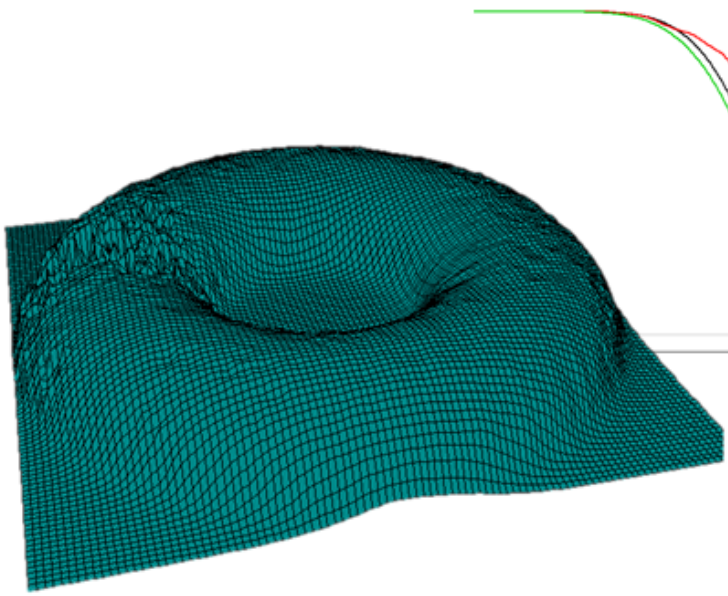
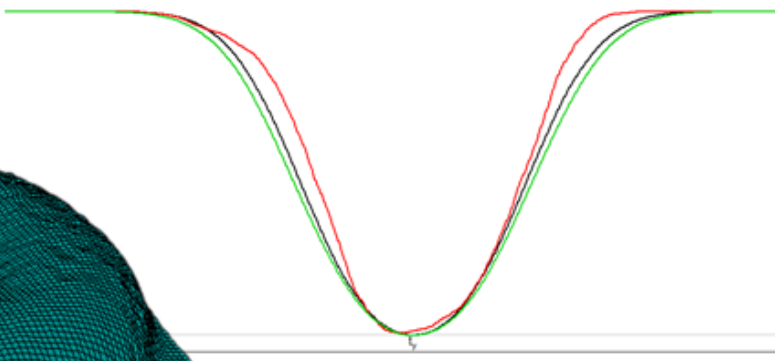
Stockholm University

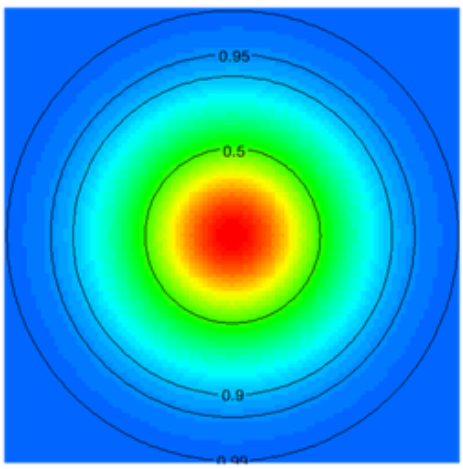
Essays on Sample Surveys

Design and Estimation

Edgar Bueno







Doctoral Thesis in Statistics at Stockholm University, Sweden 2020

Essays on Sample Surveys

Design and Estimation

Edgar Bueno

Academic dissertation for the Degree of Doctor of Philosophy in Statistics at Stockholm University to be publicly defended on Friday 4 December 2020 at 13.00 in Nordenskiöldsalen, Geovetenskapens hus, Svante Arrhenius väg 12.

Abstract

Sampling is a core stage in every survey. A sampling design carefully elaborated may imply not only a more accurate estimation of the parameters of interest, but also a reduction in the required sample size in a study. In this thesis we consider two particular but connected subjects. On the one hand, the selection of samples with probabilities proportional to some prescribed values. The first two papers are devoted to this topic. On the other hand, the choice of sampling design to implement in a given survey, which is a topic to which the last two papers are devoted.

Probability proportional to size sampling designs, often referred to as πps designs, are of practical interest due to their potential efficiency. In the literature we can find many of these designs, all having different characteristics. In the first paper we describe and compare ten πps designs with respect to several desired properties. The results suggest that the so called order sampling methods, as well as those proposed by Sunter and Chromy may be considered as good options from a practitioner's point of view.

In the second paper we introduce an algorithm for approximating a target distribution by a mixture distribution. Being a mixture, most of its properties are easy to calculate. We illustrate the use of the algorithm with several examples, both univariate and multivariate. The results indicate that the algorithm succeeds in approximating the target distribution.

The strategy that couples πps designs with the generalized regression estimator is optimal under a given superpopulation model. However, this optimality assumes that the model is correct and some of its parameters are known, which are assumptions that are hardly satisfied in practice. In the third paper we introduce a method that allows for incorporating uncertainty about the model parameters into the choice of the sampling design and then quantifying this uncertainty with a risk measure. The method is illustrated with a real dataset. The results show that the method allowed us to correctly choose the sampling design. The risk measure -as well as other functions that are useful at the planning stage of a survey- is implemented in the package optimStrat developed for R. The fourth paper in this thesis describes the functions in this package.

Keywords: *GREG estimator, mixture distribution, probability proportional to size sampling, sampling algorithms, sampling design, sampling strategy, survey sampling, stratified sampling.*

Stockholm 2020
http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-185930

ISBN 978-91-7911-268-4
ISBN 978-91-7911-269-1

Stockholm University

Department of Statistics

Stockholm University, 106 91 Stockholm

ESSAYS ON SAMPLE SURVEYS

Edgar Bueno

Stockholm
University

Essays on Sample Surveys

Design and Estimation

Edgar Bueno

ISBN print 978-91-7911-268-4
ISBN PDF 978-91-7911-269-1

Printed in Sweden by Universitetsservice US-AB, Stockholm 2020

To my Parents.

List of Papers

The following papers, referred to in the text by their Roman numerals, are included in this thesis.

PAPER I: Edgar Bueno, Dan Hedlin, A comparison of πps designs
(Submitted)

PAPER II: Edgar Bueno, Approximating prescribed distributions by mixtures
(Manuscript)

PAPER III: Edgar Bueno, Dan Hedlin, A method to find an efficient and robust sampling strategy under model uncertainty
(Accepted for publication in *Survey Methodology*)

PAPER IV: Edgar Bueno, optimStrat: An R package for assisting the choice of robust and efficient sampling strategies
(Manuscript)

Acknowledgements

I think there is some charming in traveling alone, but I have always enjoyed more traveling with someone else. Perhaps it is by the very fact that it makes things easier, but perhaps it is also because you feed from your companions' energy and it pushes you to reach places that you would never have reached by yourself.

This five years long travel is coming to an end and this is the right moment to take a look back and acknowledge to all of those who accompanied me along it. Although it is a finite population, I think it is impossible for me to exhaustively enumerate all who made this work possible. Nevertheless, there are some people who should be mentioned with probability equal to one. On the one hand, I would like to thank my supervisors, Per Gösta Andersson and Dan Hedlin, without their guidance and patience (perhaps especially their patience) it would not have been possible for me to consolidate this thesis. I have learned a lot from your knowledge and experience. I hope this thesis reflects at least some of Per Gösta's rigurosity and Dan's practicality. On the other hand, I would like to thank Monica, my wife, for her tireless support, company and love. *"Poco le queda a los hombres cercano a la libertad, yo cuento con tu sonrisa que es sincera de verdad"*.[1]

Of course there were many other travelers who crossed my path and I am also thankful to all of them: all the staff at the Department of Statistics and my PhD fellows. All of you have contributed in making of this a fruitful journey. In particular, huge thanks to Mattias Villani and Daniel Thorburn for making valuable comments on some of the papers here included.

This journey had some *sponsors*: my parents. Without their support and love it would not have been even possible to undertake this journey. I owe it all to you.

[1] 1280 Almas, Tu sonrisa.

Contents

List of Papers **iii**

Acknowledgements **v**

1 Introduction **1**

2 Sampling and estimating from a discrete population **3**
2.1 Sampling designs 3
2.2 Two unbiased estimators 8
2.2.1 The p-estimator 8
2.2.2 The π-estimator 10
2.3 Estimators using auxiliary information 15
2.3.1 The difference estimator 15
2.3.2 The ratio estimator 16
2.3.3 The generalized regression estimator 17

3 Sampling and estimating from a continuous population **23**
3.1 Sampling designs 23
3.2 Some estimators 24

4 Summary of papers **31**

5 Sammanfattning **35**

References **37**

1. Introduction

Dalenius (1985) noted that "there appears to be almost as many definitions of survey as there are people writing about the subject". We follow this tradition and define a survey in a rather wide way as follows:

"A survey is a statistical investigation whose goal is to describe a population of elements by one or more characteristics or parameters. These parameters are, in turn, functions of values of study variables associated to the elements in the population."

Our definition is based on the ones given by Dalenius (1985), Särndal et al. (1992) and Biemer and Lyberg (2003).

Consider, as a first example, a study in which a country's National Health Institute selects a sample of hospitalized patients in order to estimate the prevalence of a specific disease in the country. As a second example, consider a researcher who wants to measure the average temperature in a given region. As a last example, consider a mathematician who wants to approximate the value of the integral of a function over a given domain. Dissimilar as they may seem, the three examples fit into our definition of a survey.

In many applications an exhaustive measurement of the study variables for every element in the population is either unfeasible or simply impossible, which impedes a perfect fulfillment of the survey's goal. In those cases we have to conform ourselves to approximated values of the parameters. There are several approaches to obtain these approximations. We concentrate on one that will be referred to as a *sample survey*.

A sample survey consists of three core stages. At the first stage, a sample of elements is selected by a random mechanism. At the second stage, the values of the study variables of the selected elements are measured. At the last stage, these values are used to obtain an approximation of the desired parameters, this approximation is referred to as an estimate.

Throughout this thesis we assume that once the sample is selected, the values of the study variables are measured with no error in all the sampled elements. Therefore we concentrate on the first and the last stages, usually known as design and estimation stages.

It is the role of the statistician to choose the *sampling design* i.e. the random mechanism for selecting the sample. Also, once the sample is selected, the statistician has to choose the *estimator*, i.e. the formula that will be used for

estimating the parameters. The pair design—estimator comprises the statistical information that is required through the estimation process and it is known as the *sampling strategy*. The choice of sampling strategy should, of course, strive for obtaining estimates as close as possible to the unknown parameters taking into account the available resources for the survey.

The overall aim of the thesis is to develop statistical tools that can be used to obtain efficient estimation of parameters in sample surveys. More specifically, the contents of the thesis concentrates on the following two topics.

- It is often desired that the sampling design is such that it selects each element according to some prescribed probability. Simple as it may sound, this problem has lead to extensive research and proposals of a large number of methods. By now it seems clear that the problem is too intractable for a truly perfect solution. The first paper in this thesis describes and compares several of the methods that have been proposed in the survey literature in the context of a finite population. The second paper proposes a method to approximately fulfill this requirement in the context of continuous populations. The method approximates a prescribed density by a mixture, therefore sampling from the approximation is an easy task.

- The choice of sampling strategy is a crucial step in survey practice, as the efficiency of the resulting estimates depends on it. When choosing the sampling design no data has been collected, therefore all the information required for taking the decision on which design to employ in a given survey often relies on previous knowledge from similar surveys to the one under planning, or, on knowledge by experts. Nevertheless, it is clear that there is uncertainty about this knowledge, uncertainty that must be taken into account in the decision process. The third paper in this thesis introduces a tool that allows for quantifying this uncertainty and then choosing the sampling design to be implemented in the survey. The fourth paper describes a package developed for the statistical software R (R Core Team, 2020), which allows practitioners to employ the proposed tool.

The aim of this introduction is twofold. On the one hand, it provides background needed for reading the four papers that compose the thesis. On the other hand, it is intended to work as a motivation for the development of the papers. The contents of the introduction are arranged as follows. In chapter 2 we consider sampling designs and estimators in the context of discrete populations. In chapter 3 we consider the case of continuous populations. Finally, chapter 4 provides a summary of the papers included in the thesis.

2. Sampling and estimating from a discrete population

The definitions in this and the following chapter are based on (but not necessarily the same as) those presented in Särndal et al. (1992). The reader is referred to that source for a comprehensive treatment of the subject.

Let U be a population that consists of a countable set of elements. Without loss of generality we assume that $U = \{1, 2, \cdots, N\}$ with $N = \infty$ allowed.

2.1 Sampling designs

Definition 1. A *multiset* is an extension of the concept of subset that allows for multiple occurrences of the elements.

Definition 2. A *sample*, s, is any multiset from U. A *with-replacement sample* is a sample that includes elements having multiple occurrences.

Example 1. Let $U = \{1, 2, 3\}$. The following are examples of samples from U: $\emptyset$, $\{1\}$, $\{3\}$, $\{1, 2\}$, U, $\{1, 1, 1\}$, $\{1, 1, 2, 2, 3, 3\}$. Only the last two are with replacement samples.

Definition 3. Let Ω be the set of all multisets of U and Ω_{wo} the powerset of U. (Note that $\Omega_{\text{wo}} \subset \Omega$.) A *sampling design* is a probability distribution on Ω, denoted by $p(s)$. A *without-replacement sampling design* is a probability distribution on Ω_{wo}. A sampling design such that $p(s) > 0$ for any $s \in \Omega \backslash \Omega_{\text{wo}}$ is a *with replacement sampling design*.

Note that any sample s can be considered as an outcome of a random variable S with $P(S = s)$ given by $p(s)$.

Definition 4. The (without-replacement) sampling design that assigns a probability of one to U is called *census*.

Example 2. The following are examples of sampling designs on $U = \{1, 2, 3\}$:

$p_1(\cdot)$ is such that $p_1(\{1, 1\}) = 0.09$, $p_1(\{1, 2\}) = 0.12$, $p_1(\{1, 3\}) = 0.30$, $p_1(\{2, 2\}) = 0.04$, $p_1(\{2, 3\}) = 0.20$ and $p_1(\{3, 3\}) = 0.25$.

$p_2(\cdot)$ is such that $p_2(\{1\}) = 0.2$, $p_2(\{1,2\}) = 0.3$ and $p_2(\{1,1,2,2\}) = 0.5$.

$p_3(\cdot)$ is such that $p_3(\{1,2\}) = 0.1$, $p_3(\{1,3\}) = 0.3$ and $p_3(\{2,3\}) = 0.6$.

$p_4(\cdot)$ is such that $p_4(\{1\}) = 0.6$ and $p_4(\{1,2\}) = 0.4$.

$p_5(\cdot)$ is such that $p_5(\{1,2,3\}) = 1$.

$p_1(\cdot)$ and $p_2(\cdot)$ are with replacement sampling designs. $p_3(\cdot)$, $p_4(\cdot)$ and $p_5(\cdot)$ are without-replacement sampling designs. $p_5(\cdot)$ is a census.

A random mechanism that allows for selecting samples from a given population U is called a *sample selection scheme* or simply a scheme. It is easy to devise a scheme for selecting samples according to any of the designs in Example 2. For instance, the following scheme selects a sample according to $p_2(\cdot)$: let υ be a realization from a uniform distribution $U[0,1)$. If $\upsilon < 0.2$ then the sample $\{1\}$ is selected. If $0.2 \leq \upsilon < 0.5$ then the sample $\{1,2\}$ is selected. Otherwise, if $\upsilon \geq 0.5$, the sample $\{1,1,2,2\}$ is selected.

In the example above, the design was proposed first and then a scheme was devised for selecting a sample according to this design. Alternatively, one can first propose a sampling scheme and then identify the sampling design induced by it.

Example 3. Consider the following scheme. Let $U = \{1,2,\cdots,N\}$ be a finite population and $\lambda(1),\cdots,\lambda(N)$ a set of arbitrary values with $0 \leq \lambda(x) \leq 1$ for all $x \in U$. Let $\upsilon(1),\cdots,\upsilon(N)$ be independent realizations from a uniform distribution $U[0,1)$. The sample consists of all elements x such that $\upsilon(x) < \lambda(x)$. The sampling design induced by this scheme is known as Poisson sampling. It is a without-replacement design with $p(s) = \prod_{x \in s} \lambda(x) \prod_{x \notin s} (1 - \lambda(x))$.

Definition 5. A *statistic* is any (possibly multivariable) function from the sample, it will be denoted by $Q(S)$.

Definition 6. The *expected value* and the *variance* of a statistic $Q(S)$ under the sampling design $p(\cdot)$ are, respectively,

$$\mathrm{E}_p Q(S) = \sum_{\Omega} p(s) Q(s) \qquad \text{and} \qquad \mathrm{V}_p Q(S) = \sum_{\Omega} p(s) \left(Q(s) - \mathrm{E}_p Q(S)\right)^2$$

Definition 7. Let

$$I(x) = \begin{cases} 1 & \text{if } x \in S \\ 0 & \text{if } x \notin S \end{cases} \qquad \text{and} \qquad I(x,x') = \begin{cases} 1 & \text{if } x \in S \text{ and } x' \in S \\ 0 & \text{otherwise} \end{cases}.$$

The *inclusion probability* of x is $\pi(x) = \mathrm{E}_p I(x)$ and the *joint inclusion probability* of x and x' is $\pi(x,x') = \mathrm{E}_p I(x,x')$. A *probability sampling design* is a sampling design with $\pi(x) > 0$ for all $x \in U$. A sample selected from a probability sampling design is called a *probability sample*.

Example 4. Table 2.1 shows the inclusion probabilities induced by each design in Example 2. Note that $p_1(\cdot)$, $p_3(\cdot)$ and $p_5(\cdot)$ are probability sampling designs.

Design	$\pi(1)$	$\pi(2)$	$\pi(3)$	$\pi(1,2)$	$\pi(1,3)$	$\pi(2,3)$
$p_1(\cdot)$	0.51	0.36	0.75	0.12	0.30	0.20
$p_2(\cdot)$	1	0.8	0	0.8	0	0
$p_3(\cdot)$	0.4	0.7	0.9	0.1	0.3	0.6
$p_4(\cdot)$	1	0.4	0	0.4	0	0
$p_5(\cdot)$	1	1	1	1	1	1

Table 2.1: Inclusion probabilities of the sampling designs in Example 2

Example 5. It can be shown that the inclusion probabilities induced by Poisson sampling (Example 3) are $\pi(x) = \pi(x,x) = \lambda(x)$ for all $x \in U$ and $\pi(x,x') = \lambda(x)\lambda(x')$ for all $x \neq x'$. Therefore Poisson sampling is a probability sampling design if and only if $\lambda(x) > 0$ for all $x \in U$.

In general, sampling designs such that the inclusion probabilities $\pi(1), \cdots, \pi(N)$ coincide with some desired target probabilities $\lambda(1), \cdots, \lambda(N)$ are known as *inclusion probabilities proportional to size* designs, denoted πps(λ). For reasons that will be explained in Subsections 2.2.2 and 2.2.2, πps designs are of great interest when sampling from discrete populations.

It is worth noting that πps designs are not unique, a fact that is illustrated in the following example.

Example 6. Let $U = \{1,2,3,4\}$ and $\lambda(1) = 0.35$, $\lambda(2) = 0.45$, $\lambda(3) = 0.55$ and $\lambda(4) = 0.65$. The following are all πps(λ) designs.

Design	$\{1,2\}$	$\{1,3\}$	$\{1,4\}$	$\{2,3\}$	$\{2,4\}$	$\{3,4\}$
$p_1(\cdot)$	0.10	0.10	0.15	0.15	0.20	0.30
$p_2(\cdot)$	0.05	0.10	0.20	0.20	0.20	0.25
$p_3(\cdot)$	0.00	0.35	0.00	0.00	0.45	0.20

Unless stated otherwise we will only consider probability sampling designs, therefore "probability" will be dropped from now on.

Definition 8. Let n_s be the cardinality of a sample s. This statistic is called the *sample size*. A sampling design such that $\mathrm{V}_p n_s = 0$ is called a *fixed size sampling design*. Otherwise it is called a *random size sampling design*.

Example 7. In Example 2, the designs $p_1(\cdot)$, $p_3(\cdot)$ and $p_5(\cdot)$ are of fixed size, whereas $p_2(\cdot)$ and $p_4(\cdot)$ are of random size. Poisson sampling (Example 3) is of random size. In fact, the sample size of Poisson sampling follows a Poisson binomial distribution, which explains the name of the design.

In what remains of this subsection we introduce two important sampling designs, namely, p random sampling and simple random sampling.

Definition 9. For a finite population U, i.e. a population of size $N < \infty$, *simple random sampling without replacement* –srswo– is the sampling design that assigns probability $1/\binom{N}{n}$ to every without replacement sample of fixed size n.

A scheme for selecting srswo samples is as follows. Let $\upsilon(1), \cdots, \upsilon(N)$ be independent realizations from a uniform distribution $U[0,1)$. The sample consists of the elements associated to the n smallest υ-values.

Definition 10. Let $p(x) > 0$ for all $x \in U$ be a weight associated to the xth element, with $\sum_U p(x) = 1$. *p random sampling* –$p$rs– of size n is the design defined as

$$p(s) = p(\{x_1, x_2, \cdots, x_n\}) = \begin{cases} p(x_1)p(x_2)\cdots p(x_n) & \text{if } s \in \Omega_n \\ 0 & \text{otherwise} \end{cases}$$

where Ω_n is the set of all multisets of size n. A sample selected by this design will be called a *p random sample* of size n, or simply, a random sample.

Alternatively, a random sample can be defined as the union of n samples of size one selected by independent experiments that are identically distributed according to $p(x)$.

For reasons that will be explained in Subsection 2.2.1, in the finite population context this design is broadly known as *selection probabilities proportional to size sampling* –$p$ps–.

p random sampling is a with replacement sampling design with inclusion probabilities equal to $\pi(x) = 1 - (1 - p(x))^n$ and $\pi(x, x') = 1 - (1 - p(x))^n - (1 - p(x'))^n + (1 - p(x) - p(x'))^n$.

It remains to specify a scheme for selecting a sample with the desired probabilities. Several methods have already been proposed to this end, among them we can find the inverse–transform method, the alias method or the acceptance–rejection method. The reader is referred to Rubinstein and Kroese (2008, ch. 2), for a comprehensive description of them. We briefly describe the inverse–transform method.

Let $p_u(x) = \sum_{k=1}^{x} p(k)$ for all $x \in U$. Let also $p_l(x) = p_u(x-1)$ if $x > 1$ and $p_l(1) = 0$. In order to select a $p$rs of size one, generate one random variate

from a uniform distribution $U[0,1)$, v_1. The selected element is x_1 such that $p_l(x_1) \leq v_1 < p_u(x_1)$. In order to select a $p$rs of size $n > 1$, simply repeat the selection process n times.

Example 8. Let $U = \{1, \cdots, 10\}$. Our task is to select a $p$rs sample of size $n = 4$ according to the p-values given in the second column of Table 2.2. The third and fourth columns are the sequences $p_l(x)$ and $p_u(x)$. The $n = 4$ realizations from the uniform distribution $U[0,1)$ are $v_1 = 0.17$, $v_2 = 0.34$, $v_3 = 0.30$ and $v_4 = 0.38$. The last four columns of Table 2.2 indicate with a value of one the element that was selected in each step. The resulting sample is $\{2,3,3,3\}$.

For infinite populations, it is not possible to calculate $p_l(x)$ and $p_u(x)$ for all $x \in U$. This problem is circumvented by obtaining only the first terms in the sequence and add more terms only if required for the comparison $p_l(x_1) \leq v_i < p_u(x_1)$. According to Devroye (1986, p. 85), "An exact solution of the inversion inequalities can always be obtained in finite time".

x	$p(x)$	$p_l(x)$	$p_u(x)$	$v_1 = 0.17$	$v_2 = 0.34$	$v_3 = 0.30$	$v_4 = 0.38$
1	0.01	0.00	0.01	0	0	0	0
2	0.23	0.01	0.24	1	0	0	0
3	0.15	0.24	0.39	0	1	1	1
4	0.05	0.39	0.44	0	0	0	0
5	0.10	0.44	0.54	0	0	0	0
6	0.08	0.54	0.62	0	0	0	0
7	0.09	0.62	0.71	0	0	0	0
8	0.11	0.71	0.82	0	0	0	0
9	0.12	0.82	0.94	0	0	0	0
10	0.06	0.94	1.00	0	0	0	0

Table 2.2: Illustrating the inverse–transform method for selecting a $p$rs of size $n = 4$.

Definition 11. For a finite population U, i.e. a population of size $N < \infty$, *simple random sampling with replacement* –srs– (or uniform random sampling) is the special case of p random sampling where $p(x) = 1/N$ for all $x \in U$.

Note the explicit differentiation that we have made between a sample, a probability sample and a random sample: only samples selected from probability sampling designs are probability samples; and, only samples selected by p random sampling are random samples.

2.2 Two unbiased estimators

Let $y(x) = (y_1(x), y_2(x), \cdots, y_{J_y}(x))$ be unknown values of J_y study variables associated to the xth element in U. Let also $y = (y(1)^T, y(2)^T, \cdots, y(N)^T)^T$.

Definition 12. A parameter of U is a function $\theta_U(y)$ into $\mathbb{R}^D$ for some $D \in \mathbb{Z}^+$.

Example 9. Let U be the population of hospitalized patients in a country at a given time. Let $y_1(x) = 1$ if the xth patient has a certain disease and $y_1(x) = 0$ otherwise. Let $y_2(x)$ be the number of days that the xth patient has been hospitalized. We are interested in knowing $\theta_{U,1} = \sum_U y_1(x)$, the prevalence of the disease; $\theta_{U,2} = \sum_U y_2(x)/N$, the average number of days a patient has been hospitalized; and $\theta_{U,3} = \max(y_1(x)y_2(x))$, the maximum number of days a patient with the disease has been hospitalized. We have a three-variate parameter $\theta_U(y) = (\theta_{U,1}(y), \theta_{U,2}(y), \theta_{U,3}(y))$.

Example 10. Let $U = \{0,1,2,\cdots\}$, $y_1(x) = x \cdot 10^x e^{-10}/x!$ and $y_2(x) = x^2 \cdot 10^x e^{-10}/x!$. We are interested in knowing $\theta_{U,1} = \sum_U y_1(x)$ and $\theta_{U,2} = \sum_U y_2(x) - (\sum_U y_1(x))^2$, i.e. the expected value and the variance of a Poisson distribution with parameter $\lambda = 10$.

Definition 13. An *estimator* $\hat{\theta}_s(y)$ is a statistic that approximates the parameter $\theta_U(y)$.

In loose words, a sampling design dictates how to select a sample s from U and an estimator dictates how to use the observed sample to estimate the desired parameter. It is important to emphasize that the only source of randomness is the sampling design, everything else is assumed constant. This approach is known as the design–based approach.

Definition 14. A *sampling strategy* is the couple design and estimator, $(p(\cdot), \hat{\theta}_s(y))$. A sampling strategy is *unbiased* if and only if $\mathrm{E}_p\left[\hat{\theta}_s(y)\right] = \theta_U(y)$ for all y.

From now on, and unless otherwise stated, we will consider the case of only one study variable ($J_y = 1$) and we will focus also on a univariate parameter, namely, the total of y given by $\theta_U(y) = \sum_U y(x)$. We assume that $\theta_U(y)$ is finite.

The remaining part of this subsection is devoted to describing two estimators: the p-expanded estimator and the π-expanded estimator.

2.2.1 The p-estimator

Definition 15. If $p$rs of size n with weights $p(1), p(2), \cdots$ is implemented, then it can be shown that the following estimator is unbiased for the total

$\theta_U(y) = \sum_U y(x)$,

$$\hat{\theta}_s^p(y) = \frac{1}{n}\sum_s \frac{y(x)}{p(x)}. \tag{2.1}$$

This estimator will be called p-expanded estimator or simply p-estimator.

The following result states that the sampling strategy that couples the p-estimator with p random sampling is unbiased for the total. It also states the variance of the strategy and provides an unbiased estimator of the variance.

Result 1. The expected value and the variance of the strategy ($p$rs , $\hat{\theta}_s^p(y)$) are, respectively,

$$\mathrm{E}_{prs}\left[\hat{\theta}_s^p(y)\right] = \theta_U(y) \qquad \text{and} \qquad \mathrm{V}_{prs}\left[\hat{\theta}_s^p(y)\right] = \frac{1}{n}\sum_U p(x)\left(\frac{y(x)}{p(x)} - \theta_U(y)\right)^2. \tag{2.2}$$

The variance is defined if $\sum_U \frac{y^2(x)}{p(x)} < \infty$. An unbiased estimator of the variance is given by

$$\hat{\mathrm{V}}_{prs}\left[\hat{\theta}_s^p(y)\right] = \frac{1}{n(n-1)}\sum_s \left(\frac{y(x)}{p(x)} - \hat{\theta}_s^p(y)\right)^2. \tag{2.3}$$

Proof. Let $u(x) = y(x)/p(x)$ for all $x \in U$ and U_k the random variable denoting the value of u observed at the kth trial ($k = 1, 2, \cdots, n$). We have

$$\mathrm{E}_{prs}U_k = \sum_U p(x)u(x) = \sum_U p(x)\frac{y(x)}{p(x)} = \sum_U y(x) = \theta_U(y)$$

and

$$\mathrm{V}_{prs}U_k = \sum_U p(x)\left(u(x) - \theta_U(y)\right)^2 = \sum_U p(x)\left(\frac{y(x)}{p(x)} - \theta_U(y)\right)^2.$$

The last equality holds if $\sum_U \frac{y^2(x)}{p(x)} < \infty$. Rewriting $\hat{\theta}_s^p(y)$ as $(1/n)\sum_{k=1}^n U_k$, we have

$$\mathrm{E}_{prs}\left[\hat{\theta}_s^p(y)\right] = \mathrm{E}_{prs}\left[\frac{1}{n}\sum_{k=1}^n U_k\right] = \frac{1}{n}\sum_{k=1}^n \mathrm{E}_{prs}U_k = \theta_U(y)$$

and

$$\mathrm{V}_{prs}\left[\hat{\theta}_s^p(y)\right] = \mathrm{V}_{prs}\left[\frac{1}{n}\sum_{k=1}^n U_k\right] = \frac{1}{n}\sum_U p(x)\left(\frac{y(x)}{p(x)} - \theta_U(y)\right)^2.$$

Regarding the variance estimator, using the fact that $\frac{1}{n-1}\sum_{k=1}^n (u_k - \bar{u})^2$ is an unbiased estimator of $\mathrm{V}_{prs}U_k$, we obtain that $\hat{\mathrm{V}}_{prs}[\hat{\theta}_s^p(y)]$ given by (2.3) is an unbiased estimator of $\mathrm{V}_{prs}[\hat{\theta}_s^p(y)]$. □

The p-estimator was introduced by Hansen and Hurwitz (1943) in the finite population context. It has received different names in the literature, for example, it is called "p-expanded with replacement estimator" by Särndal et al. (1992) and "crude Monte Carlo estimator" by Rubinstein and Kroese (2008).

Let us take a look back at the variance of the strategy $(p\text{rs}, \hat{\theta}_s^p(y))$, (2.2). It is easy to see that if all $y(x)$ have the same sign, the choice $p(x) = y(x)/\sum_U y(x)$ would yield a variance equal to zero. This approach is, of course, not feasible as the y-values are typically not available and the denominator is the parameter of interest. An alternative is to make use of an auxiliary variable. An auxiliary variable is a set of values $z = (z(1), z(2), \cdots, z(N))$ that are known (either in intensional or extensional form) and such that $\sum_U z(x) < \infty$ is also known. One can consider defining $p(x) = z(x)/\sum_U z(x)$. If z is expected to be almost proportional to y, this choice is expected to yield a small variance. In finite population statistics, this approach is known as selection probabilities proportional-to-size sampling, denoted $p$ps.

2.2.2 The π-estimator

Definition 16. If a without-replacement sampling design is implemented, then it can be shown that the following estimator is unbiased for the total $\theta_U(y) = \sum_U y_k$,

$$\hat{\theta}_s^{\pi}(y) = \sum_s \frac{y(x)}{\pi(x)}. \tag{2.4}$$

This estimator, introduced by Horvitz and Thompson (1952) in the finite population context, will be called π-expanded estimator or simply π-estimator. The following result states that the sampling strategy that couples a without-replacement sampling design with the π-estimator is unbiased for the total. It also states its variance and provides an unbiased estimator of its variance.

Result 2. If $p(\cdot)$ is a without-replacement design, the expected value and the variance of the strategy $(p(\cdot), \hat{\theta}_s^{\pi}(y))$ are, respectively,

$$\mathrm{E}_p\left[\hat{\theta}_s^{\pi}(y)\right] = \theta_U(y) \qquad \text{and} \qquad \mathrm{V}_p\left[\hat{\theta}_s^{\pi}(y)\right] = \sum_U \sum_U \pi(x,x') \frac{y(x)}{\pi(x)} \frac{y(x')}{\pi(x')} - \theta_U^2(y). \tag{2.5}$$

The variance is defined whenever the first term is finite. If $\pi(x,x') > 0$ for all $x, x' \in U$, an unbiased estimator of the variance is given by

$$\hat{\mathrm{V}}_p^{(1)}\left[\hat{\theta}_s^{\pi}(y)\right] = \sum_s \sum_s \left(\frac{1}{\pi(x)\pi(x')} - \frac{1}{\pi(x,x')}\right) y(x) y(x'). \tag{2.6}$$

If, in addition, the design is of fixed sample size n, the variance can be rewritten as

$$\mathrm{V}_p\left[\hat{\theta}_s^{\pi}(y)\right] = -\frac{1}{2}\sum_U\sum_U(\pi(x,x') - \pi(x)\pi(x'))\left(\frac{y(x)}{\pi(x)} - \frac{y(x')}{\pi(x')}\right)^2 \tag{2.7}$$

and the following is also an unbiased estimator of the variance

$$\hat{\mathrm{V}}_p^{(2)}\left[\hat{\theta}_s^{\pi}(y)\right] = -\frac{1}{2}\sum_s\sum_s\left(1 - \frac{\pi(x)\pi(x')}{\pi(x,x')}\right)\left(\frac{y(x)}{\pi(x)} - \frac{y(x')}{\pi(x')}\right)^2. \tag{2.8}$$

The latter takes nonnegative values if $\pi(x,x') < \pi(x)\pi(x')$ for all $x,x' \in U$.

Proof. Regarding the expectation, we have

$$\mathrm{E}_p\left[\hat{\theta}_s^{\pi}(y)\right] = \mathrm{E}_p\left[\sum_s \frac{y(x)}{\pi(x)}\right] = \mathrm{E}_p\left[\sum_U \frac{y(x)}{\pi(x)}I(x)\right] = \sum_U \frac{y(x)}{\pi(x)}\mathrm{E}_p I(x) = \theta_U(y).$$

Regarding the variance, we have

$$\begin{aligned}
&\mathrm{V}_p\left[\hat{\theta}_s^{\pi}(y)\right] = \mathrm{V}_p\left[\sum_s \frac{y(x)}{\pi(x)}\right] = \mathrm{V}_p\left[\sum_U \frac{y(x)}{\pi(x)}I(x)\right] = \\
&\sum_U\sum_U \mathrm{Cov}_p\left[\frac{y(x)}{\pi(x)}I(x), \frac{y(x')}{\pi(x')}I(x')\right] = \sum_U\sum_U \frac{y(x)y(x')}{\pi(x)\pi(x')}\mathrm{Cov}_p\left[I(x), I(x')\right] = \\
&\qquad \sum_U\sum_U \frac{y(x)y(x')}{\pi(x)\pi(x')}\left(\mathrm{E}_p I(x)I(x') - \mathrm{E}_p I(x)\mathrm{E}_p I(x')\right) = \\
&\qquad\qquad \sum_U\sum_U \frac{y(x)y(x')}{\pi(x)\pi(x')}\left(\pi(x,x') - \pi(x)\pi(x')\right) = \\
&\qquad\qquad\qquad \sum_U\sum_U \frac{y(x)y(x')}{\pi(x)\pi(x')}\pi(x,x') - \sum_U\sum_U y(x)y(x'),
\end{aligned}$$

which is the variance (2.5). The unbiasedness of the variance estimator (2.6) follows from the fact that

$$\mathrm{E}_p\left[\left(\frac{\pi(x,x') - \pi(x)\pi(x')}{\pi(x,x')}\right)I(x)I(x')\right] = \pi(x,x') - \pi(x)\pi(x'). \tag{2.9}$$

That (2.7) is equivalent to the variance (2.5) for a fixed-size sample design is shown as follows. We can rewrite (2.7) as

$$\sum_U\sum_U \pi(x,x')\frac{y(x)}{\pi(x)}\frac{y(x')}{\pi(x')} - \theta_U^2(y) - \sum_U\sum_U(\pi(x,x') - \pi(x)\pi(x'))\frac{y^2(x)}{\pi^2(x)}.$$

But, as $\sum_{x=1}^N \pi(x,x') = n\,\pi(x')$ and $\sum_U \pi(x) = n$ for fixed-size designs, the last term is equal to zero. The unbiasedness of (2.8) follows from property (2.9). □

It is clear from (2.7) that if all $y(x)$ have the same sign the choice $\pi(x) = ny(x)/\theta_U(y)$ would yield a variance equal to zero. As in the case of the p-estimator, this is not possible, and one typically defines the inclusion probabilities π proportional to some auxiliary variable z that is expected to be, in turn, proportional to y. This yields $\pi(x) = nz(x)/\theta_U(z)$. This approach is known as inclusion probabilities proportional-to-size sampling, denoted πps.

The potential gain in efficiency due to the use of πps designs, makes them a very attractive option in survey practice. Therefore we need a sample selection scheme that allows for implementing such a design. Note, however, that there is a list of properties that one may desire from a πps scheme. First, the scheme should actually select πps(λ) samples, i.e. the inclusion probabilities $\pi(x)$ induced by the scheme should coincide with some prescribed probabilities $\lambda(x)$. Second, in order for expression (2.7) to be valid, the scheme should generate without replacement samples of fixed size. Poisson sampling (Examples 3 and 5), for instance, satisfies the first condition, but not the second one. Moreover, in order to be able to unbiasedly estimate the variance through (2.6) or (2.8), all $\pi(x,x')$ should be larger than zero and (at least those for the selected sample) known. Finally, the scheme should be easy to implement.

To the knowledge of the author there is no available scheme that fully satisfies the conditions above. Nevertheless, several methods that approximately satisfy the conditions above are available, among them we find Pareto πps (Rosén, 1997), systematic πps (see, e.g. Madow 1949) and Sampford's method (Sampford, 1967). The first paper in this thesis describes and compares several of such schemes in the finite population context.

Even when two πps(λ) designs share the first order inclusion probabilities, their joint inclusion probabilities, $\pi(x,x')$ may differ. Therefore, their variance under the π-estimator, (2.5) or (2.7), may also differ as they depend on the $\pi(x,x')$.

Example 11. In Example 6 we presented three πps(λ) designs with $\lambda(1) = 0.35$, $\lambda(2) = 0.45$, $\lambda(3) = 0.55$ and $\lambda(4) = 0.65$. Let us assume that the study variable takes the values $y(1) = 33$, $y(2) = 42$, $y(3) = 67$ and $y(4) = 62$. It is easy to show that the variance of the strategies that couple the π-estimator with each design are $\mathrm{V}_{p_1}[\hat{\theta}_s^{\pi}] = 189.9$, $\mathrm{V}_{p_2}[\hat{\theta}_s^{\pi}] = 184.3$ and $\mathrm{V}_{p_3}[\hat{\theta}_s^{\pi}] = 191.2$.

Comparing the strategies ($p$rs , p-estimator) and (πps , π-estimator)

In Section 2.2.1 we saw that the strategy that couples $p$rs with the p-estimator is expected to be efficient if the p-values are proportional to some auxiliary variable z which is, in turn, expected to be well correlated to the study variable y, i.e. if $p(x) = z(x)/\theta_U(z)$. In Section 2.2.2 we saw that we can, alternatively, use the auxiliary variable into the strategy that couples a without-replacement

fixed-size design with the π-estimator by defining the inclusion probabilities as $\pi(x) = nz(x)/\theta_U(z)$.

One might suspect that, under the same sample size n, the latter is more efficient than the former, as we are sampling without replacement. It turns out that this is, in general, true. Note that the following identity holds,

$$\mathrm{V}_{p\mathrm{rs}}\left[\hat{\theta}_s^p(y)\right] = \mathrm{V}_{\pi\mathrm{ps}}\left[\hat{\theta}_s^\pi(y)\right] + C$$

with

$$C = \frac{n-1}{n}\left(\theta_U^2(y) - \sum_{\substack{U \\ x\neq x'}}\sum_{U} p(x,x')\frac{y(x)}{p(x)}\frac{y(x')}{p(x')}\right)$$

and $p(x,x') = \frac{\pi(x,x')}{n(n-1)}$. This means that the strategy (πps , $\hat{\theta}_s^\pi(y)$) is more efficient than ($p$rs , $\hat{\theta}_s^p(y)$) if $C > 0$. This happens, for instance, if we allow the approximation $p(x,x') \approx p(x)p(x')$ (see Hajek 1981).

The following example illustrates the gain in efficiency due to the use of using πps sampling instead of $p$rs.

Example 12. This example makes use of the population `agpop` from Lohr (2009) which is available at package `SDaA` (Verbeke, 2014) in `R` (R Core Team, 2020).

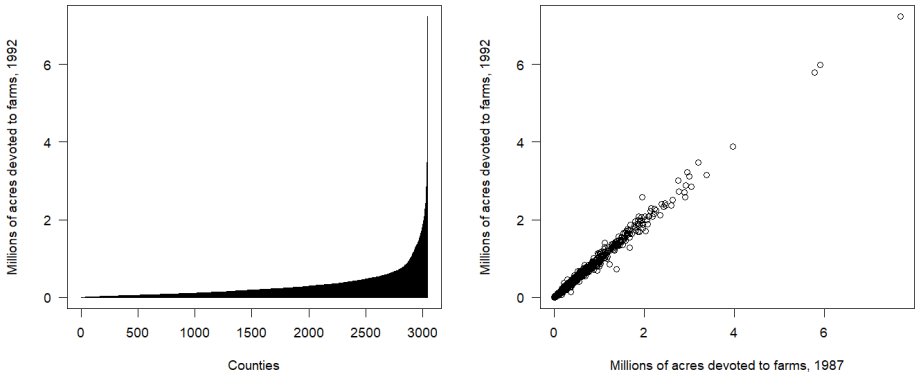


Figure 2.1: Left panel: study variable y by county x, sorted from smallest to largest. Right panel: scatter plot of the auxiliary variable z vs. study variable y.

Let U be the population of $N = 3044$ counties in the United States in 1992, $y(x) =$ acreage devoted to farms in the xth county in 1992. We want to estimate $\theta_U(y)$, the total acreage devoted to farms in the United States using $z(x) =$

acreage devoted to farms in the xth county in 1987 as auxiliary variable. The total of $z(x)$ is $\theta_U(z) = 9.62 \cdot 10^8$. The left panel of figure 2.1 shows the values of the study variable by county as vertical bars, the values have been sorted from smallest to largest. The right panel shows a scatter plot between the auxiliary variable z and the study variable y. Clearly, the z-values are not only fairly proportional to the y-values but also good approximations of them. Of course, in practice it is not possible to draw these plots. However, using subject matter knowledge, one might expect z to be highly correlated to y.

The variance of the strategy ($p$rs , p-estimator) given by (2.2) with $p(x) = z(x)/\theta_U(z)$ was computed for all sample sizes from 1 to N. The solid line in Figure 2.2 shows the coefficient of variation $\mathrm{V}^{0.5}_{p\mathrm{rs}}[\hat{\theta}^p_s(y)]/\theta_U(y)$ obtained for each sample size.

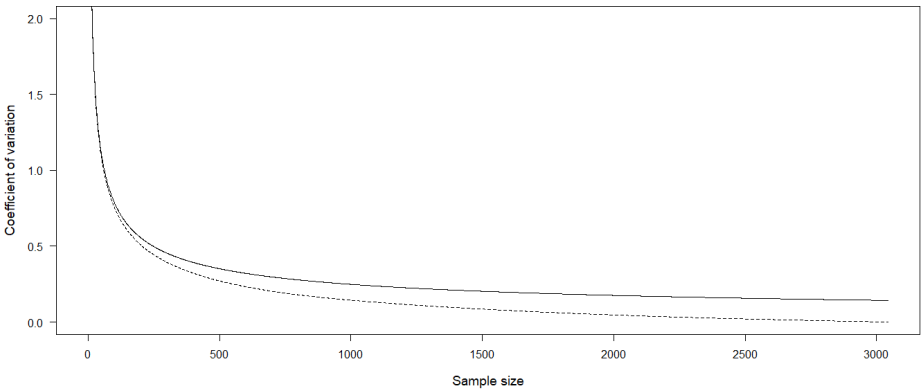


Figure 2.2: Coefficient of variation ($\times 100$) of two sampling strategies. Solid line: ($p$rs , p-estimator). Dashed line: (πps , π-estimator)

Regarding the πps design, we implemented Pareto πps (Rosén, 1997) with target inclusion probabilities given by $\lambda(x) = n z(x)/\theta_U(z)$. For some sample sizes, some λ-values obtained in this way are larger than one. This is corrected by setting them to one, and recalculating the target inclusion probabilities for the remaining values correspondingly reducing the sample size and the total.

The selection scheme of Pareto πps is as follows. Let $\upsilon(1), \cdots, \upsilon(N)$ be independent realizations from a uniform distribution $[0,1)$ and $\phi(x) = (\upsilon(x)/\lambda(x)) \cdot (1-\lambda(x))/(1-\upsilon(x))$ for all $x \in U$. The sample consists of the elements associated to the n smallest ϕ-values. The inclusion probabilities induced by this scheme are $\pi(x) \approx \lambda(x)$ and (Swensson, 1998)

$$\pi(x,x') \approx \lambda(x)\lambda(x') \left(1 - \frac{N}{N-1} \frac{(1-\lambda(x))(1-\lambda(x'))}{n - \sum_U \lambda^2(x)}\right).$$

The variance of the strategy (πps , π-estimator) given by (2.7), with the approximated $\pi(x)$ and $\pi(x,x')$, was computed for all sample sizes from 1 to N. The dashed line in Figure 2.2 shows the coefficient of variation ($\mathrm{V}[\hat{\theta}]/\theta$) in percentage units obtained for each sample size. It is clear that there is some gain in efficiency due to using without replacement sampling instead of a with replacement approach. On the one hand, it is true that (based on this example, at least) the difference between both strategies does not seem to be dramatic, especially for small sample sizes. However, in practice, even a small reduction in sample size is desirable as it represents a reduction in costs, follow-ups and response burden, among others.

2.3 Estimators using auxiliary information

In Section 2.2, we saw that under certain conditions the p- and π-estimators are unbiased. Also, we saw that if coupled with appropriate designs, the resulting strategy will yield a small variance. The use of auxiliary information is a powerful ingredient in the definition of such designs. However, up to this point, it is the design —not the estimator— the component that exploits the auxiliary information. In this subsection we introduce three estimators that make use of auxiliary information, namely, the difference estimator, the ratio estimator and the generalized regression estimator. We see that the variance may be further reduced by this technique.

The three estimators are constructed based on an unbiased estimator, either the p- or the π-estimator. To begin with, we show their construction based on the p-estimator. The alternative construction based on the π-estimator is summarized in a result at the end of the section. The difference and the ratio estimators make use of one auxiliary variable z defined as before, i.e. a set of values that are known and such that $\sum_U z(x) < \infty$ is also known. The third estimator allows for using multiple auxiliary variables.

2.3.1 The difference estimator

Definition 17. The difference estimator is defined as

$$\hat{\theta}_s^d(y) = \theta_U(z) + \hat{\theta}_s^p(e) \tag{2.10}$$

where $\theta_U(z) = \sum_U z(x)$ is the total of the auxiliary variable and $\hat{\theta}_s^p(e) = \frac{1}{n}\sum_s \frac{e(x)}{p(x)}$ is the p-estimator of the total of the differences $e(x) = y(x) - z(x)$.

In the Monte-Carlo literature, this estimator is known as the "control variate method" or "control variable method" (see, for example, Rubinstein and Kroese 2008 and Robert and Casella 2013).

The following result states that the sampling strategy that couples the difference estimator with *p*rs is also unbiased for the total. Its variance and an unbiased estimator of the variance are also provided.

Result 3. The expected value and the variance of the strategy ($p$rs , $\hat{\theta}_s^d(y)$) are, respectively,

$$\mathrm{E}_{p\mathrm{rs}}\left[\hat{\theta}_s^d(y)\right] = \theta_U(y) \qquad \text{and} \qquad \mathrm{V}_{p\mathrm{rs}}\left[\hat{\theta}_s^d(y)\right] = \frac{1}{n}\sum_U p(x)\left(\frac{e(x)}{p(x)} - \theta_U(e)\right)^2, \tag{2.11}$$

where $\theta_U(e) = \sum_U e(x)$ and $e(x)$ as defined above. The variance is defined whenever $\sum_U \frac{e^2(x)}{p(x)} < \infty$. An unbiased estimator of the variance is given by

$$\hat{\mathrm{V}}_{p\mathrm{rs}}\left[\hat{\theta}_s^d(y)\right] = \frac{1}{n(n-1)}\sum_s\left(\frac{e(x)}{p(x)} - \hat{\theta}_s^p(e)\right)^2. \tag{2.12}$$

Proof. Note that $y(x)$ can be written as $y(x) = z(x) + e(x)$, therefore $\theta_U(y) = \theta_U(z) + \theta_U(e)$. The first term on the right hand side is a known constant. The second term can be estimated using the p-estimator. The proof is completed by using result 1 with $e(x)$ instead of $y(x)$. □

If we could define $z(x) = y(x)$, the variance of the difference estimator would be equal to zero. Although this is not possible, this strategy is efficient when the z-values are close to the y-values.

2.3.2 The ratio estimator

Definition 18. The ratio estimator is defined as

$$\hat{\theta}_s^r(y) = \frac{\theta_U(z)}{\hat{\theta}_s^p(z)}\hat{\theta}_s^p(y) \tag{2.13}$$

where $\theta_U(z) = \sum_U z(x)$ is the total of the auxiliary variable, $\hat{\theta}_s^p(y) = \frac{1}{n}\sum_s \frac{y(x)}{p(x)}$ is the p-estimator of the total of y and $\hat{\theta}_s^p(z) = \frac{1}{n}\sum_s \frac{z(x)}{p(x)}$ is the p-estimator of the total of z.

The ratio estimator is, in general, biased, although its bias is small even for moderate sample sizes. Note that if we could define $z(x) \propto y(x)$, the ratio estimator would be exactly equal to $\theta_U(y)$. The (random) factor $\theta_U(z)/\hat{\theta}_s^p(z)$ in (2.13) can be seen as a correction factor: if the obtained sample is such that the p-estimator over-(under-) estimates the total of z it is reasonable that it also over-(under-) estimates the total of y, therefore it is weighted down(up) by this factor.

In the following result we present an approximation and an estimator of the variance for the strategy that couples the ratio estimator with *p*rs.

Result 4. The variance of the strategy (prs, $\hat{\theta}_s^r(y)$) can be approximated by

$$\mathrm{AV}_{prs}\left[\hat{\theta}_s^r(y)\right] = \frac{1}{n}\sum_U \frac{e^2(x)}{p(x)} \qquad \text{with} \qquad e(x) = y(x) - z(x)\frac{\theta_U(y)}{\theta_U(z)}. \quad (2.14)$$

An estimator of the variance is

$$\hat{\mathrm{V}}_{prs}\left[\hat{\theta}_s^r(y)\right] = \frac{1}{n(n-1)}\left(\frac{\theta_U(z)}{\hat{\theta}_s^p(z)}\right)^2 \sum_s \frac{e_s^2(x)}{p^2(x)} \quad \text{with} \quad e_s(x) = y(x) - z(x)\frac{\hat{\theta}_s^p(y)}{\hat{\theta}_s^p(z)}. \quad (2.15)$$

As the ratio estimator is a special case of the general regression estimator, the result will be proved when the approximation to the variance of the latter is given.

2.3.3 The generalized regression estimator

The generalized regression –GREG– estimator will be defined next. For this purpose we will consider the case of multiple auxiliary variables, i.e. for all $x \in U$, a vector of known values $z(x) = (z_1(x), z_2(x), \cdots, z_{J_z}(x))$ is associated to the xth element in U. We assume that $\theta_U(z_j) = \sum_U z_j(x)$ is finite for all $j = 1, 2, \cdots, J_z$.

The GREG estimator owes its name to the fact that a linear regression between the auxiliary variables z and the study variable y underlies its construction. There are several variations of the GREG estimator, for example, Breidt and Opsomer (2000) proposed a local polynomial variation, Montanari and Ranalli (2005) use neural network models and Rondon et al. (2012) extend the idea to generalized linear models. Our definition follows the lines of the presentation in Särndal et al. (1992).

Recall that the difference estimator is efficient if the auxiliary variable approximates the study variable well. Intuitively, a good choice for auxiliary variable would be the fitted values of a regression model of the z-variables to the study variable. As only the sampled y-values are available, this regression can only be fitted on a sample-dependent basis. The price to pay for this flexibility is that the resulting estimator is no longer unbiased, although its bias is known to be small even for moderate sample sizes.

Definition 19. More formally, the GREG estimator is defined as

$$\hat{\theta}_s^g(y) = \left(\theta_U(z) - \hat{\theta}_s^p(z)\right)\hat{B} + \hat{\theta}_s^p(y) \quad (2.16)$$

where $\hat{\theta}_s^p(y) = \frac{1}{n}\sum_s \frac{y(x)}{p(x)}$ is the p-estimator of the total of y, $\theta_U(z) = (\theta_U(z_1), \theta_U(z_2), \cdots, \theta_U(z_{J_z}))$ is the vector of totals of the auxiliary variables,

$\hat{\theta}_s^p(z) = (\hat{\theta}_s^p(z_1), \hat{\theta}_s^p(z_2), \cdots, \hat{\theta}_s^p(z_{J_z}))$ is the vector of corresponding p-estimators and

$$\hat{B} = \left(\sum_s \frac{z(x)^T z(x)}{a(x)p(x)} \right)^{-1} \sum_s \frac{z(x)^T y(x)}{a(x)p(x)}. \tag{2.17}$$

The a-values are weights defined by the statistician, they are related to the variance structure assumed in the regression model underlying the GREG estimator.

There is no closed expression for the variance of the GREG estimator. The following result provides an approximation and an estimator for the variance of the strategy that couples the GREG estimator with $p$rs.

Result 5. The variance of the strategy ($p$rs , $\hat{\theta}_s^g(y)$) can be approximated by

$$\mathrm{AV}_{p\mathrm{rs}} \left[\hat{\theta}_s^g(y) \right] = \frac{1}{n} \sum_U p(x) \left(\frac{e(x)}{p(x)} - \theta_U(e) \right)^2 \tag{2.18}$$

where $\theta_U(e) = \sum_U e(x)$ and $e(x) = y(x) - z(x)B$ with

$$B = \left(\sum_U \frac{z(x)^T z(x)}{a(x)} \right)^{-1} \sum_U \frac{z(x)^T y(x)}{a(x)}. \tag{2.19}$$

An estimator of the variance is

$$\hat{\mathrm{V}}_{p\mathrm{rs}} \left[\hat{\theta}_s^g(y) \right] = \frac{1}{n(n-1)} \sum_s \left(\frac{e_s(x)}{p(x)} - \hat{\theta}_s^p(e_s) \right)^2 \tag{2.20}$$

where $\hat{\theta}_s^p(e_s) = \frac{1}{n} \sum_s \frac{e_s(x)}{p(x)}$ and $e_s(x) = y(x) - z(x)\hat{B}$.

If, in addition to the totals of the z-variables, the matrix of cross-totals

$$\Theta_U(z) = \sum_U \frac{z(x)^T z(x)}{a(x)}$$

is available, the residuals $e_s(x)$ required for the calculation of (2.20) can be redefined as $e_s(x) = g_s(x) \left(y(x) - z(x)\hat{B} \right)$ with

$$g_s(x) = 1 + (\theta_U(z) - \hat{\theta}_s^p(z))\, \Theta_U^{-1}(z)\, z(x)^T / a(x)$$

.

The proof follows the lines of Särndal et al. (1992, p. 236). Some details will be omitted, the interested reader is therefore referred to that source.

Proof. Using the Taylor linearization method, the GREG estimator (2.16) can be approximated by $\hat{\theta}_s^g(y) \approx \theta_U(z)B + \hat{\theta}_s^p(e)$. Expression (2.18) follows by taking into account that the first term in the approximation is a constant.

Regarding the variance estimator, note that the GREG estimator can be alternatively written as $\hat{\theta}_s^g(y) = \theta_U(z)B + \hat{\theta}_s^p(g_s(x)e(x))$, therefore

$$\mathrm{V}_{prs}\left[\hat{\theta}_s^g(y)\right] = \mathrm{V}_{prs}\left[\theta_U(z)B + \hat{\theta}_s^p(g_s(x)e(x))\right] = \mathrm{V}_{prs}\left[\hat{\theta}_s^p(g_s(x)e(x))\right].$$

It is tempting to use the fact that (2.3) is an unbiased estimator of (2.2) and then simply use $g_s(x)e(x)$ instead of $y(x)$, which would lead to the estimator

$$\hat{\mathrm{V}}_{prs}\left[\hat{\theta}_s^g(y)\right] = \frac{1}{n(n-1)}\sum_s \left(\frac{g_s(x)e(x)}{p(x)} - \hat{\theta}_s^p(g_s e)\right)^2.$$

However, two problems arise with the last expression. First, the weights $g_s(x)$ are sample dependent and, second, the residuals $e(x)$ are unknown. The estimator (2.20) with the g-weighted residuals is obtained by assuming that the variance of $g_s(x)$ is negligible and using the sample residuals $e_s(x)$ instead of the population ones. The estimator (2.20) with the unweighted residuals is obtained by further letting $g_s(x) = 1$. □

In the case of a single auxiliary variable, if we let $a(x) = z(x)$, the GREG estimator becomes the ratio estimator.

As mentioned above, it is possible to construct these estimators based on the π-estimator. The following result presents the resulting estimators, approximations to their variances and variance estimators.

Result 6. Let $p(\cdot)$ be a without-replacement sampling design of fixed size n and z an auxiliary variable. The difference estimator is defined as

$$\hat{\theta}_s^d(y) = \theta_U(z) + \hat{\theta}_s^\pi(e)$$

where $\theta_U(z) = \sum_U z(x)$ is the total of the auxiliary variable and $\hat{\theta}_s^\pi(e) = \sum_s \frac{e(x)}{\pi(x)}$ is the π-estimator of the total of the differences $e(x) = y(x) - z(x)$. The variance of the strategy $(p(\cdot), \hat{\theta}_s^d(y))$ is given by

$$\mathrm{V}_p\left[\hat{\theta}_s^d(y)\right] = -\frac{1}{2}\sum_U\sum_U(\pi(x,x') - \pi(x)\pi(x'))\left(\frac{e(x)}{\pi(x)} - \frac{e(x')}{\pi(x')}\right)^2.$$

An unbiased estimator of the variance is

$$\hat{\mathrm{V}}_p\left[\hat{\theta}_s^d(y)\right] = -\frac{1}{2}\sum_s\sum_s\left(1 - \frac{\pi(x)\pi(x')}{\pi(x,x')}\right)\left(\frac{e(x)}{\pi(x)} - \frac{e(x')}{\pi(x')}\right)^2.$$

The ratio estimator is defined as

$$\hat{\theta}_s^r(y) = \frac{\theta_U(z)}{\hat{\theta}_s^\pi(z)}\hat{\theta}_s^\pi(y)$$

where $\theta_U(z) = \sum_U z(x)$ is the total of the auxiliary variable, $\hat{\theta}_s^\pi(y) = \sum_s \frac{y(x)}{\pi(x)}$ is the π-estimator of the total of y and $\hat{\theta}_s^\pi(z) = \sum_s \frac{z(x)}{\pi(x)}$ is the π-estimator of the total of z. The variance of the strategy $(p(\cdot), \hat{\theta}_s^r(y))$ is approximated by

$$\mathrm{AV}_p\left[\hat{\theta}_s^r(y)\right] = -\frac{1}{2}\sum_U\sum_U(\pi(x,x') - \pi(x)\pi(x'))\left(\frac{e(x)}{\pi(x)} - \frac{e(x')}{\pi(x')}\right)^2$$

with $e(x) = y(x) - z(x)\cdot\theta_U(y)/\theta_U(z)$. The variance can be estimated by

$$\hat{\mathrm{V}}_p\left[\hat{\theta}_s^r(y)\right] = -\frac{1}{2}\left(\frac{\theta_U(z)}{\hat{\theta}_s^\pi(z)}\right)^2\sum_s\sum_s\left(1 - \frac{\pi(x)\pi(x')}{\pi(x,x')}\right)\left(\frac{e_s(x)}{\pi(x)} - \frac{e_s(x')}{\pi(x')}\right)^2$$

with $e_s(x) = y(x) - z(x)\cdot\hat{\theta}_s^\pi(y)/\hat{\theta}_s^\pi(z)$.

Consider, again, the case of multiple auxiliary variables. The GREG estimator is defined as

$$\hat{\theta}_s^g(y) = \left(\theta_U(z) - \hat{\theta}_s^\pi(z)\right)\hat{B} + \hat{\theta}_s^\pi(y)$$

where $\theta_U(z) = (\theta_U(z_1), \theta_U(z_2), \cdots, \theta_U(z_{J_z}))$ is the vector of totals of the auxiliary variables, $\hat{\theta}_s^\pi(z) = (\hat{\theta}_s^\pi(z_1), \hat{\theta}_s^\pi(z_2), \cdots, \hat{\theta}_s^\pi(z_{J_z}))$ is the vector of corresponding π-estimators and

$$\hat{B} = \left(\sum_s \frac{z(x)^T z(x)}{a(x)\pi(x)}\right)^{-1}\sum_s \frac{z(x)^T y(x)}{a(x)\pi(x)}.$$

The a-values are weights defined by the statistician. The variance of the strategy $(p(\cdot), \hat{\theta}_s^g(y))$ is approximated by

$$\mathrm{AV}_p\left[\hat{\theta}_s^g(y)\right] = -\frac{1}{2}\sum_U\sum_U(\pi(x,x') - \pi(x)\pi(x'))\left(\frac{e(x)}{\pi(x)} - \frac{e(x')}{\pi(x')}\right)^2 \qquad (2.21)$$

with $e(x) = y(x) - z(x)B$ with B given by (2.19). The variance can be estimated by

$$\hat{\mathrm{V}}_p\left[\hat{\theta}_s^g(y)\right] = -\frac{1}{2}\sum_s\sum_s\left(1 - \frac{\pi(x)\pi(x')}{\pi(x,x')}\right)\left(\frac{e_s(x)}{\pi(x)} - \frac{e_s(x')}{\pi(x')}\right)^2$$

where $e_s(x) = y(x) - z(x)\hat{B}$. If the matrix of cross-totals $\Theta_U(z) = \sum_U z(x)^T z(x)/a(x)$ is available, the residuals can be redefined as $e_s(x) = g_s(x)\left(y(x) - z(x)\hat{B}\right)$ with $g_s(x) = 1 + (\theta_U(z) - \hat{\theta}_s^\pi(z))\,\Theta_U^{-1}(z)\,z(x)^T/a(x)$.

Due to the possibility of taking multiple auxiliary variables into account, of all the estimators presented so far, it is the GREG estimator the one having more potential for reducing variance. Moreover, sampling without replacement is expected to be more efficient than sampling with replacement. Therefore a strategy that couples the GREG estimator with a without-replacement fixed-size sampling design seems to be a good choice. In practice, however, there may be several designs fulfilling these requirements. The third paper in this thesis introduces a method for choosing between them. This method has been implemented in a package developed for the R software (R Core Team, 2020). The fourth paper describes the use of the package.

We close this chapter illustrating some of the sampling strategies introduced up to this point.

Example 13. We want to estimate the expected value of a random variable X that follows a Poisson distribution with parameter $\mu = 5$, in other words, we are interested in estimating $\theta_U(y) = \sum_U y(x)$ with $y(x) = x \cdot 5^x e^{-5}/x!$ for all $x \in U = \{0, 1, \cdots\}$. Of course we know that the expected value is $\mathrm{E}X = \mathrm{V}X = \mu = 5$.

To begin with, one sample of size $n = 5$ was selected by $p$rs with $p(x) = 5^x e^{-5}/x!$. We obtained $s = \{4, 5, 6, 7, 12\}$. Note that the p-estimator becomes

$$\hat{\theta}_s^p(y) = \frac{1}{n}\sum_s \frac{y(x)}{p(x)} = \frac{1}{n}\sum_s x = \bar{x}_s = 6.8.$$

The GREG estimator with

$$z_1(x) = \begin{cases} p(x) & \text{if } x \leq 4.5 \\ 0 & \text{otherwise} \end{cases} \qquad \text{and} \qquad z_2(x) = \begin{cases} p(x) & \text{if } x > 4.5 \\ 0 & \text{otherwise} \end{cases}$$

was also implemented. Taking into account that $\theta_U(z_1) = 0.4405$ and $\theta_U(z_2) = 0.5595$, we obtained $\hat{\theta}_s^g(y) = 6.3$, which is closer to the parameter of interest.

Alternatively, we can consider πps with target inclusion probabilities given by $\lambda(x) = np(x)$ for all $x \in U$. The selection scheme for Pareto πps described in Example 12 cannot be implemented in unbounded populations. Therefore, we opt for systematic πps. The selection scheme of systematic πps is as follows. Let $\lambda_u(x) = \sum_{k=0}^{x} \lambda(k)$ for all $x \in U$. Let also $\lambda_l(x) = \lambda_u(x-1)$ if $x > 0$ and $\lambda_l(0) = 0$. Generate one random variate from a uniform distribution $U[0,1)$, υ. The sample consists of those elements such that $\lambda_l(x) \leq \upsilon + j < \lambda_u(x)$ for $j = 0, 1, \cdots, n-1$. The inclusion probabilities induced by this scheme are $\pi(x) = \lambda(x)$.

One sample of size $n = 5$ was then selected by systematic πps, we obtained $s = \{2, 3, 5, 6, 7\}$. The π-estimator becomes

$$\hat{\theta}_s^\pi(y) = \sum_s \frac{y(x)}{\pi(x)} = \frac{1}{n}\sum_s x = \bar{x}_s = 4.6,$$

which is even closer to the parameter than the estimate obtained by the strategy ($p$rs , GREG-estimator). The GREG estimator with $z_1(x)$ and $z_2(x)$ as defined above was also implemented, we obtained $\hat{\theta}_s^g(y) = 4.5$.

Apparently, based on one sample, the strategies using πps are more efficient than those using $p$rs. In this case we can verify this fact by calculating the (approximated) variance of each strategy. For the strategy that couples $p$rs with the p-estimator we have

$$\mathrm{V}_{p\mathrm{rs}}\left[\hat{\theta}_s^p(y)\right] = \frac{1}{n}\sum_U p(x)\left(\frac{y(x)}{p(x)} - \theta_U(y)\right)^2 = \frac{1}{n}\sum_U p(x)\,(x-\mu)^2 = \frac{1}{n}\mu = 1.$$

Using (2.18), we obtain a variance of 0.43 for the strategy that couples $p$rs with the GREG estimator. Regarding the strategies using πps, in order to use (2.7) and (2.21) we need the joint inclusion probabilities $\pi(x,x')$. In general, there is no closed expression for the $\pi(x,x')$ induced by systematic sampling. However, in this case we obtain the inclusion probabilities shown in Table 2.3 for $x,x' \leq 9$. Only the upper triangular portion is shown, but the table is symmetric. The remaining inclusion probabilities are

$$\pi(x,x') = \pi(x',x) = \begin{cases} \pi(x) & \text{if } x' = 3,4,5,7 \\ 0 & \text{otherwise} \end{cases}.$$

Therefore we obtain a variance of 0.38 for the strategy that couples πps with the π-estimator, and 0.23 for the one using the GREG-estimator, which shows the gain in efficiency due to the use of the π estimator instead of the p estimator.

		x'									
	$10\,\pi(x,x')$	0	1	2	3	4	5	6	7	8	9
x	0	0.337	0.000	0.000	0.337	0.337	0.337	0.000	0.337	0.000	0.000
	1	—	1.684	0.000	1.684	1.684	0.461	1.223	1.684	0.000	0.000
	2	—	—	4.211	1.230	2.985	4.208	4.211	1.310	2.901	0.000
	3	—	—	—	7.019	5.792	5.792	4.330	5.142	0.363	1.813
	4	—	—	—	—	8.773	7.547	6.085	3.996	3.264	1.813
	5	—	—	—	—	—	8.773	6.085	3.996	3.264	1.813
	6	—	—	—	—	—	—	7.311	2.533	3.264	1.514
	7	—	—	—	—	—	—	—	5.222	0.000	0.299
	8	—	—	—	—	—	—	—	—	3.264	0.000
	9	—	—	—	—	—	—	—	—	—	1.813

Table 2.3: Joint inclusion probabilities ($\times 10$) for systematic πps in Example 13

3. Sampling and estimating from a continuous population

In this chapter we basically restate the definitions and results given in Chapter 2, adapting them to continuous populations.

3.1 Sampling designs

Let $U \subseteq \mathbb{R}^m$ be the population of interest. In the continuous case, a sample is defined in the same way as in the discrete case, i.e. a *sample*, s, is any multiset from U.

Definition 20. Let Ω be the set of all multisets of U and $\mathcal{F}$ the Borel sigma algebra on Ω. A *sampling design* is a probability measure on $(\Omega, \mathcal{F})$, and it will be denoted by $p(s)$.

The concepts of *census*, *statistic*, *expected value* and *variance* of a statistic and *sample size*, remain the same as in the discrete case (see definitions 4, 5, 6 and 8).

We limit ourselves to discuss the continuous counterpart of *p*rs and simple random sampling.

Definition 21. Let $p(x) > 0$ for all $x \in U$ be a known function such that $\int_U p(x)dx = 1$. p random sampling –*p*rs– of size n is the design defined as

$$p(s) = p(\{x_1, x_2, \cdots, x_n\}) = \begin{cases} p(x_1)p(x_2)\cdots p(x_n) & \text{if } s \subset \Omega_n \\ 0 & \text{otherwise} \end{cases}$$

where Ω_n is the set of all multisets of size n. As in the discrete case, a sample selected by this design will be called a p random sample of size n, or simply, a random sample.

For reasons that will be explained in subsection 3.2, these designs are of great interest when sampling from continuous populations.

Several schemes have been proposed for selecting samples according to the prescribed probability $p(x)$ either exactly or approximately. Among the exact

methods, we find the inverse–transform method, the composition method and the acceptance–rejection method. Among the approximate methods, we find Markov Chain Monte Carlo methods, e.g. the Metropolis–Hastings algorithm (Hastings, 1970; Metropolis et al., 1953) or Gibbs sampling (see Casella and George 1992). See Rubinstein and Kroese (2008) or Robert and Casella (2013) for a comprehensive description of them. However, each method has both advantages and disadvantages. There is no "ideal" method that allows for easily sampling from any prescribed $p(x)$ (Evans and Swartz, 1995). The second paper of this thesis introduces a method that allows for approximately sampling from $p(x)$.

Definition 22. For a bounded population U, *simple random sampling* –srs– (or uniform random sampling) is the special case of *p*rs where $p(x) = 1/N$ for all $x \in U$ and $N = \int_U dx$.

3.2 Some estimators

In this subsection we define the continuous counterparts of (some of) the estimators introduced in Chapter 2. No proofs are provided, as they are obtained in an analogous way to those already presented.

Let $y(x) = (y_1(x), y_2(x), \cdots, y_{J_y}(x))$ be unknown values of J_y study functions associated to the element x. As in the discrete case, a parameter of U is a function $\theta_U(y)$ into $\mathbb{R}^D$ for some $D \in \mathbb{Z}^+$.

The definitions of estimator and sampling strategy remain the same as in the discrete case, i.e. an *estimator* is a statistic that intends to approximate the parameter, denoted $\hat{\theta}_s(y)$, and a *sampling strategy* is the couple design and estimator, $(p(\cdot), \hat{\theta}_s(y))$.

As in the discrete case, unless stated otherwise we limit ourselves to the case of only one study variable and a univariate parameter. Regarding the parameter, we will consider the continuous counterpart of the total of the study variable, namely, the integral of the study function $\theta_U(y) = \int_U y(x)dx$. We assume that $\theta_U(y)$ is finite.

The *p*-estimator In the continuous case, the strategy that couples *p*rs of size n with the p-estimator, (2.1), is still unbiased for the integral $\theta_U(y)$.

If $\int_U \frac{y^2(x)}{p(x)}dx < \infty$, the variance of the strategy that couples the p-estimator with *p*rs is given by

$$\mathrm{V}_{p\mathrm{rs}}\left[\hat{\theta}_s^p(y)\right] = \frac{1}{n}\int_U p(x)\left(\frac{y(x)}{p(x)} - \theta_U(y)\right)^2 dx. \tag{3.1}$$

Expression (2.3) is still an unbiased estimator of the variance.

In analogy with the discrete case, if all $y(x)$ have the same sign, the choice $p(x) = y(x)/\int_U y(x)dx$ would yield a variance equal to zero. As this is clearly not feasible, we can make use of an auxiliary function. By an auxiliary function we will understand a known function $z(x)$ that can be evaluated at any point x and such that its integral over U is known and finite. One can consider defining $p(x) = z(x)/\int_U z(x)dx$. If z is almost proportional to y, this choice is expected to yield a small variance.

The potential efficiency of the strategy that couples the p-estimator with $p$rs sampling with some prescribed probabilities evidences the need for sampling schemes that allow for selecting such samples. As mentioned above, several methods have been proposed to this end. The second paper of this thesis adds one more method to the list. The method allows to approximate a target density by a mixture of densities that are easy to sample from.

The difference estimator Let $z(x)$ be an auxiliary function. The difference estimator (2.10), with $\theta_U(z) = \int_U z(x)dx$, $\hat{\theta}_s^p(e) = \frac{1}{n}\sum_s \frac{e(x)}{p(x)}$ and $e(x) = y(x) - z(x)$, is still unbiased for the integral $\theta_U(y)$.

The variance of the strategy that couples the difference estimator with $p$rs is given by

$$\mathrm{V}_{prs}\left[\hat{\theta}_s^d(y)\right] = \frac{1}{n}\int_U p(x)\left(\frac{e(x)}{p(x)} - \theta_U(e)\right)^2 dx, \tag{3.2}$$

where $\theta_U(e) = \int_U e(x)dx$ and $e(x)$ as defined above. Expression (2.12) is still an unbiased estimator of the variance.

The ratio estimator In the continuous case, the ratio estimator is defined as in (2.13) with $\theta_U(z) = \int_U z(x)dx$, $\hat{\theta}_s^p(y) = \frac{1}{n}\sum_s \frac{y(x)}{p(x)}$ and $\hat{\theta}_s^p(z) = \frac{1}{n}\sum_s \frac{z(x)}{p(x)}$.

As its discrete counterpart, the ratio estimator is biased, although its bias is usually small even for moderate sample sizes. An approximation to the variance for the strategy that couples the ratio estimator with $p$rs is given by

$$\mathrm{AV}_{prs}\left[\hat{\theta}_s^r(y)\right] = \frac{1}{n}\int_U \frac{e^2(x)}{p(x)}dx \qquad \text{with} \qquad e(x) = y(x) - z(x)\frac{\theta_U(y)}{\theta_U(z)} \tag{3.3}$$

Expression (2.15) provides an estimator of the variance.

The GREG estimator The GREG estimator makes use of several auxiliary functions $z(x) = (z_1(x), \cdots, z_{J_z}(x))$. The estimator is defined as in (2.16) where $\hat{\theta}_s^p(y) = \frac{1}{n}\sum_s \frac{y(x)}{p(x)}$ is the p-estimator of the integral of y, $\theta_U(z) =$

$(\theta_U(z_1), \theta_U(z_2), \cdots, \theta_U(z_{J_z}))$ is the vector of integrals of the auxiliary functions, $\hat{\theta}_s^p(z) = (\hat{\theta}_s^p(z_1), \hat{\theta}_s^p(z_2), \cdots, \hat{\theta}_s^p(z_{J_z}))$ is the vector of corresponding p-estimators and $\hat{B}$ is given by (2.17). As before, the a-values are weights defined by the statistician.

An approximation to the variance of the strategy that couples the GREG estimator with *p*rs is given by

$$\mathrm{AV}_{prs}\left[\hat{\theta}_s^g(y)\right] = \frac{1}{n}\int_U p(x)\left(\frac{e(x)}{p(x)} - \theta_U(e)\right)^2 dx$$

where $\theta_U(e) = \int_U e(x)dx$, $e(x) = y(x) - z(x)B$ with

$$B = \left(\int_U \frac{z(x)^T z(x)}{a(x)}dx\right)^{-1} \int_U \frac{z(x)^T y(x)}{a(x)}dx.$$

Expression (2.20) provides an estimator of the variance with $\hat{\theta}_s^p(e_s) = \frac{1}{n}\sum_s \frac{e_s(x)}{p(x)}$. The residuals $e_s(x)$ are defined as either $e_s(x) = y(x) - z(x)\hat{B}$ or $e_s(x) = g_s(x)\left(y(x) - z(x)\hat{B}\right)$ with $g_s(x) = 1 + (\theta_U(z) - \hat{\theta}_s^p(z))\,\Theta_U^{-1}(z)\,z(x)^T/a(x)$ and $\Theta_U(z) = \int_U z(x)^T z(x)/a(x)dx$.

It is worth mentioning that an extension of the π-estimator (2.4) to the continuous context has been proposed by Cordy (1993). However, this extension is not included in this thesis.

We close this chapter with an example.

Example 14. We are interested in estimating the integral of $y(x_1, x_2) = 3(x_1 + x_2) + \sin(6(x_1 + x_2))$ in the region $U = [0,1] \times [0,1]$. The left panel of Figure 3.1 represents the study function. A sample of size $n = 16$ from *p*rs with $p(x) = 1$ for all $x \in U$ (i.e. simple random sampling) was selected. The sampled points are

(0.7383, 0.9551)	(0.0007, 0.3081)	(0.4304, 0.7654)	(0.7788, 0.9905)
(0.4399, 0.3444)	(0.4027, 0.0814)	(0.5789, 0.0954)	(0.6429, 0.6894)
(0.7349, 0.4423)	(0.5075, 0.3557)	(0.6888, 0.3164)	(0.9148, 0.2129)
(0.9797, 0.1587)	(0.1994, 0.5732)	(0.1071, 0.2911)	(0.4539, 0.2954).

The estimate using the p-estimator (2.1) and the corresponding variance estimate (2.3) are $\hat{\theta}_s^p(y) = 2.8285$ and $\hat{\mathrm{V}}_{prs}[\hat{\theta}_s^p(y)] = 0.1248$. Assuming that the estimator follows a t-distribution with $n - 1 = 15$ degrees of freedom, a 95% confidence interval is $(2.0755, 3.5815)$. Using cubature methods (implemented in the `R` package `cubature`, Narasimhan et al. 2019) the integral

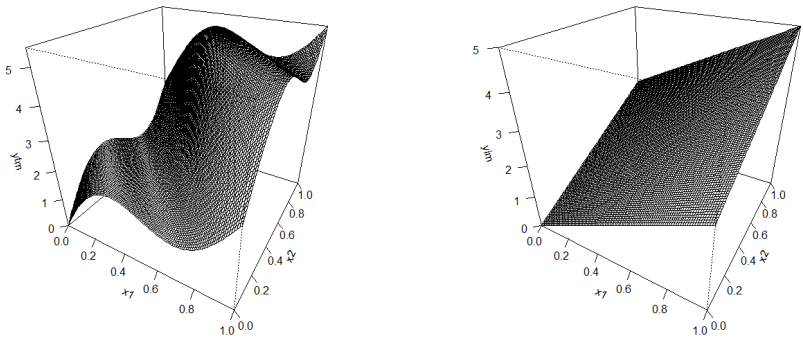


Figure 3.1: The study function $y(x)$ (left panel) and the auxiliary function $z(x)$ (right panel).

is found to be 2.9994, so the confidence interval is, in fact, covering the parameter although, due to the small sample size, the variance is quite large and therefore the confidence interval is too long.

As an auxiliary function we consider $z(x_1,x_2) = 2.5(x_1+x_2)$. It is easy to see that $\theta_U(z) = 2.5$. The right panel of Figure 3.1 represents the auxiliary function. The estimate using the difference estimator (2.10) and the corresponding variance estimate (2.12) are $\hat{\theta}_s^d(y) = 2.9107$ and $\hat{\mathrm{V}}_{prs}[\hat{\theta}_s^d(y)] = 0.0405$. Which, assuming a t-distribution again, yields the 95% confidence interval $(2.4818, 3.3396)$. Not only is the point estimate closer to the parameter than the one obtained using the p-estimator, the resulting confidence interval is almost half the length as the one obtained using the p-estimator.

Using $\hat{\theta}_s^p(z) = 2.4178$, we obtain an estimate using the ratio estimator (2.13) and the corresponding variance estimate (2.15) as $\hat{\theta}_s^r(y) = 2.9247$ and $\hat{\mathrm{V}}_{prs}[\hat{\theta}_s^r(y)] = 0.0427$, which yields a 95% confidence interval of $(2.4842, 3.3651)$. In this case, the results are similar to those obtained with the difference estimator. Note that the biased nature of the ratio estimator does not seem to affect the estimation process.

The sampling process was repeated $B = 1\,000\,000$ times. With each selected sample, estimates of the integral $\theta_U(y)$ and the corresponding variance estimates were calculated with the p-, difference and ratio estimators. Our objective is to study the bias and variance of the three estimators of $\theta_U(y)$, as well as the bias of the corresponding variance estimators and the coverage rate of the confidence intervals. To this end, we approximate each parameter by its

empirical counterpart, it is,

$$\mathrm{E}_{prs}\left[\hat{\theta}_s^i(y)\right] \approx \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_{s,b}^i(y) \quad \mathrm{E}_{prs}\hat{\mathrm{V}}_{prs}\left[\hat{\theta}_s^i(y)\right] \approx \frac{1}{B}\sum_{b=1}^{B}\hat{\mathrm{V}}_{prs,\mathrm{b}}\left[\hat{\theta}_s^i(y)\right]$$

$$\mathrm{cover}_{prs}\left[\hat{\theta}_s^i(y)\right] \approx \frac{1}{B}\sum_{b=1}^{B}I_b^i \quad \mathrm{V}_{prs}\left[\hat{\theta}_s^i(y)\right] \approx \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_{s,b}^i(y) - \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_{s,b}^i(y)\right)^2$$

where $\hat{\theta}_{s,b}^i(y)$ is the point estimate obtained with the ith estimator at the bth iteration; $\hat{\mathrm{V}}_{prs,b}[\hat{\theta}_s^i(y)]$ is the variance estimate obtained with the ith estimator at the bth iteration; and I_b^i equals one if the bth confidence interval for the ith estimator covers the parameter and equals zero otherwise ($i = p,d,r$ and $b = 1,2,\cdots,B$).

Estimator	$\mathrm{E}_{prs}[\hat{\theta}_s(y)]$	$\mathrm{V}_{prs}[\hat{\theta}_s(y)]$	$\mathrm{E}_{prs}\hat{\mathrm{V}}_{prs}[\hat{\theta}_s(y)]$	$\mathrm{cover}_{prs}[\hat{\theta}_s(y)]$
p	2.9995	0.1307	0.1308	0.9471
Difference	2.9995	0.0348	0.0348	0.9478
Ratio	2.9986	0.0323	0.0326	0.9517

Table 3.1: Results of the simulation in Example 14

Table 3.1 summarizes the results of the simulation. Respectively, the second to fifth columns show for each strategy, the empirical expected value, variance, expected value of the variance estimator and coverage. Our main conclusions are as follows:

- Regarding the bias of the estimators, we know that both the p- and the difference estimator are unbiased, the small difference with respect to the true value of the parameter is due to simulation error. Even with a moderate sample size, the bias introduced by using the ratio estimator is almost negligible.

- The variance expressions for the strategies using the p- and the difference estimators, (3.1) and (3.2), are exact. Using cubature methods, we found them to be 0.1308 and 0.0348, respectively. Again, the small differences are due to simulation errors. On the other hand, expression (3.3) is only an approximation to the true variance of the ratio estimator. Using cubature methods, we obtain 0.0312, which is only slightly smaller than the empirical variance of 0.0323.

- The use of auxiliary information in the estimator provides an additional tool for gaining efficiency. The variance of the p-estimator is almost four times larger than that for the estimators using auxiliary information.

- The variance estimators for the p- and the difference estimators, (2.3) and (2.12), are unbiased. Therefore it does not come as a surprise that their empirical expected value is so close to the actual variance. The variance estimator proposed for the ratio estimator (2.15) is not unbiased. However, its empirical expected value (0.0326) was close to the actual variance (0.0323).

- Finally, all three estimators yielded a coverage close to the desired 95%.

4. Summary of papers

Paper I. A comparison of πps designs

Sampling designs where the probability of each element being included in the sample coincides with a prescribed value are known in survey practice as probability proportional to size sampling designs, πps. The main reason why these designs are of practical interest is their potential efficiency. In particular, if we are interested in the estimation of the total of a study variable using the Horvitz–Thompson estimator, a πps design that selects elements without replacement and that assigns each element a probability proportional to the value of the variable of interest would yield perfect estimates for any possible sample. Although this approach is not feasible in practice (as the values of the study variable are unknown when selecting the sample), it makes sense to assume that if an auxiliary variable that is more or less proportional to the study variable is available at the sampling stage, selecting elements with probability proportional to such variable will be a highly efficient choice.

In survey literature there is a multitude of algorithms that allow for selecting πps samples, either exactly or approximately. In this paper we describe ten of these algorithms and we compare them in four different populations. The comparison focuses on several desirable properties, e.g. efficiency, possibility of coordination or availability of second order inclusion probabilities.

The results suggest that Sunter's (1977) and Chromy's (1979) methods are efficient alternatives if there is no need for coordinating samples. Otherwise, order sampling methods, e.g. Pareto or sequential Poisson sampling, may be considered as the best choice. They are not only efficient designs but also simple algorithms.

Paper II. Approximating prescribed distributions by mixtures

There are multiple properties that may be of interest before a given distribution, for example, its density function, distribution function, quantile function, expected value, variance, sample selection or, in the multivariate case, marginal and conditional distributions. It is true that these properties are widely known

for the most common distributions —e.g. Gaussian, beta, gamma or Poisson, just to name a few—, however, this is not true for any arbitrary distribution.

Clearly, for a given target distribution, the ideal solution would be to obtain its properties in an analytic form. Nevertheless, in many cases this may be a cumbersome —if not impossible— task. For this reason, methods yielding results in an approximate way have become more accepted in practice, e.g. Markov chain Monte Carlo methods. In this article we introduce an algorithm that allows to approximate an arbitrary distribution by a mixture distribution.

Being a mixture, most of its properties are easy to calculate —as they are "inherited" from its components—. The algorithm is flexible as it only requires the target density function to be known up to a proportionality constant. This characteristic makes it atractive for approximating the posterior distribution in Bayesian statistics.

We propose also diagnostics for measuring the convergence of the approximation to the target density. Moreover, in a first run, there may be some redundancy in the components of the mixture. We propose a method for collapsing redundant components so obtaining a more parsimonious approximation.

The implementation of the algorithm is illustrated with several examples, both univariate and multivariate. The results suggest that the algorithm succeeds in approximating the target distribution.

Paper III. A method to find an efficient and robust sampling strategy under model uncertainty

We consider the choice of sampling design to be implemented in a survey where auxiliary information is available. It has been shown that if the relation between auxiliary variables and the study variable is explained by a model —commonly known as superpopulation model— consisting of two parts (trend and spread), the optimal sampling strategy is the one using the difference estimator explaining the trend together with a probability proportional to size sampling design, πps, explaining the spread (see e.g. Cassel et al. 1976; Hájek 1959; Nedyalkova and Tillé 2008).

Although very powerful, this result has practical limitations as it assumes the model parameters to be known. Regarding the estimator, one solution is to use the generalized regression -GREG- estimator instead of the difference estimator, as it allows for estimating the trend parameters once the sample data has been collected. In fact, it can be shown that the GREG estimator is asymptotically optimal.

Regarding the sampling design, however, a common practice is to take the spread parameters from previous studies (similar to the one under planning). In

this article we show that when the superpopulation model is misspecified, the strategy πps–GREG is no longer optimal. Even mild misspecifications of the model or its parameters may lead to this strategy being inefficient. We introduce a method that allows for incorporating uncertainty about the parameters in the choice of the sampling design. In this way a measure of the risk can be obtained and the chosen sampling design should be the one that minimices this measure.

Although it is true that the uncertainty is subjective, it is also true that completely relying on the assumed parameters is subjective as well. The proposed method allows for quantifying the uncertainty with respect to the model parameters. The method is illustrated with a real dataset. The results show that it allowed us to correctly choose the sampling design.

Paper IV. optimStrat: An R package for assisting the choice of robust and efficient sampling strategies

In this paper we describe our package `optimStrat`, developed for `R` (R Core Team, 2020). The package includes a set of functions that may be used at the design stage of a survey. In particular, it includes a web-based application in `shiny` (Chang et al., 2020) that allows the user to implement the risk measure developed in the third paper and, in this way, deciding on the sampling design to be implemented in a survey.

The web-based application calculates the risk of five sampling strategies under a superpopulation model specified by the user. The five sampling strategies under comparison are: stratified simple random sampling (STSI) with the Horvitz–Thompson estimator, STSI with the poststratified estimator, STSI with the regression estimator, probability proportional to size (πps) sampling with the poststratified estimator and πps with the regression estimator.

Other functions in the package allow to stratify the population using the cum $\sqrt{f}$ rule (Dalenius and Hodges, 1959), allocate a sample to strata by means of Neyman optimal allocation, compute the variance of the generalized regression estimator, and, simulate study variables according to a superpopulation model specified by the user and a set of auxiliary variables.

5. Sammanfattning

Urvalsprocessen utgör kärnan i varje undersökning. En noggrant utformad urvalsdesign kan medföra inte bara bättre skattningar av de parametrar som är av intresse, utan också en reducering av erforderlig urvalsstorlek. I denna avhandling behandlas två särskilda ämnesområden: å ena sidan urvalsförfaranden med sannolikheter proportionella mot föreskrivna värden. De två första artiklarna ägnas åt detta, å andra sidan valet av urvalsdesign för implementering i en given undersökning, ett ämne som två sista artiklarna ägnas åt.

Urvalsdesigner med sannolikheter proportionella mot ett storleksmått —πps— är av praktisk betydelse på grund av deras potentiella effektivitet. I litteraturen kan vi hitta många av dessa designer, alla med olika karaktäristika. I den första artikeln beskriver och jämför vi tio πps-designer med avseende på flera önskvärda egenskaper. Resultaten tyder på att de så kallade ordningsurvalen såväl som dem föreslagna av Sunter och Chromy kan betraktas som bra val ur en praktisk synvinkel.

I den andra artikeln introduceras en algoritm för att approximera en given fördelning genom en blandad fördelning. Genom att vara just en blandning är de flesta av dess egenskaper enkla att bestämma. Vi illustrerar användningen av algoritmen med flera exempel, både univariata och multivariata. Resultaten indikerar att algoritmen lyckas med att approximera den avsedda fördelningen.

Strategin som kopplar samman πps-designer med den generaliserade regressionsskattningen är optimal given en superpopulationsmodell. Denna optimalitet är emellertid avhängig av att modellen är korrekt, ett antagande som knappast är uppfyllt i verkligheten. I den tredje artikeln introducerar vi en metod som möjliggör en inkorporering av osäkerhet rörande parametrar med valet av urvalsdesign, för att sedan kvantifiera denna osäkerhet genom ett riskmått. Resultaten visar att metoden klarar av att välja korrekt urvalsdesign. Detta riskmått —såväl som andra funktioner som är användbara i planeringsstadiet av en undersökning— är implementerat i paketet optimStrat som Ãr utvecklat för `R`. Den fjärde artikeln i avhandlingen beskriver funktionerna i detta paket.

References

Biemer, P. and Lyberg, L. (2003). *Introduction to survey quality*, volume 335. John Wiley & Sons. 1

Breidt, F. J. and Opsomer, J. D. (2000). Local polynomial regresssion estimators in survey sampling. *The annals of statistics*, 28(4):1026–1053. 17

Casella, G. and George, E. (1992). Explaining the Gibbs sampler. *The American Statistician*, 46(3):167–174. 24

Cassel, C. M., Särndal, C. E., and Wretman, J. H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63(3):615–620. 32

Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J. (2020). *shiny: Web Application Framework for R*. R package version 1.4.0.2. 33

Chromy, J. (1979). Sequential sample selection methods. In *Proceedings of the Survey Research Methods Section of the American Statistical Association*, pages 401–406. 31

Cordy, C. B. (1993). An extension of the Horvitz-Thompson theorem to point sampling from a continuous universe. *Statistics & Probability Letters*, 18(5):353–362. 26

Dalenius, T. (1985). *Elements of Survey Sampling: Notes Prepared for the Swedish Agency for Research Countries (SAREC)*. Sarec. 1

Dalenius, T. and Hodges, J. (1959). Minimum variance stratification. *Journal of the American Statistical Association*, 54(285):88–101. 33

Devroye, L. (1986). *Nonuniform random variate generation*. Springer-Verlag. 7

Evans, M. and Swartz, T. (1995). Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems. *Statistical science*, 10(3):254–272. 24

Hájek, J. (1959). Optimal strategy and other problems in probability sampling. *Časopis pro pěstování matematiky*, 84(4):387–423. 32

Hajek, J. (1981). *Sampling from a finite population*. Marcel Dekker, Incorporated. 13

Hansen, M. H. and Hurwitz, W. N. (1943). On the theory of sampling from finite populations. *The Annals of Mathematical Statistics*, 14(4):333–362. 10

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109. 24

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685. 10

Lohr, S. (2009). *Sampling: Design and Analysis*. Cengage learning. 13

Madow, W. (1949). On the theory of systematic sampling, II. *The Annals of Mathematical Statistics*, 20(3):333–354. 12

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092. 24

Montanari, G. E. and Ranalli, M. G. (2005). Nonparametric model calibration estimation in survey sampling. *Journal of the American Statistical Association*, 100(472):1429–1442. 17

Narasimhan, B., Johnson, S., Hahn, T., Bouvier, A., and Kiêu, K. (2019). *cubature: Adaptive Multivariate Integration over Hypercubes*. R package version 2.0.4. 26

Nedyalkova, D. and Tillé, Y. (2008). Optimal sampling and estimation strategies under the linear model. *Biometrika*, 95(3):521–537. 32

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. 2, 13, 21, 33

Robert, C. and Casella, G. (2013). *Monte Carlo statistical methods*. Springer Science & Business Media. 15, 24

Rondon, L. M., Vanegas, L. H., and Ferraz, C. (2012). Finite population estimation under generalized linear model assistance. *Computational Statistics & Data Analysis*, 56(3):680–697. 17

Rosén, B. (1997). On sampling with probability proportional to size. *Journal of Statistical Planning and Inference*, 62(2):159–191. 12, 14

Rubinstein, R. Y. and Kroese, D. P. (2008). *Simulation and the Monte Carlo method*. John Wiley & Sons. 6, 10, 15, 24

Sampford, M. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika*, 54(3-4):499–513. 12

Särndal, C., Swensson, B., and Wretman, J. (1992). *Model assisted survey sampling*. Springer Science & Business Media. 1, 3, 10, 17, 18

Sunter, A. (1977). List sequential sampling with equal or unequal probabilities without replacement. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 26(3):261–268. 31

Swensson, B. (1998). On Pareto πps sampling. *Unpublished manuscript*. 14

Verbeke, T. (2014). *SDaA: Sampling: Design and Analysis*. R package version 0.1-3. 13