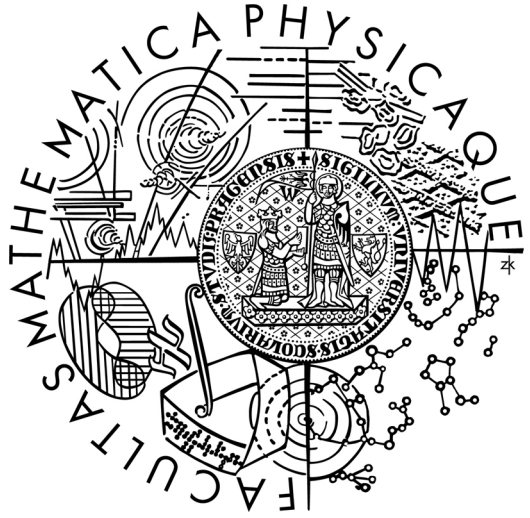
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta
Katedra pravděpodobnosti a matematické statistiky

DIZERTAČNÍ PRÁCE



Michaela Šedová

Odhad parametru při dvoufázovém stratifikovaném a skupinovém výběru

Vedoucí dizertační práce: Doc. Mgr. Michal Kulich, Ph.D.
Studijní obor: Pravděpodobnost a matematická statistika

Prohlašuji, že jsem tuto dizertační práci vypracovala samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V Praze dne 31. 3. 2011

Acknowledgement

Firstly, I would like to thank my supervisor doc. Michal Kulich. Not only did his valuable advice and comments contribute greatly to this thesis, but during various discussions I had the opportunity to learn much from him about modern statistical thinking and approaches to problem solving. I am also grateful for his understanding, patience and flexibility, especially during my stay abroad. I would like to thank all those members of the Department of Probability and Mathematical Statistics who either taught me or gave me guidance on my own teaching experience. I also appreciate that I had an opportunity to spend the first year of my PhD program studying at the Center for Statistics of the Hasselt University in Belgium. Last, but not least, I would like to express my thanks to those friends and colleagues, who provided me with valuable feedback, or supported me otherwise, throughout the period of my PhD studies.

Abstrakt

Název práce: Odhad parametru při dvoufázovém stratifikovaném a skupinovém výběru
Autor: Mgr. Michaela Šedová
Katedra: Katedra pravděpodobnosti a matematické statistiky
Vedoucí dizertační práce: Doc. Mgr. Michal Kulich, Ph.D.
Abstrakt: V této práci se věnujeme metodám odhadu parametru při dvoufázovém stratifikovaném a skupinovém výběru. Narozdíl od klasické teorie výběrových šetření se nezabýváme parametry charakterizujícími konečnou populaci, ale soustředíme se na situaci, kdy jsou pozorování považována za realizace náhodné veličiny. Nás pak zajímají parametry modelu, který tuto náhodnou veličinu popisuje. Přesto však teorie výběrových šetření využíváme, neboť musíme zohlednit dané výběrové schéma. Uvedené metody můžeme tedy chápat jako kombinaci obou přístupů. Pro obě výběrová schémata pracujeme s konceptem, kdy je populace považována za výběr získaný v první fázi, z něhož v druhé fázi obdržíme podvýběr. Sledovaná veličina je pozorovaná pouze pro jedince z podvýběru. Věnujeme se odhadu střední hodnoty, včetně jeho statistických vlastností, a popisujeme, jak je možné najít přesnější odhad v případě, že je k dispozici pomocná veličina známá pro celou populaci a korelovaná se sledovanou veličinou. Tuto metodu rozšiřujeme také na obecný problém odhadu regresního parametru.
Klíčová slova: dvoufázový výběr, Horvitz-Thompsonův odhad, odhad parametrů modelu, skupinový výběr, stratifikovaný výběr

Abstract

Title: Parameter Estimation under Two-phase Stratified and Cluster Sampling
Author: Mgr. Michaela Šedová
Department: Department of Probability and Mathematical Statistics
Supervisor: Doc. Mgr. Michal Kulich, Ph.D.
Abstract: In this thesis we present methods of parameter estimation under two-phase stratified and cluster sampling. In contrast to classical sampling theory, we do not deal with finite population parameters, but focus on model parameter inference, where the observations in a population are considered to be realisations of a random variable. However, we consider the sampling schemes used, and thus we incorporate much of survey sampling theory. Therefore, the presented methods of the parameter estimation can be understood as a combination of the two approaches. For both sampling schemes, we deal with the concept where the population is considered to be the first-phase sample, from which a subsample is drawn in the second phase. The target variable is then observed only for the subsampled subjects. We present the mean value estimation, including the statistical properties of the estimator, and show how this estimation can be improved if some auxiliary information, correlated with the target variable, is observed for the whole population. We extend the method to the regression problem.
Keywords: cluster sampling, Horvitz-Thompson estimation, model-based inference, two-phase sampling, stratified sampling

Contents

Chapter 1

Introduction

1.1 Motivation

Motivation to investigate the topic of this thesis originally arose from Project ACCEPT [1], a phase III randomized controlled trial of HIV prevention in Africa and Thailand. The complexity and uniqueness of the study lead to many challenging statistical questions, one of which we chose to study in more detail here. While the main objectives and design of Project ACCEPT will be described below and will serve as an example, the results of the thesis are far more general and can be applied in other areas and in different contexts.

In countries hit by HIV epidemic, a lot of different strategies to reduce the incidence of HIV have been and also will be applied. However, it is very hard to quantify and assess their effectiveness. Project ACCEPT is the first randomized controlled phase III trial to determine the efficacy of a behavioral and social science intervention with an HIV incidence endpoint in the developing world. In this trial, 34 communities in Africa and 14 communities in Thailand were randomized to either intervention arm, consisting of community based HIV voluntary counseling and testing (CBVCT) plus standard clinic-based VCT (SVCT), or to a control arm, consisting of SVCT alone. The primary objective of this study is to test the hypothesis that communities receiving 3 years of CBVCT, relative to communities receiving 3 years of SVCT, will have significantly lower prevalence of recent HIV infection.

The CBVCT involves strategies which are designed to change community norms and reduce risk for HIV infection among all community members, irrespective of whether they participated directly in the intervention. Thus, the assessment of efficacy is based on changes in the communities' risk behaviors, using repeat cross sectional data collected using household probability samples, as well as community-level prevalence of recent HIV infection determined approximately 3 years after services are introduced in each community. This is different from traditional study designs, namely that the individuals receiving the two different kinds of VCT are studied. Rather, the interest is in the impact of community-based VCT on the entire community, relative to standard VCT.

Baseline and post-intervention assessments are being conducted using the same house-

hold probability sampling technique. Households are selected at random, and then an eligible member of the household is selected at random and offered participation in the assessment. Such sampling procedure implies that the selected households are represented by one member, irrespective of their size. Thus, a simple random sample is not available. Rather, the probability of a community member being included in the sample depends on the size of the household to which he/she belongs. Consequently, people from "small" households contribute to the obtained sample more and people from "big" households contribute less compared to reality. If the measured endpoint is associated with the size of the household, it might bring along a challenge for a statistician.

For example, in Project ACCEPT we could be interested in alcohol consumption as a potential risk behavior. More precisely, we might want to estimate a mean monthly amount of consumed alcohol per person in a given community. It might occur that members of bigger households come from a different social background and have different habits in drinking of alcohol. If we estimated the mean of alcohol consumed per person as an average of the values observed for the selected individuals, we would get biased result, since we would ignore the fact that people from bigger households are underrepresented in the sample.

In summary, we face the problem of making inference based on a sample which was not produced by simple random sampling. Instead, some more complex sampling scheme was involved resulting in unequal probabilities of inclusion for different subjects. While this situation is usually not considered in classical statistics, it is very common in survey sampling theory. For this reason, in the next section we describe basic concepts of the two statistical areas, differences between them and our position in between the two.

As we have already mentioned, baseline as well as post-intervention assessments are being conducted using household probability sampling. Only one member from each household will be asked to participate in the assessment. However, during the visit of the project staff in the household, basic data about the rest of members as well as about the household as a whole will also be collected. This information might be valuable and associated with the outcome which would be obtained were all the household members assessed. Therefore, the main focus of this thesis will be on making use of such auxiliary information.

1.2 Classical versus Sample Survey Inference

In classical sampling theory, a finite population is a basic object of interest. It is often a group of people living in the same geographical area (e.g. country), sometimes specified by other characteristics (e.g. age). The target of inference is the finite population parameter, e.g. the total or the mean of N fixed values. For example, we could be interested in the mean income of an adult person in the Czech republic. Since the whole population cannot be observed, a sample is drawn (involving predefined sampling scheme) and based on this sample the inference about the target parameter is made. As we deal with fixed values, the only source of randomness lies in the sampling process.

In classical statistics, however, understanding of reality is different. Observations in

a population are considered to be realisations of a random variable. The target of inference is then a parameter which characterizes its distribution (e.g. expectation). The source of randomness consists in the stochastic model generating the observations.

A graphical representation of the two concepts can be found in Figure 1.1. The sample survey approach is more sensible e.g. in administrative applications. For addressing scientific questions, however, the latter understanding of the data is typically more appropriate. For example, in Project ACCEPT, we would like to generalize results for other similar populations or for the same population, but in different time. Another example would be comparing two group means. While it is of interest to ask if the expectations in the two groups are equal, there is no point to ask whether finite population means are equal. [7] On the other hand, the classical statistical methods usually assume that a simple random sample is available. As we have already mentioned, it is not always possible especially in the case of extensive epidemiologic studies. In some situations, well designed more complex sampling can be actually also more efficient.

The above suggests an idea to combine the two approaches. It means to draw inference about parameters associated with the stochastic model generating the data, while taking into account the sampling scheme. In sample survey literature, this concept is not entirely new. Some authors describe it with the help of "superpopulation", a hypothetical infinite population from which a finite population is sampled. It is also referred to as a "model-based" approach as opposed to a classical "design-based" approach. Graubard and Korn devoted two papers to this topic ([7],[8]), where they focus on comparison of variance estimation under the two approaches. A recently published book from Fuller [6] puts much more emphasis on links between standard survey techniques and classical statistical methods than the "traditional" literature such as [16]. The last chapter called "Analytic studies" deals with the use of survey data for the estimation of a model parameter. In epidemiologic applications, this concept was used in situations when subsampling from a large cohort is required to obtain additional more detailed information ([3], [4]).

1.3 Two Phase Sampling

The concept of superpopulation is useful in the unification of the classical and survey statistics. We can see the final sample as a result of two-phase sampling (see Figure 1.1). In the first phase, a simple random sample is drawn from a hypothetical superpopulation and a large finite population is obtained. This phase can also be understood as generation of observations by a model. In phase two, a possibly more complex sampling scheme is employed to draw a random subsample for a measurement of the target variable [3]. Thus, when considering properties of estimators, we must incorporate the two sources of variability. The first one (denoted by subscript I) results from the generation of the finite population by a model, the second one (denoted by subscript II) stems from subsampling. It implies that for the estimator $\hat{\theta}$ of the parameter θ, we can write

$$\mathrm{E}\,\hat{\theta} = \mathrm{E}_I(\mathrm{E}_{II}(\hat{\theta}))$$

and

$$\operatorname{var}\hat{\theta} = \operatorname{var}_I(\mathrm{E}_{II}(\hat{\theta})) + E_I(\operatorname{var}_{II}(\hat{\theta})). \tag{1.1}$$

In other words, the first component of the estimator represents the model-based variance of the usual estimates which would be obtained if the full data were available for the entire finite population. The second component then results from observing only a subsample.

We can illustrate this with a simple example (see e.g. [6], pg. 344). Let us assume that the finite population is a realization of N iid (independent identically distributed) random variables with expectation μ and variance σ^2. From the finite population, a simple random sample of size n is drawn. The usual estimator from survey literature, $\bar{y}_n$, estimates the finite population average, $\bar{y}_N$. The finite population average estimates expectation (i.e. model parameter). Thus we have

$$\operatorname{var}_I(\mathrm{E}_{II}(\bar{y}_n)) = \operatorname{var}_I(\bar{y}_N) = \frac{\sigma^2}{N}.$$

The design-based variance of $\bar{y}_n$ presented in classical survey literature is

$$\operatorname{var}_{II}(\bar{y}_n) = \frac{1-f}{n} S_N^2, \quad \text{where } f = \frac{n}{N} \text{ and } S_N^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y}_N)^2.$$

We get

$$E_I(\operatorname{var}_{II}(\hat{\theta})) = \frac{1-f}{n}\sigma^2$$

and therefore

$$\operatorname{var}_I(\mathrm{E}_{II}(\bar{y}_n)) + E_I(\operatorname{var}_{II}(\bar{y}_n)) = \frac{1}{N}\sigma^2 + \Big(\frac{1}{n} - \frac{1}{N}\Big)\sigma^2 = \frac{\sigma^2}{n},$$

which is the model-based variance of estimator $\bar{y}_n$.

1.4 Auxiliary Information

In certain situations, we have access to some auxiliary information. It means that one or more variables, closely correlated with the variable of interest, are observed for the entire population. Use of such information is a very common topic in survey literature. The auxiliary variables can be employed at design or estimation stage. A typical example of the first case is so called *probability proportional-to-size sampling* [16], where the elements are selected with probability proportional to the auxiliary variable. This approach is often used in the context of Bernoulli sampling.

One well known technique which involves the vector of auxiliary variables $\boldsymbol{x}$ in the estimation stage is the *regression estimator*. In brief, it replaces the unobserved values of the target variable y by values predicted by the regression model with $\boldsymbol{x}$ serving as

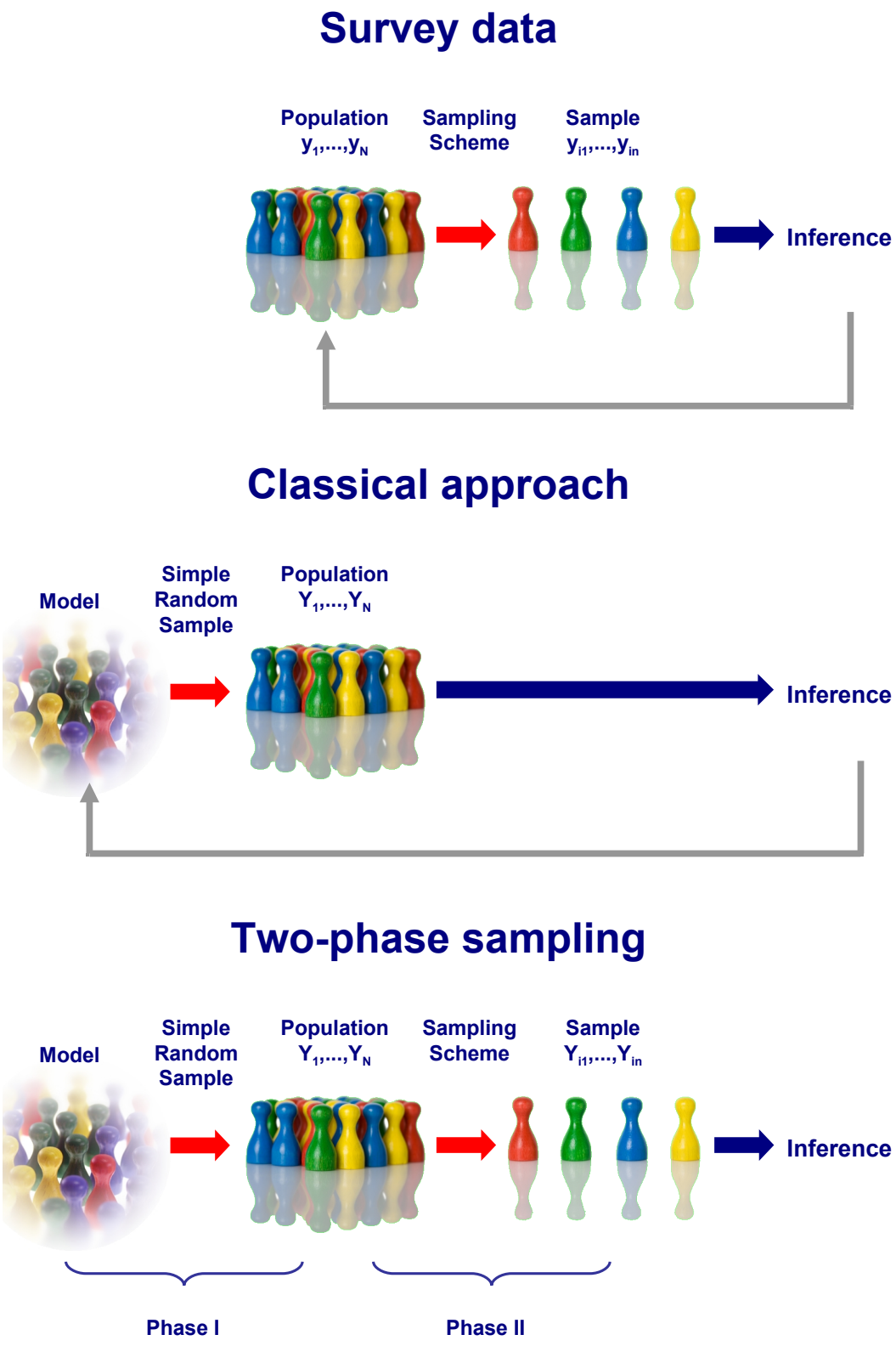


Figure 1.1: Survey sampling, classical and two-phase sampling concepts

explanatory variables (see [6]). It is a member of more general class of estimators linear in y, i.e. the estimators of the form

$$\sum_{i \in s} w_i y_i,$$

where s denotes the subsample from the finite population. There are a lot of different options how the weights w_i can be defined. An important example is the requirement to satisfy *the calibration property*

$$\sum_{i \in s} w_i \boldsymbol{x}_i = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{x_i}. \tag{1.2}$$

One way to construct weights with calibration property is to minimize a function of the weights subject to the restriction (1.2).

In this thesis, we will focus on the adjustment of the weights which is more common in biostatistics and epidemiology (see e.g. [15]). More specifically, we will use weights estimated from a parametric model

$$\pi_i = P(\xi_i | \boldsymbol{x}_i),$$

where ξ_i is a sampling indicator. This approach is closely related to a general missing data problem, where *inverse probability estimators* are often considered as one of the possible solutions. While in our case data are missing by design (as a result of a prespecified sampling scheme), in other applications we encounter data missing by chance.

The purpose of the techniques employing auxiliary information is to reduce the variance of the resulting estimator. As we have already mentioned, in the context of superpopulation inference the variance has two components, see expression (1.1). Apparently, only the second one is amenable to improvement by the auxiliary information available for the finite population. Thus the ideal case occurs when the auxiliary variable is identical to the target variable and the variance reaches its lower limit, that is the value of the first "model-based" component.

The main focus of this thesis is on estimating a model parameter under the two most common sampling schemes: stratified and cluster sampling. We summarize and where needed clarify published results related to model-based inference and relate them to finite-population inference. Our interest is mainly in the use of auxiliary variables to improve the properties of the estimators. In the case of stratified sampling the idea has been extensively studied, considering also semiparametric models. Especially Breslow et al. made a lot of effort to present this approach to the broad community of biostatisticians and epidemiologists ([3], [4]) and encourage its use. They also provided links to the `survey` package in freely available statistical software `R` [14], where the methods are implemented ([11], [12]).

The use of auxiliary variables in stratified sampling led us to question whether a similar method could be applied in case of cluster sampling. To our knowledge, this idea was never studied in published literature. The original results of this thesis therefore relate to the

use of auxiliary information in cluster or combined stratified cluster sampling. While the presented method of employing the auxiliary information under the cluster sampling is similar to the one used under the stratified sampling, it differs in important aspects; especially in the derivation of the asymptotic properties. Some of the results described in chapters 2 and 3 have been published in [17, 18, 19].

All statistical analyses and simulations presented in this thesis were carried out in the statistical package `R`, version 2.9.1 [14].

1.5 Estimation of Expectation under a Sampling Scheme

Let us describe a basic idea which is used in different modifications throughout this thesis. Although it will seem to be a little bit overcomplicated for the following simple example, it illustrates the reasoning behind its applications in a more complex setting.

Let Y be a random variable with mean $\mathrm{E}\,Y = \theta$, where $\theta < \infty$. From a population of size N, a sample is drawn. Let ξ_i be a dichotomous random variable indicating whether an individual i was sampled or not and let π_i be his or her sampling probability. We assume that variables ξ_i are independent. Consequently, the sample size is random. Such design is called *Bernoulli* or *Poisson Sampling* [16].

The *Horvitz-Thompson* estimator of parameter θ is defined as

$$\tilde{\theta} = \frac{1}{N} \sum_{i=1}^{N} \frac{\xi_i}{\pi_i} Y_i. \tag{1.3}$$

Let us assume that all the sampling probabilities are equal, $\pi = \pi_1 = \pi_2 = \cdots = \pi_N$. The estimator of θ using (1.3) is

$$\tilde{\theta} = \frac{1}{N\pi} \sum_{i=1}^{N} \xi_i Y_i.$$

It follows that

$$\sqrt{N}(\tilde{\theta} - \theta) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left(\frac{\xi_i}{\pi} Y_i - \theta \right) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} Q_i, \text{ where } Q_i = \frac{\xi_i}{\pi} Y_i - \theta.$$

The sampling probabilities are independent of Y and thus we can write

$$\begin{aligned} \mathrm{E}\, Q_i &= \mathrm{E}\, \frac{\xi_i}{\pi}\, \mathrm{E}\, Y_i - \theta = 0 \\ \Sigma_{\tilde{\theta}} &= \operatorname{var} Q_i = \mathrm{E}\, Q_i^2 = \frac{1}{\pi} \operatorname{var} Y_i + \frac{1-\pi}{\pi} \theta^2. \end{aligned}$$

Since Q_i are independent identically distributed (iid) variables, according to the Central limit theorem

$$\sqrt{N}(\tilde{\theta} - \theta) \xrightarrow{d} N(0, \Sigma_{\tilde{\theta}}).$$

The disadvantage of this estimator is that its asymptotic variance depends on θ^2, meaning that it is not invariant to location.

One could also use another estimator with estimated sampling probabilities, defined as

$$\hat{\theta} = \frac{1}{N\hat{\pi}} \sum_{i=1}^{N} \xi_i Y_i, \quad \text{where } \hat{\pi} = \frac{1}{N} \sum_{i=1}^{N} \xi_i.$$

We have

$$\sqrt{N}(\hat{\theta} - \theta) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{\xi_i}{\pi} Y_i + \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \Big(\frac{1}{\hat{\pi}} - \frac{1}{\pi}\Big) \xi_i Y_i - \sqrt{N}\theta.$$

By the Taylor expansion of $\frac{1}{\hat{\pi}}$ around $\frac{1}{\pi}$, we get

$$\frac{1}{\hat{\pi}} - \frac{1}{\pi} = -\frac{1}{\pi^2}(\hat{\pi} - \pi) + o_p\Big(\frac{1}{\sqrt{N}}\Big) = -\frac{1}{\pi^2} \frac{1}{N} \sum_{i=1}^{N} (\xi_i - \pi) + o_p\Big(\frac{1}{\sqrt{N}}\Big).$$

Thus (see the proof of the Theorem 1)

$$\begin{aligned}
\sqrt{N}(\hat{\theta} - \theta) &= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{\xi_i}{\pi} Y_i - \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{1}{N\pi^2} \sum_{j=1}^{N} (\xi_j - \pi) \xi_i Y_i - \sqrt{N}\theta + o_p(1) \\
&= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} Q_i + o_p(1),
\end{aligned}$$

where

$$Q_i = \frac{\xi_i}{\pi} Y_i - \frac{\xi_i - \pi}{\pi} \theta - \theta$$

are iid, $\mathrm{E}\, Q_i = 0$ and

$$\Sigma_{\hat{\theta}} = \operatorname{var} Q_i = \frac{1}{\pi} \operatorname{var} Y.$$

According to the Central limit theorem,

$$\sqrt{N}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \Sigma_{\hat{\theta}}).$$

We can see that replacing the known sampling probability π by its estimate $\hat{\pi}$ results in an estimator $\hat{\theta}$ with better properties. Its asymptotic variance is invariant to location and moreover always smaller than or equal to the variance of estimator $\tilde{\theta}$, i.e. $\Sigma_{\hat{\theta}} \leq \Sigma_{\tilde{\theta}}$.

Chapter 2

Two-phase Stratified Sampling

2.1 Introduction

Stratified Sampling

In *stratified sampling*, the population of N elements is divided in K subpopulations of $N_1, N_2, \ldots, N_K$ elements, respectively. These subpopulations are called strata and they comprise the whole of the population, $N_1 + N_2 + \cdots + N_K = N$. A sample is drawn independently from each stratum.

Stratification is a common technique in survey sampling. On top of administrative and logistical advantages it can also produce a gain in precision in the estimates of the population parameters. It may be possible to divide a heterogenous population into subpopulations, each of which is internally homogenous. Then each stratum mean is estimated with high precision and these estimates can be combined into a precise estimate for the whole population. [5]

If no information about the strata is available at the outset, two-phase sampling is used. A large first-phase sample is selected and stratified with the aid of the auxiliary characteristics observed, at low cost, for the elements in the first-phase sample. The second phase is then carried out as stratified sampling, with a considerably smaller sample size and the variable of interest observed for this small sample.[16] This sampling design was proposed by Neyman [13].

Two-phase designs have also been suggested in epidemiology. They are particularly valuable when a large cohort (i.e. first-phase sample) is under surveillance for a disease event of interest and sampling from the cohort is required to obtain information on additional covariates. Standard case-control designs stratify second phase sampling on disease status, while it was also proposed [20] to stratify on both disease status and exposure with the aim to gain efficiency. [3]

In these epidemilogic applications, the target of inference are not a finite population parameters, but superpopulation parameters which should help to answer some scientific question. Thus, the cohort is understood as a first-phase sample from the infinite superpopulation. Then information about the strata is identified, and based on it, the second

phase sample is drawn. The analysis is performed on the subsample.

As we have already mentioned, this chapter mainly summarizes results contained in biostatistical and epidemiological literature, see e.g. [3], [4]. However, since we felt that sometimes they do not provide enough details and insight, we chose to present our own form of statements and their proofs. It will also help to build a link to the results of the next chapters.

Statistical Formulation

A statistical formulation of this problem is as follows. Let Y be a random variable of interest and W a discrete random variable taking on values from $\{1, 2, \ldots, K\}$. In our case, W corresponds to the stratum in a population. We write

$$\mathrm{I}_{ik} = \begin{cases} 1 & \text{if } W_i = k \\ 0 & \text{otherwise.} \end{cases}$$

Let N_k denotes the number of individuals in stratum k, $\sum_{k=1}^{K} N_k = N$. Let ξ_i be the sampling indicator and π_k is a sampling probability of an individual who belongs to stratum k. We assume that variables ξ_i are independent, which means that within each stratum we perform Bernoulli sampling with constant probabilities

$$\mathrm{E}\,(\xi_i | W_i = k) = \mathrm{P}(\xi_i = 1 | W_i = k) = \pi_k, \quad \text{for} \quad i = 1, 2, \ldots, N.$$

We also assume that W_i is observed for all N members of the population, while Y_i is observed only for the selected individuals, i.e. when $\xi_i = 1$.

2.2 Estimation of the Mean

We denote

$$\theta = \mathrm{E}\, Y = \sum_{k=1}^{K} p_k \theta_{[k]}, \tag{2.1}$$

where

$$\theta_{[k]} = \mathrm{E}_{\,k} Y = \mathrm{E}\,(Y | W = k) \quad \text{and} \quad p_k = \mathrm{P}(W = k).$$

We define the estimator of the parameter θ as

$$\hat{\theta} = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} \Big(\frac{1}{\hat{\pi}_k} \xi_i Y_i \Big) \mathrm{I}_{ik}, \quad \text{where} \quad \hat{\pi}_k = \frac{1}{N_k} \sum_{i=1}^{N} \xi_i \mathrm{I}_{ik}, \tag{2.2}$$

and N_k is the number of individuals belonging to stratum k. The estimator $\hat{\theta}$ has a form of weighted average of the observations, where the weights are the reciprocal values of the empirical sampling probabilities. For a given stratum k, the sampling probability is estimated as the number of the sampled individuals divided by the total number of individuals in stratum k.

Theorem 1. *Assume that vectors* (Y_i, W_i, ξ_i) *are iid and* ξ_i *is independent of* Y_i *given* W_i*, for* $i = 1, 2, \ldots, N$*. Assume that* $\operatorname{var} Y_i < \infty$*. Then the following holds:*

$$\sqrt{N}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \Sigma), \tag{2.3}$$

where

$$\Sigma = \operatorname{var} Y_i + \sum_{k=1}^{K} p_k \frac{1 - \pi_k}{\pi_k} \operatorname{var}_k Y_i \tag{2.4}$$

and

$$\operatorname{var}_k(Y_i) = \operatorname{var}(Y_i | W_i = k).$$

Proof Since for given k

$$\mathrm{E}\Big(\frac{\xi_i Y_i}{\pi_k} \mathrm{I}_{ik}\Big) = p_k \mathrm{E}_k \frac{\xi_i Y_i}{\pi_k} = p_k \theta_{[k]},$$

according to the weak law of large numbers we have

$$\frac{1}{N_k} \sum_{i=1}^{N} \frac{\xi_i Y_i}{\pi_k} \mathrm{I}_{ik} = \frac{1}{N} \sum_{i=1}^{N} \frac{N}{N_k} \frac{\xi_i Y_i}{\pi_k} \mathrm{I}_{ik} \xrightarrow{P} \theta_{[k]}. \tag{2.5}$$

We can write

$$\sqrt{N}(\hat{\theta} - \theta) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \sum_{k=1}^{K} \Big(\frac{\xi_i Y_i}{\pi_k} + \Big(\frac{1}{\hat{\pi}_k} - \frac{1}{\pi_k} \Big) \xi_i Y_i - \theta \Big) \mathrm{I}_{ik}.$$

By the Taylor expansion of $\frac{1}{\hat{\pi}_k}$ around $\frac{1}{\pi_k}$, we get

$$\begin{aligned} \frac{1}{\hat{\pi}_k} - \frac{1}{\pi_k} &= -\frac{1}{\pi_k^2}(\hat{\pi}_k - \pi_k) + \frac{1}{2} \frac{1}{\pi_k^{*3}} (\hat{\pi}_k - \pi_k)^2 \\ &= -\frac{1}{\pi_k^2} \frac{1}{N_k} \sum_{i=1}^{N} (\xi_i - \pi_k) \mathrm{I}_{ik} + \frac{1}{2} \frac{1}{\pi_k^{*3}} \Big(\frac{1}{N_k} \sum_{i=1}^{N} (\xi_i - \pi_k) \mathrm{I}_{ik} \Big)^2, \end{aligned}$$

where π_k^* is between $\hat{\pi}_k$ and π_k. Thus

$$\begin{aligned} \sqrt{N}(\hat{\theta} - \theta) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \sum_{k=1}^{K} \Big(\frac{\xi_i Y_i}{\pi_k} + \Big[&- \frac{1}{\pi_k^2} \frac{1}{N_k} \sum_{j=1}^{N} (\xi_j - \pi_k) \mathrm{I}_{jk} \\ &+ \frac{1}{2} \frac{1}{\pi_k^{*3}} \Big(\frac{1}{N_k} \sum_{j=1}^{N} (\xi_j - \pi_k) \mathrm{I}_{jk} \Big)^2 \Big] \xi_i Y_i - \theta \Big) \mathrm{I}_{ik}. \end{aligned}$$

For the third expression in the above sum we have

$$\frac{1}{2\pi_k^{*3}}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\sum_{k=1}^{K}\Big(\frac{1}{N_k}\sum_{j=1}^{N}(\xi_j-\pi_k)\mathrm{I}_{jk}\Big)^2\xi_i Y_i\mathrm{I}_{ik}$$
$$=\frac{1}{2\pi_k^{*3}}\sum_{k=1}^{K}\Big(\frac{1}{\sqrt{N}}\sum_{j=1}^{N}\sqrt{\frac{N}{N_k}}(\xi_j-\pi_k)\mathrm{I}_{jk}\Big)^2\frac{1}{\sqrt{N}}\frac{1}{N_k}\sum_{i=1}^{N}\xi_i Y_i\mathrm{I}_{ik}=\sum_{k=1}^{K}O_p(1)o_p(1)=o_p(1),$$

since $\frac{1}{\sqrt{N}}\frac{1}{N_k}\sum_{i=1}^{N}\xi_i Y_i\mathrm{I}_{ik}\xrightarrow{P}0$ and $\frac{1}{\sqrt{N}}\sum_{j=1}^{N}\sqrt{\frac{N}{N_k}}(\xi_j-\pi_k)\mathrm{I}_{jk}$ converges in distribution to a normal random variable. Then we get

$$\sqrt{N}(\hat\theta-\theta)=\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\sum_{k=1}^{K}\Big(\frac{\xi_i Y_i}{\pi_k}-\Big(\frac{1}{\pi_k^2}\frac{1}{N_k}\sum_{j=1}^{N}(\xi_j-\pi_k)\mathrm{I}_{jk}\Big)\xi_i Y_i-\theta\Big)\mathrm{I}_{ik}+o_p(1)$$
$$=\frac{1}{\sqrt{N}}\Big[\sum_{i=1}^{N}\sum_{k=1}^{K}\Big(\frac{\xi_i Y_i}{\pi_k}-\theta\Big)\mathrm{I}_{ik}-\sum_{k=1}^{K}\frac{1}{\pi_k}\frac{1}{N_k}\Big(\sum_{i=1}^{N}\frac{\xi_i Y_i}{\pi_k}\mathrm{I}_{ik}\Big)\Big(\sum_{j=1}^{N}(\xi_j-\pi_k)\mathrm{I}_{jk}\Big)\Big]+o_p(1)$$
$$\overset{(2.5)}{=}\frac{1}{\sqrt{N}}\Big[\sum_{i=1}^{N}\sum_{k=1}^{K}\Big(\frac{\xi_i Y_i}{\pi_k}-\theta\Big)\mathrm{I}_{ik}-\sum_{k=1}^{K}\frac{1}{\pi_k}\theta_{[k]}\sum_{j=1}^{N}(\xi_j-\pi_k)\mathrm{I}_{jk}\Big]+o_p(1)$$
$$=\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\sum_{k=1}^{K}\Big(\frac{\xi_i Y_i}{\pi_k}-\frac{\xi_i-\pi_k}{\pi_k}\theta_{[k]}-\theta\Big)\mathrm{I}_{ik}+o_p(1). \tag{2.6}$$

We can write

$$\sqrt{N}(\hat\theta-\theta)=\frac{1}{\sqrt{N}}\sum_{i=1}^{N}Q_i+o_p(1),$$

where

$$Q_i=\sum_{k=1}^{K}\Big(\frac{\xi_i Y_i}{\pi_k}-\frac{\xi_i-\pi_k}{\pi_k}\theta_{[k]}-\theta\Big)\mathrm{I}_{ik}$$

are iid variables and

$$\mathrm{E}\,Q_i=\mathrm{E}\left(\mathrm{E}\left[Q_i|W_i\right]\right)=\mathrm{E}\Big(\frac{1}{\pi_k}\mathrm{E}_k\xi_i\mathrm{E}_k Y_i\Big)-\theta=0$$
$$\Sigma=\operatorname{var}Q_i=\mathrm{E}\left(\mathrm{E}\left[Q_i^2|W_i\right]\right)=\mathrm{E}\Big(\mathrm{E}\Big[\Big(Y_i-\theta+\frac{\xi_i-\pi_k}{\pi_k}(Y_i-\theta_{[k]})\Big)^2|W_i\Big]\Big)$$
$$=\mathrm{E}\left(\mathrm{E}\left[(Y_i-\theta)^2|W_i\right]\right)+\mathrm{E}\Big(\mathrm{E}\Big[\Big(\frac{\xi_i-\pi_k}{\pi_k}\Big)^2(Y_i-\theta_{[k]})^2|W_i\Big]\Big)$$
$$=\operatorname{var}Y_i+\sum_{k=1}^{N}p_k\frac{1-\pi_k}{\pi_k}\operatorname{var}_k Y_i.$$

According to the Central limit theorem for iid random variables

$$\sqrt{N}(\hat\theta-\theta)\xrightarrow{d}N(0,\Sigma).$$

□

Remark In the proofs of the theorems which follow later in this text, some steps will be very similar to the ones shown in proof of Theorem 1 and thus will not be repeated in detail.

Variance Σ (2.4) can be estimated as

$$\widehat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\frac{\xi_i}{\hat{\pi}_k}(Y_i - \hat{\theta})^2 \mathrm{I}_{ik} + \sum_{k=1}^{K}\frac{N_k}{N}\frac{1-\hat{\pi}_k}{\hat{\pi}_k}\Big(\frac{1}{N_k}\sum_{i=1}^{N}\xi_i(Y_i - \hat{\theta}_{[k]})^2\mathrm{I}_{ik}\Big). \tag{2.7}$$

The variance of $\hat{\theta}$ has an interesting interpretation. As we have already mentioned, the first term corresponds to the variance of $\bar{Y}_N$ which would be used to estimate θ if the entire population could be observed. The second term can be understood as a penalty for observing only a sample. If there is only one stratum in the population (i.e. W is a constant) and the sampling probability is π, we get

$$\Sigma = \operatorname{var} Y_i + \frac{1-\pi}{\pi}\operatorname{var} Y_i = \frac{1}{\pi}\operatorname{var} Y_i.$$

If there are K strata in the population and the sampling probabilities $\pi_1 = \cdots = \pi_K = \pi$ are equal, we get

$$\Sigma = \operatorname{var} Y_i + \frac{1-\pi}{\pi}\sum_k p_k \operatorname{var}_k Y_i.$$

Here the sampling penalty takes the form of $\frac{1-\pi}{\pi}$ times the weighted average of the within-strata variances weighted by the stratum probability. If the variable of interest has a constant value within each stratum, the within-stratum variance is always zero and the sampling penalty vanishes. Obviously, if we knew that all the individuals in the stratum have the same value, we would not lose any information by observing just one individual from each stratum. In this case, the stratum variable W_i contains all the information on the value of Y_i. If the information on Y_i is only partial, i.e. we can a priori split the population into strata with similar values within each of them, it decreases each stratum variance, compared to the situation when W_i contains no information on the value of Y_i. The overall variance (2.4) of the estimator $\hat{\theta}$ then decreases as well. If W_i contains no information on the value of Y_i, splitting population into strata is random, which results in the same variance within each stratum as it is in the whole population. In other words, the stratification of population cannot increase the asymptotic variance of the estimator $\hat{\theta}$ over that of the unstratified estimator (with $K = 1$). This result applies when N_k is large for all $k = 1, 2, \ldots K$.

Remark One might argue that in case of awareness that the population is heterogenous, it might be more appropriate to perform conditional estimation, i.e. to estimate the expectation given stratum. It might certainly be true in some situations, but in other cases the marginal inference is more appropriate. It always depends on the scientific question to be addressed. Marginal inference is for example very common in clinical trials. Although

the population of patients is known to be heterogenous and it is expected that some patients will benefit from a drug under investigation more than others, depending on other (e.g. health, region, genetic, etc.) conditions, still the goal is to evaluate whether the drug works overall. Another example is a situation when the stratification is performed based on the outcome or on an intermediate variable on the causal pathway leading to outcome. We can have two strata of patients suffering from a disease which is accompanied by high fever: patients whose body temperature returned back to normal after certain intervention and patients who did not respond. To conclude whether a patient was finally cured from the disease, depending to which stratum he/she belongs (i.e. whether his temperature decreased back to normal or not), would probably be very straightforward, but would not tell us much about what role the studied intervention played in curing the disease.

2.3 Note on Estimators Presented in Survey Literature

The connection between results presented in Theorem 1 and those which can be found in the survey literature can be seen through the two phases of sampling, as was explained in section 1.3.

Under stratified sampling, the "survey estimator" of the population mean is identical to our estimator $\hat{\theta}$. It estimates the population average, which in turn estimates the expectation θ. The variance of the population average is obviously

$$\mathrm{var}_I(\mathrm{E}_{II}(\hat{\theta})) = \mathrm{var}_I(\bar{y}_N) = \frac{1}{N}\mathrm{var}Y_i.$$

The estimator of the design-based variance of $\hat{\theta}$ (see e.g. [16], pg. 103) is

$$\widehat{\mathrm{var}}_{II}(\hat{\theta}) = \frac{1}{N^2}\sum_{k=1}^{K} N_k^2 \frac{1-f_k}{n_k} s_k^2, \quad \text{where } f_k = \frac{n_k}{N_k},$$

n_k denotes number of individuals sampled from stratum k and

$$s_k^2 = \frac{1}{n_k - 1}\sum_{i=1}^{N} \xi_i (y_i - \hat{\theta}_{[k]})\mathrm{I}_{ik}.$$

The expression $\widehat{\mathrm{var}}_{II}(\hat{\theta})$ estimates

$$E_I(\mathrm{var}_{II}(\hat{\theta})) = \frac{1}{N}\sum_{k=1}^{K} p_k \frac{1-\pi_k}{\pi_k}\mathrm{var}_k Y_i.$$

We get

$$\mathrm{var}_I(\mathrm{E}_{II}(\hat{\theta})) + E_I(\mathrm{var}_{II}(\hat{\theta})) = \frac{1}{N}\Sigma,$$

which is the asymptotic model-based variance of $\hat{\theta}$.

2.4 Use of Auxiliary Variables

When some auxiliary variables, correlated with the target variable, are observed for the whole population, the previously described estimator can be improved. Let us assume an s-dimensional vector of auxiliary variables $\boldsymbol{X}_i$ and denote

$$\boldsymbol{Z}_i^T = (\mathrm{I}_{i1}, \ldots, \mathrm{I}_{iK}, \boldsymbol{X}_i^T).$$

We adjust the sampling weights in the Horvitz-Thompson estimator by fitting the logistic regression model for the sampling probabilities, where stratum indicator as well as the vector of auxiliary variables are included as explanatory variables

$$\log\big(\frac{\pi_i}{1-\pi_i}\big) = \boldsymbol{\gamma}^T \boldsymbol{z}_i. \tag{2.8}$$

Since the sampling probabilities depend only on the stratum and not on the auxiliary variables, the only nonzero components of vector $\boldsymbol{\gamma}$ will be the components reflecting the stratum. The part of the vector of parameters pertaining to the auxiliary variables $\boldsymbol{\gamma}_x = \mathbf{0}$. However, its estimate will never be exactly equal to $\mathbf{0}$ and we obtain the vector of the estimated sampling probabilities

$$\tilde{\pi}_i = \pi_i(\tilde{\boldsymbol{\gamma}}^T \boldsymbol{z}_i) = \frac{\exp(\tilde{\boldsymbol{\gamma}}^T \boldsymbol{z}_i)}{1 + \exp(\tilde{\boldsymbol{\gamma}}^T \boldsymbol{z}_i)}, \tag{2.9}$$

where $\tilde{\boldsymbol{\gamma}}$ is the estimate of parameter $\boldsymbol{\gamma}$.

The estimator of θ has the following form

$$\tilde{\theta} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\tilde{\pi}_i} \xi_i Y_i.$$

Theorem 2. *Assume that vectors $(Y_i, W_i, \boldsymbol{X_i}, \xi_i)$ are iid and ξ_i is independent of Y_i and $\boldsymbol{X_i}$ given W_i, for $i = 1, 2, \ldots, N$. Assume that $\operatorname{var} Y_i < \infty$ and $\operatorname{var} X_{ij} < \infty$ for each component $j = 1, 2, \ldots, s$. Then the following holds:*

$$\sqrt{N}(\tilde{\theta} - \theta) \xrightarrow{d} N(0, \Sigma_z), \tag{2.10}$$

where

$$\Sigma_z = \operatorname{var} Y_i + \sum_k p_k \frac{1-\pi_k}{\pi_k} \operatorname{var}_k Y_i - \boldsymbol{c}^T V^{-1} \boldsymbol{c}, \tag{2.11}$$

and

$$\boldsymbol{c} = \sum_{k=1}^{K} p_k (1-\pi_k) \operatorname{cov}_k(\boldsymbol{X}_i, Y_i), \quad V = \sum_{k=1}^{K} p_k \pi_k (1-\pi_k) \operatorname{var}_k \boldsymbol{X}_i.$$

Proof We can write

$$\sqrt{N}(\tilde{\theta}-\theta)=\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\Big(\frac{\xi_i Y_i}{\pi_i}+\Big(\frac{1}{\tilde{\pi}_i}-\frac{1}{\pi_i}\Big)\xi_i Y_i\Big)-\sqrt{N}\theta.$$

For each $j=1,2,\ldots,K+s$ it holds

$$\frac{\partial}{\partial\gamma_j}\Big(\frac{1}{\pi_i(\boldsymbol{\gamma}^T\boldsymbol{Z}_i)}\Big)=\frac{\partial}{\partial\gamma_j}\Big(\frac{1+\exp(\boldsymbol{\gamma}^T\boldsymbol{Z}_i)}{\exp\left(\gamma^T\boldsymbol{Z}_i\right)}\Big)=-\frac{Z_{ij}}{\exp(\boldsymbol{\gamma}^T\boldsymbol{Z}_i)} \tag{2.12}$$

and we have

$$\tilde{\boldsymbol{\gamma}}-\boldsymbol{\gamma}=\frac{1}{N}J_\gamma^{-1}\sum_{i=1}^{N}\boldsymbol{U}_i(\boldsymbol{\gamma})+o_p\Big(\frac{1}{\sqrt{N}}\Big)=\frac{1}{N}J_\gamma^{-1}\sum_{i=1}^{N}(\xi_i-\pi_i(\boldsymbol{\gamma}^T\boldsymbol{Z}_i))\boldsymbol{Z}_i+o_p\Big(\frac{1}{\sqrt{N}}\Big), \tag{2.13}$$

where $\boldsymbol{U}_i$ is the score and J_γ is the Fisher information matrix pertaining to the model (2.8)

$$J_\gamma=\mathrm{E}\,\pi_i(1-\pi_i)\boldsymbol{Z}_i\boldsymbol{Z}_i^T.$$

By the Taylor expansion of $\frac{1}{\tilde{\pi}_i}$ around $\frac{1}{\pi_i}$, using (2.12) and (2.13) we get

$$\begin{aligned}\frac{1}{\pi_i(\tilde{\boldsymbol{\gamma}}^T\boldsymbol{Z}_i)}-\frac{1}{\pi_i(\boldsymbol{\gamma}^T\boldsymbol{Z}_i)}&=-(\tilde{\boldsymbol{\gamma}}-\boldsymbol{\gamma})^T\frac{\boldsymbol{Z}_i}{\exp(\boldsymbol{\gamma}^T\boldsymbol{Z}_i)}+o_p\Big(\frac{1}{\sqrt{N}}\Big)\\&=-\frac{1}{N}\frac{1}{\exp(\boldsymbol{\gamma}^T\boldsymbol{Z}_i)}\Big[\sum_{j=1}^{N}(\xi_j-\pi_j)\boldsymbol{Z}_j^TJ_\gamma^{-1}\Big]\boldsymbol{Z}_i+o_p\Big(\frac{1}{\sqrt{N}}\Big).\end{aligned}$$

Then

$$\begin{aligned}&\sqrt{N}(\tilde{\theta}-\theta)=\\&=\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\Big(-\frac{1}{N}\frac{1}{\exp(\boldsymbol{\gamma}^T\boldsymbol{Z}_i)}\Big[\sum_{j=1}^{N}(\xi_j-\pi_j)\boldsymbol{Z}_j^TJ_\gamma^{-1}\Big]\boldsymbol{Z}_i\xi_iY_i+\frac{\xi_iY_i}{\pi_i}\Big)-\sqrt{N}\theta+o_p(1)\\&=\frac{1}{\sqrt{N}}\sum_{j=1}^{N}(\xi_j-\pi_j)\boldsymbol{Z}_j^TJ_\gamma^{-1}\Big(-\frac{1}{N}\sum_{i=1}^{N}\frac{\xi_i\boldsymbol{Z}_iY_i}{\exp(\boldsymbol{\gamma}^T\boldsymbol{Z}_i)}\Big)+\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{\xi_iY_i}{\pi_i}-\sqrt{N}\theta+o_p(1).\end{aligned}$$

Since

$$-\frac{1}{N}\sum_{i=1}^{N}\frac{\xi_i\boldsymbol{Z}_iY_i}{\exp(\boldsymbol{\gamma}^T\boldsymbol{Z}_i)}=-\frac{1}{N}\sum_{i=1}^{N}\xi_i\frac{1-\pi_i}{\pi_i}\boldsymbol{Z}_iY_i\xrightarrow{P}\boldsymbol{q}=\mathrm{E}\,(1-\pi_i)\boldsymbol{Z}_iY_i,$$

we have

$$\sqrt{N}(\tilde{\theta}-\theta)=\frac{1}{\sqrt{N}}\sum_{i=1}^{N}Q_i+o_p(1),$$

where Q_i are iid random variables

$$Q_i = \frac{\xi_i Y_i}{\pi_i} + (\xi_i - \pi_i)\boldsymbol{Z}_i^T J_\gamma^{-1}\boldsymbol{q} - \theta.$$

It holds

$$\begin{aligned}\Sigma_z = \operatorname{var} Q_i = \mathrm{E}\, Q_i^2 &= \mathrm{E}\left((Y_i - \theta) + \frac{\xi_i - \pi_i}{\pi_i}(Y_i - \pi_i \boldsymbol{Z}_i J_\gamma^{-1}\boldsymbol{q})\right)^2 \\ &= \operatorname{var} Y_i + \mathrm{E}\,\frac{1-\pi_i}{\pi_i} Y_i^2 - \boldsymbol{q}^T J_\gamma^{-1}\boldsymbol{q}.\end{aligned}$$

Again, according to the Central limit theorem for iid random variables

$$\sqrt{N}(\tilde{\theta} - \theta) \xrightarrow{d} N(0, \Sigma_z).$$

The last step is to show (2.11). We have

$$J_\gamma = \sum_{k=1}^K p_k \pi_k (1-\pi_k) \mathrm{E}_k \boldsymbol{Z}_i \boldsymbol{Z}_i^T = \begin{pmatrix} A & B \\ B^T & D \end{pmatrix},$$

where

$$\begin{aligned} A_{k\times k} &= \operatorname{Diag}\{p_k \pi_k (1-\pi_k)\}_{k=1}^K \\ B_{k\times s} &= [p_k \pi_k (1-\pi_k) \mathrm{E}_k \boldsymbol{X}_i^T]_{k=1}^K \\ D_{s\times s} &= \sum_{k=1}^K p_k \pi_k (1-\pi_k) \mathrm{E}_k \boldsymbol{X}_i \boldsymbol{X}_i^T. \end{aligned}$$

Then

$$J_\gamma^{-1} = \begin{pmatrix} A^{-1} + A^{-1}BP^{-1}B^T A^{-1} & -A^{-1}BP^{-1} \\ -P^{-1}B^T A^{-1} & P^{-1} \end{pmatrix}, \text{ where } P = D - B^T A^{-1} B,$$

and

$$\begin{aligned} A^{-1}B &= [\mathrm{E}_k \boldsymbol{X}_i^T]_{k=1}^K \\ B^T A^{-1} B &= \sum_{k=1}^K p_k \pi_k (1-\pi_k) \mathrm{E}_k \boldsymbol{X}_i \mathrm{E}_k \boldsymbol{X}_i^T \\ P &= \sum_{k=1}^K p_k \pi_k (1-\pi_k) \operatorname{var}_k \boldsymbol{X}_i. \end{aligned}$$

We denote $\boldsymbol{q}^T = (\boldsymbol{q}_1^T, \boldsymbol{q}_2^T)$ for

$$\boldsymbol{q}_1 = [p_k(1-\pi_k)\mathrm{E}_k Y_i]_{k=1}^K, \qquad \boldsymbol{q}_2 = \sum_{k=1}^K p_k(1-\pi_k)\mathrm{E}_k(Y_i \boldsymbol{X}_i),$$

and

$$\boldsymbol{a}^T = \boldsymbol{q}_1^T A^{-1} B = \sum_{k=1}^K p_k(1-\pi_k)\mathrm{E}_k Y_i \mathrm{E}_k \boldsymbol{X}_i^T.$$

Then we get

$$\begin{aligned}\boldsymbol{q}^T J_\gamma^{-1} \boldsymbol{q}^T &= \boldsymbol{q}_1^T A^{-1} \boldsymbol{q}_1 + \boldsymbol{a}^T P^{-1} \boldsymbol{a} - \boldsymbol{q}_2^T P^{-1} \boldsymbol{a} - \boldsymbol{a}^T P^{-1} \boldsymbol{q}_2 + \boldsymbol{q}_2^T P^{-1} \boldsymbol{q}_2 \\ &= \boldsymbol{q}_1^T A^{-1} \boldsymbol{q}_1 + (\boldsymbol{q}_2 - \boldsymbol{a})^T P^{-1} (\boldsymbol{q}_2 - \boldsymbol{a}).\end{aligned}$$

Since

$$\boldsymbol{q}_1^T A^{-1} \boldsymbol{q}_1 = \sum_{k=1}^K p_k \frac{1-\pi_k}{\pi_k} (\mathrm{E}_k Y_i)^2, \qquad \boldsymbol{q}_2 - \boldsymbol{a} = \sum_{k=1}^K p_k (1-\pi_k) \mathrm{cov}_k(\boldsymbol{X}_i, Y_i),$$

we obtain (2.11).

□

The variance Σ_z (2.11) can be estimated similarly to Σ (2.7), with the usual estimators of covariance and variance matrices $\widehat{\mathrm{cov}}_k(\boldsymbol{X}, Y)$ and $\widehat{\mathrm{var}}_k \boldsymbol{X}$ for $k = 1, \ldots, K$.

Since V is a linear combination of variance matrices, V^{-1} is a positive semidefinite matrix and $\boldsymbol{c}^T V^{-1} \boldsymbol{c} \geq 0$. This important observation shows that $\Sigma \geq \Sigma_z$, which means that use of auxiliary variables to adjust weights in the Horvitz-Thompson estimator can never increase the asymptotic variance of the estimator. In the worst case, when auxiliary variables and the target variable are independent, the asymptotic variances of the original estimator $\hat{\theta}$ taking into account only the stratum variable and the estimator $\tilde{\theta}$ with weights adjusted for auxiliary variables are equal. When auxiliary variables are correlated with the variable of interest, the variance of the estimator $\tilde{\theta}$ is lower.

We also do not have to use the auxiliary variable in its original form, but rather seek a transformation which would be most correlated with the target variable Y. If we choose to use as an auxiliary variable $\boldsymbol{X}_i/\pi_k$ for the observation from stratum k ($W_i = k$), we obtain

$$\Sigma_z = \mathrm{var}\, Y_i + \sum_k p_k \frac{1-\pi_k}{\pi_k} \mathrm{var}_k Y_i - \boldsymbol{c}^T V^{-1} \boldsymbol{c},$$

where

$$\boldsymbol{c} = \sum_{k=1}^K p_k \frac{1-\pi_k}{\pi_k} \mathrm{cov}_k(\boldsymbol{X}_i, Y_i), \quad V = \sum_{k=1}^K p_k \frac{1-\pi_k}{\pi_k} \mathrm{var}_k \boldsymbol{X}_i.$$

If we have a variable X for which $\mathrm{cor}_k(X, Y) = 1$, then $c^T V^{-1} c = \sum_k p_k \frac{1-\pi_k}{\pi_k} \mathrm{var}_k Y_i$ and the variance of the estimator $\tilde{\theta}$ reaches its lower limit, $\Sigma_z = \mathrm{var}\, Y_i$. In this sense, the transformation of the auxiliary variable defined as $\boldsymbol{X}_i/\pi_k$, where $W_i = k$, is optimal.

2.5 Example

The above presented results are illustrated by the following hypothetical example. We are interested in the mean serum uric acid concentration (UC). By "mean" we mean expectation of the stochastic distribution of the uric acid concentrations in a general population.

Let us assume a representative population of $N_m = 500$ men and $N_w = 500$ women. From this population, a stratified sample (based on gender) was drawn, where men were sampled with probability $\pi_m = 0.6$ and women with probability $\pi_w = 0.4$. For n_m selected men and n_w selected women, blood samples were taken and the serum uric acid concentration measured. The triacylglycerols levels (Tgl) were available for all the population and thus can be considered as the auxiliary variable.

The sampling probabilities are estimated as $\hat{\pi}_m = n_m/N_m$ and $\hat{\pi}_w = n_w/N_w$. The standard estimator of expectation, which reflects only the stratum, has the following form

$$\hat{\theta} = \frac{1}{1000}\Big(\sum_{i\in s_w}\frac{1}{\hat{\pi}_w}Y_i + \sum_{i\in s_m}\frac{1}{\hat{\pi}_m}Y_i\Big), \tag{2.14}$$

where s_m and s_w denote the set of sampled men and women, resp.

The improved estimator takes into account Tgl of the ith patient

$$\tilde{\theta} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\tilde{\pi}_i}\xi_i Y_i, \tag{2.15}$$

where $\tilde{\pi}_i$ denotes the weights adjusted for triacylglycerols levels divided by the sampling probabilities, see (2.9).

Simulation To demonstrate how these two estimators differ with respect to precision, we carried out a small simulation. It was conducted as follows. We generated a population of 1000 individuals, assigning gender with equal probabilities for men and women. We assumed that the mean value of uric acid concentration in men is $\theta_m = 320$ μmol/L and the mean value in women is $\theta_w = 240$ μmol/L. It implies that in the general population, the mean is

$$\theta = 0.5 * 320 + 0.5 * 240 = 280.$$

Normally, the correlation between Tgl and UA is around 0.3, but for illustration we assumed also correlations equal to 0.6 and 0.9 (within the stratum). The values of UA and Tgl were generated from multivariate normal distribution. In the next step, a subset of individuals was sampled using the stratified sampling described above. The standard estimate (2.14) as well as the estimate taking into account Tgl (2.15) were calculated. This procedure was repeated 1000 times. The results are displayed in Table 2.1. We can see that for each type of estimator, the estimate of the expectation was very close to its real value 280. The estimate of variance was very close to the calculated asymptotic variance of the estimator. The empirical variance of the estimator was in all cases a little bit higher than the calculated and the estimated asymptotic variance, but the difference was acceptable.

Table 2.1: Example of stratified sampling; Results of simulation

Estimator	cor(UC, Tgl)	Average estimate of θ	Variance of estimator		
			asympt.	empirical	estimate*
$\hat{\theta}$		279.868	11.808	12.265	11.830
$\tilde{\theta}$	0.3	279.898	11.331	11.669	11.340
	0.6	279.954	9.897	10.070	9.900
	0.9	280.026	7.509	7.586	7.501
Full population		280.013	6.500	6.725	6.495

*Average of estimates

Chapter 3

Cluster Sampling

3.1 Introduction

In *single stage cluster sampling*, the finite population is partitioned into n subpopulations, called clusters. From this population of clusters, a sample is selected and all elements in the selected clusters are surveyed. Due to the tendency for elements in the cluster to resemble each other and to control the costs, researches often perform a subsample within the selected clusters. In classical survey terminology, this is called *two-stage element sampling*. Two conditions are usually required. *Invariance* means that the subsampling must be independent of the sample of clusters, i.e. every time the ith cluster is included in a first-stage sample, the same subsampling scheme must be used. *Independence* means that the subsampling within each selected cluster is carried out independently of subsampling in any other cluster. An important aspect of cluster sampling is that the variability of estimators consists of two components; variability between and within the clusters. In order to estimate the latter one, at least two members from each cluster must be selected.

The household probability sampling technique used in Project ACCEPT is an example of such two-stage element sampling with households having the role of clusters. However, as it was already explained, we are interested in the model parameters. Therefore, the first-stage sample of households is not considered to be a subsample from a finite population of households, but rather to be generated by a model. We assume that a simple random sample of predetermined size is drawn within each household. This covers the specific case when only one member from each household is selected (like in Project ACCEPT). A graphical representation of such a sampling scheme can be found in Figure 3.1. As it was mentioned above, in the context of survey sampling, at least two members from each household would be required. We will see that the model-based inference does not have this limitation. Results presented in this chapter are consistent with similar results found in [8].

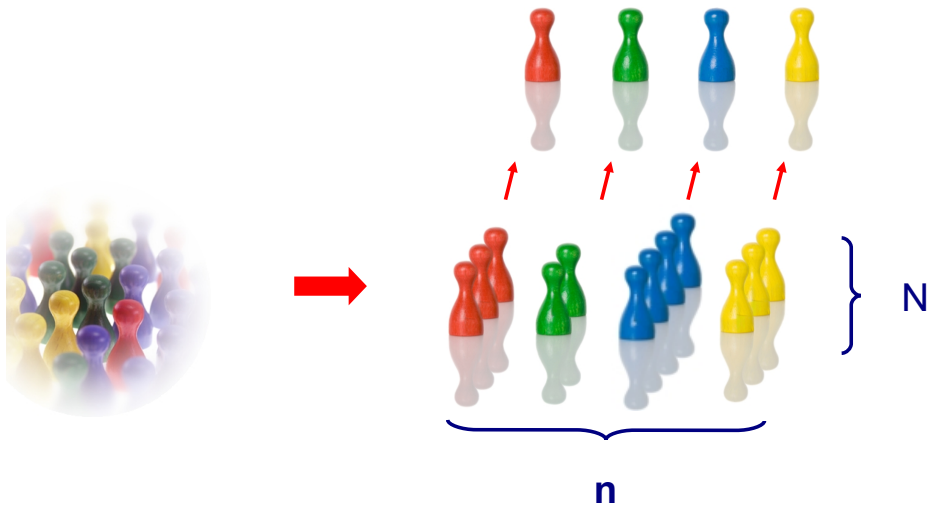


Figure 3.1: Household sampling

3.2 Estimation of the Mean

Let us assume a population of households. The total size of the population is n. It can be considered as a simple random sample from an infinite population of households. Let the random variable M_i represent the number of members in the ith household. Its density is denoted as $f(m)$ and the expectation as μ. The total number of individuals in the population is $N = \sum_{i=1}^{n} M_i$. Further, let Y_{ir} be the target variable for the rth member of the ith household, ξ_{ir} be a random variable for which

$$\xi_{ir} = \begin{cases} 1 & \text{if the } r\text{th member of the } i\text{th household is included in the sample} \\ 0 & \text{otherwise,} \end{cases}$$

m_i be the prespecified number of members selected from the ith household and $\pi_{ir} = \mathrm{E}\,(\xi_{ir}|M_i) = \frac{m_i}{M_i}$ be the probability that the rth member of the ith household is included in the sample, given M_i. The variable Y_{ir} is observed only for the sampled individuals, i.e. for $\xi_{ir} = 1$.

In the ith household, Y_{ir}, $r = 1, \ldots, M_i$, are iid random variables and their distribution depends on the size of the household M_i and a random parameter $\boldsymbol{b}_i$. The density of the random variable Y_{ir} is

$$f(y|m, \boldsymbol{b}).$$

The expectation of Y_{ir} in the ith household is denoted by

$$\theta_i = \int y f(y|m, \boldsymbol{b})\, dy.$$

The density of the variable Y_{ir} in any household of size m is

$$f(y|m) = \int f(y|m, \boldsymbol{b}) f(\boldsymbol{b}|m) d\boldsymbol{b},$$

where $f(\boldsymbol{b}|m)$ is the density of the parameter $\boldsymbol{b}_i$, given M_i. The population density of Y_{ir} is then

$$f(y) = \frac{\int m f(y|m) f(m) d\mu(m)}{\int m f(m)\, d\mu(m)} = \frac{1}{\mu} \int m \Big[\int f(y|m, \boldsymbol{b}) f(\boldsymbol{b}|m)\, d\boldsymbol{b} \Big] f(m)\, d\mu(m)$$
$$= \frac{1}{\mu} \iint m f(y|m, \boldsymbol{b}) f(m, \boldsymbol{b})\, d\boldsymbol{b} d\mu(m), \tag{3.1}$$

where $f(m, \boldsymbol{b})$ is the joint density of M_i and $\boldsymbol{b}_i$. We can write for the expectation of the variable Y_{ir}

$$\theta = \mathrm{E}\, Y_{ir} = \frac{1}{\mu} \iint m \Big[\int y f(y|m, \boldsymbol{b}) dy \Big] f(m, \boldsymbol{b})\, d\boldsymbol{b} d\mu(m) = \frac{1}{\mu} \mathrm{E}\, M_i \theta_i. \tag{3.2}$$

Similarly, the variance of the variable Y_{ir} is

$$\mathrm{var}\, Y_{ir} = \frac{1}{\mu} \iint m \Big[\int (y - \theta)^2 f(y|m, \boldsymbol{b}) dy \Big] f(m, \boldsymbol{b})\, d\boldsymbol{b} d\mu(m)$$
$$= \frac{1}{\mu} \mathrm{E}\, M_i \mathrm{E}_i (Y_{ir} - \theta)^2 = \frac{1}{\mu} \big[\mathrm{E}\, M_i \mathrm{var}_i Y_{ir} + \mathrm{E}\, M_i (\theta_i - \theta)^2 \big]. \tag{3.3}$$

This setup, including the form of the marginal density (3.1) of the random variable Y_{ir}, will be assumed in the rest of this thesis unless specified otherwise.

The estimator of the parameter θ is defined as

$$\hat{\theta} = \frac{\sum_{i=1}^n \sum_{r=1}^{M_i} \frac{\xi_{ir}}{\pi_{ir}} Y_{ir}}{\sum_{i=1}^n M_i} = \frac{\sum_{i=1}^n M_i \bar{Y}_i}{\sum_{i=1}^n M_i}, \text{ where } \bar{Y}_i = \frac{1}{m_i} \sum_{r=1}^{M_i} \xi_{ir} Y_{ir}.$$

Theorem 3. *Let $M_i, i = 1, 2, \ldots, n$, be iid random variables. Let $Y_{ir}, i = 1, 2, \ldots, n$ and $r = 1, 2, \ldots, M_i$, be random variables with marginal distribution given by (3.1) and $\mathrm{var}\, Y_{ir} < \infty$. Let $(Y_{i1}, Y_{i2} \ldots Y_{iM_i})$ and also $(\xi_{i1}, \xi_{i2} \ldots \xi_{iM_i})$ be independent random vectors for $i = 1, 2, \ldots n$. We assume $\sum_{r=1}^{M_i} \xi_{ir} = m_i$ and ξ_{ir} is independent from Y_{ir}, given M_i. Then*

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \Sigma_{\hat{\theta}}),$$

where

$$\Sigma_{\hat{\theta}} = \frac{1}{\mu^2} \mathrm{E}\, M_i^2 (\bar{Y}_i - \theta)^2. \tag{3.4}$$

Proof For $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n M_i$, by the Taylor expansion $\frac{1}{\hat{\mu}}$ around $\frac{1}{\mu}$, we get

$$\left(\frac{1}{\hat{\mu}} - \frac{1}{\mu} \right) = -\frac{1}{n\mu} \sum_{i=1}^n \left(\frac{M_i}{\mu} - 1 \right) + o_p\Big(\frac{1}{\sqrt{n}} \Big).$$

Then

$$\begin{aligned}\sqrt{n}(\hat{\theta}-\theta) &= \frac{1}{\sqrt{n}\hat{\mu}}\sum_{i=1}^{n} M_i\bar{Y}_i - \sqrt{n}\theta \\ &= \frac{1}{\sqrt{n}\mu}\sum_{i=1}^{n} M_i\bar{Y}_i - \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\Big[\frac{1}{n\mu}\sum_{i=1}^{n} M_i\bar{Y}_i\Big]\Big(\frac{M_j}{\mu}-1\Big) - \sqrt{n}\theta + o_p(1).\end{aligned}$$

Since

$$\frac{1}{\mu n}\sum_{i=1}^{n} M_i\bar{Y}_i \xrightarrow{P} \frac{1}{\mu}\mathrm{E}\, M_i\theta_i = \theta,$$

we have

$$\sqrt{n}(\hat{\theta}-\theta) = \frac{1}{\sqrt{n}\mu}\sum_{i=1}^{n} M_i\bar{Y}_i - \frac{1}{\sqrt{n}}\theta\sum_{i=1}^{n}\left(\frac{M_i}{\mu}-1\right) - \sqrt{n}\theta + o_p(1) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Q_i + o_p(1),$$

where $Q_i = \frac{1}{\mu}M_i(\bar{Y}_i-\theta)$ are iid random variables. According to the Central limit theorem $\sqrt{n}(\hat{\theta}-\theta) \xrightarrow{d} N(0,\Sigma_{\hat{\theta}})$, because

$$\mathrm{E}\, Q_i = \mathrm{E}\left(\mathrm{E}\left(Q_i|i\right)\right) = \frac{1}{\mu}\mathrm{E}\, M_i\theta_i - \theta = 0$$

and

$$\Sigma_{\hat{\theta}} = \mathrm{var}\, Q_i = \frac{1}{\mu^2}\mathrm{E}\, M_i^2(\bar{Y}_i-\theta)^2.$$

□

From a sample of n households, variance $\Sigma_{\hat{\theta}}$ can be estimated as

$$\hat{\Sigma}_{\hat{\theta}} = \frac{1}{\hat{\mu}^2}\frac{1}{n}\sum_{i=1}^{n} M_i^2(\bar{Y}_i-\hat{\theta})^2, \quad \text{where} \quad \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} M_i. \tag{3.5}$$

When only one member from each household is sampled (i.e. $m_i = 1$ for $i = 1, 2, \ldots, n$), the asymptotic variance of the normalized estimator $\hat{\theta}$ is $\Sigma_{\hat{\theta}} = \frac{1}{\mu^2}\mathrm{E}\, M_i^2(Y_i-\theta)^2$, where Y_i is the value for the sampled individual in the ith household. Based on (3.2), we can also write

$$\Sigma_{\hat{\theta}} = \frac{1}{\mu}\mathrm{E}\, M_i(Y_{ir}-\theta)^2. \tag{3.6}$$

In some situations, there is a direct relationship between the variance of a random variable and the variance of the estimator of its expectation. For example, for simple random sample of n iid random variables $X_i, i = 1, \ldots, n$, we have $\mathrm{var}(\widehat{\mathrm{E}X_i}) = \mathrm{var}(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n}\mathrm{var}X_i$. However, in this case

$$\Sigma_{\hat{\theta}} = \frac{1}{\mu^2}\mathrm{E}\, M_i^2(\bar{Y}_i-\theta)^2 = \frac{1}{\mu^2}\big[\mathrm{E}\,\frac{M_i^2}{m_i}\mathrm{var}_i Y_{ir} + \mathrm{E}\, M_i^2(\theta_i-\theta)^2\big], \tag{3.7}$$

from which we can see that the variance of the estimator $\hat{\theta}$ cannot be directly linked to the variance of Y_{ir} (3.3).

The knowledge of the variance of the estimator $\hat{\theta}$ allows us to address the following question: which design leads to the smaller variance of the estimator?

a) One member from each of n households is sampled.

b) All members from $\frac{n}{\mu}$ households are sampled.

The number of households selected in b) is determined so that the mean total number of selected individuals $\sum_{i \in s} M_i$ would be equal to the total number of selected individuals in situation a).

The variance of the estimator $\hat{\theta}$ is

a) $\mathrm{var}_a(\hat{\theta}) = \frac{1}{n}\frac{1}{\mu^2}\left[\mathrm{E}\, M_i^2 \mathrm{var}_i Y_{ir} + \mathrm{E}\, M_i^2(\theta_i - \theta)^2\right]$

b) $\mathrm{var}_b(\hat{\theta}) = \frac{1}{n}\frac{1}{\mu}\ \left[\mathrm{E}\, M_i\ \mathrm{var}_i Y_{ir} + \mathrm{E}\, M_i^2(\theta_i - \theta)^2\right]$.

When the number of members in all households is the same, then

$$\mathrm{var}_a(\hat{\theta}) \leq \mathrm{var}_b(\hat{\theta}).$$

In other situations, the answer is not clear. Let us assume that the variance within the households is equal, i.e. $\mathrm{var}_i Y_{ir} = \sigma^2$ for all i. Then in case there is no variability between the households, i.e. $\mathrm{E}\, M_i^2(\theta_i - \theta)^2 = 0$, we have $\sigma^2 \frac{1}{\mu}\mathrm{E}\, M_i^2 \geq \sigma^2 \mathrm{E}\, M_i$ and thus

$$\mathrm{var}_a(\hat{\theta}) \geq \mathrm{var}_b(\hat{\theta}).$$

Otherwise, if there is big variability between the households and small variability within households, $\mathrm{var}_a(\hat{\theta}) \leq \mathrm{var}_b(\hat{\theta})$ and vice versa.

In conclusion, design a) is more appropriate (in terms of precision given by the variance of the estimator) in situations where we expect high correlation within households and big differences between households. When households are similar and the correlation between observations within households is low, design b) leads to more efficient estimation.

3.3 Comparison with Bernoulli Sampling

If we sample only one member from each household, we avoid having correlated data in the resulting sample, in other words we obtain a sample which consists of independent values. In this situation a natural question arises - whether the variance of the estimator of the parameter θ obtained from cluster sampling is comparable with the variance of the estimator coming from a Bernoulli sample. The difference between household and Bernoulli sampling is displayed in Figure 3.2. While the idea to treat data from cluster sampling as if it came from Bernoulli sampling is often applied in practical situations, it has not been shown to be a correct approach. In this section we will address this question.

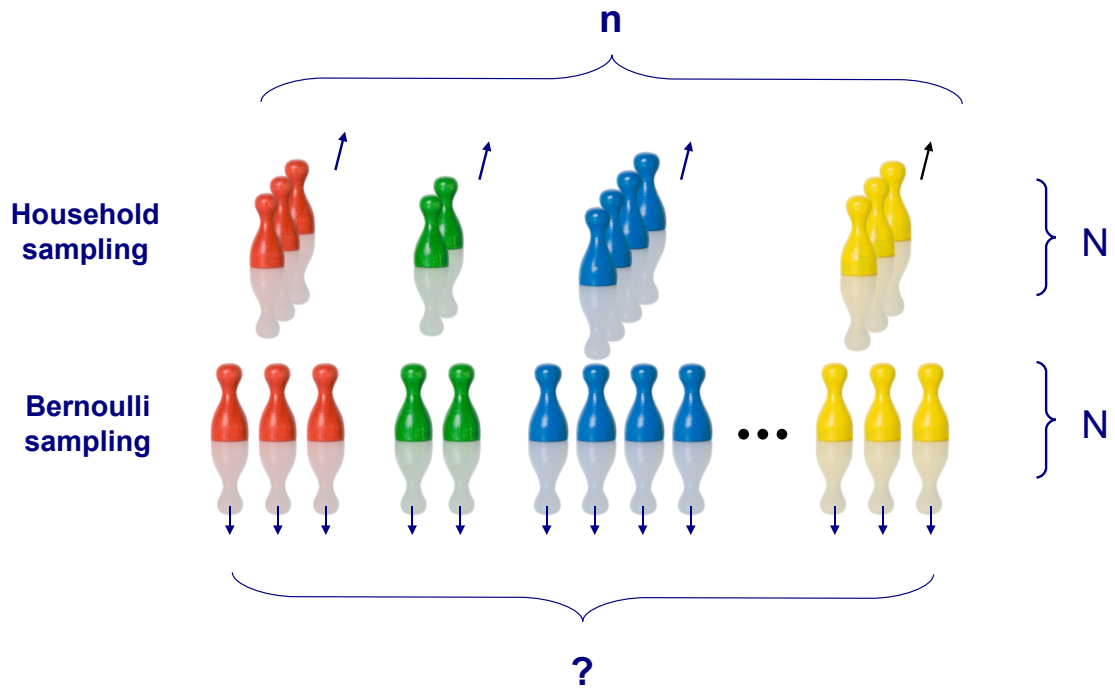


Figure 3.2: Household versus Bernoulli sampling

Let us assume that each of N individuals is sampled independently of the others with probability $1/size\ of\ household$. Consequently, the size of the sample is random with mean value n. The notation used in the previous chapters is more appropriate in this case. Let Y_j be the random variable of interest for the jth individual, ξ_j be its sampling indicator and $\pi_j = \mathrm{E}\,(\xi_j|M_j) = \frac{1}{M_j}$ be the sampling probability, where M_j is the size of the household, to which the jth individual belongs. Note that π_j is a random variable. The variable Y_j is observed only for sampled individuals, i.e. for $\xi_j = 1$.

The estimator of parameter θ is defined as

$$\tilde{\theta} = \frac{\sum_{j=1}^{N} \frac{\xi_j}{\pi_j} Y_j}{\sum_{j=1}^{N} \frac{\xi_j}{\pi_j}}.$$

Theorem 4. *Let $(Y_j, \xi_j, M_j), j = 1 \ldots N$, be iid random variables and ξ_j is independent of Y_j, given M_j. Assume that $\operatorname{var} Y_j < \infty$. Then*

$$\sqrt{N}(\tilde{\theta} - \theta) \xrightarrow{d} N(0, \Sigma_{\tilde{\theta}}),$$

where

$$\Sigma_{\tilde{\theta}} = \mathrm{E}\, M_i (Y_i - \theta)^2.$$

Proof For $\hat{N} = \sum_{i=1}^{N} \frac{\xi_i}{\pi_i}$, by the Taylor expansion $\frac{1}{\hat{N}}$ around $\frac{1}{N}$ we get

$$\sqrt{N}\left(\frac{1}{\hat{N}} - \frac{1}{N}\right) = -N^{-\frac{3}{2}}\left(\hat{N} - N\right) + o_p(1) = -N^{-\frac{3}{2}} \sum_{i=1}^{N}\left(\frac{\xi_i}{\pi_i} - 1\right) + o_p(1).$$

Then

$$\begin{aligned}
\sqrt{N}(\tilde{\theta} - \theta) &= \sqrt{N}\Big(\frac{1}{N}\sum_{i=1}^{N}\frac{\xi_i}{\pi_i}Y_i + \Big(\frac{1}{\hat{N}} - \frac{1}{N}\Big)\sum_{i=1}^{N}\frac{\xi_i}{\pi_i}Y_i - \theta\Big) \\
&= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{\xi_i}{\pi_i}Y_i - \theta\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\big(\frac{\xi_i}{\pi_i} - 1\big) - \sqrt{N}\theta + o_p(1) \\
&= \frac{1}{\sqrt{N}}\sum_{i=1}^{N} Q_i + o_p(1),
\end{aligned}$$

where $Q_i = \frac{\xi_i}{\pi_i}(Y_i - \theta)$ are iid random variables. We have $\mathrm{E}\, Q_i = 0$ and

$$\operatorname{var} Q_i = \mathrm{E}\, Q_i^2 = \mathrm{E}\,\frac{1}{\pi_i}(Y_i - \theta)^2.$$

According to the Central limit theorem $\sqrt{n}(\tilde{\theta} - \theta) \xrightarrow{d} N(0, \Sigma_{\tilde{\theta}})$.

□

While in the case of sampling from households the asymptotics is based on the increasing number of households, i.e. $n \to \infty$, for Bernoulli sampling an increasing number of individuals is needed, i.e. $N \to \infty$. For a comparable sample size the two estimators have the same asymptotic variance

$$\begin{aligned}
\operatorname{var}\tilde{\theta} &= \frac{1}{N}\mathrm{E}\, M_i(Y_i - \theta)^2 = \frac{1}{n\frac{1}{n}\sum_{j=1}^{n} M_j}\mathrm{E}\, M_i(Y_i - \theta)^2 \\
&\xrightarrow{P} \frac{1}{n\mu}\mathrm{E}\, M_i(Y_i - \theta)^2 = \operatorname{var}\hat{\theta} \qquad \text{(refer to 3.6)}.
\end{aligned}$$

Remark In the case of Bernoulli sampling (and thus in Theorem 4) we assume that the total number of individuals N, from which we sample, is fixed. By contrast, in the case of household sampling (and thus in Theorem 3) we assume that the fixed quantity is the number of households n and the total number of individuals $\sum_{i=1}^{n} M_i$ is a random variable. If we looked for a more precise analogy to the household sampling, we would have to consider the number of individuals N to be also random in the case of Bernoulli sampling. That would indeed lead to a higher variance of the estimator. Another imprecision is that Bernoulli sampling assumes independence of the sampled variables, which again in a strict analogy to the household sampling does not hold.

Nevertheless the result presented above has an important practical consequence. It implies that the statistical methods developed for Bernoulli sampling are also valid for data analysis based on cluster sampling with only one element drawn from each cluster.

3.4 Note on Estimators Presented in Survey Literature

Here again a connection between the survey literature and the results presented in Theorem 3 will be shown, referring to section 1.3.

Under cluster sampling, the "survey estimator" of the population mean is identical to our estimator $\hat{\theta}$. It estimates the population average, which can be written as

$$\frac{\sum_{i=1}^{N} M_i \bar{Y}_i}{\sum_{i=1}^{N} M_i}, \quad \text{where } \bar{Y}_i = \frac{1}{M_i} \sum_{r=1}^{M_i} Y_{ir}.$$

This population average estimates the expectation θ. We can calculate its asymptotic variance based on Theorem 3 applied to the case when all members from N households are observed. We get

$$\begin{aligned}\mathrm{var}_I(\mathrm{E}_{II}(\hat{\theta})) &= \mathrm{var}_I\Big(\frac{\sum_{i=1}^{N} M_i \bar{Y}_i}{\sum_{i=1}^{N} M_i}\Big) \cong \frac{1}{N}\frac{1}{\mu^2}\mathrm{E}\, M_i^2(\bar{Y}_i - \theta)^2 = \frac{1}{N}\frac{1}{\mu^2}\mathrm{E}\, M_i^2(\bar{Y}_i - \theta_i + \theta_i - \theta)^2 \\ &= \frac{1}{N}\frac{1}{\mu^2}\Big[\mathrm{E}\, M_i^2(\bar{Y}_i - \theta_i)^2 + \mathrm{E}\, M_i^2(\theta_i - \theta)^2\Big] = \frac{1}{N}\frac{1}{\mu^2}\Big[\mathrm{E}\, M_i\sigma_i^2 + \mathrm{E}\, M_i^2(\theta_i - \theta)^2\Big],\end{aligned}$$

where σ_i^2 is the variance of the variable Y_{ir} in the ith household.

The estimator of the design-based variance of $\hat{\theta}$ (see e.g. [16], pg. 315) is

$$\widehat{\mathrm{var}}_{II}(\hat{\theta}) = \frac{(1-f)s^2 + \frac{1}{N}\sum_{i=1}^{n}(1-f_i)M_i^2\frac{s_i^2}{m_i}}{n\big(\frac{1}{n}\sum_{i=1}^{n} M_i\big)^2},$$

where $f = \frac{n}{N}$, $f_i = \frac{m_i}{M_i}$,

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} M_i^2(\bar{y}_i - \hat{\theta})^2 \quad \text{and} \quad s_i^2 = \frac{1}{m_i - 1}\sum_{r=1}^{M_i}\xi_{ir}(y_{ir} - \bar{y}_i)^2.$$

Expression $\widehat{\mathrm{var}}_{II}(\hat{\theta})$ estimates

$$E_I(\mathrm{var}_{II}(\hat{\theta})) = \frac{1}{\mu^2}\frac{(1-\frac{n}{N})}{n}\Big[\mathrm{E}\,\frac{M_i^2}{m_i}\sigma_i^2 + \mathrm{E}\, M_i^2(\theta_i - \theta)^2\Big] + \frac{1}{N}\frac{1}{\mu^2}\mathrm{E}\,\Big(1 - \frac{m_i}{M_i}\Big)\frac{M_i^2}{m_i}\sigma_i^2.$$

We get

$$\begin{aligned}\mathrm{var}_I(\mathrm{E}_{II}(\hat{\theta})) + E_I(\mathrm{var}_{II}(\hat{\theta})) &= \frac{1}{\mu^2}\Big[\frac{1}{N}\mathrm{E}\, M_i\sigma_i^2 + \frac{1}{N}\mathrm{E}\, M_i^2(\theta_i - \theta)^2 \\ &\quad + \frac{1}{n}\mathrm{E}\,\frac{M_i^2}{m_i}\sigma_i^2 - \frac{1}{N}\mathrm{E}\, M_i\sigma_i^2 + \Big(\frac{1}{n} - \frac{1}{N}\Big)\mathrm{E}\, M_i^2(\theta_i - \theta)^2\Big] \\ &= \frac{1}{n}\frac{1}{\mu^2}\Big(\mathrm{E}\,\frac{M_i^2}{m_i}\sigma_i^2 + \mathrm{E}\, M_i^2(\theta_i - \theta)^2\Big),\end{aligned}$$

which is the asymptotic model-based variance of $\hat{\theta}$.

Chapter 4

Stratified Cluster Sampling

4.1 Introduction

In *stratified cluster sampling*, a population of households is divided in K strata which comprise the whole of the population. A sample of households is drawn independently from each stratum and then one or more members are sampled independently from each selected household. In this section, we will describe the estimation of a parameter in the case of stratified cluster sampling, and examine whether and how the idea of improving the estimation by auxiliary information can be extended to this situation. As in the previous chapter, we will assume that the size of the household is correlated with the variable of interest and therefore it must be taken into account.

The notation will stay similar. Let Y_{ir} be the random variable of interest for the rth member from the ith household of size M_i. Let W_i be a discrete random variable taking on values from $\{1, 2, \ldots, K\}$, corresponding to a stratum. We denote

$$\mathrm{I}_{ik} = \begin{cases} 1 & \text{if } W_i = k \\ 0 & \text{otherwise.} \end{cases}$$

Let ξ_i and ζ_{ir} be the sampling indicators, i.e.

$$\xi_i = \begin{cases} 1 & \text{if the } i\text{th household is included in the sample} \\ 0 & \text{otherwise,} \end{cases}$$

$$\zeta_{ir} = \begin{cases} 1 & \text{if the } r\text{th member of the } i\text{th household is included in the sample} \\ 0 & \text{otherwise.} \end{cases}$$

Let $\pi_k = \mathrm{E}\,(\xi_i | W_i = k)$ be the sampling probability of each household from stratum k, m_i is the prespecified number of members to be selected from the ith household and $\pi_{ir} = \frac{m_i}{M_i}$ is the probability that the rth member of the ith selected household is included in the sample, given M_i.

As in (3.2), the expectation of the variable Y_{ir} is

$$\theta = \mathrm{E}\, Y_{ir} = \frac{1}{\mu} \mathrm{E}\, M_i \theta_i,$$

where θ_i is the expectation of Y_{ir} in the ith household. We denote a stratum-specific mean value by

$$\theta_{[k]} = \mathrm{E}\,(Y_{ir}|W_i = k) = \frac{1}{\mu_k}\mathrm{E}\,(M_i\theta_i|W_i = k), \text{ where } \mu_k = \mathrm{E}\,(M_i|W_i = k).$$

However, it should be mentioned that the relationship (2.1) does not hold in this case. Each stratum can have a different distribution of household size, which determines the contribution of stratum-specific mean values to the marginal mean value. The following holds

$$\begin{aligned}\theta = \mathrm{E}\,Y_{ir} &= \frac{1}{\mu}\mathrm{E}\,M_i\theta_i = \frac{1}{\mu}\sum_{k=1}^{K}\mathrm{E}\,(M_i\theta_i|W_i = k)\mathrm{P}(W_i = k)\\ &= \sum_{k=1}^{K}\frac{\mu_k}{\mu}\frac{\mathrm{E}\,(M_i\theta_i|W_i = k)}{\mu_k}\mathrm{P}(W_i = k) = \sum_{k=1}^{K}\frac{\mu_k}{\mu}\theta_{[k]}\mathrm{P}(W_i = k).\end{aligned}$$

The estimator of the parameter θ is defined as

$$\hat{\theta} = \frac{\sum_{i=1}^{n}\sum_{k=1}^{K}\big(\frac{1}{\hat{\pi}_k}\xi_i\sum_{r=1}^{M_i}\frac{\zeta_{ir}}{\pi_{ir}}Y_{ir}\big)I_{ik}}{\sum_{i=1}^{n}\sum_{k=1}^{K}\big(\frac{1}{\hat{\pi}_k}\xi_i M_i\big)I_{ik}} = \frac{1}{\hat{\mu}}\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\big(\frac{1}{\hat{\pi}_k}\xi_i M_i\bar{Y}_i\big)I_{ik},$$

where

$$\bar{Y}_i = \frac{1}{m_i}\sum_{r=1}^{M_i}\zeta_{ir}Y_{ir}, \quad \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\big(\frac{1}{\hat{\pi}_k}\xi_i M_i\big)I_{ik}, \quad \hat{\pi}_k = \frac{1}{n_k}\sum_{i=1}^{n}I_{ik}$$

and n_k is the number of households belonging to stratum k.

Theorem 5. *Let $(M_i, W_i, \xi_i), i = 1, 2, \ldots, n$, be iid random vectors. Let $Y_{ir}, i = 1, 2, \ldots, n$ and $r = 1, 2, \ldots, M_i$, be random variables with marginal distribution given by density (3.1), with $\mathrm{var}\,Y_{ir} < \infty$, and form independent random vectors $(Y_{i1}, Y_{i2}\ldots Y_{iM_i})$. Assume ξ_i is independent of M_i and $Y_{ir}, r = 1, 2, \ldots, M_i$, given W_i. Assume also that $(\zeta_{i1}, \zeta_{i2}\ldots\zeta_{iM_i})$ are independent random vectors, $\sum_{r=1}^{M_i}\zeta_{ir} = m_i$, and ζ_{ir} is independent from Y_{ir}, given M_i. Then*

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \Sigma_{\hat{\theta}}),$$

where

$$\Sigma_{\hat{\theta}} = \frac{1}{\mu^2}\Big(\mathrm{E}\,M_i^2(\bar{Y}_i - \theta)^2 + \sum_{k=1}^{K}\mathrm{P}(W_i = k)\frac{1-\pi_k}{\pi_k}\mathrm{var}_k[M_i(\bar{Y}_i - \theta)]\Big) \tag{4.1}$$

and

$$\mathrm{var}_k[M_i(\bar{Y}_i - \theta)] = \mathrm{var}[M_i(\bar{Y}_i - \theta)|W_i = k].$$

Proof It holds

$$\sqrt{n}(\hat{\theta}-\theta)=\frac{1}{\sqrt{n}}\Big[\frac{1}{\mu}\sum_{i=1}^{n}\sum_{k=1}^{K}(\frac{1}{\hat{\pi}_k}\xi_i M_i\bar{Y}_i)I_{ik}+\Big(\frac{1}{\hat{\mu}}-\frac{1}{\mu}\Big)\sum_{i=1}^{n}\sum_{k=1}^{K}(\frac{1}{\hat{\pi}_k}\xi_i M_i\bar{Y}_i)I_{ik}-\theta\Big]. \quad (4.2)$$

Similarly as in (2.6), we can write for the first expression in (4.2)

$$\frac{1}{\sqrt{n}\mu}\sum_{i=1}^{n}\sum_{k=1}^{K}(\frac{1}{\hat{\pi}_k}\xi_i M_i\bar{Y}_i)I_{ik}=\frac{1}{\sqrt{n}\mu}\sum_{i=1}^{n}\sum_{k=1}^{K}\Big(\frac{\xi_i M_i\bar{Y}_i}{\pi_k}-\frac{\xi_i-\pi_k}{\pi_k}\mathrm{E}_k(M_i\bar{Y}_i)\Big)I_{ik}+o_p(1). \quad (4.3)$$

By the Taylor expansion of $\frac{1}{\hat{\mu}}$ around $\frac{1}{\mu}$, we get

$$\begin{aligned}\frac{1}{\sqrt{n}}\Big(\frac{1}{\hat{\mu}}-\frac{1}{\mu}\Big)&=-\frac{1}{\sqrt{n}\mu^2}\Big(\sum_{i=1}^{n}\sum_{k=1}^{K}(\frac{1}{\hat{\pi}_k}\xi_i M_i)I_{ik}-\mu\Big)+o_p(1)\\&\overset{(2.6)}{=}-\frac{1}{\sqrt{n}\mu^2}\sum_{i=1}^{n}\sum_{k=1}^{K}\Big(\frac{\xi_i M_i}{\pi_k}-\frac{\xi_i-\pi_k}{\pi_k}\mu_k-\mu\Big)I_{ik}+o_p(1).\end{aligned}$$

Also,

$$\frac{1}{n\mu}\sum_{i=1}^{n}\sum_{k=1}^{K}(\frac{1}{\hat{\pi}_k}\xi_i M_i\bar{Y}_i)I_{ik}\xrightarrow{P}\theta.$$

Thus the second term in (4.2) is

$$\frac{1}{\sqrt{n}}\Big(\frac{1}{\hat{\mu}}-\frac{1}{\mu}\Big)\sum_{i=1}^{n}\sum_{k=1}^{K}(\frac{1}{\hat{\pi}_k}\xi_i M_i\bar{Y}_i)I_{ik}=-\theta\frac{1}{\sqrt{n}\mu}\sum_{i=1}^{n}\sum_{k=1}^{K}\Big(\frac{\xi_i M_i}{\pi_k}-\frac{\xi_i-\pi_k}{\pi_k}\mu_k-\mu\Big)I_{ik}+o_p(1). \quad (4.4)$$

By (4.3) and (4.4), we get

$$\sqrt{n}(\hat{\theta}-\theta)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}Q_i+o_p(1),$$

where

$$\begin{aligned}Q_i&=\frac{1}{\mu}\sum_{k=1}^{K}\Big(\frac{\xi_i M_i\bar{Y}_i}{\pi_k}-\frac{\xi_i-\pi_k}{\pi_k}\mathrm{E}_k(M_i\bar{Y}_i)-\frac{\xi_i M_i}{\pi_k}\theta+\frac{\xi_i-\pi_k}{\pi_k}\mu_k\theta\Big)I_{ik}\\&=\frac{1}{\mu}\sum_{k=1}^{K}\Big(\frac{\xi_i}{\pi_k}M_i(\bar{Y}_i-\theta)-\frac{\xi_i-\pi_k}{\pi_k}\mathrm{E}_k M_i(\bar{Y}_i-\theta)\Big)I_{ik}\end{aligned}$$

are iid variables and

$$E(E[Q_i|W_i])=\frac{1}{\mu}\mathrm{E}\left(\mathrm{E}_k M_i\theta_i-\mu_k\theta\right)=0$$

$$\begin{aligned}\mathrm{var}Q_i = \mathrm{E}\left(\mathrm{E}\left[Q_i^2|W_i\right]\right) &= \frac{1}{\mu^2}\mathrm{E}\Big(\frac{1}{\pi_k}\mathrm{E}_k M_i^2(\bar{Y}_i-\theta)^2 + \frac{1-\pi_k}{\pi_k}(\mathrm{E}_k M_i(\bar{Y}_i-\theta))^2 \\ &\quad - 2\frac{1-\pi_k}{\pi_k}(\mathrm{E}_k M_i(\bar{Y}_i-\theta))^2\Big) \\ &= \frac{1}{\mu^2}\mathrm{E}\Big(\mathrm{E}_k M_i^2(\bar{Y}_i-\theta)^2 + \frac{1-\pi_k}{\pi_k}M_i^2(\bar{Y}_i-\theta)^2 - \frac{1-\pi_k}{\pi_k}(\mathrm{E}_k M_i(\bar{Y}_i-\theta))^2\Big) = \Sigma_{\hat{\theta}}.\end{aligned}$$

According to the Central limit theorem for iid random variables,

$$\sqrt{n}(\hat{\theta}-\theta) \xrightarrow{d} N(0,\Sigma_{\hat{\theta}}).$$

□

Variance $\Sigma_{\hat{\theta}}$ can be estimated as

$$\hat{\Sigma}_{\hat{\theta}} = \frac{1}{\hat{\mu}^2}\Big(\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\frac{\xi_i}{\hat{\pi}_k}M_i^2(\bar{Y}_i-\hat{\theta})^2 I_{ik} + \sum_{k=1}^{K}\frac{n_k}{n}\frac{1-\hat{\pi}_k}{\hat{\pi}_k}\widehat{\mathrm{var}}_k\big[M_i(\bar{Y}_i-\hat{\theta})\big]\Big), \tag{4.5}$$

where $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\frac{\xi_i}{\hat{\pi}_k}M_i I_{ik}$ and $\widehat{\mathrm{var}}_k\big[M_i(\bar{Y}_i-\hat{\theta})\big]$ is the usual estimate of variance of the random variable $M_i(\bar{Y}_i-\hat{\theta})$ within stratum k.

4.2 Auxiliary Variables in Stratified Cluster Sampling

Suppose there is auxiliary information available for all the households in the population. For example, it could be some basic characteristics of the household available in the registry, such as size or location. Can we use this information to improve the precision of an estimator? Is an analogy of the procedure described in the previous chapter for the stratified sampling applicable in the case of stratified cluster sampling?

The notation will stay the same as in section 2.4. Let us assume an s-dimensional vector of auxiliary variables $\boldsymbol{X}_i$ related to the household i and denote

$$\boldsymbol{Z}_i^T = (\mathrm{I}_{i1},\ldots,\mathrm{I}_{iK},\boldsymbol{X}_i^T).$$

We adjust the sampling weights in the Horvitz-Thompson estimator by fitting the logistic regression model for the sampling probabilities as defined by (2.8). The estimated sampling probabilities are denoted $\tilde{\pi}_i$, see (2.9).

The estimator of θ has the following form:

$$\tilde{\theta} = \frac{1}{\tilde{\mu}}\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\tilde{\pi}_i}\xi_i M_i\bar{Y}_i, \quad \text{where } \tilde{\mu} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\tilde{\pi}_i}\xi_i M_i.$$

Theorem 6. *Let* $(M_i, W_i, \boldsymbol{X_i}, \xi_i), i = 1, 2, \ldots, n$, *be iid random vectors,* $\operatorname{var} X_{ij} < \infty$ *for each component* $j = 1, 2, \ldots, s$. *Let* $Y_{ir}, i = 1, 2, \ldots, n$ *and* $r = 1, 2, \ldots, M_i$, *be random variables with marginal distribution given by density (3.1), with* $\operatorname{var} Y_{ir} < \infty$, *and form independent random vectors* $(Y_{i1}, Y_{i2} \ldots Y_{iM_i})$. *Assume* ξ_i *is independent of* M_i *and* $Y_{ir}, r = 1, 2, \ldots, M_i$, *given* W_i. *Assume also that* $(\zeta_{i1}, \zeta_{i2} \ldots \zeta_{iM_i})$ *are independent random vectors,* $\sum_{r=1}^{M_i} \zeta_{ir} = m_i$ *and* ζ_{ir} *is independent from* Y_{ir}, *given* M_i. *Then*

$$\sqrt{n}(\tilde{\theta} - \theta) \xrightarrow{d} N(0, \Sigma_z),$$

where

$$\Sigma_z = \Sigma_{\hat{\theta}} - \boldsymbol{c}^T V^{-1} \boldsymbol{c} \tag{4.6}$$

and

$$\boldsymbol{c} = \sum_{k=1}^{K} p_k (1 - \pi_k) \operatorname{cov}_k(\boldsymbol{X}_i, M_i(\bar{Y}_i - \theta)), \quad V = \sum_{k=1}^{K} p_k \pi_k (1 - \pi_k) \operatorname{var}_k \boldsymbol{X}_i.$$

Proof We have

$$\sqrt{n}(\tilde{\theta} - \theta) = \frac{1}{\sqrt{n}} \Big(\frac{1}{\mu} \sum_{i=1}^{n} \frac{1}{\tilde{\pi}_k} \xi_i M_i \bar{Y}_i + \Big(\frac{1}{\tilde{\mu}} - \frac{1}{\mu} \Big) \sum_{i=1}^{n} \frac{1}{\tilde{\pi}_k} \xi_i M_i \bar{Y}_i - \theta \Big). \tag{4.7}$$

By the same argument as in the proof of Theorem 2, we can write for the first expression in (4.7)

$$\frac{1}{\sqrt{n}\mu} \sum_{i=1}^{n} \frac{1}{\tilde{\pi}_k} \xi_i M_i \bar{Y}_i = \frac{1}{\sqrt{n}\mu} \sum_{i=1}^{n} \Big(\frac{\xi_i M_i \bar{Y}_i}{\pi_i} + (\xi_i - \pi_i) \boldsymbol{Z}_i^T J_\gamma^{-1} \mathrm{E}\,(1 - \pi_i) \boldsymbol{Z}_i M_i \bar{Y}_i \Big) + o_p(1) \tag{4.8}$$

By the Taylor expansion of $\frac{1}{\tilde{\mu}}$ around $\frac{1}{\mu}$ and by the same argument as in the proof of Theorem 2, we get

$$\begin{aligned} \Big(\frac{1}{\tilde{\mu}} - \frac{1}{\mu} \Big) &= -\frac{1}{\mu^2} \Big(\frac{1}{n} \sum_{i=1}^{n} \frac{1}{\tilde{\pi}_i} \xi_i M_i - \mu \Big) + o_p\Big(\frac{1}{\sqrt{n}} \Big) \\ &= -\frac{1}{\mu^2} \Big(\frac{1}{n} \sum_{i=1}^{n} \big(\frac{\xi_i M_i}{\pi_i} + (\xi_i - \pi_i) \boldsymbol{Z}_i^T J_\gamma^{-1} \mathrm{E}\,(1 - \pi_i) \boldsymbol{Z}_i M_i \big) - \mu \Big) + o_p\Big(\frac{1}{\sqrt{n}} \Big) \end{aligned} \tag{4.9}$$

By analogy to Theorem 5, from (4.8) and (4.9) we get

$$\sqrt{n}(\tilde{\theta} - \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Q_i + o_p(1),$$

where

$$\begin{aligned} Q_i = \frac{1}{\mu} \Big(& \frac{\xi_i M_i \bar{Y}_i}{\pi_i} - (\xi_i - \pi_i) \boldsymbol{Z}_i^T J_\gamma^{-1} \mathrm{E}\,(1 - \pi_i) \boldsymbol{Z}_i M_i \bar{Y}_i \\ & - \theta \frac{\xi_i M_i}{\pi_i} + \theta (\xi_i - \pi_i) \boldsymbol{Z}_i^T J_\gamma^{-1} \mathrm{E}\,(1 - \pi_i) \boldsymbol{Z}_i M_i \Big) \\ = \frac{1}{\mu} \Big(& \frac{\xi_i}{\pi_i} M_i (\bar{Y}_i - \theta) - (\xi_i - \pi_i) \boldsymbol{Z}_i^T J_\gamma^{-1} \mathrm{E}\,(1 - \pi_i) \boldsymbol{Z}_i M_i (\bar{Y}_i - \theta) \Big) \end{aligned}$$

are iid variables and

$$\mathrm{E}\, Q_i = 0 \quad \text{and} \quad \mathrm{var} Q_i = \Sigma_z.$$

According to the Central limit theorem for iid random variables, $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \Sigma_z)$.

□

The variance Σ_z (4.6) can be estimated analogously to (4.5) with the usual estimates of variance and covariance within the strata.

This theorem says that, as in the case of stratified sampling (Theorem 2), the use of auxiliary variables to adjust the weights in the Horvitz-Thompson estimator can never increase the asymptotic variance of the estimator. We can also see that the optimal trasformation of the auxiliary variable is $\boldsymbol{X}_i/\pi_k$, see section 2.4. However, in the situation of stratified sampling, the extent to which the precision of the estimator could have been improved depended on the correlation between the auxiliary vector $\boldsymbol{X}_i$ and the target variable Y_i. Here, the improvement of the estimator depends on the correlation between $\boldsymbol{X}_i$ and $M_i(\bar{Y}_i - \theta)$. Apparently, it makes the application in a real situation more complex: there are several questions to consider.

One of the obvious candidates for an auxiliary variable is the size of the households. In some situations, this would improve the variance. Nevertheless, under certain circumstances the correlation between M_i and $M_i(\bar{Y}_i - \theta)$ might not be very strong. For example, let us assume that the size of the household has a uniform distribution, i.e. small and big households are represented equally in the population. Consequently, the proportion of big household members in the population is greater and they contribute more to the population expectation θ. Thus the expression $(\bar{Y}_i - \theta)$ is in general smaller for big households than for small households. This implies that the expression $M_i(\bar{Y}_i - \theta)$ could be similar for big and small households and therefore not strongly correlated with the size of the household.

Suppose (rather unlikely situation) that there is another auxiliary variable D_{ir} available for all the members in all households. What transformation of these auxiliary variables would be most correlated with $M_i(\bar{Y}_i - \theta)$? Would it be the sum of the values observed for all members of the household, i.e. $\sum_{r=1}^{M_i} D_{ir}$, or only for the sampled ones, i.e. $\sum_{r=1}^{M_i} \zeta_{ir} D_{ir}$? There is no clear answer to this question. Depending on the nature of the auxiliary information, one or the other should be preferred.

However, the scenario described in the previous paragraph is not very likely to occur. More typically, the auxiliary information would be available only for members of the sampled households. In such case we recommend a different approach described in the next chapter.

4.3 Example

Let us assume that there are two strata in a population of households, A and B. 60 % of households belong to stratum A. Households in stratum A have either 2 members (60 %)

or 3 members (40 %). In stratum B, there are households of size 7 (60 %) or 8 (40 %). The mean value of the variable of interest Y is denoted as θ_m for a household of size m:

$$\theta_2 = 100 \quad \theta_3 = 200 \quad \theta_7 = 300 \quad \theta_8 = 400.$$

The mean value of Y is then

$$\theta = \frac{2 * 100 * 0.6 * 0.6 + 3 * 200 * 0.4 * 0.6 + 7 * 300 * 0.6 * 0.4 + 8 * 400 * 0.4 * 0.4}{2 * 0.6 * 0.6 + 3 * 0.4 * 0.6 + 7 * 0.6 * 0.4 + 8 * 0.4 * 0.4} = 280.$$

A value of the target variable for the rth member in the ith household of size m is

$$Y_{ir} = \theta_m + \delta_i + \epsilon_{ir}, \tag{4.10}$$

where $\delta_i \sim N(0, 1600)$ and $\epsilon_{ir} \sim N(0, 1600)$ are iid random variables, mutually independent of each other. The random variable δ_i represents a household specific component (and thus reflects the fact that observations within one household are not independent) and ϵ_{ir} is a subject specific component. A household is included in a sample with probability $\pi_A = 0.8$ (for stratum A) or $\pi_B = 0.6$ (for stratum B). From each selected household, one member is included in a sample.

Simulation To illustrate the results from previous sections, we performed a small simulation. We considered the following situations:

- No further auxiliary information is available, parameter θ is estimated by the estimator $\hat{\theta}$.
- The size of each household in a population is known and thus it can be used as auxiliary information to improve the estimation, estimator $\tilde{\theta}$ is used.
- Another auxiliary variable is available for each household in a population, it is used to improve the estimation, estimator $\tilde{\theta}$ is used.

The simulation procedure was conducted as follows. First, we generated a population of 1000 households. The size and the stratum was assigned to each household with the above defined frequency. The values of Y_{ir} were generated for each member according to (4.10). In the second step, a sample of households was drawn and then one member from each household was sampled at random. Based on this sample, the parameter θ was estimated by the estimator $\hat{\theta}$, not taking into account any auxiliary variable, as well as by the estimator $\tilde{\theta}$, using the size of the households as an auxiliary variable. The size of the households was used in two different ways; either as a continuous variable, i.e. using its actual value, or as a categorical variable.

In section 4.2 we showed that the improvement of the estimator depends on the correlation between the auxiliary variable and $M_i(\bar{Y}_i - \theta)$. In order to demonstrate this observation, the second auxiliary variable was defined as

$$X_i = M_i(\bar{Y}_i - \theta) + \eta_i,$$

Table 4.1: Example of stratified cluster sampling; Results of simulation

Estimator	Auxiliary variable	Average estimate of θ	Variance of estimator		Coverage of CI (%)
			empirical	estimate*	
$\hat{\theta}$	No	280.02	20.687	20.050	94.5
$\tilde{\theta}$	Size of household	280.04	18.693	18.366	94.7
$\tilde{\theta}$	Size of household (as categorical variable)	280.01	17.735	17.775	94.9
$\tilde{\theta}$	X_{ir}	280.02	17.473	18.173	95.4
All households included in a sample		279.94	15.729		

*Average of estimates

where $\eta_i \sim N(0, 13000)$ for households from stratum A and $\eta_i \sim N(0, 190000)$ for households from stratum B. Let us remark that $\bar{Y}_i$ is the average of the values for the selected individuals from the ith household. When only one member is sampled, $\bar{Y}_i$ corresponds to the value of the selected member. In the definition of X_i for the households not included in the sample, one member was chosen randomly and used for the derivation of X_i. The variance of the random variable η_i was chosen so that the correlation between X_i and $M_i(\bar{Y}_i - \theta)$ within each stratum would be approximately 0.8.

This procedure was repeated 1000 times. The average estimate of the parameter was calculated, as well as the average of the estimate of the asymptotic variance. This latter quantity was compared to the empirical variance of the 1000 obtained estimates. Using the estimated variance, the 95% confidence intervals were derived. The percentage of confidence intervals which covered the true value θ was obtained in order to assess whether their coverage was close to the desired 95 %. The results are displayed in Table 4.1.

We can see that all the estimators provided unbiased estimates of the parameter θ. The average estimates of the asymptotic variance were very close to the empirical variance of the estimators and the coverage of the confidence intervals was aproximately 95 % in all cases. The empirical variance of the estimator which did not take into account any auxiliary variable was equal to 20.687. When all the households were included in the sample and one member was selected from each, the empirical variance of the estimator of θ was 15.729. This is in fact a limit to which the variance of the estimator could be possibly decreased by taking into account auxiliary information, when not all the households are observed. The size of the household (as a categorical variable) as well as the variable X_{ir} considered as the auxiliary variable helped to decrease the empirical variance of the estimator to 17.735 and 17.473, respectively, which is approximately 60 % of the possible gain.

Chapter 5

Use of Auxiliary Variables in Cluster Sampling

5.1 Use of Auxiliary Variables

Chapter 3 was devoted to two-stage element sampling. In this sampling scheme, first a sample of clusters is obtained and then one or more elements from each cluster are selected. The analysis is performed based on the subsample of elements. However, in some situations, auxiliary information may be available for all the elements in the selected clusters. This is also the case in Project ACCEPT. Detailed assessments are performed only on the selected members of households, but basic information (such as age and gender) is collected for each member of a household. In this chapter we will explore whether additional data about non-selected members of the households can improve the precision of the estimation as was the case with stratified sampling. For simplicity we will assume that only one member is selected from each household and that auxiliary information is available for all the other members of the household. Some households in the sample may have only one member. Obviously, such households do not contribute to the improvement of variance and thus we assume that the number of households of size bigger than one is non negligible.

We keep the same notation as in the previous chapters. Furthermore, $\mathrm{E}_i X_{ir}$ and $\mathrm{var}_i X_{ir}$ denotes expectation and variance, respectively, of variable X_{ir} in the ith household and $\mathrm{cov}_i(X_{ir}, X_{is})$ denotes covariance of variables X_{ir} and X_{is} in the ith household.

Now, let us denote $\boldsymbol{X}_{ir}$ the vector of auxiliary variables for the rth member of the ith household. We will assume that $(\boldsymbol{X}_{i1}, \boldsymbol{X}_{i2} ... \boldsymbol{X}_{iM_i})$ are independent random matrices for $i = 1, 2, \ldots, n$, where each component has a finite variance. We set

$$\boldsymbol{Z}_{ir}^T = (\mathbf{I}_{ir}^T, \boldsymbol{X}_{ir}^T),$$

where $\mathbf{I}_{ir}$ is a vector of dummy variables for the size of household represented in the model as a factor variable (and thus the same for all individuals within one household). Following a similar approach as described previously for stratified sampling, the logistic regression

for the sampling probabilities has the form

$$\log\left(\frac{\pi_{ir}}{1-\pi_{ir}}\right) = \boldsymbol{\gamma}^T \boldsymbol{z}_{ir}. \tag{5.1}$$

The sampling probabilities depend only on the size of the given household and not on the auxiliary variables, and that is why the components reflecting the size will be the only nonzero components of $\boldsymbol{\gamma}$. The part $\boldsymbol{\gamma}_d$ of the vector of parameters pertaining to the auxiliary variables is $\mathbf{0}$.

The sampling probabilities within one household are correlated, because when one member of the household is included in the sample, the others are certainly not. However, as we are not interested in the variance of the estimator of the parameter $\boldsymbol{\gamma}$, we can still fit the model with estimating equations

$$\sum_{i=1}^{n}\sum_{r=1}^{M_i}(\xi_{ir} - \pi_{ir})\boldsymbol{z_{ir}} = \mathbf{0}.$$

The estimator of the parameter θ has the form

$$\tilde{\theta} = \frac{\sum_{i=1}^{n}\sum_{r=1}^{M_i}\frac{\xi_{ir}}{\tilde{\pi}_{ir}}Y_{ir}}{\sum_{i=1}^{n} M_i},$$

where $\tilde{\pi}_{ir}$ are the sampling probabilities predicted by the model (5.1).

Theorem 7. *Let $M_i, i = 1, 2, \ldots, n$, be iid random variables. Let $Y_{ir}, i = 1, 2, \ldots, n$ and $r = 1, 2, \ldots, M_i$, be random variables with marginal distribution given by (3.1) and $\operatorname{var} Y_{ir} < \infty$. Let $(Y_{i1}, Y_{i2} \ldots Y_{iM_i})$ and also $(\xi_{i1}, \xi_{i2} \ldots \xi_{iM_i})$ be independent random vectors for $i = 1, 2, \ldots n$. Let $\sum_{r=1}^{M_i}\xi_{ir} = 1$ and ξ_{ir} is independent from Y_{ir}, given M_i. Then*

$$\sqrt{n}(\tilde{\theta} - \theta) \xrightarrow{d} N(0, \Sigma_{\tilde{\theta}}),$$

where

$$\Sigma_{\tilde{\theta}} = \Sigma_{\hat{\theta}} + \boldsymbol{q}^T J_\gamma^{-1}(J_\gamma - L_\gamma)J_\gamma^{-1}\boldsymbol{q} - 2(\boldsymbol{q}^T - \boldsymbol{t}^T)J_\gamma^{-1}\boldsymbol{q}, \tag{5.2}$$

$\Sigma_{\hat{\theta}}$ is the variance obtained without taking into account the auxiliary variables (see 3.4)

$$\Sigma_{\hat{\theta}} = \frac{1}{\mu^2}\mathrm{E}\, M_i^2\Big(\sum_{r=1}^{M_i}\xi_{ir}Y_{ir} - \theta\Big)^2,$$

and

$$\boldsymbol{q} = \frac{1}{\mu}\mathrm{E}\left[(M_i - 1)\mathrm{E}_i(Y_{ir}\boldsymbol{Z}_{ir})\right] \qquad \boldsymbol{t} = \frac{1}{\mu}\mathrm{E}\left[(M_i - 1)\mathrm{E}_i(Y_{ir}\boldsymbol{Z}_{is})\right]$$

$$J_\gamma = \mathrm{E}\left[\Big(1 - \frac{1}{M_i}\Big)\mathrm{E}_i(\boldsymbol{Z}_{ir}\boldsymbol{Z}_{ir}^T)\right] \qquad L_\gamma = \mathrm{E}\left[\Big(1 - \frac{1}{M_i}\Big)\mathrm{E}_i(\boldsymbol{Z}_{ir}\boldsymbol{Z}_{is}^T)\right].$$

Proof If we denote $\boldsymbol{R}_i = \sum_{r=1}^{M_i}(\xi_{ir} - \pi_{ir})\boldsymbol{Z}_{ir}$, we can write

$$\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma} = J_\gamma^{-1}\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{R}_i + o_p\Big(\frac{1}{\sqrt{n}}\Big),$$

where $\tilde{\boldsymbol{\gamma}}$ is the estimate of the true l-dimensional parameter $\boldsymbol{\gamma}$ and

$$J_\gamma = -\mathrm{E}\,\frac{\partial}{\partial\boldsymbol{\gamma}}\boldsymbol{R}_i = \mathrm{E}\,\Big[\Big(1 - \frac{1}{M_i}\Big)\mathrm{E}_i(\boldsymbol{Z}_{ir}\boldsymbol{Z}_{ir}^T)\Big].$$

Symbol $\frac{\partial}{\partial\boldsymbol{\gamma}}\boldsymbol{R}_i$ denotes the value of matrix $(\frac{\partial}{\partial\gamma_1}\boldsymbol{R}_i(\boldsymbol{\gamma}), \frac{\partial}{\partial\gamma_2}\boldsymbol{R}_i(\boldsymbol{\gamma}), \ldots, \frac{\partial}{\partial\gamma_l}\boldsymbol{R}_i(\boldsymbol{\gamma}))^T$ for the true parameter $\boldsymbol{\gamma}$. It holds that

$$\frac{1}{\tilde{\pi}_{ir}} - \frac{1}{\pi_{ir}} = -\frac{1-\pi_{ir}}{\pi_{ir}}(\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma})^T\boldsymbol{Z}_{ir} + o_p\Big(\frac{1}{\sqrt{n}}\Big).$$

We can write

$$\begin{aligned}
\sqrt{n}(\tilde{\theta} - \theta) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big[\frac{1}{\mu} + \Big(\frac{1}{\hat{\mu}} - \frac{1}{\mu}\Big)\Big]\sum_{r=1}^{M_i}\Big[\frac{1}{\pi_{ir}} + \Big(\frac{1}{\tilde{\pi}_{ir}} - \frac{1}{\pi_{ir}}\Big)\Big]\xi_{ir}Y_{ir} - \sqrt{n}\theta \\
&= \frac{1}{\sqrt{n}\mu}\sum_{i=1}^{n}\sum_{r=1}^{M_i}\frac{\xi_{ir}}{\pi_{ir}}Y_{ir} + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(\frac{1}{\hat{\mu}} - \frac{1}{\mu}\Big)\sum_{r=1}^{M_i}\frac{\xi_{ir}}{\pi_{ir}}Y_{ir} \\
&+ \frac{1}{\sqrt{n}\mu}\sum_{i=1}^{n}\sum_{r=1}^{M_i}\Big(\frac{1}{\tilde{\pi}_{ir}} - \frac{1}{\pi_{ir}}\Big)\xi_{ir}Y_{ir} + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(\frac{1}{\hat{\mu}} - \frac{1}{\mu}\Big)\sum_{r=1}^{M_i}\Big(\frac{1}{\tilde{\pi}_{ir}} - \frac{1}{\pi_{ir}}\Big)\xi_{ir}Y_{ir} - \sqrt{n}\theta
\end{aligned}$$

and since

$$\begin{aligned}
&\frac{1}{\sqrt{n}\mu}\sum_{i=1}^{n}\sum_{r=1}^{M_i}\Big(\frac{1}{\tilde{\pi}_{ir}} - \frac{1}{\pi_{ir}}\Big)\xi_{ir}Y_{ir} = \\
&\qquad = \frac{1}{\sqrt{n}\mu}\sum_{i=1}^{n}\sum_{r=1}^{M_i}\Big(-\frac{1-\pi_{ir}}{\pi_{ir}}\big(\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{R}_j^T J_\gamma^{-1}\big)\boldsymbol{Z}_{ir}\Big)\xi_{ir}Y_{ir} + o_p(1) \\
&\qquad = -\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\boldsymbol{R}_j^T J_\gamma^{-1}\frac{1}{n\mu}\sum_{i=1}^{n}\sum_{r=1}^{M_i}\Big(\frac{\xi_{ir}}{\pi_{ir}}(1-\pi_{ir})\boldsymbol{Z}_{ir}Y_{ir}\Big) + o_p(1) \\
&\qquad = -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{R}_i^T J_\gamma^{-1}\boldsymbol{q} + o_p(1),
\end{aligned}$$

we get

$$\begin{aligned}
\sqrt{n}(\tilde{\theta} - \theta) =& \frac{1}{\sqrt{n}\mu}\sum_{i=1}^{n}M_i\sum_{r=1}^{M_i}\xi_{ir}Y_{ir} - \frac{1}{\sqrt{n}}\theta\sum_{i=1}^{n}\Big(\frac{M_i}{\mu} - 1\Big) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{R}_i^T J_\gamma^{-1}\boldsymbol{q} - \sqrt{n}\theta + o_p(1) \\
=& \frac{1}{\sqrt{n}}\sum_{i=1}^{n}Q_i + o_p(1),
\end{aligned}$$

where $Q_i = \frac{1}{\mu} M_i \sum_{r=1}^{M_i} \xi_{ir} Y_{ir} - \theta \frac{M_i}{\mu} + \boldsymbol{R}_i^T J_\gamma^{-1} \boldsymbol{q}$.

Since

$$\begin{aligned}
\mathrm{E}\, \boldsymbol{R}_i \boldsymbol{R}_i^T &= \mathrm{E}\Big[\sum_{r=1}^{M_i} (\xi_{ir} - \pi_{ir}) \boldsymbol{Z}_{ir} \sum_{s=1}^{M_i} (\xi_{is} - \pi_{is}) \boldsymbol{Z}_{is}^T \Big] \\
&= \mathrm{E}\Big[\sum_{r=1}^{M_i} (\xi_{ir} - \pi_{ir})^2 \boldsymbol{Z}_{ir} \boldsymbol{Z}_{ir}^T + \sum_r \sum_{s \neq r} (\xi_{ir} - \pi_{ir})(\xi_{is} - \pi_{is}) \boldsymbol{Z}_{ir} \boldsymbol{Z}_{is}^T \Big] \\
&= \mathrm{E}\Big[\Big(1 - \frac{1}{M_i}\Big) \mathrm{E}_i (\boldsymbol{Z}_{ir} \boldsymbol{Z}_{ir}^T - \boldsymbol{Z}_{ir} \boldsymbol{Z}_{is}^T) \Big] = J_\gamma - L_\gamma
\end{aligned}$$

and similarly

$$\begin{aligned}
\frac{1}{\mu} \mathrm{E}\Big[\Big(M_i \sum_{r=1}^{M_i} \xi_{ir} Y_{ir} \Big) \boldsymbol{R}_i^T \Big] &= \frac{1}{\mu} \mathrm{E}\Big[\Big(M_i \sum_{r=1}^{M_i} \xi_{ir} Y_{ir} \Big) \sum_{s=1}^{M_i} (\xi_{is} - \pi_{is}) \boldsymbol{Z}_{is}^T \Big] \\
&= \frac{1}{\mu} \mathrm{E}\Big[\Big(M_i \sum_{r=1}^{M_i} \xi_{ir} (\xi_{ir} - \pi_{ir}) Y_{ir} \boldsymbol{Z}_{ir}^T \Big) + \sum_r \sum_{s \neq r} M_i \xi_{ir} (\xi_{is} - \pi_{is}) Y_{ir} \boldsymbol{Z}_{is}^T \Big] \\
&= \frac{1}{\mu} \mathrm{E}\left[(M_i - 1) \mathrm{E}_i (Y_{ir} \boldsymbol{Z}_{ir}^T - Y_{ir} \boldsymbol{Z}_{is}^T) \right] = \boldsymbol{q}^T - \boldsymbol{t}^T,
\end{aligned}$$

we have

$$\Sigma_{\tilde{\theta}} = \operatorname{var} Q_i = \Sigma_{\hat{\theta}} + \boldsymbol{q}^T J_\gamma^{-1} (J_\gamma - L_\gamma) J_\gamma^{-1} \boldsymbol{q} - 2(\boldsymbol{q}^T - \boldsymbol{t}^T) J_\gamma^{-1} \boldsymbol{q}.$$

According to the Central limit theorem for iid random variables, $\sqrt{n}(\tilde{\theta} - \theta) \xrightarrow{d} N(0, \Sigma_{\tilde{\theta}})$.

□

In a general situation, it does not necessarily hold that $\Sigma_{\tilde{\theta}} \leq \Sigma_{\hat{\theta}}$, or in other words, the use of the auxiliary variable could increase the variance of the estimator. However, we will show that under certain conditions the variance of the estimator will be improved. The conditions are:

- $\mathrm{E}_i \boldsymbol{X}_{ir} = 0$ (5.3)
- vectors $(Y_{ir}, \boldsymbol{X}_{ir})$ and $(Y_{is}, \boldsymbol{X}_{is})$ for $r \neq s$ are independent, given household. (5.4)

If the first condition is fulfilled, the variance $\Sigma_{\tilde{\theta}}$ (5.2) simplifies to

$$\Sigma_{\tilde{\theta}} = \Sigma_{\hat{\theta}} + \boldsymbol{q}_\delta^T J_\delta^{-1} (J_\delta - L_\delta) J_\delta^{-1} \boldsymbol{q}_\delta - 2(\boldsymbol{q}_\delta^T - \boldsymbol{t}_\delta^T) J_\delta^{-1} \boldsymbol{q}_\delta, \tag{5.5}$$

where

$$\begin{aligned}
\boldsymbol{q}_\delta &= \frac{1}{\mu} \mathrm{E}\left[(M_i - 1) \mathrm{E}_i (Y_{ir} \boldsymbol{X}_{ir}) \right] & \boldsymbol{t}_\delta &= \frac{1}{\mu} \mathrm{E}\left[(M_i - 1) \mathrm{E}_i (Y_{ir} \boldsymbol{X}_{is}) \right] \\
J_\delta &= \mathrm{E}\Big[\Big(1 - \frac{1}{M_i}\Big) \mathrm{E}_i (\boldsymbol{X}_{ir} \boldsymbol{X}_{ir}^T) \Big] & L_\delta &= \mathrm{E}\Big[\Big(1 - \frac{1}{M_i}\Big) \mathrm{E}_i (\boldsymbol{X}_{ir} \boldsymbol{X}_{is}^T) \Big].
\end{aligned}$$

If both conditions are fulfilled, the variance $\Sigma_{\tilde{\theta}}$ (5.2) is

$$\Sigma_{\tilde{\theta}} = \Sigma_{\hat{\theta}} - \boldsymbol{q}_\delta^T J_\delta^{-1} \boldsymbol{q}_\delta.$$

Because J_δ^{-1} is a positive semidefinite matrix, $\boldsymbol{q}_\delta^T J_\delta^{-1} \boldsymbol{q}_\delta \geq 0$. This implies that $\Sigma_{\hat{\theta}} \geq \Sigma_{\tilde{\theta}}$, which means that the use of auxiliary variables to adjust the weights in the estimator of the parameter θ can never increase the asymptotic variance of the estimator. In the worst case, when the auxiliary variables and the target variable are independent, the asymptotic variances of the original estimator $\hat{\theta}$ and the estimator $\tilde{\theta}$ are equal. When the auxiliary variables are correlated with the variable of interest, the variance of the estimator $\tilde{\theta}$ is lower.

We saw in chapter 3 (expression 3.7) that

$$\Sigma_{\hat{\theta}} = \frac{1}{\mu^2}\left[\mathrm{E}\,\frac{M_i^2}{m_i}\mathrm{var}_i Y_{ir} + \mathrm{E}\,M_i^2(\theta_i - \theta)^2\right].$$

If one member from each family is sampled ($m_i = 1$), we obtain

$$\Sigma_{\hat{\theta}} = \frac{1}{\mu^2}\left[\mathrm{E}\,M_i^2\mathrm{var}_i Y_{ir} + \mathrm{E}\,M_i^2(\theta_i - \theta)^2\right],$$

where only the first part of the expression, i.e. $\mathrm{E}\,M_i^2\mathrm{var}_i Y_{ir}$, is amenable to improvement by the use of an auxiliary variable. The lower limit for the variance of an improved estimator is the variance which would be obtained if all the individuals were used for the estimation, i.e.

$$\Sigma_{\hat{\theta}} = \frac{1}{\mu^2}\left[\mathrm{E}\,M_i\mathrm{var}_i Y_{ir} + \mathrm{E}\,M_i^2(\theta_i - \theta)^2\right]. \tag{5.6}$$

Transformation of Auxiliary Variable

While in many practical situations the second of the two conditions will be fulfilled, the first one will rarely be true. We can define a transformation of the auxiliary variable X_{ir} as

$$D_{ir} = X_{ir} - \mathrm{E}_i(X_{ir}), \tag{5.7}$$

the discrepancy of X_{ir} from its expectation in the ith household. Then $\mathrm{E}_i D_{ir} = 0$. The sampling probabilities are estimated as described above, denoted $\tilde{\pi}_{ir}^e$. The estimator of the parameter θ has the form

$$\tilde{\theta}_e = \frac{\sum_{i=1}^n \sum_{r=1}^{M_i} \frac{\xi_{ir}}{\tilde{\pi}_{ir}^e} Y_{ir}}{\sum_{i=1}^n M_i}.$$

As the household expectation is rarely known, we could not apply this transformation in practice. Nevertheless, we still use it to formulate important theoretical results and to get a better insight into the problem. The implementation in practical situations and its consequences will be presented later. The idea described above is summarized in Theorem 8.

Theorem 8. *Let $M_i, i = 1, 2, \ldots, n$, be iid random variables. Let $Y_{ir}, i = 1, 2, \ldots, n$ and $r = 1, 2, \ldots, M_i$, be random variables with marginal distribution given by (3.1) and* $\operatorname{var} Y_{ir} < \infty$. *Let $(Y_{i1}, Y_{i2} \ldots Y_{iM_i})$ and also $(\xi_{i1}, \xi_{i2} \ldots \xi_{iM_i})$ be independent random vectors for $i = 1, 2, \ldots n$. Let $\sum_{r=1}^{M_i} \xi_{ir} = 1$ and ξ_{ir} is independent from Y_{ir}, given M_i. Then*

$$\sqrt{n}(\tilde{\theta}_e - \theta) \xrightarrow{d} N(0, \Sigma_e),$$

where

$$\Sigma_e = \Sigma_{\hat{\theta}} - \boldsymbol{q}_\delta^T J_\delta^{-1} \boldsymbol{q}_\delta \tag{5.8}$$

for

$$\boldsymbol{q}_\delta = \frac{1}{\mu} \mathrm{E}\left[(M_i - 1)\mathrm{cov}_i(\boldsymbol{D}_{ir}, Y_{ir})\right] \quad \textit{and} \quad J_\delta = \mathrm{E}\left[\left(1 - \frac{1}{M_i}\right)\mathrm{var}_i \boldsymbol{D}_{ir}\right].$$

From a sample of n households, the variance Σ_e can be estimated as

$$\hat{\Sigma}_e = \hat{\Sigma}_{\hat{\theta}} - \hat{\boldsymbol{q}}_\delta^T \hat{J}_\delta^{-1} \hat{\boldsymbol{q}}_\delta, \tag{5.9}$$

where

$$\hat{\boldsymbol{q}}_\delta = \frac{1}{\hat{\mu}} \frac{1}{n} \sum_{i=1}^{n} (M_i - 1) \sum_{r=1}^{M_i} \xi_{ir} Y_{ir} \boldsymbol{D}_{ir} \quad \text{and} \quad \hat{J}_\delta = \frac{1}{n} \sum_{i=1}^{n} \left(1 - \frac{1}{M_i}\right) \frac{1}{M_i} \sum_{r=1}^{M_i} \boldsymbol{D}_{ir} \boldsymbol{D}_{ir}^T, \tag{5.10}$$

$\hat{\Sigma}_{\hat{\theta}}$ and $\hat{\mu}$ are given in (3.5).

It is useful to see how the situation simplifies when all households have equal size, $M_i = M$. Then we can write for the variance (5.8)

$$\begin{aligned} \Sigma_e &= \Sigma_{\hat{\theta}} - \left(1 - \frac{1}{M}\right) \mathrm{E}\,\mathrm{cov}_i(Y_{ir}, \boldsymbol{D}_{ir}) (\mathrm{E}\,\mathrm{var}_i \boldsymbol{D}_{ir})^{-1} \mathrm{E}\,\mathrm{cov}_i(\boldsymbol{D}_{ir}, Y_{ir}) \\ &= \Sigma_{\hat{\theta}} - \left(1 - \frac{1}{M}\right) \mathrm{E}\,\mathrm{cov}_i(Y_{ir}, \boldsymbol{X}_{ir}) (\mathrm{E}\,\mathrm{var}_i \boldsymbol{X}_{ir})^{-1} \mathrm{E}\,\mathrm{cov}_i(\boldsymbol{X}_{ir}, Y_{ir}). \end{aligned} \tag{5.11}$$

Similarly as in the case of stratified sampling, we can use as an auxiliary variable $D_{ir}/\pi_{ir} = M_i D_{ir}$. Then we get

$$\Sigma_e = \Sigma_{\hat{\theta}} - \boldsymbol{q}_\delta^T J_\delta^{-1} \boldsymbol{q}_\delta,$$

where

$$\boldsymbol{q}_\delta = \frac{1}{\mu} \mathrm{E}\left[M_i(M_i - 1)\mathrm{cov}_i(\boldsymbol{D}_{ir}, Y_{ir})\right] \quad \text{and} \quad J_\delta = \mathrm{E}\left[M_i(M_i - 1)\mathrm{var}_i \boldsymbol{D}_{ir}\right].$$

If it holds for $i = 1, 2, \ldots, n$ that

$$\mathrm{cov}_i(\boldsymbol{X}_{ir}, Y_{ir}) = \boldsymbol{c}, \quad \mathrm{var}_i \boldsymbol{X}_{ir} = V, \quad \mathrm{var}_i Y_{ir} = \sigma^2, \tag{5.12}$$

we get

$$\Sigma_e = \frac{1}{\mu^2} \left[\mathrm{E}\, M_i^2 \sigma^2 + \mathrm{E}\, M_i^2 (\theta_i - \theta)^2 - \boldsymbol{c}^T V^{-1} \boldsymbol{c}\, \mathrm{E}\, M_i(M_i - 1)\right]. \tag{5.13}$$

If we have a variable X_{ir} for which $\mathrm{cor}_i(Y_{ir}, X_{ir}) = 1$, we obtain

$$\Sigma_e = \frac{1}{\mu^2}[\sigma^2(\mathrm{E}\, M_i^2 - \mathrm{E}\, M_i^2 + \mathrm{E}\, M_i) + \mathrm{E}\, M_i^2(\theta_i - \theta)^2] = \frac{1}{\mu^2}[\mathrm{E}\, M_i \mathrm{var}_i Y_{ir} + \mathrm{E}\, M_i^2(\theta_i - \theta)^2],$$

and the variance of the estimator $\tilde{\theta}_e$ reaches its lower limit, see (5.6). In this sense, the optimal transformation of the auxiliary variable $\boldsymbol{X}_{ir}$ is defined as $M_i[X_{ir} - \mathrm{E}_i(X_{ir})]$.

Since we can hardly assume that the household expectation would be known in practice, the transformation (5.7) of the auxiliary variable cannot usually be obtained, and thus Theorem 8 has purely theoretical relevance. A logical step is to replace the expectation in (5.7) by its estimate, i.e. the average of the observations in the household. However, since the number of members in one household is small, we cannot rely on asymptotics and must investigate the consequences of such replacement in more detail.

Let us denote $\bar{X}_i = \frac{1}{M_i}\sum_{r=1}^{M_i} X_{ir}$ and define a transformation of the auxiliary variable X_{ir} as

$$H_{ir} = X_{ir} - \bar{X}_i.$$

All the other steps and notation will stay the same as in the previous subsection. Condition (5.3) is still fulfilled, since $\mathrm{E}_i H_{ir} = \mathrm{E}_i X_{ir} - \mathrm{E}_i \bar{X}_i = 0$. The main change lies in the fact that we lose the independence of H_{ir} and H_{is}, for $r \neq s$, within a household.

We will denote the estimated sampling probabilities by $\tilde{\pi}_{ir}^a$ and the estimator of the parameter θ by

$$\tilde{\theta}_a = \frac{\sum_{i=1}^n \sum_{r=1}^{M_i} \frac{\xi_{ir}}{\tilde{\pi}_{ir}^a} Y_{ir}}{\sum_{i=1}^n M_i}.$$

Theorem 9. *Let $M_i, i = 1, 2, \ldots, n$, be iid random variables. Let $Y_{ir}, i = 1, 2, \ldots, n$ and $r = 1, 2, \ldots, M_i$, be random variables with marginal distribution given by (3.1) and $\mathrm{var}\, Y_{ir} < \infty$. Let $(Y_{i1}, Y_{i2} \ldots Y_{iM_i})$ and also $(\xi_{i1}, \xi_{i2} \ldots \xi_{iM_i})$ be independent random vectors for $i = 1, 2, \ldots n$. Let $\sum_{r=1}^{M_i} \xi_{ir} = 1$ and ξ_{ir} is independent from Y_{ir}, given M_i. Then*

$$\sqrt{n}(\tilde{\theta}_a - \theta) \xrightarrow{d} N(0, \Sigma_a),$$

where

$$\Sigma_a = \Sigma_{\hat{\theta}} + \boldsymbol{q}_\delta^T J_\delta^{-1} \mathrm{E}\,[\mathrm{var}_i \boldsymbol{H}_{ir}] J_\delta^{-1} \boldsymbol{q}_\delta - 2\frac{1}{\mu}\mathrm{E}\,[M_i \mathrm{cov}_i(Y_{ir}, \boldsymbol{H}_{ir})] J_\delta^{-1} \boldsymbol{q}_\delta, \tag{5.14}$$

for

$$\boldsymbol{q}_\delta = \frac{1}{\mu}\mathrm{E}\,[(M_i - 1)\mathrm{cov}_i(\boldsymbol{H}_{ir}, Y_{ir})] \quad \textit{and} \quad J_\delta = \mathrm{E}\left[\left(1 - \frac{1}{M_i}\right)\mathrm{var}_i \boldsymbol{H}_{ir}\right].$$

Here, the variance Σ_a differs from the variance Σ_e in Theorem 8 in the fact that the matrix $\mathrm{E}\,[\mathrm{var}_i \boldsymbol{H}_{ir}]$ does not cancel with J_δ and $\frac{1}{\mu}\mathrm{E}\,[M_i \mathrm{cov}_i(\boldsymbol{H}_{ir}, Y_{ir})]$ is not equal to $\boldsymbol{q}_\delta$.

Proof We refer to the proof of Theorem 7 and highlight only the specific steps. Note that

$$\mathrm{var}_i H_{ir} = \mathrm{E}_i (X_{ir} - \bar{X}_i)(X_{ir} - \bar{X}_i)^T = \Big(1 - \frac{1}{M_i}\Big)\mathrm{var}_i X_{ir}, \tag{5.15}$$

$$\mathrm{cov}_i(Y_{ir}, H_{ir}) = \mathrm{cov}_i(Y_{ir}, X_{ir} - \bar{X}_i) = \Big(1 - \frac{1}{M_i}\Big)\mathrm{cov}_i(Y_{ir}, X_{ir}). \tag{5.16}$$

Since

$$\begin{aligned}
&\mathrm{E}\Big[\Big(1 - \frac{1}{M_i}\Big)\mathrm{E}_i[\boldsymbol{H}_{ir}\boldsymbol{H}_{ir}^T - \boldsymbol{H}_{ir}\boldsymbol{H}_{is}^T]\Big] \\
&\quad = \mathrm{E}\Big[\Big(1 - \frac{1}{M_i}\Big)\mathrm{E}_i[(\boldsymbol{X}_{ir} - \bar{\boldsymbol{X}}_i)(\boldsymbol{X}_{ir} - \bar{\boldsymbol{X}}_i)^T - (\boldsymbol{X}_{ir} - \bar{\boldsymbol{X}}_i)(\boldsymbol{X}_{is} - \bar{\boldsymbol{X}}_i)^T]\Big] \\
&\quad = \mathrm{E}\Big[\Big(1 - \frac{1}{M_i}\Big)[E_i(\boldsymbol{X}_{ir}\boldsymbol{X}_{ir}^T) - E_i\boldsymbol{X}_{ir}E_i\boldsymbol{X}_{is}]\Big] = \mathrm{E}\Big[\mathrm{E}_i[(\boldsymbol{X}_{ir} - \bar{\boldsymbol{X}}_i)(\boldsymbol{X}_{ir} - \bar{\boldsymbol{X}}_i)^T]\Big] \\
&\quad = \mathrm{E}\Big[\mathrm{E}_i(\boldsymbol{H}_{ir}\boldsymbol{H}_{ir}^T)\Big] = \mathrm{E}\Big[\mathrm{var}_i \boldsymbol{H}_{ir}\Big],
\end{aligned}$$

we get

$$\mathrm{E}\,\boldsymbol{R}_i\boldsymbol{R}_i^T = \mathrm{E}\Big[\Big(1 - \frac{1}{M_i}\Big)\mathrm{E}_i[\boldsymbol{Z}_{ir}\boldsymbol{Z}_{ir}^T - \boldsymbol{Z}_{ir}\boldsymbol{Z}_{is}^T]\Big] = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathrm{E}\,[\mathrm{var}_i H_{ir}] \end{pmatrix}.$$

Similarly,

$$\begin{aligned}
&\frac{1}{\mu}\mathrm{E}\Big[(M_i - 1)E_i[Y_{ir}(\boldsymbol{H}_{ir} - \boldsymbol{H}_{is})]\Big] \\
&\qquad = \frac{1}{\mu}\mathrm{E}\Big[(M_i - 1)E_i(Y_{ir}[(\boldsymbol{X}_{ir} - \bar{\boldsymbol{X}}_i) - (\boldsymbol{X}_{is} - \bar{\boldsymbol{X}}_i)])\Big] = \frac{1}{\mu}\mathrm{E}\,[M_i\mathrm{cov}_i(Y_{ir}, \boldsymbol{H}_{ir})]
\end{aligned}$$

and thus

$$\frac{1}{\mu}\mathrm{E}\Big[\Big(M_i\sum_{r=1}^{M_i}\xi_{ir}Y_{ir}\Big)\boldsymbol{R}_i\Big] = \frac{1}{\mu}\mathrm{E}\Big[(M_i - 1)E_i[Y_{ir}(\boldsymbol{Z}_{ir} - \boldsymbol{Z}_{is})]\Big] = \big(\mathbf{0}, \frac{1}{\mu}\mathrm{E}\,[M_i\mathrm{cov}_i(Y_{ir}, \boldsymbol{H}_{ir})]\big)$$

□

From a sample of n households, the variance Σ_a can be estimated as

$$\hat{\Sigma}_a = \hat{\Sigma}_{\hat{\theta}} + \hat{\boldsymbol{q}}_\delta^T\hat{J}_\delta^{-1}\Big(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{M_i}\sum_{r=1}^{M_i}\boldsymbol{H}_{ir}\boldsymbol{H}_{ir}^T\Big)\hat{J}_\delta^{-1}\hat{\boldsymbol{q}}_\delta - 2\frac{1}{\hat{\mu}}\Big(\frac{1}{n}\sum_{i=1}^{n}M_i\sum_{r=1}^{M_i}\xi_{ir}Y_{ir}\boldsymbol{H}_{ir}^T\Big)\hat{J}_\delta^{-1}\hat{\boldsymbol{q}}_\delta, \tag{5.17}$$

where $\hat{\boldsymbol{q}}_\delta$ and $\hat{J}_\delta$ are estimated analogously to (5.10), $\hat{\Sigma}_{\hat{\theta}}$ and $\hat{\mu}$ are given in (3.5).

If all the households have the same number of members $M_i = M$, the variance (5.14) can be rewritten as

$$\begin{aligned}
\Sigma_a &= \Sigma_{\hat{\theta}} - \mathrm{E}\,\mathrm{cov}_i(Y_{ir}, \boldsymbol{H}_{ir})(\mathrm{E}\,\mathrm{var}_i\boldsymbol{H}_{ir})^{-1}\mathrm{E}\,\mathrm{cov}_i(\boldsymbol{H}_{ir}, Y_{ir}) \\
&= \Sigma_{\hat{\theta}} - \Big(1 - \frac{1}{M}\Big)\mathrm{E}\,\mathrm{cov}_i(Y_{ir}, \boldsymbol{X}_{ir})(\mathrm{E}\,\mathrm{var}_i\boldsymbol{X}_{ir})^{-1}\mathrm{E}\,\mathrm{cov}_i(\boldsymbol{X}_{ir}, Y_{ir}).
\end{aligned}$$

The second equality follows from (5.15) and (5.16). The expression for the variance Σ_a is equal to (5.11), from which we can conclude that in case households have the same size, the replacement of the household expectation by its estimate does not affect the asymptotic variance of the estimator.

Similarly as before, we use the transformation $M_i H_{ir}$ of the auxiliary variable. Then we have

$$\Sigma_a = \Sigma_{\hat{\theta}} + \boldsymbol{q}_\delta^T J_\delta^{-1} \mathrm{E}\,[M_i^2 \mathrm{var}_i \boldsymbol{H}_{ir}] J_\delta^{-1} \boldsymbol{q}_\delta - 2\frac{1}{\mu} \mathrm{E}\,[M_i^2 \mathrm{cov}_i(Y_{ir}, \boldsymbol{H}_{ir})] J_\delta^{-1} \boldsymbol{q}_\delta,$$

where

$$\boldsymbol{q}_\delta = \frac{1}{\mu} \mathrm{E}\,[M_i(M_i-1)\mathrm{cov}_i(\boldsymbol{H}_{ir}, Y_{ir})] \quad \text{and} \quad J_\delta = \mathrm{E}\,[M_i(M_i-1)\mathrm{var}_i \boldsymbol{H}_{ir}].$$

If we assume equal covariance and variance matrices in all households as specified in (5.12) and consider (5.15) and (5.16), we get

$$\begin{aligned} \Sigma_a &= \Sigma_{\hat{\theta}} + \frac{1}{\mu^2}\Big[\mathrm{E}\, M_i^2\Big(1 - \frac{1}{M_i}\Big)c^T V^{-1} c - 2\mathrm{E}\, M_i^2\Big(1 - \frac{1}{M_i}\Big)c^T V^{-1} c\Big] \\ &= \Sigma_{\hat{\theta}} - \frac{1}{\mu^2} \mathrm{E}\, M_i(M_i-1) c^T V^{-1} c, \end{aligned} \tag{5.18}$$

which is in this specific case equal to Σ_e (see 5.13), and thus the replacement of household expectation by its estimate does not affect the asymptotic variance of the estimator. Moreover, if $\mathrm{cor}_i(Y_{ir}, X_{ir}) = 1$, the variance of estimator $\tilde{\theta}_a$ reaches its lower limit.

5.2 Example

This example illustrates the statements of theorems 8 and 9. Let us assume two kinds of households; small (of size $M = 2$) and large (of size $M = 6$), equally represented in a population, $p_2 = \mathrm{P}(M = 2) = p_6 = \mathrm{P}(M = 6) = 0.5$. The mean value of the variable of interest Y in small households $\theta_{(2)} = 100$, in large households $\theta_{(6)} = 150$. Let us assume that the value of the target variable for the rth member in the ith household is

$$Y_{ir} = \theta_{(m)} + \delta_i + \epsilon_{ir}, \tag{5.19}$$

and the auxiliary variable for this individual is

$$X_{ir} = \theta_{(m)} + \delta_i + \epsilon_{ir} + \eta_{ir}, \tag{5.20}$$

where m is the size of the ith household and $\delta_i \sim (0, \sigma_\delta^2)$, $\epsilon_{ir} \sim (0, \sigma_\epsilon^2)$ and $\eta_{ir} \sim (0, \sigma_\eta^2)$ are iid random variables, mutually independent of each other. The random variable δ_i represents a household-specific component (and thus reflects the fact that observations within one household are not independent), ϵ_{ir} is subject-specific component and η_{ir} stands

for the difference between the variable of interest and the auxiliary variable. Let $\sigma_\delta^2 = 1200$, $\sigma_\epsilon^2 = 900$ and $\sigma_\eta^2 = 400$.

The mean household size is

$$\mu = p_2\theta_{(2)} + p_6\theta_{(6)} = 4.$$

The mean value of the target variable Y is

$$\theta = \mathrm{E}\, Y_{ir} = \frac{\mathrm{E}\, M_i\theta_i}{\mu} = (p_2 M_2\theta_{(2)} + p_6 M_6\theta_{(6)})/\mu = 137.50$$

When we observe only one member from each household, the asymptotic variance of the normalized estimator $\hat{\theta}$ (i.e. the asymptotic variance of $\sqrt{n}(\hat{\theta} - \theta)$) is

$$\begin{aligned}\Sigma_{\hat{\theta}} &= \frac{1}{\mu^2}\mathrm{E}\, M_i^2(\bar{Y}_i - \theta)^2 = \frac{1}{\mu^2}\mathrm{E}\,\mathrm{E}\left(M_i^2[(Y_i - \theta_{(m)})^2 + (\theta_{(m)} - \theta)^2]|M_i\right)\\ &= \frac{1}{\mu^2}\left(p_2 M_2^2[(\sigma_\delta^2 + \sigma_\epsilon^2) + (\theta_{(2)} - \theta)^2] + p_6 M_6^2[(\sigma_\delta^2 + \sigma_\epsilon^2) + (\theta_{(6)} - \theta)^2]\right) = 2976.\end{aligned}$$

Now, we would like to see how this estimator can be improved with the help of the auxiliary variable X. If we observed all the members in each household, the asymptotic variance of the normalized estimator $\hat{\theta}$ would be

$$\begin{aligned}\Sigma_{\hat{\theta}} &= \frac{1}{\mu^2}\mathrm{E}\, M_i^2(\bar{Y}_i - \theta)^2 = \frac{1}{\mu^2}\mathrm{E}\,\mathrm{E}\left(M_i^2[(Y_i - \theta_{(m)})^2 + (\theta_{(m)} - \theta)^2]|M_i\right)\\ &= \frac{1}{\mu^2}\left(p_2 M_2^2[(\sigma_\delta^2 + \frac{1}{2}\sigma_\epsilon^2) + (\theta_{(2)} - \theta)^2] + p_6 M_6^2[(\sigma_\delta^2 + \frac{1}{6}\sigma_\epsilon^2) + (\theta_{(6)} - \theta)^2]\right) = 2076.\end{aligned} \tag{5.21}$$

This is the lowest variance of the estimator that could be obtained with this data. However, we can observe only one member from each household.

Let us now apply the estimator $\tilde{\theta}_e$. We use the transformation

$$M_i D_{ir} = M_i(X_{ir} - E_i X_{ir}) = M_i(\epsilon_{ir} + \eta_{ir}).$$

We have

$$\begin{aligned}\mathrm{cov}_i(Y_{ir}, D_{ir}) &= \mathrm{E}_i(\theta_{(m)} + \delta_i + \epsilon_{ir})(\epsilon_{ir} + \eta_{ir}) = \sigma_\epsilon^2\\ \mathrm{var}_i D_{ir} &= \mathrm{E}_i D_{ir}^2 = \sigma_\epsilon^2 + \sigma_\eta^2\end{aligned}$$

and thus

$$\begin{aligned}q_\delta &= \frac{1}{\mu}\sigma_\epsilon^2\mathrm{E}\, M_i(M_i - 1)\\ J_\delta &= (\sigma_\epsilon^2 + \sigma_\eta^2)\mathrm{E}\, M_i(M_i - 1).\end{aligned}$$

This implies that the asymptotic variance of the normalized estimator $\tilde{\theta}_e$ is

$$\Sigma_e = \Sigma_{\hat{\theta}} - q_\delta^2/J_\delta = \Sigma_{\hat{\theta}} - \frac{1}{\mu^2}\mathrm{E}\, M_i(M_i - 1)\frac{\sigma_\epsilon^4}{(\sigma_\epsilon^2 + \sigma_\eta^2)} = 2353. \tag{5.22}$$

Table 5.1: Example of cluster sampling; Results of simulation

Estimator	Use of auxiliary variable	Average estimate of θ	Variance of estimator		
			asympt.	empirical	estimate*
$\hat{\theta}$	No	137.56	2.976	3.003	2.968
$\tilde{\theta}_e$	$X_{ir} - \mathrm{E}_i X_{ir}$	137.56	2.353	2.407	2.305
$\tilde{\theta}_a$	$X_{ir} - \bar{X}_i$	137.57	2.353	2.409	2.296

*Average of estimates

If the auxiliary variable is identical to the target variable, i.e. $\sigma_\eta^2 = 0$, the variance of the estimator $\tilde{\theta}_e$ reaches its lower limit (5.21)

$$\Sigma_e = \Sigma_{\hat{\theta}} - \frac{1}{\mu^2}\mathrm{E}\, M_i(M_i - 1)\sigma_\epsilon^2 = 2076. \tag{5.23}$$

As previously mentioned, the estimator $\tilde{\theta}_e$ is not applicable in reality. Instead, we must use the estimator $\tilde{\theta}_a$, which replaces the household mean value in the transformation of the auxiliary variable by its estimate. The transformed auxiliary variable is now

$$M_i H_{ir} = M_i(X_{ir} - \bar{X}_{ir}) = M_i(\epsilon_{ir} + \eta_{ir} - \bar{\epsilon}_i - \bar{\eta}_i).$$

From (5.18) we know that in this example

$$\Sigma_a = \Sigma_e = 2353.$$

Simulation The results can also be illustrated by a simulation. The procedure was conducted as follows. First, we generated a population of 1000 households, with size assigned randomly to each household, being 2 or 6 with probability 0.5. Then values of the target and the auxiliary variable were generated for each member according to (5.19) and (5.20). We assumed a normal distribution of the variables δ,ϵ and η. In the second step, one member from each household was sampled at random and based on this sample, the parameter θ was estimated by the three suggested estimators; i.e. either not taking into account the auxiliary variable X ($\hat{\theta}$) or using the information contained in X as described above ($\tilde{\theta}_e$ and $\tilde{\theta}_a$). The variance of the estimator was estimated as suggested in (3.5), (5.9) and (5.17).

This procedure was repeated 1000 times. The results are displayed in Table 5.1. The average estimate of parameter θ was in all three cases very close to the real value 137.5. We can see a common pattern in the variance of the estimators. While the empirical variance was on average slightly higher than the calculated asymptotic variance, the estimate of the asymptotic variance was on average slightly lower. However, the difference between the empirical variance and the estimated variance of the estimator was acceptable in all three cases.

5.3 Simulation Study

All the results presented in this chapter were related to the asymptotic variance of the estimates, with the hope that an estimate of the asymptotic variance could be used for inference about the parameter of interest. To show that this approach is appropriate in different situations, a simulation study was conducted. The aim was to simulate different scenarios and to compare the empirical and the estimated asymptotic variances of the estimator.

We investigated the effect of 3 possible factors which could influence the variance of the estimator and its estimate. First, we focused on the size of households. Second, we inspected the correlation of the target and the auxiliary variable and the variability of observations within households with respect to the variability between households. Third, we considered different distributions generating the observations to make sure that the results were not specific for normally distributed random variables.

The construction of simulations investigating the first two factors was very similar to the previous example. We assumed that the value of the target variable for the rth member in the ith household is

$$Y_{ir} = \theta_{(m)} + \delta_i + \epsilon_{ir}, \tag{5.24}$$

and the auxiliary variable for this individual is

$$X_{ir} = \theta_{(m)} + \delta_i + \epsilon_{ir} + \eta_{ir}, \tag{5.25}$$

where m is the size of the ith household and $\delta_i \sim (0, \sigma_\delta^2)$, $\epsilon_{ir} \sim (0, \sigma_\epsilon^2)$ and $\eta_{ir} \sim (0, \sigma_\eta^2)$ are iid random variables, mutually independent of each other.

The simulation procedure comprised of the following steps. First, we generated a population of 1000 households, with size assigned randomly to each household. Based on the size of the household, the household-specific mean value was generated. Around this mean value, observations for the target variable (as specified by (5.24)) were simulated. Then values for the auxiliary variable were generated, as defined by (5.25). Again, we assumed a normal distribution of variables δ, ϵ and η. In the second step, one member of each household was sampled at random. Based on this sample, the parameter θ was estimated by the three suggested estimators. The variance of the estimate was estimated according to (3.5), (5.9) and (5.17). Using the estimated variance, the 95% confidence interval (CI) was also derived. This procedure was repeated 3000 times. The average estimate of the parameter was calculated, as well as the average estimate of the asymptotic variance. This latter quantity was compared to the empirical variance of 3000 simulated estimates. The percentage of confidence intervals which covered the true value θ was obtained in order to assess whether their coverage was close to the desired 95 %. The estimate of the parameter θ based on the full population and its empirical variance was also calculated; it represents the lower limit of the variance which could be possibly obtained. If not specified otherwise, the default options were:

- households of size from 2 to 6 members, each size equally represented in the population
- for household of size m
$$\theta_{(m)} = 50 + 25m$$
- $\sigma_\delta^2 = 1200$, $\sigma_\epsilon^2 = 900$ and $\sigma_\eta^2 = 400$

Size of Households

First, we addressed the situation where all the households have the same size. To inspect the performance of the estimators in case the households are very small, we assumed that all the households have only 2 members, where for one of them the target variable Y is observed, while for the other one only the value of the auxiliary variable is known and used to improve the estimate of the mean value of Y. We also considered a scenario of all households having an intermediate size, i.e. 6, and a large size, i.e. 12. In all three cases, the mean value θ was 150. The results are displayed in Table 5.2.

We can see that in all three cases, the average estimate of the parameter is very close to the true value $\theta = 150$. The estimate of the asymptotic variance is slightly lower than the empirical variance of the estimator. However, the difference is acceptable. This is also confirmed by the coverage of the confidence intervals which is very close to 95 %. The results illustrate the fact that the more members the households have, the lower is the variance of the estimator that takes the auxiliary information into account. In all three cases, the variances of $\tilde{\theta}_e$ and $\tilde{\theta}_a$ are essentially equal. This confirms the theoretical results from the end of section 5.1, where we saw that for households of an equal size, the replacement of the household expectation by its estimate does not affect the asymptotic variance of the estimator.

Let us now consider households of variable size. We focus on the following scenarios:

- Small households (from 2 to 6 members), each size is equally represented in the population. For a household of the size m,
$$\theta_{(m)} = 50 + 25m.$$
- Large households (from 8 to 12 members), each size is equally represented in the population. For a household of the size m,
$$\theta_{(m)} = 50 + 25(m - 6).$$
- The household size has a shifted Poisson distribution
$$size\ of\ household \sim Poiss(4) + 2,$$
so that the minimal household size is 2 and the mean household size is 6. For a household of the size m,
$$\theta_{(m)} = 80 + 10 \min(m, 12).$$

Table 5.2: Simulation results: households of equal size

Estimator	Use of auxiliary variable	Average estimate of θ	Variance of estimator*		Coverage of CI (%)
			empirical	estimate**	
Household size 2					
$\hat{\theta}$	No	150.00	2161	2098	94.83
$\tilde{\theta}_e$	$X_{ir} - \mathrm{E}_i X_{ir}$	150.00	1832	1772	94.83
$\tilde{\theta}_a$	$X_{ir} - \bar{X}_i$	150.00	1836	1761	94.37
Full pop.	–	150.01	1683		
Household size 6					
$\hat{\theta}$	No	150.01	2138	2096	94.73
$\tilde{\theta}_e$	$X_{ir} - \mathrm{E}_i X_{ir}$	150.01	1578	1554	94.80
$\tilde{\theta}_a$	$X_{ir} - \bar{X}_i$	150.01	1581	1551	94.33
Full pop.	–	150.02	1370		
Household size 12					
$\hat{\theta}$	No	150.02	2109	2099	94.90
$\tilde{\theta}_e$	$X_{ir} - \mathrm{E}_i X_{ir}$	150.02	1538	1501	94.57
$\tilde{\theta}_a$	$X_{ir} - \bar{X}_i$	150.02	1538	1501	94.60
Full pop.	–	150.03	1278		

*Estimate of the asymptotic variance of $\sqrt{n}(\check{\theta}-\theta)$, where $\check{\theta}$ is the corresponding estimator.
**Average of estimates

- The household size has a shifted Poisson distribution

$$size\ of\ household \sim Poiss(8) + 2,$$

so that the minimal household size is 2 and the mean household size is 10. For a household of the size m,

$$\theta_{(m)} = 80 + 10 \min(m, 18).$$

The theoretical value of the parameter θ is different for each of these options and thus it is presented in Table 5.3 on page 59 together with the results of the simulation.

As in the previous table, we observe that the average estimate of the parameter θ is very close to its true value. In most cases, the average estimate of the asymptotic variance is slightly lower than the empirical variance of the estimator. Still, the difference is acceptable. The coverage of the confidence intervals is also very good.

Variability and Correlation

Next, we simulated data for different values of σ^2_δ, σ^2_ϵ and σ^2_η. The values of σ^2_ϵ and σ^2_δ control the variability of the target variable within households and between households. The values of σ^2_η determine the correlation between the target variable and the auxiliary variable. The true expectation of the target variable was $\theta = 162.50$. The results are displayed in Table 5.4 on page 60.

The table shows that the average estimate of the asymptotic variance tends to be lower than the empirical variance of the estimator, but the difference is acceptable and the coverage of the confidence intervals is again very close to 95 %. The results also demonstrate that the weaker the correlation between the target and the auxiliary variable, the smaller the gain in the efficiency of the estimators $\tilde{\theta}_e$ and $\tilde{\theta}_a$. In the last case ($\sigma^2_\delta = 900$, $\sigma^2_\epsilon = 400$, $\sigma^2_\eta = 1200$), where the correlation between Y and X within households and the variability of Y within households are both small, the variance of the estimator using the auxiliary variable was improved only by a very small amount. It is also noteworthy that the variance of the estimators $\tilde{\theta}_e$ or $\tilde{\theta}_a$ was never bigger than the variance of the estimator $\hat{\theta}$.

Different distribution

To demonstrate that the results can be generalized beyond normally distributed random variables, the following data were simulated. We assumed a linear predictor for the ith household

$$\nu_i = \theta_{(m)} + \delta_i,$$

where m is the size of the ith household and δ_i are iid random variables. We assumed that the mean value for the ith household is

$$\mathrm{E}_i Y_{ir} = \exp(\nu_i)$$

and generated the observations Y_{ir} from

- the Poisson distribution
- the gamma distribution with the shape parameter k and the scale parameter $\psi_i = \frac{\mathrm{E}_i Y_{ir}}{k}$.

The auxiliary variable was defined as

$$X_{ir} = Y_{ir} + \eta_{ir}, \text{ where } \eta_{ir} \sim N(0, \sigma^2_\eta).$$

We considered two different distributions of the household-specific component δ_i

- the normal distribution $N(0, \sigma^2_\delta)$
- the skew-normal distribution with location μ, scale σ and shape α.

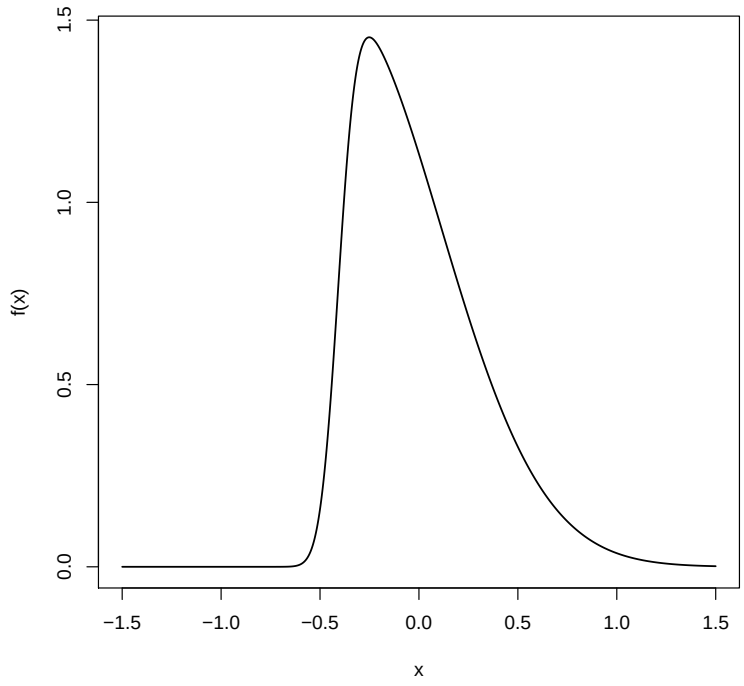


Figure 5.1: The density of the skew-normal distribution SN(−0.41,0.52,7)

Remark The skew-normal distribution [2] with the parameters $(\mu,\ \sigma,\ \alpha)$ is defined by the density function

$$f(x) = \frac{1}{\sigma\pi}\exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big)\int_{-\infty}^{\alpha(\frac{x-\mu}{\sigma})}\exp\Big(-\frac{t^2}{2}\Big)dt, \quad x\in(-\infty,\infty).$$

The parameter α determines the skewness of the distribution, for $\alpha = 0$ we obtain the normal distribution $N(\mu,\sigma^2)$. It holds

$$\mathrm{E}\,X = \mu + \sigma\delta\sqrt{\frac{2}{\pi}},\ \text{where}\ \delta = \frac{\alpha}{\sqrt{1+\alpha^2}}$$
$$\mathrm{var}X = \sigma^2\Big(1-\frac{2\delta^2}{\pi}\Big).$$

Figure 5.1 displays the density of the skew-normal distribution with parameters $\mu = -0.41,\ \sigma = 0.52,\ \alpha = 7$.

The simulation procedure was carried out analogously to the previously described scenarios. Households had size from 2 to 6 members, each size equally represented in the population. The parameters were:

- for the Poisson distribution

$$\theta_{(m)} = 2 + 0.2(m-2) \quad \text{and} \quad \sigma_\eta^2 = 16$$

- for the gamma distribution

$$\theta_{(m)} = 4.5 + 0.2(m-2)$$

and one of the following

$$k = 2 \quad \text{and} \quad \sigma_\eta^2 = 1600$$
$$k = 9 \quad \text{and} \quad \sigma_\eta^2 = 800$$

- in both cases for the distribution of δ_i

$$N(0, 0.1) \quad \text{or} \quad SN(-0.41, 0.52, 7).$$

The parameters for the skew-normal distribution were chosen so that $\mathrm{E}\, X = 0$ and $\mathrm{var} X = 0.1$.

The results are presented in Table 5.5 on page 61. In the case of the Gamma distribution with shape $k = 2$, we obtained a negative estimate of the variance of $\tilde{\theta}_e$ or $\tilde{\theta}_a$ in 27 (normal distribution of δ) and 42 (skew-normal distribution of δ) out of 3000 repetitions. These cases were excluded from the presentation of the results in Table 5.5. In practice, the estimator $\hat{\theta}$ would be used when $\tilde{\theta}$ cannot be calculated or has a negative variance estimate. However, the frequency of such cases was very small and the estimator worked well overall. The parameter estimates were close to the true values, the estimates of the asymptotic variance were very close to the empirical variance and the coverage of the confidence intervals was never less than 93 %.

Summary

In this simulation study, we inspected three factors which could influence the variance of the estimator and its estimate. In all three cases, the estimate of the symptotic variance performed relatively well. In most cases, it was slightly lower then the empirical variance of the estimate, however the difference between the two was acceptable. The coverage of the 95% confidence intervals was always close to the desired level of 95 %. In all situations, the loss of precision caused by the replacement of the household expectation by its estimate was negligible.

Table 5.3: Simulation results: normal distribution, varying distributions of household sizes

Estimator	Use of auxiliary variable	Average estimate of θ	Variance of estimator*		Coverage of CI (%)
			empirical	estimate**	
Small households (size 2-6), uniform distribution, $\theta = 162.5$					
$\hat{\theta}$	No	162.50	3453	3428	94.87
$\tilde{\theta}_e$	$X_{ir} - \mathrm{E}_i X_{ir}$	162.51	2985	2836	93.73
$\tilde{\theta}_a$	$X_{ir} - \bar{X}_i$	162.51	2989	2826	93.70
Full pop.	–	162.50	2725		
Large households (size 8-12), uniform distribution, $\theta = 155$					
$\hat{\theta}$	No	155.06	3406	3354	94.57
$\tilde{\theta}_e$	$X_{ir} - \mathrm{E}_i X_{ir}$	155.04	2792	2741	94.13
$\tilde{\theta}_a$	$X_{ir} - \bar{X}_i$	155.04	2793	2739	94.17
Full pop.	–	155.02	2547		
Size of household $\sim$ Poiss(4)+2, $\theta = 146.53$					
$\hat{\theta}$	No	146.63	2840	2835	95.17
$\tilde{\theta}_e$	$X_{ir} - \mathrm{E}_i X_{ir}$	146.60	2323	2194	93.63
$\tilde{\theta}_a$	$X_{ir} - \bar{X}_i$	146.61	2326	2188	93.73
Full pop.	–	146.61	2039		
Size of household $\sim$ Poiss(8)+2, $\theta = 187.82$					
$\hat{\theta}$	No	187.91	3267	3183	94.60
$\tilde{\theta}_e$	$X_{ir} - \mathrm{E}_i X_{ir}$	187.90	2637	2507	93.53
$\tilde{\theta}_a$	$X_{ir} - \bar{X}_i$	187.91	2640	2505	93.47
Full pop.	–	187.90	2357		

*Estimate of the asymptotic variance of $\sqrt{n}(\check{\theta} - \theta)$, where $\check{\theta}$ is the corresponding estimator.
**Average of estimates

Table 5.4: Simulation results: normal distribution, varying variability and correlation

Estimator	Use of auxiliary variable	Average estimate of θ	Variance of estimator*		Coverage of CI (%)
			empirical	estimate**	
$\sigma_\delta^2 = 1200$, $\sigma_\epsilon^2 = 900$, $\sigma_\eta^2 = 400$					
$\hat{\theta}$	No	162.50	3453	3428	94.87
$\tilde{\theta}_e$	$X_{ir} - \mathrm{E}_i X_{ir}$	162.51	2985	2836	93.73
$\tilde{\theta}_a$	$X_{ir} - \bar{X}_i$	162.51	2989	2826	93.70
Full pop.	–	162.50	2725		
$\sigma_\delta^2 = 1200$, $\sigma_\epsilon^2 = 900$, $\sigma_\eta^2 = 900$					
$\hat{\theta}$	No	162.50	3453	3428	94.87
$\tilde{\theta}_e$	$X_{ir} - \mathrm{E}_i X_{ir}$	162.51	3128	2988	94.03
$\tilde{\theta}_a$	$X_{ir} - \bar{X}_i$	162.51	3132	2978	93.87
Full pop.	–	162.50	2725		
$\sigma_\delta^2 = 400$, $\sigma_\epsilon^2 = 900$, $\sigma_\eta^2 = 1200$					
$\hat{\theta}$	No	162.50	2515	2529	95.13
$\tilde{\theta}_e$	$X_{ir} - \mathrm{E}_i X_{ir}$	162.50	2225	2146	94.53
$\tilde{\theta}_a$	$X_{ir} - \bar{X}_i$	162.51	2230	2136	94.10
Full pop.	–	162.50	1783		
$\sigma_\delta^2 = 900$, $\sigma_\epsilon^2 = 400$, $\sigma_\eta^2 = 1200$					
$\hat{\theta}$	No	162.50	2590	2529	94.53
$\tilde{\theta}_e$	$X_{ir} - \mathrm{E}_i X_{ir}$	162.50	2534	2395	94.13
$\tilde{\theta}_a$	$X_{ir} - \bar{X}_i$	162.51	2537	2385	94.03
Full pop.	–	162.50	2270		

*Estimate of the asymptotic variance of $\sqrt{n}(\check{\theta} - \theta)$, where $\check{\theta}$ is the corresponding estimator.
**Average of estimates

Table 5.5: Simulation results: non-normal distributions

Estimator	Use of auxiliary variable	Average estimate of θ	Variance of estimator*		Coverage of CI (%)
			empirical	estimate**	
Poisson distribution, Normal distribution of δ, $\theta = 13.242$					
$\hat{\theta}$	No	13.244	52.435	52.189	94.80
$\tilde{\theta}_e$	$X_{ir} - \mathrm{E}_i X_{ir}$	13.245	46.124	45.696	94.80
$\tilde{\theta}_a$	$X_{ir} - \bar{X}_i$	13.246	46.204	45.603	94.73
Full pop.	–	13.243	38.996		
Poisson distribution, Skew-normal distribution of δ, $\theta = 13.312$					
$\hat{\theta}$	No	13.311	60.599	61.541	95.00
$\tilde{\theta}_e$	$X_{ir} - \mathrm{E}_i X_{ir}$	13.308	54.929	54.829	95.03
$\tilde{\theta}_a$	$X_{ir} - \bar{X}_i$	13.309	55.053	54.678	95.10
Full pop.	–	13.309	48.310		
Gamma distribution with $k = 2$, Normal distribution of δ, $\theta = 161.33$					
$\hat{\theta}$	No	161.23	23654	24704	95.19
$\tilde{\theta}_e$	$X_{ir} - \mathrm{E}_i X_{ir}$	161.25	10808	10082	93.31
$\tilde{\theta}_a$	$X_{ir} - \bar{X}_i$	161.24	10837	10037	93.17
Full pop.	–	161.28	9219		
Gamma distribution with $k = 2$, Skew-normal distribution of δ, $\theta = 162.15$					
$\hat{\theta}$	No	162.00	25076	26992	95.47
$\tilde{\theta}_e$	$X_{ir} - \mathrm{E}_i X_{ir}$	162.18	11937	11567	93.75
$\tilde{\theta}_a$	$X_{ir} - \bar{X}_i$	162.17	11921	11538	93.54
Full pop.	–	162.18	10558		
Gamma distribution with $k = 9$, Normal distribution of δ, $\theta = 161.33$					
$\hat{\theta}$	No	161.33	9257	9667	95.60
$\tilde{\theta}_e$	$X_{ir} - \mathrm{E}_i X_{ir}$	161.34	6655	6673	94.00
$\tilde{\theta}_a$	$X_{ir} - \bar{X}_i$	161.34	6673	6639	93.80
Full pop.	–	161.34	6018		
Gamma distribution with $k = 9$, Skew-normal distribution of δ, $\theta = 162.16$					
$\hat{\theta}$	No	162.16	10844	11243	95.13
$\tilde{\theta}_e$	$X_{ir} - \mathrm{E}_i X_{ir}$	162.14	7987	8027	95.03
$\tilde{\theta}_a$	$X_{ir} - \bar{X}_i$	162.14	8001	7995	94.73
Full pop.	–	162.14	7359		

*Estimate of the asymptotic variance of $\sqrt{n}(\check{\theta} - \theta)$, where $\check{\theta}$ is the corresponding estimator.
**Average of estimates

5.4 Application to Project ACCEPT Data

In this chapter, we will show an application of some of our results to a subset of Project ACCEPT data collected during the baseline behavioral assessment [10]. The baseline survey consisted of administering demographic and behavioral questionnaires to a two-stage probability sample of eligible community residents. At the first stage, households were selected with equal probability from a listing of all households in a community. At the second stage, one eligible household member was sampled from each household that was selected at the first stage, and included at least one eligible individual. The demographic and behavioral questionnaires were administered to each selected individual; household-level data were collected on a separate form. Members of the households were eligible to participate in the baseline survey if they are were aged 18-32 years, had lived in the community at least 4 months in the past year, and slept regularly in their household at least 2 nights per week. In this analysis, we will consider data collected in Vulindlela, South Africa.

The behavioral questionnaires included questions related to HIV risk behaviors (sexual life, alcohol and drug use), HIV testing, HIV-related stigma, social norms and similar aspects. Here, we will focus on the following items.

- **Drug use** – answer to the question: "Have you ever used any drugs in your lifetime?" (yes/no)
- **Number of sexual partners** – answer to the question: "In your lifetime, with how many different people have you had sex?"(numeric)
- **HIV testing history** – answer to the question: "Have you ever been voluntarily tested for HIV?" (Dichotomized yes = voluntary or non-voluntary test/ no = never)
- **Social norms score** derived from responses to 6 questions regarding community norms concerning HIV testing. Each response was evaluated on the scale 0 – 3 (3 = strongly agree; 2 = agree; 1 = disagree; 0 = strongly disagree, reversed in negatively phrased questions) and the scores were summed. Thus, the range of the Social norms score is between 0 and 18, with a high score interpreted as a positive outcome.

The household-level data included information about age and gender of each eligible member of each selected household. Therefore, age and gender can serve as auxiliary information in the analysis of behavioral assessments.

Population

A total of 2596 households with at least one eligible member were selected. There were 4969 eligible individuals living in these households. Table 5.6 displays basic characteristics of the *first stage population*, i.e. all individuals living in the selected households, and the *second stage population*, i.e. individuals selected for the assessments.

Table 5.6: Basic characteristics of the study populations

	First stage population	Second stage population
Households	2596	2596
Individuals	4969	2596
Males	2214 (44.56 %)	1075 (41.41 %)
Females	2755 (55.44 %)	1521 (58.59 %)
Age		
Mean	23.95	23.79
Median	23	23

Table 5.7: Distribution of household size

Size	1	2	3	4	5	6	7	8
Number of households	1246	728	363	167	53	30	7	2
Percent of households	48.00	28.04	13.98	6.43	2.04	1.16	0.27	0.08

There were slightly fewer males in the second stage population than in the first stage population (41.41 % versus 44.56 %). The mean age was comparable in the two populations (23.95 and 23.79 years). Table 5.7 summarizes the distribution of household size. There were 1246 (48 %) households with only one eligible member, the rest of the households had from 2 to 8 eligible members.

Results

Table 5.8 presents for each variable of interest (Drug use, Number of sexual partners, HIV testing history, Social norms score) either average or proportion of positive answers, as appropriate, calculated by gender. For simplicity, we always used a subset of data with non-missing values for a given variable.

Our goal was to estimate the expectation (having the meaning of the probability of the answer "yes", where appropriate). For each variable, we calculated the average, the estimate of the expectation taking into account household size and estimates of the expectation taking into account both household size and the auxiliary information. We also calculated the estimates of the asymptotic variance of the three estimators. The results are presented in Table 5.9 on page 66. It displayes the variances of the normalized estimators, i.e. variances of $\sqrt{n}(\hat{\theta} - \theta)$ for the estimator $\hat{\theta}$. Also in the following description of the results we always refer to the asymptotic variance of the normalized estimator.

Note that 48 % of households had only one eligible member. Indeed, only the remaining 52 % of households play a role in increasing the precision of the estimates.

Table 5.8: Summary of the selected responses

	Male	Female
Drug use (yes)	295 (27.52 %)	30 (1.97 %)
Number of sexual partners	7.58	2.47
HIV testing history (yes)	178 (16.62 %)	651 (42.91 %)
Social norms score	6.82	6.40

Drug use The probability of having used a drug was estimated to be 0.1218, taking into account household size. From Table 5.8 we can see that drug use strongly depends on gender, with males more likely to have had an experience with drugs. It suggests that using the information about gender, known for the first stage population, could improve the precision of the estimation.

The estimate taking into account gender was 0.1280, slightly higher than the estimate that ignores the auxiliary information. This is due to the fact that there were lower percentage of males in the second stage population than in the first stage population. In the estimating procedure, higher weights were assigned to males compared to females and thus the estimate was shifted more towards the values in males, which are in average higher than the values in females.

By taking into account the information about gender, the estimated asymptotic variance of the estimate decreased from 0.1436 to 0.1349 (by 6 %). At first glance we might expect more gain in precision. However, we should keep in mind that the extent to which the precision can be improved depends on the correlation of the target and auxiliary variables *within* households. We can imagine this as the ability of the auxiliary variable to distinguish members of the given household in terms of the variable of interest. For a dichotomous auxiliary variable this ability is limited.

Accounting for age had a negligible influence on the precision of the estimate.

Number of sexual partners The estimated mean number of sexual partners was 4.6092, taking into account the size of the household. The estimate of the mean number of sexual partners also taking into account gender was 4.7782, again slightly higher than the estimate not considering any auxiliary information. The estimated asymptotic variance of the estimate decreased from 163.15 to 160.04 by adjusting for gender. The estimate adjusted for gender and age was 4.7782, its variance was 159.85.

HIV testing history The estimated probability of having been tested for HIV (taking into account household size) was 0.3052. The estimate of this probability taking into account gender was 0.2996, reflecting that females, who undergo HIV testing more commonly, were represented slightly more in the second stage population than in the first stage population. The estimated asymptotic variance of the estimate decreased from 0.2796 to 0.2603. Accounting for age had a negligible influence on the precision of the estimate.

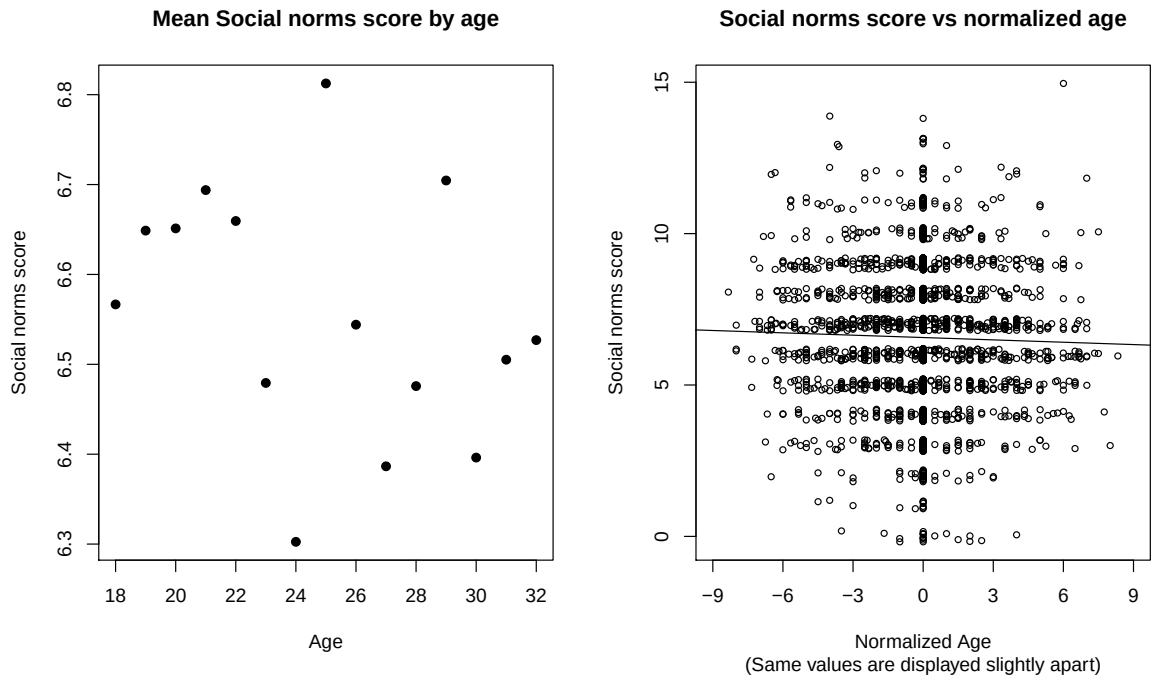


Figure 5.2: Social norms score against age

Social norms score The estimated mean social norms score, taking into account size of household, was 6.6186. Figure 5.2 displays average social norms score by age groups. We can observe a slight tendency of the score to decrease with age. The same figure also displays the original scores plotted against the "normalized" age, i.e. the age from which the mean household value was subtracted. There is a tendency - although almost negligible - that older members of households report lower social norms score than the younger ones.

The estimate of the mean social norms score taking into account age was 6.6237. The variance decreased from 6.2425 to 4.9803. That is, knowledge of age for the full population helped to improve the precision of the estimate by 20 %. The estimate taking into account age and gender was 6.6461, with the estimated variance 4.7804. Thus, employing the available auxiliary information decreased the variance of the estimator by 23 %.

Discussion

We have seen that in all four cases, the estimate taking into account gender as auxiliary information reflected a slight underrepresentation of males in the second stage population compared to the first stage population. The observation that the estimates were shifted towards values in males refers to the particular realisation of the estimating procedure and does not contradict the statement that both estimates (whether taking into account auxiliary information or not) are asymptotically unbiased. The shift of the estimate towards the

Table 5.9: Project ACCEPT data; Estimation of expectation of the selected variables

	Average	Taking into account household size and			
Drug use		–	gender	age	gender & age
$\hat{\theta}$	0.1254	0.1218	0.1280	0.1214	0.1284
$n * \widehat{\text{var}}\,\hat{\theta}^1$		0.1436	0.1349	0.1427	0.1341
Nbr of sexual partners					
$\hat{\theta}$	4.5598	4.6092	4.7655	4.6189	4.7782
$n * \widehat{\text{var}}\,\hat{\theta}^1$		163.15	160.04	162.99	159.85
HIV testing					
$\hat{\theta}$	0.3203	0.3052	0.2996	0.3102	0.3050
$n * \widehat{\text{var}}\,\hat{\theta}^1$		0.2796	0.2603	0.2795	0.2603
Social norms score					
$\hat{\theta}$	6.5721	6.6186	6.6434	6.6237	6.6461
$n * \widehat{\text{var}}\,\hat{\theta}^1$		6.2425	6.0575	4.9803	4.7804

[1] $n * \widehat{\text{var}}\,\hat{\theta}$ refers to the estimate of the asymptotic variance of $\sqrt{n}(\hat{\theta} - \theta)$.

values in males could go in the wrong direction, i.e. farther from the expectation. However, we have shown that the variance of the estimator adjusting for the auxiliary variable is smaller, which implies that the shift goes "more often" in the correct direction, i.e. closer to the expectation. While the first three variables seemed to be strongly dependent on gender, the use of gender as the auxiliary information resulted in a relatively small decrease in the variance of the estimator. We ascribe this to the fact that gender is a dichotomous variable and thus its ability to distinguish members of a given household in terms of the variable of interest is limited. On the other hand, considering age as the auxiliary variable lead to a 20% (and together with gender to a 23%) decrease in the variance of the estimate of the mean Social norms score. For this effect, an overall association of Social norms score with age would not be enough. The gain in precision depends on the role which age plays within households. To depict this within households correlation, a graph of the social norms score against the normalized age was plotted (see Figure 5.2). Surprisingly, the correlation seems to be very weak, we can observe only a very slight tendency that within one household, older members express slightly lower social norms than younger members. However, in this case even weak correlation seems to have an impact on the precision of the estimate.

Chapter 6

Extension to the Regression Problem

6.1 Introduction

In this chapter we will study extensions of the methods investigated in the previous chapters to the regression problem. We will assume a response variable Y and a vector of explanatory variables $\boldsymbol{X} = (X_1, \ldots, X_p)$. The aim will be to estimate the regression parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ which links the explanatory variables to the expectation of Y. More specifically, we will suppose that there exists a strictly monotone and twice differentiable link function g for which

$$\mathrm{E}\, Y = g^{-1}(\boldsymbol{x}^T\boldsymbol{\theta}). \tag{6.1}$$

A typical situation, which has been studied for example by Breslow et. al [3] and on which we will focus as well, happens when the response is known for the whole population but some explanatory variables are observed only for a subsample. Normally, the estimation procedure would be based only on the subsample where the full information is available. However, we will show that if there are auxiliary variables correlated with the regressors, they can be used to improve the efficiency of the estimators.

In order to present the estimation methods in a broader context, we will first provide a brief overview of the most often used concepts. In the framework of the *Generalized Linear Models* (GLM) it is required that the conditional density of Y, given $\boldsymbol{X} = \boldsymbol{x}$, is from the exponential family, i.e.

$$f(y|\boldsymbol{x}) = \exp\Big(\frac{yt - b(t)}{\phi}a + c(y, \phi)\Big), \tag{6.2}$$

where $t \in (-\infty, \infty)$, $\phi > 0$, $a > 0$ and $b(t)$ is twice continuously differentiable function. The parameter t depends on $\boldsymbol{x}$ through the linear predictor $\eta = \boldsymbol{x}^T\boldsymbol{\theta}$. We assume that $\eta = g(\mathrm{E}\, Y)$ and thus we have (6.1). The target is to draw inference about the parameter $\boldsymbol{\theta}$. Let

$$\boldsymbol{U}_i(\boldsymbol{\theta}) = \Big(\frac{\partial f_i(\boldsymbol{\theta}|y_i)}{\partial\theta_j}\Big)_{j=1}^{p} = \Big(\frac{y_i - \mathrm{E}\, Y_i}{\phi v(\mathrm{E}\, Y_i)}\frac{\partial g^{-1}(\eta_i)}{\partial\eta_i}x_{ij}\Big)_{j=1}^{p}, \tag{6.3}$$

where $\phi v(\mathrm{E}\, Y_i) = \mathrm{var} Y_i$, denotes the usual parametric likelihood score for an individual i. The parameter is estimated by solving the estimating equations

$$\sum_{i=1}^{N} \boldsymbol{U}_i(\boldsymbol{\theta}) = \mathbf{0}. \tag{6.4}$$

The likelihood equations depend on the assumed distribution for Y_i only through $\mathrm{E}\, Y_i$ and $\mathrm{var}(Y_i)$. Based on this observation, the *quasi-likelihood* estimation was proposed. It only specifies the mean-variance relationship rather than the whole density of Y_i. The mean has the form (6.1) and the variance is

$$\mathrm{var}(Y_i) = \phi v(\mathrm{E}\, Y_i)$$

for some chosen variance function v. The equations that determine the quasi-likelihood estimates are the same as the likelihood equations (6.4) for GLMs. They are not likelihood equations, however, without the additional assumption that Y_i has density (6.2). For example, the purpose of the quasi-likelihood method might be to encompass a greater variety of cases than it corresponds to some distribution from the exponential family. The quasi-likelihood estimator has the same asymptotic variance as the ordinary GLM estimator.

When $\mathrm{E}\, Y_i$ satisfies (6.1) but the variance is misspecified, the quasi-likelihood estimating equations still provide a consistent estimator of the regression parameters, but their asymptotic variance is different. This property is utilized in the *pseudo-likelihood* method. The equations that determine the parameter estimates are still the same as the likelihood equations (6.4), but it is no longer assumed that $v(\mathrm{E}\, Y_i)$ is the correct variance function of Y_i. The asymptotic variance is estimated by the so-called *sandwich estimator*, see (6.10) and (6.11) below. For simplicity, the expression $\boldsymbol{U}_i$ will be called *score* in all situations.

In the rest of this chapter we modify the quasi-likelihood estimating equations to stratified sampling and cluster sampling and show how to use auxiliary information to improve the asymptotic variance of the regression parameter estimator. We will address each sampling design separately. For stratified sampling, we derive the results previously published in [3], but we present them in a slightly different form and provide more detailed proofs. The analogous problem for cluster sampling has not been previously studied.

6.2 Stratified Sampling

Let us assume the stratified sampling design defined in chapter 2. If the model with mean value given by (6.1) was valid for Y in each stratum, i.e. independently of W, then the sampling stratum would not have to be taken into account, and the estimation methods described above would also be valid for data collected under the stratified sampling scheme.

We will deal with a more general situation where conditional distribution of Y given $\boldsymbol{X}$ varies between strata in a certain way. However, we are interested in modeling the conditional mean of Y given $\boldsymbol{X}$ in the sense of (6.1), where stratum is *not* included between

the explanatory variables. Such models are sometimes called *population-averaged*. They are very common in certain situations, for example when analyzing repeated measurements data. In that case, the population-averaged model is used when the research question pertains to the marginal distribution, but the correlation between the observations from one subject must be taken into account.

The marginal density of Y given $\boldsymbol{X}$ is a mixture of the stratum-specific densities $f(y|\boldsymbol{x}, W = k)$, i.e.

$$f(y|\boldsymbol{x}) = \sum_{k=1}^{K} p_k f(y|\boldsymbol{x}, W = k).$$

We assume that (6.1) holds. We consider the score equations (6.4) as an estimating equations derived from the pseudo-likehood principle and take it as a basis for the estimation of $\boldsymbol{\theta}$ with completely observed data. When the complete data are only available for a subsample obtained by stratified random sampling, we modify the estimating equations in the same way as in Section 2.2

$$\boldsymbol{V}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \sum_{k=1}^{K} \frac{\xi_i}{\hat{\pi}_k} \boldsymbol{U}_i(\boldsymbol{\theta}) \mathrm{I}_{ik} = \boldsymbol{0}, \quad \text{where} \quad \hat{\pi}_k = \frac{1}{N_k} \sum_{i=1}^{N} \xi_i \mathrm{I}_{ik}. \tag{6.5}$$

The estimate of the parameter $\boldsymbol{\theta}$, denoted as $\hat{\boldsymbol{\theta}}$, is obtained by solving (6.5).

Theorem 10. *Assume that* $(Y_i, \boldsymbol{X_i}, W_i, \xi_i)$ *are iid random vectors,* $\mathrm{E}\, Y_i = g^{-1}(\boldsymbol{x}_i^T \boldsymbol{\theta})$ *for some known link function* g *and* $\operatorname{var} Y_i < \infty$. *Let* ξ_i *be independent of* Y_i *and* $\boldsymbol{X_i}$ *given* W_i, *for* $i = 1, 2, \ldots, N$. *Then the following holds:*

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N(0, D(\boldsymbol{\theta})^{-1} \Sigma D(\boldsymbol{\theta})^{-1}), \tag{6.6}$$

where

$$\Sigma = J(\boldsymbol{\theta}) + \sum_{k=1}^{K} p_k \frac{1 - \pi_k}{\pi_k} J_k(\boldsymbol{\theta}), \tag{6.7}$$

$D(\boldsymbol{\theta}) = (-\mathrm{E}\, \frac{\partial}{\partial \theta_j} \boldsymbol{U}_i^T)_{j=1}^{p}$, $J(\boldsymbol{\theta}) = \operatorname{var} \boldsymbol{U}_i(\boldsymbol{\theta})$ *(not conditioning on* k*) and* $J_k(\boldsymbol{\theta}) = \operatorname{var}_k \boldsymbol{U}_i(\boldsymbol{\theta})$.

Proof First we will show that

$$\frac{1}{\sqrt{N}} \boldsymbol{V}(\boldsymbol{\theta}) \xrightarrow{d} N(0, \Sigma). \tag{6.8}$$

By the Taylor expansion of $\frac{1}{\hat{\pi}_k}$ around $\frac{1}{\pi_k}$, we get

$$\frac{1}{\hat{\pi}_k} - \frac{1}{\pi_k} = -\frac{1}{\pi_k^2} \frac{1}{N_k} \sum_{i=1}^{N} (\xi_i - \pi_k) \mathrm{I}_{ik} + o_p\Big(\frac{1}{\sqrt{N_k}}\Big)$$

and we can write

$$\frac{1}{\sqrt{N}}\boldsymbol{V}(\boldsymbol{\theta}) = \frac{1}{\sqrt{N}}\Big[\sum_{i=1}^{N}\sum_{k=1}^{K}\frac{\xi_i \boldsymbol{U}_i(\boldsymbol{\theta})}{\pi_k}\mathrm{I}_{ik} - \\ - \sum_{k=1}^{K}\frac{1}{\pi_k N_k}\Big(\sum_{i=1}^{N}\frac{\xi_i \boldsymbol{U}_i(\boldsymbol{\theta})}{\pi_k}\mathrm{I}_{ik}\Big)\Big(\sum_{j=1}^{N}(\xi_j - \pi_k)\mathrm{I}_{jk}\Big)\Big] + o_p(1).$$

If we denote $\boldsymbol{S}_k = \mathrm{E}_k \boldsymbol{U}_i(\boldsymbol{\theta})$, we obtain

$$\frac{1}{\sqrt{N}}\boldsymbol{V}(\boldsymbol{\theta}) = \frac{1}{\sqrt{N}}\Big[\sum_{i=1}^{N}\sum_{k=1}^{K}\frac{\xi_i \boldsymbol{U}_i(\boldsymbol{\theta})}{\pi_k}\mathrm{I}_{ik} - \sum_{i=1}^{N}\sum_{k=1}^{K}\frac{(\xi_i - \pi_k)}{\pi_k}\boldsymbol{S}_k \mathrm{I}_{ik}\Big] + o_p(1).$$

Thus

$$\frac{1}{\sqrt{N}}\boldsymbol{V}(\boldsymbol{\theta}) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\boldsymbol{Q}_i(\boldsymbol{\theta}) + o_p(1),$$

where

$$\boldsymbol{Q}_i(\boldsymbol{\theta}) = \sum_{k=1}^{K}\left[\frac{\xi_i \boldsymbol{U}_i(\boldsymbol{\theta})}{\pi_k} - \frac{(\xi_i - \pi_k)}{\pi_k}\boldsymbol{S}_k\right]\mathrm{I}_{ik}$$

are iid random variables and $\mathrm{E}\,\boldsymbol{Q}_i = \mathrm{E}\,\boldsymbol{U}_i(\boldsymbol{\theta}) = \boldsymbol{0}$. After a short calculation we get

$$\operatorname{var}\boldsymbol{Q}_i = \mathrm{E}\,\boldsymbol{Q}_i\boldsymbol{Q}_i^T = J(\boldsymbol{\theta}) + \sum_k p_k \frac{1-\pi_k}{\pi_k}J_k(\boldsymbol{\theta}).$$

According to the Central limit theorem for iid random variables

$$\frac{1}{\sqrt{N}}\boldsymbol{V}(\boldsymbol{\theta}) \xrightarrow{d} N(\boldsymbol{0}, \operatorname{var}\boldsymbol{Q}_i).$$

By the Taylor expansion of $\boldsymbol{V}(\hat{\boldsymbol{\theta}})$ around $\boldsymbol{V}(\boldsymbol{\theta})$

$$\boldsymbol{V}(\hat{\boldsymbol{\theta}}) - \boldsymbol{V}(\boldsymbol{\theta}) = \Big(\frac{\partial}{\partial\theta_j}\boldsymbol{V}(\boldsymbol{\theta})^T\Big)_{j=1}^{p}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) + o_p\Big(\frac{1}{\sqrt{N}}\Big),$$

and therefore

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = -\Big[\frac{1}{N}\Big(\frac{\partial}{\partial\theta_j}\boldsymbol{V}(\boldsymbol{\theta})^T\Big)_{j=1}^{p}\Big]^{-1}\frac{1}{\sqrt{N}}\boldsymbol{V}(\boldsymbol{\theta}) + o_p(1).$$

We have

$$-\frac{1}{N}\Big(\frac{\partial}{\partial\theta_j}\boldsymbol{V}(\boldsymbol{\theta})^T\Big)_{j=1}^{p} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\frac{\xi_i}{\hat{\pi}_k}\Big(\frac{\partial}{\partial\theta_j}\boldsymbol{U}_i(\boldsymbol{\theta})^T\Big)_{j=1}^{p}\mathrm{I}_{ik} \xrightarrow{P} D(\boldsymbol{\theta}). \tag{6.9}$$

From (6.9) and (6.8) we get

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N(0, D(\boldsymbol{\theta})^{-1}\Sigma D(\boldsymbol{\theta})^{-1}).$$

□

From a sample of n households, the asymptotic variance of $\hat{\boldsymbol{\theta}}$ is estimated by the so-called sandwich estimator. The estimation of $D(\boldsymbol{\theta})$ is motivated by (6.9), i.e.

$$\hat{D}(\boldsymbol{\theta}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\frac{\xi_i}{\hat{\pi}_k}\Big(\frac{\partial}{\partial\theta_j}\boldsymbol{U}_i(\hat{\boldsymbol{\theta}})^T\Big)_{j=1}^{p}\mathrm{I}_{ik} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\frac{\xi_i}{\hat{\pi}_k}\Big(\frac{\partial^2}{\partial\theta_j\partial\theta_l}f(y_i,\hat{\boldsymbol{\theta}})\Big)_{j,l=1}^{p}\mathrm{I}_{ik}, \tag{6.10}$$

while the estimator of Σ is based on the empirical variance of the scores

$$\hat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\frac{\xi_i}{\hat{\pi}_i}\hat{\boldsymbol{U}}_i\hat{\boldsymbol{U}}_i^T\mathrm{I}_{ik} + \sum_{k=1}^{K}\hat{p}_k\frac{1-\pi_k}{\pi_k}\frac{\sum_{i=1}^{N}\xi_i(\hat{\boldsymbol{U}}_i - \hat{\boldsymbol{S}}_k)(\hat{\boldsymbol{U}}_i - \hat{\boldsymbol{S}}_k)^T\mathrm{I}_{ik}}{\sum_{i=1}^{N}\xi_i\mathrm{I}_{ik}}, \tag{6.11}$$

where $\hat{\boldsymbol{S}}_k$ is the average of $\hat{\boldsymbol{U}}_i$ in the stratum k.

Auxiliary Variables

As in the estimation of the expectation, auxiliary variables observed for the whole population can improve the estimator. Let us assume a vector of auxiliary variables $\boldsymbol{Z}_i$, which represents the components of $\boldsymbol{X}_i$, that are observed for the whole population, and also includes other variables correlated with those components of $\boldsymbol{X}_i$, that are observed only for the sampled individuals. We denote the sampling weights adjusted for the auxiliary variables and stratum as $\tilde{\pi}_i$ (see 2.9). The estimate $\tilde{\theta}$ is obtained by solving the estimating equations

$$\boldsymbol{V}(\boldsymbol{\theta}) = \sum_{i=1}^{N}\frac{\xi_i}{\tilde{\pi}_i}\boldsymbol{U}_i(\boldsymbol{\theta}) = \mathbf{0}. \tag{6.12}$$

Theorem 11. *Assume that vectors* $(Y_i, W_i, \boldsymbol{X_i}, \boldsymbol{Z_i}, \xi_i)$ *are iid and that* $\mathrm{E}\,Y_i = g^{-1}(\boldsymbol{x}_i^T\boldsymbol{\theta})$ *for some known link function* g. *Assume that* $\mathrm{var}\,Y_i < \infty$ *and* $\mathrm{var}\,Z_{ij} < \infty$ *for each component* $j = 1, 2, \ldots, s$. *Let* ξ_i *be independent of* Y_i, $\boldsymbol{Z_i}$ *and* $\boldsymbol{X_i}$ *given* W_i, *for* $i = 1, 2, \ldots, N$. *Then the following holds:*

$$\sqrt{N}(\tilde{\theta} - \theta) \xrightarrow{d} N(0, D(\boldsymbol{\theta})^{-1}\Sigma_z D(\boldsymbol{\theta})^{-1}), \tag{6.13}$$

where

$$\Sigma_z = J(\boldsymbol{\theta}) + \sum_{k=1}^{K}p_k\frac{1-\pi_k}{\pi_k}J_k(\boldsymbol{\theta}) - C^TV^{-1}C, \tag{6.14}$$

and

$$C = \sum_{k=1}^{K}p_k(1-\pi_k)\mathrm{cov}_k(\boldsymbol{Z}_i, \boldsymbol{U}_i), \quad V = \sum_{k=1}^{K}p_k\pi_k(1-\pi_k)\mathrm{var}_k\boldsymbol{Z}_i.$$

Proof Follows exactly the same steps as theorems 10 and 2.

□

The variance Σ_z (6.14) can be estimated similarly to Σ (6.10, 6.11), with the usual estimators of covariance and variance matrices $\widehat{\text{cov}}_k(\boldsymbol{Z}, \boldsymbol{U})$ and $\widehat{\text{var}}_k \boldsymbol{Z}$ for $k = 1, \ldots, K$.

As before, we have $\Sigma \geq \Sigma_z$, which means that the use of the auxiliary variables cannot increase the asymptotic variance of the estimator. From the form of Σ_z we can see that the amount by which the asymptotic variance decreases depends on the correlation of $\boldsymbol{Z}_i$ with the scores $\boldsymbol{U}_i$ within the strata. The transformation of the auxiliary variable $\boldsymbol{Z}_i$ that maximizes the correlation with the scores is the conditional mean value of scores, given $\boldsymbol{Z}_i$. Together with the reasoning explained in section 2.4, it implies that the optimal transformation of the auxiliary variables is $\boldsymbol{z}_i^{opt} = \frac{1}{\pi_k} \mathrm{E}\left(\boldsymbol{U}_i | \boldsymbol{Z_i} = \boldsymbol{z}_i\right)$, for an observation from the stratum k ($W_i = k$). While the sampling probabilities π_k are known, the conditional mean value of scores is obviously generally unknown. Breslow et al. [3] recommend to use the "plug-in" method to estimate the scores, suggested by Kulich and Lin [9]. The steps are as follows:

- Develop regression models using inverse probability weighted estimation (6.5) to predict each variable observed only for the sampled individuals given the variables $\boldsymbol{z}_i$ observed for the whole population.
- Use these models to predict $\hat{\boldsymbol{x}}_i = \mathrm{E}\left(\boldsymbol{X}_i | \boldsymbol{Z_i} = \boldsymbol{z}_i\right)$ for all individuals from the population. The fully observed variables will be used in their original form.
- Fit the model of interest to the whole population using the values $\hat{\boldsymbol{x}}_i$.
- Based on this model, estimate the scores $\boldsymbol{U}_i$ and divide them by the corresponding sampling probabilities π_k. These are the estimates of the optimal transformation of the auxiliary variables, denoted as $\hat{\boldsymbol{z}}_i^{opt}$.
- Estimate θ using weights adjusted for the auxiliary variables $\hat{\boldsymbol{z}}_i^{opt}$.

Breslow et al. [3] note that this method is likely to be most useful when only one or two regressors are not observed for the whole population.

6.3 Cluster Sampling

Now we assume the cluster sampling as defined in chapter 3. The marginal density $f(y|\boldsymbol{x})$ resembles (3.1)

$$f(y|\boldsymbol{x}) = \frac{1}{\mathrm{E}\left[M|\boldsymbol{x}\right]} \int\int m f(y|\boldsymbol{x}, m, \boldsymbol{b}) f(m, \boldsymbol{b}|\boldsymbol{x}) \, d\boldsymbol{b} d\mu(m). \tag{6.15}$$

The aim is to estimate the parameter $\boldsymbol{\theta}$ in the marginal model for the mean value (6.1). We adopt the pseudo-likelihood approach and, analogously to the estimation of the expectation, we modify the estimating equations (6.4) as follows

$$\sum_{i=1}^{n}\sum_{r=1}^{M_i}\frac{\xi_{ir}}{\pi_{ir}}\boldsymbol{U}_{ir}(\boldsymbol{\theta})=\mathbf{0}. \tag{6.16}$$

The estimate of the parameter $\boldsymbol{\theta}$, denoted as $\hat{\boldsymbol{\theta}}$, is obtained by solving (6.16).

Theorem 12. *Let $M_i, i=1,2,\ldots,n$, be iid random variables. Let $Y_{ir}, i=1,2,\ldots,n$ and $r=1,2,\ldots,M_i$, be random variables with $\mathrm{E}\,Y_i = g^{-1}(\boldsymbol{x}_i^T\boldsymbol{\theta})$ for some known link function g. Let $(Y_{i1}, Y_{i2}\ldots Y_{iM_i})$ and also $(\xi_{i1}, \xi_{i2}\ldots\xi_{iM_i})$ be independent random vectors for $i=1,2,\ldots n$. We assume $\sum_{r=1}^{M_i}\xi_{ir}=m_i$ and ξ_{ir} is independent from Y_{ir} and $\boldsymbol{X}_{ir}$, given M_i. Then*

$$\sqrt{n}(\hat{\theta}-\theta)\xrightarrow{d}N(0, D(\boldsymbol{\theta})^{-1}\Sigma_{\hat{\theta}}D(\boldsymbol{\theta})^{-1}),$$

where

$$\Sigma_{\hat{\theta}}=\frac{1}{\mu^2}\mathrm{E}\,M_i^2\bar{\boldsymbol{U}}_i\bar{\boldsymbol{U}}_i^T, \qquad \bar{\boldsymbol{U}}_i=\frac{1}{m_i}\sum_{r=1}^{M_i}\xi_{ir}\boldsymbol{U}_{ir} \tag{6.17}$$

and $D(\boldsymbol{\theta})=(-\mathrm{E}\,\frac{\partial}{\partial\theta_j}\boldsymbol{U}_{ir}^T)_{j=1}^{p}$.

Proof Follows the same steps as the proofs of Theorems 3 and 10.

□

From a sample of n households, $\Sigma_{\hat{\theta}}$ can be estimated as

$$\hat{\Sigma}_{\hat{\theta}}=\frac{1}{\hat{\mu}^2}\frac{1}{n}\sum_{i=1}^{n}M_i^2\bar{\boldsymbol{U}}_i(\hat{\boldsymbol{\theta}})\bar{\boldsymbol{U}}_i(\hat{\boldsymbol{\theta}})^T, \quad \text{where} \quad \hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}M_i. \tag{6.18}$$

The estimation of $D(\boldsymbol{\theta})$ follows the logic of the inverse probability weighting

$$\hat{D}(\boldsymbol{\theta})=-\frac{\sum_{i=1}^{n}\sum_{r=1}^{M_i}\frac{\xi_{ir}}{\pi_{ir}}\left(\frac{\partial}{\partial\theta_j}\boldsymbol{U}_{ir}(\hat{\boldsymbol{\theta}})^T\right)_{j=1}^{p}}{\sum_{i=1}^{n}M_i}. \tag{6.19}$$

Examples for linear and logistic regression models will be shown below.

Auxiliary Variables

Auxiliary variables can be used to improve the efficiency of the estimator. Let us assume a vector of auxiliary variables $\boldsymbol{Z}_{ir}$ for the rth member of the ith household. It represents the components of $\boldsymbol{X}_{ir}$ that are observed for all members of the households, and it also includes other variables correlated with those components of $\boldsymbol{X}_{ir}$ that are observed only for

the sampled individuals. We assume that each component of $\boldsymbol{Z}_{ir}$ has a finite variance and that $(Y_{ir}, \boldsymbol{Z}_{ir})$ are independent within households. To ensure that the use of the auxiliary variables will not lead to a loss of efficiency, we utilize centered auxiliary variables

$$\boldsymbol{H}_{ir} = \boldsymbol{Z}_{ir} - \bar{\boldsymbol{Z}}_i, \quad \text{where} \quad \bar{\boldsymbol{Z}}_i = \frac{1}{M_i}\sum_{r=1}^{M_i} \boldsymbol{Z}_{ir}.$$

The estimated sampling probabilities are denoted as $\tilde{\pi}^a_{ir}$ and the parameter θ is estimated by solving the estimating equations

$$\sum_{i=1}^{n}\sum_{r=1}^{M_i} \frac{\xi_{ir}}{\tilde{\pi}^a_{ir}} \boldsymbol{U}_{ir} = \boldsymbol{0}. \tag{6.20}$$

An analogy of theorem 9 holds.

Theorem 13. *Let $M_i, i = 1, 2, \ldots, n$, be iid random variables. Let $Y_{ir}, i = 1, 2, \ldots, n$ and $r = 1, 2, \ldots, M_i$, be random variables with $\mathrm{E}\, Y_i = g^{-1}(\boldsymbol{x}_i^T\boldsymbol{\theta})$ for some known link function g. Let $(Y_{i1}, Y_{i2} \ldots Y_{iM_i})$ and also $(\xi_{i1}, \xi_{i2} \ldots \xi_{iM_i})$ be independent random vectors for $i = 1, 2, \ldots n$. We assume $\sum_{r=1}^{M_i} \xi_{ir} = m_i$ and ξ_{ir} is independent from Y_{ir} and $\boldsymbol{X}_{ir}$, given M_i. Then*

$$\sqrt{n}(\tilde{\theta}_a - \theta) \xrightarrow{d} N(0, D(\boldsymbol{\theta})^{-1}\Sigma_a D(\boldsymbol{\theta})^{-1}),$$

where

$$\Sigma_a = \Sigma_{\hat{\theta}} + \boldsymbol{q}_\delta^T J_\delta^{-1} \mathrm{E}\left[\mathrm{var}_i \boldsymbol{H}_{ir}\right] J_\delta^{-1}\boldsymbol{q}_\delta - 2\frac{1}{\mu}\mathrm{E}\left[M_i \mathrm{cov}_i(\boldsymbol{U}_{ir}, \boldsymbol{H}_{ir})\right] J_\delta^{-1}\boldsymbol{q}_\delta, \tag{6.21}$$

for

$$\boldsymbol{q}_\delta = \frac{1}{\mu}\mathrm{E}\left[(M_i - 1)\mathrm{cov}_i(\boldsymbol{H}_{ir}, \boldsymbol{U}_{ir})\right], \qquad J_\delta = \mathrm{E}\left[\left(1 - \frac{1}{M_i}\right)\mathrm{var}_i \boldsymbol{H}_{ir}\right]$$

and $D(\boldsymbol{\theta}) = (-\mathrm{E}\,\frac{\partial}{\partial\theta_j}\boldsymbol{U}_{ir}^T)_{j=1}^p$.

Proof Follows the same steps as the proofs of Theorems 9 and 10.

□

The estimator of the asymptotic variance Σ_a is constructed analogously to (6.18) and (5.17).

As with stratified sampling, from the form of Σ_a we can see that the amount by which the variance of the estimator decreases depends on the correlation of $\boldsymbol{H}_{ir}$ with the scores $\boldsymbol{U}_{ir}$ within the household. The transformation of the auxiliary variable $\boldsymbol{H}_{ir}$ that maximizes the correlation with the scores is the conditional mean value of scores, given $\boldsymbol{H}_{ir}$. Based on the same reasoning as in section 5.1 we conclude that the optimal transformation of the original auxiliary variable $\boldsymbol{Z}_{ir}$ is $\boldsymbol{h}_{ir}^{opt} = M_i \mathrm{E}\,(\boldsymbol{U}_{ir}|\boldsymbol{H}_{\boldsymbol{ir}} = \boldsymbol{h}_{ir})$, where $\boldsymbol{H}_{ir} = \boldsymbol{Z}_{ir} - \bar{\boldsymbol{Z}}_i$. Therefore, the "plug-in" method to find the estimates of the scores and use them

as auxiliary variables can also be applied here. We will illustrate the whole estimation procedure by a few examples in the next section.

So far we have focused on the situation when the response variable was known for all individuals, and some of the explanatory variables were missing for a part of the population. However, we could also face the opposite situation; the explanatory variables are known, but the response is observed only for the subsample. In that case, if the available auxiliary information is correlated with the response, the analogous estimation procedure can be applied. The only difference is that this time the expectation of the response, given the auxiliary variables, must be estimated, while the explanatory variables are used in their original form. The predicted response is then employed in the "plug-in" method to find the estimate of the optimal auxiliary variable $\boldsymbol{h}_{ir}^{opt}$.

6.4 Cluster Sampling - Examples

Linear Model

We are interested in the relationship between the variable Y and two explanatory variables x and w. We assume that

$$\mathrm{E}\, Y = \beta_0 + \beta_1 x + \beta_2 w, \tag{6.22}$$

and the objective is to estimate the parameter $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$.

There are two kinds of households; small (of size $M = 3$) and large (of size $M = 6$), equally represented in the population, $p_3 = \mathrm{P}(M = 3) = p_6 = \mathrm{P}(M = 6) = 0.5$. The value of the target variable Y for the rth member in the ith household is given by the following relationship

$$Y_{ir} = 3x_{ir} + 3\mathrm{I}[M_i = 6]x_{ir} + 1.5w_{ir} + 1.5\mathrm{I}[M_i = 6]w_{ir} + \delta_i + \epsilon_{ir}. \tag{6.23}$$

This implies that the relationship between the response and the explanatory variables depends on the size of household. The parameter δ_i represents the correlation within the households and the parameter ϵ_{ir} is a random error. The variables Y_{ir} and w_{ir} are known for all the members of the households, while the variable x_{ir} is known only for one randomly selected member of the household. However, there is an auxiliary variable

$$Z_{ir} = x_{ir} + \eta_{ir},$$

which is known for all individuals.

Even though the relationship between Y and x depends on the household size, we are interested in the marginal model (6.22), i.e. without conditioning on the household size. From (6.23) we can see that $\beta_0 = 0$. Weighting by the household size we obtain

$$\beta_1 = \frac{3 * 3 + 6 * 6}{9} = 5 \tag{6.24}$$

$$\beta_2 = \frac{1.5 * 3 + 3 * 6}{9} = 2.5. \tag{6.25}$$

The estimating equations (6.16) can be directly used to estimate the parameter $\boldsymbol{\beta}$ without taking into account the auxiliary variable. Let us denote this estimate as $\hat{\boldsymbol{\beta}}$. The asymptotic variance matrix of the estimator $\hat{\boldsymbol{\beta}}$ was presented in Theorem 12. It can be estimated by replacing the unknown quantities by their estimates. Specifically, the scores $\boldsymbol{U}_{is} = (U_{is1}, U_{is2}, U_{is3})$ are estimated as

$$\begin{aligned}
\hat{U}_{is1} &= Y_{is} - \hat{\beta}_0 - \hat{\beta}_1 x_{is} - \hat{\beta}_2 w_{is} \\
\hat{U}_{is2} &= (Y_{is} - \hat{\beta}_0 - \hat{\beta}_1 x_{is} - \hat{\beta}_2 w_{is}) x_{is} \\
\hat{U}_{is3} &= (Y_{is} - \hat{\beta}_0 - \hat{\beta}_1 x_{is} - \hat{\beta}_2 w_{is}) w_{is},
\end{aligned} \tag{6.26}$$

where s denotes the selected member from the household i, i.e. the member for whom x_{is} is known. From a sample of n households,

$$\hat{\Sigma}_{\hat{\beta}} = \frac{1}{\hat{\mu}^2} \frac{1}{n} \sum_{i=1}^{n} M_i^2 \hat{\boldsymbol{U}}_{is} \hat{\boldsymbol{U}}_{is}^T, \quad \text{where} \quad \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} M_i \tag{6.27}$$

$$\hat{D}(\boldsymbol{\beta}) = \frac{1}{n\hat{\mu}} X^T \mathrm{Diag}\Big\{ \frac{1}{\pi_{i1}} \Big\}_{i=1}^{n} X, \quad \text{where} \quad X = (\mathbf{1}, \boldsymbol{x}, \boldsymbol{w}). \tag{6.28}$$

Then the asymptotic variance of $\hat{\boldsymbol{\beta}}$ is estimated as $\frac{1}{n}\hat{D}(\boldsymbol{\beta})^{-1}\hat{\Sigma}_{\hat{\beta}}\hat{D}(\boldsymbol{\beta})^{-1}$.

To incorporate the auxiliary information, we need to find an appropriate transformation of the auxiliary variable Z. We use the "plug-in" method described in the previous section. First, the expectation of the variable X_{ir}, given z_{ir} and w_{ir}, is estimated from the linear model

$$\mathrm{E}\, X_{ir} = \alpha_0 + \alpha_1 z_{ir} + \alpha_2 w_{ir}$$

and using the weights unadjusted for the auxiliary variables. Based on the estimate of the parameter $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2)$, the values x_{ir} for all $i = 1, \ldots, n$, $r = 1, \ldots, M_i$ are predicted and denoted $\tilde{x}_{ir}$. The scores are estimated as follows

$$\begin{aligned}
\tilde{U}_{ir1} &= Y_{ir} - \hat{\beta}_0 - \hat{\beta}_1 \tilde{x}_{ir} - \hat{\beta}_2 w_{ir} \\
\tilde{U}_{ir2} &= (Y_{ir} - \hat{\beta}_0 - \hat{\beta}_1 \tilde{x}_{ir} - \hat{\beta}_2 w_{ir}) \tilde{x}_{ir} \\
\tilde{U}_{ir3} &= (Y_{ir} - \hat{\beta}_0 - \hat{\beta}_1 \tilde{x}_{ir} - \hat{\beta}_2 w_{ir}) w_{ir}.
\end{aligned}$$

To satisfy condition (5.3), the estimated scores should be centered, i.e. for $j = (1, 2, 3)$

$$H_{irj} = \tilde{U}_{irj} - \bar{U}_{ij}, \quad \text{where} \quad \bar{U}_{ij} = \frac{1}{M_i} \sum_{r=1}^{M_i} \tilde{U}_{irj}.$$

We write $\boldsymbol{H}_{ir} = (H_{ir1}, H_{ir2}, H_{ir3})$.

The sampling probabilities are estimated from the logistic regression model

$$\log\left(\frac{\pi_{ir}}{1 - \pi_{ir}}\right) = \gamma_0 + \gamma_1 \mathrm{I}[M_i = 6] + \boldsymbol{\gamma}_2^T M_i \boldsymbol{H}_{ir}.$$

Table 6.1: Example of linear model; Results of simulation

Estimator	Parameter	Average estimate	Variance of estimator*		Coverage of CI (%)
			empirical	estimate**	
	β_0	0.010	2267	2337	95.6
$\hat{\beta}$	β_1	4.993	9.685	9.395	94.2
(no aux. info)	β_2	2.498	6.831	6.748	94.5
	β_0	0.008	1818	1868	95.1
$\tilde{\beta}$	β_1	4.997	6.478	5.938	93.4
(x part. missing)	β_2	2.500	4.238	3.917	93.5
	β_0	-0.002	1529	1636	95.4
$\check{\beta}$	β_1	4.997	4.635	4.471	93.6
(Y part. missing)	β_2	2.500	2.940	2.770	94.6
	β_0	0.007	1146		
Full population	β_1	4.998	3.595		
	β_2	2.500	1.995		

*Estimate of the asymptotic variance of $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ for $\hat{\boldsymbol{\beta}}$, for $\tilde{\boldsymbol{\beta}}$ and $\check{\boldsymbol{\beta}}$ analogously.
**average of estimates

The desired estimate of the parameter $\boldsymbol{\beta}$ is obtained by solving the estimating equations (6.20), and denoted as $\tilde{\boldsymbol{\beta}}$.

The variance matrix of the estimator $\tilde{\boldsymbol{\beta}}$ was presented in Theorem 13. Again, it can be estimated by replacing the unknown quantities by their estimates. The estimation of the scores is similar to (6.26), with $\hat{\boldsymbol{\beta}}$ replaced by $\tilde{\boldsymbol{\beta}}$. The rest of the procedure is analogous to (6.27), (6.28) and (5.10).

Simulation To illustrate this method, we performed a simulation. It was conducted as follows. We assumed that there are 1000 households, and assigned the size of 3 or 6 members to each of them randomly with equal probability. Then the data were generated from the following distributions

$$x \sim N(0, 400) \quad w \sim N(0, 400) \quad \delta \sim N(0, 900) \quad \epsilon \sim N(0, 400) \quad \eta \sim N(0, 100).$$

The variable Y was calculated according to (6.23). From each household, one member was selected at random. The estimates of the parameter $\boldsymbol{\beta}$ were calculated as described above. This procedure was repeated 1000 times. The results can be found in Table 6.1. We can see that the average estimate of the parameter $\boldsymbol{\beta}$ is very close to its true value in all cases. Also, the estimates of the asymptotic variance of the parameter estimators are very close to the empirical variance. When the auxiliary information was employed, the asymptotic variance of the normalized estimate $\hat{\beta}_1$ decreased from 9.685 to 6.478 (by 33 %) and the

variance of the normalized estimate $\hat{\beta}_2$ decreased from 6.831 to 4.238 (by 38 %), which is a noticeable improvement. Indeed, the amount by which the variance decreases depends on the correlation between x and Z, i.e. the variance of the variable η in this specific case.

We also inspected the situation when the response is available only for the subsample. If we considered the same example and assumed that the response variable was known only for the sampled individuals, the variance of the estimator taking into account the auxiliary information was not better than the variance of the estimator which does not consider the auxiliary data (results not shown). This is not surprising since the variable Z was correlated strongly with the variable x, but apart from that it did not contain any additional information about the response.

Then we changed the setting of the simulation and assumed that the explanatory variables were available for the whole population, the response was known only for a subsample, but the auxiliary variable was correlated with the response. More precisely, we considered the auxiliary variable

$$Z_{ir} = Y_{ir} + \eta_{ir}, \qquad \text{where} \qquad \eta_{ir} \sim N(0, 500).$$

The results are presented in Table 6.1 ($\check{\beta}$). The improvement in the asymptotic variances was remarkable.

Logistic Regression

The following example is in principle very similar to the previous one. We consider an analogy to a logistic regression model with one explanatory variable. More precisely, we are interested in the relationship between a dichotomous variable Y and a continuous explanatory variable x. We assume the following model

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

and the objective is to estimate the parameter $\boldsymbol{\beta} = (\beta_0, \beta_1)$.

For the rth member in the ith household, the variable Y_{ir} is equal to 1 with probability p_{ir}, otherwise it is equal to 0. The probability p_{ir} is given by the following relationship

$$\log\left(\frac{p_{ir}}{1-p_{ir}}\right) = 3x_{ir} + 3\mathrm{I}[M_i = 6]x_{ir} + \delta_i. \tag{6.29}$$

Similarly as in the previous example, this indicates that the relationship between the response and the explanatory variable in fact depends on the size of household. The variable Y_{ir} is known for all members of the households, while the variable x_{ir} is known only for one randomly selected member of the household. The auxiliary variable

$$Z_{ir} = x_{ir} + \eta_{ir}$$

is known for all individuals.

We can see from (6.29) that $\beta_0 = 0$. To calculate the true value of the parameter β_1 in the marginal model is not as straightforward as in the case of the linear model, that is why it was obtained via simulations as $\beta_1 = 2.961$. When the auxiliary variable is ignored the parameter $\boldsymbol{\beta}$ is estimated based on estimating equations (6.16). To determine its variance matrix, we estimate the scores $\boldsymbol{U}_{is} = (U_{is1}, U_{is2})$

$$\hat{U}_{is1} = Y_{is} - \hat{p}_{is}$$
$$\hat{U}_{is2} = (Y_{is} - \hat{p}_{is})x_{is},$$

where

$$\hat{p}_{is} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_{is})}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_{is})}$$

and s denotes the selected member from household i. We get

$$\hat{\Sigma}_{\hat{\beta}} = \frac{1}{\hat{\mu}^2}\frac{1}{n}\sum_{i=1}^{n} M_i^2 \hat{\boldsymbol{U}}_{is}\hat{\boldsymbol{U}}_{is}^T, \quad \text{where} \quad \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} M_i, \tag{6.30}$$

and

$$\hat{D}(\boldsymbol{\beta}) = \frac{1}{n\hat{\mu}} X_p^T \mathrm{Diag}\Big\{\frac{1}{\pi_{i1}}\Big\}_{i=1}^{n} X, \tag{6.31}$$

where $X_p = (\hat{p}_{is}(1-\hat{p}_{is}), \hat{p}_{is}(1-\hat{p}_{is})x_{is})_{i=1}^n$ and $X = (\mathbf{1}, \boldsymbol{x})$. Then the estimate of the asymptotic variance $\widehat{\mathrm{var}}\hat{\boldsymbol{\beta}}$ is $\frac{1}{n}\hat{D}(\boldsymbol{\beta})^{-1}\hat{\Sigma}_{\hat{\beta}}\hat{D}(\boldsymbol{\beta})^{-1}$.

The "plug-in" method to find the appropriate transformation of the auxiliary variable Z can be applied. The expectation of the variable X_{ir}, given z_{ir}, is estimated from the linear model

$$\mathrm{E}\, X_{ir} = \alpha_0 + \alpha_1 z_{ir},$$

using the weights unadjusted for the auxiliary variable. Based on the estimate of the parameter $\boldsymbol{\alpha} = (\alpha_0, \alpha_1)$, the values x_{ir} for all $i = 1, \ldots, n$, $r = 1, \ldots, M_i$ are predicted and denoted $\tilde{x}_{ir}$. The scores are estimated as follows

$$\tilde{U}_{ir1} = Y_{ir} - \tilde{p}_{ir}$$
$$\tilde{U}_{ir2} = (Y_{ir} - \tilde{p}_{ir})\tilde{x}_{ir},$$

where

$$\tilde{p}_{ir} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 \tilde{x}_{ir})}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 \tilde{x}_{ir})}.$$

As in the previous example, the centered scores are denoted $\boldsymbol{H}_{ir} = (H_{ir1}, H_{ir2})$. The sampling probabilities are estimated from the logistic regression model

$$\log\Big(\frac{\pi_{ir}}{1-\pi_{ir}}\Big) = \gamma_0 + \gamma_1 \mathrm{I}[M_i = 6] + \boldsymbol{\gamma}_2^T M_i \boldsymbol{H}_{ir}.$$

and the estimate of the parameter $\boldsymbol{\beta}$ is obtained by solving the estimating equations (6.20). The variance matrix estimation is analogous to the procedure described above.

Table 6.2: Example of logistic regression model; Results of simulation

Estimator	Parameter	Average estimate	Variance of estimator*		Coverage of CI (%)
			empirical	estimate**	
$\hat{\beta}$	β_0	-0.003	6.675	6.705	94.8
(no aux. info)	β_1	2.976	34.241	35.899	95.7
$\tilde{\beta}$	β_0	-0.001	2.713	2.690	94.3
(x part. missing)	β_1	2.965	11.078	10.873	94.6
$\check{\beta}$	β_0	< 0.001	5.254	5.127	94.5
(Y part. missing)	β_1	2.9625	30.872	29.799	94.4
Full population	β_0	-0.001	2.679		
	β_1	2.963	9.991		

*Estimate of the asymptotic variance of $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ for $\hat{\boldsymbol{\beta}}$, for $\tilde{\boldsymbol{\beta}}$ and $\check{\boldsymbol{\beta}}$ analogously.
**average of estimates

Simulation The results are illustrated by a simulation. Again we had 1000 households with randomly assigned size of 3 or 6 members. Then the variables were generated from the following distributions

$$x \sim N(0, 0.4) \quad \eta \sim N(0, 0.004) \quad \delta \sim N(0, 4) \quad \epsilon \sim N(0, 1).$$

The variable Y was calculated from (6.29). From each household, one member was selected at random and the estimates of the parameter $\boldsymbol{\beta}$ and their variance were calculated as described above. This procedure was repeated 1000 times. The results are summarized in Table 6.2. The average estimate of the parameter $\boldsymbol{\beta}$ was very close to its true value in both cases. The estimate of the asymptotic variance and the empirical variance of the estimator were very similar. When the auxiliary information was considered, the estimated variance of the normalized estimate $\hat{\beta}_1$ decreased from 34.241 to 11.078, which is very close to the variance of the estimator based on the full population. As we mentioned in the previous example, the improvement in the variance depends on the correlation between x and Z, which was set up rather high.

The simulation of the scenario when the response is available only for the subsample showed results similar to the linear model. If the setting was left unchanged but in addition we assumed that the response variable was available only for the sampled individuals, the auxiliary information did not improve the estimation (results not shown). We also considered the situation where the explanatory variables were available for the whole population and the response only for the subsample. The auxiliary variable Z_{ir} had a binary distribution with probability of success

$$p_{ir} = 0.9 * \mathrm{I}[Y_{ir} = 1] + 0.1 * \mathrm{I}[Y_{ir} = 0].$$

The results are displayed in Table 6.2 ($\check{\beta}$).

In summary, we can say that the estimating procedure performed very well in both examples. The results supported the statement that both estimators are asymptotically unbiased. We could also see that the asymptotic variance was very close to the empirical variance, which shows that the estimator converges to its asymptotic distribution with an acceptable speed. We also inspected examples where the correlation of the explanatory (x) and the auxiliary (Z) variable was very weak. In no case was the asymptotic variance increased when the estimation procedure taking into account the auxiliary information was applied.

6.5 Application to Project ACCEPT data

Unfortunately, the data from Project ACCEPT are not very appropriate for a useful application of the method, since only age and gender are known for the whole population. We will only show a simple example for illustrative purposes.

The study design and the data were described in section 5.4. Now we would like to know whether the HIV testing history depends on the social norms score and gender. We assume the following model

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \mathrm{I}[g = \text{"Male"}] + \beta_2 sns,$$

where p is the probability that an individual has been tested for HIV, sns is the social norms score and g represents gender. To estimate the parameters, exactly the same steps as shown in the previous examples were followed. The only difference was that in order to create an appropriate transformation of the auxiliary variables, we had to predict not only the social norms score, but also the variable representing HIV testing history. While for social norms score the linear regression model was used, HIV testing history was predicted based on the logistic regression model. Both prediction models included age and gender as the explanatory variables. The results of both estimation procedures (taking and not taking into account age as the auxiliary variable) are presented in Table 6.3.

We can see that in this case, the use of the auxiliary information improved the asymptotic variance of the parameter by a small, but not negligible amount. The asymptotic variance of the normalized estimate of β_1, representing gender, decreased from 46.148 to 39.867 (by 13.5 %). For the other two components of $\boldsymbol{\beta}$, the gain was smaller. This result was expected since not only the explanatory variable but also the response variable were not available for the whole population.

Significance testing, i.e. testing of $H_0 : \beta_j = 0$, was performed using the Wald test. It is based on the test statistic $\frac{\hat{\beta}_j}{\sqrt{\widehat{\mathrm{var}}\hat{\beta}_j}}$, which has an asymptotically normal distribution. Both estimating procedures lead to the same conclusion. While social norms score had no statistically significant effect, the effect of gender was highly significant. In general, women had $\exp(1.234) = 3.435$ (or $\exp(1.328) = 3.773$, based on $\hat{\beta}$) times higher odds of being tested for HIV than men.

Table 6.3: Logistic regression model for HIV testing history

Estimator	Parameter	Param. estimate	As. variance*	p-value**
$\hat{\beta}$	Intercept	-0.191	701.771	0.715
	Gender (Male)	-1.328	46.148	< 0.001
	Soc. norms	-0.025	14.318	0.739
$\tilde{\beta}$	Intercept	-0.187	663.391	0.712
	Gender (Male)	-1.234	39.867	< 0.001
	Soc. norms	-0.023	13.461	0.754

*Estimate of the asymptotic variance of $\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})$ and $\sqrt{n}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})$, respectively, $n = 2586$.
**Based on the Wald test.

6.6 Discussion

The method of parameter estimation described in this chapter is likely to be most useful when only one or two explanatory variables are missing and they are highly correlated with the auxiliary variables. Such auxiliary information can be for example a less precise measurement of the same parameter, screening results, etc. If the auxiliary variables do not contain too much information about the explanatory variables, we cannot expect that they would lead to increased precision in the estimation.

We mainly addressed the situation when the response variable was known for the whole population and some of the explanatory variables were available only for the subsample. We also touched upon the opposite scenario when the explanatory variables are observed, but the response is known only for the sampled individuals. We mentioned that the analogous estimation method can be applied. However, it is important to consider carefully the nature of the auxiliary information. Indeed, it should not be the explanatory variables or variables closely related to them, but rather variables bearing the information related to the response. For example in the context of Project ACCEPT, when the response variable known for the sampled individuals would be the result of HIV testing and the explanatory variables would be region, gender, age and intervention, the auxiliary information could be an extensive questionnaire including information about relatives' HIV status, previous and present sexual behaviour, drug and alcohol use and other risk factors providing a good guess as to how likely the interviewee is to be HIV positive, even without the blood test. In general, the auxiliary variables are likely to improve the estimation only if they are highly correlated with the response. Again, it is often some less precise or preliminary measurement of the response.

We could also consider the third scenario, when the response as well as some explanatory variables are known only for the sampled individuals. The same approach to the estimation of the parameters would be applicable here too. However, although it cannot be ruled out, it is unlikely that it would lead to a substantial improvement of the estimator. It might work in a situation where some auxiliary variables would be closely related to the explanatory variables while others would be correlated with the response, but in that case we would rather recommend to reconsider the design of the study.

Chapter 7

Summary and Conclusion

In this thesis we presented methods of parameter estimation under two-phase stratified sampling and cluster sampling. In contrast to classical sampling theory, we did not deal much with finite population parameters, but rather focused on model parameters inference, which is more appropriate in scientific applications. However, we had to consider the sampling schemes employed and consequently incorporated much of the survey sampling theory as well. Therefore, the methods used for parameter estimation could be considered as a combination or unification of the two approaches.

In stratified sampling, we addressed the situation where the full population is divided into strata and subsampled within each stratum, not necessarily with the same sampling probabilities. The target variable is then observed only for the selected individuals. We presented the mean value estimation, including the statistical properties of the estimator, and showed how this estimation can be improved if some auxiliary information, correlated with the target variable, is observed for the whole population.

This led us to an idea that the same approach might be adapted for the case of cluster sampling, although the two situations are not completely analogous. While in the case of stratified sampling the subsamples are drawn within a small number of strata, in the case of cluster sampling the subsampling is performed within a "large" number of clusters, where "large" means that the number of clusters increases with increasing size of the population. Nevertheless, the estimation procedure (also using auxiliary information) can be modified for this sampling scheme and its detailed description together with simulations supporting the theoretical results were presented. We considered two scenarios: when the auxiliary variable is available for all households (including non-selected), and when the auxiliary variable is known for all members (including non-selected) from the sampled households.

We addressed not only the estimation of the expectation, but also extended the method in the context of the GLM. We described in detail the estimation procedure which makes use of the auxiliary variables, including the derivation of the appropriate transformation of the auxiliary variable, and illustrated the process with several examples. The methodology can be easily extended to other types of regression models (censored data regression, quantile regression, etc.).

We have shown that the use of auxiliary information can never increase the variance of

the estimator. In the worst case, it will be equal to the variance of the estimator which does not incorporate the auxiliary variables. The extent to which the precision can be improved depends on the correlation between the auxiliary variable and the target variable (when estimating the expectation) or the scores (in GLM). As we have seen in some examples, this is not the same as an "obvious association", which might be observed in the data. In the case of cluster sampling, the critical property is the correlation within the cluster, i.e. the ability of the auxiliary variable to distinguish the observations within the cluster. This ability might be limited, for example, for dichotomous variables.

The basic assumption was that the marginal inference was appropriate to answer the scientific question. Indeed, if the inference conditional on stratum or household, respectively, was more adequate, the researcher could in most situations apply the classical methods which do not take into account the sampling scheme. For example, in the case of cluster sampling, one might utilize the random effects model. Still, we believe that marginal inference is suitable in a lot of situations, an example of which is in Project ACCEPT.

The studied subject is related to a general missing data problem. We assumed that the data were missing by design, while they could be also missing by chance. Thus, as a further step, it would be of interest to study whether an extension of the presented methods under a more general missingness pattern would be possible. This would for example include the situation where several auxiliary variables are available, but none of them is known for the whole population. Instead, only a few of the auxiliary variables are observed for the non-sampled individuals.

Bibliography

[1] Project ACCEPT Study Group (2007) *Project ACCEPT: A Phase III Randomized Controlled Trial of Community Mobilization, Mobile Testing, Same-Day Results, and Post-Test Support for HIV in Sub-Saharan Africa and Thailand.* `http://www.cbvct.med.ucla.edu/protocol.pdf`

[2] Azzalini A. (2005) *The skew-normal distribution and related multivariate families.* Scand J Stat **32**, 159–188.

[3] Breslow N.E., Lumley T., Ballantyne C.M., et al (2009) *Improved Horvitz-Thompson Estimation of Model Parameters from Two-phase Stratified Samples: Applications in Epidemiology.* Stat Biosci. **1**, 1–32.

[4] Breslow N.E., Lumley T., Ballantyne C.M., et al (2009) *Using the whole cohort in the analysis of casecohort data.* Am J Epidemiol. **169**, 1398–1405.

[5] Cochran W. G. (1977) Sampling Techniques. John Wiley & Sons, New York.

[6] Fuller W. A. (2009) Sampling Statistics. John Wiley & Sons, Hoboken, New Jersey.

[7] Graubard B. I., Korn E. L. (2002) *Inference for Superpopulation Parameters Using Sample Surveys.* Statistical Science **17**, 73–96.

[8] Korn E. L., Graubard B. I. (1998) *Variance estimation for superpopulation parameters.* Statistica Sinica **8**, 1131–1151.

[9] Kulich M., Lin D.Y. (2004)*Improving the efficiency of relative-risk estimation in case-cohort studies.* J Am Stat Assoc. **99**, 832–844.

[10] Kulich M. (2006) *Project ACCEPT; Baseline Data documentation* Internal document of Project ACCEPT Statistical Center

[11] Lumley T. (2004) *Analysis of complex survey samples.* J Stat Softw **9**, 1–19.

[12] Lumley T. (2010) Complex Surveys: a guide to analysis using R. John Wiley & Sons, Hoboken, New Jersey.

[13] Neyman J. (1938) *Contribution to the theory of sampling human populations.* J Am Stat Assoc. **33**, 101–116.

[14] R Development Core Team (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org.

[15] Robins J. M., Rotnitzky A. and Zhao L.P. (1994) *Estimation of Regression Coefficients When Some Regressors Are Not Always Observed*. J Am Stat Assoc. **89**, 846–866.

[16] Särndal C. E., Swensson B. and Wretman J. (1991) Model Assisted Survey Sampling. Springer-Verlag, New York.

[17] Šedová M., Kulich M. (2007) *Statistical Methods for Analysis of Survey Data.* in WDS'07 Proceedings of Contributed Papers: Part I – Mathematics and Computer Sciences Prague, Matfyzpress, 181 – 186.

[18] Šedová M., Kulich M. (2009) *Maximálně věrohodné odhady a lineární regrese ve výběrových šetřeních*, ROBUST 2008, Sborník prací 15. letní školy JČMF, Praha, 409–416.

[19] Šedová M., Kulich M. (2010) *Dvoustupňové náhodné výběry ve výběrových šetřeních*, ROBUST 2010, Sborník prací 16. zimní školy JČMF, Praha, 109–114.

[20] White J.E. (1982) *A two stage design for the study of the relationships between rare exposure and a rare disease*. Am J Epidemiol. **115**, 119–128.