{n_i}{N_i}\frac{n_j}{N_j}\frac{N_j}{N}\frac{n}{m} \\ &= \frac{n}{N}\frac{m}{N_j}\frac{N_j}{N}\frac{n}{m} \\ &= \frac{n}{N}\frac{n}{N}\end{aligned} \tag{3.35}$$

This case is similar to the case where U_i is non-self-selecting and U_j is self-selecting.

(v) Suppose U_i and U_j are both non-self-selecting municipalities, then the second-order inclusion probability for elements k and l is

$$\begin{aligned}\pi_{kl} &= \mathbb{P}\left(k,l \in S \;\middle|\; U_i, U_j \in S_I\right) \cdot \mathbb{P}\left(U_i, U_j \in S_I\right) \\ &= \mathbb{P}\left(k \in S \mid U_i \in S_I\right) \cdot \mathbb{P}\left(l \in S \;\middle|\; U_j \in S_I\right) \cdot \mathbb{P}\left(U_i, U_j \in S_I\right) \\ &= \frac{n_i}{N_i}\frac{n_j}{N_j} \cdot \pi_{U_iU_j} \\ &= \frac{m}{N_i}\frac{m}{N_j} \cdot \pi_{U_iU_j}\end{aligned} \tag{3.36}$$

where an approximation of $\pi_{U_iU_j}$ is given by Equation (3.26). So an approximation for the second-order inclusion probability is

$$\begin{aligned}
\pi^B_{kl} &= \frac{m}{N_i}\frac{m}{N_j} \cdot \pi^B_{U_iU_j} \\
&= \frac{m}{N_i}\frac{m}{N_j} \cdot CN_iN_j \\
&= m^2 \frac{1}{F^2}\left(1 - \frac{1}{\tilde{n}_I}\right)\left(1 - \frac{1}{\sum_{g=1}^{\widetilde{I}} \frac{N_g^2}{\widetilde{N}^2}}\right) \\
&= \frac{n}{N}\frac{n}{N}\left(1 - \frac{1}{\tilde{n}_I}\right)\left(1 - \frac{1}{\sum_{g=1}^{\widetilde{I}} \frac{N_g^2}{\widetilde{N}^2}}\right)
\end{aligned} \tag{3.37}$$

Note that if $\tilde{n}_I$ and $\sum_{g=1}^{\widetilde{I}} \frac{N_g^2}{\widetilde{N}^2}$ are large we have that $\pi^B_{kl} \approx \frac{n}{N}\frac{n}{N}$.

3.6. Conclusion

The self-weighting two-stage sampling design that is introduced in this chapter, is similar to simple random sampling without replacement. In both designs, inhabitants are selected with equal probabilities and without replacement. However, the second-order inclusion probabilities of the two-stage sampling design are not equal to those of SRSWR. This is because the two-stage sampling design ensures that at least m inhabitants are selected per municipality, and the number of inhabitants per municipality is proportional to the population size of the municipality.

For simplicity, we sometimes use simple random sampling without replacement instead of the two-stage sampling design in the next chapters. We have shown that when the cluster size is equal to 1 and the sample size is large enough, most municipalities are self-selecting. Note that if most municipalities are self-selecting, the second-order inclusion probabilities are comparable to the second-order inclusion probabilities of SRSWR. So if the cluster size is equal to 1 (which is usually the case nowadays) and if the sample size is large enough, SRSWR can be used as an approximation to the two-stage sampling design.

4

Screening the sample

The sample that is selected by the self-weighting two-stage sampling design described in Section 3 is not the sample that is used for data collection. The sample that was selected by the sampling design first undergoes a screening procedure, which eliminates some elements from the sample. This is mainly done to make sure the surveys of Statistics Netherlands are equally spread among the Dutch households and to ensure that inhabitants who cannot or presumably will not participate in the survey do not receive a request. In this section, we will discuss the different reasons to apply a screening procedure and the different steps of the screening procedure in detail.

The inhabitants that are removed from the sample due to the screening, can be grouped into three categories:

(i) Inhabitants whose information cannot be used due to confidentiality.

(ii) Inhabitants whose address already occurred in another sample of Statistics Netherlands in the past twelve months.

(iii) Inhabitants that are removed by other reasons.

We will refer to the inhabitants that are removed from the sample as inhabitants which are *not eligible after screening* and the inhabitants that are still in the sample after screening as *eligible after screening*. In the next sections, we will describe these categories in detail.

4.1. Confidential information

In Section 2.1 we discussed that Statistics Netherlands is in possession of a sampling frame that covers the whole Dutch population. This means that for every inhabitant of the Netherlands, information on how to contact that person is available.

Dutch inhabitants can request their municipality to not share their data in the Personal Records Database with any third parties. Consequently, the data of these inhabitants are not shared with organisations such as sports clubs, institutional health care organisations and charity organisations [11]. Although Statistics Netherlands is provided with data of these inhabitants by law [12], it has been

decided to do not approach these inhabitants for any of its surveys. This fulfils the desires of these inhabitants and moreover, it is assumed that those inhabitants will presumably not respond to any survey of Statistics Netherlands, so including them would not be beneficial.

Approximately 2% of the Dutch inhabitants have declared that their data in the Personal Records Database may not be shared with any third parties. After a sample is selected by the two-stage sampling design of Section 3, the most recent data from the Personal Records Database is used to see if there are any inhabitants in the sample that have declared that their data should not be shared with any third parties. Those inhabitants are then removed from the sample by the reason that we denote as *confidential information*. The inhabitants are then marked as not eligible after screening.

In Appendix C the distributions of several auxiliary variables for the population is plotted. Here a distinction is made based on whether inhabitants have a confidentiality indicator or not. As far as we can draw conclusions from these plots, the distribution is different for the two parts of the population. For example, for the auxiliary variable gender, we can conclude that relatively more women have declared that their data is confidential. It seems like there are some relationships between those auxiliary variables and the confidentiality of data.

4.2. Occurrence of an address

Statistics Netherlands wants to spread their surveys among the Dutch inhabitants as good as possible. To lower the response burden of surveys on the Dutch households, a screening procedure is applied to the sample. It is assumed that inhabitants, who live on the same address and who receive multiple requests to participate in a survey within a short period of time experience a high response burden. This makes it less likely that the inhabitants respond to the survey, which is not desirable.

To lower the response burden on the Dutch inhabitants, the screening procedure ensures that the sample after screening meets the following requirements:

(i) If two or more inhabitants who live on the same address are selected in the sample, then only one of them is eligible.

(ii) If the address of an inhabitant in the sample already occurs in another sample of Statistics Netherlands in the past twelve months[1], then the inhabitant is not eligible.

If an inhabitant is not eligible by one of these requirements, then we will denote it as not eligible by the *occurrence of an address*.

Remark 4.1. Applying a screening procedure to ensure that the sample meets these requirements is not the only regulation Statistics Netherlands made to lower the response burden. One example of a regulation is that the Dutch population is divided into different non-overlapping subpopulations. Each year a different subpopulation is considered for selecting samples. Consequently, inhabitants cannot be selected in a sample two years in a row. We assume that these regulations do not have any effects on the sampling design.

4.2.1. Inclusion probability after screening on the occurrence of an address

Let S^* denote the set-random variable of the sample of inhabitants that are eligible after screening. In this section, we will compute the inclusion probability of inhabitant k in S^*. The inclusion probability is computed under the following simplifying assumptions:

[1] Only self-weighting samples that are selected from the whole population are included.

(**A1**) Simple random sampling without replacement is applied to select the sample, instead of the sampling design of Section 3.

(**A2**) All samples of Statistics Netherlands for one year are selected at the same moment. This means that we will treat all different samples that are usually selected throughout the year as one big sample that is selected at the same moment.

(**A3**) Only screening on the occurrence of an address is applied.

By making these assumptions, we derive an approximation of the inclusion probability $\mathbb{P}\left(k \in S^*\right)$. By assuming (**A2**) we approximate the second requirement of the screening on the occurrence of an address by ensuring that if two or more inhabitants with the same address are selected in the sample, only one of them can be eligible. This means that within the sample of one year, no inhabitants can have the same address.

Let $U = \{1, 2, \ldots, N\}$ denote the target population of size N and let S denote the random variable of the sample that was selected from the population by SRSWR. Let n be the total sample size of all samples of Statistics Netherlands in one year. Under the assumptions (**A1**) - (**A3**), we will compute the inclusion probability of inhabitant k after the screening on the occurrence of an address, that is the probability $\mathbb{P}\left(k \in S^*\right)$.

The probability that inhabitant k is eligible after screening on the occurrence of an address is dependent on the number of people that live on the address of inhabitant k. Let $a_k \geq 1$ denote the number of people that have the same address as inhabitant k (including inhabitant k). This number is known for all Dutch inhabitants at Statistics Netherlands. The inclusion probability of inhabitant k in S^* is

$$\mathbb{P}\left(k \in S^*\right) = \mathbb{P}\left(k \in S^* \mid k \in S\right) \cdot \mathbb{P}\left(k \in S\right) \tag{4.1}$$

The event $\left\{k \in S^* \mid k \in S\right\}$ is dependent on the number of inhabitants in S that have the same address as inhabitant k, i.e.

$$\begin{aligned}\mathbb{P}\left(k \in S^* \mid k \in S\right) = \sum_{i=0}^{a_k-1} &\mathbb{P}\left(k \in S^* \mid k \in S \text{ with } i \text{ other inhabitants with same address}\right) \\ &\cdot \mathbb{P}\left(k \in S \text{ with } i \text{ other inhabitants with same address}\right)\end{aligned} \tag{4.2}$$

The conditional probability that inhabitant k is in the sample S together with exactly i $(i = 0, 1, 2, \ldots, a_k - 1)$ other inhabitants with the same address is

$$\mathbb{P}\left(k \in S \text{ with } i \text{ other inhabitants with same address}\right) = \binom{a_k - 1}{i}\binom{N - a_k}{n - i - 1}\frac{1}{\binom{N-1}{n-1}} \tag{4.3}$$

Given that k is in the sample with exactly i other inhabitants with the same address, the probability that k is in sample S^* is

$$\mathbb{P}\left(k \in S^* \mid k \in S \text{ with } i \text{ other inhabitants with same address}\right) = \frac{1}{i+1} \tag{4.4}$$

Consequently, the conditional probability in Equation (4.1) equals

$$\mathbb{P}\left(k \in S^* \mid k \in S\right) = \sum_{i=0}^{a_k-1} \frac{1}{i+1}\binom{a_k - 1}{i}\binom{N - a_k}{n - i - 1}\frac{1}{\binom{N-1}{n-1}} \tag{4.5}$$

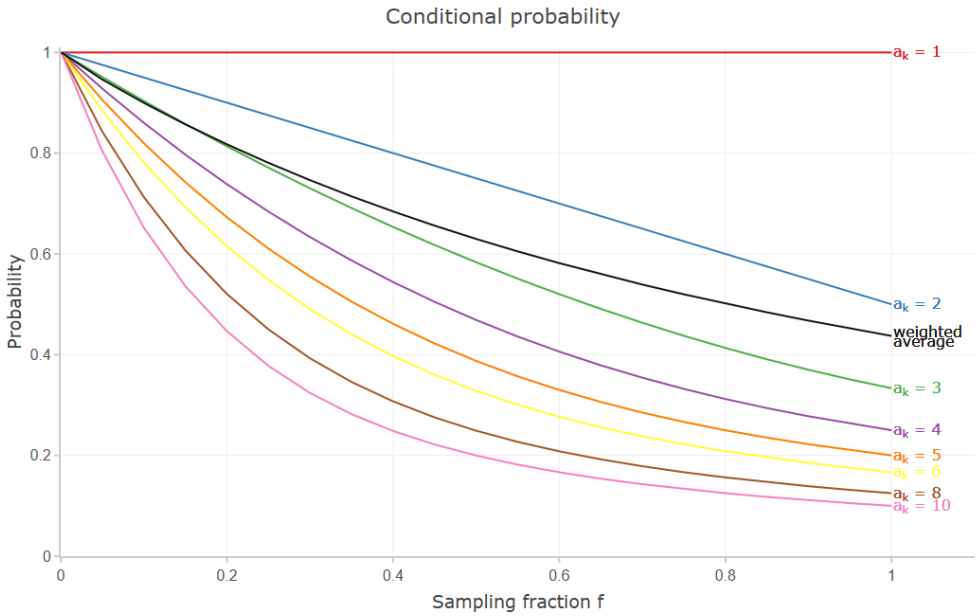


Figure 4.1: Plot of the conditional probability in Equation (4.5) as a function of the sampling fractions f and a_k (the number of people that live on the address of element k).

The conditional probability in Equation (4.5) is dependent on the population size N, the sample size n and the number of inhabitants on the address a_k. The probability is computed for different values of a_k and different values of the sampling fraction $f = \frac{n}{N}$, see Figure 4.1. Note that for a sampling fraction of $f = 1$, we have that the conditional probability equals $\frac{1}{a_k}$. The figure illustrates that the more inhabitants live on the same address, the smaller the conditional probability is, which matches the intuitive expectations.

Furthermore, Figure 4.1 illustrates that the effects of the screening procedure become larger as the sampling fraction increases. For a larger sampling fraction, the difference between the conditional probability for the different values of a_k are larger.

Using Equation (4.5), the inclusion probability of element k in S^* is

$$\begin{aligned}\mathbb{P}\left(k \in S^*\right) &= \mathbb{P}\left(k \in S^* \mid k \in S\right) \cdot \mathbb{P}\left(k \in S\right) \\ &= \sum_{i=0}^{a_k-1} \frac{1}{i+1} \frac{1}{\binom{N-1}{n-1}} \binom{a_k-1}{i} \binom{N-a_k}{n-i-1} \frac{n}{N} \\ &= \sum_{i=0}^{a_k-1} \frac{1}{i+1} \frac{1}{\binom{N}{n}} \binom{a_k-1}{i} \binom{N-a_k}{n-i-1}\end{aligned} \tag{4.6}$$

The inclusion probability after screening on the occurrence of an address is plotted in Figure 4.2 for different values of a_k and for different sampling fractions. Note that in case $a_k = 1$, the probability in Equation (4.6) equals the first-order inclusion probability of simple random sampling without replacement.

In Figures 4.1 and 4.2 the weighted average of the probabilities is plotted. This line represents the probability for an average inhabitant of the Netherlands. This probability can be used to obtain

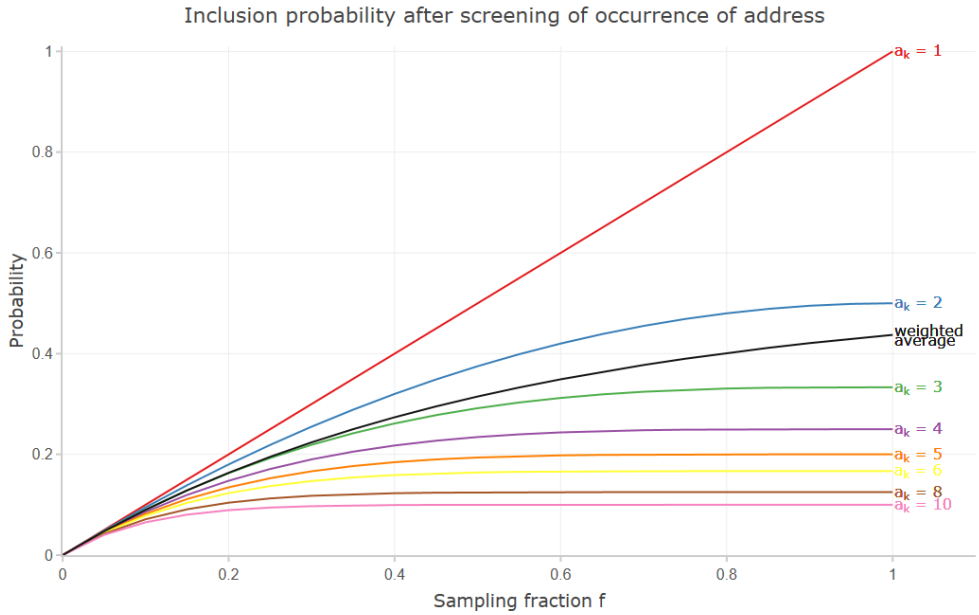


Figure 4.2: Plot of the inclusion probability in Equation (4.6) for different sampling fractions f and values of a_k (the number of people that live on the address of element k).

an estimation for the amount of people that become not eligible by screening on the occurrence of an address.

Example 4.1 (The Dutch Crime Victimisation Survey)**.** One survey of Statistics Netherlands is the The Dutch Crime Victimisation Survey[2]. The aim of the survey is to estimate the safety perception, victimisation and criminality among the Dutch inhabitants [13]. Municipalities and regional police units have the possibility to participate in the safety monitor to obtain more data from their own regions [13]. As a consequence sampling fraction are different among regions. The selected samples of these regions are then screened on the occurrence of an address. In Figure 4.3 the results of the screening on the occurrence of an address is shown for the different regions. The probability of Equation (4.5) is plotted for comparison. Note that in general the realisations follow the theoretical probability quite well. However, there are some outliers, which can be explained by the distribution of number of people on an address in each region. The theoretical probability is based on the distribution of number of people on an address in the whole population, but this distribution may be different in some regions.

4.3. Other reasons

Inhabitants can be removed from the sample by confidential information or another inhabitant with the same address is already selected in the same or another sample in the past twelve months. Furthermore, inhabitants who meet the following conditions are not eligible after screening:

(i) Inhabitants who passed away or emigrated recently

(ii) Inhabitants that belong to an institutional household

[2]In Dutch better known as the Veiligheidsmonitor.

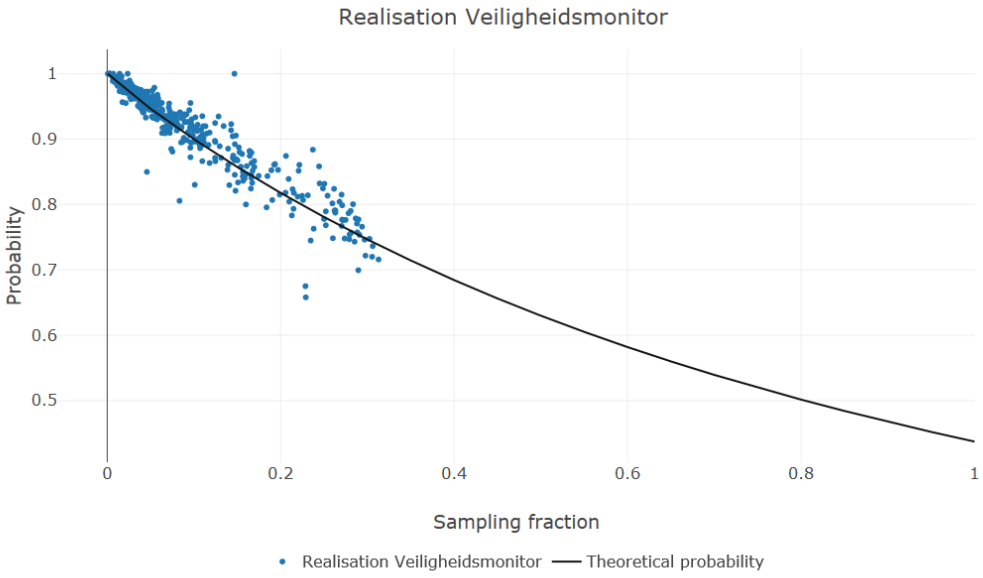


Figure 4.3: Results of the screening on the occurrence of an address for the samples of the Dutch crime victimisation survey in the different regions. Every blue dot represents a region and the probability denotes the percentage of inhabitants that are still egligible after screening. The theoretical probability is the probability of Equation (4.5) for an average Dutch inhabitant

(iii) Inhabitants whose name and address details cannot be obtained for other reasons, because for example contact information is missing.

The main reason those inhabitants are not included in the sample after screening is that those inhabitants cannot or presumably will not participate in the survey.

4.4. Screening in 2018

The screening procedure is applied to all self-weighting samples that are selected from the whole population. In 2018, approximately 92% of the inhabitants that were included in the sample before screening are eligible after screening, see Figure 4.4. In other words, in 2018 approximately 8% of the inhabitants that were included in the sample are not eligible after screening. Approximately 3 out of 4 inhabitants that are not eligible after screening, are not eligible by the occurrence of an address, see Figure 4.5. Almost all other non-eligible inhabitants are not eligible by confidentiality. Note that only a negligible amount of inhabitants are removed from the sample by other reasons.

The amount of inhabitants that become non-eligible by the screening procedure has increased over the last years, as the sample sizes have increased. By contrast, in 2007 approximately 2% of the inhabitants became not eligible by the screening procedure, whereas this is approximately 8% nowadays.

4.5. Conclusion

Inhabitants that are selected in the sample before screening can become not eligible after the screening by three different reasons. We have seen that most non-eligible inhabitants are not eligible because of screening on the occurrence of an address.

From this section we can conclude that one should be cautious with assuming that inhabitants have an equal probability to be eligible after screening. The probability that an inhabitant is eligible after

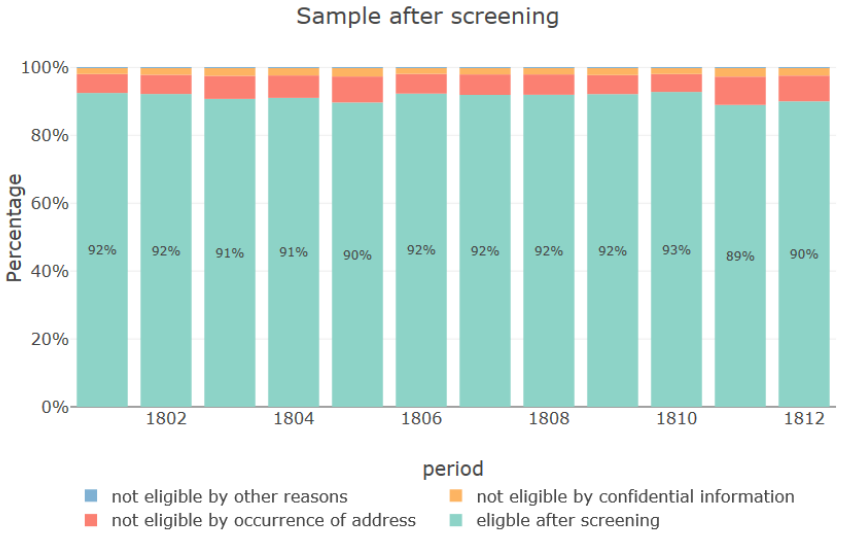


Figure 4.4: Representation of the sample after screening per period. The eligible and non-eligible inhabitants by different reasons of the screening procedure are plotted for the samples of all months of 2018.

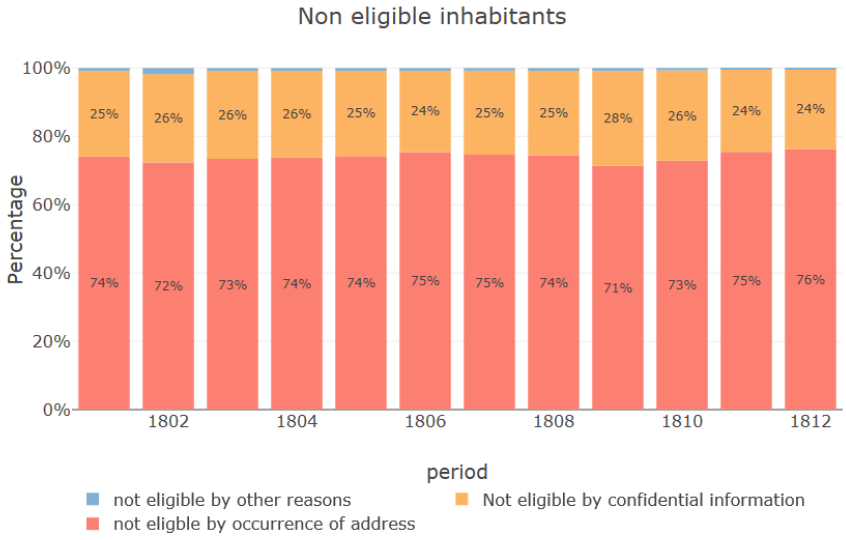


Figure 4.5: Representation of the non-eligible inhabitants for all months in 2018. Approximately 75% of the non-eligible inhabitants are not eligible by the screening on occurrence of an address.

screening on the occurrence of an address is strongly dependent on the number of people living on the same address, see Equation (4.5). Furthermore, we have seen that the approximated probability is dependent on the sampling fraction. As the sampling fraction increases, the effects of the screening on the occurrence of an address become larger. Moreover, we have seen that the distribution of auxiliary variables is different for inhabitants with a confidentiality indicator than for inhabitants with no confidentiality indicator. It seems that there is a relation between the confidentiality indicator and those auxiliary variables. This means that during the screening procedure not every inhabitant has the same

probability of becoming not eligible by confidentiality.

Furthermore, we have seen that the effects of the screening become larger as the sample size increases.

5

Statistical testing

To identify the effects of the screening procedure (see Section 4) on the sample, the distribution of auxiliary variables in the population and the sample after screening are compared. For each auxiliary variable, we determine if the sample after screening is representative for the population.

A sample is said to be *representative* with respect to a variable if its relative distribution in the sample is equal to its relative distribution on the population [3]. For example, a sample is representative with respect to gender if the percentages of males and females in the sample are in expectation equal to the percentages of males and females in the population.

The intention of selecting a self-weighting sample is that the selected sample is representative for the population with respect to all relevant auxiliary variables. In Section 4 we have seen that the inclusion probability in the sample after screening is not equal for all inhabitants. The question arises whether the sample after screening is representative for the population with respect to relevant auxiliary variables. In addition to comparing the distributions of the sample after screening with the population, we determine if the sample after screening is representative for the sample before screening with respect to the relevant auxiliary variables.

In this section, we discuss two statistical tests, one to determine whether the sample after screening is representative for the population and one to determine if the sample after screening is representative for the sample before screening, with respect to a given auxiliary variable. Those tests are applied to several auxiliary variables, which are available for the whole population at Statistics Netherlands.

5.1. The multivariate hypergeometric distribution

Suppose we observe a categorical auxiliary variable with K categories $\mathcal{C}_1, \ldots, \mathcal{C}_K$. Let the population U have size N and let $c_1, \ldots, c_K$ denote the number of inhabitants in categories $\mathcal{C}_1, \ldots, \mathcal{C}_K$ respectively, so that

$$c_1 + \cdots + c_K = N \tag{5.1}$$

So c_j denotes the number of inhabitants beloning to category $\mathcal{C}_j$ $(j = 1, \ldots, K)$. The vector of inhabitants per category in the population for the observed auxiliary variable is known at Statistics Netherlands.

Let S_{SRSWR} be the set-random variable denoting the sample of size n that is selected from the population U by simple random sampling without replacement, see Section 2.1.1. Let $\mathbf{P} = (P_1, \dots, P_K)$ denote the random number of inhabitants in each category in the sample such that

$$P_1 + \cdots + P_K = n \quad \text{and} \quad P_j \leq c_j \; \forall \; j = 1, \dots, K \tag{5.2}$$

Let $\mathbf{p} = (p_1, \dots, p_K)$ denote a realisation of the vector $\mathbf{P}$ for a given sample s_{SRSWR}. Then the joint distribution of $\mathbf{P} = (P_1, \dots, P_K)$ has probability mass function

$$\mathbb{P}(\mathbf{P} = \mathbf{p}) = \mathbb{P}\left(\bigcap_{j=1}^{K} \left\{P_j = p_j\right\}\right) = \frac{\prod_{j=1}^{K} \binom{c_j}{p_j}}{\binom{N}{n}} \tag{5.3}$$

The distribution in Equation (5.3) is called a *multivariate hypergeometric distribution* with parameters $(n, c_1, \dots, c_K)$ [14]. We will denote this distribution by

$$\mathbf{P} = (P_1, \dots, P_K) \sim \text{Mult. Hypgeom.}\,(n, c_1, \dots, c_K) \tag{5.4}$$

The marginal distribution of P_j is univariate hypergeometric with paramaters $\left(n, c_j, N\right)$ [14], that is

$$\mathbb{P}\left(P_j = p_j\right) = \frac{\binom{c_j}{p_j}\binom{N-c_j}{n-p_j}}{\binom{N}{n}} \tag{5.5}$$

The vector of expected values of $\mathbf{P}$ is

$$\mathbb{E}(\mathbf{P}) = \left(\frac{n}{N}c_1, \dots, \frac{n}{N}c_K\right) \tag{5.6}$$

The variance of P_j is given by [14]

$$\begin{aligned} \mathbb{V}\left(P_j\right) &= \frac{n(n-1)}{N(N-1)}c_j\left(c_j - 1\right) + \frac{n}{N}c_j - \left(\frac{n}{N}c_j\right)^2 \\ &= \frac{n(N-n)}{N(N-1)}c_j\left(N - c_j\right) \end{aligned} \tag{5.7}$$

Let the sample S of size n denote the sample that is selected from the population by the sampling design that was described in Section 3. After the screening process described in Section 4 we have a sample S^* of size n^* containing only eligible elements, such that $S^* \subseteq S$ and $n^* \leq n$. Let the vector $\mathbf{M} = (M_1, \dots, M_K)$ denote the random vector of number of inhabitants per category in the sample S and let $\mathbf{M}^* = \left(M_1^*, \dots, M_K^*\right)$ denote the random vector of number of inhabitants per category in the sample after screening S^*, so that

$$M_1 + \cdots + M_K = n \quad \text{and} \quad M_1^* + \cdots + M_K^* = n^* \tag{5.8}$$

If the process of screening the sample is a random process with equal probabilities, the sample after screening S^* should be representative for both the population U and the sample S with respect to all relevant auxiliary variables. We formulate two different hypotheses:

Hypothesis 1 The sample after screening S^* is representative for the population U with respect to a given auxiliary variable, i.e. $\mathbf{M}^*$ follows a multivariate hypergeometric distribution with parameters $\left(n^*, c_1, \dots, c_K\right)$.

Hypothesis 2 The sample after screening S^* is representative for the sample before screening S with respect to a given auxiliary variable, i.e. $\mathbf{M}^*|\mathbf{M}$ follows a multivariate hypergeometric distribution with parameters $\left(n^*, M_1, \ldots, M_K\right)$.

We define two different statistical tests for these hypotheses in the next two sections.

5.2. Testing hypothesis 1

The first question that arises is whether the distribution of a given auxiliary variables in the sample after screening S^* is similar to the distribution of the auxiliary variable in the population U. Define the null- and alternative hypothesis as

$$\begin{aligned} H_0 &: \mathbf{M}^* \sim \text{Mult. Hypgeom.}\left(n^*, c_1, \ldots, c_K\right) \\ H_a &: \mathbf{M}^* \nsim \text{Mult. Hypgeom.}\left(n^*, c_1, \ldots, c_K\right) \end{aligned} \tag{5.9}$$

Note that since the observed auxiliary variable is known for the whole population, the vector $(c_1, \ldots, c_K)$ is given.

Under the null hypothesis, the distribution of the auxiliary variable in the sample after screening is a realisation from the multivariate hypergeometric distribution with parameters $\left(n^*, c_1, \ldots, c_K\right)$. Thus, we can compare the expected values for the number of inhabitants per category in the sample after screening with the observed values, i.e. we compare

$$\left\| \mathbb{E}\left(\mathbf{M}^*\right) - \mathbf{m}^* \right\|^2 \tag{5.10}$$

where $\mathbf{m}^* = \left(m_1^*, \ldots, m_K^*\right)$ represents the realisation of the vector of inhabitants per category in the sample after screening. The expected value is known under the null-hypothesis, namely

$$\mathbb{E}\left(\mathbf{M}^*\right) = \left(\frac{n^*}{N}c_1, \ldots, \frac{n^*}{N}c_K\right) \tag{5.11}$$

The squared difference between the observed and expected values is then equal to

$$\sum_{j=1}^{K}\left(\mathbb{E}\left(M_j^*\right) - m_j^*\right)^2 = \sum_{j=1}^{K}\left(\frac{n^*}{N}c_j - m_j^*\right)^2 = \left(n^*\right)^2 \sum_{j=1}^{K}\left(\frac{c_j}{N} - \frac{m_j^*}{n^*}\right)^2 \tag{5.12}$$

We define the test statistic for this test as

$$T_1 = \sum_{j=1}^{K}\left(\frac{c_j}{N} - \frac{m_j^*}{n^*}\right)^2 \tag{5.13}$$

Let t_1^* denote the value of this test statistic for the observed sample after screening. The distribution of T_1 is unknown, so it is not possible to compute the probability $\mathbb{P}\left(T_1 \geq t^*\right)$ directly. Using *parametric bootstrap* we can find an approximate value for this probability. Parametric bootstrapping is a technique for obtaining estimates of the properties of statistical estimators by assuming the underlying data is distributed as some specified parametric distribution [15]. The main idea of this test is to independently draw R realisations from the distribution Mult. Hypgeom. $\left(n^*, c_1, \ldots, c_K\right)$. For each realisation, we can compute the value of the test statistic T_1. Let the values of the test statistic for these samples be denoted by the R-vector $\left(T_{1,1}, \ldots, T_{1,R}\right)$. Then the $\mathbf{p}$-value corresponding to this test can be approximated by

counting the number of times that $T_{1,i}$ is greater than or equal to t_1^*, for $i = 1, \dots, R$, that is

$$\mathbf{p} = \mathbb{P}\left(T_1 \geq t_1^*\right) \approx \frac{1}{R} \sum_{i=1}^{R} \mathbb{1}\left\{T_{1,i} \geq t_1^*\right\} \tag{5.14}$$

The procedure that should be followed to apply this test is summarised in Algorithm 5.1.

Algorithm 5.1: Procedure that describes the parametric bootstrap approach to test the null-hypothesis that the vector of the number of inhabitants per category in the sample after screening is selected from the multivariate hypergeometric distribution with parameters $\left(n^*, c_1, \dots, c_K\right)$.

STATISTICAL TEST 1

1. Derive $(c_1, \dots, c_K)$ from U
2. Compute $t^* = \sum_{j=1}^{K} \left(\frac{c_j}{N} - \frac{m_j^*}{n^*}\right)^2$ for the population U and the sample after screening s^*
3. **For** i from 1 to R
4. Draw $\mathbf{M}^* \sim$ Mult. Hypgeom. $\left(n^*, c_1, \dots, c_K\right)$
5. Compute $T_{1,i} = \sum_{j=1}^{K} \left(\frac{c_j}{N} - \frac{m_j^*}{n^*}\right)^2$
6. Compute $\mathbf{p} \approx \frac{1}{R} \sum_{i=1}^{R} \mathbb{1}\left\{T_{1,i} \geq t^*\right\}$

5.3. Testing hypothesis 2

The second hypothesis is that the distribution of a given auxiliary variable in the sample after screening is similar to the distribution of the auxiliary variable in the sample before screening. Recall that $\mathbf{M}$ denotes the random vector of number of inhabitants per category in the sample before screening S and $\mathbf{M}^*$ denotes the random vector of number of inhabitants per category in the sample after screening S^*. Define the null- and alternative hypothesis for this test as

$$\begin{aligned} H_0 &: \mathbf{M}^* | \mathbf{M} \sim \text{Mult. Hypgeom.}\left(n^*, M_1, \dots, M_K\right) \\ H_a &: \mathbf{M}^* | \mathbf{M} \nsim \text{Mult. Hypgeom.}\left(n^*, M_1, \dots, M_K\right) \end{aligned} \tag{5.15}$$

Note that $\mathbf{M}$ is a random vector and is therefore not known in advance.

Under the null-hypothesis the vector $\mathbf{M}^*$ given the vector $\mathbf{M}$ is a realisation from the multivariate hypergeometric distribution with parameters $\left(n^*, M_1, \dots, M_K\right)$. We compare the expected values for the number of inhabitants per category in S^* given the vector $\mathbf{M}$ with the observed vector of inhabitants per category in the sample after screening S^*, i.e.

$$\begin{aligned} \left\| \mathbb{E}\left(\mathbf{M}^* | \mathbf{M}\right) - \mathbf{m}^* \right\|^2 &= \sum_{j=1}^{K} \left(\mathbb{E}\left(M_j^* | \mathbf{M}\right) - m_j^* \right)^2 = \sum_{j=1}^{K} \left(\frac{n^*}{n} M_j - m_j^* \right)^2 \\ &= \left(n^*\right)^2 \sum_{j=1}^{K} \left(\frac{M_j}{n} - \frac{m_j^*}{n^*} \right)^2 \end{aligned} \tag{5.16}$$

where $\mathbf{m}^* = \left(m_1^*, \ldots, m_K^*\right)$ represents the observed realisation of the vector $\mathbf{M}^*$. Define the test statistic for this test as

$$T_2 = \sum_{j=1}^{K} \left(\frac{M_j}{n} - \frac{m_j^*}{n^*} \right)^2 \tag{5.17}$$

Let t_2^* denote the value of this test statistic for the observed sample before screening and sample after screening. The distribution of T_2 is unknown, so we need to approximate the $\mathbf{p}$-value for this test using a parametric bootstrap approach.

The vector $\mathbf{M}$ is a random vector that can be obtained from the sample S after the sampling design is applied. Consequently, the vector $\mathbf{M}$ is not known in advance and is different for each different sample S.

To approach this test using bootstrap, we randomly draw the sample S from the sampling design from Section 3 R times. Then we obtain R different vectors $\mathbf{M} = (M_1, \ldots, M_K)$. For each of such a vector $\mathbf{M}$ we can draw vector $\mathbf{M}^*$ from the multivariate hypergeometric distribution with parameters $\left(n^*, M_1, \ldots, M_K\right)$. So we obtain R different vectors $\mathbf{M}^*$. For each pair of vectors $\mathbf{M}$ and $\mathbf{M}^*$ the test statistic is computed. The procedure that is followed to apply this test is summarised in Algorithm 5.2.

Algorithm 5.2: Procedure that describes a parametric bootstrap approach to test the null-hypothesis that the vector of the number of inhabitants per category in the sample after screening is selected from the multivariate hypergeometric distribution with parameters $\left(n^*, M_1, \ldots, M_K\right)$.

STATISTICAL TEST 2

1. Compute $t^* = \sum_{j=1}^{K} \left(\frac{M_j}{n} - \frac{M_j^*}{n^*} \right)^2$ from the observed samples s and s^*
2. **For** i from 1 to R
3. Draw sample S from U according to the sampling design
4. Derive $\mathbf{M} = (M_1, \ldots, M_K)$ from S
5. Draw $\mathbf{M}^* | \mathbf{M} \sim$ Mult. Hypgeom. $\left(n^*, M_1, \ldots, M_K\right)$
6. Compute $T_{2,i} = \sum_{j=1}^{K} \left(\frac{M_j}{n} - \frac{M_j^*}{n^*} \right)^2$
7. Compute $\mathbf{p} \approx \frac{1}{R} \sum_{i=1}^{R} \mathbb{1}\left\{T_{2,i} \geq t^*\right\}$

5.4. Results

The statistical tests for Hypothesis 1 and Hypothesis 2 are applied to the sample of the mobility survey of April 2019. This sample was selected from the population by the self-weighting two-stage sampling design of Section 3. Subsequently, this sample was screened according to the procedure that was described in Section 4.

The sample size of the mobility survey from April 2019 is 8427. The target population for the mobility survey is all Dutch inhabitants that are 6 years or older. In Table 5.3 it is shown how many elements are eligible and not eligible after screening. For example, 449 of the 8427 elements in the sample are not eligible by the screening on the occurrence of an address. The main goal of the test corresponding to hypothesis 1 is to determine whether the sample after screening is representative for the population with

Eligible after screening	7807	92.6%
Not eligible after screening	620	7.4%
Not eligible by occurrence of address	449	5.3%
Not eligible by confidential information	151	1.8%
Not eligible by other reason	20	0.2%
Total sample	8427	100 %

Table 5.3: Sample of the mobility survey of April 2019.

respect to relevant auxiliary variables. The aim of the second test is to determine whether the sample after screening is representative for the sample before screening with respect to relevant auxiliary variables. Furthermore, it is also interesting to see if the sample containing inhabitants that are not eligible after screening is representative for the population and the sample before screening with respect to those auxiliary variables. Therefore the tests are not only applied to the sample after screening but also to:

(i) The sample after screening (only eligible inhabitants)

(ii) The non-eligible inhabitants by screening on the occurrence of an address

(iii) The non-eligible inhabitants by screening on confidential information

(iv) The non-eligible inhabitants (all reasons of screening)

Furthermore, the first test is also applied to the sample before screening. The sample before screening was selected by the two-stage self-weighting sampling design, which means that every inhabitant has an equal inclusion probability. Consequently, the sample before screening should be representative for the population with respect to all auxiliary variables.

The auxiliary variables that are used for the tests are gender, marital status, age, ethnicity, place in household, type of household, number of people in household and number of people on address. These are all categorical variables and the definitions for the categories can be found in Appendix B. In Appendix D the distributions of the auxiliary variables for the different samples of the mobility survey of April 2019 are plotted. This allows the reader to compare the distribution of an auxiliary variable in the population with the distribution in the sample after screening. In this Appendix we have plotted for all auxiliary variables the distribution in the population against the sample before screening, the sample after screening, the non-eligible inhabitants by occurrence of address, the non-eligible inhabitants by confidential information and the non-eligible inhabitants. Furthermore, we have plotted the sample before screening against the sample after screening, the non-eligible inhabitants by occurrence of address, the non-eligible inhabitants by confidential information and the non-eligible inhabitants. The results of the two tests applied to the mobility survey of April 2019 are presented in Table 5.4 and Table 5.5.

We reject the null-hypothesis for small values of the $\mathbf{p}$-value. First of all note that we can conclude that for all auxiliary variables the sample before screening is representative for the population. This is as expected, because the sampling design is self-weighting. For each auxiliary variable, we briefly discuss on the results.

	Hypothesis 1		Hypothesis 2	
	Test statistic	**p**-value	Test statistic	**p**-value
Gender				
Sample before screening	$6.418 \cdot 10^{-5}$	0.29442	–	–
Sample after screening	$1.047 \cdot 10^{-4}$	0.19826	$4.940 \cdot 10^{-6}$	0.30234
Not eligible by occurrence of address	$5.604 \cdot 10^{-3}$	0.02340	$6.868 \cdot 10^{-3}$	0.01086
Not eligible by confidential information	$1.993 \cdot 10^{-2}$	0.01094	$1.773 \cdot 10^{-2}$	0.01906
Not eligible	$3.984 \cdot 10^{-4}$	0.46881	$7.824 \cdot 10^{-4}$	0.30600
Marital status				
Sample before screening	$2.011 \cdot 10^{-5}$	0.77950	–	–
Sample after screening	$1.922 \cdot 10^{-4}$	0.08414	$1.105 \cdot 10^{-4}$	$< 1 \cdot 10^{-5}$
Not eligible by occurrence of address	$2.302 \cdot 10^{-2}$	$< 1 \cdot 10^{-5}$	$2.376 \cdot 10^{-2}$	$< 1 \cdot 10^{-5}$
Not eligible by confidential information	$4.211 \cdot 10^{-2}$	0.00015	$4.282 \cdot 10^{-2}$	0.00013
Not eligible	$1.674 \cdot 10^{-2}$	$< 1 \cdot 10^{-5}$	$1.745 \cdot 10^{-2}$	$< 1 \cdot 10^{-5}$
Age				
Sample before screening	$1.151 \cdot 10^{-4}$	0.33673	–	–
Sample after screening	$2.183 \cdot 10^{-4}$	0.05405	$6.086 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$
Not eligible by occurrence of address	$1.824 \cdot 10^{-2}$	$< 1 \cdot 10^{-5}$	$1.857 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$
Not eligible by confidential information	$9.462 \cdot 10^{-3}$	0.11356	$1.031 \cdot 10^{-2}$	0.07757
Not eligible	$9.221 \cdot 10^{-3}$	$< 1 \cdot 10^{-5}$	$9.639 \cdot 10^{-3}$	$< 1 \cdot 10^{-5}$
Ethnicity				
Sample before screening	$2.370 \cdot 10^{-5}$	0.57609	–	–
Sample after screening	$8.182 \cdot 10^{-5}$	0.18744	$1.751 \cdot 10^{-5}$	0.01226
Not eligible by occurrence of address	$9.794 \cdot 10^{-4}$	0.30826	$1.308 \cdot 10^{-3}$	0.19884
Not eligible by confidential information	$7.448 \cdot 10^{-3}$	0.06097	$8.311 \cdot 10^{-3}$	0.04302
Not eligible	$2.286 \cdot 10^{-3}$	0.03238	$2.773 \cdot 10^{-3}$	0.01351
Place in household				
Sample before screening	$5.284 \cdot 10^{-5}$	0.74249	–	–
Sample after screening	$2.313 \cdot 10^{-4}$	0.05489	$1.145 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$
Not eligible by occurrence of address	$3.956 \cdot 10^{-2}$	$< 1 \cdot 10^{-5}$	$3.987 \cdot 10^{-2}$	$< 1 \cdot 10^{-5}$
Not eligible by confidential information	$2.684 \cdot 10^{-2}$	0.00034	$2.872 \cdot 10^{-2}$	0.00014
Not eligible	$1.739 \cdot 10^{-2}$	$< 1 \cdot 10^{-5}$	$1.814 \cdot 10^{-2}$	$< 1 \cdot 10^{-5}$
Type of household				
Sample before screening	$4.636 \cdot 10^{-5}$	0.74441	–	–
Sample after screening	$1.982 \cdot 10^{-4}$	0.08801	$9.490 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$
Not eligible by occurrence of address	$4.399 \cdot 10^{-2}$	$< 1 \cdot 10^{-5}$	$4.399 \cdot 10^{-2}$	$< 1 \cdot 10^{-5}$
Not eligible by confidential information	$5.123 \cdot 10^{-2}$	$< 1 \cdot 10^{-5}$	$5.364 \cdot 10^{-2}$	$< 1 \cdot 10^{-5}$
Not eligible	$1.436 \cdot 10^{-2}$	$< 1 \cdot 10^{-5}$	$1.503 \cdot 10^{-2}$	$< 1 \cdot 10^{-5}$

Table 5.4: Results of the statistical test for hypothesis 1 and hypothesis 2 ($R = 100.000$) for the auxiliary variables: gender, marital status, age, ethnicity, place in household and type of household.

	Hypothesis 1		Hypothesis 2	
	Test statistic	**p**-value	Test statistic	**p**-value
Number of people in household				
Sample before screening	$7.820 \cdot 10^{-5}$	0.48276	–	–
Sample after screening	$2.828 \cdot 10^{-4}$	0.02458	$1.259 \cdot 10^{-4}$	$< 1 \cdot 10^{-5}$
Not eligible by occurrence of address	$3.984 \cdot 10^{-2}$	$< 1 \cdot 10^{-5}$	$4.074 \cdot 10^{-2}$	$< 1 \cdot 10^{-5}$
Not eligible by confidential information	$1.265 \cdot 10^{-2}$	0.04658	$1.331 \cdot 10^{-2}$	0.03533
Not eligible	$1.902 \cdot 10^{-2}$	$< 1 \cdot 10^{-5}$	$1.994 \cdot 10^{-2}$	$< 1 \cdot 10^{-5}$
Number of people on address				
Sample before screening	$7.663 \cdot 10^{-5}$	0.50860	–	–
Sample after screening	$3.313 \cdot 10^{-4}$	0.01107	$1.779 \cdot 10^{-4}$	$< 1 \cdot 10^{-5}$
Not eligible by occurrence of address	$6.015 \cdot 10^{-2}$	$< 1 \cdot 10^{-5}$	$6.125 \cdot 10^{-2}$	$< 1 \cdot 10^{-5}$
Not eligible by confidential information	$1.040 \cdot 10^{-2}$	0.08949	$1.092 \cdot 10^{-2}$	0.07141
Not eligible	$2.728 \cdot 10^{-2}$	$< 1 \cdot 10^{-5}$	$2.817 \cdot 10^{-2}$	$< 1 \cdot 10^{-5}$

Table 5.5: Results of the statistical test for hypothesis 1 and hypothesis 2 ($R = 100.000$) for the auxiliary variables: number of people in household and number of people on address.

Gender For the auxiliary variable gender we cannot reject both null-hypotheses. This means that for the variable gender, the sample after screening can be considered representative for the population and for the sample before screening. The inhabitants that are not eligible can also be considered representative for the population and for the sample before screening. However, the inhabitants that are not eligible by screening on the occurrence of an address and by screening on confidential information are not representative for the population. This is remarkable, but can be explained by looking at the plots in Appendix D. Apparently, for this specific sample men are deleted more often by occurrence of address while women are more often non-eligible by confidential information.

Marital status For a significance level of $\alpha = 0.05$ we cannot reject the null-hypothesis that the sample after screening is representative for the population. However, for a significance level of $\alpha = 0.1$ we could reject this hypothesis. Furthermore we can conclude that with respect to marital status the sample after screening is not representative for the sample before screening. When looking at the plots, it seems that unmarried inhabitants become non-eligible by the screening procedure more often than other inhabitants. No parts of the sample can be considered representative for the sample before screening.

Age For a significance level of $\alpha = 0.1$ we can reject that the sample after screening is representative for the population with respect to age. However, for a significance level of $\alpha = 0.05$ we cannot reject this. We can reject the hypothesis that the sample after screening is representative for the sample before screening. Also the non-eligible inhabitants cannot be considered representative for the population or the sample before screening, yet the non-eligible inhabitants by confidential information can be considered representative for both the population and the sample before screening. From the plots in Appendix D we can conclude that the young inhabitants (under age 25) become non-eligible by screening on the occurrence of an address relatively often. Apparently inhabitants from age 50 to 64 have a confidentiality indicator relatively often.

Ethnicity For ethnicity, we can only reject the hypothesis that the non-eligible inhabitants are repre-

sentative for the population. So with respect to ethnicity we can consider the sample after screening representative for the population, but the sample after screening cannot be considered representative for the sample before screening. From the plots we can conclude that native Dutch inhabitants become not eligible by the screening procedure less often than non-native Dutch habitants.

Place in Household For a significance level of $\alpha = 0.1$ we can reject all hypotheses. But although the $\mathbf{p}$-value is low, for a significance level of $\alpha = 0.05$ we cannot reject the hypothesis that the sample after screening is representative for the population. When looking at the plots in Appendix D we can conclude that children (Category 1) or partners with children (Category 6) become non-eligible by the screening on the occurrence of an address relatively often, which seems understandable since those inhabitants presumably live with multiple people on one address.

Type of Household The results with respect to the variable type of household are similar to the results with respect to the variable place in household. We can almost reject all hypothesis, except for the hypothesis that the sample after screening is representative for the population. From the plots we can conclude that inhabitants in a household consisting of a married couple with children become non-eligible by screening on the occurrence of an address relatively often. Furthermore, it seems that inhabitants in single-person households and single-parent household become not eligible by confidentiality relatively often.

Number of people in household Based on the variable number of people in household, we can conclude that the sample after screening is neither representative for the population, nor for the sample before screening. For the screening on occurrence of address this seems reasonable, since there is presumably a highly correlation between the number of people in a household and the number of people on an address. What is remarkable is that inhabitants with a 1-person household become non-eligible by confidential information relatively often.

Number of people on address This variable probably has the most obvious results, since we showed in Chapter 4 that the screening on occurrence of address is directly dependent on the number of people on an address.

For the sample of the mobility survey of April 2019, we have discussed the results of the two statistical tests in detail. The results for other samples may be different because of sampling fluctuations. In Appendix E the results for testing hypothesis 1 for different periods are presented. Performing the second test is computationally intensive, so we did not perform those tests for the other samples.

From these results we see that for each realisation, there is at least one auxiliary variable for which we can reject the null-hypothesis. This means that in general the sample after screening cannot be considered as representative for the population. Furthermore, we can conclude that the elements that are not eligible by screening are not representative for the population with respect to all auxiliary variables, except for gender. This holds for all realisations we have seen.

5.5. Conclusion

In general we can conclude that the inhabitants that become not eligible by the screening are not representative for the population. For some samples, this may cause that the sample after screening is not representative for the population, but for some samples the sample after screening can still be considered as representative for the population. For the sample of the mobility survey of April 2019, we have seen that for all auxiliary variables except gender, the sample after screening is not representative for the sample before screening.

6

Simulation study

In Section 5 two different statistical tests were applied to samples from the mobility survey. Recall from Section 4 that the effects of the screening become larger as the sample size increases. In this section we investigate the consequences of the screening procedure if the total sample size of all samples in one year becomes larger. In the current situation the amount of inhabitants that are selected in a self-weighting sample that is selected from the total Dutch population is approximately 500.000, see Figure 6.1. We will suppose an extra survey is selected every month with a sample size of 30.000 inhabitants each month. This would mean that in one year approximately 5% of the Dutch inhabitants will be selected in a self-weighting sample of Statistics Netherlands. This situation is comparable with the situation that is likely to happen at Statistics Netherlands, where the Labour Force Survey (LFS)[1] becomes a self-weighting survey sample that is selected from the whole Dutch population. Nowadays, the Labour Force Survey has a different sampling design, so it is not included in the screening procedure. Note from Figure 6.1 that there is a positive trend in the sampling fraction of the sample size of all samples in the previous year.

Because the sample size is different every period, we are not able to rely on the data of the of survey samples that are selected by Statistics Netherlands anymore. Therefore a simulation study is needed. We will simulate the current situation (case A) and the situation where sample sizes are increased (case B).

6.1. Simulating the current situation

To reproduce a similar situation as in Section 5, we need all samples that are selected from the population from the previous year. First, note that the population is continuously changing, since inhabitants are born, inhabitants pass away or immigrate or emigrate. Once a month, we have access to the most recent population at that time. This means that for each month, a dataset containing the inhabitants in the Dutch population at that time is available.

Furthermore, note that each sample that is selected is screened with respect to all samples that were selected in the past twelve months. This is impossible to reproduce, since this requires selecting samples for multiple years in the past. Therefore we try to approximate the screening procedure as

[1] In Dutch known as the *Enquête Beroepsbevolking (EBB).*

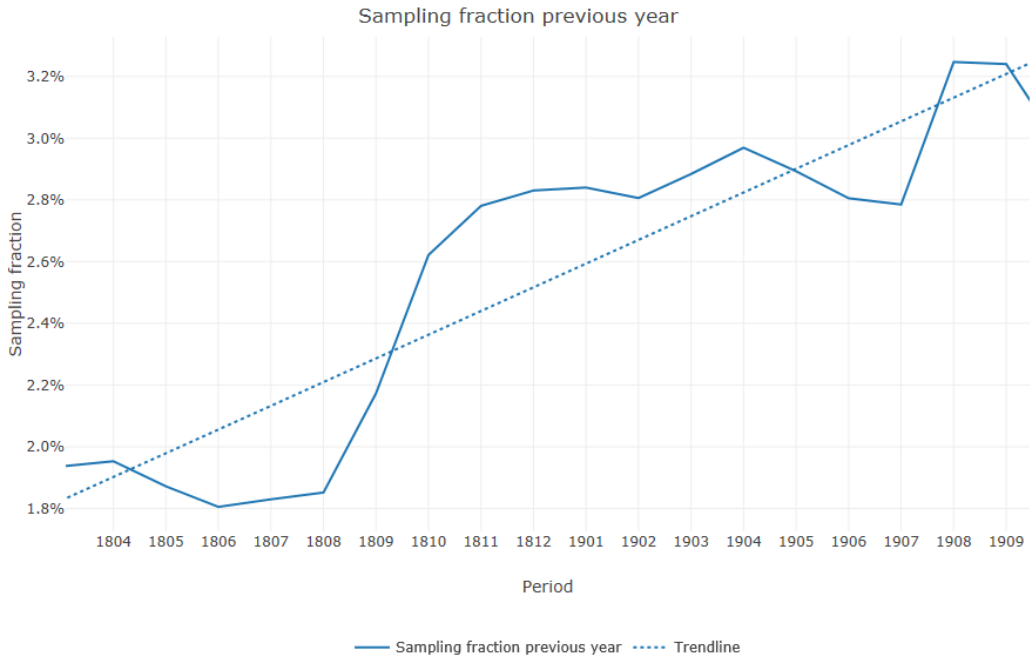


Figure 6.1: Sampling fractions of the sample size of all samples in the previous year.

good as possible. We start with selecting samples for 12 months without screening them yet. Let this sample be denoted by $s_{\text{A,year}}$. Subsequently, all selected samples are screened with respect to each other. Note that by this approximation each sample is screened with respect to 12 months of samples. We are only able to apply screening on the occurrence of an address and on confidential information. Most inhabitants that become not eligible by other reasons are inhabitants that have passed away or emigrated recently. This is difficult to reproduce and can be considered negligible. This results in a sample after screening, which we will denote by $s^*_{\text{A,year}}$. All inhabitants that are eligible after screening would hypothetically have received a letter from Statistics Netherlands with the request to participate in one of the surveys. Note that this basically means that $s^*_{\text{A,year}}$ contains all unique addresses of the inhabitants that do not have a confidentiality indicator in $s_{\text{A,year}}$. Next we can draw a sample s that has the same size as the mobility survey of April 2019 and screen this sample with respect to the samples after screening $s^*_{\text{A,year}}$.

Note that we are considering sampling without replacement, which means that an inhabitant can only be selected once in the sample. Since we use a new version of the population each month we need to keep track on the inhabitants that have already been selected in the sample and remove those from the updated version of the population.

6.2. Simulating the situation with increased sample sizes

Simulating this situation is similar to the simulation of the current situation, but we will use larger sample sizes fpr all samples in the past twelve months. Let the sample in this situation be denoted by $s_{\text{B,year}}$ and the sample after screening by $s^*_{\text{B,year}}$. To be able to compare the sample after screening in both situations, we make sure that $s_{\text{A,year}}$ is a subset of $s_{\text{B,year}}$. This allows us to select one sample s and screen it against both samples $s^*_{\text{A,year}}$ and $s^*_{\text{B,year}}$.

6.3. Results

We have selected the samples $s_{\text{A,year}}$ and $s_{\text{B,year}}$ and screened them with respect to each other to obtain the samples $s^*_{\text{A,year}}$ and $s^*_{\text{B,year}}$. Then we select a sample s from the population and we screen it against $s^*_{\text{A,year}}$ and $s^*_{\text{B,year}}$ to obtain the screened samples s^*_{A} and s^*_{B} respectively. Next, we can apply the tests from Section 5 to both samples.

We have applied the statistical test corresponding to Hypothesis 1 (see Section 5.2) to four different realisations of the sample s. The results are presented in Tables 6.1 and 6.2.

If we look at the $\mathbf{p}$-values of the sample after screening we see that in most cases the $\mathbf{p}$-values in case B is lower than in case A. This confirms that the effects of the screening procedure becomes larger if sample sizes become larger. If we consider the results in Tables 6.1 and 6.2 it occurs regularly that in case A we cannot reject that the sample after screening is representative for the population with respect to an auxiliary variable, but we can reject this in case B.

6.4. Conclusion

The effects of screening the sample become larger as sample sizes increase. As sample sizes increase, more inhabitants become non-eligible by the screening on the occurrence of an address. As a consequence, the sample after screening is not representative for the population with respect to an auxiliary variable more often.

We have seen that samples sizes have become larger in the previous months. If this trend continues, the problem of applying the screening procedure becomes bigger.

	Simulation 1		Simulation 2		Simulation 3		Simulation 4	
	Case A	Case B	Case A	Case B	Case A	Case B	Case A	Case B
Gender								
Sample before screening	0.35271	0.35451	0.47499	0.47704	0.82618	0.82600	0.39352	0.39489
Sample after screening	0.63376	0.95244	0.67402	0.60037	0.90766	0.61582	0.33465	0.28578
Not eligible by occurrence of address	0.01808	0.00127	0.03324	0.19825	0.77317	0.19155	0.12355	0.15828
Not eligible by confidential information	0.25929	0.25840	0.15002	0.15078	0.14276	0.14234	0.04638	0.04693
Not eligible	0.09376	0.00611	0.23222	0.49162	0.68506	0.44055	0.74679	0.62536
Marital status								
Sample before screening	0.97571	0.97667	0.78452	0.78335	0.84147	0.84261	0.22727	0.22903
Sample after screening	0.35919	0.13180	0.25980	0.14599	0.11130	0.04927	0.04451	0.01069
Not eligible by occurrence of address	0.00003	0.00001	0.02191	0.00139	0.00006	0.00007	0.00089	0.00006
Not eligible by confidential information	0.00003	$< 1 \cdot 10^{-5}$	0.00001	0.00002	0.00022	0.00008	0.00029	0.00048
Not eligible	0.00002	$< 1 \cdot 10^{-5}$	0.00628	0.00366	$< 1 \cdot 10^{-5}$	0.00003	0.00124	0.00009
Age								
Sample before screening	0.02354	0.02407	0.33158	0.33051	0.53265	0.53022	0.09508	0.09334
Sample after screening	0.04868	0.00577	0.01871	0.00198	0.21947	0.05374	0.05354	0.00129
Not eligible by occurrence of address	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	0.00003	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$
Not eligible by confidential information	0.05143	0.04978	0.00067	0.00076	0.02950	0.02902	0.01312	0.01304
Not eligible	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	0.00001	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$
Ethnicity								
Sample before screening	0.97448	0.97415	0.49613	0.49793	0.96033	0.96094	0.68670	0.68394
Sample after screening	0.07630	0.04706	0.49291	0.47252	0.22359	0.23988	0.18631	0.10713
Not eligible by occurrence of address	0.00002	0.00019	0.75240	0.45046	0.01262	0.12532	0.10347	0.05685
Not eligible by confidential information	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	0.10340	0.10484	0.00003	0.00003	0.00472	0.00488
Not eligible	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	0.23715	0.17116	0.00003	0.00169	0.00481	0.00427

Table 6.1: The $\mathbf{p}$-values of the statistical test for hypothesis 1 ($R = 100.000$) for the auxiliary variables gender, marital status, age and ethnicity for four different realisations of the simulations.

	Simulation 1		Simulation 2		Simulation 3		Simulation 4	
	Case A	Case B	Case A	Case B	Case A	Case B	Case A	Case B
Place in household								
Sample before screening	0.15938	0.16033	0.98394	0.98463	0.05065	0.04896	0.09234	0.09333
Sample after screening	0.89812	0.32712	0.25966	0.00215	0.22510	0.05138	0.00496	0.00003
Not eligible by occurrence of address	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$
Not eligible by confidential information	$< 1 \cdot 10^{-5}$	0.00002	0.00004	0.00003	0.00005	0.00003	0.00010	0.00009
Not eligible	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$
Type of household								
Sample before screening	0.02919	0.02889	0.83245	0.83237	0.06248	0.06207	0.07304	0.07366
Sample after screening	0.58955	0.35376	0.07037	0.00017	0.09057	0.00556	0.00266	0.00001
Not eligible by occurrence of address	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$
Not eligible by confidential information	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$
Not eligible	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$
Number of people in household								
Sample before screening	0.11992	0.11739	0.19006	0.18875	0.04504	0.04554	0.04427	0.04490
Sample after screening	0.66977	0.18769	0.21623	0.00967	0.03636	0.00258	0.01154	0.00062
Not eligible by occurrence of address	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$
Not eligible by confidential information	0.09581	0.09526	0.03242	0.03209	0.69978	0.69738	0.05796	0.05805
Not eligible	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	0.00002	$< 1 \cdot 10^{-5}$
Ethnicity								
Sample before screening	0.09684	0.09643	0.21487	0.21101	0.10279	0.10361	0.05249	0.05343
Sample after screening	0.14961	0.00281	0.04904	0.00040	0.08818	0.00110	0.00091	$< 1 \cdot 10^{-5}$
Not eligible by occurrence of address	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$
Not eligible by confidential information	0.05547	0.05318	0.11082	0.11066	0.45186	0.44967	0.10199	0.10142
Not eligible	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$

Table 6.2: The $\mathbf{p}$-values of the statistical test for hypothesis 1 ($R = 100.000$) for the auxiliary variables place in household, type of household, number of people in household and number of people on address for four different realisations of the simulations.

7

Estimation of population characteristics

The aim of survey sampling is to obtain an estimate of a population characteristic that lies close to the true value of the population characteristic. This estimate has to be based on information that is collected from the inhabitants in the sample. In Section 2.3.1 we have introduced the Horvitz-Thompson estimator. We have seen that it is an unbiased estimator, and that it makes no explicit use of auxiliary information. Explicit use of auxiliary variables can be used to define an estimator that has a smaller variance. Statistics Netherlands therefore makes use of the *generalised regression estimator* to obtain estimates.

In this chapter, we first define this estimator and subsequently we show that the approximated variance of the generalised regression estimator is smaller than or equal to the variance of the Horvitz-Thompson estimator. In the next section we prove that the generalised regression estimator is consistent and asymptotically unbiased under specific conditions. We end this chapter by describing the effects of the screening procedure on the generalised regression estimator.

7.1. The generalised regression estimator

Let the population be denoted by $U = \{1, \ldots, N\}$ where N is the population size. Suppose a sample s of size n is selected from U by a sampling design $p(\cdot)$ with inclusion probabilities $\pi_k > 0$. Suppose we are interested in estimating the population mean for a certain variable y. Suppose there are p auxiliary variables $x_1, \ldots, x_p$ available for the whole population. Let $\mathbf{X}$ be the $N \times p$-matrix of all auxiliary variables,

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \ldots & x_{Np} \end{pmatrix} \tag{7.1}$$

where the $k-$th row of $\mathbf{X}$ is denoted by the vector $\mathbf{x}_k^T = (x_{k1}, x_{k2}, \ldots, x_{kp})$ of auxiliary variables for element k. The p-vector of population means for all auxiliary variables is denoted by

$$\overline{\mathbf{x}}_U = \frac{1}{N} \sum_{k=1}^{N} \mathbf{x}_k \tag{7.2}$$

Let $y_1, \ldots, y_N$ be the target values for all elements in the population. Suppose one is interested to estimate the population mean

$$\overline{y}_U = \frac{1}{N} \sum_{k=1}^{N} y_k \tag{7.3}$$

We follow the model that is described by Wright [16] and Särndal et al. [4]. If the auxiliary variables are correlated with the target variable y, then it is possible to write the vector $(y_1, \ldots, y_N)$ as a function of $\mathbf{X}$. One possible model to use is linear regression, which means that there exists a suitably chosen vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ of regression coefficients of a best linear fit of y on $\mathbf{X}$. The residuals $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_N)^T$ are then defined by

$$\varepsilon_k = y_k - \mathbf{x}_k^T \boldsymbol{\beta} \tag{7.4}$$

The residuals ε_k are random variables satisfying

$$\mathbb{E}(\varepsilon_k) = 0 \quad \text{and} \quad \mathbb{E}(\varepsilon_k \varepsilon_l) = \begin{cases} \sigma^2 v_k & \text{if } k = l \\ 0 & \text{if } k \neq l \end{cases} \tag{7.5}$$

Here $\sigma^2 > 0$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ are the parameters of the linear model and v_k is a positive auxiliary variable, that is not contained in $\mathbf{X}$ [16]. Suppose that $\sum_{k \in s} \mathbf{x}_k \mathbf{x}_k^T 1/v_k$ is nonsingular for all possible samples $s \in \mathcal{S}$. Based on the whole population U, the weighted ordinary least squares method gives an estimator $\boldsymbol{\beta}_U$ for $\boldsymbol{\beta}$ [4]

$$\boldsymbol{\beta}_U = \left(\sum_{k=1}^{N} \mathbf{x}_k \mathbf{x}_k^T \frac{1}{\sigma^2 v_k} \right)^{-1} \left(\sum_{k=1}^{N} \mathbf{x}_k y_k \frac{1}{\sigma^2 v_k} \right) \tag{7.6}$$

Let $E_k = y_k - \mathbf{x}_k^T \boldsymbol{\beta}_U$ be the population fit residuals. In case $\sum_{k \in s} \mathbf{x}_k \mathbf{x}_k^T 1/v_k$ is singular, one can use a generalised inverse of $\sum_{k \in s} \mathbf{x}_k \mathbf{x}_k^T 1/v_k$ instead.

In general, the vector $\boldsymbol{\beta}_U$ will not be known, particularly because y_k is unknown for all elements. Note that the unknown parameter σ^2 cancels out in Equation (7.6). Let $\mathbf{T}_{\mathbf{xx}}$ and $\mathbf{t}_{\mathbf{x}y}$ denote

$$\mathbf{T}_{\mathbf{xx}} = \sum_{k=1}^{N} \mathbf{x}_k \mathbf{x}_k^T \frac{1}{v_k} \quad \text{and} \quad \mathbf{t}_{\mathbf{x}y} = \sum_{k=1}^{N} \mathbf{x}_k y_k \frac{1}{v_k} \tag{7.7}$$

Then $\boldsymbol{\beta}_U = \mathbf{T}_{\mathbf{xx}}^{-1} \mathbf{t}_{\mathbf{x}y}$. The vector $\boldsymbol{\beta}_U$ can be estimated from the sample using the Horvitz-Thompson estimators for respectively $\mathbf{T}_{\mathbf{xx}}$ and $\mathbf{t}_{\mathbf{x}y}$, which Särndal et al. [4] and Nieuwenbroek and Boonstra [17] define as

$$\hat{\mathbf{T}}_{\mathbf{xx},\mathrm{HT}} = \sum_{k \in s} \frac{1}{\pi_k} \mathbf{x}_k \mathbf{x}_k^T \frac{1}{v_k} \quad \text{and} \quad \hat{\mathbf{t}}_{\mathbf{x}y,\mathrm{HT}} = \sum_{k \in s} \frac{1}{\pi_k} \mathbf{x}_k y_k \frac{1}{v_k} \tag{7.8}$$

The vector $\hat{\boldsymbol{\beta}}$ defined by

$$\hat{\boldsymbol{\beta}} = \left(\hat{\mathbf{T}}_{\mathbf{xx},\mathrm{HT}} \right)^{-1} \hat{\mathbf{t}}_{\mathbf{x}y,\mathrm{HT}} \tag{7.9}$$

is an estimator for $\boldsymbol{\beta}_U$. The regression model is used to define the generalised regression estimator for $\overline{y}_U$. Under the linear regression model, the population mean can be written by

$$\overline{y}_U = \frac{1}{N} \sum_{k=1}^{N} y_k = \frac{1}{N} \sum_{k=1}^{N} \mathbf{x}_k^T \hat{\boldsymbol{\beta}} + \frac{1}{N} \sum_{k=1}^{N} \left(y_k - \mathbf{x}_k^T \hat{\boldsymbol{\beta}} \right) \tag{7.10}$$

The first part in Equation (7.10) is known, whereas the second part is unknown, since y_k is only known for $k \in s$. By making use of the Horvitz Thompson estimator, an unbiased estimator for the second part can be found, which gives that the generalised regression estimator for $\overline{y}_U$ is defined as

$$\begin{aligned}\hat{\overline{y}}_{\text{GREG}} &= \frac{1}{N}\sum_{k=1}^{N}\mathbf{x}_k^T\hat{\boldsymbol{\beta}} + \frac{1}{N}\sum_{k\in s}\left(\frac{1}{\pi_k}y_k - \frac{1}{\pi_k}\mathbf{x}_k^T\hat{\boldsymbol{\beta}}\right)\\ &= \frac{1}{N}\sum_{k\in s}\frac{1}{\pi_k}y_k + \frac{1}{N}\sum_{k=1}^{N}\mathbf{x}_k^T\hat{\boldsymbol{\beta}} - \frac{1}{N}\sum_{k\in s}\frac{1}{\pi_k}\mathbf{x}_k^T\hat{\boldsymbol{\beta}}\\ &= \hat{\overline{y}}_{\text{HT}} + \left(\overline{\mathbf{x}}_U - \hat{\overline{\mathbf{x}}}_{\text{HT}}\right)^T\hat{\boldsymbol{\beta}}\end{aligned} \tag{7.11}$$

where $\hat{\overline{y}}_{\text{HT}}$ and $\hat{\overline{\mathbf{x}}}_{\text{HT}}$ are the Horvitz-Thompson estimators for $\overline{y}_U$ and $\overline{\mathbf{x}}_U$ respectively, i.e.

$$\hat{\overline{y}}_{\text{HT}} = \frac{1}{N}\sum_{k\in s}\frac{1}{\pi_k}y_k \quad \text{and} \quad \hat{\overline{\mathbf{x}}}_{\text{HT}} = \frac{1}{N}\sum_{k\in s}\frac{1}{\pi_k}\mathbf{x}_k \tag{7.12}$$

7.1.1. Alternative expressions for the generalised regression estimator

The generalised regression estimator can be expressed as a weighted sample sum of the target variable, which makes calculations more intuitively, i.e.

$$\begin{aligned}\hat{\overline{y}}_{\text{GREG}} &= \hat{\overline{y}}_{\text{HT}} + \left(\overline{\mathbf{x}}_U - \hat{\overline{\mathbf{x}}}_{\text{HT}}\right)^T\hat{\boldsymbol{\beta}}\\ &= \hat{\overline{y}}_{\text{HT}} + \left(\overline{\mathbf{x}}_U - \hat{\overline{\mathbf{x}}}_{\text{HT}}\right)^T\left(\hat{\mathbf{T}}_{\mathbf{xx},\text{HT}}\right)^{-1}\hat{\mathbf{t}}_{\mathbf{x}y,\text{HT}}\\ &= \frac{1}{N}\sum_{k\in s}\frac{1}{\pi_k}y_k + \left(\overline{\mathbf{x}}_U - \hat{\overline{\mathbf{x}}}_{\text{HT}}\right)^T\left(\hat{\mathbf{T}}_{\mathbf{xx},\text{HT}}\right)^{-1}\sum_{k\in s}\frac{1}{\pi_k}\mathbf{x}_k y_k\frac{1}{\upsilon_k}\\ &= \sum_{k\in s}\left(\frac{1}{N}\frac{1}{\pi_k}y_k + \left(\overline{\mathbf{x}}_U - \hat{\overline{\mathbf{x}}}_{\text{HT}}\right)^T\left(\hat{\mathbf{T}}_{\mathbf{xx},\text{HT}}\right)^{-1}\frac{1}{\pi_k}\mathbf{x}_k y_k\frac{1}{\upsilon_k}\right)\\ &= \sum_{k\in s}\frac{1}{\pi_k}\left(\frac{1}{N} + \left(\overline{\mathbf{x}}_U - \hat{\overline{\mathbf{x}}}_{\text{HT}}\right)^T\left(\hat{\mathbf{T}}_{\mathbf{xx},\text{HT}}\right)^{-1}\mathbf{x}_k\frac{1}{\upsilon_k}\right)y_k\\ &= \sum_{k\in s}w_k y_k\end{aligned} \tag{7.13}$$

where the weights are defined by

$$w_k = \frac{1}{\pi_k}\left(\frac{1}{N} + \left(\overline{\mathbf{x}}_U - \hat{\overline{\mathbf{x}}}_{\text{HT}}\right)^T\left(\hat{\mathbf{T}}_{\mathbf{xx},\text{HT}}\right)^{-1}\mathbf{x}_k\frac{1}{\upsilon_k}\right) \tag{7.14}$$

Writing the generalised regression in this way, means that each inhabitant in the sample obtains a weight that is used for estimation. The estimated value is then obtained by computing the weighted average of the target variables.

Under specific conditions, the generalised regression estimator can be simplified [18], [19], [20]. If there exists a p-vector $\mathbf{c}$ such that

$$\mathbf{x}_k^T\mathbf{c} = \upsilon_k \quad \text{for all } k = 1, \ldots, N \tag{7.15}$$

then the weights of the generalised regression estimator can be rewritten. Note that if there exists such

a p-vector $\mathbf{c}$, then $\mathbf{c}^T \mathbf{x}_k \frac{1}{v_k} = 1$, which gives

$$\begin{aligned}
\left(\hat{\bar{\mathbf{x}}}_{\text{HT}}\right)^T \left(\hat{\mathbf{T}}_{\mathbf{xx},\text{HT}}\right)^{-1} \mathbf{x}_k \frac{1}{v_k} &= \left(\frac{1}{N}\sum_{k\in s}\frac{1}{\pi_k}\cdot 1\cdot \mathbf{x}_k^T\right)\left(\sum_{k\in s}\frac{1}{\pi_k}\mathbf{x}_k\mathbf{x}_k^T\frac{1}{v_k}\right)^{-1}\mathbf{x}_k\frac{1}{v_k} \\
&= \left(\frac{1}{N}\sum_{k\in s}\frac{1}{\pi_k}\cdot \mathbf{c}^T\mathbf{x}_k\frac{1}{v_k}\cdot \mathbf{x}_k^T\right)\left(\sum_{k\in s}\frac{1}{\pi_k}\mathbf{x}_k\mathbf{x}_k^T\frac{1}{v_k}\right)^{-1}\mathbf{x}_k\frac{1}{v_k} \\
&= \frac{1}{N}\mathbf{c}^T\left(\sum_{k\in s}\frac{1}{\pi_k}\mathbf{x}_k\mathbf{x}_k^T\frac{1}{v_k}\right)\left(\sum_{k\in s}\frac{1}{\pi_k}\mathbf{x}_k\mathbf{x}_k^T\frac{1}{v_k}\right)^{-1}\mathbf{x}_k\frac{1}{v_k} \\
&= \frac{1}{N}\mathbf{c}^T\mathbf{x}_k\frac{1}{v_k} = \frac{1}{N}
\end{aligned} \tag{7.16}$$

Under the condition in Equation 7.15, the weights of the generalised regression estimator can be written as

$$\begin{aligned}
w_k &= \frac{1}{\pi_k}\left(\frac{1}{N} + \left(\bar{\mathbf{x}}_U - \hat{\bar{\mathbf{x}}}_{\text{HT}}\right)^T \left(\hat{\mathbf{T}}_{\mathbf{xx},\text{HT}}\right)^{-1}\mathbf{x}_k\frac{1}{v_k}\right) \\
&= \frac{1}{\pi_k}\left(\frac{1}{N} + \bar{\mathbf{x}}_U^T\left(\hat{\mathbf{T}}_{\mathbf{xx},\text{HT}}\right)^{-1}\mathbf{x}_k\frac{1}{v_k} - \left(\hat{\bar{\mathbf{x}}}_{\text{HT}}\right)^T\left(\hat{\mathbf{T}}_{\mathbf{xx},\text{HT}}\right)^{-1}\mathbf{x}_k\frac{1}{v_k}\right) \\
&= \frac{1}{\pi_k}\bar{\mathbf{x}}_U^T\left(\hat{\mathbf{T}}_{\mathbf{xx},\text{HT}}\right)^{-1}\mathbf{x}_k\frac{1}{v_k}
\end{aligned} \tag{7.17}$$

Consequently, under the condition that there exists a p-vector $\mathbf{c}$ such that $\mathbf{x}_k^T\mathbf{c} = v_k$, the generalised regression estimator is

$$\begin{aligned}
\hat{\bar{y}}_{\text{GREG}} &= \sum_{k\in s} w_k y_k \\
&= \sum_{k\in s}\frac{1}{\pi_k}\bar{\mathbf{x}}_U^T\left(\hat{\mathbf{T}}_{\mathbf{xx},\text{HT}}\right)^{-1}\mathbf{x}_k\frac{1}{v_k}y_k \\
&= \bar{\mathbf{x}}_U^T\left(\hat{\mathbf{T}}_{\mathbf{xx},\text{HT}}\right)^{-1}\sum_{k\in s}\frac{1}{\pi_k}\mathbf{x}_k y_k\frac{1}{v_k} \\
&= \bar{\mathbf{x}}_U^T\left(\hat{\mathbf{T}}_{\mathbf{xx},HT}\right)^{-1}\hat{\mathbf{t}}_{\mathbf{x}y,\text{HT}} \\
&= \bar{\mathbf{x}}_U^T\hat{\boldsymbol{\beta}}
\end{aligned} \tag{7.18}$$

7.2. Variance of the generalised regression estimator

Statistics Netherlands uses the generalised regression estimator because its approximated variance is smaller than the variance of the Horvitz-Thompson estimator. We prove this for simple random sampling without replacement.

Because of the complex nature of the generalised regression estimator it is not possible to compute the variance exactly. We derive an approximation of the variance by using Taylor linearisation.

Theorem 7.1 (Approximated variance generalised regression estimator)**.** The generalised regression estimator $\hat{\bar{y}}_{\text{GREG}}$ for $\bar{y}_U$ has approximated variance

$$\mathbb{AV}\left(\hat{\bar{y}}_{\text{GREG}}\right) = \frac{1}{N^2}\sum_{k=1}^{N}\sum_{l=1}^{N}\left(\pi_{kl} - \pi_k\pi_l\right)\breve{E}_k\breve{E}_l \tag{7.19}$$

where $E_k = y_k - \mathbf{x}_k^T \boldsymbol{\beta}_U$ denote the population fit residuals, and as was introduced in Section 2.3 the $\breve{}$ denotes the operation of dividing a value variable by the inclusion probability, i.e. $\breve{E}_k = \frac{1}{\pi_k} E_k$.

Proof. The proof can be found in Appendix A.3.

Theorem 7.2 (Approximated variance generalised regression estimator for fixed size designs)**.** If the sample size n of the sampling design $p(\cdot)$ is fixed, the approximated variance of the generalised regression estimator in Equation (7.19) can be written alternatively as

$$\mathbb{AV}\left(\hat{\bar{y}}_{\text{GREG}}\right) = -\frac{1}{2N^2} \sum_{k=1}^{N} \sum_{l=1}^{N} (\pi_{kl} - \pi_k \pi_l) \left(\breve{E}_k - \breve{E}_l\right)^2 \tag{7.20}$$

Proof. The proof can be found in Appendix A.4.

Remark 7.1. Under the same condition that is made for obtaining the simplified expression for the generalised regression estimator in Equation (7.18), the sum of the population fit residuals is zero. In other words, if there exists a p-vector $\mathbf{c}$ such that $\mathbf{x}_k^T \mathbf{c} = v_k$, then $\sum_{k=1}^{N} E_k = 0$.

Proof.

$$\begin{aligned}
\sum_{k=1}^{N} E_k &= \sum_{k=1}^{N} \left(y_k - \mathbf{x}_k^T \boldsymbol{\beta}_U\right) \\
&= \sum_{k=1}^{N} y_k - \sum_{k=1}^{N} \mathbf{x}_k^T \boldsymbol{\beta}_U \\
&= \sum_{k=1}^{N} y_k - \left(\sum_{k=1}^{N} 1 \cdot \mathbf{x}_k^T\right) \left(\sum_{k=1}^{N} \mathbf{x}_k \mathbf{x}_k^T \frac{1}{v_k}\right)^{-1} \left(\sum_{k=1}^{N} \mathbf{x}_k y_k \frac{1}{v_k}\right) \\
&= \sum_{k=1}^{N} y_k - \left(\sum_{k=1}^{N} \mathbf{c}^T \mathbf{x}_k \frac{1}{v_k} \cdot \mathbf{x}_k^T\right) \left(\sum_{k=1}^{N} \mathbf{x}_k \mathbf{x}_k^T \frac{1}{v_k}\right)^{-1} \left(\sum_{k=1}^{N} \mathbf{x}_k y_k \frac{1}{v_k}\right) \\
&= \sum_{k=1}^{N} y_k - \mathbf{c}^T \left(\sum_{k=1}^{N} \mathbf{x}_k y_k \frac{1}{v_k}\right) \\
&= \sum_{k=1}^{N} y_k - \sum_{k=1}^{N} y_k = 0
\end{aligned} \tag{7.21}$$

□

The approximated variance of the regression estimator for simple random sampling can be derived

from Equation (7.20). We use Remark 7.1 and Equation (2.35).

$$\begin{aligned}
\mathbb{AV}\left(\hat{\bar{y}}_{\text{GREG}}\right) &= -\frac{1}{2N^2}\sum_{k=1}^{N}\sum_{l=1}^{N}(\pi_{kl}-\pi_k\pi_l)\left(\breve{E}_k-\breve{E}_l\right)^2 \\
&= -\frac{1}{2N^2}\sum_{k=1}^{N}\sum_{k\neq l=1}^{N}(\pi_{kl}-\pi_k\pi_l)\left(\breve{E}_k-\breve{E}_l\right)^2 - \frac{1}{2N^2}\sum_{k=1}^{N}(\pi_{kk}-\pi_k\pi_k)\left(\breve{E}_k-\breve{E}_k\right)^2 \\
&= -\frac{1}{2N^2}\frac{-f(1-f)}{N-1}\sum_{k=1}^{N}\sum_{l=1}^{N}\left(\frac{N}{n}(E_k-E_l)\right)^2 \\
&= -\frac{1}{2N^2}\frac{-f(1-f)}{N-1}\frac{1}{f^2}\sum_{k=1}^{N}\sum_{l=1}^{N}(E_k-E_l)^2 \\
&= \frac{1}{2N^2}\frac{1-f}{N-1}\frac{1}{f}\sum_{k=1}^{N}\sum_{l=1}^{N}\left(E_k^2-2E_kE_l+E_l^2\right) \\
&= \frac{1}{2N^2}\frac{1-f}{N-1}\frac{1}{f}\left(2N\sum_{k=1}^{N}E_k^2-2\sum_{k=1}^{N}\left(E_k\sum_{l=1}^{N}E_l\right)\right) \\
&= \frac{1}{2N^2}\frac{1-f}{N-1}\frac{1}{f}2N\sum_{k=1}^{N}E_k^2 \\
&= \frac{1-f}{n}\frac{1}{N-1}\sum_{k=1}^{N}E_k^2 \\
&= \frac{1-f}{n}S_{yU}^2\left(1-Q^2\right)
\end{aligned} \tag{7.22}$$

where

$$Q^2 = 1-\frac{1}{S_{yU}^2}\frac{1}{N-1}\sum_{k=1}^{N}E_k^2 \tag{7.23}$$

and S_{yU}^2 is the population variance, which is given by Equation (2.39). Recall that the variance of the Horvitz-Thompson estimator for simple random sampling without replacement is given by (see Equation 2.38)

$$\mathbb{V}\left(\hat{\bar{y}}_{\text{HT}}\right) = \frac{1-f}{n}S_{yU}^2 \tag{7.24}$$

Then the ratio of approximated variance of the generalised regression estimator and the variance of the Horvitz-Thompson estimator for SRSWR is

$$\frac{\mathbb{AV}\left(\hat{\bar{y}}_{\text{GREG}}\right)}{\mathbb{V}\left(\hat{\bar{y}}_{\text{HT}}\right)} = \frac{\frac{1-f}{n}S_{yU}^2\left(1-Q^2\right)}{\frac{1-f}{n}S_{yU}^2} = 1-Q^2 \tag{7.25}$$

Then we conclude that

$$\mathbb{AV}\left(\hat{\bar{y}}_{\text{GREG}}\right) \leq \mathbb{V}\left(\hat{\bar{y}}_{\text{HT}}\right) \text{ if and only if } Q^2 \geq 0 \tag{7.26}$$

Consequently, the generalised regression estimator is by approximation likely to be an improvement over the Horvitz-Thompson estimator if $Q^2 > 0$.

7.3. Consistency and asymptotically design unbiasedness

Although the approximated variance of the generalised regression estimator is smaller than the variance of the Horvitz-Thompson estimator (if $Q^2 > 0$), the generalised regression estimator is not an unbiased estimator whereas the Horvitz-Thompson estimator is an unbiased estimator. However, the generalised regression estimator is *asymptotically unbiased*. In practice, this means that the generalised regression estimator can be considered approximately unbiased when the sample size n is large enough [4]. Moreover, the generalised regression estimator is a *consistent* estimator, which means that the sampling distribution of $\hat{\bar{y}}_{\text{GREG}}$ can be considered tightly concentrated around $\bar{y}_U$, when n is large enough [4]. In this section, we give more formal and precise definitions for asymptotically unbiasedness and consistency. Subsequently, we prove that the generalised regression estimator is asymptotically unbiased and consistent under specific conditions.

To introduce definitions for consistency and asymptotic unbiasedness of estimators, a more comprehensive framework known as the *super-population model* is required. The main idea is to assume that we can treat the finite population of size N as if it were a iid sample of an infinite super-population [20], [21], [22]. The super-population model allows us to investigate properties of estimators as the sample size and the population size become large.

Thus far we have treated the sampling design $p(s)$ as the probability that a sample outcome s is selected by the design. The target variables y_k are considered non-stochastic but unknown numbers and probability statements arise from selection of units in the sample [22]. This limits our inference to the reference population only.

By contrast, the super-population view regards the finite population of interest as a sample of size N from an infinite population [21]. Consequently, under the super-population model the sample of size n is generated by a two step procedure:

STEP 1: Draw an iid sample of size N from an infinite super population

STEP 2: Draw a sample of size $n < N$ from the sample obtained from STEP 1.

Note that the first step is an imaginary step. Usually it is assumed that the resulting sample elements are independent and identically distributed. In terms of the super-population model, the finite population sampling may be regarded as being based on the conditional distribution given a particular outcome of drawing a sample of size N from an infinite population [21].

Rubin-Bleuer et al. [22] state that the finite population U is associated with a super-population that consists of a probability space $(\Omega, \mathcal{F}, P)$ and random vectors $(Y_k, \mathbf{X}_k)$, $Y_k : \Omega \to \mathbb{R}$, $\mathbf{X}_k : \Omega \to \mathbb{R}^p$, such that $Y_k(\omega_0) = y_k$ and $\mathbf{X}_k(\omega_0) = \mathbf{x}_k$ for some $\omega_0 \in \Omega$. Here Y_k represents the random variable of the target variable and $\mathbf{X}_k \in \mathbb{R}^p$ the random vector of the auxiliary information. Rubin-Bleuer et al. [22] write the vectors $\mathbf{Y}^N$ and $\mathbf{X}^N$ as

$$\mathbf{Y}^N = (y_k)_{k=1,\dots,N} \quad \text{and} \quad \mathbf{X}^N = (\mathbf{x}_k)_{k=1,\dots,N} \tag{7.27}$$

Then any distribution of $\left(\mathbf{Y}^N, \mathbf{X}^N\right)$ that is given a priori is called a super-population model and the finite population U is a realisation of the super-population.

To allow us to investigate the properties of the generalised regression estimator for large population- and sample sizes, consider an infinite sequence of populations $U_1, U_2, U_3, \dots$ where U_τ consists of the first N_τ elements from the infinite sequence of populations, that is $U_\tau = \{1, 2, \dots, N_\tau\}$. Assume that $U_1 \subset U_2 \subset U_3 \subset \cdots$ and hence $N_1 < N_2 < N_3 < \cdots$ [20].

For each population U_τ consider a probability sampling design $p_\tau(\cdot)$ that assigns a probability $p_\tau(s_\tau)$ to each possible sample s_τ of elements of U_τ. Let $\pi_{k\tau}$ and $\pi_{kl\tau}$ $(k,l = 1,2,\ldots,N_\tau)$ denote the first- and second-order inclusion probabilities determined by the sampling design $p_\tau(\cdot)$. Let the sample size n_τ of sample s_τ be fixed and such that $n_\tau \leq N_\tau$ for all τ. Assume that $n_1 < n_2 < n_3 < \cdots$. Now $\tau \to \infty$ means that both $n_\tau \to \infty$ and $N_\tau \to \infty$, but it is not required that n_τ increases as fast as N_τ [23].

Let $I_{k\tau}$ be an indicator stating whether element k from population U_τ is in the sample s_τ or not. Let $\mathbf{Y}_\tau = \left(Y_1,\ldots,Y_{N_\tau}\right)$ denote the vector of independent random variables and let ξ denote the probability distribution of the infinite sequence of random variables $Y_1, Y_2, \ldots$. This setup is also known as the super-population model, which is also denoted by ξ for short. Let $t_{\mathbf{Y}_\tau}$ be the a population parameter of population U_τ, which means that $t_{\mathbf{Y}_\tau}$ is a function of $Y_1,\ldots,Y_{N_\tau}$. Let $\hat{t}_{\mathbf{Y}_\tau}$ be an estimator for $t_{\mathbf{Y}_\tau}$.

Definition 7.1. The estimator $\hat{t}_{\mathbf{Y}_\tau}$ is *asymptotically design unbiased* for $t_{\mathbf{Y}_\tau}$ if

$$\lim_{\tau\to\infty} \mathbb{E}_{p_\tau}\left(\hat{t}_{\mathbf{Y}_\tau}\middle|\,\mathbf{Y}_\tau\right) - t_{\mathbf{Y}_\tau} = 0 \quad \xi\text{- almost surely} \tag{7.28}$$

Definition 7.2. The estimator $\hat{t}_{\mathbf{Y}_\tau}$ is *consistent* for $t_{\mathbf{Y}_\tau}$ if $\forall\varepsilon > 0$

$$\lim_{\tau\to\infty} \mathbb{P}_{p_\tau}\left(\left|\hat{t}_{\mathbf{Y}_\tau} - t_{\mathbf{Y}_\tau}\right| > \varepsilon\right) = 0 \quad \xi\text{- almost surely} \tag{7.29}$$

In this framework, the population mean for population U_τ is written as $\overline{y}_{U_\tau} = \frac{1}{N_\tau}\sum_{k=1}^{N_\tau} Y_k$ and the generalised regression estimator from Equation (7.11) for $\overline{y}_{U_\tau}$ is written as

$$\begin{aligned}
\hat{\overline{y}}_{\tau,\text{GREG}} &= \hat{\overline{y}}_{\tau,\text{HT}} + \left(\overline{\mathbf{x}}_U - \hat{\overline{\mathbf{x}}}_{\text{HT}}\right)^T \hat{\boldsymbol{\beta}}_\tau \\
&= \frac{1}{N_\tau}\sum_{k=1}^{N_\tau}\frac{I_{k\tau}}{\pi_{k\tau}}Y_k + \left(\frac{1}{N_\tau}\sum_{k=1}^{N_\tau}\mathbf{x}_k - \frac{1}{N_\tau}\sum_{k=1}^{N_\tau}\frac{I_{k\tau}}{\pi_{k\tau}}\mathbf{x}_k\right)^T \hat{\boldsymbol{\beta}}_\tau \\
&= \frac{1}{N_\tau}\sum_{k=1}^{N_\tau}\frac{I_{k\tau}}{\pi_{k\tau}}Y_k + \sum_{j=1}^{p}\left(\left(\frac{1}{N_\tau}\sum_{k=1}^{N_\tau}x_{kj} - \frac{1}{N_\tau}\sum_{k=1}^{N_\tau}\frac{I_{k\tau}}{\pi_{k\tau}}x_{kj}\right)\left(\hat{\boldsymbol{\beta}}_\tau\right)_j\right) \\
&= \frac{1}{N_\tau}\sum_{k=1}^{N_\tau}\left(\frac{I_{k\tau}}{\pi_{k\tau}}Y_k + \sum_{j=1}^{p}\left(1 - \frac{I_{k\tau}}{\pi_{k\tau}}\right)x_{kj}\left(\hat{\boldsymbol{\beta}}_\tau\right)_j\right)
\end{aligned} \tag{7.30}$$

where $\left(\hat{\boldsymbol{\beta}}_\tau\right)_j$ represents the j-th element of the vector of regression coefficients $\hat{\boldsymbol{\beta}}_\tau$

$$\hat{\boldsymbol{\beta}}_\tau = \left(\sum_{k=1}^{N_\tau}\frac{I_{k\tau}}{\pi_{k\tau}}\mathbf{x}_k\mathbf{x}_k^T\frac{1}{v_k}\right)^{-1}\left(\sum_{k=1}^{N_\tau}\frac{I_{k\tau}}{\pi_{k\tau}}\mathbf{x}_k Y_k\frac{1}{v_k}\right) \tag{7.31}$$

Before we prove asymptotically unbiasedness and consistency, a few conditions on the auxiliary- and target variables and on the inclusion probabilities are imposed.

(C0) $\sum_{k\in s}\mathbf{x}_k\mathbf{x}_k^T\frac{1}{v_k}$ is nonsingular for all possible samples $s \in \mathcal{S}$

(C1) $\limsup_{\tau\to\infty}\frac{1}{N_\tau}\sum_{k=1}^{N_\tau}x_{kj}^2 < \infty$ for $j = 1,\ldots,p$

(C2) $\limsup_{\tau\to\infty}\frac{1}{N_\tau}\sum_{k=1}^{N_\tau}Y_k^2 < \infty \quad \xi$- almost surely

(C3) $\limsup_{\tau\to\infty} \mathbb{E}_\tau\left(\sum_{j=1}^{p}\left(\left(\hat{\boldsymbol{\beta}}_\tau\right)_j\right)^2\middle|\mathbf{Y}_\tau\right) < \infty \quad \xi$- almost surely

(C4) $\liminf_{\tau\to\infty} N_\tau \min_{1\le k\le N_\tau} \pi_{k\tau} = \infty$

(C5) $\lim_{\tau\to\infty} \max_{1\le k\ne l\le N_\tau}\left|\frac{\pi_{kl\tau}}{\pi_{k\tau}\pi_{l\tau}} - 1\right| = 0$

The most important result from Robinson and Särndal [23] is formulated in the following theorem.

Theorem 7.3. Under conditions (C0) to to (C5), the generalised regression estimator $\hat{\bar{y}}_{\tau,\text{GREG}}$ from Equation (7.30) is asymptotically design unbiased and consistent for $\bar{y}_{U_\tau}$.

Proof. This proof follows the proof that is given by Robinson and Särndal [23]. We use some important inequalities that are introduced in Appendix A.5.

Using the Markov inequality (Lemma A.1) it follows that to prove asymptotically design unbiasedness and consistency, it is sufficient to prove

$$\lim_{\tau\to\infty} \mathbb{E}_\tau\left(\left|\hat{\bar{y}}_{\tau,\text{GREG}} - \bar{y}_{U_\tau}\right|\middle|\mathbf{Y}_\tau\right) = 0 \tag{7.32}$$

ξ -almost surely. Here $\mathbb{E}_\tau$ denotes the expected value with respect to τ. Using the triangle inequality and the generalised regression estimator from Equation (7.30) we have that

$$\begin{aligned}
\left|\hat{\bar{y}}_{\tau,\text{GREG}} - \bar{y}_{U_\tau}\right| &= \left|\frac{1}{N_\tau}\sum_{k=1}^{N_\tau}\left(\frac{I_{k\tau}}{\pi_{k\tau}}Y_k + \sum_{j=1}^{p}\left(1-\frac{I_{k\tau}}{\pi_{k\tau}}\right)x_{kj}\left(\hat{\boldsymbol{\beta}}_\tau\right)_j\right) - \frac{1}{N_\tau}\sum_{k=1}^{N_\tau}Y_k\right| \\
&= \left|\frac{1}{N_\tau}\sum_{k=1}^{N_\tau}\left(\left(\frac{I_{k\tau}}{\pi_{k\tau}}-1\right)Y_k + \sum_{j=1}^{p}\left(1-\frac{I_{k\tau}}{\pi_{k\tau}}\right)x_{kj}\left(\hat{\boldsymbol{\beta}}_\tau\right)_j\right)\right| \\
&\le \left|\frac{1}{N_\tau}\sum_{k=1}^{N_\tau}\left(\frac{I_{k\tau}}{\pi_{k\tau}}-1\right)Y_k\right| + \left|\frac{1}{N_\tau}\sum_{k=1}^{N_\tau}\sum_{j=1}^{p}\left(1-\frac{I_{k\tau}}{\pi_{k\tau}}\right)x_{kj}\left(\hat{\boldsymbol{\beta}}_\tau\right)_j\right|
\end{aligned} \tag{7.33}$$

Consequently

$$\begin{aligned}
\mathbb{E}_\tau\left(\left|\hat{\bar{y}}_{\tau,\text{GREG}} - \bar{y}_{U_\tau}\right|\middle|\mathbf{Y}_\tau\right) \le{}& \mathbb{E}_\tau\left(\left|\frac{1}{N_\tau}\sum_{k=1}^{N_\tau}\left(\frac{I_{k\tau}}{\pi_{k\tau}}-1\right)Y_k\right|\middle|\mathbf{Y}_\tau\right) \\
&+ \mathbb{E}_\tau\left(\left|\frac{1}{N_\tau}\sum_{k=1}^{N_\tau}\sum_{j=1}^{p}\left(1-\frac{I_{k\tau}}{\pi_{k\tau}}\right)x_{kj}\left(\hat{\boldsymbol{\beta}}_\tau\right)_j\right|\middle|\mathbf{Y}_\tau\right)
\end{aligned} \tag{7.34}$$

Applying the Lyapounov inequality (Lemma A.2) with $p = 1$ and $r = 2$ to the first part, gives that

$$\mathbb{E}_\tau\left(\left|\frac{1}{N_\tau}\sum_{k=1}^{N_\tau}\left(\frac{I_{k\tau}}{\pi_{k\tau}}-1\right)Y_k\right|\middle|\mathbf{Y}_\tau\right) \le \underbrace{\left(\mathbb{E}_\tau\left(\left|\frac{1}{N_\tau}\sum_{k=1}^{N_\tau}\left(\frac{I_{k\tau}}{\pi_{k\tau}}-1\right)Y_k\right|^2\middle|\mathbf{Y}_\tau\right)\right)^{\frac{1}{2}}}_{\text{I}} \tag{7.35}$$

For the second part we have that

$$\mathbb{E}_\tau\left(\left|\frac{1}{N_\tau}\sum_{k=1}^{N_\tau}\sum_{j=1}^{p}\left(1-\frac{I_{k\tau}}{\pi_{k\tau}}\right)x_{kj}\left(\hat{\boldsymbol{\beta}}_\tau\right)_j\right|\middle|\mathbf{Y}_\tau\right) = \mathbb{E}_\tau\left(\left|\sum_{j=1}^{p}\frac{1}{N_\tau}\sum_{k=1}^{N_\tau}\left(1-\frac{I_{k\tau}}{\pi_{k\tau}}\right)x_{kj}\left(\hat{\boldsymbol{\beta}}_\tau\right)_j\right|\right) \tag{7.36}$$

Applying the Cauchy-Schwarz inequality (Lemma A.3) gives that

$$\begin{aligned}
&\mathbb{E}_\tau\left(\left|\sum_{j=1}^{p}\frac{1}{N_\tau}\sum_{k=1}^{N_\tau}\left(1-\frac{I_{k\tau}}{\pi_{k\tau}}\right)x_{kj}\left(\hat{\boldsymbol{\beta}}_\tau\right)_j\right|\right)\\
&\quad\leq\underbrace{\left(\mathbb{E}_\tau\left(\sum_{j=1}^{p}\left|\frac{1}{N_\tau}\sum_{k=1}^{N_\tau}\left(1-\frac{I_{k\tau}}{\pi_{k\tau}}\right)x_{kj}\right|^2\middle|\mathbf{Y}_\tau\right)\right)^{\frac{1}{2}}}_{\textbf{II}}\cdot\underbrace{\left(\mathbb{E}_\tau\left(\sum_{j=1}^{p}\left(\left(\hat{\boldsymbol{\beta}}_\tau\right)_j\right)^2\middle|\mathbf{Y}_\tau\right)\right)^{\frac{1}{2}}}_{\textbf{III}}
\end{aligned}\tag{7.37}$$

We evaluate the parts **I**, **II** and **III** separately as $\tau\to\infty$. For the first part, we start by developing the square, which gives

$$\begin{aligned}
&\lim_{\tau\to\infty}\mathbb{E}_\tau\left(\left|\frac{1}{N_\tau}\sum_{k=1}^{N_\tau}\left(\frac{I_{k\tau}}{\pi_{k\tau}}-1\right)Y_k\right|^2\middle|\mathbf{Y}_\tau\right)\\
&\quad=\lim_{\tau\to\infty}\mathbb{E}_\tau\left(\frac{1}{N_\tau^2}\sum_{k=1}^{N_\tau}\sum_{l=1}^{N_\tau}\left(\frac{I_{k\tau}}{\pi_{k\tau}}-1\right)\left(\frac{I_{l\tau}}{\pi_{l\tau}}-1\right)Y_kY_l\middle|\mathbf{Y}_\tau\right)\\
&\quad=\lim_{\tau\to\infty}\mathbb{E}_\tau\left(\frac{1}{N_\tau^2}\sum_{k=1}^{N_\tau}\left(\frac{I_{k\tau}}{\pi_{k\tau}}-1\right)^2Y_k^2\middle|\mathbf{Y}_\tau\right)\\
&\qquad+\lim_{\tau\to\infty}\mathbb{E}_\tau\left(\frac{1}{N_\tau^2}\sum_{k=1}^{N_\tau}\sum_{k\neq l=1}^{N_\tau}\left(\frac{I_{k\tau}}{\pi_{k\tau}}-1\right)\left(\frac{I_{l\tau}}{\pi_{l\tau}}-1\right)Y_kY_l\middle|\mathbf{Y}_\tau\right)
\end{aligned}\tag{7.38}$$

Next, we use that the expected value of a sum equals the sum of expectations, which gives

$$\begin{aligned}
&\lim_{\tau\to\infty}\mathbb{E}_\tau\left(\left|\frac{1}{N_\tau}\sum_{k=1}^{N_\tau}\left(\frac{I_{k\tau}}{\pi_{k\tau}}-1\right)Y_k\right|^2\middle|\mathbf{Y}_\tau\right)\\
&\quad=\lim_{\tau\to\infty}\frac{1}{N_\tau^2}\sum_{k=1}^{N_\tau}\mathbb{E}_\tau\left(\left(\frac{I_{k\tau}}{\pi_{k\tau}}-1\right)^2Y_k^2\middle|\mathbf{Y}_\tau\right)\\
&\qquad+\lim_{\tau\to\infty}\frac{1}{N_\tau^2}\sum_{k=1}^{N_\tau}\sum_{k\neq l=1}^{N_\tau}\mathbb{E}_\tau\left(\left(\frac{I_{k\tau}}{\pi_{k\tau}}-1\right)\left(\frac{I_{l\tau}}{\pi_{l\tau}}-1\right)Y_kY_l\middle|\mathbf{Y}_\tau\right)\\
&\quad=\lim_{\tau\to\infty}\frac{1}{N_\tau^2}\sum_{k=1}^{N_\tau}\mathbb{E}_\tau\left(\frac{I_{k\tau}I_{k\tau}}{\pi_{k\tau}\pi_{k\tau}}-2\frac{I_{k\tau}}{\pi_{k\tau}}+1\right)Y_k^2\\
&\qquad+\lim_{\tau\to\infty}\frac{1}{N_\tau^2}\sum_{k=1}^{N_\tau}\sum_{k\neq l=1}^{N_\tau}\mathbb{E}_\tau\left(\frac{I_{k\tau}I_{l\tau}}{\pi_{k\tau}\pi_{l\tau}}-\frac{I_{k\tau}}{\pi_{k\tau}}-\frac{I_{l\tau}}{\pi_{l\tau}}+1\right)Y_kY_l
\end{aligned}\tag{7.39}$$

Recall that $\mathbb{E}_\tau(I_{k\tau})=\pi_{k\tau}$, $\mathbb{E}_\tau(I_{k\tau}I_{l\tau})=\pi_{kl\tau}$ if $k\neq l$ and $\mathbb{E}_\tau(I_{k\tau}I_{l\tau})=\pi_{k\tau}$ if $k=l$. We use this to

compute the expected values. This gives

$$\begin{aligned}\lim_{\tau\to\infty}\mathbb{E}_\tau&\left(\left|\frac{1}{N_\tau}\sum_{k=1}^{N_\tau}\left(\frac{I_{k\tau}}{\pi_{k\tau}}-1\right)Y_k\right|^2\middle|\mathbf{Y}_\tau\right)\\&=\lim_{\tau\to\infty}\frac{1}{N_\tau^2}\sum_{k=1}^{N_\tau}\left(\frac{1}{\pi_{k\tau}}-1\right)Y_k^2+\lim_{\tau\to\infty}\frac{1}{N_\tau^2}\sum_{k=1}^{N_\tau}\sum_{k\neq l=1}^{N_\tau}\left(\frac{\pi_{kl\tau}}{\pi_{k\tau}\pi_{l\tau}}-1\right)Y_kY_l\end{aligned}\tag{7.40}$$

The first term in Equation (7.40) is dominated by

$$\begin{aligned}\lim_{\tau\to\infty}\frac{1}{N_\tau^2}\sum_{k=1}^{N_\tau}\left(\frac{1}{\pi_{k\tau}}-1\right)Y_k^2&\leq\lim_{\tau\to\infty}\frac{1}{N_\tau^2}\sum_{k=1}^{N_\tau}\frac{1}{\pi_{k\tau}}Y_k^2\\&\leq\lim_{\tau\to\infty}\left(N_\tau\min_{1\leq k\leq N_\tau}\pi_{k\tau}\right)^{-1}\frac{1}{N_\tau}\sum_{k=1}^{N_\tau}Y_k^2\end{aligned}\tag{7.41}$$

and the second term in Equation (7.40) is dominated by

$$\begin{aligned}\lim_{\tau\to\infty}\frac{1}{N_\tau^2}\sum_{k=1}^{N_\tau}\sum_{k\neq l=1}^{N_\tau}\left(\frac{\pi_{kl\tau}}{\pi_{k\tau}\pi_{l\tau}}-1\right)Y_kY_l&\leq\lim_{\tau\to\infty}\max_{1\leq k\neq l\leq N_\tau}\left|\frac{\pi_{kl\tau}}{\pi_{k\tau}\pi_{l\tau}}-1\right|\frac{1}{N_\tau^2}\sum_{k=1}^{N_\tau}\sum_{k\neq l=1}^{N_\tau}Y_kY_l\\&\leq\lim_{\tau\to\infty}\max_{1\leq k\neq l\leq N_\tau}\left|\frac{\pi_{kl\tau}}{\pi_{k\tau}\pi_{l\tau}}-1\right|\left(\frac{1}{N_\tau}\sum_{k=1}^{N_\tau}Y_k\right)^2\end{aligned}\tag{7.42}$$

Applying the Cauchy-Schwarz inequality (Lemma A.3) to $\left(\frac{1}{N_\tau}\sum_{k=1}^{N_\tau}Y_k\right)^2$ gives

$$\left(\frac{1}{N_\tau}\sum_{k=1}^{N_\tau}Y_k\right)^2=\left(\sum_{k=1}^{N_\tau}\frac{1}{N_\tau}Y_k\right)^2\leq\left(\sum_{k=1}^{N_\tau}\frac{1}{N_\tau^2}\right)\left(\sum_{k=1}^{N_\tau}Y_k^2\right)=\frac{1}{N_\tau}\sum_{k=1}^{N_\tau}Y_k^2\tag{7.43}$$

Consequently, part **I** is dominated by

$$\begin{aligned}\lim_{\tau\to\infty}\mathbb{E}_\tau&\left(\left|\frac{1}{N_\tau}\sum_{k=1}^{N_\tau}\left(\frac{I_{k\tau}}{\pi_{k\tau}}-1\right)Y_k\right|^2\middle|\mathbf{Y}_\tau\right)\\&\leq\lim_{\tau\to\infty}\left(N_\tau\min_{1\leq k\leq N_\tau}\pi_{k\tau}\right)^{-1}\frac{1}{N_\tau}\sum_{k=1}^{N_\tau}Y_k^2\\&\quad+\lim_{\tau\to\infty}\max_{1\leq k\neq l\leq N_\tau}\left|\frac{\pi_{kl\tau}}{\pi_{k\tau}\pi_{l\tau}}-1\right|\frac{1}{N_\tau}\sum_{k=1}^{N_\tau}Y_k^2\end{aligned}\tag{7.44}$$

By using Conditions (C2), (C4) and (C5) we obtain the desired result of part **I** as $\tau\to\infty$

$$\lim_{\tau\to\infty}\left(\mathbb{E}_\tau\left(\left|\frac{1}{N_\tau}\sum_{k=1}^{N_\tau}\left(\frac{I_{k\tau}}{\pi_{k\tau}}-1\right)Y_k\right|^2\middle|\mathbf{Y}_\tau\right)\right)^{\frac{1}{2}}=0\quad\xi\text{- almost surely}\tag{7.45}$$

Obtaining the result for part **II** is quite similar to part **I**. By similarly developing the square, computing the expected values and by using the same type of dominations we obtain that

$$\begin{aligned}
&\lim_{\tau\to\infty} \mathbb{E}_\tau\left(\sum_{j=1}^{p}\left|\frac{1}{N_\tau}\sum_{k=1}^{N_\tau}\left(1-\frac{I_{k\tau}}{\pi_{k\tau}}\right)x_{kj}\right|^2\middle|\mathbf{Y}_\tau\right)\\
&\quad\leq \lim_{\tau\to\infty}\left(N_\tau \min_{1\leq k\leq N_\tau}\pi_{k\tau}\right)^{-1}\frac{1}{N_\tau}\sum_{j=1}^{p}\sum_{k=1}^{N_\tau}x_{kj}^2\\
&\quad\quad+\lim_{\tau\to\infty}\max_{1\leq k\neq l\leq N_\tau}\left|\frac{\pi_{kl\tau}}{\pi_{k\tau}\pi_{l\tau}}-1\right|\left(\frac{1}{N_\tau}\sum_{j=1}^{p}\sum_{k=1}^{N_\tau}x_{kj}\right)^2
\end{aligned}\tag{7.46}$$

Now applying the Cauchy-Schwarz inequality (Lemma A.3) to $\left(\frac{1}{N_\tau}\sum_{j=1}^{p}\sum_{k=1}^{N_\tau}x_{kj}\right)^2$ gives

$$\begin{aligned}
\left(\frac{1}{N_\tau}\sum_{j=1}^{p}\sum_{k=1}^{N_\tau}x_{kj}\right)^2&=\left(\sum_{k=1}^{N_\tau}\left(\frac{1}{N_\tau}\sum_{j=1}^{p}x_{kj}\right)\right)^2\\
&\leq\left(\sum_{k=1}^{N_\tau}\frac{1}{N_\tau^2}\right)\left(\sum_{k=1}^{N_\tau}\left(\sum_{j=1}^{p}x_{kj}\right)^2\right)\\
&\leq\frac{1}{N_\tau}\sum_{j=1}^{p}\sum_{k=1}^{N_\tau}x_{kj}^2
\end{aligned}\tag{7.47}$$

Then we have that

$$\begin{aligned}
&\lim_{\tau\to\infty} \mathbb{E}_\tau\left(\sum_{j=1}^{p}\left|\frac{1}{N_\tau}\sum_{k=1}^{N_\tau}\left(1-\frac{I_{k\tau}}{\pi_{k\tau}}\right)x_{kj}\right|^2\middle|\mathbf{Y}_\tau\right)\\
&\quad\leq \lim_{\tau\to\infty}\left(N_\tau \min_{1\leq k\leq N_\tau}\pi_{k\tau}\right)^{-1}\frac{1}{N_\tau}\sum_{j=1}^{p}\sum_{k=1}^{N_\tau}x_{kj}^2\\
&\quad\quad+\lim_{\tau\to\infty}\max_{1\leq k\neq l\leq N_\tau}\left|\frac{\pi_{kl\tau}}{\pi_{k\tau}\pi_{l\tau}}-1\right|\frac{1}{N_\tau}\sum_{j=1}^{p}\sum_{k=1}^{N_\tau}x_{kj}^2
\end{aligned}\tag{7.48}$$

And conditions (C1), (C4) and (C5) give the desired result for part **III**:

$$\lim_{\tau\to\infty}\left(\mathbb{E}_\tau\left(\sum_{j=1}^{p}\left|\frac{1}{N_\tau}\sum_{k=1}^{N_\tau}\left(1-\frac{I_{k\tau}}{\pi_{k\tau}}\right)x_{kj}\right|^2\middle|\mathbf{Y}_\tau\right)\right)^{\frac{1}{2}}=0\quad\xi\text{- almost surely}\tag{7.49}$$

Note that part **III** immediately follows from Condition (C3)

$$\lim_{\tau\to\infty}\left(\mathbb{E}_\tau\left(\sum_{j=1}^{p}\left(\left(\hat{\boldsymbol{\beta}}_\tau\right)_j\right)^2\middle|\mathbf{Y}_\tau\right)\right)<\infty\quad\xi\text{- almost surely}\tag{7.50}$$

Combining all three parts now gives that

$$\lim_{\tau\to\infty}\mathbb{E}_\tau\left(\left|\hat{\bar{y}}_{\tau,\mathrm{GREG}}-\bar{y}_{U_\tau}\right|\middle|\mathbf{Y}_\tau\right)=0\quad\xi\text{- almost surely}\tag{7.51}$$

which completes the proof. □

7.4. The generalised regression estimator and the screening procedure

To discuss the effects of the screening procedure on the generalised regression estimator, we consider two different sampling designs: simple random sampling without replacement (see Section 2.1.1) and the two-stage self-weighting sampling design that was described in Section 3. We have seen that the generalised regression estimator is consistent and asymptotically unbiased under Conditions (C0) to (C5). For both sampling designs, we will check if the conditions are met and subsequently, we discuss the effects of the screening procedure on the generalised regression estimator.

Recall from Section 3 that the first-order inclusion probabilities are equal for the two sampling designs, but the second-order inclusion probabilities are different. Since only Condition (C5) is dependent on the second-order inclusion probabilities, we discuss Conditions (C0) until (C4) in general.

Condition (C0) The first condition is related to the singularity of the matrix $\sum_{k\in s} \mathbf{x}_k \mathbf{x}_k^T \frac{1}{v_k}$, which is dependent on the auxiliary variables that are used. In general, the auxiliary variables that are used at Statistics Netherlands are mostly categorical. Note that any categorical (or quantitive variable) can be replaced by a set of dummy variables. A dummy variable is equal to 1 of the inhabitant is in a category and 0 otherwise. If the matrix $\sum_{k\in s} \mathbf{x}_k \mathbf{x}_k^T \frac{1}{v_k}$ is singular, one can skip some dummy variables to prevent that $\sum_{k\in s} \mathbf{x}_k \mathbf{x}_k^T$ is singular [17].

Condition (C1) In general, mostly categorical variables are used at Statistics Netherlands. This means that each variable has a finite number of categories and that each inhabitant belongs to exactly one category. By using solely categorical variables, it is assured that the auxiliary variables are bounded, which gives that Condition (C1) is met. If quantitative auxiliary variables (variables that measure a phenomenon at a numerical scale) are used the condition is also met if the auxiliary variables are bounded. Furthermore, note that quantitative variables can always be replaced by a set of categorical variables.

Condition (C2) This condition is related to the target variables. At Statistics Netherlands target variables are always bounded, and usually categorical. Consequently, Condition (C2) is met.

Condition (C3) First, note that Condition (C3) can be written as

$$\limsup_{\tau\to\infty} \mathbb{E}_\tau \left(\left|\left| \hat{\boldsymbol{\beta}}_\tau \right|\right|^2 \middle| \mathbf{Y}_\tau \right) < \infty \quad \xi - \text{almost surely} \tag{7.52}$$

where

$$\left|\left| \hat{\boldsymbol{\beta}}_\tau \right|\right|^2 = \left|\left| \left(\hat{\mathbf{T}}_{\mathbf{xx},\mathsf{HT}}\right)^{-1} \hat{\mathbf{t}}_{\mathbf{x}y,\mathsf{HT}} \right|\right|^2 = \hat{\mathbf{t}}_{\mathbf{x}y,\mathsf{HT}}^T \left(\hat{\mathbf{T}}_{\mathbf{xx},\mathsf{HT}}\right)^{-1} \left(\hat{\mathbf{T}}_{\mathbf{xx},\mathsf{HT}}\right)^{-1} \hat{\mathbf{t}}_{\mathbf{x}y,\mathsf{HT}} \tag{7.53}$$

The norm of $\hat{\boldsymbol{\beta}}_\tau$ can be further rewritten as

$$\begin{aligned} \left|\left| \hat{\boldsymbol{\beta}}_\tau \right|\right|^2 &= \frac{\left|\left| \hat{\boldsymbol{\beta}}_\tau \right|\right|^2}{\left|\left| \frac{1}{N_\tau} \hat{\mathbf{t}}_{\mathbf{x}y,\mathsf{HT}} \right|\right|^2} \cdot \left|\left| \frac{1}{N_\tau} \hat{\mathbf{t}}_{\mathbf{x}y,\mathsf{HT}} \right|\right|^2 \\ &= \frac{\frac{1}{N_\tau} \hat{\mathbf{t}}_{\mathbf{x}y,\mathsf{HT}}^T \left(\frac{1}{N_\tau} \hat{\mathbf{T}}_{\mathbf{xx},\mathsf{HT}}\right)^{-1} \left(\frac{1}{N_\tau} \hat{\mathbf{T}}_{\mathbf{xx},\mathsf{HT}}\right)^{-1} \frac{1}{N_\tau} \hat{\mathbf{t}}_{\mathbf{x}y,\mathsf{HT}}}{\frac{1}{N_\tau} \hat{\mathbf{t}}_{\mathbf{x}y,\mathsf{HT}}^T \frac{1}{N_\tau} \hat{\mathbf{t}}_{\mathbf{x}y,\mathsf{HT}}} \cdot \left|\left| \frac{1}{N_\tau} \hat{\mathbf{t}}_{\mathbf{x}y,\mathsf{HT}} \right|\right|^2 \end{aligned} \tag{7.54}$$

We use this expression for the norm later. First we make a remark on the consistency of the estimator $\frac{1}{N_\tau}\hat{\mathbf{T}}_{\mathbf{xx},\mathsf{HT}}$ for the matrix $\frac{1}{N}\mathbf{T}_{\mathbf{xx}}$.

Remark 7.2 (Consistency of $\frac{1}{N_\tau}\hat{\mathbf{T}}_{\mathbf{xx},\mathsf{HT}}$)**.** The estimator $\frac{1}{N_\tau}\hat{\mathbf{T}}_{\mathbf{xx},\mathsf{HT}}$ is consistent for the matrix $\frac{1}{N_\tau}\mathbf{T}_{\mathbf{xx}}$.

Proof. The proof can be found in Appendix A.6.

Let the largest eigenvalue of the matrix $\mathbf{A}$ be denoted by $\lambda_{\mathsf{max}}(\mathbf{A})$, then [24]

$$\lambda_{\mathsf{max}}(\mathbf{A}) = \max_{||\mathbf{v}||^2=1} \mathbf{v}^T\mathbf{A}\mathbf{v} \tag{7.55}$$

This gives that the norm of $\hat{\boldsymbol{\beta}}_\tau$ is bounded by

$$\left|\left|\hat{\boldsymbol{\beta}}_\tau\right|\right|^2 \leq \lambda_{\mathsf{max}}\left(\left(\frac{1}{N_\tau}\hat{\mathbf{T}}_{\mathbf{xx},\mathsf{HT}}\right)^{-1}\left(\frac{1}{N_\tau}\hat{\mathbf{T}}_{\mathbf{xx},\mathsf{HT}}\right)^{-1}\right)\cdot\left|\left|\frac{1}{N_\tau}\hat{\mathbf{t}}_{\mathbf{x}y,\mathsf{HT}}\right|\right|^2 \tag{7.56}$$

Recall that if λ is an eigenvalue of $\mathbf{A}$ then λ^2 is an eigenvalue for $\mathbf{A}^2$ [24]. And for a matrix $\mathbf{A}$, the largest eigenvalue of $\mathbf{A}^{-1}$ is equal to the 1 divided by smallest eigenvalue of $\mathbf{A}$ [24]. This gives that

$$\left|\left|\hat{\boldsymbol{\beta}}_\tau\right|\right|^2 \leq \frac{1}{\left(\lambda_{\mathsf{min}}\left(\frac{1}{N_\tau}\hat{\mathbf{T}}_{\mathbf{xx},\mathsf{HT}}\right)\right)^2}\cdot\left|\left|\frac{1}{N_\tau}\hat{\mathbf{t}}_{\mathbf{x}y,\mathsf{HT}}\right|\right|^2 \tag{7.57}$$

We impose a new condition (C0*) on the positivity of smallest eigenvalue of the matrix $\frac{1}{N_\tau}\mathbf{T}_{\mathbf{xx}}$, i.e.

(C0*) $\exists\ \epsilon > 0$ such that $\liminf_{\tau\to\infty}\lambda_{\mathsf{min}}\left(\frac{1}{N_\tau}\mathbf{T}_{\mathbf{xx}}\right) \geq \epsilon$

Note that condition (C0*) implies condition (C0).

Condition (C0*) and the consistency of the estimator $\frac{1}{N_\tau}\hat{\mathbf{T}}_{\mathbf{xx},\mathsf{HT}}$ for the matrix $\frac{1}{N_\tau}\mathbf{T}_{\mathbf{xx}}$ (see Remark 7.2) imply that $\exists\ \epsilon > 0$ such that

$$\liminf_{\tau\to\infty}\lambda_{\mathsf{min}}\left(\frac{1}{N_\tau}\hat{\mathbf{T}}_{\mathbf{xx},\mathsf{HT}}\right) > \epsilon \tag{7.58}$$

Consequently,

$$\limsup_{\tau\to\infty}\mathbb{E}_\tau\left(\left|\left|\hat{\boldsymbol{\beta}}_\tau\right|\right|^2\middle|\mathbf{Y}_\tau\right) \leq \limsup_{\tau\to\infty}\mathbb{E}_\tau\left(\frac{1}{\epsilon^2}\cdot\left|\left|\frac{1}{N_\tau}\hat{\mathbf{t}}_{\mathbf{x}y,\mathsf{HT}}\right|\right|^2\middle|\mathbf{Y}_\tau\right) \tag{7.59}$$

We compute the expected value with respect to the design of the norm of the vector $\frac{1}{N_\tau}\hat{\mathbf{t}}_{\mathbf{x}y,\mathsf{HT}}$.

$$\mathbb{E}_\tau\left(\left|\left|\frac{1}{N_\tau}\hat{\mathbf{t}}_{\mathbf{x}y,\mathsf{HT}}\right|\right|^2\middle|\mathbf{Y}_\tau\right) = \mathbb{E}_\tau\left(\sum_{j=1}^{p}\left(\frac{1}{N_\tau}\sum_{k=1}^{N_\tau}\frac{I_{k\tau}}{\pi_{k\tau}}x_{kj}Y_k\frac{1}{\upsilon_k}\right)^2\middle|\mathbf{Y}_\tau\right) \tag{7.60}$$

By applying the Cauchy-Schwarz inequality (see Lemma A.3) to the right-hand side, we obtain

$$\begin{aligned}\mathbb{E}_\tau\left(\left|\left|\frac{1}{N_\tau}\hat{\mathbf{t}}_{\mathbf{x}y,\mathsf{HT}}\right|\right|^2\middle|\mathbf{Y}_\tau\right) &\leq \sum_{j=1}^{p}\left(\mathbb{E}_\tau\left(\frac{1}{N_\tau}\sum_{k=1}^{N_\tau}\frac{I_{k\tau}I_{k\tau}}{\pi_{k\tau}\pi_{k\tau}}x_{kj}^2\frac{1}{\upsilon_k}\middle|\mathbf{Y}_\tau\right)\cdot\mathbb{E}_\tau\left(\frac{1}{N_\tau}\sum_{k=1}^{N_\tau}Y_k^2\frac{1}{\upsilon_k}\middle|\mathbf{Y}_\tau\right)\right)\\ &= \left(\frac{1}{N_\tau}\sum_{k=1}^{N_\tau}Y_k^2\frac{1}{\upsilon_k}\right)\sum_{j=1}^{p}\mathbb{E}_\tau\left(\frac{1}{N_\tau}\sum_{k=1}^{N_\tau}\frac{I_{k\tau}}{\pi_{k\tau}^2}x_{kj}^2\frac{1}{\upsilon_k}\right)\end{aligned} \tag{7.61}$$

where

$$\mathbb{E}_\tau\left(\frac{1}{N}\sum_{k=1}^{N}\frac{I_{k\tau}}{\pi_{k\tau}^2}x_{kj}^2\frac{1}{v_k}\right)=\frac{1}{N}\sum_{k=1}^{N}\frac{1}{\pi_{k\tau}}x_{kj}^2\frac{1}{v_k}\leq\frac{1}{\liminf_{\tau\to\infty}\pi_{k\tau}}\frac{1}{N}\sum_{k=1}^{N}x_{kj}^2\frac{1}{v_k} \tag{7.62}$$

Recall in the introduction of the model (see Equation (7.5)) we have assumed that the auxiliary variable v_k is positive [16]. Consequently, if Condition (C1) holds, we also have that

$$\limsup_{\tau\to\infty}\frac{1}{N_\tau}\sum_{k=1}^{N_\tau}x_{kj}^2\frac{1}{v_k}<\infty\quad\text{for } j=1,\dots,p \tag{7.63}$$

Similarly, we can conclude

$$\limsup_{\tau\to\infty}\frac{1}{N_\tau}\sum_{k=1}^{N_\tau}Y_k^2\frac{1}{v_k}<\infty\quad\xi\text{-almost surely} \tag{7.64}$$

Assume that the infimum of the inclusion probabilities is strictly greater than zero as $\tau\to\infty$, i.e.

$$\liminf_{\tau\to\infty}\pi_{k\tau}>0 \tag{7.65}$$

Under this assumption and conditions (C0*), (C1), (C1*) (see Appendix A.6), (C2), (C4) and (C5) we can conclude that

$$\limsup_{\tau\to\infty}\mathbb{E}_\tau\left(\left|\left|\frac{1}{N_\tau}\hat{\mathbf{t}}_{\mathbf{x}y,\mathrm{HT}}\right|\right|^2\middle|\mathbf{Y}_\tau\right)<\infty \tag{7.66}$$

and hence that

$$\limsup_{\tau\to\infty}\mathbb{E}_\tau\left(\left|\left|\hat{\boldsymbol{\beta}}_\tau\right|\right|^2\middle|\mathbf{Y}_\tau\right)<\infty\quad\xi-\text{almost surely} \tag{7.67}$$

To sum up, we have shown that Condition (C3) is met under some extra modest conditions.

Condition (C4) In Section 3.1 we have seen that the first order inclusion probability is equal to the sampling fraction for all inhabitants for both sampling designs. In the super-population model this is denoted by

$$\pi_{k\tau}=f_\tau=\frac{n_\tau}{N_\tau} \tag{7.68}$$

Then for condition (C4) we have

$$\liminf_{\tau\to\infty}N_\tau\min_{1\leq k\leq N_\tau}\pi_{k\tau}=\liminf_{\tau\to\infty}N_\tau\frac{n_\tau}{N_\tau}=\liminf_{\tau\to\infty}n_\tau=\infty \tag{7.69}$$

since in the super-population model $\tau\to\infty$ means that $n_\tau\to\infty$.

7.4.1. Simple random sampling without replacement

It remains so check if Condition (C5) is met for SRSWR.

Condition (C5) Using the second-order inclusion probability we have computed in Section 2.1.1, we find

$$\left|\frac{\pi_{kl\tau}}{\pi_{k\tau}\pi_{l\tau}}-1\right|=\left|\frac{N_\tau}{n_\tau}\frac{N_\tau}{n_\tau}\frac{n_\tau(n_\tau-1)}{N_\tau(N_\tau-1)}-1\right|=\left|\frac{n_\tau-1}{n_\tau}\frac{N_\tau}{N_\tau-1}-1\right| \tag{7.70}$$

Recall that $\tau\to\infty$ means that $n_\tau\to\infty$ and $N_\tau\to\infty$, so it follows that Condition (C5) is met

$$\lim_{\tau\to\infty}\max_{1\leq k\neq l\leq N_\tau}\left|\frac{\pi_{kl\tau}}{\pi_{k\tau}\pi_{l\tau}}-1\right|=0 \tag{7.71}$$

We have seen that under extra modest assumptions for simple random sampling without replacement all conditions are met. This means that the generalised regression estimator is consistent and asymptotically unbiased for simple random sampling without replacement.

By applying the screening procedure on the sample, second-order inclusion probabilities are zero for inhabitants who live on the same address. This would mean that by applying the screening procedure Condition (C5) is not met and consequently, we cannot conclude that the generalised regression estimator is consistent and asymptotically unbiased in this case.

7.4.2. Two-stage self-weighting sampling design

We check if Condition (C5) is met for the two-stage self-weighting sampling design.

Condition (C5) Recall that the second order inclusion probability of two inhabitants is dependent on the municipalities those inhabitants live in, see Section 3.5.2. We assume that the size of municipalities increases to infinity but the number of municipalities does not increase as $\tau \to \infty$, i.e.

$$\lim_{\tau\to\infty} N_{i\tau} = \infty \quad \forall\, i = 1, \dots, I \tag{7.72}$$

In Remark 3.2 we have explained that if the cluster size m is equal to 1 and if the sample size is large enough almost all municipalities become self-selecting. So since $n_\tau \to \infty$ as $\tau \to \infty$ we can assume that all municipalities are self-selecting in the super-population model. So for verifying Condition (C5), we assume that inhabitants k and l are in self-selecting municipalities.

First suppose that inhabitants $k \neq l$ are not in the same municipality U_i. Recall from Section 3.5 that if inhabitants $k \neq l$ are not in the same municipality we have that the second-order inclusion probability is equal to

$$\pi_{kl\tau} = \frac{n_\tau}{N_\tau}\frac{n_\tau}{N_\tau} \tag{7.73}$$

Then we simply have

$$\frac{\pi_{kl\tau}}{\pi_{k\tau}\pi_{l\tau}} = \frac{N_\tau}{n_\tau}\frac{N_\tau}{n_\tau}\frac{n_\tau}{N_\tau}\frac{n_\tau}{N_\tau} = 1 \tag{7.74}$$

It easily follows that Condition (C5) is met.

Suppose k and l are in the same municipality U_i. In the super-population model, the second-order inclusion probability is

$$\pi_{kl\tau} = \frac{n_{i\tau}\,(n_{i\tau}-1)}{N_{i\tau}\,(N_{i\tau}-1)} \tag{7.75}$$

where $n_{i\tau} = N_{i\tau}\frac{n_\tau}{N_\tau}$. Note that the sample size $n_{i\tau}$ is chosen such that $\frac{n_{i\tau}}{N_{i\tau}} = \frac{n_\tau}{N_\tau}$. Consequently, we have

$$\frac{\pi_{kl\tau}}{\pi_{k\tau}\pi_{l\tau}} = \frac{N_\tau}{n_\tau}\frac{N_\tau}{n_\tau}\frac{n_{i\tau}\,(n_{i\tau}-1)}{N_{i\tau}\,(N_{i\tau}-1)} = \frac{N_\tau}{n_\tau}\frac{n_{i\tau}-1}{N_{i\tau}-1} = \frac{N_{i\tau}-\frac{N_\tau}{n_\tau}}{N_{i\tau}-1} \tag{7.76}$$

Then

$$\left|\frac{\pi_{kl\tau}}{\pi_{k\tau}\pi_{l\tau}} - 1\right| = \left|\frac{N_{i\tau}-\frac{N_\tau}{n_\tau}}{N_{i\tau}-1} - 1\right| = \left|\frac{N_{i\tau}-\frac{N_\tau}{n_\tau}-N_{i\tau}+1}{N_{i\tau}-1}\right| = \left|\frac{1-\frac{N_\tau}{n_\tau}}{N_{i\tau}-1}\right| \tag{7.77}$$

Assume that the population size divided by the sample size is bounded as $\tau \to \infty$ i.e.

$$\lim_{\tau\to\infty} \frac{N_\tau}{n_\tau} < \infty \tag{7.78}$$

Under these assumptions stated in Equation (7.72) and in Equation (7.78), we have

$$\lim_{\tau\to\infty} \max_{1\le k\neq l\le N_\tau} \left|\frac{\pi_{kl\tau}}{\pi_{k\tau}\pi_{l\tau}} - 1\right| = 0 \tag{7.79}$$

To sum up, it follows that Condition (C5) is met for the two-stage self-weighting sampling design if we assume that all municipalities are self-selecting as $\tau \to \infty$. Consequently, under the assumptions that all municipalities become self-selecting the generalised regression estimator is consistent and asymptotically unbiased for the two-stage self-weighting sampling design of Section 3. If there are still non-self-selecting municipalities Condition (C5) is not met.

Again, by applying the screening procedure to the sample, second order inclusion probabilities become zero for inhabitants that live on the same address. This would mean that Condition (C5) is not met. Hence if the screening procedure is applied, we cannot conclude that the generalised regression estimator is consistent and asymptotically unbiased anymore.

7.5. Conclusion

Statistics Netherlands uses the generalised regression estimator to estimate unknown population characteristics, because its approximated variance is smaller than or equal to the variance of the Horvitz-Thompson estimator. However, the generalised regression estimator is not unbiased. We have shown that under specific conditions, the generalised regression estimator is consistent and asymptotically design unbiased. We have shown that if simple random sampling without replacement is used for selecting the samples, all conditions are met. Furthermore we have shown that under some extra modest assumptions, the conditions are met if the self-weighting two-stage sampling design is used for selecting a sample.

Applying the screening procedure to the sample causes that Condition (C5) is not met, despite the sampling design that is used. This means that when the screening procedure is applied, we cannot conclude that the generalised regression estimator is consistent and asymptotically unbiased.

8

Correcting for the screening procedure during estimation

The main goal of survey sampling is to estimate properties of the population that are unknown. For example, suppose the goal is to obtain an accurate estimate for the population mean of target variable y. The value for this variable y is only known for the elements that are included in the sample s. We will assume that every element in the sample s responds to the survey, so nonresponse does not occur. There are different types of estimators that can be used to estimate the mean of y; the Horvitz-Thompson estimator and the generalised regression estimator are two examples.

In the previous sections, we have seen that not all inhabitants have the same probability to be eligible in the sample after screening. Thus far, it is assumed that the effects of the screening procedure are negligible and hence during the estimation of population characteristics the screening procedure is not taken into consideration.

The accuracy of the conclusions of a survey is to a large extent based on the choice of the sampling design and the estimator [3]. One combination of an estimator and a sampling design may lead to more precise estimates than other combinations. For example, the advantages of a well-chosen sampling design can be undone by a badly chosen estimator. On the other hand, a badly chosen sampling design can be undone by using an effective estimator [3].

In this section we discuss two possibilities for estimators that can be used to undo the effects of the screening procedure. We focus on the screening on the occurrence of an address. We use simulated target variables to show how the Horvitz-Thompson estimator and the generalised regression estimator can be used to correct for the screening on the occurrence of an address during the estimation of population characteristics. The estimation is done using the `survey`-package in `R` [25].

8.1. Simulating target variables

In Chapter 4.2.1 we have showed that the probability that an inhabitant is eligible after screening on the occurrence of an address is dependent on the number of people on an address. If the target variable y is related to the number of people on address, the screening procedure will affect the estimated population characteristic.

Let $\mathbf{a} = (a_1, \ldots, a_N)$ denote the vector of the number of elements on an address for each element in the population. We will generate six different variables that all have a different positive correlation with $\mathbf{a}$. We use those variables to show the effects of screening on the estimated values for the population mean. Furthermore, we use those variables to show how the Horvitz-Thompson estimator and the generalised regression estimator can be used to correct for the screening on the occurrence of an address.

Let ρ be the correlation between $\mathbf{a}$ and the target variable $\mathbf{V}_\rho$. Let $\tilde{\mathbf{a}}$ denote the standardised vector of $\mathbf{a}$ such that the sum of $\tilde{\mathbf{a}}$ is zero. Let $\mathbf{Z}$ be a vector containing N realisations of a standard normally distributed variable $Z \sim N(0, 1)$. If the vectors $\mathbf{Z}$ and $\tilde{\mathbf{a}}$ are not correlated, then the vector of target variables for all elements in the population is generated by [26]

$$\mathbf{V}_\rho = \rho \cdot \tilde{\mathbf{a}} + \sqrt{1 - \rho^2} \cdot \mathbf{Z} \tag{8.1}$$

Then Cor $\left(\mathbf{a}, \mathbf{V}_\rho\right) = \rho$. We can now rescale the vector $\mathbf{V}_\rho$ with any mean μ and any standard deviation σ. For each variable, we choose $\mu = 10$ and $\sigma = 3$. Note that rescaling the vector does not change the correlation with $\tilde{\mathbf{a}}$. We generate six different target variables, such that the correlation with the number of people on an address are approximately $0, 0.2, 0.4, 0.6, 0.8$ and 1 respectively.

8.2. The Horvitz-Thompson estimator

The Horvitz-Thompson estimator makes no explicit use of auxiliary information, but it only makes use of the inclusion probabilities. Under the assumption that the effects of the screening are negligible, the Horvitz-Thompson estimator with equal inclusion probabilities will give an accurate estimate for the population mean. However, we have seen that one should be cautious with assuming that the screening procedure is negligible, because inhabitants have unequal probabilities to be eligible in the sample after screening.

In Chapter 4.2.1 we have computed the inclusion probability of an inhabitant in the sample after screening if only screening on the occurrence of an address is applied. We can use these adjusted inclusion probabilities to undo the effects of the screening on the occurrence of an address during the estimation of population characteristics.

Note that for equal inclusion probabilities, the Horvitz-Thompson estimator of the population mean is equal to the sample mean. For unequal probabilities the Horvitz-Thompson estimator is a weighted average, where the weight of an inhabitant is $\frac{1}{\pi_k}$. By using the adjusted inclusion probabilities from Section 4.2.1, inhabitants who live with multiple people on one address have a relatively lower inclusion probability in the sample after screening. Consequently, the weight of inhabitants who live with multiple people on an address is relatively higher.

We have generated the six target variables $\mathbf{V}_0, \mathbf{V}_{0.2}, \mathbf{V}_{0.4}, \mathbf{V}_{0.6}, \mathbf{V}_{0.8}$ and $\mathbf{V}_1$ for the population that was used to select the mobility survey of April 2019. This is the same sample that we have discussed in Section 5.4. For each target variable, we have computed the population mean and the mean in the sample after screening with equal and the adjusted inclusion probabilities. The obtained estimates for this sample are presented in Table 8.1. For all variables, the population mean is approximately 10, but the mean in the sample after screening with equal probabilities is lower than 10, especially if the correlation with the number of people on an address is high. If the correlation with the number of people on an address is high, the estimates obtained from the Horvitz-Thompson estimator with adjusted probabilities is close to 10. This shows that by using the adjusted inclusion probabilities from Section 4.2.1, the obtained estimates are closer to the value to be estimated than by using equal

inclusion probabilities. This means that if the correlation between the target variable and the number of people on an address is high, we can at least partly undo the effects of the screening procedure by using the adjusted inclusion probabilities during the estimation of population characteristics.

	$\mathbf{V}_0$	$\mathbf{V}_{0.2}$	$\mathbf{V}_{0.4}$	$\mathbf{V}_{0.6}$	$\mathbf{V}_{0.8}$	$\mathbf{V}_1$
Mean population	9.9998	9.9945	9.9991	9.9998	9.9992	10.0000
HT-estimator equal probabilities	9.9547	9.9861	9.9920	9.9365	9.9450	9.9311
HT-estimator adjusted probabilities	9.9479	9.9995	10.0174	9.9685	9.9874	9.9925

Table 8.1: Results of estimating the population mean of the mobility survey of April 2019 by using the Horvitz-Thompson estimator with equal probabilities and the adjusted probabilities.

Note that results are different for different samples due to sampling fluctuations. We have applied the Horvitz-Thompson estimator with equal and adjusted probabilities to several samples of the mobility survey. These results are presented in Appendix F. These results imply that if the correlation between the target variable and the number of people on an address is not zero, the use of the adjusted inclusion probabilities improves the estimated values of the population mean with respect to the Horvitz-Thompson estimates with equal inclusion probabilities.

8.3. The generalised regression estimator

Instead of using the adjusted inclusion probabilities, we can also make use of auxiliary information to undo the effects of the screening procedure during estimation of population characteristics. The generalised regression estimator allows us to correct for screening using auxiliary information. We have applied the generalised regression estimator to the mobility survey of April 2019 for several different weighting models each consisting of one auxiliary variable. The results are presented in Table 8.2. Recall from Section 7.1 that the generalised regression estimator requires the input of the inclusion probabilities. To see the effects of the different auxiliary variables, we have used equal inclusion probabilities for obtaining these estimates. Note that it also possible to use the adjusted inclusion probabilities from Section 4.2.1.

From Table 8.2 we can conclude that by using the number of people on an address as an auxiliary variable for the generalised regression estimator, the obtained estimates are close to the true values, if the correlation between the target variable and the number of people on an address is high. Note that the use of other variables result in estimates with a larger error.

Again, results are different for other samples. In Appendix F the results for different samples of the mobility survey are presented. Results are similar to the results we have presented for the mobility survey of April 2019. The best estimates are obtained when using the number of people on an address as an auxiliary variable for the generalised regression estimator.

8.4. Horvitz-Thompson estimator vs. generalised regression estimator

The simulation study we did in Section 6 allows us to apply the Horvitz-Thompson estimator and the generalised regression estimators to many different samples. We have selected 7,000 different samples by the simulation that are similar to the mobility survey of April 2019. Each sample undergoes the screening procedure and subsequently, the population mean of the variables $\mathbf{V}_0, \mathbf{V}_{0.2}, \mathbf{V}_{0.4}, \mathbf{V}_{0.6}, \mathbf{V}_{0.8}, \mathbf{V}_1$

	$\mathbf{V}_0$	$\mathbf{V}_{0.2}$	$\mathbf{V}_{0.4}$	$\mathbf{V}_{0.6}$	$\mathbf{V}_{0.8}$	$\mathbf{V}_1$
Mean population	9.9998	9.9945	9.9991	9.9998	9.9992	10.0000
Mean sample after screening	9.9547	9.9861	9.9920	9.9365	9.9450	9.9311
Gender	9.9547	9.9859	9.9917	9.9367	9.9446	9.9311
Marital status	9.9562	9.9864	9.9929	9.9369	9.9450	9.9317
Age	9.9541	9.9861	9.9938	9.9371	9.9464	9.9344
Ethnicity	9.9552	9.9862	9.9917	9.9368	9.9456	9.9316
Place Household	9.9563	9.9871	9.9943	9.9383	9.9468	9.9337
Type Household	9.9566	9.9863	9.9931	9.9378	9.9464	9.9332
Number of people in household	9.9683	9.9692	10.0010	10.0137	9.9355	9.9352
Number of people in household 11[a]	9.9681	9.9692	10.0009	10.0137	9.9354	9.9353
Number of people on address	9.9472	9.9995	10.0214	9.9731	9.9942	10.0000
Number of people on address 11[b]	9.9517	9.9898	9.9969	9.9429	9.9566	9.9449

Table 8.2: Estimated means by the generalised regression estimator for different auxiliary variables. Based on the sample of the mobility survey of April 2019.
[a] This is a categorical variable of the number of people in a household consisting of eleven categories: 1 to 10 and 11 or more.
[b] This is a categorical variable of the number of people on an address consisting of eleven categories: 1 to 10 and 11 or more.

are estimated by the Horvitz-Thompson estimator with equal and adjusted probabilities, and by the generalised regression estimator with the number of people on an address as an auxiliary variable. The results are represented in Table 8.3 and Figure 8.1. In Table 8.3 we have presented the average of the 7,000 estimated values for the population mean for the three different estimators.

These results imply that the use of the number of people on an address as an auxiliary variable for the generalised regression estimator undo the effects of the screening better than using the adjusted inclusion probabilities for the Horvitz-Thompson estimator. We presume this is because the adjusted inclusion probabilities are an approximation for the actual probability. We have not investigated this any further.

	$\mathbf{V}_0$	$\mathbf{V}_{0.2}$	$\mathbf{V}_{0.4}$	$\mathbf{V}_{0.6}$	$\mathbf{V}_{0.8}$	$\mathbf{V}_1$
Population mean	10.0021	9.9985	10.0004	9.9961	9.9968	10.0000
HT-estimator equal probabilities	10.0012	9.9841	9.9704	9.9526	9.9382	9.9271
HT-estimator adjusted probabilities	10.0011	9.9898	9.9823	9.9699	9.9614	9.9565
GREG-estimator	10.0009	9.9980	9.9999	9.9961	9.9957	10.0000

Table 8.3: Average of the estimated values for the population mean based on 7,000 different samples. Here 'HT-estimator' is an abbreviation for Horvitz-Thompson estimator and 'GREG' is an abbreviation for generalised regression estimator.

8.5. Conclusion

We have showed how the Horvitz-Thompson estimator and the generalised regression estimator can be used to undo the effects from the screening on the occurrence of an address during the estimation of population characteristics. By using the adjusted inclusion probabilities we have derived in Section 4.2.1 we can use the Horvitz-Thompson estimator to undo the effects of the screening procedure. Our simulation study showed that better results are obtained when using the generalised regression

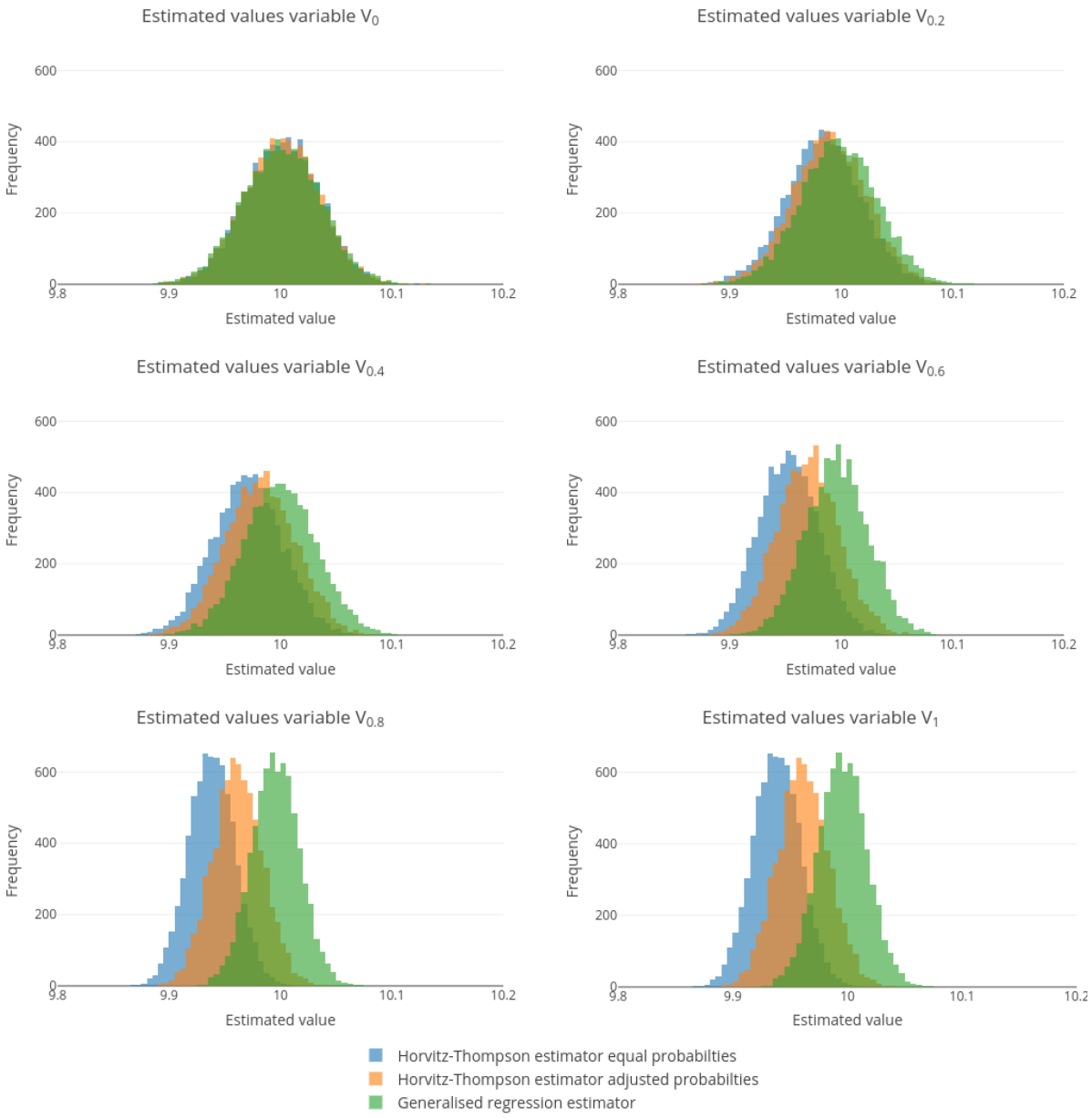


Figure 8.1: Estimated values of the population mean for the six simulated target variables for 7,000 different realisations that were obtained by the simulation study that was described in Section 6.

estimator with the number of people on an address as an auxiliary variable.

It is important to make some comments on these conclusions though. We have only imposed a way to undo the effects of the screening on the occurrence of an address, but not for the other parts of the screening. Furthermore we have assumed that nonresponse does not occur, whereas in reality nonresponse does occur. The rate of response differs for the surveys and is dependent on the type of interviewing that is applied. For example, there are surveys for which Computer-Assisted Web Interviewing is used that only have a response rate of approximately 30%. The nonresponse does presumably have more effects than the screening procedure.

9

Conclusion and discussion

In this thesis, we have presented the consequences of the screening procedure that is applied at Statistics Netherlands. After describing the problem, we provided the reader with necessary background information about survey sampling and the sampling design that is used by Statistics Netherlands for selecting samples. We have seen that the self-weighting two-stage sampling design that is used by Statistics Netherlands is in many aspects similar to simple random sampling without replacement. First-order inclusion probabilities are equal for both sampling designs, but the second-order inclusion probabilities are different. We assumed that if the cluster size is equal to one and the sample size is large enough, simple random sampling without replacement can be used as an approximation for the self-weighting two-stage sampling design. We did not further investigate the effects of using simple random sampling without replacement as an approximation to the self-weighting two-stage sampling design.

The screening procedure is applied to the selected samples to make sure the surveys are equally spread among the Dutch households. Most inhabitants that become not eligible by the screening procedure are not eligible by the screening on the occurrence of an address. Furthermore, inhabitants can become not eligible by the screening on confidential information or by other reasons that can be considered negligible.

Under simplifying assumptions, we derived an approximation for the conditional probability that an inhabitant is eligible after the screening on the occurrence of an address, given that the inhabitant was included in the sample before screening. This probability is dependent on the number of people on an address, which gives that the probability is not equal for all inhabitants in the Dutch population. Consequently, the inclusion probability for an inhabitant in the sample after screening is not equal for all inhabitants. Moreover, we have seen that the conditional probability is dependent on the sampling fraction, meaning that if the sampling fraction increases, the effects of the screening procedure become larger.

Based on the figures, there seems to be a relationship between the confidentiality and the auxiliary variables. We have not investigated the relationship in depth. The figures suggest that inhabitants do not have an equal probability to become not eligible by confidential information.

To identify the effects of the screening procedure on existing samples, we have compared the distribution of auxiliary variables in the population, in the sample before screening and in the sample after screening. We have developed two different statistical tests that allow us to determine whether the sample after screening can be considered representative for the population and for the sample before screening, with respect to a given auxiliary variable.

In general, we concluded that the inhabitants that are not eligible after screening are not representative for the population with respect to relevant auxiliary variables. For some samples this may cause that the sample after screening is not representative for the population, but for others the sample after screening can still be considered representative for the population. If the sample size increases, the effects of the screening procedure become larger and the sample after screening is not representative for the population with respect to the observed auxiliary variables more often.

Furthermore, for the samples we applied the second test to, the sample after screening cannot be considered representative for the sample before screening with respect to most auxiliary variables. Because applying the second test is computationally intensive, we did not perform this to a lot of samples. Additionally, more experiments will have to be conducted to derive solid conclusions on the second hypothesis.

The main aim of survey sampling is to obtain accurate estimates for unknown population characteristics. Statistics Netherlands uses the generalised regression estimator for obtaining these estimates. The approximated variance of the generalised regression estimator is smaller than or equal to the variance of the Horvitz-Thompson estimator, but the generalised regression estimator is not unbiased. We have shown that under modest conditions, the generalised regression estimator is consistent and asymptotically unbiased. Under extra modest assumptions, these conditions are met for the sampling design that is used by Statistics Netherlands.

Applying the screening procedure to the sample causes that Condition (C5) is not met. Consequently, if the screening procedure is applied, we cannot conclude that the generalised regression estimator is consistent and asymptotically unbiased.

During the estimation of population characteristics, the Horvitz-Thompson estimator and the generalised regression estimator can be used to undo the effects of the screening on the occurrence of an address. By using the adjusted inclusion probabilities we have computed in Section 4.2.1 the inhabitants that have a lower probability to be selected in the sample after screening obtain a higher weight that is used for estimation by the Horvitz-Thompson estimator. Furthermore, the number of people on an address can be used as an auxiliary variable for the generalised regression estimator.

Under the assumption that nonresponse does not occur, we have investigated the effects of the screening procedure on six different simulated target variables. We introduced two ways to undo the effects of the screening on the occurrence of an address, but we did not consider the other parts of the screening procedure.

To sum up, we can conclude that it is not fair to assume that the effects of the screening procedure are negligible. First of all, we have seen that the probability that an inhabitant is selected in the sample after screening is not equal for all inhabitants. Secondly, we have shown that for existing samples the sample after screening is often not representative for the population with respect to relevant auxiliary variables and that the effects of the screening procedure become larger as the sample size increases. Finally, we showed that if screening is applied, we cannot conclude that the generalised regression

estimator is consistent and asymptotically unbiased.

However, we should not forget that throughout this thesis we have assumed nonresponse does not occur. In many cases the amount of inhabitants that do not respond to the questionnaire is greater than the amount of inhabitants that become not eligible by the screening procedure. Consequently, the nonresponse may have greater effects on the sample than the screening procedure.

We believe that our work makes useful contributions to the evaluation of the screening procedure. Our approach may not be perfect in various aspects and many interesting directions for future work remain to be explored.

10

Future work

In the previous chapters, we have discussed the effects of the screening procedure. Nonetheless, there are still many improvements that can be made. In this chapter, we will present out suggestions for future work on the matter.

10.1. Sampling design

We assumed that if the cluster size is equal to one and the sample size is large enough, simple random sampling without replacement can be used as an approximation for the self-weighting two-stage sampling design we described in Section 3. The differences of these two sampling designs could be further investigated. Moreover, if the two sampling designs turns out to be similar, one could investigate if simple random sampling without replacement can be used instead of the self-weighting two-stage sampling design. This would simplify further calculations and implementations, which would probably outweigh the advantages of the self-weighting two-stage sampling design.

10.2. Approximation adjusted inclusion probabilities

Under simplifying assumptions we have derived an approximation of the probability that an inhabitant becomes not eligible by the screening on the occurrence of an address. Potential improvements on this matter are:

- Although plots might imply that there are some relationships between the auxiliary variables and the confidentiality indicator, the specific patterns are not investigated. We could estimate the probability that an inhabitant becomes not eligible by confidential information. We presume that logistic regression methods can be used for this.

- We have chosen derive an approximation under simplifying assumptions, to make calculations easier. It is interesting to examine the probability that an inhabitant become not eligible by the occurrence of an address under different assumptions. For example, it may be interesting to compute the probability under the self-weighting two-stage sampling design.

10.3. Statistical testing

In the execution of the statistical tests corresponding to hypothesis 1 and hypothesis 2, we assumed that the distributions of the test statistics are unknown. Consequently, we have used a parametric bootstrap approach to obtain an estimate of the $\mathbf{p}$-values. We have assumed that the underlying data is distributed according to the multivariate hypergeometric distribution. From Norman L. Johnson [14] we know that the variance and the covariance of the univariate hypergeometric distribution are known. Hence one can presumably derive the distribution of the test statistics, which would make the use of parametric bootstrap unnecessary.

10.4. Correcting for the screening procedure during estimation

In Chapter 8, we have discussed two possibilities that can be used to undo the effects of the screening procedure. The results of the simulation study imply that using the number of people on an address as an auxiliary variable for the generalised regression estimator gives better estimates for the population mean than using the adjusted inclusion probabilities for the Horvitz-Thompson estimator. We did not investigate this any further. To be able to draw any further conclusions, it is important to be able to explain these differences.

10.5. Nonresponse

Throughout our work we assumed that nonresponse does not occur. It should be investigated how large effects of nonresponse are relative to the effects of the screening procedure. Statistics Netherlands uses several methods to undo the effects of nonresponse already. It is interesting to investigate if the same methods can be used to undo the effects of the screening procedure.

Bibliography

[1] Statistics Netherlands (CBS). Organisation. https://www.cbs.nl/en-gb/about-us/organisation, aug 2019.

[2] Statistics Netherlands (CBS). Onderzoeksbeschrijving odin 2018. *Internal document Statistics Netherlands (CBS)*, 2019.

[3] Jelke Bethlehem. *Applied Survey Methods - A Statistical Perspective*. Series in survey methodology. John Wiley and Sons Inc., Hoboken New Jersey, 2009.

[4] Carl Erik Särndal, Bengt Swensson, and Jan Wretman. *Model Assisted Survey Sampling*. Springer Series in Statistics, 1992.

[5] Wayne A. Fuller. *Sampling Statistics*. Series in survey methodology. John Wiley and Sons Inc., Iowa State University, 2009.

[6] Statistics Netherlands (CBS). Benaderingsstrategieën, algemene beschrijving en uni-mode designs. *Statische Methoden (10002)*, 2010.

[7] D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):pp. 663–685, Dec 1952.

[8] Statistics Netherlands (CBS). Statline. https://opendata.cbs.nl/statline/, sep 2019.

[9] Ministerie van Binnenlandse Zaken en Koninkrijksrelaties. Regioatlas. https://www.regioatlas.nl/indelingen/, okt 2019.

[10] E. Vondenhoff and C.A.M. van Berkel. Approximations of second order inclusion probabilities for dollar unit sampling. *Statistics Netherlands, Division of Data-Collection*. Internal CBS document.

[11] Rijksoverheid. Waar vraag ik het niet doorgeven van persoonsgegevens uit de brp (geheimhouding) aan?, aug 2019. URL `https://www.rijksoverheid.nl/onderwerpen/privacy-en-persoonsgegevens/vraag-en-antwoord/waar-vraag-ik-geheimhouding-van-mijn-persoonsgegevens-aan`.

[12] Rijksoverheid. Wet op het centraal bureau voor de statistiek, okt 2019. URL `https://wetten.overheid.nl/BWBR0015926/2018-07-01`.

[13] Statistics Netherlands (CBS). Over de veiligheidsmonitor, okt 2019. URL `http://www.veiligheidsmonitor.nl/Veiligheidsmonitor/Over_de_Veiligheidsmonitor`.

[14] N. Balakrishnan Norman L. Johnson, Samuel Kotz. *Discrete Multivariate Distributions*. John Wiley & Sons, Inc., 1996.

[15] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

[16] Roger L. Wright. *Finite Population Sampling With Multivariate Auxiliary Information*, volume 78 of *384*. Taylor & Francis, Ltd. on behalf of the American Statistical Association, https://www.jstor.org/stable/2288199, dec 1983. pp. 879-884.

[17] Nico Nieuwenbroek and Harm Jan Boonstra. Bascula 4.0 reference manual. *Statistics Netherlands Division Technology and Facilities Department of Methods and Informatics*, 2001. Project no.: RSM-80810, BPA no.: 3554-99-RSM.

[18] Jelke G. Bethlehem and Wouter J. Keller. Linear weighting of sample survey data. *Journal of official Statistics*, 3(2):141–153, 1987.

[19] Carl Erik Särndal. On π-inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika*, 67(3):639–650, 1980.

[20] Cary T. Isaki and Wayne A. Fuller. Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77(377):89–96, 1982.

[21] H.O. Hartley and Robert L. Sielken Jr. A" super-population viewpoint" for finite population sampling. *Biometrics*, pages 411–422, 1975.

[22] Susana Rubin-Bleuer, Ioana Schiopu Kratina, et al. On the two-phase framework for joint model and design-based inference. *The Annals of Statistics*, 33(6):2789–2810, 2005.

[23] P. M. Robinson and Carl Erik Särndal. *Asymptotic Properties of the Generalized Regression Estimator in Probability Sampling*, volume 45 of *The Indian Journal of Statistics, Series B (1960-2002)*. Indian Statistical Institute, https://www.jstor.org/stable/25052292, Aug., 1983. pp. 240-248.

[24] John B. Fraleigh and Raymond A. Beauregard. *Linear Algebra*. Addison-Wesley Publishing Company, 3rd edition, 1995.

[25] Thomas Lumley. *Complex surveys: a guide to analysis using R*, volume 565. John Wiley & Sons, 2011.

[26] Henry F. Kaiser and Kern Dickman. Sample and population score matrices and sample correlation matrices from an arbitrary population correlation matrix. *Psychometrika*, 27(2):179–182, 1962.

[27] Allan Gut. *Probability: A Graduate Course*. Springer Text in Statistics, second edition, 2013. ISBN 978-1-4614-4707-8.

A

Proofs

A.1. Proof of Theorem 2.2

We follow the proof that was given by Särndal et al. [4].

The variance of the Horvitz-Thompson estimator is given by

$$\begin{aligned}
\mathbb{V}\left(\hat{\bar{y}}_{\mathsf{HT}}\right) &= \mathbb{V}\left(\frac{1}{N}\sum_{k=1}^{N} I_k \frac{y_k}{\pi_k}\right) \\
&= \frac{1}{N^2}\sum_{k=1}^{N}\sum_{l=1}^{N} \mathsf{Cov}\left(I_k \frac{y_k}{\pi_k}, I_l \frac{y_l}{\pi_l}\right) \\
&= \frac{1}{N^2}\sum_{k=1}^{N}\sum_{l=1}^{N} \mathsf{Cov}\left(I_k, I_l\right) \breve{y}_k \breve{y}_l \\
&= \frac{1}{N^2}\sum_{k=1}^{N} \mathbb{V}\left(I_k\right) \breve{y}_k^2 + \frac{1}{N^2}\sum_{k=1}^{N}\sum_{k\neq l=1}^{N} \mathsf{Cov}\left(I_k, I_l\right) \breve{y}_k \breve{y}_l
\end{aligned} \tag{A.1}$$

In Equations (2.11) and (2.12) we have computed the variance and the covariance of the sample membership indicators, which gives

$$\begin{aligned}
\mathbb{V}\left(\hat{\bar{y}}_{\mathsf{HT}}\right) &= \frac{1}{N^2}\sum_{k=1}^{N} \pi_k\left(1-\pi_k\right) \breve{y}_k^2 + \frac{1}{N^2}\sum_{k=1}^{N}\sum_{k\neq l=1}^{N} \left(\pi_{kl} - \pi_k\pi_l\right) \breve{y}_k \breve{y}_l \\
&= \frac{1}{N^2}\sum_{k=1}^{N}\sum_{l=1}^{N} \left(\pi_{kl} - \pi_k\pi_l\right) \breve{y}_k \breve{y}_l
\end{aligned} \tag{A.2}$$

□

A.2. Proof of Theorem 2.3

We follow the proof that was given by Särndal et al. [4].

Developing the square in Equation (2.33) gives the expression for the variance that was stated in Theorem 2.2.

$$\begin{aligned}
\mathbb{V}\left(\hat{\bar{y}}_{\mathsf{HT}}\right) &= -\frac{1}{2N^2}\sum_{k=1}^{N}\sum_{l=1}^{N}\left(\pi_{kl}-\pi_k\pi_l\right)\left(\breve{y}_k-\breve{y}_l\right)^2 \\
&= -\frac{1}{2N^2}\sum_{k=1}^{N}\sum_{l=1}^{N}\left(\pi_{kl}-\pi_k\pi_l\right)\left(\breve{y}_k{}^2-2\breve{y}_k\breve{y}_l+\breve{y}_l{}^2\right) \\
&= \frac{1}{N^2}\sum_{k=1}^{N}\sum_{l=1}^{N}\left(\pi_{kl}-\pi_k\pi_l\right)\breve{y}_k\breve{y}_l-\frac{1}{N^2}\sum_{k=1}^{N}\sum_{l=1}^{N}\left(\pi_{kl}-\pi_k\pi_l\right)\breve{y}_k{}^2 \\
&= \frac{1}{N^2}\sum_{k=1}^{N}\sum_{l=1}^{N}\left(\pi_{kl}-\pi_k\pi_l\right)\breve{y}_k\breve{y}_l-\frac{1}{N^2}\sum_{k=1}^{N}\left(\breve{y}_k{}^2\sum_{l=1}^{N}\left(\pi_{kl}-\pi_k\pi_l\right)\right)
\end{aligned} \tag{A.3}$$

Note that since for a fixed sample design it holds that $\sum_{k=1}^{N} I_k = n$ (see Equation (2.13)), we have

$$\begin{aligned}
\sum_{l=1}^{N}\pi_{kl} &= \pi_{kk}+\sum_{k\neq l=1}^{N}\pi_{kl} \\
&= \pi_k+\sum_{k\neq l=1}^{N}\mathbb{E}\left(I_kI_l\right) \\
&= \pi_k+\mathbb{E}\left(I_k\left(\sum_{k\neq l=1}^{N}I_l\right)\right) \\
&= \pi_k+(n-1)\mathbb{E}\left(I_k\right)=n\pi_k
\end{aligned} \tag{A.4}$$

This gives that the variance of the Horvitz-Thompson estimator for any fixed-size sampling design is

$$\begin{aligned}
\mathbb{V}\left(\hat{\bar{y}}_{\mathsf{HT}}\right) &= \frac{1}{N^2}\sum_{k=1}^{N}\sum_{l=1}^{N}\left(\pi_{kl}-\pi_k\pi_l\right)\breve{y}_k\breve{y}_l-\frac{1}{N^2}\sum_{k=1}^{N}\breve{y}_k{}^2\left(n\pi_k-n\pi_k\right) \\
&= \frac{1}{N^2}\sum_{k=1}^{N}\sum_{l=1}^{N}\left(\pi_{kl}-\pi_k\pi_l\right)\breve{y}_k\breve{y}_l
\end{aligned} \tag{A.5}$$

□

A.3. Proof of Theorem 7.1

We follow the proof that was given by Särndal et al. [4]. The generalised regression estimator can be written as a nonlinear function of estimators

$$\begin{aligned}
\hat{\bar{y}}_{\mathsf{GREG}} &= \hat{\bar{y}}_{\mathsf{HT}}+\left(\bar{\mathbf{x}}_U-\hat{\bar{\mathbf{x}}}_{\mathsf{HT}}\right)^T\hat{\boldsymbol{\beta}} \\
&= \hat{\bar{y}}_{\mathsf{HT}}+\left(\bar{\mathbf{x}}_U-\hat{\bar{\mathbf{x}}}_{\mathsf{HT}}\right)^T\left(\hat{\mathbf{T}}_{\mathbf{xx},\mathsf{HT}}\right)^{-1}\hat{\mathbf{t}}_{\mathbf{x}y,\mathsf{HT}} \\
&= f\left(\hat{\bar{y}}_{\mathsf{HT}},\hat{\bar{\mathbf{x}}}_{\mathsf{HT}},\hat{\mathbf{T}}_{\mathbf{xx},\mathsf{HT}},\hat{\mathbf{t}}_{\mathbf{x}y,\mathsf{HT}}\right)
\end{aligned} \tag{A.6}$$

Using Taylor linearisation, this function can be approximated by a linear function [4].

$$
\begin{aligned}
f\left(\hat{\bar{y}}_{\mathsf{HT}}, \hat{\bar{\mathbf{x}}}_{\mathsf{HT}}, \hat{\mathbf{T}}_{\mathbf{xx},\mathsf{HT}}, \hat{\mathbf{t}}_{\mathbf{x}y,\mathsf{HT}}\right) \approx\ & f\left(\bar{y}_U, \bar{\mathbf{x}}_U, \mathbf{T}_{\mathbf{xx}}, \mathbf{t}_{\mathbf{x}y}\right) \\
& + \left.\frac{\partial f}{\partial \hat{\bar{y}}_{\mathsf{HT}}}\right|_{\left(\bar{y}_U, \bar{\mathbf{x}}_U, \mathbf{T}_{\mathbf{xx}}, \mathbf{t}_{\mathbf{x}y}\right)} \left(\hat{\bar{y}}_{\mathsf{HT}} - \bar{y}_U\right) \\
& + \sum_{j=1}^{p} \left.\frac{\partial f}{\partial \left(\hat{\bar{\mathbf{x}}}_{\mathsf{HT}}\right)_j}\right|_{\left(\bar{y}_U, \bar{\mathbf{x}}_U, \mathbf{T}_{\mathbf{xx}}, \mathbf{t}_{\mathbf{x}y}\right)} \left(\hat{\bar{\mathbf{x}}}_{\mathsf{HT}} - \bar{\mathbf{x}}_U\right)_j \\
& + \sum_{j=1}^{p} \sum_{i \le j} \left.\frac{\partial f}{\partial \left(\hat{\mathbf{T}}_{\mathbf{xx},\mathsf{HT}}\right)_{j,i}}\right|_{\left(\bar{y}_U, \bar{\mathbf{x}}_U, \mathbf{T}_{\mathbf{xx}}, \mathbf{t}_{\mathbf{x}y}\right)} \left(\hat{\mathbf{T}}_{\mathbf{xx},\mathsf{HT}} - \mathbf{T}_{\mathbf{xx}}\right)_{j,i} \\
& + \sum_{j=1}^{p} \left.\frac{\partial f}{\partial \left(\hat{\mathbf{t}}_{\mathbf{x}y,\mathsf{HT}}\right)_j}\right|_{\left(\bar{y}_U, \bar{\mathbf{x}}_U, \mathbf{T}_{\mathbf{xx}}, \mathbf{t}_{\mathbf{x}y}\right)} \left(\hat{\mathbf{t}}_{\mathbf{x}y,\mathsf{HT}} - \mathbf{t}_{\mathbf{x}y}\right)_j
\end{aligned}
\tag{A.7}
$$

where $\left(\hat{\bar{\mathbf{x}}}_{\mathsf{HT}}\right)_j$ denotes the j-th component of $\hat{\bar{\mathbf{x}}}_{\mathsf{HT}}$ and $\left(\hat{\mathbf{T}}_{\mathbf{xx},\mathsf{HT}}\right)_{j,i}$ denotes the value in position (j,i) in the matrix $\hat{\mathbf{T}}_{\mathbf{xx},\mathsf{HT}}$. The partial derivatives in Equation (A.7) are

$$\frac{\partial f}{\partial \hat{\bar{y}}_{\mathsf{HT}}} = 1 \tag{A.8}$$

$$\frac{\partial f}{\partial \left(\hat{\bar{\mathbf{x}}}_{\mathsf{HT}}\right)_j} = -\hat{\boldsymbol{\beta}}_j, \quad j = 1, \ldots, p \tag{A.9}$$

$$\frac{\partial f}{\partial \left(\hat{\mathbf{T}}_{\mathbf{xx},\mathsf{HT}}\right)_{j,i}} = \left(\bar{\mathbf{x}}_U - \hat{\bar{\mathbf{x}}}_{\mathsf{HT}}\right)^T \left(-\left(\hat{\mathbf{T}}_{\mathbf{xx},\mathsf{HT}}\right)^{-1} \boldsymbol{\Lambda}_{j,i} \left(\hat{\mathbf{T}}_{\mathbf{xx},\mathsf{HT}}\right)^{-1}\right) \hat{\mathbf{t}}_{\mathbf{x}y,\mathsf{HT}}, \quad i \le j = 1, \ldots, p \tag{A.10}$$

$$\frac{\partial f}{\partial \left(\hat{\mathbf{t}}_{\mathbf{x}y,\mathsf{HT}}\right)_j} = \left(\bar{\mathbf{x}}_U - \hat{\bar{\mathbf{x}}}_{\mathsf{HT}}\right)^T \left(\hat{\mathbf{T}}_{\mathbf{xx},\mathsf{HT}}\right)^{-1} \lambda_j, \quad j = 1, \ldots, \tag{A.11}$$

where $\boldsymbol{\Lambda}_{j,i}$ is a $p \times p$-matrix with value 1 in positions (j,i) and (i,j) and 0 on other positions; and λ_j a p-vector with a 1 in the j-th position and zero's elsewhere.

Evaluating these partial derivatives at the expected value point gives

$$
\begin{aligned}
& f\left(\hat{\bar{y}}_{\mathsf{HT}}, \hat{\bar{\mathbf{x}}}_{\mathsf{HT}}, \hat{\mathbf{T}}_{\mathbf{xx},\mathsf{HT}}, \hat{\mathbf{t}}_{\mathbf{x}y,\mathsf{HT}}\right) \\
& \quad \approx \bar{y}_U + 1\left(\hat{\bar{y}}_{\mathsf{HT}} - \bar{y}_U\right) \\
& \qquad + \quad \sum_{j=1}^{p} -\hat{\boldsymbol{\beta}}_j \left(\hat{\bar{\mathbf{x}}}_{\mathsf{HT}} - \bar{\mathbf{x}}_U\right)_j \\
& \qquad + \quad \sum_{j=1}^{p} \sum_{i \leq j} \left(\bar{\mathbf{x}}_U - \bar{\mathbf{x}}_U\right)^T \left(-\left(\mathbf{T}_{\mathbf{xx}}\right)^{-1} \boldsymbol{\Lambda}_{j,i} \left(\mathbf{T}_{\mathbf{xx}}\right)^{-1}\right) \mathbf{t}_{\mathbf{x}y} \left(\hat{\mathbf{T}}_{\mathbf{xx},\mathsf{HT}} - \mathbf{T}_{\mathbf{xx}}\right)_{j,i} \\
& \qquad + \quad \sum_{j=1}^{p} \left(\bar{\mathbf{x}}_U - \bar{\mathbf{x}}_U\right)^T \left(\mathbf{T}_{\mathbf{xx}}\right)^{-1} \boldsymbol{\lambda}_j \left(\hat{\mathbf{t}}_{\mathbf{x}y,\mathsf{HT}} - \mathbf{t}_{\mathbf{x}y}\right)_j \\
& \quad = \bar{y}_U + \hat{\bar{y}}_{\mathsf{HT}} - \bar{y}_U + \sum_{j=1}^{p} -\boldsymbol{\beta}_j \left(\hat{\bar{\mathbf{x}}}_{\mathsf{HT}} - \bar{\mathbf{x}}_U\right)_j \\
& \quad = \hat{\bar{y}}_{\mathsf{HT}} + \left(\bar{\mathbf{x}}_U - \hat{\bar{\mathbf{x}}}_{\mathsf{HT}}\right)^T \boldsymbol{\beta}
\end{aligned}
\tag{A.12}
$$

We can rewrite this approximation as

$$
\begin{aligned}
f\left(\hat{\bar{y}}_{\mathsf{HT}}, \hat{\bar{\mathbf{x}}}_{\mathsf{HT}}, \hat{\mathbf{T}}_{\mathbf{xx},\mathsf{HT}}, \hat{\mathbf{t}}_{\mathbf{x}y,\mathsf{HT}}\right) &\approx \hat{\bar{y}}_{\mathsf{HT}} + \left(\bar{\mathbf{x}}_U - \hat{\bar{\mathbf{x}}}_{\mathsf{HT}}\right)^T \boldsymbol{\beta} \\
&= \frac{1}{N} \sum_{k=1}^{N} \mathbf{x}_k^T \boldsymbol{\beta} + \frac{1}{N} \sum_{k \in s} \left(\frac{1}{\pi_k} y_k - \frac{1}{\pi_k} \mathbf{x}_k^T \boldsymbol{\beta}\right) \\
&= \frac{1}{N} \sum_{k=1}^{N} \mathbf{x}_k^T \boldsymbol{\beta} + \frac{1}{N} \sum_{k \in s} \breve{E}_k
\end{aligned}
\tag{A.13}
$$

where $\breve{E}_k = \frac{1}{\pi_k} E_k$ and $E_k = y_k - \mathbf{x}_k^T \boldsymbol{\beta}_U$. The approximated variance of the generalised regression estimator equals the variance of the Taylor approximation of $\hat{\bar{y}}_{\mathsf{GREG}}$, so

$$
\begin{aligned}
\mathbb{AV}\left(\hat{\bar{y}}_{\mathsf{GREG}}\right) &= \mathbb{V}\left(\frac{1}{N} \sum_{k=1}^{N} \mathbf{x}_k^T \boldsymbol{\beta} + \frac{1}{N} \sum_{k \in s} \breve{E}_k\right) \\
&= \mathbb{V}\left(\frac{1}{N} \sum_{k=1}^{N} I_k \breve{E}_k\right) \\
&= \frac{1}{N^2} \sum_{k=1}^{N} \sum_{l=1}^{N} \breve{E}_k \breve{E}_l \, \mathsf{Cov}\left(I_k, I_l\right) \\
&= \frac{1}{N^2} \sum_{k=1}^{N} \sum_{l=1}^{N} \breve{E}_k \breve{E}_l \left(\pi_{kl} - \pi_k \pi_l\right)
\end{aligned}
\tag{A.14}
$$

□

A.4. Proof of Theorem 7.2

This proof is similar to the proof of Theorem 2.3, see Appendix A.2. Developing the square of Equation (7.20) gives

$$\begin{aligned}
\mathbb{AV}\left(\hat{\bar{y}}_{\text{GREG}}\right) &= -\frac{1}{2N^2}\sum_{k=1}^{N}\sum_{l=1}^{N}\left(\pi_{kl}-\pi_k\pi_l\right)\left(\breve{E}_k-\breve{E}_l\right)^2 \\
&= -\frac{1}{2N^2}\sum_{k=1}^{N}\sum_{l=1}^{N}\left(\pi_{kl}-\pi_k\pi_l\right)\left(\breve{E}_k^2-2\breve{E}_k\breve{E}_l+\breve{E}_l^2\right) \\
&= \frac{1}{N^2}\sum_{k=1}^{N}\sum_{l=1}^{N}\left(\pi_{kl}-\pi_k\pi_l\right)\breve{E}_k\breve{E}_l-\frac{1}{N^2}\sum_{k=1}^{N}\sum_{l=1}^{N}\left(\pi_{kl}-\pi_k\pi_l\right)\breve{E}_k^2 \\
&= \frac{1}{N^2}\sum_{k=1}^{N}\sum_{l=1}^{N}\left(\pi_{kl}-\pi_k\pi_l\right)\breve{E}_k\breve{E}_l-\frac{1}{N^2}\sum_{k=1}^{N}\left(\breve{E}_k^2\sum_{l=1}^{N}\left(\pi_{kl}-\pi_k\pi_l\right)\right) \\
&= \frac{1}{N^2}\sum_{k=1}^{N}\sum_{l=1}^{N}\left(\pi_{kl}-\pi_k\pi_l\right)\breve{E}_k\breve{E}_l-\frac{1}{N^2}\sum_{k=1}^{N}\breve{E}_k^2\left(n\pi_k-n\pi_k\right) \\
&= \frac{1}{N^2}\sum_{k=1}^{N}\sum_{l=1}^{N}\left(\pi_{kl}-\pi_k\pi_l\right)\breve{E}_k\breve{E}_l
\end{aligned} \tag{A.15}$$

□

A.5. Useful Lemma's for the proof of Theorem 7.3

The next Lemma's from Gut [27] are used for the proof of Theorem 7.3.

Lemma A.1 (Markov's inequality)**.** Suppose $\mathbb{E}\left(|X|^r\right)<\infty$ for some $r>0$ and let $a>0$. Then

$$\mathbb{P}\left(|X|>a\right)\leq\frac{\mathbb{E}\left(|X|^r\right)}{a} \tag{A.16}$$

Lemma A.2 (The Lyapounov inequality)**.** For $0<p\leq r$,

$$\left(\mathbb{E}\left(|X|^p\right)\right)^{\frac{1}{p}}\leq\left(\mathbb{E}\left(|X|^r\right)\right)^{\frac{1}{r}} \tag{A.17}$$

Lemma A.3 (The Cauchy-Schwarz inequality)**.** Suppose that X and Y have finite variances. Then

$$\left|\mathbb{E}\left(XY\right)\right|\leq\mathbb{E}\left(|XY|\right)\leq||X||_2\,||Y||_2=\left(\mathbb{E}\left(X^2\right)\mathbb{E}\left(Y^2\right)\right)^{\frac{1}{2}} \tag{A.18}$$

For a summation, this means

$$\left(\mathbb{E}\left(\sum_{k=1}^{N}x_ky_k\right)\right)^2\leq\left(\sum_{k=1}^{N}\mathbb{E}\left(x_k^2\right)\right)\left(\sum_{k=1}^{N}\mathbb{E}\left(y_k^2\right)\right) \tag{A.19}$$

A.6. Proof of Remark 7.2

Recall from Definition 7.2 that the estimator $\frac{1}{N_\tau}\hat{\mathbf{T}}_{\mathbf{xx},\text{HT}}$ is consistent for $\frac{1}{N_\tau}\mathbf{T}_{\mathbf{xx}}$ if $\forall\varepsilon>0$

$$\lim_{\tau\to\infty}\mathbb{P}_{p_\tau}\left(\left|\frac{1}{N_\tau}\mathbf{T}_{\mathbf{xx}}-\frac{1}{N_\tau}\hat{\mathbf{T}}_{\mathbf{xx},\text{HT}}\right|>\varepsilon\right)=0\quad\xi\text{-almost surely} \tag{A.20}$$

Using the Markov inequality (Lemma A.1) it follows that it is sufficient to prove that

$$\lim_{\tau\to\infty} \mathbb{E}_{p_\tau}\left(\left|\frac{1}{N_\tau}\mathbf{T}_{\mathbf{xx}} - \frac{1}{N_\tau}\hat{\mathbf{T}}_{\mathbf{xx},\mathsf{HT}}\right|\right) = 0 \tag{A.21}$$

The expressions for $\mathbf{T}_{\mathbf{xx}}$ and $\hat{\mathbf{T}}_{\mathbf{xx},\mathsf{HT}}$ from Equations (7.7) and (7.8) give that

$$\begin{aligned}\left|\frac{1}{N_\tau}\mathbf{T}_{\mathbf{xx}} - \frac{1}{N_\tau}\hat{\mathbf{T}}_{\mathbf{xx},\mathsf{HT}}\right| &= \left|\frac{1}{N_\tau}\sum_{k=1}^{N_\tau}\mathbf{x}_k\mathbf{x}_k^T\frac{1}{v_k} - \frac{1}{N_\tau}\sum_{k=1}^{N_\tau}\frac{I_{k\tau}}{\pi_{k\tau}}\mathbf{x}_k\mathbf{x}_k^T\frac{1}{v_k}\right| \\ &= \left|\frac{1}{N_\tau}\sum_{k=1}^{N_\tau}\left(1-\frac{I_{k\tau}}{\pi_{k\tau}}\right)\mathbf{x}_k\mathbf{x}_k^T\frac{1}{v_k}\right|\end{aligned} \tag{A.22}$$

The element (i,j) , $i,j=1,\ldots,p$, of the matrix is then denoted by

$$\left|\frac{1}{N_\tau}\mathbf{T}_{\mathbf{xx}} - \frac{1}{N_\tau}\hat{\mathbf{T}}_{\mathbf{xx},\mathsf{HT}}\right|_{(i,j)} = \left|\frac{1}{N_\tau}\sum_{k=1}^{N_\tau}\left(1-\frac{I_{k\tau}}{\pi_{k\tau}}\right)x_{ki}x_{kj}\frac{1}{v_k}\right| \tag{A.23}$$

We show that expectation of the (i,j)-th element of this matrix is zero as $\tau\to\infty$. By the Lyapuanov inequality (Lemma A.2) we have

$$\lim_{\tau\to\infty}\mathbb{E}_{p_\tau}\left(\left|\frac{1}{N_\tau}\sum_{k=1}^{N_\tau}\left(1-\frac{I_{k\tau}}{\pi_{k\tau}}\right)x_{ki}x_{kj}\frac{1}{v_k}\right|\right) \le \lim_{\tau\to\infty}\left(\mathbb{E}_{p_\tau}\left(\left|\frac{1}{N_\tau}\sum_{k=1}^{N_\tau}\left(1-\frac{I_{k\tau}}{\pi_{k\tau}}\right)x_{ki}x_{kj}\frac{1}{v_k}\right|^2\right)\right)^{\frac{1}{2}}$$

Similar steps as were taken in the proof of Theorem 7.3 give that

$$\begin{aligned}&\mathbb{E}_{p_\tau}\left(\left|\frac{1}{N_\tau}\sum_{k=1}^{N_\tau}\left(1-\frac{I_{k\tau}}{\pi_{k\tau}}\right)x_{ki}x_{kj}\frac{1}{v_k}\right|^2\right) \\ &\quad= \frac{1}{N_\tau^2}\sum_{k=1}^{N_\tau}\left(\frac{1}{\pi_{k\tau}}-1\right)x_{ki}^2x_{kj}^2\frac{1}{v_k^2} + \frac{1}{N_\tau^2}\sum_{k=1}^{N_\tau}\sum_{k\neq l=1}^{N_\tau}\left(\frac{\pi_{kl\tau}}{\pi_{k\tau}\pi_{l\tau}}-1\right)x_{ki}x_{kj}x_{li}x_{lj}\frac{1}{v_k}\frac{1}{v_l}\end{aligned} \tag{A.24}$$

The first part of Equation (A.24) is bounded by

$$\begin{aligned}\lim_{\tau\to\infty}\frac{1}{N_\tau^2}\sum_{k=1}^{N_\tau}\left(\frac{1}{\pi_{k\tau}}-1\right)x_{ki}^2x_{kj}^2\frac{1}{v_k^2} &\le \lim_{\tau\to\infty}\frac{1}{N_\tau^2}\sum_{k=1}^{N_\tau}\frac{1}{\pi_{k\tau}}x_{ki}^2x_{kj}^2\frac{1}{v_k^2} \\ &\le \lim_{\tau\to\infty}\left(N_\tau\min_{1\le k\le N_\tau}\pi_{kt}\right)^{-1}\frac{1}{N_\tau}\sum_{k=1}^{N_\tau}x_{ki}^2x_{kj}^2\frac{1}{v_k^2}\end{aligned} \tag{A.25}$$

We impose a new Condition (C1*) that is similar to Condition (C1)

(C1*) $\frac{1}{N_\tau}\sum_{k=1}^{N_\tau}x_{ki}^2x_{kj}^2\frac{1}{v_k^2} < \infty$ for $i,j=1,\ldots,p$

Then under Conditions (C1*) and (C4) we obtain the desired result for the first part of Equation (A.24).

The second part of Equation (A.24) is bounded by

$$\begin{aligned}
&\lim_{\tau\to\infty} \frac{1}{N_\tau^2} \sum_{k=1}^{N_\tau} \sum_{k\neq l=1}^{N_\tau} \left(\frac{\pi_{kl\tau}}{\pi_{k\tau}\pi_{l\tau}} - 1 \right) x_{ki}x_{kj}x_{li}x_{lj} \frac{1}{v_k} \frac{1}{v_l} \\
&\quad \leq \lim_{\tau\to\infty} \max_{1\leq k\neq l\leq N_\tau} \left| \frac{\pi_{kl\tau}}{\pi_{k\tau}\pi_{l\tau}} - 1 \right| \frac{1}{N_\tau^2} \sum_{k=1}^{N_\tau} \sum_{k\neq l=1}^{N_\tau} x_{ki}x_{kj}x_{li}x_{lj} \frac{1}{v_k} \frac{1}{v_l} \\
&\quad \leq \lim_{\tau\to\infty} \max_{1\leq k\neq l\leq N_\tau} \left| \frac{\pi_{kl\tau}}{\pi_{k\tau}\pi_{l\tau}} - 1 \right| \left(\frac{1}{N_\tau} \sum_{k=1}^{N_\tau} x_{ki}x_{kj} \frac{1}{v_k} \right)^2 \\
&\quad \leq \lim_{\tau\to\infty} \max_{1\leq k\neq l\leq N_\tau} \left| \frac{\pi_{kl\tau}}{\pi_{k\tau}\pi_{l\tau}} - 1 \right| \frac{1}{N_\tau} \sum_{k=1}^{N_\tau} x_{ki}^2 x_{kj}^2 \frac{1}{v_k^2}
\end{aligned} \tag{A.26}$$

Under Conditions (C1*) and (C5) the desired result is obtained, which concludes the proof. □

B

Definitions auxiliary variables

	Gender
M	Male
V	Female

	Marital status
1	Unmarried
2	Married or partnership
3	Widowed after marriage or partnership
4	Divorced after marriage or partnership

	Age
1	0 - 5 years
2	6 - 11 years
3	12 - 17 years
4	18 - 24 years
5	25 - 29 years
6	30 - 39 years
7	40 - 49 years
8	50 - 64 years
9	65 - 74 years
10	75 - 125 years

	Ethnicity
0	Native Dutch
1	From non-Western countries
2	From other Western countries

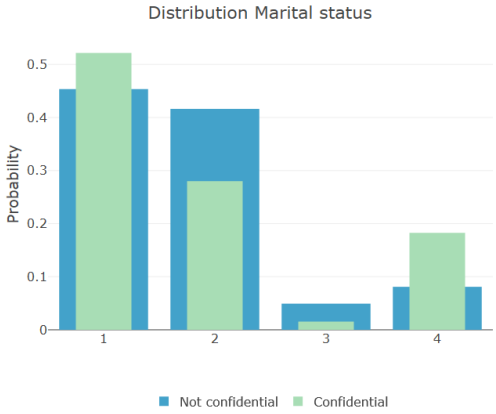
	Place in household
1	Child
2	Single person
3	Partner in non-married couple without children
4	Partner in married couple without children
5	Partner in non-married couple with children
6	Partner in married couple with children
7	Parent in single-parent household
8	Reference person in other household
9	Other member of a household
10	Member of an institutional household

	Type of household
1	Single person household
2	Non-married couple without children
3	Married couple without children
4	Non-married couple with children
5	Married couple with children
6	Singe-parent household
7	Other type of household
8	Institutional household

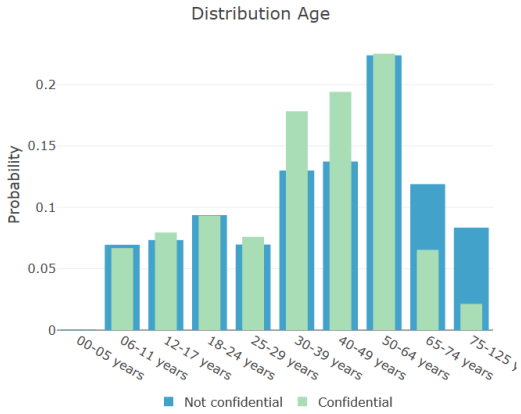
C

Additional figures to Section 4

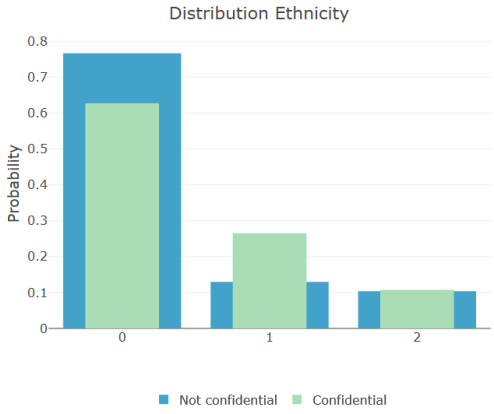
(a) Gender



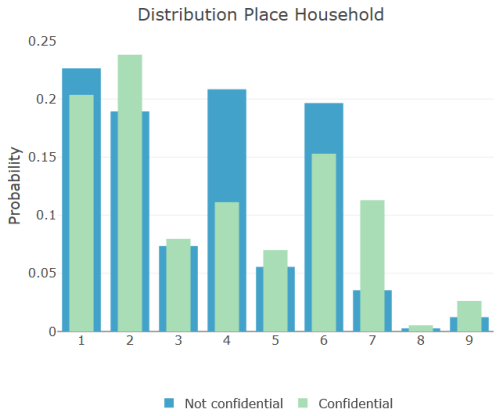
(b) Marital status



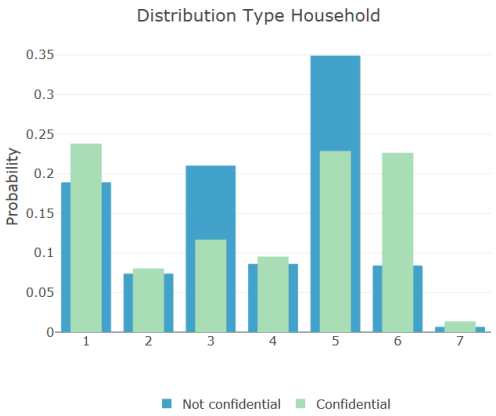
(c) Age



(d) Ethnicity

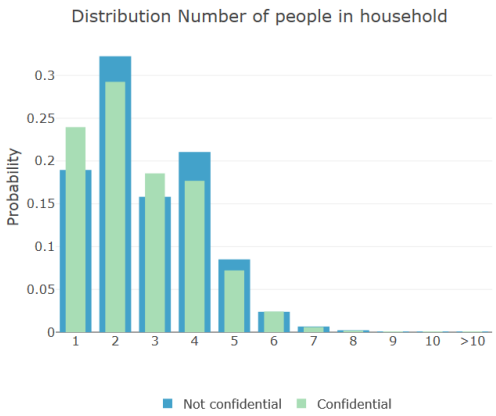


(e) Place of household



(f) Type of household

(g) Number of people on address

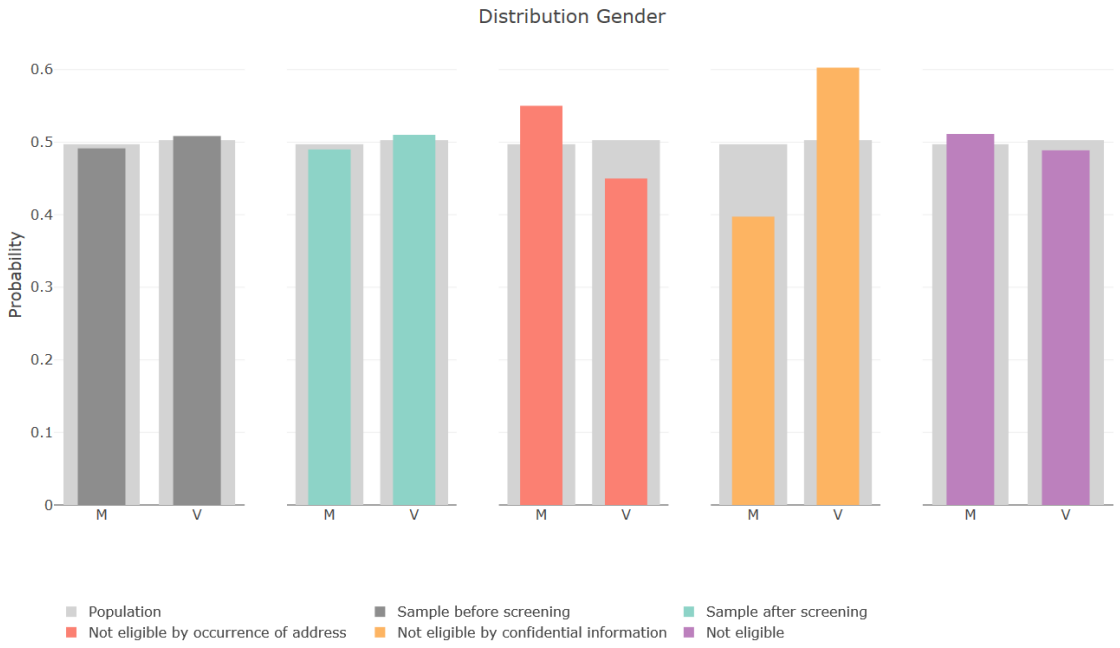


(h) Number of people in household

D

Additional figures to Section 5

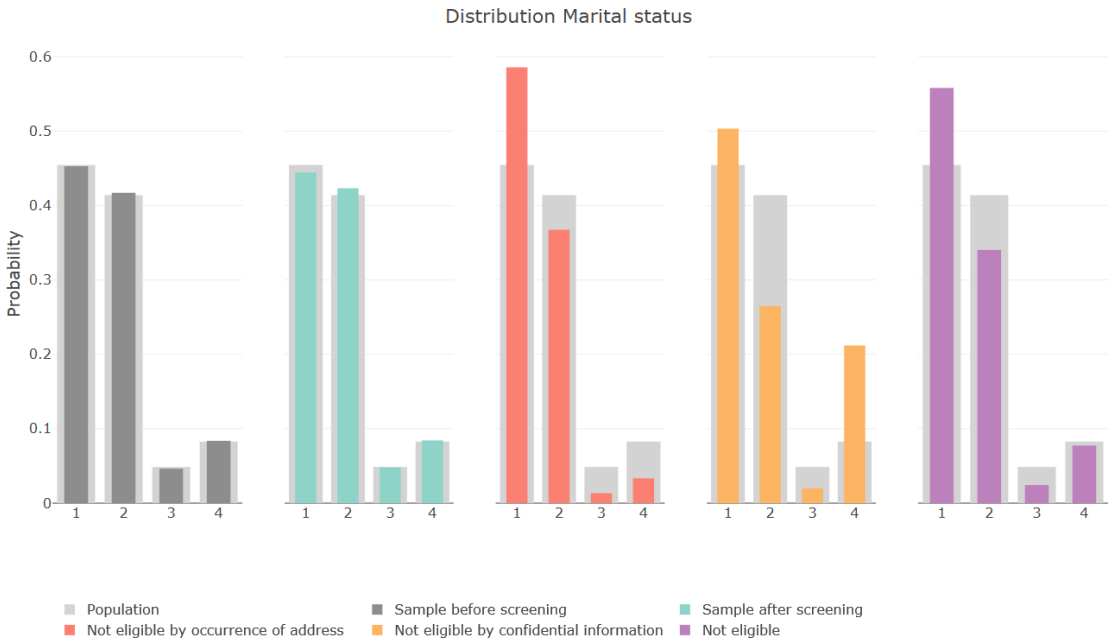
D.1. Gender



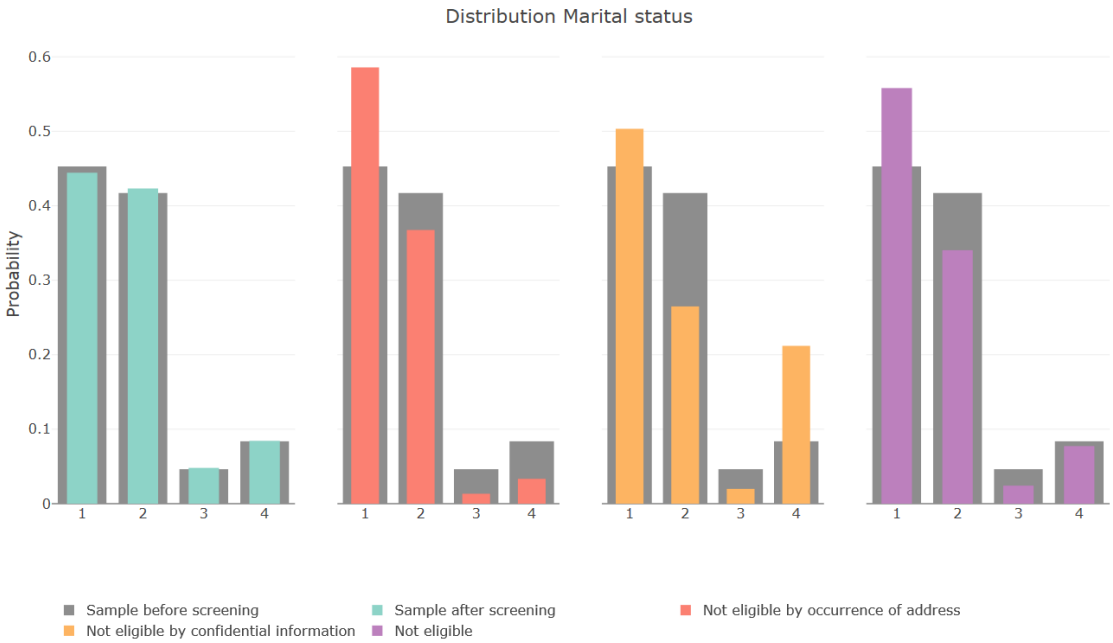
(a) Hypothesis 1

(b) Hypothesis 2

D.2. Marital status

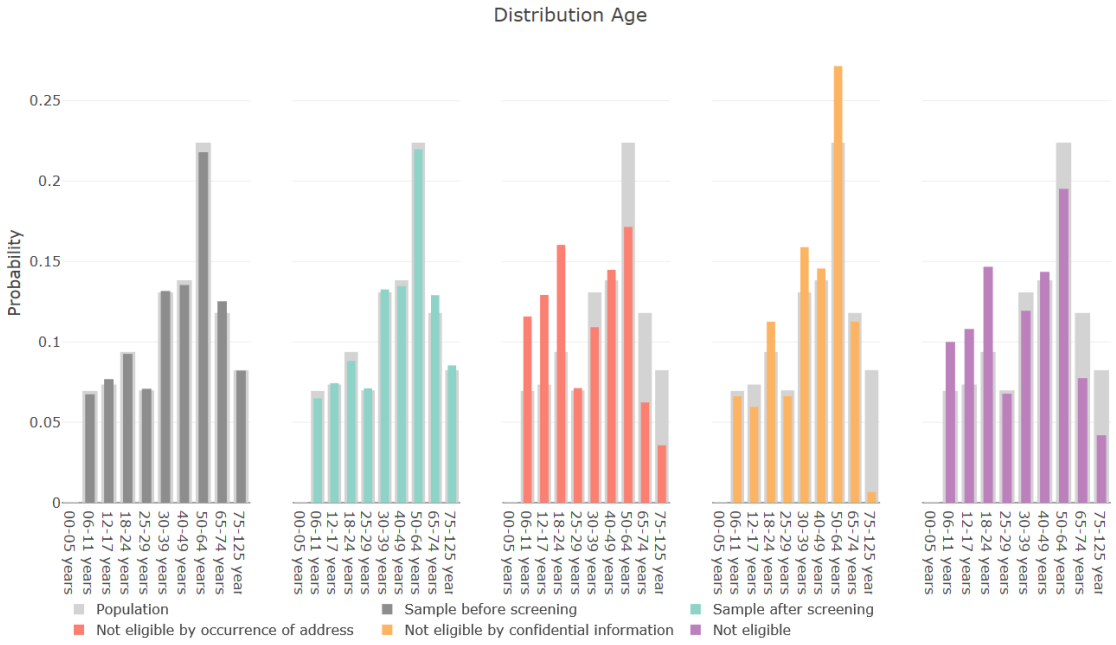


(a) Hypothesis 1

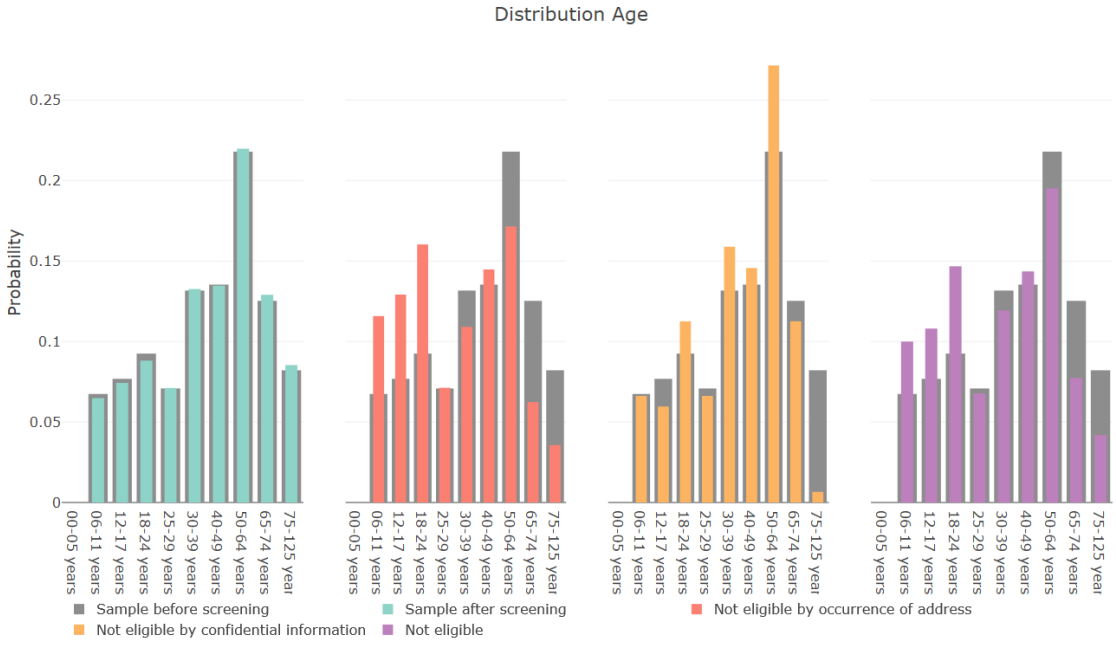


(b) Hypothesis 2

D.3. Age

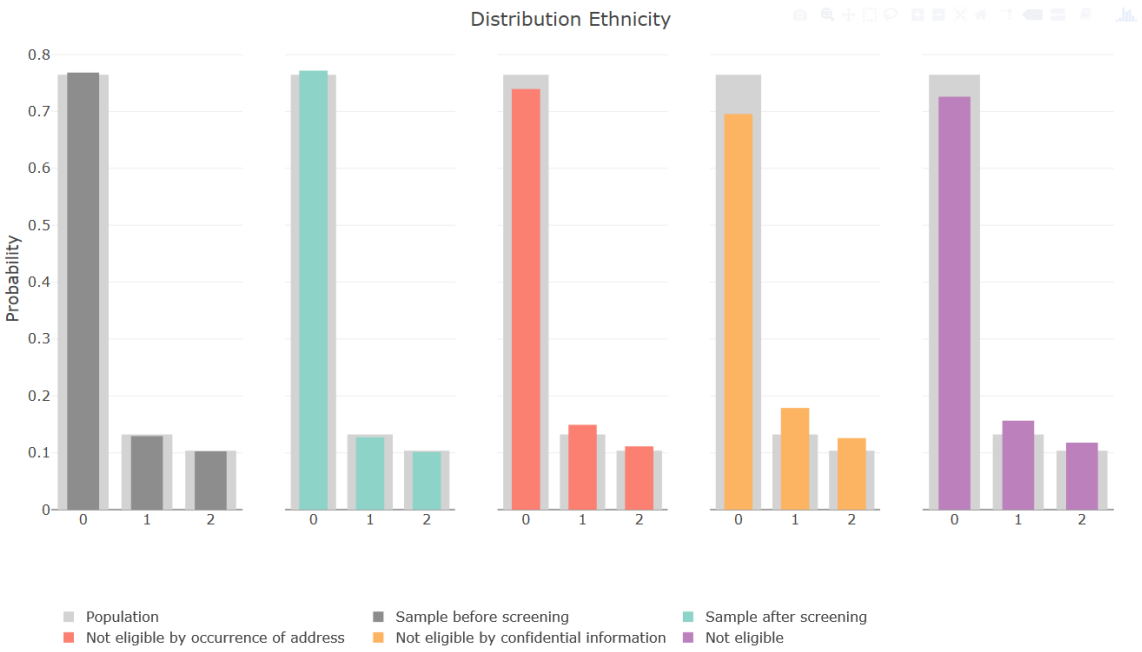


(a) Hypothesis 1

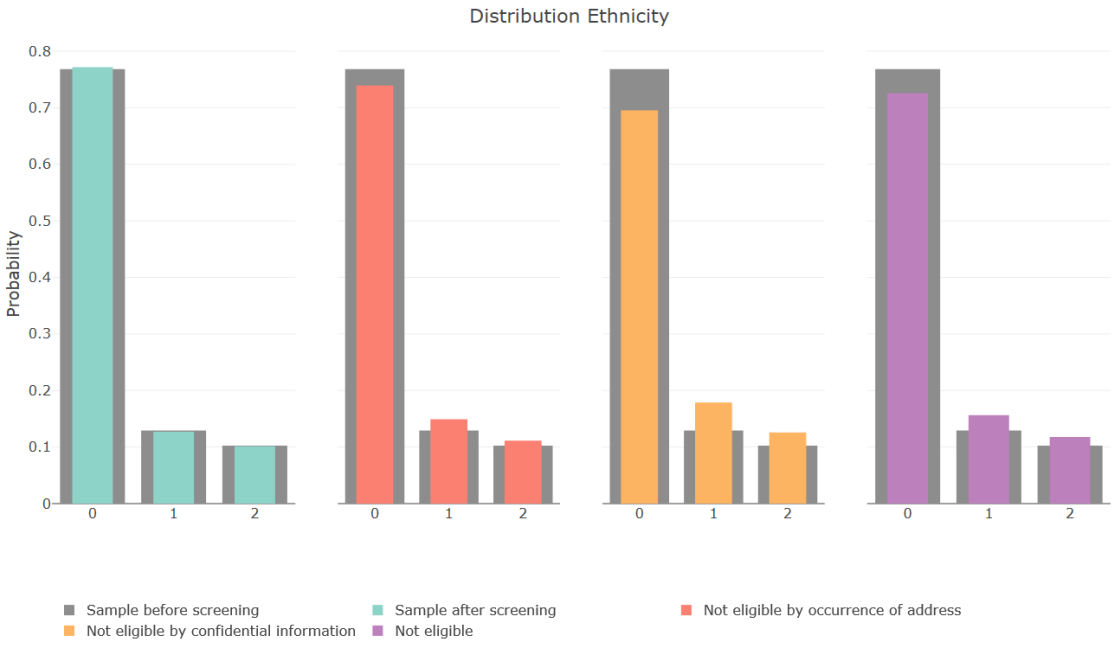


(b) Hypothesis 2

D.4. Ethnicity



(a) Hypothesis 1



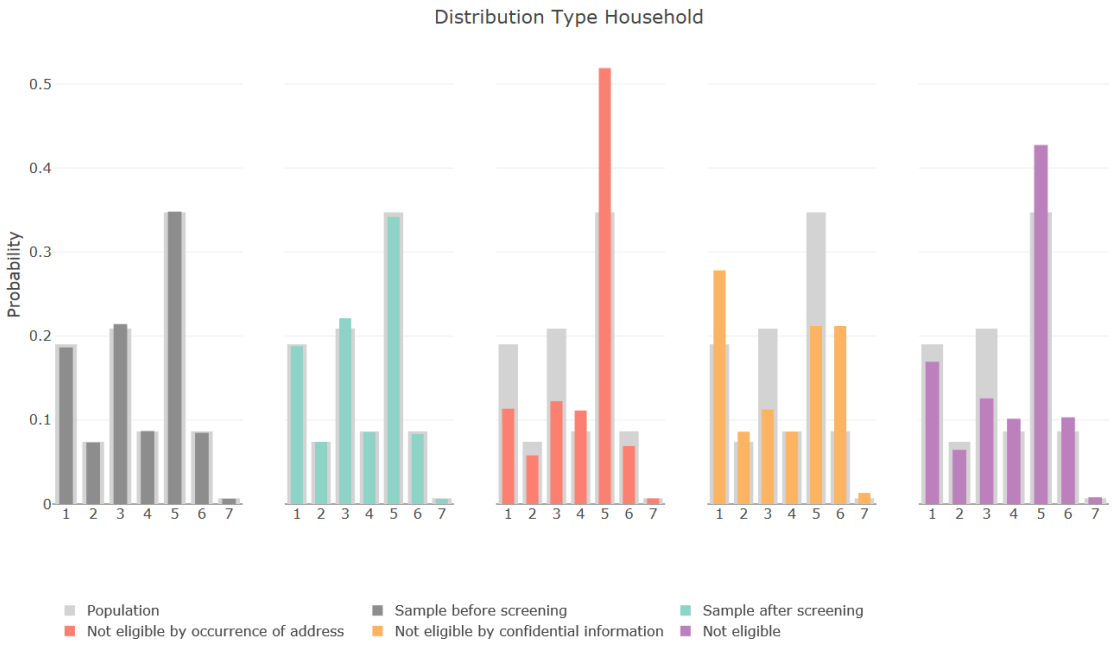
(b) Hypothesis 2

D.5. Place in Household

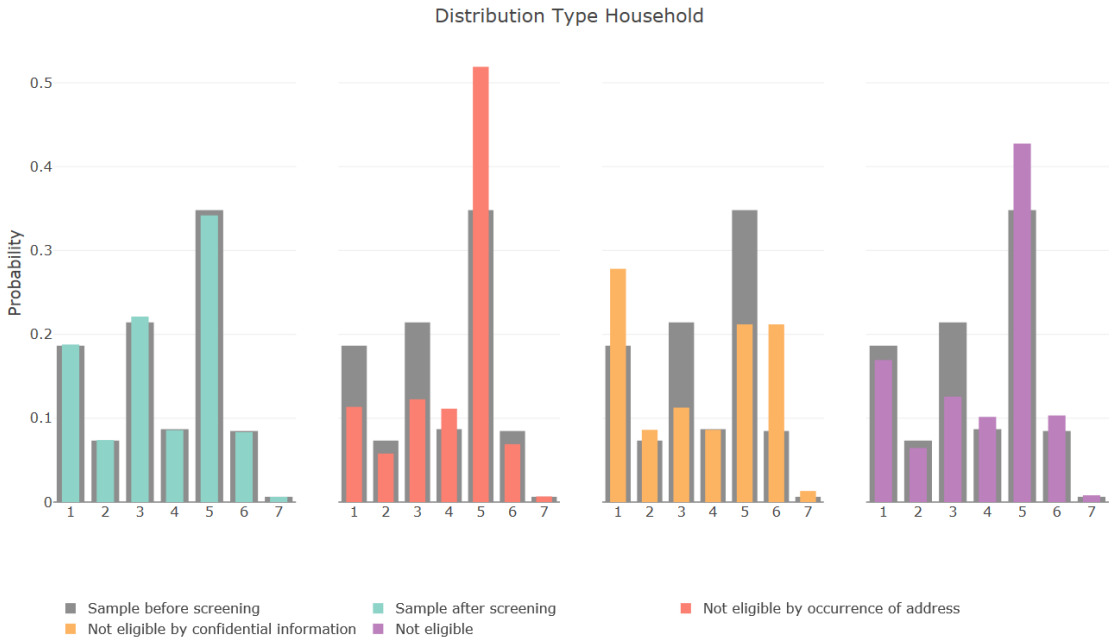
(a) Hypothesis 1

(b) Hypothesis 2

D.6. Type of household

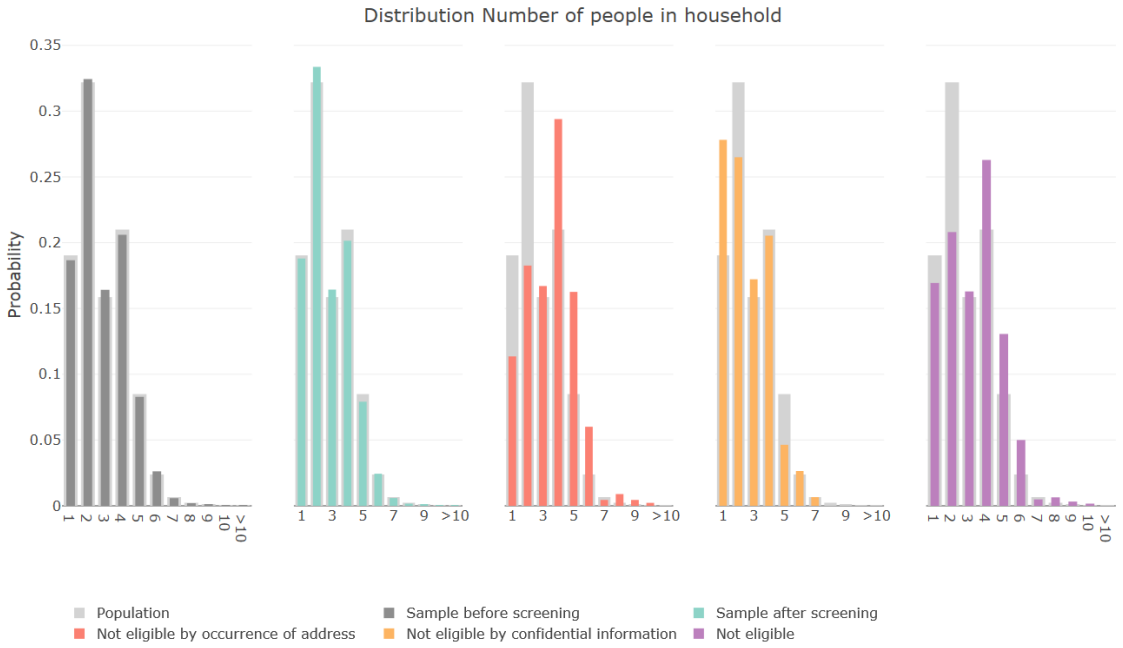


(a) Hypothesis 1

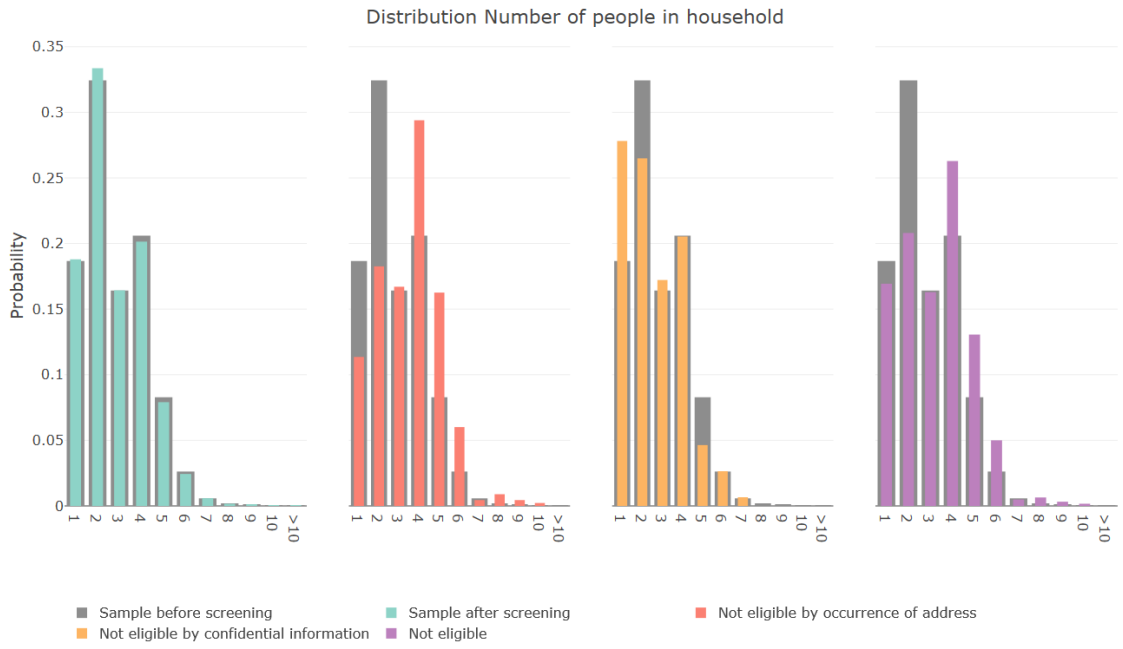


(b) Hypothesis 2

D.7. Number of people in household

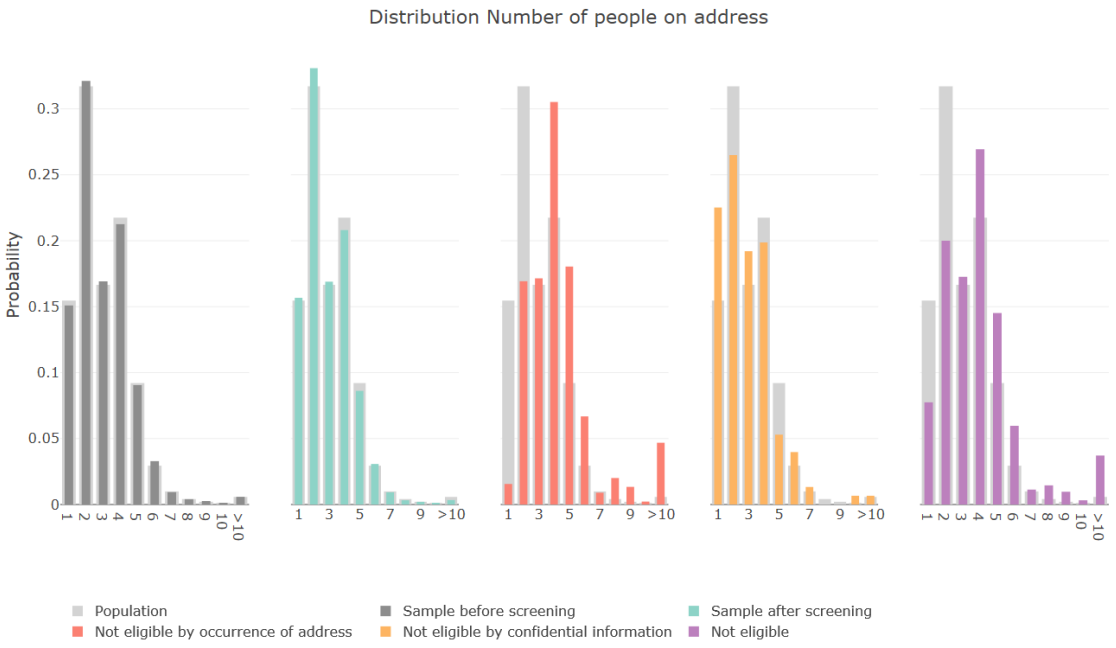


(a) Hypothesis 1

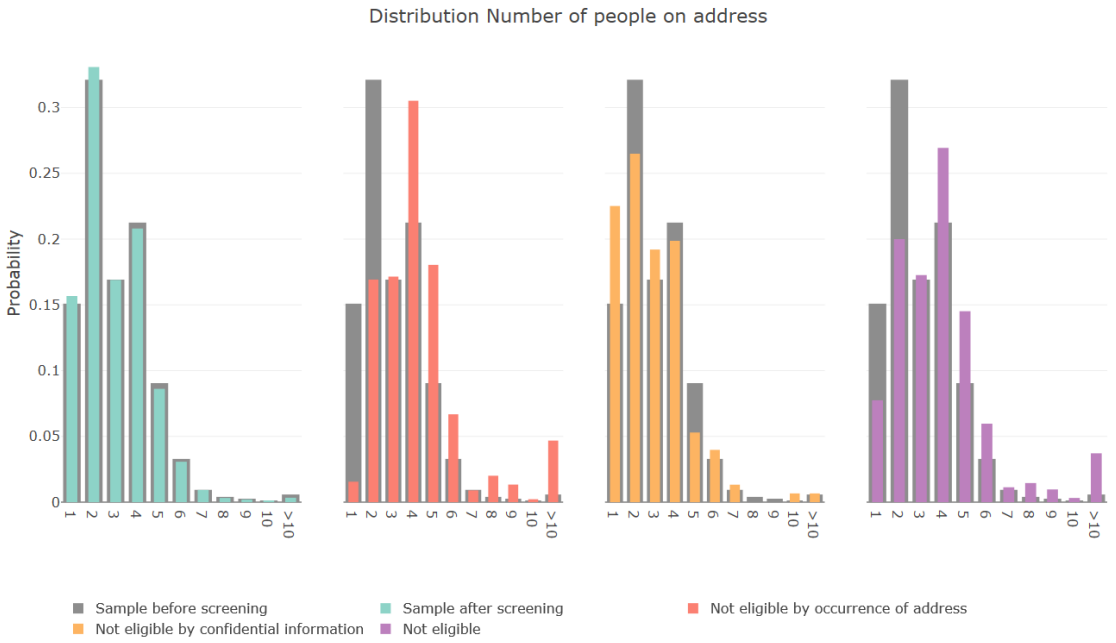


(b) Hypothesis 2

D.8. Number of people on address



(a) Hypothesis 1



(b) Hypothesis 2

E

Additional results to Section 5

	2018-04	2018-05	2018-06	2018-07	2018-08	2018-09	2018-10	2018-11	2018-12
Gender									
Sample before screening	0.08643	0.95503	0.25669	0.24097	0.62341	0.00276	0.81754	0.21282	0.01396
Sample after screening	0.03101	0.67049	0.27265	0.22008	0.81855	0.00423	0.75792	0.18710	0.03333
Not eligible by occurrence of address	0.51210	0.46838	0.44089	0.34752	0.78918	0.10219	0.11932	0.70924	0.14338
Not eligible by confidential information	0.00011	0.41325	0.10524	0.04137	0.15326	0.76934	0.22151	0.41013	0.87598
Not eligible	0.16012	0.22464	0.75381	0.97435	0.33518	0.34793	0.06024	0.91191	0.12446
Marital status									
Sample before screening	0.36901	0.67819	0.82990	0.79266	0.64694	0.37776	0.70423	0.53991	0.86261
Sample after screening	0.07127	0.16311	0.55197	0.12639	0.25523	0.05114	0.15644	0.27762	0.26956
Not eligible by occurrence of address	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	0.00041	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	0.00356	0.00003	0.00094
Not eligible by confidential information	0.01161	$< 1 \cdot 10^{-5}$	0.00042	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	0.00401	$< 1 \cdot 10^{-5}$	0.00286
Not eligible	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	0.00007	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	0.00190	$< 1 \cdot 10^{-5}$	0.00072
Age									
Sample before screening	0.91026	0.29789	0.81170	0.49706	0.72451	0.10812	0.79878	0.15878	0.89129
Sample after screening	0.31070	0.03563	0.39177	0.10581	0.92039	0.04056	0.72007	0.03700	0.27698
Not eligible by occurrence of address	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$
Not eligible by confidential information	$< 1 \cdot 10^{-5}$	0.00337	0.00167	0.00001	0.00005	0.00652	0.01581	0.06081	0.00153
Not eligible	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$
Ethnicity									
Sample before screening	0.45088	0.93256	0.46286	0.61478	0.57071	0.89992	0.35444	0.92685	0.43296
Sample after screening	0.94472	0.04706	0.10344	0.08544	0.17883	0.29975	0.99762	0.10260	0.02350
Not eligible by occurrence of address	0.12090	0.00037	0.69875	0.34114	0.00018	0.00055	0.12329	0.00342	0.00100
Not eligible by confidential information	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	0.00002	0.00008	0.23534	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	0.00004
Not eligible	0.00002	$< 1 \cdot 10^{-5}$	0.01173	0.00035	0.00002	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$

Table E.1: The $\mathbf{p}$-values of the statistical test for hypothesis 1 ($R = 100.000$) for the auxiliary variables: gender, marital status, age and ethnicity for the samples of the mobility survey of April 2018 until December 2018.

	2018-04	2018-05	2018-06	2018-07	2018-08	2018-09	2018-10	2018-11	2018-12
Place in household									
Sample before screening	0.24346	0.86518	0.89310	0.84237	0.44939	0.11531	0.68152	0.76531	0.80097
Sample after screening	0.00290	0.03375	0.56344	0.03043	0.68326	0.04303	0.11781	0.07882	0.16784
Not eligible by occurrence of address	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$
Not eligible by confidential information	0.00103	0.00114	0.00682	$< 1 \cdot 10^{-5}$	0.00001	$< 1 \cdot 10^{-5}$	0.00592	$< 1 \cdot 10^{-5}$	0.01783
Not eligible	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$
Type of household									
Sample before screening	0.05574	0.50695	0.99769	0.84035	0.46651	0.11469	0.71535	0.70787	0.74028
Sample after screening	0.00030	0.08325	0.29029	0.04612	0.82522	0.15930	0.07341	0.12310	0.08061
Not eligible by occurrence of address	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$
Not eligible by confidential information	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	0.00003	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	0.00045	$< 1 \cdot 10^{-5}$	0.00001
Not eligible	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$
Number of people in household									
Sample before screening	0.58627	0.94890	0.96198	0.45114	0.18693	0.00847	0.07134	0.73162	0.28096
Sample after screening	0.03034	0.18431	0.60662	0.04563	0.23179	0.01052	0.00557	0.09047	0.22208
Not eligible by occurrence of address	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$
Not eligible by confidential information	0.09080	0.56586	0.95897	0.17157	0.04216	0.05572	0.03325	0.01663	0.11089
Not eligible	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$
Number of people on address									
Sample before screening	0.64393	0.95516	0.92686	0.47401	0.42242	0.02795	0.15823	0.81175	0.41264
Sample after screening	0.02576	0.04687	0.22598	0.00490	0.14088	0.00141	0.00229	0.02561	0.01733
Not eligible by occurrence of address	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$
Not eligible by confidential information	0.06763	0.31117	0.86602	0.14890	0.01147	0.19944	0.02274	0.07133	0.03656
Not eligible	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$	$< 1 \cdot 10^{-5}$

Table E.2: The $\mathbf{p}$-values of the statistical test for hypothesis 1 ($R = 100.000$) for the auxiliary variables: place in household, type of household, number of people in household and number of people on an address for the samples of the mobility survey of April 2018 until December 2018.

F

Additional results to Section 8

	$\mathbf{V}_0$	$\mathbf{V}_{0.2}$	$\mathbf{V}_{0.4}$	$\mathbf{V}_{0.6}$	$\mathbf{V}_{0.8}$	$\mathbf{V}_1$
2018-04						
Mean population	9.9976	9.9992	9.9952	10.0002	10.0018	10.0000
HT-estimator equal prob.	10.0484	9.9831	9.9057	9.9574	9.9353	9.8967
HT-estimator adjusted prob.	10.0537	10.0069	9.9564	10.0339	10.0374	10.0310
2018-05						
Mean population	9.9978	10.0036	10.00021	10.00031	10.0036	10.0000
HT-estimator equal prob.	9.9927	9.9595	9.9314	9.9678	9.9084	9.8959
HT-estimator adjusted prob.	9.9967	9.9759	9.9672	10.0202	9.9813	9.9880
2018-06						
Mean population	9.9970	10.0032	10.0000	10.0065	9.9978	10.0000
HT-estimator equal prob.	9.9883	10.0272	9.9707	9.9486	9.9040	9.8907
HT-estimator adjusted prob.	9.9889	10.0459	9.9989	9.9862	9.9544	9.9557
2018-07						
Mean population	10.0023	10.0011	9.9983	10.0000	10.0016	10.0000
HT-estimator equal prob.	9.9816	9.9987	9.9134	9.9364	9.9052	9.8923
HT-estimator adjusted prob.	9.9785	10.0275	9.9647	10.0212	10.0247	10.0375
2018-08						
Mean population	10.0062	9.9998	9.9970	9.9993	10.0019	10.0000
HT-estimator equal prob.	9.9667	9.9534	9.9931	9.9254	9.8944	9.8860
HT-estimator adjusted prob.	9.9640	9.9650	10.0205	9.9655	9.9478	9.9530
2018-09						
Mean population	9.9960	9.9971	9.9992	10.0011	10.0005	10.0000
HT-estimator equal prob.	10.0183	9.9635	9.9996	9.9179	9.9148	9.8867
HT-estimator adjusted prob.	10.0247	9.9841	10.0360	9.9714	9.9903	9.9805
2018-10						
Mean population	10.0016	10.0022	10.0021	9.9996	10.0024	10.0000
HT-estimator equal prob.	10.0259	10.0087	9.9360	9.8764	9.8950	9.8873
HT-estimator adjusted prob.	10.0279	10.0404	9.9936	9.9626	10.0025	10.0242
2018-11						
Mean population	9.9978	9.9952	9.9981	9.9979	9.9994	10.0000
HT-estimator equal prob.	9.9526	9.9713	9.9553	9.9552	9.9247	9.8891
HT-estimator adjusted prob.	9.9449	9.9982	10.0025	10.0243	10.0128	9.9984
2018-12						
Mean population	10.0007	10.0061	10.0002	10.0029	10.0025	10.0000
HT-estimator equal prob.	9.9433	9.9903	9.9524	9.9013	9.9098	9.8873
HT-estimator adjusted prob.	9.9458	9.9995	9.9736	9.9386	9.9621	9.9520

Table F.1: Results of estimating the population mean by using the Horvitz-Thompson estimator with equal probabilities and the adjusted probabilities.

	$\mathbf{V}_0$	$\mathbf{V}_{0.2}$	$\mathbf{V}_{0.4}$	$\mathbf{V}_{0.6}$	$\mathbf{V}_{0.8}$	$\mathbf{V}_1$
			2018-04			
Mean population	9.9976	9.9992	9.9952	10.0002	10.0018	10.0000
Mean sample after screening	10.0484	9.9831	9.9057	9.9574	9.9353	9.8967
Gender	10.0491	9.9831	9.9059	9.9575	9.9344	9.8962
Marital status	10.0474	9.9838	9.9055	9.9572	9.9350	9.8967
Age	10.0477	9.9854	9.9080	9.9623	9.9408	9.9045
Ethnicity	10.0484	9.9831	9.9060	9.9575	9.9354	9.8968
Place Household	10.0475	9.9859	9.9110	9.9644	9.9428	9.9068
Type Household	10.0464	9.9862	9.9107	9.9652	9.9433	9.9077
Number of people in household	10.0478	9.9849	9.9102	9.9643	9.9439	9.9084
Number of people in household 11[a]	10.0472	9.9857	9.9113	9.9650	9.9450	9.9099
Number of people on address	10.0517	10.0011	9.9458	10.0164	10.0136	10.0000
Number of people on address 11[b]	10.0483	9.9877	9.9230	9.9815	9.9673	9.9383
			2018-05			
Mean population	9.9978	10.0036	10.00021	10.00031	10.0036	10.0000
Mean sample after screening	9.9927	9.9595	9.9314	9.9678	9.9084	9.8959
Gender	9.9926	9.9592	9.9314	9.9678	9.9084	9.8959
Marital status	9.9933	9.9595	9.9325	9.9698	9.9112	9.8991
Age	9.9945	9.9595	9.9333	9.9714	9.9127	9.9021
Ethnicity	9.9911	9.9612	9.9322	9.9686	9.9093	9.8977
Place Household	9.9930	9.9592	9.9353	9.9724	9.9151	9.9040
Type Household	9.9920	9.9609	9.9343	9.9713	9.9134	9.9027
Number of people in household	9.9916	9.9617	9.9369	9.9746	9.9163	9.9063
Number of people in household 11[a]	9.9908	9.9633	9.9395	9.9735	9.9177	9.9075
Number of people on address	9.9967	9.9779	9.9722	10.0272	9.9902	10.0000
Number of people on address 11[b]	9.9920	9.9658	9.9450	9.9875	9.9363	9.9303
			2018-06			
Mean population	9.9970	10.0032	10.0000	10.0065	9.9978	10.0000
Mean sample after screening	9.9883	10.0272	9.9707	9.9486	9.9040	9.8907
Gender	9.9887	10.0275	9.9712	9.9492	9.9040	9.8909
Marital status	9.9877	10.0276	9.9721	9.9502	9.9058	9.8932
Age	9.9875	10.0278	9.9731	9.9519	9.9073	9.8959
Ethnicity	9.9884	10.0268	9.9719	9.9497	9.9046	9.8922
Place Household	9.9879	10.0305	9.9737	9.9524	9.9099	9.8978
Type Household	9.9880	10.0292	9.9740	9.9528	9.9106	9.8984
Number of people in household	9.9888	10.0293	9.9747	9.9549	9.9142	9.9011
Number of people in household 11[a]	9.9888	10.0293	9.9750	9.9554	9.9143	9.9015
Number of people on address	9.9897	10.0578	10.0193	10.0123	9.9885	10.0000
Number of people on address 11[b]	9.9876	10.0375	9.9841	9.9656	9.9249	9.9201

Table F.2: Estimated means by the generalised regression estimator for different auxiliary variables. Based on the sample of the mobility survey of April, May and June 2018.

[a] This is a categorical variable of the number of people in a household consisting of eleven categories: 1 to 10 and 11 or more.

[b] This is a categorical variable of the number of people on an address consisting of eleven categories: 1 to 10 and 11 or more.

	$\mathbf{V}_0$	$\mathbf{V}_{0.2}$	$\mathbf{V}_{0.4}$	$\mathbf{V}_{0.6}$	$\mathbf{V}_{0.8}$	$\mathbf{V}_1$
			2018-07			
Mean population	10.0023	10.0011	9.9983	10.0000	10.0016	10.0000
Mean sample after screening	9.9816	9.9987	9.9134	9.9364	9.9052	9.8923
Gender	9.9815	9.9986	9.9134	9.9361	9.9048	9.8919
Marital status	9.9814	9.9983	9.9146	9.9381	9.9069	9.8949
Age	9.9822	10.0009	9.9187	9.9415	9.9115	9.9009
Ethnicity	9.9827	9.9994	9.9135	9.9375	9.9065	9.8944
Place Household	9.9798	10.0015	9.9191	9.9436	9.9151	9.9043
Type Household	9.9795	10.0014	9.9181	9.9435	9.9150	9.9038
Number of people in household	9.9801	10.0009	9.9168	9.9446	9.9152	9.9047
Number of people in household 11[a]	9.9801	10.0011	9.9174	9.9453	9.9157	9.9055
Number of people on address	9.9788	10.0211	9.9511	9.9998	9.9932	10.0000
Number of people on address 11[b]	9.9791	10.0076	9.9303	9.9664	9.9467	9.9424
			2018-08			
Mean population	10.0062	9.9998	9.9970	9.9993	10.0019	10.0000
Mean sample after screening	9.9667	9.9534	9.9931	9.9254	9.8944	9.8860
Gender	9.9666	9.9533	9.9931	9.9255	9.8945	9.8860
Marital status	9.9657	9.9547	9.9955	9.9283	9.8985	9.8909
Age	9.9667	9.9560	9.9950	9.9288	9.8983	9.8905
Ethnicity	9.9667	9.9543	9.9938	9.9263	9.8961	9.8878
Place Household	9.9675	9.9546	9.9946	9.9278	9.8976	9.8897
Type Household	9.9668	9.9542	9.9948	9.9279	9.8976	9.8901
Number of people in household	9.9684	9.9558	9.9961	9.9307	9.9035	9.8968
Number of people in household 11[a]	9.9685	9.9549	9.9955	9.9306	9.9045	9.8975
Number of people on address	9.9620	9.9750	10.0389	9.9943	9.9858	10.0000
Number of people on address 11[b]	9.9651	9.9583	10.0038	9.9440	9.9201	9.9192
			2018-09			
Mean population	9.9960	9.9971	9.9992	10.0011	10.0005	10.0000
Mean sample after screening	10.0183	9.9635	9.9996	9.9179	9.9148	9.8867
Gender	10.0190	9.9644	9.9997	9.9182	9.9153	9.8876
Marital status	10.0180	9.9661	10.0022	9.9225	9.9207	9.8940
Age	10.0184	9.9656	10.0041	9.9232	9.9207	9.8949
Ethnicity	10.0182	9.9630	10.0003	9.9184	9.9159	9.8883
Place Household	10.0174	9.9660	10.0024	9.9217	9.9202	9.8926
Type Household	10.0173	9.9654	10.0012	9.9202	9.9194	9.8912
Number of people in household	10.0178	9.9666	10.0028	9.9254	9.9254	9.8992
Number of people in household 11[a]	10.0179	9.9663	10.0031	9.9261	9.9262	9.9001
Number of people on address	10.0260	9.9889	10.0434	9.9826	10.0055	10.0000
Number of people on address 11[b]	10.0211	9.9739	10.0125	9.9399	9.9468	9.9265

Table F.3: Estimated means by the generalised regression estimator for different auxiliary variables. Based on the sample of the mobility survey of July, August and September 2018.

[a] This is a categorical variable of the number of people in a household consisting of eleven categories: 1 to 10 and 11 or more.

[b] This is a categorical variable of the number of people on an address consisting of eleven categories: 1 to 10 and 11 or more.

	$\mathbf{V}_0$	$\mathbf{V}_{0.2}$	$\mathbf{V}_{0.4}$	$\mathbf{V}_{0.6}$	$\mathbf{V}_{0.8}$	$\mathbf{V}_1$
			2018-10			
Mean population	10.0016	10.0022	10.0021	9.9996	10.0024	10.0000
Mean sample after screening	10.0259	10.0087	9.9360	9.8764	9.8950	9.8873
Gender	10.0259	10.0088	9.9360	9.8763	9.8950	9.8872
Marital status	10.0262	10.0099	9.9377	9.8791	9.8984	9.8916
Age	10.0259	10.0103	9.9401	9.8813	9.9010	9.8951
Ethnicity	10.0258	10.0087	9.9360	9.8764	9.8951	9.8874
Place Household	10.0261	10.0113	9.9392	9.8822	9.9015	9.8961
Type Household	10.0268	10.0119	9.9399	9.8829	9.9016	9.8969
Number of people in household	10.0248	10.0109	9.9407	9.8845	9.9045	9.8992
Number of people in household 11[a]	10.0252	10.0108	9.9410	9.8850	9.9051	9.8999
Number of people on address	10.0285	10.0357	9.9825	9.9469	9.9837	10.0000
Number of people on address 11[b]	10.0256	10.0179	9.9506	9.8987	9.9254	9.9256
			2018-11			
Mean population	9.9978	9.9952	9.9981	9.9979	9.9994	10.0000
Mean sample after screening	9.9526	9.9713	9.9553	9.9552	9.9247	9.8891
Gender	9.9527	9.9713	9.9545	9.9551	9.9246	9.8886
Marital status	9.9513	9.9722	9.9570	9.9583	9.9286	9.8936
Age	9.9510	9.9714	9.9579	9.9594	9.9302	9.8966
Ethnicity	9.9526	9.9705	9.9562	9.9565	9.9263	9.8913
Place Household	9.9516	9.9751	9.9586	9.9636	9.9325	9.9000
Type Household	9.9515	9.9739	9.9581	9.9620	9.9316	9.8979
Number of people in household	9.9515	9.9734	9.9595	9.9636	9.9333	9.9007
Number of people in household 11[a]	9.9516	9.9733	9.9597	9.9641	9.9346	9.9014
Number of people on address	9.9455	9.9968	10.0033	10.0258	10.0130	10.0000
Number of people on address 11[b]	9.9509	9.9791	9.9719	9.9775	9.9538	9.9252
			2018-12			
Mean population	10.0007	10.0061	10.0002	10.0029	10.0025	10.0000
Mean sample after screening	9.9433	9.9903	9.9524	9.9013	9.9098	9.8873
Gender	9.9426	9.9897	9.9525	9.9017	9.9096	9.8868
Marital status	9.9430	9.9901	9.9529	9.9026	9.9120	9.8900
Age	9.9441	9.9926	9.9559	9.9039	9.9138	9.8928
Ethnicity	9.9419	9.9907	9.9523	9.9037	9.9116	9.8895
Place Household	9.9432	9.9911	9.9545	9.9067	9.9160	9.8952
Type Household	9.9428	9.9906	9.9548	9.9063	9.9159	9.8951
Number of people in household	9.9451	9.9909	9.9549	9.9069	9.9167	9.8951
Number of people in household 11[a]	9.9450	9.9913	9.9555	9.9073	9.9170	9.8954
Number of people on address	9.9482	10.0079	9.9900	9.9658	10.0012	10.0000
Number of people on address 11[b]	9.9469	9.9943	9.9649	9.9210	9.9370	9.9191

Table F.4: Estimated means by the generalised regression estimator for different auxiliary variables. Based on the sample of the mobility survey of October, November and December 2018.

[a] This is a categorical variable of the number of people in a household consisting of eleven categories: 1 to 10 and 11 or more.

[b] This is a categorical variable of the number of people on an address consisting of eleven categories: 1 to 10 and 11 or more.