DISSERTATIONES MATHEMATICAE UNIVERSITATIS TARTUENSIS
54

DISSERTATIONES MATHEMATICAE UNIVERSITATIS TARTUENSIS
54

RESTRICTION ESTIMATOR FOR DOMAINS

KAJA SÕSTRA

TARTU UNIVERSITY
PRESS

Faculty of Mathematics and Computer Science, University of Tartu, Tartu

Dissertation is accepted for the commencement of the degree of Doctor of Philosophy (Ph.D.) in mathematical statistics on October 22, 2007, by the Council of the Faculty of Mathematics and Computer Science, University of Tartu

Supervisor:
Associate Professor, Cand. Sc. Imbi Traat
University of Tartu
Tartu, Estonia

Opponents:
Professor, Ph.D. Risto Lehtonen
University of Helsinki
Helsinki, Finland

Associate Professor, Cand. Sc. Ebu Tamm
Tallinn University of Technology
Tallinn, Estonia

The public defence will take place on December 21, 2007

ISSN 1024−4212
ISBN 978−9949−11−749−9 (trükis)
ISBN 978−9949−11−750−5 (pdf)

Tartu Ülikooli Kirjastus
www.tyk.ee
Tellimus nr. 481

Contents

List of original publications

Papers in refereed journals

1. Sõstra, K. (2004) Comparison of Small Area Estimation Methods: Simulation Study in EURAREA Project. *Acta et Commentationes Universitatis Tartuensis de Mathematica*, 8, 243−252

2. Eamets, R., Varblane, U., Sõstra, K. (2003) External Macroeconomic Shocks and the Estonian economy: How did the Russian Financial Crisis affect Estonian Unemployment and Foreign Trade? *Baltic Journal of Economics*, Vol 3, No 2 Spring/Summer 2003, pp 5−24

3. Kurvits, M., Sõstra, K., Traat, I. (2002) The Estonian Household Sample Surveys - Focus on the Labour Force Survey. *Statistics in Transition*, 5(4), 605−616

4. Traat, I., Meister, K., Sõstra, K. (2001) Statistical inference in sampling theory. *Theory of Stochastic Processes*, vol. 7(23), 301−316

5. Traat, I., Kukk, A., Sõstra, K. (2000) Sampling and Estimation Methods in the Estonian Household Budget Survey. *Statistics in Transition*, vol. 4, No. 6, pp. 1029−1046

Other publications

1. Ollila, P., Laaksonen, S., Sõstra, K., Berger, Y., Boonstra, H-J., van den Brakel, J., Davison, A., Sardy, S. Magg, K., Münnich, R., Ohly, D. (2004) *Evaluation of Software for Variance Estimation in Complex Surveys.* Data Quality in Complex Surveys within the New European Information Society (DACSEIS) Project Research Papers under Workpackage 4. URL http://www.dacseis.de. - IST-2000-26057-DACSEIS Reports

2. Münnich, R., Magg, K., Sõstra, K., Schmidt, K., Wiegert, R. (2004) *Variance Estimation for Small Area Estimates.* DACSEIS Project Research Papers under Workpackage 10. URL http://www.dacseis.de. - IST-2000-26057-DACSEIS Reports

3. EURAREA Consortium with K. Sõstra among the members (2004) *Enhancing Small Area Estimation Techniques to meet European Needs* (*EURAREA project*) Final Reference Vol.1−3
https://www.statistics.gov.uk/eurarea/default.asp

4. Leetmaa, R., Võrk, A., Eamets, R., Sõstra, K. (2003) *Evaluation of Active Labor Market Policies in Estonia* (*in Estonian with English summary*). Center for Policy Studies PRAXIS, Tallinn

K. Sõstra has published five more methodological papers of official statistics in publications of Statistics Estonia (2004−2007) and has been a co-author of six publications of the Estonian Labour Force Survey (1997−2000) and of the publication Earnings 2005 (2006).

Acknowledgements

I would like to express my gratitude to my supervisor Imbi Traat for her advice and support during the whole process of my PhD studies and during all phases writing this Thesis.

I am also grateful to my family and friends for their support and understanding during my studies.

I would like to thank all my colleagues in Statistics Estonia and Statistics Finland for their support and encouragement. Many thanks to Professor Gunnar Kulldorff at University of Umeå for his support during my studies.

Introduction

Domain or population subgroup estimation has become an important area in survey sampling. Growing demand for reliable domain statistics has forced rapid developments in theory. Main concern has been the small sample size in some domains, called small areas. Different small area estimation methods have been developed to improve the precision of estimates (Rao, 2003). Several research papers have been written (Lehtonen et al. 2003, 2005) and large research projects have been carried out to compare the performance of different methods in real sampling situation (EURAREA consortium, 2004).

Another problem with domain estimation, being the topic of the present thesis, is the lack of consistency between estimates. It is known that the domain and population parameters satisfy certain relationships, e.g. domain totals have to sum up to population total, a different decomposition of domains demands that certain relationships between the new and old domains hold. These relationships often do not hold for domain estimates, i.e. estimates are not consistent with each other. Inconsistent estimates are not acceptable for statistics users. Consistency of statistical data is also an important quality aspect in the European Statistical System (European Commission, 2005, Principle 14). Inconsistencies occur due to several reasons: estimators are random, domain estimators are not additive, different estimation methods are used for different domains, estimates are taken from different surveys, only some domain or population parameters are known, others are to be estimated. Described problem occurs both for small and large domains. Usually simple ad hoc methods are used in statistical agencies to achieve the consistency of domain estimates. For example, in Statistics Estonia the problem arose with Foreign affiliate trade statistics (FATS) where the survey results were published already and the additional domain estimates were needed. In the present case the adjustment of domain estimates using ratios of new and published estimates was used. In Statistics Netherlands the repeated weighting method is developed for solving the consistency problem (van Duin, Snijders, 2003 and Houbiers, Knottnerus et al. 2003). The raking ratio method (Deming and Stephan 1940) helps to solve

iteratively the consistency problem for frequency tables. Usually no statements are made that these methods are the best possible (in the sense of estimator variances) among many alternatives. Often the variance formulae are even not known. In addition, aiming to achieve the consistency of estimates in several levels of domains makes this problem mathematically quite complex.

The main goals of this thesis are:

- to develop the domain estimators that satisfy known restrictions in some simple but important practical cases and being optimal in some class of estimators;
- to develop the variance/covariance expressions of the estimators;
- to test and illustrate the theoretical results in simulation studies.

Recently, a general restriction estimator is presented by Knottnerus (2003) which handles the consistency problem of estimates in sample surveys. The linear and nonlinear restrictions are allowed. In fact, estimation under restrictions is an old problem in mathematics and statistics which can be dated back even to the works of Gauss in the 19th century (Hald, 1998, chapter 21). Knottnerus seems to be the first one who uses these ideas in sample surveys. He briefly presents his estimator and some of its properties, gives relationships with other well-known estimators, also involving restrictions but only on certain auxiliary variables (Deville and Särndal 1992, Montanari 1987). He also gives some examples of applications. However, there are many other fields in sample surveys where the restriction estimator can be applied but which are not covered in this book. Domain estimation is also such a case.

In present thesis the restriction estimator of Knottnerus is developed for domain estimation with the aim to satisfy known relationships between estimates. Important minimum variance property of the Knottnerus restriction estimator passes to our new domain estimators. Therefore the domain estimators developed in present thesis are the best possible among all other estimators constructed on the same initial estimators and which satisfy the same restrictions.

The approach used in present thesis is the design-based one, i.e. the properties of estimators such as expectation and variance/covariance are determined by the sampling design. The estimators have different properties under different sampling designs. Two special sampling designs are considered here, simple random sampling without replacement (SI) and hypergeometric (HG)

sampling. The SI-design represents an equal probability design (all population units have an equal probability of inclusion into sample). The HG-design represents an unequal probability design. It also represents with-replacement design (WR). Both are widely used sampling designs.

The design-based approach is elaborated in the sampling vector framework (Traat, 2000). This framework allows handling the without-replacement (WOR) and the WR designs jointly in a unified manner. Therefore, many results of this thesis are very general and hold thus both for WOR and WR designs.

The sample sizes in domains are assumed to be not too small, i.e. small area estimation methods are not considered in this thesis. Instead, two direct domain estimators are considered, a linear estimator and a ratio estimator (a nonlinear estimator). The estimator called linear is known as the Horvitz-Thompson estimator under WOR designs (Horvitz, Thompson, 1952) and as the Hansen-Hurwitz estimator under WR designs (Hansen, Hurvitz, Madow, 1953). Both are direct domain estimators in the sense that they use data only from that domain. Ratio estimator uses additionally an auxiliary information which is a known domain total of the auxiliary variable. These estimators are considered as initial estimators when constructing general restricted domain estimators.

The thesis is organised in the following way.

Chapter 1 gives an overview of known yet necessary for this thesis results. However, presentation in the sampling vector framework makes the results more general, valid simultaneously both for WOR and WR sampling designs. In this way, results known from literature become special cases of ours. The two estimators of the population total, namely the linear and the ratio estimator are introduced. Their general variance and covariance formulae are given. For the ratio estimator these formulae are approximate, derived from the linear part of the Taylor expansion. These results are used later for domain estimators by suitable modification of the involved variables. Three special sampling designs are described: SI-, HG- and latters' approximation, multinomial sampling design. Covariance formulae of the observed estimators are derived under these designs from general results. The covariance formulae under HG-design are novel.

In Chapter 2 attention is turned to domains. Linear and ratio estimators are modified to estimate domain totals. General variance and covariance formulae are derived for domain estimators, also their special cases for SI- and HG-designs are derived. Dependence of domain estimators is an interesting issue, but so far, it has found only little attention in the sampling literature. For

WOR case the covariance of domain ratio estimators is given in Särndal et al. (1992, pp. 395 and 413), the covariance matrix of the ratio estimator is given in Lehtonen, Pahkinen (1995), for complex sampling design. Our formulae allow to make several interesting conclusions on the dependence between domain estimators. They are commented in the work. In this thesis the covariance matrix of domain estimators is an important building block of restricted domain estimators, constructed later. Two examples on a small population are given to illustrate the dependence structure of domain estimators numerically. The results of this chapter are novel in the sense that they are expressed in the unifying sampling vector framework, they hold both for WOR and WR-designs. The results for the HG-design are new.

Chapter 3 gives main results of this thesis. The results given in Theorems 3.2−3.4, as well as in the Corollaries are new. First of all an overview on the concept of the general restriction (GR) estimator (Knottnerus, 2003) together with its main properties is given. The GR-estimator solves the estimation problem under known restrictions on the parameters. In this thesis, GR-estimators are derived for domain as well as population totals. Corresponding variance/covariance formulae are also derived. Results are presented in three cases: population total is known, population total is estimated from the same or another survey, population total is estimated but kept conditionally fixed. When developing domain GR-estimators, linear and ratio initial estimators are assumed, i.e. respective covariance matrices are used in GR-estimators. Special cases under SI- and HG-designs are presented. Besides satisfying restrictions, the importance of domain GR-estimators stands in the minimal variance property (excluding conditional restriction estimator). In this way, the domain GR-estimator may serve as a benchmark when evaluating other (ad hoc) domain estimators under restrictions. Study of the analytical form of domain GR-estimators reveals how to construct other simpler restriction estimators (without using variances/covariances of initial estimators) that are still close to the optimal. Remarks on this issue are made.

Chapter 4 presents results of simulations. The aim is to evaluate the performance of the GR-estimator in a practical situation where SI- and HG-designs with sample sizes on the average of 200 persons were carried out in the population of 2000 persons with three domains. It is shown that the restrictions are satisfied for the GR-estimator and the variance of GR-estimator is smaller than the one of the initial estimator. The dependence between domain estimators both of the initial and of the GR-estimators is illustrated by correlations. Another aim is to check the derived variance/covariance formulae with the emphasis on the asymptotic ones. It is demonstrated that the variance of the conditional restriction estimator can be bigger than the one of the initial es-

timator; the components of the variance are illustrated on the figure. All the derived formulae work well and describe adequately real situation (despite the modest sample size for the asymptotic results).

Chapter 5 summarizes main results and contributions of the present Thesis.

Chapter 1

Preliminaries

1.1 Estimation of population parameters

Let $U = (1, 2, \ldots, N)$ denote a finite population of N units. Let a random vector (design vector) $\mathbf{I} = (I_1, I_2, \ldots, I_N)$ describe the sampling process on U. Elements I_i show the number of possible selections of the unit $i \in U$, whereas $I_i \in \{0, 1\}$ for without-replacement (WOR) designs and $I_i \in \{0, 1, 2, \ldots\}$ for with-replacement (WR) designs. The distribution of $\mathbf{I}$ is sampling design, $p(\mathbf{k}) = Pr(\mathbf{I} = \mathbf{k})$, where $\mathbf{k} = (k_1, k_2, \ldots, k_N)$ is an outcome of $\mathbf{I}$ (Traat et al. 2004, Tillé, 2006). The moments of $\mathbf{I}$, such as $E(I_i)$, $V(I_i)$ and $Cov(I_i, I_j)$ play a crucial role in finite population estimation theory. It is assumed that $E(I_i) > 0, \forall i$ for any sampling design. In the case of WOR design, the inclusion indicator I_i is a random variable with a Bernoulli distribution, $I_i \sim B(1, \pi_i)$. In this case

$$\begin{aligned} E(I_i) &= \pi_i, \qquad V(I_i) = \pi_i(1 - \pi_i), \\ Cov(I_i, I_j) &= \pi_{ij} - \pi_i \pi_j, \end{aligned}$$

where $\pi_i = Pr(I_i = 1)$ and $\pi_{ij} = Pr(I_i = 1, I_j = 1)$ are the first- and second-order inclusion probabilities respectively.

Hereafter, unless a special need occurs, a shorter form for sums is used. A sum in the form $\sum_B a_i$ means that index i takes all the values in B, $\sum_B a_i = \sum_{i \in B} a_i$. Similarly, $\sum \sum_B a_{ij} = \sum_{i \in B} \sum_{j \in B} a_{ij}$.

The unbiased estimator for the population total $Y = \sum_U y_i$ under any sampling design, and corresponding variance formulae are known. For WOR case

they are given e.g. in Särndal et al. (1992, p. 43). Throughout this thesis a more general presentation, covering both WOR and WR cases, is used (Traat, 2000, Traat, Meister, Sõstra, 2001, Tillé, 2006, Meister, 2004). Correspondingly, the unbiased estimator $\hat{Y}$ of Y and its variance are:

$$\hat{Y} = \sum\nolimits_U I_i \breve{y}_i = \sum\nolimits_U \omega_i y_i, \tag{1.1}$$

$$V(\hat{Y}) = \sum \sum\nolimits_U \Delta_{ij} \breve{y}_i \breve{y}_j, \tag{1.2}$$

where $\breve{y}_i = y_i / E(I_i)$ and $\Delta_{ij} = Cov(I_i, I_j)$. Provided that $E(I_i I_j) > 0$, $\forall i \neq j$, the unbiased estimator of variance (1.2) is

$$\hat{V}(\hat{Y}) = \sum \sum\nolimits_U \breve{\Delta}_{ij} \omega_i y_i \omega_j y_j, \tag{1.3}$$

where $\breve{\Delta}_{ij} = \Delta_{ij} / E(I_i I_j)$. The quantity ω_i is a design weight

$$\omega_i = I_i / E(I_i). \tag{1.4}$$

Through this thesis the estimator (1.1) is referred to as a linear estimator. In the case of WOR designs, it is known as the Horvitz-Thompson estimator, in the case of WR designs as the Hansen-Hurwitz estimator. Since $\omega_i = 0$ for nonsampled elements, all the sums over U involving weights are, in fact, sample sums. For a sampled unit i the weight is $\omega_i = 1/E(I_i)$ under WOR designs and $\omega_i = k_i / E(I_i)$ under WR designs (k_i is the number of selections of unit i).

It holds for fixed size n sampling designs:

$$\sum\nolimits_U I_i = n, \quad \sum\nolimits_U E(I_i) = n, \tag{1.5}$$

$$\sum_{j \in U} \Delta_{ij} = E\left\{ [I_i - E(I_i)] \sum_{j \in U} [I_j - E(I_j)] \right\} = 0, \sum_{i \in U} \Delta_{ij} = 0. \tag{1.6}$$

Using (1.5) − (1.6) it is easy to see that for fixed size sampling designs, the variance (1.2) of $\hat{Y}$ can be written alternatively:

$$V(\hat{Y}) = -\frac{1}{2} \sum \sum\nolimits_U \Delta_{ij} (\breve{y}_i - \breve{y}_j)^2. \tag{1.7}$$

In the case of $E(I_i I_j) > 0$, $i, j \in U$, an obvious unbiased estimator of $V(\hat{Y})$ under fixed size sampling design is

$$\hat{V}(\hat{Y}) = -\frac{1}{2} \sum \sum\nolimits_U I_i I_j \breve{\Delta}_{ij} (\breve{y}_i - \breve{y}_j)^2. \tag{1.8}$$

The estimator (1.8) is called Sen-Yates-Grundy (SYG) variance estimator (Sen, A. R., 1953, Yates, F., Grundy, P. M., 1953). It is more stable than (1.3) applied under fixed size sampling designs.

The covariance of two linear unbiased estimators $\hat{Y} = \sum_U \omega_i y_i$ and $\hat{X} = \sum_U \omega_i x_i$ has a similar expression to the variance. In WOR case it is given in e.g Särndal et al (1992, p. 170). The SYG form of covariance in WOR case is given in Knottnerus (2003, p. 307). Below we present these results more generally for both the WOR and WR designs.

Theorem 1.1. The covariance of $\hat{Y} = \sum_U \omega_i y_i$ and $\hat{X} = \sum_U \omega_i x_i$ is

$$Cov(\hat{Y}, \hat{X}) = \sum\sum_U \Delta_{ij} \breve{y}_i \breve{x}_j. \tag{1.9}$$

Its unbiased estimator under any design with $E(I_i I_j) > 0$ is:

$$\widehat{Cov}(\hat{Y}, \hat{X}) = \sum\sum_U \breve{\Delta}_{ij} \omega_i y_i \omega_j x_j \tag{1.10}$$

and the SYG-type estimator under fixed size sampling designs is

$$\widehat{Cov}(\hat{Y}, \hat{X}) = -\frac{1}{2} \sum\sum_U I_i I_j \breve{\Delta}_{ij} (\breve{y}_i - \breve{x}_j)^2. \tag{1.11}$$

Proof: The covariance of two estimators $\hat{Y} = \sum_U \omega_i y_i$ and $\hat{X} = \sum_U \omega_i x_i$ is by definition

$$Cov(\hat{Y}, \hat{X}) = E[(\hat{Y} - Y)(\hat{X} - X)], \tag{1.12}$$

where $X = \sum_U x_i$. Using the alternative forms $\omega_i y_i = I_i \breve{y}_i$ and $\omega_i x_i = I_i \breve{x}_i$ in $\hat{Y}$ and $\hat{X}$ respectively, we get

$$\begin{aligned} \hat{Y} - Y &= \sum_U I_i \breve{y}_i - \sum_U y_i = \sum_U [I_i - E(I_i)] \breve{y}_i, \\ \hat{X} - X &= \sum_U [I_i - E(I_i)] \breve{x}_i. \end{aligned}$$

Now (1.12) takes the form:

$$\begin{aligned} Cov(\hat{Y}, \hat{X}) &= E\left\{ \sum\sum_U [I_i - E(I_i)] \breve{y}_i [I_j - E(I_j)] \breve{x}_j \right\} \\ &= \sum\sum_U E[I_i - E(I_i)][I_j - E(I_j)] \breve{y}_i \breve{x}_j. \end{aligned}$$

Since $E[I_i - E(I_i)][I_j - E(I_j)] = Cov(I_i, I_j) = \Delta_{ij}$ we get (1.9). The unbiasedness of (1.10) for (1.9) can be immediately seen by replacing $\omega_i = I_i / E(I_i)$ in (1.10) and taking expectations.

The alternative covariance formula for fixed size sampling designs is

$$Cov(\hat{Y}, \hat{X}) = -\frac{1}{2} \sum\sum_U \Delta_{ij} (\breve{y}_i - \breve{x}_j)^2. \tag{1.13}$$

One can see that it equals to (1.9) by opening the brackets and applying (1.5) – (1.6) to the terms. Obviously, (1.11) is an unbiased estimator of (1.13).

□

Correlation of the two estimators is by definition

$$Cor(\hat{Y}, \hat{X}) = \frac{Cov(\hat{Y}, \hat{X})}{\sqrt{V(\hat{Y})V(\hat{X})}}. \tag{1.14}$$

One of the most important parameters in sample surveys is a population ratio $R = Y/H$, where Y and H are population totals of y and h variables respectively. For example, population means and proportions can be seen as ratios. The ratio R is estimated by $\hat{R} = \hat{Y}/\hat{H}$, where $\hat{Y} = \sum_U \omega_i y_i$ and $\hat{H} = \sum_U \omega_i h_i$ are unbiased estimators of Y and H. If the total H is known (auxiliary information) then another estimator of Y, called ratio estimator, can be constructed:

$$\hat{Y}^r = \hat{R}H, \tag{1.15}$$

The estimator $\hat{Y}^r$ is nonlinear. Usually Taylor expansion is used to find its properties. Särndal et al. (1992, p. 178) gives a linear part of the Taylor expansion of $\hat{R}$:

$$\hat{R} \approx R + \frac{1}{H}(\hat{Y} - R\hat{H}). \tag{1.16}$$

From here the linear part for $\hat{Y}^r$ is:

$$\hat{Y}^r \approx Y + (\hat{Y} - R\hat{H}). \tag{1.17}$$

The expansions (1.16) – (1.17) are used to derive approximate variance formulae. For WOR designs they are given in many sources, including Särndal et al. (1992, p. 178-179). Here we derive the approximate covariance formulae of two ratio estimators. We do it in a general level covering both the WOR and WR designs. The variance formula follows from that result as a special case. Ratio estimator is a special case of the general regression estimator (Särndal, et al., 1992). Rajaleid (2004) has derived the covariance matrix of a vector of GREG-estimators in the sampling vector framework, but not the estimator of that covariance matrix.

Theorem 1.2. Let Y and X be two totals under estimation. Let $R_y = Y/H$ and $R_x = X/H$. The approximate covariance of two ratio estimators $\hat{Y}^r = \hat{R}_y H$ and $\hat{X}^r = \hat{R}_x H$ with $\hat{R}_y = \hat{Y}/\hat{H}$ and $\hat{R}_x = \hat{X}/\hat{H}$ is

$$ACov(\hat{Y}^r, \hat{X}^r) = \sum\sum\nolimits_U \Delta_{ij} \breve{u}_i \breve{v}_j, \tag{1.18}$$

where

$$u_i = y_i - R_y h_i \text{ and } v_i = x_i - R_x h_i. \tag{1.19}$$

An estimator of (1.18) under any design with $E(I_i I_j) > 0$ is

$$\widehat{Cov}(\hat{Y}^r, \hat{X}^r) = \sum \sum\nolimits_U \breve{\Delta}_{ij} \omega_i \tilde{u}_i \omega_j \tilde{v}_j, \tag{1.20}$$

where

$$\tilde{u}_i = y_i - \hat{R}_y h_i \text{ and } \tilde{v}_i = x_i - \hat{R}_x h_i. \tag{1.21}$$

The SYG-type estimator for fixed size sampling designs is

$$\widehat{Cov}(\hat{Y}^r, \hat{X}^r) = -\frac{1}{2} \sum \sum\nolimits_U I_i I_j \breve{\Delta}_{ij} (\breve{\tilde{u}}_i - \breve{\tilde{v}}_j)^2. \tag{1.22}$$

Proof: Inserting estimators $\hat{Y} = \sum_U \omega_i y_i$, $\hat{X} = \sum_U \omega_i x_i$ and $\hat{H} = \sum_U \omega_i h_i$ in the linear parts of Taylor expansions, they can be presented in the following form:

$$\begin{aligned} \hat{Y}^r &\approx Y + (\hat{Y} - R_y \hat{H}) = Y + \sum\nolimits_U \omega_i u_i, \\ \hat{X}^r &\approx X + (\hat{X} - R_x \hat{H}) = X + \sum\nolimits_U \omega_i v_i. \end{aligned}$$

Now Y and X are fixed numbers and do not affect the covariance. Consequently, we have to find the covariance of two linear estimators which is done in Theorem 1.1. The formula (1.18) follows directly from (1.9). Direct application of (1.10) gives:

$$\widehat{Cov}(\hat{Y}^r, \hat{X}^r) = \sum \sum\nolimits_U \breve{\Delta}_{ij} \omega_i u_i \omega_j v_j.$$

Since u_i and v_j include population values R_y and R_x which are not known, they will be replaced by $\hat{R}_y$ and $\hat{R}_x$ and so the formula (1.20) follows. The formula (1.22) comes analogically from Theorem 1.1.

□

Remark 1.1. If $y_i \equiv x_i$ then $\hat{Y}^r \equiv \hat{X}^r$ in which case Theorem 1.2 gives an approximate variance formulae of $\hat{Y}^r$. They hold both for WOR and WR designs. In some sources (e.g. Särndal, 1992) the variance estimator formulae include the coefficient $(H/\hat{H})^2$ which is close to one for large sample sizes. This coefficient is obtained first deriving the variance of $\hat{R}$ and then the variance of $\hat{Y}^r = \hat{R}H$. There is empirical evidence that the coefficient $(H/\hat{H})^2$ makes variance estimator more stable.

1.2 Special sampling designs and estimation

Three sampling designs are briefly introduced here: simple random sampling (SI), hypergeometric (HG) and multinomial (M) sampling designs. They are considered later in this thesis when developing special cases of the results. These are common designs in official statistics. For example, simple random sampling and stratified simple random sampling are used for sample surveys of businesses. Businesses are stratified according to number of employees and economic activity and SI-design is used in every strata. Hypergeometric design describes selection mechanism in social surveys where we select individuals from population register and include all persons of their households into the sample. Multinomial design is used as an approximation of the HG-design, which is justified in usual survey situation. Formulae under multinomial sampling design are simpler than under HG-design.

1.2.1 Simple random sampling

Under SI-design all samples with fixed size are equally probable. Characteristics of the SI-design with population size N, sample size n and sampling fraction $f = n/N$ are for all $i, j \in U$ (Särndal, 1991, p. 66-72, Cochran, 1977, p. 28-29):

$$E(I_i) = f,\ V(I_i) = \Delta_{ii} = f\,(1-f)\,, \tag{1.23}$$

$$E(I_i I_j) = f\frac{(n-1)}{(N-1)},\ i \neq j, \tag{1.24}$$

$$\Delta_{ij} = -f(1-f)\frac{1}{(N-1)},\ i \neq j. \tag{1.25}$$

These formulae are crucial when developing design-based properties of estimators under SI-design. This design is well studied in the literature, though the covariance formulae have got little attention. Here we bring the covariance formulae of linear estimator under SI-design. Under our presentation they follow from Theorem 1.1 by using characteristics (1.23) − (1.25). We formulate the result as a Theorem since it is often referred to in this thesis.

Theorem 1.3. The covariance of two linear estimators $\hat{Y}$ and $\hat{X}$ and its estimator under SI-design are:

$$Cov(\hat{Y}, \hat{X}) = N^2(1-f)S_{yx}/n, \tag{1.26}$$

$$\widehat{Cov}(\hat{Y}, \hat{X}) = N^2(1-f)s_{yx}/n, \tag{1.27}$$

where

$$S_{yx} = \frac{1}{N-1}\left[\sum_U y_i x_i - N\bar{Y}\bar{X}\right] \tag{1.28}$$

is the population variance of y and x and

$$s_{yx} = \frac{1}{n-1}\left[\sum_U I_i y_i x_i - n\bar{y}\bar{x}\right] \tag{1.29}$$

is sample variance of y and x. The quantities $\bar{Y} = Y/N$ and $\bar{X} = X/N$ are population means, their unbiased estimators are sample means $\bar{y} = \sum_U I_i y_i/n$ and $\bar{x} = \sum_U I_i x_i/n$.

Proof: Covariance (1.9) of two estimators $\hat{Y}$ and $\hat{X}$ takes under SI-design the form:

$$\begin{aligned}
Cov(\hat{Y},\hat{X}) &= \sum_U \Delta_{ii}\frac{y_i x_i}{E(I_i)^2} + \sum\sum_{U,i\neq j}\Delta_{ij}\frac{y_i}{E(I_i)}\frac{x_j}{E(I_j)} \\
&= \frac{1-f}{f}\sum_U y_i x_i - \frac{1-f}{f(N-1)}\sum\sum_{U,i\neq j} y_i x_j \\
&= \frac{1-f}{f(N-1)}\left[(N-1)\sum_U y_i x_i - (\sum_U y_i)(\sum_U x_i) + \sum_U y_i x_i\right] \\
&= N^2(1-f)\frac{1}{N-1}\left[\sum_U y_i x_i - N\bar{Y}\bar{X}\right]/n \\
&= N^2(1-f)S_{yx}/n.
\end{aligned}$$

Covariance estimator can be derived analogously from (1.10) by using design characteristics (1.23) – (1.25).

□

Remark 1.2. The variance formulae $V(\hat{Y})$ and $\hat{V}(\hat{Y})$ follow from (1.26) – (1.29) if y-variable equals to the x-variable, $y_i \equiv x_i$.

Obviously, the correlation between two linear estimators under SI-design is

$$Cor(\hat{Y},\hat{X}) = \frac{S_{yx}}{\sqrt{S_{yy}S_{xx}}},$$

thus, being equal to the correlation of the variables y and x in the population.

Let us now consider two ratio estimators $\hat{Y}^r = H\hat{Y}/\hat{H}$ and $\hat{X}^r = H\hat{X}/\hat{H}$ under SI-design. On the basis of Theorems 1.1–1.3 we can give covariance expressions of ratio estimators.

Corollary 1.1. The approximate covariance of $\hat{Y}^r$ and $\hat{X}^r$, and its unbiased estimator under SI-design are:

$$ACov(\hat{Y}^r, \hat{X}^r) = N^2(1-f)S_{uv}/n, \quad (1.30)$$

$$\widehat{Cov}(\hat{Y}^r, \hat{X}^r) = N^2(1-f)s_{uv}/n, \quad (1.31)$$

where

$$S_{uv} = \frac{1}{N-1}\sum\nolimits_U u_i v_i,$$

$$s_{uv} = \frac{1}{n-1}\sum\nolimits_U I_i \tilde{u}_i \tilde{v}_i,$$

with u_i, v_i, $\tilde{u}_i$ and $\tilde{v}_i$ defined in Theorem 1.2.

Proof: The formulae (1.30) – (1.31) follow from Theorem 1.2 by noticing that the variance and its SYG estimator in that theorem have similar expressions to the respective formulae in Theorem 1.1. Only the variables are denoted differently. Theorem 1.1 was used to derive SI-formulae in Theorem 1.3. Consequently, replacing variables y and x in Theorem 1.3 by u and v defined in (1.19) and (1.21), we get formulae of our corollary. Formulae for S_{uv} and s_{uv} follow from (1.28) – (1.29) by noting that $\bar{U} = \sum_U u_i/N = 0$, $\bar{V} = \sum_U v_i/N = 0$ and $\bar{\tilde{u}} = \sum_U I_i\tilde{u}_i/n = 0$, $\bar{\tilde{v}} = \sum_U I_i\tilde{v}_i/n = 0$.

□

1.2.2 Hypergeometric and multinomial sampling

The HG-design is an unequal probability sampling design. Selection mechanism under HG-design can be described for households/persons sampling situation as follows. SI-sampling of n persons is carried through in the list of M persons. Each selected person brings his/her household into sample. Sample of households is a HG-sample from the population of households. Let $\mathbf{I}$ be sampling vector in the population of households. Then its distribution is a multivariate hypergeometric distribution $\mathbf{I} \sim HG(M, n, m_1, m_2, ..., m_N)$, where m_i is the number of persons in the household i.

The characteristics of the HG-design are that of the HG-distribution of $\mathbf{I}$ (Johnson et al. 1997, Traat, Ilves, 2007):

$$E(I_i) = np_i, \quad (1.32)$$

$$V(I_i) = c\,np_i(1-p_i), \quad (1.33)$$

$$E(I_iI_j) = n(n-1)\frac{M}{(M-1)}p_ip_j,\ i \neq j, \tag{1.34}$$

$$E(I_i^2) = np_i(c\,(1-p_i) + np_i), \tag{1.35}$$

$$\Delta_{ij} = -c\,np_ip_j,\ i \neq j, \tag{1.36}$$

$$\breve{\Delta}_{ij} = -c\,\frac{M-1}{M(n-1)},\ i \neq j, \tag{1.37}$$

where in household/person terminology I_i is a selection variable of the household i, $i \in U$, $p_i = m_i/M$ is selection probability of household i, $M = \sum m_i$ is a number of persons in the frame (list of persons) and

$$c = \frac{M-n}{M-1}.$$

Instead of households one could think about sampling of other units through the list of smaller units comprising them.

The hypergeometric sampling design is usually not considered in sampling literature, rather its approximation, multinomial design, is considered. Therefore we derive here the estimation formulae under HG-design.

Theorem 1.4. The linear estimator of the population total $Y = \sum_U y_i$ under HG-design is

$$\hat{Y} = \sum\nolimits_U I_iy_i/(np_i). \tag{1.38}$$

Variance of $\hat{Y}$ is

$$V(\hat{Y}) = \frac{c}{n}\sum\nolimits_U \left(\frac{y_i}{p_i} - Y\right)^2 p_i \tag{1.39}$$

and its unbiased SYG variance estimator is

$$\hat{V}(\hat{Y}) = \frac{M-1}{M}\frac{c}{n(n-1)}\sum\nolimits_U I_i\left(\frac{y_i}{p_i} - \hat{Y}\right)^2. \tag{1.40}$$

Alternative forms of (1.39) and (1.40) are

$$V(\hat{Y}) = \frac{c}{n}\left(\sum\nolimits_U \frac{y_i^2}{p_i} - Y^2\right), \tag{1.41}$$

$$\hat{V}(\hat{Y}) = \frac{M-1}{M}\frac{c}{n(n-1)}\left(\sum\nolimits_U I_i\frac{y_i^2}{p_i^2} - n\hat{Y}^2\right). \tag{1.42}$$

Proof: The estimator (1.38) follows from the general form of linear estimator (1.1) by using (1.32):

$$\hat{Y} = \sum\nolimits_U I_iy_i/E(I_i) = \sum\nolimits_U I_iy_i/(np_i).$$

Variance of $\hat{Y}$ under HG-design follows from general form (1.2) by using (1.32) – (1.37):

$$\begin{aligned} V(\hat{Y}) &= \sum\nolimits_U \Delta_{ii} \left(\frac{y_i}{E(I_i)}\right)^2 + \sum \sum_{U,i\neq j} \Delta_{ij} \frac{y_i}{E(I_i)} \frac{y_j}{E(I_j)} \\ &= \sum\nolimits_U \frac{c\, np_i(1-p_i) y_i^2}{(np_i)^2} - \sum \sum_{U,i\neq j} c\, np_i p_j \frac{y_i}{np_i} \frac{y_j}{np_j} \\ &= \frac{c}{n} \left(\sum\nolimits_U \frac{y_i^2}{p_i} - \sum\nolimits_U y_i^2 - \sum \sum_{U,i\neq j} y_i y_j \right) \\ &= \frac{c}{n} \left(\sum\nolimits_U \frac{y_i^2}{p_i} - Y^2 \right). \end{aligned}$$

The received formula is just an alternative presentation of (1.39), which can be seen by opening brackets in (1.39) and using $\sum_U p_i = 1$:

$$V(\hat{Y}) = \frac{c}{n} \sum\nolimits_U \left(\frac{y_i^2}{p_i^2} - 2\frac{y_i}{p_i} Y + Y^2 \right) p_i = \frac{c}{n} \left(\sum\nolimits_U \frac{y_i^2}{p_i} - Y^2 \right).$$

The unbiased variance estimator (1.40) follows from the SYG formula (1.8). The use of SYG formula is justified since the design is a fixed size design. Noting that the terms having $i = j$ equal 0 in (1.8), we insert $\breve{\Delta}_{ij}$ for $i \neq j$ and get:

$$\begin{aligned} \hat{V}(\hat{Y}) &= \frac{M-1}{2M} \frac{c}{(n-1)} \sum \sum\nolimits_U I_i I_j \left(\frac{y_i}{np_i} - \frac{y_j}{np_j} \right)^2 \\ &= \frac{M-1}{2M} \frac{c}{(n-1)} \sum \sum\nolimits_U I_i I_j \left(\frac{y_i^2}{(np_i)^2} - \frac{2y_i y_j}{n^2 p_i p_j} + \frac{y_j^2}{(np_j)^2} \right) \\ &= \frac{M-1}{2M} \frac{c}{(n-1)} \left(n \sum\nolimits_U I_i \frac{y_i^2}{(np_i)^2} + n \sum\nolimits_U I_j \frac{y_j^2}{(np_j)^2} - 2\hat{Y}^2 \right) \\ &= \frac{M-1}{M} \frac{c}{n(n-1)} \left(\sum\nolimits_U I_i \frac{y_i^2}{p_i^2} - n\hat{Y}^2 \right). \end{aligned}$$

The formula (1.42) is an alternative presentation of (1.40), which can be seen by opening brackets in (1.40) and using $\sum_U I_i = n$:

$$\sum\nolimits_U \frac{y_i^2}{p_i} - Y^2 = \sum\nolimits_U I_i \left(\frac{y_i^2}{p_i^2} - 2\frac{y_i}{p_i} \hat{Y} + \hat{Y}^2 \right) = \sum\nolimits_U I_i \frac{y_i^2}{p_i^2} - n\hat{Y}^2.$$

□

Theorem 1.5. The covariance of estimators $\hat{Y}$ and $\hat{X}$ under HG-design is

$$Cov(\hat{Y},\hat{X}) \;=\; \frac{c}{n}\left(\sum\nolimits_U \frac{y_i x_i}{p_i} - YX\right). \tag{1.43}$$

Its unbiased SYG-type covariance estimator is

$$\widehat{Cov}(\hat{Y},\hat{X}) \;=\; \frac{M-1}{M}\frac{c}{n(n-1)}\left(\sum\nolimits_U I_i \frac{y_i x_i}{p_i^2} - n\hat{Y}\hat{X}\right). \tag{1.44}$$

Proof: Covariance of $\hat{Y}$ and $\hat{X}$ under HG-design follows from general form (1.9) by using (1.32) − (1.37):

$$\begin{aligned}
Cov(\hat{Y},\hat{X}) &= \sum\nolimits_U \Delta_{ii}\frac{y_i x_i}{E(I_i)^2} + \sum\sum_{U,i\neq j}\Delta_{ij}\frac{y_i}{E(I_i)}\frac{x_j}{E(I_j)} \\
&= \sum\nolimits_U \frac{c\,np_i(1-p_i)y_i x_i}{(np_i)^2} - \sum\sum_{U,i\neq j} c\,np_i p_j \frac{y_i}{np_i}\frac{x_j}{np_j} \\
&= \frac{c}{n}\left(\sum\nolimits_U \frac{y_i x_i}{p_i} - \sum\nolimits_U y_i x_i - \sum\sum_{U,i\neq j} y_i x_j\right) \\
&= \frac{c}{n}\left(\sum\nolimits_U \frac{y_i x_i}{p_i} - YX\right).
\end{aligned}$$

Estimator (1.44) is constructed using the analogy with SYG variance estimator (1.42). Note that (1.42) follows from (1.44) for $\hat{Y}=\hat{X}$. To show unbiasedness, we use $E(\hat{Y}\hat{X}) = Cov(\hat{Y},\hat{X}) + E(\hat{Y})E(\hat{X})$. Now, using unbiasedness of $\hat{Y}$ and $\hat{X}$ and the formula (1.43) for covariance, it follows

$$\begin{aligned}
E[\widehat{Cov}(\hat{Y},\hat{X})] &= \frac{M-1}{M}\frac{c}{n(n-1)}\left[\sum\nolimits_U \frac{np_i y_i x_i}{p_i^2} - nCov(\hat{Y},\hat{X}) - nYX\right] \\
&= \frac{M-1}{M}\frac{c}{(n-1)}\left[\sum\nolimits_U \frac{y_i x_i}{p_i} - Cov(\hat{Y},\hat{X}) - YX\right] \\
&= \frac{M-1}{M}\frac{c}{(n-1)}\left[\frac{n}{c}Cov(\hat{Y},\hat{X}) - Cov(\hat{Y},\hat{X})\right] \\
&= Cov(\hat{Y},\hat{X}).
\end{aligned}$$

□

Corollary 1.2. In case $m_i = 1, \forall i$, the HG-design is SI-design. The HG-formulae in Theorems 1.4 and 1.5 reduce to SI-design formulae in this case.

Proof: We show it for covariance estimator. By assumptions $M = \sum_U m_i = N$, $p_i = 1/N$, $c = \frac{N-n}{N-1}$. Now (1.44) takes the form:

$$\begin{aligned}\widehat{Cov}(\hat{Y}, \hat{X}) &= \frac{N-1}{N}\frac{N-n}{N-1}\frac{1}{n(n-1)}\left(\sum_U I_i \frac{y_i x_i}{1/N^2} - n\hat{Y}\hat{X}\right) \\ &= \frac{N^2(1-f)}{n(n-1)}\left(\sum_U I_i y_i x_i - n\bar{y}\bar{x}\right) \\ &= N^2(1-f)s_{yx}/n.\end{aligned}$$

□

Corollary 1.3. The approximate covariance of $\hat{Y}^r$ and $\hat{X}^r$, and its unbiased estimator under HG-design are:

$$ACov(\hat{Y}^r, \hat{X}^r) = \frac{c}{n}\sum_U \frac{u_i v_i}{p_i}, \tag{1.45}$$

$$\widehat{Cov}(\hat{Y}^r, \hat{X}^r) = \frac{M-1}{M}\frac{c}{n(n-1)}\sum_U I_i \frac{\tilde{u}_i \tilde{v}_i}{p_i^2}, \tag{1.46}$$

where u_i, v_i are given in (1.19) and $\tilde{u}_i$, $\tilde{v}_i$ in (1.21).

Proof: Analogously to Corollary 1.1, the formulae (1.45) − (1.46) follow from the Theorem 1.2. The formulae in Theorem 1.2 use variables u and v, otherwise these formulae are similar to the ones in Theorem 1.1. These latter formulae were elaborated for HG-design in Theorem 1.5. Consequently, replacing y and x variables in Theorem 1.5 by u and v variables defined in (1.19), and further on for covariance estimator, by $\tilde{u}$ and $\tilde{v}$ variables ((1.21) will

$$\begin{aligned}ACov(\hat{Y}^r, \hat{X}^r) &= \frac{c}{n}\left(\sum_U \frac{u_i v_i}{p_i} - UV\right), \\ \widehat{Cov}(\hat{Y}^r, \hat{X}^r) &= \frac{M-1}{M}\frac{c}{n(n-1)}\left(\sum_U I_i \frac{\tilde{u}_i \tilde{v}_i}{p_i^2} - n\hat{U}\hat{V}\right).\end{aligned}$$

Taking into account that the population totals of variables u and v and their estimators are equal to zero, for example,

$$\begin{aligned}\hat{U} &= \sum_U \frac{I_i \tilde{u}_i}{E(I_i)} \\ &= \sum_U \frac{I_i(y_i - \hat{R}_y h_i)}{np_i} \\ &= \sum_U \frac{I_i y_i}{np_i} - \frac{\hat{Y}}{\hat{H}}\sum_U \frac{I_i h_i}{np_i} = \hat{Y} - (\hat{Y}/\hat{H})\hat{H} = 0,\end{aligned}$$

we get the formulae of corollary.

□

Multinomial sampling design is a classical WR-design. Whenever the WR-design is assumed in sampling literature actually the multinomial design is meant, though the name multinomial is usually not used. However, there are infinitely many WR-designs, multinomial and HG-design are just two examples in this work. In our households/persons example, the selection mechanism for households is multinomial if n persons are selected by *with replacement* simple random sampling and each person brings his/her household into sample. The distribution of sampling vector for households is multinomial $I \sim M(n, p_1, p_2, ..., p_M)$.

Characteristics of multinomial design follow as special cases of the HG-design formulae in the limit $c \to 1, M \to \infty$:

$$
\begin{aligned}
E(I_i) &= np_i, && (1.47)\\
V(I_i) &= np_i(1-p_i), && (1.48)\\
E(I_iI_j) &= n(n-1)p_ip_j,\ i \neq j, && (1.49)\\
E(I_i^2) &= np_i(1-p_i+np_i), && (1.50)\\
\Delta_{ij} &= -np_ip_j,\ i \neq j, && (1.51)\\
\breve{\Delta}_{ij} &= -\frac{1}{(n-1)},\ i \neq j. && (1.52)
\end{aligned}
$$

These multinomial formulae can be used as approximations for HG-design when M is big compared to n, $M \gg n$.

The estimation formulae under multinomial design are special cases of the ones for HG-design. The formulae follow from Theorems 1.4 and 1.5 by taking $c = 1$ and $(M-1)/M = 1$. They can also be developed from general formulae (1.2) and (1.8) and Theorem 1.1 by using characteristics (1.47) − (1.52). The variance formula for multinomial design are given in Särndal et al. (1992, p. 51-52).

Chapter 2

Domain estimation

In this chapter we consider estimation of domain parameters. Domains could be geographical areas (county, municipality) or socio-economic groups (age-sex-education group) or other sub-populations (economic activity and size class of enterprises). Two direct estimators are considered for domain estimation: the linear estimator and the ratio estimator. The domain estimator is called direct if it uses the study variable values only from the observed domain. It can incorporate auxiliary information outside domain.

According to Rao (2003, p. 1) domain or area is counted as large if the sample size of domain is large enough for reliable direct estimates. A domain is regarded as small if sample size is too small for reliable direct estimates. In this thesis the domains are observed where the sample size is neither small nor empty. The estimators are considered in the design-based framework. The two estimators, linear and ratio, form a basis for building restriction estimators later. Attention is paid to covariances of domain estimators which are needed for restriction estimators and which are not much considered in sampling literature.

2.1 Definitions

Let us assume that population U consists of D domains (d=1,2,...,D). Let $U_d \subset U$ be a domain with size N_d, $\sum_{d=1}^{D} N_d = N$. Let the sampling design in U be given by $\mathbf{I} \sim p(k)$. Part of the design vector $\mathbf{I}$ where index $i \in U_d$ describes sampling in domain U_d. Sample size in U_d is $n_d = \sum_{U_d} I_i$, which is

usually random even if the overall sample size $n = \sum_{d=1}^{D} n_d$ is fixed. Expected sample size in U_d is

$$E(n_d) = \sum\nolimits_{U_d} E(I_i).$$

Let us define a domain indicator z_i^d:

$$z_i^d = \begin{cases} 1, \ i \in U_d, \\ 0, \ \text{otherwise}, \end{cases} \tag{2.1}$$

and create the new variable y_i^d:

$$y_i^d = z_i^d y_i = \begin{cases} y_i, \ i \in U_d, \\ 0, \ \text{otherwise}. \end{cases} \tag{2.2}$$

The defined variables (2.1) – (2.2) have a key role in domain estimation. Using them we can directly apply all earlier brought general formulae for estimation of population parameters. It is important to note that the domain total,

$$Y_d = \sum\nolimits_{U_d} y_i,$$

can be presented as a total of the new variable y_i^d over the entire population

$$Y_d = \sum\nolimits_{U} y_i^d = \sum\nolimits_{U} z_i^d y_i. \tag{2.3}$$

2.2 Estimators

Two domain estimators are introduced in this section: the linear estimator and the ratio estimator. They are both direct estimators which use only the study variable values collected or known for the units of the particular domain. The ratio estimator in addition uses auxiliary information which is the known domain total of an auxiliary variable for each domain.

The unbiased estimator of the domain total (2.3) follows directly from (1.1):

$$\hat{Y}_d \ = \ \sum\nolimits_{U} \omega_i y_i^d = \sum\nolimits_{U} \omega_i z_i^d y_i = \sum\nolimits_{U_d} \omega_i y_i. \tag{2.4}$$

As before, the design weight ω_i is determined by $\omega_i = I_i / E(I_i)$. In the case of WOR-designs the estimator (2.4) is the well known Horvitz-Thompson estimator for a domain.

Noting that domain size is the total of a special variable, $N_d = \sum_{U_d} y_i$, where $y_i \equiv 1$, we get its linear estimator as

$$\hat{N}_d \ = \ \sum\nolimits_{U} \omega_i z_i^d y_i = \sum\nolimits_{U_d} \omega_i. \tag{2.5}$$

The variability of the linear estimator (2.4) is relatively large, especially for small domains. A more effective estimator is the ratio estimator which uses an auxiliary variable h_i:

$$\hat{Y}_d^r = H_d \hat{R}_d, \tag{2.6}$$

where

$$\begin{aligned} H_d &= \sum\nolimits_{U_d} h_i, & (2.7)\\ \hat{R}_d &= \hat{Y}_d / \hat{H}_d, \\ R_d &= Y_d / H_d. & (2.8) \end{aligned}$$

Here $\hat{Y}_d$ is given in (2.4) and, similarly $\hat{H}_d = \sum_{U_d} \omega_i h_i$. In the special case $h_i \equiv 1$, we have $\hat{H}_d = \hat{N}_d$ and consequently, $\hat{R}_d$ estimates domain mean $R_d = Y_d/N_d$, in which case the ratio estimator $\hat{Y}_d^r$ is:

$$\hat{Y}_d^r = N_d \frac{\hat{Y}_d}{\hat{N}_d}. \tag{2.9}$$

2.3 Covariance of estimators

The dependence of domain estimators is a complex issue. The estimators are independent if the domains are strata and sampled independently. The estimators can be independent, or at least uncorrelated, also in other cases, as will be seen later. Generally, the correlation between domain estimators depends on different things, among them the sampling design, the analytic form of the estimator, the observed study variable and the domain size. It is known that the estimators of two domains are usually dependent, unless these domains are sampled independently (like strata). Here we derive general forms of covariances between domain estimators and between estimators of the domain and of the population total. The linear estimator and ratio estimator are considered. The design is assumed to be arbitrary with a condition $E(I_i I_j) > 0, \forall i, j$. The covariances for special sampling designs follow from the general results.

Let us consider linear estimators of two domain totals $\hat{Y}_d = \sum_{U_d} \omega_i y_i$ and $\hat{Y}_g = \sum_{U_g} \omega_i y_i$. They can be presented with new variables y_i^d and y_i^g as estimators of population totals $\hat{Y}_d = \sum_U \omega_i y_i^d$ and $\hat{Y}_g = \sum_U \omega_i y_i^g$, where variables y_i^d and y_i^g are defined by (2.2) for domains d and g, respectively. With this representation the covariance and its estimator follow from Theorem 1.1. They are given in the next corollary.

Corollary 2.1. The covariance between two linear domain estimators is

$$Cov(\hat{Y}_d, \hat{Y}_g) = \sum\sum\nolimits_U \Delta_{ij} \breve{y}_i^d \breve{y}_j^g = \sum_{i \in U_d} \sum_{j \in U_g} \Delta_{ij} \breve{y}_i \breve{y}_j \tag{2.10}$$

An unbiased estimator of covariance is

$$\widehat{Cov}(\hat{Y}_d, \hat{Y}_g) = \sum\sum\nolimits_U \breve{\Delta}_{ij} \omega_i y_i^d \omega_j y_j^g. \tag{2.11}$$

For fixed size designs the SYG-estimator is

$$\widehat{Cov}(\hat{Y}_d, \hat{Y}_g) = -\frac{1}{2} \sum\sum\nolimits_U I_i I_j \breve{\Delta}_{ij} (\breve{y}_i^d - \breve{y}_j^g)^2. \tag{2.12}$$

Remark 2.1. The variance $V(\hat{Y}_d)$ and its estimator follow from Corollary 2.1, if $d = g$.

Similarly we get the covariance between the estimators of population total, $\hat{Y}$, and domain total, $\hat{Y}_d$**.** We use variables y_i and y_i^d in Theorem 1.1.

Corollary 2.2. The covariance between estimators $\hat{Y} = \sum_U \omega_i y_i$ and $\hat{Y}_d = \sum_U \omega_i y_i^d$ is

$$Cov(\hat{Y}, \hat{Y}_d) = \sum\sum\nolimits_U \Delta_{ij} \breve{y}_i \breve{y}_j^d = \sum_{i \in U} \sum_{j \in U_d} \Delta_{ij} \breve{y}_i \breve{y}_j. \tag{2.13}$$

The unbiased covariance estimator is

$$\widehat{Cov}(\hat{Y}, \hat{Y}_d) = \sum\sum\nolimits_U \breve{\Delta}_{ij} \omega_i y_i \omega_j y_j^d. \tag{2.14}$$

The SYG estimator is

$$\widehat{Cov}(\hat{Y}, \hat{Y}_d) = -\frac{1}{2} \sum\sum\nolimits_U I_i I_j \breve{\Delta}_{ij} (\breve{y}_i - \breve{y}_j^d)^2. \tag{2.15}$$

Let us now consider ratio estimators of two domains $\hat{Y}_d^r = H_d \hat{R}_d$ and $\hat{Y}_g^r = H_g \hat{R}_g$, in obvious notation. Their covariance follows as a corollary from Theorem 1.2.

Corollary 2.3. The approximate Taylor expansion based covariance between estimators $\hat{Y}_d^r$ and $\hat{Y}_g^r$ is

$$ACov(\hat{Y}_d^r, \hat{Y}_g^r) = \sum\sum\nolimits_U \Delta_{ij} \breve{u}_i^d \breve{u}_j^g, \tag{2.16}$$

where

$$u_i^d = y_i^d - R_d h_i^d \text{ and } u_i^g = y_i^g - R_g h_i^g. \tag{2.17}$$

The covariance estimator is

$$\widehat{Cov}(\hat{Y}_d^r, \hat{Y}_g^r) = \sum\sum\nolimits_U \breve{\Delta}_{ij}\omega_i\tilde{u}_i^d\omega_j\tilde{u}_j^g, \tag{2.18}$$

where

$$\tilde{u}_i^d = y_i^d - \hat{R}_d h_i^d \text{ and } \tilde{u}_j^g = y_i^g - \hat{R}_g h_i^g, \tag{2.19}$$

and the SYG-type estimator is

$$\widehat{Cov}(\hat{Y}_d^r, \hat{Y}_g^r) = -\frac{1}{2}\sum\sum\nolimits_U I_i I_j \breve{\Delta}_{ij}(\breve{\tilde{u}}_i^d - \breve{\tilde{u}}_j^g)^2.$$

An alternative linearization based covariance estimator formula of domain ratio estimators is given in Lehtonen, Pahkinen (1995, p. 180). We preferred the form which uses new variables u_i^d and u_i^g. For WOR case the approximate covariance formula is given in Särndal et al. (1992, pp. 395 and 413).

Remark 2.2. The approximate variance and variance estimator of ratio estimator $\hat{Y}_d^r$ follows from Corollary 2.3 if $d = g$.

It is interesting to note that for some designs the ratio estimators are approximately uncorrelated.

Corollary 2.4. For the designs with $\frac{\Delta_{ij}}{E(I_i)E(I_j)} = C$ (constant) for $i \neq j$ the ratio estimators of two domains $\hat{Y}_d^r$ and $\hat{Y}_g^r$ are approximately uncorrelated:

$$ACov(\hat{Y}_d^r, \hat{Y}_g^r) = 0.$$

Proof: Since $\breve{u}_i^d = 0$ outside U_d and $\breve{u}_i^g = 0$ outside U_g, we get from (2.16)

$$ACov(\hat{Y}_d^r, \hat{Y}_g^r) = \sum_{i\in U_d}\sum_{j\in U_g}\Delta_{ij}\breve{u}_i^d\breve{u}_j^g = \sum_{i\in U_d}\sum_{j\in U_g}\frac{\Delta_{ij}}{E(I_i)E(I_j)}u_i^d u_j^g.$$

Now due to assumption,

$$\begin{aligned} ACov(\hat{Y}_d^r, \hat{Y}_g^r) &= C\sum\nolimits_{U_d}(y_i - R_d h_i)\sum\nolimits_{U_g}(y_j - R_g h_j) \\ &= C(Y_d - R_d H_d)(Y_g - R_g H_g) = 0. \end{aligned}$$

□

The assumption $\frac{\Delta_{ij}}{E(I_i)E(I_j)} = C$, $i \neq j$, holds for the designs considered in this thesis. For SI-design we get from (1.23) and (1.25):

$$C = -\frac{1-f}{f(N-1)} = -\frac{N-n}{n(N-1)}.$$

For HG-design we get from (1.32) and (1.36),

$$C = -\frac{M-n}{n(M-1)},$$

and for multinomial design

$$C = -\frac{1}{n}.$$

Analogically, one can get the covariance formulae for the ratio estimators of the population total $\hat{Y}^r = H\hat{R}$ and of a domain total $\hat{Y}_d^r = H_d\hat{R}_d$ from Theorem 1.2.

Corollary 2.5. The approximate covariance of estimators $\hat{Y}^r$ and $\hat{Y}_d^r$ is

$$ACov(\hat{Y}^r, \hat{Y}_d^r) = \sum\sum\nolimits_U \Delta_{ij}\breve{u}_i\breve{u}_j^d.$$

The covariance estimator is

$$\widehat{Cov}(\hat{Y}^r, \hat{Y}_d^r) = \sum\sum\nolimits_U \breve{\Delta}_{ij}\omega_i\tilde{u}_i\omega_j\tilde{u}_j^d,$$

and the SYG-type estimator is

$$\widehat{Cov}(\hat{Y}^r, \hat{Y}_d^r) = -\frac{1}{2}\sum\sum\nolimits_U I_iI_j\breve{\Delta}_{ij}(\breve{\breve{u}}_i - \breve{\breve{u}}_j^d)^2,$$

where $\tilde{u}_i^d$ is given in (2.19) and $\tilde{u}_i$ in (1.19).

It is worthwhile to note that the covariance formulae in the Corollary 2.5 reduce to the variance formulae of $\hat{Y}_d^r$ for some designs.

Corollary 2.6. For $R_d = R$ and for the designs with $\frac{\Delta_{ij}}{E(I_i)E(I_j)} = C$ for $i \neq j$, it holds:

$$ACov(\hat{Y}^r, \hat{Y}_d^r) = AV(\hat{Y}_d^r).$$

Proof: We can write

$$\begin{aligned} ACov(\hat{Y}^r, \hat{Y}_d^r) &= \sum\sum\nolimits_U \Delta_{ij}\breve{u}_i\breve{u}_j^d. \\ &= \sum\sum\nolimits_{U_d} \Delta_{ij}\breve{u}_i\breve{u}_j^d + \sum_{i\in U\setminus U_d}\sum_{j\in U_d} \Delta_{ij}\breve{u}_i\breve{u}_j^d, \end{aligned}$$

where $U \setminus U_d$ means complement of U_d. Due to the assumption the second term can be written in the form:

$$\sum_{i \in U \setminus U_d} \sum_{j \in U_d} \Delta_{ij} \breve{u}_i \breve{u}_j^d = C \sum_{i \in U \setminus U_d} u_i \sum_{j \in U_d} u_j^d.$$

This term is zero since $\sum_{U_d} u_j^d = \sum_{U_d} (y_j - R_d h_j) = 0$. The first term can be written as

$$\begin{aligned}
\sum\sum_{U_d} \Delta_{ij} \breve{u}_i \breve{u}_j^d &= \sum\sum_{U_d} \Delta_{ij} (\breve{y}_i - R\breve{h}_i)(\breve{y}_j - R_d \breve{h}_j) \\
&= \sum\sum_{U_d} \Delta_{ij} (\breve{y}_i - R_d \breve{h}_i + (R_d - R)\breve{h}_i)(\breve{y}_j - R_d \breve{h}_j) \\
&= \sum\sum_{U_d} \Delta_{ij} \breve{u}_i^d \breve{u}_j^d \\
&+ (R_d - R) \sum\sum_{U_d} \Delta_{ij} \breve{h}_i (\breve{y}_j - R_d \breve{h}_j).
\end{aligned}$$

Due to $R_d = R$, this is approximate variance of $\hat{Y}_d^r$ (see Remark 2.2).

□

Remark 2.3. Corollary 2.6 states that under given assumptions the approximate covariance between $\hat{Y}^r$ and $\hat{Y}_d^r$ is always nonnegative.

2.4 Covariance under SI-design

Corollary 2.7. Under SI-design the covariance of two linear domain estimators $\hat{Y}_d$ and $\hat{Y}_g$ is

$$Cov(\hat{Y}_d, \hat{Y}_g) = -N^2(1-f)\frac{1}{f(N-1)} \bar{Y}^d \bar{Y}^g, \tag{2.20}$$

where $\bar{Y}^d$ is the population mean of the variable y_i^d

$$\bar{Y}^d = \sum_U y_i^d / N,$$

and similarly $\bar{Y}^g$ is the population mean of variable y_i^g.

The covariance estimator is

$$\widehat{Cov}(\hat{Y}_d, \hat{Y}_g) = -N^2(1-f)\frac{1}{f(n-1)} \bar{y}^d \bar{y}^g, \tag{2.21}$$

where $\bar{y}^d = \sum_U I_i y_i^d / n$, and similarly $\bar{y}^g$.

Proof: Covariance of two linear domain estimators is given in (2.10). We know that under SI-design this formula simplifies to (1.26) with appropriate modification of notation. Consequently,

$$Cov(\hat{Y}_d, \hat{Y}_g) = N^2(1-f)S_{y^d y^g}/n, \tag{2.22}$$

where

$$S_{y^d y^g} = \frac{1}{N-1}\left[\sum\nolimits_U y_i^d y_i^g - N\bar{Y}^d\bar{Y}^g\right]. \tag{2.23}$$

Note that the first sum of (2.23) is equal to 0, because by definition (2.2) $y_i^g = 0$, if $i \in U_d$ and $y_i^d = 0$, if $i \in U_g$. Outside domains U_d and U_g both variables are equal to zero. Therefore we get from (2.22) the covariance formula (2.20) of the corollary. Covariance estimator follows analogically from (1.27).

□

Using Corollary 2.7 we get the correlation of two linear domain estimators under SI-design:

$$Cor(\hat{Y}_d, \hat{Y}_g) = -\frac{N}{N-1}\frac{\bar{Y}^d\bar{Y}^g}{\sqrt{S_{y^d y^d}S_{y^g y^g}}},$$

where $S_{y^d y^d}$ and $S_{y^g y^g}$ are population variances of the variables y^d and y^g, respectively (given in (2.23) with obvious modifications). For nonnegative y-variable, the estimators are negatively correlated.

Corollary 2.8. Under SI-design the covariance of linear estimators of the population total, $\hat{Y}$, and of the domain total, $\hat{Y}_d$, and the covariance estimator are:

$$\begin{aligned} Cov(\hat{Y}, \hat{Y}_d) &= N^2(1-f)S_{yy^d}/n, \\ \widehat{Cov}(\hat{Y}, \hat{Y}_d) &= N^2(1-f)s_{yy^d}/n, \end{aligned} \tag{2.24}$$

where

$$S_{yy^d} = \frac{1}{N-1}\left[\sum\nolimits_U (y_i^d)^2 - N\bar{Y}\bar{Y}^d\right] \tag{2.25}$$

is the covariance of y_i and y_i^d in the population, and

$$s_{yy^d} = \frac{1}{n-1}\left[\sum\nolimits_U I_i(y_i^d)^2 - n\bar{y}\bar{y}^d\right]$$

is a sample covariance of y_i and y_i^d.

Proof: Covariance formulae follow directly from (1.26) and (1.27). The sum of $(y_i^d)^2$ in the formulae above follows from the fact that variable $y_i^d = y_i$ inside the domain U_d and $y_i^d = 0$ outside the domain U_d, thus $\sum_U y_i y_i^d = \sum_U (y_i^d)^2$.

□

Correlation of estimators $\hat{Y}$ and $\hat{Y}_d$ takes the form:

$$Cor(\hat{Y}, \hat{Y}_d) = \frac{S_{yy^d}}{\sqrt{S_{yy} S_{y^d y^d}}},$$

where S_{yy} is population variance of the variable y_i.

Corollary 2.9. Under $R_d = R$ and SI-design the approximate covariance of ratio estimators of the population total, $\hat{Y}^r$, and domain total, $\hat{Y}_d^r$, and its estimator are:

$$\begin{aligned} ACov(\hat{Y}^r, \hat{Y}_d^r) &= N^2(1-f) S_{u^d u^d}/n, \\ \widehat{Cov}(\hat{Y}^r, \hat{Y}_d^r) &= N^2(1-f) s_{u^d u^d}/n. \end{aligned}$$

where

$$\begin{aligned} S_{u^d u^d} &= \frac{1}{N-1} \sum\nolimits_U (u_i^d)^2, \\ s_{u^d u^d} &= \frac{1}{n-1} \sum\nolimits_U I_i (\tilde{u}_i^d)^2, \end{aligned}$$

with u_i^d and $\tilde{u}_i^d$ given in (2.17) and (2.19).

Proof: Covariance formulae follow from Corollary 2.6,

$$ACov(\hat{Y}^r, \hat{Y}_d^r) = AV(\hat{Y}_d^r),$$

and Corollary 1.2.

□

Remark 2.4. In the special case $h_i \equiv 1$, the condition $R_d = R$ is not needed for the Corollary 2.9 to hold. Looking at the proof of Corollary 2.6, it can be easily verified that under SI-design,

$$\sum \sum\nolimits_{U_d} \Delta_{ij} \breve{h}_i (\breve{y}_j - R_d \breve{h}_j) = 0.$$

Now, irrespective of the value of $R_d - R$, it holds,

$$ACov(\hat{Y}^r, \hat{Y}_d^r) = AV(\hat{Y}_d^r).$$

The expressions for this special case are given in Corollary 2.9 with $u_i^d = y_i^d - (Y_d/N_d)z_i^d$ and $\tilde{u}_i^d = y_i^d - (\hat{Y}_d/\hat{N}_d)z_i^d$.

The following example illustrates dependence characteristics of linear domain estimators in a small population under SI-design. The asymptotic results of the ratio estimator can not be illustrated in such a small population. They will be considered in a simulation study of this thesis.

Example 2.1. Let us have a population with size $N = 8$ persons with two domains and one study variable y. The table 2.1 presents our population involving study variable y, the two domain variables y^1, y^2 and the totals and standard deviations (S) of the variables. Covariances of variables are $S_{yy^1} = 0.107$, $S_{yy^2} = 0.304$, $S_{y^1y^2} = -1.250$.

Table 2.1: Population and its parameters.

Domain	Individual ID	Household ID	y	y^1	y^2
1	1	1	2	2	0
1	2	1	1	1	0
1	3	1	2	2	0
1	4	2	2	2	0
1	5	2	3	3	0
2	6	3	2	0	2
2	7	4	3	0	3
2	8	4	2	0	2
Σ			17	10	7
S			1.357	1.554	0.411

All possible 28 samples with size $n = 6$ were formed from the population. The samples have equal probability 1/28 under SI-design. The linear estimators $\hat{Y}$, $\hat{Y}_1$ and $\hat{Y}_2$ were evaluated on each sample. Some parameters of the calculated estimates are presented in Table 2.2.

Table 2.2: Mean, minimum and maximum of estimates

Estimator	Mean	Minimum	Maximum
$\hat{Y}$	17.000	14.667	18.667
$\hat{Y}_1$	10.000	6.667	13.333
$\hat{Y}_2$	7.000	2.667	9.333

Denote $\mathbf{b}'(\mathbf{k}) = \left(\hat{Y}(\mathbf{k}), \hat{Y}_1(\mathbf{k}), \hat{Y}_2(\mathbf{k})\right)$ the vector of estimators evaluated on sample $\mathbf{k}$. Denote $\mathbf{b}' = (17, 10, 7)$ the vector of totals. Then the design-based

covariance matrix is

$$\mathbf{V} = \sum\nolimits_k [\mathbf{b}(\mathbf{k}) - \mathbf{b}][\mathbf{b}(\mathbf{k}) - \mathbf{b}]' p(\mathbf{k}),$$

where $p(\mathbf{k})$ is the probability of sample $\mathbf{k}$ and the sum is over all samples. As a result, the design-based covariance matrix of estimators is

$$\mathbf{V} = \begin{bmatrix} 1.095 & 0.286 & 0.810 \\ 0.286 & 3.619 & -3.333 \\ 0.810 & -3.333 & 4.142 \end{bmatrix}.$$

which is the same as theoretical covariance matrix based on formulae (2.20) and (2.24). The correlation matrix of vector $\mathbf{b}(\mathbf{k})$ is

$$Cor = \begin{bmatrix} 1.000 & 0.144 & 0.380 \\ 0.144 & 1.000 & -0.861 \\ 0.380 & -0.861 & 1.000 \end{bmatrix}.$$

Domain estimators are strongly negatively correlated. Correlation between the estimators of the domain and of the population totals is positive.

2.5 Covariance under HG-design

Corollary 2.10. Under HG-design the covariance of two linear domain estimators $\hat{Y}_d$ and $\hat{Y}_g$ and the covariance estimator are:

$$Cov(\hat{Y}_d, \hat{Y}_g) = -\frac{c}{n} Y_d Y_g, \tag{2.26}$$

$$\widehat{Cov}(\hat{Y}_d, \hat{Y}_g) = -\frac{M-n}{M(n-1)} \hat{Y}_d \hat{Y}_g, \tag{2.27}$$

where $c = \frac{M-n}{M-1}$.

Proof: Covariance of two linear domain estimators is given in (2.10). Under HG-design this general form simplifies to (1.43), and so, using appropriate notation we get:

$$Cov(\hat{Y}_d, \hat{Y}_g) = \frac{c}{n} \left(\sum\nolimits_U \frac{y_i^d y_i^g}{p_i} - Y_d Y_g \right). \tag{2.28}$$

Note that the sum in (2.28) is equal to zero by definition (2.2) for y_i^d and y_i^g. Therefore we get from (2.28) the first formula of corollary. Covariance estimator follows analogously from (1.44).

Corollary 2.11. Under HG-design the covariance formulae of linear estimators of the population total and of the domain total are:

$$\begin{aligned} Cov(\hat{Y}, \hat{Y}_d) &= \frac{c}{n}\left(\sum\nolimits_U \frac{(y_i^d)^2}{p_i} - YY_d\right). \qquad (2.29) \\ \widehat{Cov}(\hat{Y}, \hat{Y}_d) &= \frac{M-1}{M}\frac{c}{n(n-1)}\left(\sum\nolimits_U I_i \frac{(y_i^d)^2}{p_i^2} - n\hat{Y}\hat{Y}_d\right). \end{aligned}$$

Proof: Covariance formulae follow directly from (1.45) and (1.46). The sum of $(y_i^d)^2$ in the formulae above follows from the fact that variable $y_i^d = y_i$ inside the domain U_d and $y_i^d = 0$ outside the domain U_d, thus $\sum_U y_i y_i^d = \sum_U (y_i^d)^2$.

□

Corollary 2.12. Under $R_d = R$ and HG-design the approximate covariance of ratio estimators of the population total and of the domain total with covariance estimator are:

$$\begin{aligned} ACov(\hat{Y}^r, \hat{Y}_d^r) &= \frac{c}{n}\sum\nolimits_U \frac{(u_i^d)^2}{p_i}, \\ \widehat{Cov}(\hat{Y}^r, \hat{Y}_d^r) &= \frac{M-n}{Mn(n-1)}\sum\nolimits_U I_i \frac{(\tilde{u}_i^d)^2}{p_i^2}. \end{aligned}$$

with u_i^d and $\tilde{u}_i^d$ given in (2.17) and (2.19).

Proof: Covariance formulae follow from the result of Corollary 2.6:

$$ACov(\hat{Y}^r, \hat{Y}_d^r) = AV(\hat{Y}_d^r),$$

and Corollary 1.4.

□

Remark 2.5. The next corollary shows that for some auxiliary information the condition $R_d = R$ is not needed under HG-design.

Corollary 2.13. Under HG-design, $\mathbf{I} \sim HG(M, n, m_1, m_2, \ldots, m_N)$, it holds for the ratio estimators of a domain and population total which use the auxiliary variable m_i:

$$ACov(\hat{Y}^r, \hat{Y}_d^r) = AV(\hat{Y}_d^r).$$

Proof: Looking on the proof of Corollary 2.6, let us develop its last double sum for our case:

$$\begin{aligned} A &= \sum\sum\nolimits_{U_d} \Delta_{ij}\breve{h}_i(\breve{y}_j - R_d\breve{h}_j) = \sum\sum\nolimits_{U_d} \Delta_{ij}\frac{m_i}{np_i}\cdot\frac{y_j - R_d m_j}{np_j} \\ &= \sum_{j\in U_d}\left(\frac{y_j - R_d m_j}{np_j}\sum_{i\in U_d}\Delta_{ij}\frac{m_i}{np_i}\right). \end{aligned}$$

Writing the second sum separately for $i = j$ and $i \neq j$, then using the HG-formulae (1.32)−(1.37) and the relation $p_i = nm_i/M$, gives:

$$\begin{aligned} \sum_{i\in U_d}\Delta_{ij}\frac{m_i}{np_i} &= \frac{cnp_j(1-p_j)}{np_j}m_j + \sum_{i\in U_d, i\neq j}\frac{-cnp_ip_j}{np_i}m_i \\ &= cp_j(\frac{M}{n} - M_d), \end{aligned}$$

where $M_d = \sum_{U_d} m_i$. Finally, putting the expressons together and taking use of $R_d = Y_d/M_d$, we get

$$A = \sum_{j\in U_d}\frac{y_j - R_d m_j}{np_j}\cdot cp_j(\frac{M}{n} - M_d) = \frac{c}{n}(\frac{M}{n} - M_d)\sum_{j\in U_d}(y_j - R_d m_j) = 0$$

Consequently, with auxiliary variable m_i, the relation $ACov(\hat{Y}^r, \hat{Y}^r_d) = AV(\hat{Y}^r_d)$ holds under HG-design, irrespective of the values $R_d - R$, i.e. $Y_d/M_d - Y/M$ in our case.

□

The covariance formulae are given in Corollary 2.12, where $u^d_i = y^d_i - (Y_d/M_d)m^d_i$ and $\tilde{u}^d_i = y^d_i - (\hat{Y}_d/\hat{M}_d)m^d_i$.

Example 2.2. Consider the same data as in the Example 2.1. But let the population now consist of $N = 4$ households, involving $M = 8$ persons. The Table 2.3 presents our population of households and its parameters: m_i is the number of persons in the household i, $p_i = m_i/M$. The variables y, y^1 and y^2 are now for households, they are totals over household members.

Constant $c = (M - n)/(M - 1) = 0.286$, for HG-sampling.

For the HG-sampling from the population of households, the list of all $M = 8$ persons was observed and SI-sampling was performed in that list. All possible 28 samples with size $n = 6$ persons were formed from that list. Household of selected person was included into sample. The design for households is HG and

Table 2.3: Population and its parameters.

Domain	Household ID	m_i	p_i	y	y^1	y^2
1	1	3	0.375	5	5	0
1	2	2	0.250	5	5	0
2	3	1	0.125	2	0	2
2	4	2	0.250	5	0	5
Σy_i		8	1.000	17	10	7
$\Sigma(y_i)^2/p_i$				299	167	132

Table 2.4: Sample probability

HH ID	m_i	$\mathbf{k}$	samples (in columns)								
1	3	k_1	1	2	2	2	3	3	3	3	3
2	2	k_2	2	1	2	2	0	1	1	2	2
3	1	k_3	1	1	0	1	1	0	1	0	1
4	2	k_4	2	2	2	1	2	2	1	1	0
	$p(\mathbf{k})$		$\frac{3}{28}$	$\frac{6}{28}$	$\frac{3}{28}$	$\frac{6}{28}$	$\frac{1}{28}$	$\frac{2}{28}$	$\frac{4}{28}$	$\frac{2}{28}$	$\frac{1}{28}$

there are 9 possible samples $\mathbf{k}$ of households (see Table 2.4). The probability depends on the size m_i of household i and the number of selections k_i of the household i, and is calculated by the formula (Johnson et. al., 1997):

$$p(\mathbf{k}) = \frac{\binom{m_1}{k_1}\binom{m_2}{k_2}\binom{m_3}{k_3}\binom{m_4}{k_4}}{\binom{N}{n}}.$$

According to the linear estimator (1.38) the values of $\hat{Y}$, $\hat{Y}_1$ and $\hat{Y}_2$ were calculated on each sample. Some parameters of estimates are presented in Table 2.5.

Table 2.5: Mean, minimum and maximum of estimates

Estimator	Mean	Minimum	Maximum
$\hat{Y}$	17.000	16.000	18.222
$\hat{Y}_1$	10.000	6.667	13.333
$\hat{Y}_2$	7.000	2.667	9.333

Design-based covariance matrix of estimators was calculated as in example 2.1:

$$\mathbf{V} = \begin{bmatrix} 0.460 & -0.159 & 0.619 \\ -0.159 & 3.174 & -3.333 \\ 0.619 & -3.333 & 3.952 \end{bmatrix}.$$

It coincides with the theoretical covariance matrix based on formulae (2.28) and (2.29). Correlation matrix of the vector of estimators is:

$$Cor = \begin{bmatrix} 1.000 & -0.131 & 0.459 \\ -0.131 & 1.000 & -0.941 \\ 0.459 & -0.941 & 1.000 \end{bmatrix}.$$

The domain estimators are more strongly negatively correlated compared to the SI-design. We see that in case of linear estimators, the estimators of the domain and of the population total can be positively as well as negatively correlated.

Chapter 3

General restriction estimator for domains

In practical situations it may occur that the same population parameter is estimated in different surveys. Often the estimates from different surveys have to obey a set of restrictions. For example the total net income of households from wage labour estimated in the household budget survey has to be equal to the total net wages estimated in the labour force survey. Similarly different estimators from one survey have to satisfy some conditions (estimated totals of sub-populations have to sum up to the estimated population total). One solution of the described problem is to use the general restriction (GR) estimator proposed by Knottnerus (2003). The advantages of that estimator are the variance minimizing property and the explicit analytical form of that variance. In this chapter we first introduce the GR-estimator in general. We then proceed with working out the GR-estimator for domains.

3.1 General form of GR-estimator

Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_k)'$ be a parameter vector under study. Consider a k-dimensional vector of unbiased estimators from one or more samples denoted by $\hat{\boldsymbol{\theta}} = \left(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2, \ldots, \hat{\boldsymbol{\theta}}_k\right)'$. Denote the nonsingular covariance matrix of $\hat{\boldsymbol{\theta}}$ by $\mathbf{V}$. It is assumed that the parameters have to obey the following set of r linear restrictions:

$$\mathbf{R}\boldsymbol{\theta} = \mathbf{c}, \tag{3.1}$$

where $\mathbf{R}$ is an $r \times k$ matrix of rank r and $\mathbf{c}$ is the r-dimensional vector of constants.

Theorem 3.1. (Knottnerus, 2003, p. 328-329) The general restriction estimator $\hat{\boldsymbol{\theta}}^{GR}$ that satisfies restrictions (3.1), and the covariance matrix $\mathbf{V}^{GR}$ of the general restriction estimator are:

$$\hat{\boldsymbol{\theta}}^{GR} = \hat{\boldsymbol{\theta}} + \mathbf{K}(\mathbf{c} - \mathbf{R}\hat{\boldsymbol{\theta}}), \tag{3.2}$$

$$\mathbf{V}^{GR} \equiv \mathrm{Cov}(\hat{\boldsymbol{\theta}}^{GR}) = (\mathbf{I}_k - \mathbf{KR})\mathbf{V}, \tag{3.3}$$

where $\mathbf{I}_k$ is the identity matrix and

$$\mathbf{K} = \mathbf{VR}'(\mathbf{RVR}')^{-1}. \tag{3.4}$$

Knottnerus derives the estimator (3.2) by first assuming the normality of the vector $\hat{\boldsymbol{\theta}}$ and then minimizing corresponding log-likelihood function,

$$l(\boldsymbol{\theta}, \mathbf{V}) = -\frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \mathbf{V}^{-1}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}^{GR}) + const,$$

with respect to $\boldsymbol{\theta}$ under the restrictions $\mathbf{R}\boldsymbol{\theta} = \mathbf{c}$. The Lagrange's multipliers method is used. Later he shows that the normality assumption of $\hat{\boldsymbol{\theta}}$ is unnecessary.

□

Knottnerus shows that the GR-estimator is a linear minimum variance estimator of $\boldsymbol{\theta}$, given $\hat{\boldsymbol{\theta}}$, and given the information that $\mathbf{R}\boldsymbol{\theta} = \mathbf{c}$ (Knottnerus, 2003, p. 332). Based on results (3.2) − (3.4) we bring the following corollaries.

Corollary 3.1. The estimator $\hat{\boldsymbol{\theta}}^{GR}$ satisfies restrictions (3.1).

Proof: Multiplying $\hat{\boldsymbol{\theta}}^{GR}$ by $\mathbf{R}$ and using (3.4) we get:

$$\begin{aligned} \mathbf{R}\hat{\boldsymbol{\theta}}^{GR} &= \mathbf{R}\hat{\boldsymbol{\theta}} + \mathbf{RKc} - \mathbf{RKR}\hat{\boldsymbol{\theta}} \\ &= \mathbf{R}\hat{\boldsymbol{\theta}} + \mathbf{RVR}'(\mathbf{RVR}')^{-1}\mathbf{c} - \mathbf{RVR}'(\mathbf{RVR}')^{-1}\mathbf{R}\hat{\boldsymbol{\theta}} \\ &= \mathbf{R}\hat{\boldsymbol{\theta}} + \mathbf{c} - \mathbf{R}\hat{\boldsymbol{\theta}} = \mathbf{c}. \end{aligned}$$

□

Definition 3.1. Let two estimators, the vectors $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$ estimate unbiasedly the parameter vector $\boldsymbol{\theta}$. We say that $\hat{\boldsymbol{\theta}}_1$ is more effective than $\hat{\boldsymbol{\theta}}_2$ if $Cov(\hat{\boldsymbol{\theta}}_1) < Cov(\hat{\boldsymbol{\theta}}_2)$ in the sense of Löwner ordering.

Definition 3.2. The Löwner ordering $\mathbf{A} < \mathbf{B}$ of matrices $\mathbf{A}$ and $\mathbf{B}$ means that matrix $\mathbf{B} - \mathbf{A}$ is positive definite (see Lütkepohl, 1996).

Corollary 3.2. The GR-estimator $\hat{\boldsymbol{\theta}}^{GR}$ is more effective than the initial estimator $\hat{\boldsymbol{\theta}}$, $\mathbf{V}^{GR} < \mathbf{V}$.

Proof: Instead $\mathbf{V}^{GR} < \mathbf{V}$ we show that $\mathbf{V} - \mathbf{V}^{GR} > 0$. Inserting (3.4) into (3.3) we get

$$\begin{aligned} \mathbf{V} - \mathbf{V}^{GR} &= \mathbf{V}\mathbf{R}'(\mathbf{R}\mathbf{V}\mathbf{R}')^{-1}\mathbf{R}\mathbf{V} \\ &= (\mathbf{R}\mathbf{V})'(\mathbf{R}\mathbf{V}\mathbf{R}')^{-1}\mathbf{R}\mathbf{V}. \end{aligned}$$

This matrix has a form $\mathbf{A}\mathbf{A}'$. This type of matrix is positive definite if $\mathbf{A}$ is of full rank which is true by assumptions here: nonsingularity of $\mathbf{V}$ and full rank of $\mathbf{R}$ (Lütkepohl, 1996).

□

Corollary 3.3. If the restriction (3.1) is satisfied for the initial estimator $\hat{\boldsymbol{\theta}}$, $\mathbf{R}\hat{\boldsymbol{\theta}} = \mathbf{c}$, the GR-estimator is equal to the initial estimator, $\hat{\boldsymbol{\theta}}^{GR} = \hat{\boldsymbol{\theta}}$.

Proof: Since by assumption $\mathbf{c} - \mathbf{R}\hat{\boldsymbol{\theta}} = 0$, we get from (3.2):

$$\hat{\boldsymbol{\theta}}^{GR} = \hat{\boldsymbol{\theta}} + \mathbf{K}(\mathbf{c} - \mathbf{R}\hat{\boldsymbol{\theta}}) = \hat{\boldsymbol{\theta}}.$$

□

Corollary 3.4. The alternative representation of the general restriction estimator is through covariance matrices:

$$\begin{aligned} \mathbf{K} &= Cov(\hat{\boldsymbol{\theta}}, \mathbf{R}\hat{\boldsymbol{\theta}})Cov^{-1}(\mathbf{R}\hat{\boldsymbol{\theta}}), \\ Cov(\hat{\boldsymbol{\theta}}^{GR}) &= Cov(\hat{\boldsymbol{\theta}}, \mathbf{R}\hat{\boldsymbol{\theta}})Cov^{-1}(\mathbf{R}\hat{\boldsymbol{\theta}})Cov(\hat{\boldsymbol{\theta}}, \mathbf{R}\hat{\boldsymbol{\theta}}). \end{aligned}$$

Proof: The covariance matrix of two different random vectors is by definition:

$$\begin{aligned} Cov(\hat{\boldsymbol{\theta}}, \mathbf{R}\hat{\boldsymbol{\theta}}) &= E[\hat{\boldsymbol{\theta}} - E(\hat{\boldsymbol{\theta}})][\mathbf{R}\hat{\boldsymbol{\theta}} - \mathbf{R}E(\hat{\boldsymbol{\theta}})]' \\ &= \left\{E[\hat{\boldsymbol{\theta}} - E(\hat{\boldsymbol{\theta}})][\hat{\boldsymbol{\theta}} - E(\hat{\boldsymbol{\theta}})]'\right\}\mathbf{R}' \\ &= \mathbf{V}\mathbf{R}'. \end{aligned}$$

Using the property of covariance matrix we get:

$$\begin{aligned} Cov(\mathbf{R}\hat{\boldsymbol{\theta}}) &= \mathbf{R}Cov(\hat{\boldsymbol{\theta}})\mathbf{R}' \\ &= \mathbf{R}\mathbf{V}\mathbf{R}'. \end{aligned}$$

Consequently, the formulae of the corollary can be developed to the form of Theorem 3.1.

□

Replacing covariance matrix $\mathbf{V}$ in the formulae of GR-estimator with another matrix $\mathbf{B}$ we still get an estimator that satisfies restrictions.

Corollary 3.5: The estimator

$$\hat{\boldsymbol{\theta}}^{GRB} = \hat{\boldsymbol{\theta}} + \mathbf{K}(\mathbf{c} - \mathbf{R}\hat{\boldsymbol{\theta}}),$$

with $\mathbf{K} = \mathbf{B}\mathbf{R}'(\mathbf{R}\mathbf{B}\mathbf{R}')^{-1}$, where $\mathbf{B}$ is a matrix such that $\mathbf{R}\mathbf{B}\mathbf{R}'$ can be inverted, satisfies restrictions $\mathbf{R}\hat{\boldsymbol{\theta}}^{GRB} = \mathbf{c}$.

Proof: Analogically with Corollary 3.1 we multiply $\hat{\boldsymbol{\theta}}^{GR}$ by $\mathbf{R}$ and use (3.4). We get that the restrictions are satisfied:

$$\begin{aligned} \mathbf{R}\hat{\boldsymbol{\theta}}^{GRB} &= \mathbf{R}\hat{\boldsymbol{\theta}} + \mathbf{R}\mathbf{B}\mathbf{R}'(\mathbf{R}\mathbf{B}\mathbf{R}')^{-1}\mathbf{c} - \mathbf{R}\mathbf{B}\mathbf{R}'(\mathbf{R}\mathbf{B}\mathbf{R}')^{-\mathbf{1}}\mathbf{R}\hat{\boldsymbol{\theta}} \\ &= \mathbf{R}\hat{\boldsymbol{\theta}} + \mathbf{c} - \mathbf{R}\hat{\boldsymbol{\theta}} = \mathbf{c}. \end{aligned}$$

□

Variance of this estimator is higher than the variance of GR-estimator.

In practical situations the covariance matrix $\mathbf{V}$ is not known but can be estimated. The Corollary 3.5 showed that using $\hat{\mathbf{V}}$ instead of $\mathbf{V}$ gives still an estimator that satisfies restrictions.

Remark 3.1. Replacing covariance matrix $\mathbf{V}$ with its estimator $\hat{\mathbf{V}}$ we get the GR-estimator in the form:

$$\hat{\hat{\boldsymbol{\theta}}}^{GR} = \hat{\boldsymbol{\theta}} + \hat{\mathbf{V}}\mathbf{R}'(\mathbf{R}\hat{\mathbf{V}}\mathbf{R}')^{-1}(\mathbf{c} - \mathbf{R}\hat{\boldsymbol{\theta}}). \tag{3.5}$$

The estimator $\hat{\hat{\boldsymbol{\theta}}}^{GR}$ is not unbiased and is less effective than the estimator $\hat{\boldsymbol{\theta}}^{GR}$. However, for big samples where $\hat{\mathbf{V}}$ is close to $\mathbf{V}$, the estimator $\hat{\hat{\boldsymbol{\theta}}}^{GR}$ is close to $\hat{\boldsymbol{\theta}}^{GR}$. Also the linear term of the Taylor expansion of $\hat{\hat{\boldsymbol{\theta}}}^{GR}$ equals the $\hat{\boldsymbol{\theta}}^{GR}$ itself (Lepik, Sõstra, Traat, 2007). Thus, asymptotically the bias of $\hat{\hat{\boldsymbol{\theta}}}^{GR}$ vanishes and the covariance matrix of $\hat{\hat{\boldsymbol{\theta}}}^{GR}$ is equal to $Cov(\hat{\boldsymbol{\theta}}^{GR})$. Our simulation results show the slight increase of variance if covariance matrix $\mathbf{V}$ is replaced by its estimator $\hat{\mathbf{V}}$. No visible effect on bias was discovered.

In our derivations we continue to handle the restriction estimator that uses theoretical covariance matrix $\mathbf{V}$ of initial estimators. Our theoretical results of domains show in which way the restriction estimator depends on the elements of $\mathbf{V}$. In some cases this leads us to the construction of simpler estimators that do not use the matrix $\mathbf{V}$ at all and are still quite effective.

Example 3.1 As an example of GR-estimator let us consider two independent samples from the same population. Estimates for the population total Y are obtained from both samples. Denote the two parameters θ_1 and θ_2 which stand for population total Y from the first and second samples respectively. Restriction for parameters can be defined as $\theta_1 - \theta_2 = 0$ and (3.1) takes the form

$$\mathbf{R}\boldsymbol{\theta} \;=\; \begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} = 0.$$

Let the unbiased estimators of parameters θ_1 and θ_2 be $\hat{\theta}_1$ and $\hat{\theta}_2$. Assuming independent samples the covariance matrix of estimators is a diagonal matrix, $\mathbf{V} = diag(V_1, V_2)$, where $V_1 = V(\hat{\theta}_1)$ and $V_2 = V(\hat{\theta}_2)$ are variances of the estimators. In this case (3.4) can be developed to the form

$$\mathbf{K} \;=\; \frac{1}{V_1 + V_2} \begin{pmatrix} V_1 \\ -V_2 \end{pmatrix},$$

and the general restriction estimator of $\boldsymbol{\theta}$ is

$$\hat{\boldsymbol{\theta}}^{GR} = \begin{pmatrix} \hat{\theta}_1^{GR} \\ \hat{\theta}_2^{GR} \end{pmatrix} = \begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{pmatrix} - \frac{\hat{\theta}_1 - \hat{\theta}_2}{V_1 + V_2} \begin{pmatrix} V_1 \\ -V_2 \end{pmatrix}. \tag{3.6}$$

The restriction $\theta_1 - \theta_2 = 0$ is satisfied for this estimator:

$$\hat{\theta}_1^{GR} - \hat{\theta}_2^{GR} \;=\; \hat{\theta}_1 - \frac{V_1(\hat{\theta}_1 - \hat{\theta}_2)}{V_1 + V_2} - \hat{\theta}_2 - \frac{V_2(\hat{\theta}_1 - \hat{\theta}_2)}{V_1 + V_2} = 0.$$

The covariance matrix of $\hat{\boldsymbol{\theta}}^{GR}$ follows from (3.3):

$$\mathbf{V}^{GR} = \frac{V_1 V_2}{V_1 + V_2} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}. \tag{3.7}$$

It can be seen from (3.7) that $Cor(\hat{\theta}_1^{GR}, \hat{\theta}_2^{GR}) = 1$. It is a consequence of linear relationship, $\hat{\theta}_1^{GR} = \hat{\theta}_2^{GR}$. The relative change of variance,

$$\frac{V_1^{GR} - V_1}{V_1} = \left(\frac{V_1 V_2}{V_1 + V_2} - V_1 \right) / V_1 = -\frac{V_1}{V_1 + V_2},$$

depends on the proportion of the variance of initial estimator in total variance. The variance of general restriction estimator is smaller than the variance of the corresponding initial estimator.

□

Let $\hat{\theta}_1 = \hat{Y}_1$ and $\hat{\theta}_2 = \hat{Y}_2$ be linear estimators of Y. In case of SI-design with sample sizes n_1 and n_2 the variances V_1 and V_2 are given by (1.30). Inserting them into (3.6) and (3.7) the general restriction estimator of population total and its variance take the forms:

$$\begin{aligned} \hat{Y}_1^{GR} &= \hat{Y}_1 - \frac{(\hat{Y}_1 - \hat{Y}_2)(1-f_1)/n_1}{(1-f_1)/n_1 + (1-f_2)/n_2}, \\ V_1^{GR} = V_2^{GR} &= \frac{N^2(1-f_1)(1-f_2)S_{yy}/n_1 n_2}{(1-f_1)/n_1 + (1-f_2)/n_2}, \end{aligned}$$

where $f_i = n_i/N$, $i = 1, 2$ and S_{yy} is the population variance of study variable y, (1.28). The relative change of variance in the case of SI-design depends on the sample sizes:

$$\frac{V_1^{GR} - V_1}{V_1} = -\frac{(1-f_1)/n_1}{(1-f_1)/n_1 + (1-f_2)/n_2}.$$

In case of equal sample sizes, $n_1 = n_2$, the decrease of the variance is $1/2$ of the initial variance V_1.

3.2 General form of conditional GR-estimator

Above it was assumed that vector $\mathbf{c}$ in GR-estimator is constant, known from external sources. In practical situations $\mathbf{c}$ in the restrictions may be the vector of estimates from earlier surveys or of estimates from the same survey using different estimators. Here the aim is to find the restriction estimator so that the estimates in $\mathbf{c}$ do not change. For example the need for this situation occurs, when some estimates are already published and cannot be changed, but additional domain estimates are needed which should be consistent with published ones.

We assume that some population parameters are previously unbiasedly estimated by $\hat{\boldsymbol{\theta}}_1$. We want to consider $\hat{\boldsymbol{\theta}}_1$ fixed. We have initial unbiased estimators $\hat{\boldsymbol{\theta}}_2$ for population parameters $\boldsymbol{\theta}_2$. The task is to solve consistency problem for $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$ by finding conditional GR-estimator so that

$$\mathbf{R}\hat{\boldsymbol{\theta}}_2^{GR} = \hat{\boldsymbol{\theta}}_1. \tag{3.8}$$

However, when finding covariance matrix of $\hat{\boldsymbol{\theta}}_2^{GR}$, the estimator $\hat{\boldsymbol{\theta}}_1$ should be considered as random. Let $\mathbf{V_1} = \mathrm{Cov}(\hat{\boldsymbol{\theta}}_1)$, $\mathbf{V_2} = \mathrm{Cov}(\hat{\boldsymbol{\theta}}_2)$ and $\mathbf{V_{21}} = \mathrm{Cov}(\hat{\boldsymbol{\theta}}_2, \hat{\boldsymbol{\theta}}_1)$ be covariance matrices of initial estimators.

Theorem 3.2. The conditional GR-estimator of parameters in $\boldsymbol{\theta}_2$, that satisfies the restrictions (3.8) is

$$\hat{\boldsymbol{\theta}}_2^{GR} = \hat{\boldsymbol{\theta}}_2 + \mathbf{K}(\hat{\boldsymbol{\theta}}_1 - \mathbf{R}\hat{\boldsymbol{\theta}}_2), \tag{3.9}$$

where $\mathbf{K} = \mathbf{V_2 R'(RV_2 R')^{-1}}$. Its unconditional covariance matrix is

$$\mathrm{Cov}(\hat{\boldsymbol{\theta}}_2^{GR}) \;=\; \mathbf{PV_2} + \mathbf{KV_1K'} + \mathbf{PV_{21}K'} + (\mathbf{PV_{21}K'})', \tag{3.10}$$

where

$$\mathbf{P} = \mathbf{I} - \mathbf{KR}. \tag{3.11}$$

Proof: It is straightforward to check that (3.8) is satisfied. Let us look the estimator (3.9) in the form

$$\hat{\boldsymbol{\theta}}_2^{GR} = \mathbf{K}\hat{\boldsymbol{\theta}}_1 + \mathbf{P}\hat{\boldsymbol{\theta}}_2.$$

The result (3.10) follows by using the properties of covariance matrix of a random vector (Srivastava, 2002, pp. 21-24):

$$\begin{aligned} Cov(\hat{\boldsymbol{\theta}}_2^{GR}) &= Cov(\mathbf{K}\hat{\boldsymbol{\theta}}_1) + Cov(\mathbf{P}\hat{\boldsymbol{\theta}}_2) \\ &+ Cov(\mathbf{K}\hat{\boldsymbol{\theta}}_1, \mathbf{P}\hat{\boldsymbol{\theta}}_2) + Cov(\mathbf{P}\hat{\boldsymbol{\theta}}_2, \mathbf{K}\hat{\boldsymbol{\theta}}_1). \end{aligned}$$

The covariance matrix of two different random vectors, being certain linear transformations, is

$$\begin{aligned} Cov(\mathbf{K}\hat{\boldsymbol{\theta}}_1, \mathbf{P}\hat{\boldsymbol{\theta}}_2) &= E[\mathbf{K}\hat{\boldsymbol{\theta}}_1 - \mathbf{K}E(\hat{\boldsymbol{\theta}}_1)][\mathbf{P}\hat{\boldsymbol{\theta}}_2 - \mathbf{P}E(\hat{\boldsymbol{\theta}}_2)]' \\ &= \mathbf{K}\left\{E[\hat{\boldsymbol{\theta}}_1 - E(\hat{\boldsymbol{\theta}}_1)][\hat{\boldsymbol{\theta}}_2 - E(\hat{\boldsymbol{\theta}}_2)]'\right\}\mathbf{P}' \\ &= \mathbf{K}Cov(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)\mathbf{P}' \\ &= \mathbf{KV_{12}P'} = \mathbf{PV_{21}K'}. \end{aligned}$$

Analogically, covariance matrix of one linearly transformed random vector is:

$$Cov(\mathbf{K}\hat{\boldsymbol{\theta}}_1) = \mathbf{K}Cov(\hat{\boldsymbol{\theta}}_1)\mathbf{K}' = \mathbf{KV_1K'}.$$

□

3.3 GR-estimator for domains when population total is known

Let $\boldsymbol{\theta} = (Y_1, Y_2, ..., Y_D)'$ be the vector of unknown domain totals. Let us assume that design-based estimation method is used for Y_d, $d \in \mathfrak{D} = \{1, 2, \ldots, D\}$. Denote the vector of estimators by $\hat{\boldsymbol{\theta}} = \left(\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_D\right)'$.

If nonoverlapping domains, $d \in \mathfrak{D}$, cover all the population then the domain totals Y_d sum up to the population total Y,

$$\sum\nolimits_{\mathfrak{D}} Y_d = Y.$$

Restriction (3.1) takes the form:

$$\mathbf{R}\boldsymbol{\theta} = \begin{pmatrix} 1 & \cdots & 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} Y_1 \\ \cdots \\ Y_d \\ \cdots \\ Y_D \end{pmatrix} = Y, \tag{3.12}$$

with

$$\mathbf{R} = \begin{pmatrix} 1 & \cdots & 1 & \cdots & 1 \end{pmatrix} = \mathbf{1}'_D \tag{3.13}$$

being a D-vector of ones. Population total can be obtained from external sources with certainty or from other survey with high precision or estimated from the same survey using different estimation method.

Here we assume that the population total is known, e.g. from external sources. For calculating restriction estimators for domains the covariance matrix of initial estimators $\hat{Y}_i$, $i \in \mathfrak{D}$, is needed. Let $\mathbf{V} = (V_{ij})$ be a $D \times D$ matrix with $V_{ij} = Cov(\hat{Y}_i, \hat{Y}_j)$, $i, j \in \mathfrak{D}$. Naturally, $V_{ii} = V(\hat{Y}_i)$. Denote the sum of estimated domain totals by $\hat{Y}_*$,

$$\hat{Y}_* = \sum\nolimits_{\mathfrak{D}} \hat{Y}_i.$$

We assume that $\hat{Y}_*$ is a random variable, in which case its variance is strictly positive and consequently:

$$\sum\sum\nolimits_{\mathfrak{D}} V_{ij} = V(\hat{Y}_*) > 0. \tag{3.14}$$

Theorem 3.3. Let the initial domain estimators be $\hat{Y}_i$, $i \in \mathfrak{D}$ with nonsingular covariance matrix $\mathbf{V} = (V_{ij})$. Let the restrictions be given by (3.12). Then the general restriction estimator of domain U_d, $\hat{Y}_d^{GR}$, and its variance, V_{dd}^{GR}, are:

$$\hat{Y}_d^{GR} = \hat{Y}_d + \frac{Y - \hat{Y}_*}{\sum\sum_{\mathfrak{D}} V_{ij}} \cdot \sum_{i \in \mathfrak{D}} V_{di}, \tag{3.15}$$

$$V_{dd}^{GR} = V_{dd} - \frac{\left(\sum\limits_{i \in \mathfrak{D}} V_{di}\right)^2}{\sum\sum_{\mathfrak{D}} V_{ij}}. \tag{3.16}$$

The covariance of domain estimators $\hat{Y}_d^{GR}$ and $\hat{Y}_g^{GR}$ is

$$V_{dg}^{GR} = V_{dg} - \frac{\left(\sum\limits_{i \in \mathfrak{D}} V_{di}\right)\left(\sum\limits_{i \in \mathfrak{D}} V_{gi}\right)}{\sum\sum_{\mathfrak{D}} V_{ij}}. \tag{3.17}$$

Assuming unbiased initial estimators for domain totals, $E(\hat{Y}_i) = Y_i$, $i \in \mathfrak{D}$, the GR-estimator $\hat{Y}_d^{GR}$ is unbiased.

Proof: We use formulae of general restriction estimator in Theorem 3.1. With $\mathbf{R}$ in (3.13) we get from (3.4):

$$\mathbf{K} = \mathbf{V} \cdot \mathbf{1}_D (\mathbf{1}_D' \mathbf{V} \mathbf{1}_D)^{-1} = \frac{1}{\sum\sum_{\mathfrak{D}} V_{ij}} \begin{pmatrix} \sum\limits_{i \in \mathfrak{D}} V_{1i} \\ \cdots \\ \sum\limits_{i \in \mathfrak{D}} V_{di} \\ \cdots \\ \sum\limits_{i \in \mathfrak{D}} V_{Di} \end{pmatrix}. \tag{3.18}$$

Noting that constant $\mathbf{c}$ in (3.2) is equal to known population total, $\mathbf{c} = Y$, we can develop restriction into form:

$$\mathbf{c} - \mathbf{R}\hat{\boldsymbol{\theta}} = Y - \sum\nolimits_{\mathfrak{D}} \hat{Y}_i = Y - \hat{Y}_*. \tag{3.19}$$

Inserting (3.18) and (3.19) into (3.2) we get restriction estimator (3.15) for domain U_d.

Inserting (3.18) into (3.3) we get

$$\mathbf{V}^{GR} = [\mathbf{I}_D - \mathbf{V}\mathbf{1}_D(\mathbf{1}_D'\mathbf{V}\mathbf{1}_D)^{-1}\mathbf{1}_D']\mathbf{V}.$$

Taking the scalar $(\mathbf{1}_D'\mathbf{V}\mathbf{1}_D)^{-1} = (\sum\sum_{\mathfrak{D}} V_{ij})^{-1}$ out, and using $\mathbf{1}_D'\mathbf{V} = (\mathbf{V}\mathbf{1}_D)'$, we have

$$\mathbf{V}^{GR} = \mathbf{V} - (\mathbf{V}\mathbf{1}_D)(\mathbf{V}\mathbf{1}_D)'(\sum\sum\nolimits_{\mathfrak{D}} V_{ij})^{-1}. \tag{3.20}$$

The covariance V_{dg}^{GR} is the (d, g) element of the matrix $\mathbf{V}^{GR}$. Noting that element d of the vector $\mathbf{V1}_D$ is $\sum_{i\in\mathfrak{D}} V_{di}$, the covariance of two GR-estimators (3.17) follows. Variance (3.16) of GR-estimator follows from (3.20) as its diagonal element for $d = g$.

To show unbiasedness of $\hat{Y}_d^{GR}$ note that $E(\hat{Y}_d) = Y_d$ by assumption. Since this holds for all $d \in \mathfrak{D}$, we have

$$E(Y - \hat{Y}_*) = E(Y - \sum\nolimits_{\mathfrak{D}} \hat{Y}_i) = Y - \sum\nolimits_{\mathfrak{D}} Y_i = 0.$$

The last equality comes from the condition for totals $\sum_{\mathfrak{D}} Y_i = Y$. Therefore the expectation of GR-estimator equals the true total, $E(\hat{Y}_d^{GR}) = Y_d$.

□

Example 3.2. Let us check that the variance of the GR-estimator for domain U_d is not greater than the variance of corresponding initial estimator:

$$0 \le V_{dd}^{GR} \le V_{dd}. \tag{3.21}$$

Instead of the conditions (3.21) we check the alternative conditions received by inserting (3.16) into (3.21):

$$0 \le \frac{\left(\sum\limits_{i\in\mathfrak{D}} V_{di}\right)^2}{\sum\sum_{\mathfrak{D}} V_{ij}} \le V_{dd} \tag{3.22}$$

Obviously $\left(\sum\limits_{i\in\mathfrak{D}} V_{di}\right)^2 \ge 0$ because of the square. The denominator is strictly positive due to (3.14). Therefore the left inequality of (3.22) is satisfied.

The right inequality of (3.22) follows from Cauchy-Schwartz inequality (Casella, Berger, 1990, p. 180):

$$[Cov(\hat{Y}_d, \hat{Y}_*)]^2 \le V(\hat{Y}_d) \cdot V(\hat{Y}_*). \tag{3.23}$$

Noting that $Cov(\hat{Y}_d, \hat{Y}_*) = Cov(\hat{Y}_d, \sum_{\mathfrak{D}} \hat{Y}_i) = \sum_{i\in\mathfrak{D}} V_{di}$, and using (3.14) we get alternatively

$$\left(\sum_{i\in\mathfrak{D}} V_{di}\right)^2 \le (\sum\sum\nolimits_{\mathfrak{D}} V_{ij}) V_{dd}.$$

Dividing both sides of inequality by positive quantity $\sum \sum_{\mathfrak{D}} V_{ij}$ we get the right inequality of (3.22).

□

If $\hat{Y}_d$ and $\hat{Y}_*$ are not linearly dependent then strict inequality holds in (3.23) as well in $V_{dd}^{GR} < V_{dd}$.

Remark 3.2. We see from (3.15) that the initial domain estimator $\hat{Y}_d$ is adjusted by the proportion of the difference $Y - \hat{Y}_*$. The proportion is defined by $Cov(\hat{Y}_d, \hat{Y}_*)/V(\hat{Y}_*)$, i.e. depends on domain U_d.

Corollary 3.6. In case of uncorrelated unbiased domain estimators $\hat{Y}_i$, $i \in \mathfrak{D}$, the GR-estimator is

$$\hat{Y}_d^{GR} = \hat{Y}_d + \frac{Y - \hat{Y}_*}{\sum_{\mathfrak{D}} V_{ii}} \cdot V_{dd} \tag{3.24}$$

and its variance is

$$V_{dd}^{GR} = V_{dd} - \frac{V_{dd}^2}{\sum_{\mathfrak{D}} V_{ii}}. \tag{3.25}$$

The covariance of two GR-estimators is

$$Cov(\hat{Y}_d^{GR}, \hat{Y}_g^{GR}) = -\frac{V_{dd} V_{gg}}{\sum_{\mathfrak{D}} V_{ii}}. \tag{3.26}$$

Proof: Since for uncorrelated domain estimators the covariance $V_{ij} = 0$, $i \neq j$, $i, j \in \mathfrak{D}$, the GR-estimator (3.24) and its variance (3.25) follow from (3.15) and (3.16) respectively. The covariance (3.26) follows from (3.17).

□

Remark 3.3. In spite of the uncorrelated initial estimators the restriction estimators of two domains U_d and U_g are correlated. The correlation is always negative.

Remark 3.4. In Corollary 3.6, the adjustment constant $a_d = V_{dd} / \sum_{\mathfrak{D}} V_{ii}$ is used in restriction estimator. It is obvious that

$$0 \leq a_d \leq 1 \text{ and } \sum\nolimits_{\mathfrak{D}} a_d = 1. \tag{3.27}$$

This leads to a practical simplification of the restriction estimator in case V_{ii}, $i \in \mathfrak{D}$ is unknown and difficult to estimate. One should choose some constants a'_d, satisfying (3.27). The restrictions are satisfied for

$$\hat{Y}_d^{GR} = \hat{Y}_d + a'_d(Y - \hat{Y}_*).$$

The simplest choice is $a'_d = 1/D$, where D is the number of domains. Since the variance of $\hat{Y}_d^{GR}$ is minimized for $a'_d = a_d$, one should try to choose a'_d close to a_d. Here certain knowledge of the variability, V_{dd}, of domain estimators is helpful. The discussion here concerned the uncorrelated domain estimators, like e.g. ratio estimator. Similar discussion can be developed for more general case based on Remark 3.2.

3.3.1 GR-estimator under SI-design

Let the sampling design in the population U be SI with sample size n. Let the domain totals Y_d, $d \in \mathfrak{D}$, be estimated by linear estimator $\hat{Y}_d = \sum_{U_d} \omega_i y_i$. Linear estimator is additive, i.e. the domain estimators sum up to the linear estimator of the population total:

$$\sum\nolimits_{\mathfrak{D}} \hat{Y}_d = \sum\nolimits_{\mathfrak{D}} \sum\nolimits_{U_d} \omega_i y_i = \sum\nolimits_{U} \omega_i y_i = \hat{Y}. \tag{3.28}$$

We assume that the true population total Y is known. Next we formulate the GR-estimator under this special case.

Corollary 3.7. Assuming SI-design and the linear domain estimators, the GR-estimator and its variance are:

$$\hat{Y}_d^{GR} = \hat{Y}_d + \frac{Y - \hat{Y}}{S_{yy}} \cdot S_{yy^d}, \tag{3.29}$$

$$V_{dd}^{GR} = \frac{N^2(1-f)}{n} \left(S_{y^d y^d} - \frac{S_{yy^d}^2}{S_{yy}} \right), \tag{3.30}$$

where S_{yy} is the variance of the variable y_i, $S_{y^d y^d}$ is the variance of the variable y_i^d, and S_{yy^d} is a covariance of variables y_i and y_i^d, given in (1.28). The covariance of two GR-estimators under SI-design is

$$V_{dg}^{GR} = \frac{N^2(1-f)}{n} \left(S_{y^d y^g} - \frac{S_{yy^d} S_{yy^g}}{S_{yy}} \right), \tag{3.31}$$

where $S_{y^d y^g}$ is the covariance of variables y_i^d and y_i^g, given in (2.23).

Proof: We use general results in Theorem 3.3. Due to (3.28) $\hat{Y}_* = \hat{Y} = \sum_U \omega_i y_i$, which is the linear estimator of the population total. Consequently $V(\hat{Y}_*) = V(\hat{Y}) = \sum \sum_{\mathfrak{D}} V_{ij}$, and thus the denominator in the formulae (3.15) – (3.17) is

$$\sum \sum\nolimits_{\mathfrak{D}} V_{ij} = N^2(1-f) S_{yy}/n.$$

Noting that $\sum_{i\in\mathfrak{D}} V_{di} = Cov(\hat{Y}_d, \sum_{\mathfrak{D}} \hat{Y}_i) = Cov(\hat{Y}_d, \hat{Y})$, we have

$$\sum_{i\in\mathfrak{D}} V_{di} = N^2(1-f)S_{yy^d}/n.$$

Inserting these results into (3.15) − (3.17) we get (3.29) − (3.31).

□

From (3.29) and (3.31) we can see that the GR-estimator and its variance depend on the population variance and covariance of variables y_i and y_i^d. The covariance of variables y_i and y_i^d is relatively larger for large domains. Thus, the initial estimator of larger domains is changed relatively more compared to smaller ones when calculating GR-estimators. Also, the decrease of variance of GR-estimator compared to initial estimator is larger for large domains.

3.3.2 GR-estimator under HG-design

Let the sampling design in the population U of size N be HG with sample size n, i.e. HG$(M, n, m_1, m_2, ..., m_N)$. Let the domain totals Y_d, $d \in \mathfrak{D}$, be estimated by linear estimator $\hat{Y}_d = \sum_{U_d} \omega_i y_i$. Linear estimator is additive, which means (3.28) holds. We assume that the true population total Y is known. Next we formulate the GR-estimator under this special case.

Corollary 3.8. Assuming HG-design in the population and the linear domain estimators, the GR-estimator and its variance are:

$$\hat{Y}_d^{GR} = \hat{Y}_d + (Y - \hat{Y})\frac{v_d}{v}, \tag{3.32}$$

$$V_{dd}^{GR} = \sum_U \frac{(y_i^d)^2}{p_i} - Y_d^2 - \frac{c}{n}\frac{v_d^2}{v}, \tag{3.33}$$

where

$$v_d = \sum_U \frac{(y_i^d)^2}{p_i} - YY_d$$

$$v = \sum_U \frac{y_i^2}{p_i} - Y^2$$

$$p_i = m_i/M,\ c = \frac{M-n}{M-1}$$

The covariance of two GR-estimators under HG-design is

$$V_{dg}^{GR} = -\frac{c}{n}\left(Y_d Y_g + \frac{v_d v_g}{v}\right). \tag{3.34}$$

Proof: We use general results in Theorem 3.3. Due to additivity, $\hat{Y}_* = \hat{Y} = \sum_U \omega_i y_i$, being the linear estimator of the population total. Consequently $V(\hat{Y}_*) = V(\hat{Y}) = \sum\sum_{\mathfrak{D}} V_{ij}$, and thus the denominator in the formulae (3.15) – (3.17) is given by the variance of $\hat{Y}$ in case HG-design (1.41):

$$\sum\sum_{\mathfrak{D}} V_{ij} = \frac{c}{n}\left(\sum_U \frac{y_i^2}{p_i} - Y^2\right) = \frac{c}{n}v.$$

Noting that $\sum_{i\in\mathfrak{D}} V_{di} = Cov(\hat{Y}_d, \sum_{\mathfrak{D}} \hat{Y}_i) = Cov(\hat{Y}_d, \hat{Y})$, we have due to (2.29):

$$\sum_{i\in\mathfrak{D}} V_{di} = \frac{c}{n}\left(\sum_U \frac{(y_i^d)^2}{p_i} - YY_d\right) = \frac{c}{n}v_d.$$

Inserting these results into (3.15) – (3.17) we get (3.32) – (3.34).

□

3.4 GR-estimator for domains when population total is estimated

Let us assume that population total Y_0 is estimated from a survey by $\hat{Y}_0$, and domain totals Y_i, $i \in \mathfrak{D}$, are estimated from the same or from other survey by $\hat{Y}_i$. The parameter vector $\boldsymbol{\theta} = (\ Y_0,\ \ Y_1,\ \ Y_2,\ \ \ldots,\ \ Y_D\ )'$ is estimated by $\hat{\boldsymbol{\theta}} = (\ \hat{Y}_0,\ \ \hat{Y}_1,\ \ \hat{Y}_2,\ \ \ldots,\ \ \hat{Y}_D\ )'$. Let us denote the variance of population estimator by $V_{00} = V(\hat{Y}_0)$, of domain estimators by $V_{dd} = V(\hat{Y}_d)$, covariance between estimators of domain U_d and population total by $V_{d0} = Cov(\hat{Y}_d, \hat{Y}_0)$, and covariance between two domain estimators by $V_{dg} = Cov(\hat{Y}_d, \hat{Y}_g)$. Then covariance matrix of vector $\hat{\boldsymbol{\theta}}$ takes the form:

$$\mathbf{V} = \begin{pmatrix} V_{00} & V_{01} & \cdots & V_{0D} \\ V_{10} & V_{11} & \cdots & V_{1D} \\ \cdots & \cdots & \ddots & \cdots \\ V_{D0} & V_{D1} & \cdots & V_{DD} \end{pmatrix}. \tag{3.35}$$

Let $\mathbf{V}$ be nonsingular, i.e. it is positive definite. Let the vector of restriction coefficients be

$$\mathbf{R} = \ (-1 \quad 1 \quad 1 \quad \ldots \quad 1 \quad \ldots \quad 1) \ = \begin{pmatrix} -1 \\ \mathbf{1}_D \end{pmatrix}'. \tag{3.36}$$

Let us use the restriction:

$$\mathbf{R}\boldsymbol{\theta} = -Y_0 + \sum\nolimits_{\mathfrak{D}} Y_i = -Y_0 + Y_* = 0, \tag{3.37}$$

where $Y_* = \sum_{\mathfrak{D}} Y_i$. The restriction postulates that the domain totals have to sum up to population total.

Theorem 3.4. The GR-estimators for domain total Y_d and population total Y_0, based on initial survey estimators $\hat{Y}_i$, $i \in \mathfrak{D}$, and $\hat{Y}_0$ under restriction (3.37), are

$$\hat{Y}_d^{GR} = \hat{Y}_d + (\hat{Y}_0 - \hat{Y}_*)(\sum_{i\in\mathfrak{D}} V_{di} - V_{d0})/v, \tag{3.38}$$

$$\hat{Y}_0^{GR} = \hat{Y}_0 + (\hat{Y}_0 - \hat{Y}_*)(\sum_{i\in\mathfrak{D}} V_{i0} - V_{00})/v, \tag{3.39}$$

where $\hat{Y}_* = \sum_{\mathfrak{D}} \hat{Y}_i$ and $v = \sum\sum_{\mathfrak{D}} V_{ij} + V_{00} - 2\sum\limits_{i\in\mathfrak{D}} V_{i0}$ and $\mathfrak{D} = \{1, 2, ..., D\}$. If $\hat{Y}_0$ and $\hat{Y}_i$ are unbiased estimators of Y and Y_i, $i \in \mathfrak{D}$, respectively, then corresponding GR-estimators are unbiased.

The variances and covariances of GR-estimators are:

$$V_{dd}^{GR} = V_{dd} - \left(\sum_{i\in\mathfrak{D}} V_{di} - V_{d0}\right)^2 /v. \tag{3.40}$$

$$V_{00}^{GR} = V_{00} - \left(\sum_{i\in\mathfrak{D}} V_{i0} - V_{00}\right)^2 /v, \tag{3.41}$$

$$V_{dg}^{GR} = V_{dg} - \left(\sum_{i\in\mathfrak{D}} V_{di} - V_{d0}\right)\left(\sum_{i\in\mathfrak{D}} V_{gi} - V_{g0}\right) /v, \tag{3.42}$$

$$V_{d0}^{GR} = V_{d0} - \left(\sum_{i\in\mathfrak{D}} V_{i0} - V_{00}\right)\left(\sum_{i\in\mathfrak{D}} V_{di} - V_{d0}\right) /v. \tag{3.43}$$

Proof: With $\mathbf{R}$ as in (3.36) we get from (3.4):

$$\mathbf{K} = \mathbf{V}\begin{pmatrix} -1 \\ \mathbf{1}_D \end{pmatrix}\left[\begin{pmatrix} -1 \\ \mathbf{1}_D \end{pmatrix}' \mathbf{V}\begin{pmatrix} -1 \\ \mathbf{1}_D \end{pmatrix}\right]^{-1}. \tag{3.44}$$

Inserting matrix $\mathbf{V}$ in (3.35) into (3.44) we get:

$$\mathbf{K} = \frac{1}{v} \begin{pmatrix} \sum\limits_{i \in \mathfrak{D}} V_{i0} - V_{00} \\ \sum\limits_{i \in \mathfrak{D}} V_{1i} - V_{10} \\ \cdots \\ \sum\limits_{i \in \mathfrak{D}} V_{Di} - V_{D0} \end{pmatrix}. \tag{3.45}$$

Noting that in our case the constant $\mathbf{c}$ in (3.2) is equal to zero, $\mathbf{c} = 0$, we can develop:

$$\mathbf{c} - \mathbf{R}\hat{\boldsymbol{\theta}} = -(-\hat{Y}_0 + \sum\nolimits_{\mathfrak{D}} \hat{Y}_i) = \hat{Y}_0 - \hat{Y}_*. \tag{3.46}$$

Inserting (3.45) and (3.46) into general formula of restriction estimator (3.2), we get (3.38) and (3.39).

Inserting (3.44) into (3.3) we get

$$\mathbf{V}^{GR} = \left\{ \mathbf{I}_D - \mathbf{V} \begin{pmatrix} -1 \\ \mathbf{1}_D \end{pmatrix} \left[\begin{pmatrix} -1 \\ \mathbf{1}_D \end{pmatrix}' \mathbf{V} \begin{pmatrix} -1 \\ \mathbf{1}_D \end{pmatrix} \right]^{-1} \begin{pmatrix} -1 \\ \mathbf{1}_D \end{pmatrix}' \right\} \mathbf{V}.$$

Since $\mathbf{V}$ is positive definite, the following scalar in the above expression is strictly positive:

$$v = \left[\begin{pmatrix} -1 \\ \mathbf{1}_D \end{pmatrix}' \mathbf{V} \begin{pmatrix} -1 \\ \mathbf{1}_D \end{pmatrix} \right] = \left(\sum \sum\nolimits_{\mathfrak{D}} V_{ij} + V_{00} - 2 \sum_{i \in \mathfrak{D}} V_{i0} \right) > 0.$$

Note that alternatively

$$v = V(\hat{Y}_0 - \hat{Y}_*). \tag{3.47}$$

Taking v^{-1} out, and using $\begin{pmatrix} -1 \\ \mathbf{1}_D \end{pmatrix}' \mathbf{V} = \left[\mathbf{V} \begin{pmatrix} -1 \\ \mathbf{1}_D \end{pmatrix} \right]'$, we have

$$\mathbf{V}^{GR} = \mathbf{V} - \left[\mathbf{V} \begin{pmatrix} -1 \\ \mathbf{1}_D \end{pmatrix} \right] \left[\mathbf{V} \begin{pmatrix} -1 \\ \mathbf{1}_D \end{pmatrix} \right]' v^{-1}. \tag{3.48}$$

The covariance V_{dg}^{GR} is the $(d+1, g+1)$ element and V_{d0}^{GR} is the $(d, 1)$ element of the matrix $\mathbf{V}^{GR}$. Noting that the first element of the vector $\mathbf{V} \begin{pmatrix} -1 \\ \mathbf{1}_D \end{pmatrix}$ is $\sum_{i \in \mathfrak{D}} V_{0i} - V_{00}$ and $d+1$ element is $\sum_{i \in \mathfrak{D}} V_{di} - V_{d0}$, the covariances of two GR-estimators (3.42) and (3.43) follow. Variances (3.40) and (3.41) of GR-estimators follow from (3.48) as its diagonal elements (0,0) and $(d+1, d+1)$.

To show unbiasedness of $\hat{Y}_d^{GR}$ and $\hat{Y}_0^{GR}$ note that $E(\hat{Y}_d) = Y_d$ and $E(\hat{Y}_0) = Y$ by assumption. Since this holds for all $d \in \mathfrak{D}$, we have

$$E(\hat{Y}_0 - \hat{Y}_*) = E(\hat{Y}_0 - \sum\nolimits_{\mathfrak{D}} \hat{Y}_i) = Y - \sum\nolimits_{\mathfrak{D}} Y_i = 0.$$

The last equality comes from condition for totals $\sum_{\mathfrak{D}} Y_i = Y$. Therefore the expectation of GR-estimator is equal to the true total, $E(\hat{Y}_d^{GR}) = Y_d$.

□

The next corollary sheds further light on the interpretation of restriction estimator.

Corollary 3.9. The restriction estimator of domain and population totals in Theorem 3.4 can be presented alternatively in the form:

$$\begin{aligned} \hat{Y}_d^{GR} &= \hat{Y}_d + T \cdot \frac{Cov(\hat{Y}_d, T)}{V(T)}, \\ \hat{Y}_0^{GR} &= \hat{Y}_0 + T \cdot \frac{Cov(\hat{Y}_0, T)}{V(T)}, \end{aligned}$$

where $T = \hat{Y}_* - \hat{Y}_0$.

Proof: The corollary follows by noting that

$$\begin{aligned} v &= V(\hat{Y}_* - \hat{Y}_0), \\ \sum_{i \in \mathfrak{D}} V_{di} - V_{d0} &= Cov(\hat{Y}_d, \hat{Y}_* - \hat{Y}_0), \\ \sum_{i \in \mathfrak{D}} V_{i0} - V_{00} &= Cov(\hat{Y}_0, \hat{Y}_* - \hat{Y}_0). \end{aligned}$$

□

Remark 3.5. We see from Corollary 3.9 that the restriction $\sum_{\mathfrak{D}} \hat{Y}_i^{GR} - \hat{Y}_0^{GR} = 0$ is satisfied for any such numbers $a_d, a_0, d \in \mathfrak{D}$, instead of $Cov(\hat{Y}_d, T)/V(T)$ and $Cov(\hat{Y}_0, T)/V(T)$, for which

$$\sum\nolimits_{\mathfrak{D}} a_d - a_0 = 1.$$

The resulting estimators $\hat{Y}_d^{GR} = \hat{Y}_d + a_d T$ and $\hat{Y}_0^{GR} = \hat{Y}_0 + a_0 T$ are still restriction estimators, but not optimal (in the sense of variance). However the practitioner being able to choose a_d, a_0 close to the true values, simplifies

the restriction estimator considerably. The simplest possibility is to choose $a_d = a_0 = \frac{1}{D-1}$, where D is the number of domains.

Example 3.3. Let us check that in the case of additive domain estimators, i.e. in the case

$$\hat{Y}_* = \sum\nolimits_{\mathfrak{D}} \hat{Y}_i = \hat{Y}_0, \tag{3.49}$$

the restrictions do not change the initial estimators:

$$\hat{Y}_d^{GR} = \hat{Y}_d,\ d = 0, 1, 2, ..., \mathfrak{D}, \tag{3.50}$$
$$V_{dg}^{GR} = V_{dg},\ d, g = 0, 1, 2, ..., \mathfrak{D}. \tag{3.51}$$

Using assumption (3.49) the equality (3.50) follows directly from (3.38) and (3.39). Noting that

$$\sum_{i \in \mathfrak{D}} V_{di} - V_{d0} = Cov(\hat{Y}_d, (\hat{Y}_* - \hat{Y}_0)) \tag{3.52}$$

and

$$\sum_{i \in \mathfrak{D}} V_{i0} - V_{00} = Cov(\hat{Y}_*, \hat{Y}_0) - V_{00}, \tag{3.53}$$

which both equal zero under assumption (3.49). Correspondingly (3.51) follows.

□

Theorem 3.4 allows dependence of estimators. It simplifies if the estimators are uncorrelated. We can construct the uncorrelated domain estimators (e.g. ratio estimators are approximately uncorrelated). The population estimator is independent from domain estimators if it is estimated from another survey.

Corollary 3.10. Let $\hat{Y}_d$, $d \in \mathfrak{D}$, be uncorrelated between themselves and with $\hat{Y}_0$ as well. Then under restriction (3.37) the GR-estimators for domain total Y_d and population total Y_0 are:

$$\hat{Y}_d^{GR} = \hat{Y}_d + (\hat{Y}_* - \hat{Y}_0)V_{dd}/v, \tag{3.54}$$
$$\hat{Y}_0^{GR} = \hat{Y}_0 - (\hat{Y}_* - \hat{Y}_0)V_{00}/v, \tag{3.55}$$

where $v = \sum_{\mathfrak{D}} V_{ii} + V_{00}$.

The variances of GR-estimators are:

$$V_{dd}^{GR} = V_{dd} - V_{dd}^2/v, \tag{3.56}$$
$$V_{00}^{GR} = V_{00} - V_{00}^2/v. \tag{3.57}$$

The covariances of GR-estimators are:

$$V_{dg}^{GR} = -V_{dd}V_{gg}/v, \tag{3.58}$$
$$V_{d0}^{GR} = V_{00}V_{dd}/v. \tag{3.59}$$

Proof: By assumption, all estimators are uncorrelated: $V_{ij} = 0, i \neq j, i, j \in \mathfrak{D}$, and $V_{i0} = 0, i \in \mathfrak{D}$. Therefore the covariance matrix (3.35) of $\hat{\boldsymbol{\theta}}$ takes the form:

$$\mathbf{V} = \begin{pmatrix} V_{00} & 0 & \cdots & 0 \\ 0 & V_{11} & \cdots & 0 \\ \cdots & \cdots & \ddots & \cdots \\ 0 & 0 & \cdots & V_{DD} \end{pmatrix},$$

and the denominator v in (3.38) – (3.43) is:

$$v = \sum\sum\nolimits_{\mathfrak{D}} V_{ij} + V_{00} - 2\sum_{i\in\mathfrak{D}} V_{i0} = \sum\nolimits_{\mathfrak{D}} V_{ii} + V_{00}.$$

GR-estimators (3.38) and (3.39) reduce to (3.54) and (3.55) respectively. Variances and covariances (3.56) – (3.59) follow from (3.40) – (3.43) respectively.

□

The relative decrease of variance is

$$\frac{V_{dd}^{GR} - V_{dd}}{V_{dd}} = -\frac{V_{dd}}{v}.$$

It depends on the ratio of the variance of initial estimator to the sum of variances of initial estimators. The relative decrease of variance is higher if the initial estimators are more variable and if the variability of population total decreases.

From (3.58) we see that in the case of uncorrelated initial estimators the corresponding GR-estimators of domains U_d and U_g are negatively correlated. The GR-estimator of domain U_d and of the population total is positively correlated (see 3.59).

Corollary 3.11. Assume that the population total Y and domain totals $Y_d, d \in \mathfrak{D}$, are estimated so that domain estimators are uncorrelated. Assuming restriction (3.37) the GR-estimators for domain U_d and population total are

$$\hat{Y}_d^{GR} = \hat{Y}_d + (\hat{Y}_0 - \hat{Y}_*)(V_{dd} - V_{d0})/v, \tag{3.60}$$
$$\hat{Y}_0^{GR} = \hat{Y}_0 + (\hat{Y}_0 - \hat{Y}_*)(\sum_{i\in\mathfrak{D}} V_{i0} - V_{00})/v, \tag{3.61}$$

where $v = \sum_{\mathfrak{D}} V_{ii} + V_{00} - 2 \sum_{i \in \mathfrak{D}} V_{i0}$.

The variances and covariances of GR-estimators are:

$$V_{dd}^{GR} = V_{dd} - (V_{dd} - V_{d0})^2 / v, \tag{3.62}$$

$$V_{00}^{GR} = V_{00} - \left(\sum_{i \in \mathfrak{D}} V_{i0} - V_{00} \right)^2 / v, \tag{3.63}$$

$$V_{dg}^{GR} = -(V_{dd} - V_{d0})(V_{gg} - V_{g0}) / v, \tag{3.64}$$

$$V_{d0}^{GR} = V_{d0} - \left(\sum_{i \in \mathfrak{D}} V_{i0} - V_{00} \right)(V_{dd} - V_{d0}) / v. \tag{3.65}$$

Proof: Due to the assumption $V_{ij} = 0$, $i \neq j$, $i, j \in \mathfrak{D}$, the covariance matrix (3.35) takes the form

$$\mathbf{V} = \begin{pmatrix} V_{00} & V_{01} & V_{02} & \cdots & V_{0D} \\ V_{10} & V_{11} & 0 & \cdots & 0 \\ V_{20} & 0 & V_{22} & \cdots & 0 \\ \cdots & \cdots & \cdots & \ddots & \cdots \\ V_{D0} & 0 & 0 & \cdots & V_{DD} \end{pmatrix},$$

and the sums simplify to $\sum \sum_{\mathfrak{D}} V_{ij} = \sum_{\mathfrak{D}} V_{ii}$ and $\sum_{i \in \mathfrak{D}} V_{di} = V_{dd}$. Consequently the denominator v takes the form:

$$\begin{aligned} v &= \sum \sum_{\mathfrak{D}} V_{ij} + V_{00} - 2 \sum_{i \in \mathfrak{D}} V_{i0} \\ &= \sum_{\mathfrak{D}} V_{ii} + V_{00} - 2 \sum_{i \in \mathfrak{D}} V_{i0}. \end{aligned}$$

Inserting this into (3.38) – (3.43) we get (3.60) –(3.65).

□

The gain in efficiency of the GR-estimator depends on the variance of initial estimator and the covariance between estimator of domain U_d and population estimator:

$$V_{dd}^{GR} - V_{dd} = -(V_{dd} - V_{d0})^2 / v.$$

Since the numerator is nonpositive due to square and denominator $v > 0$ (see theorem 3.3) the variance of GR-estimator is not higher than the variance of corresponding initial estimator, $V_{dd}^{GR} \leq V_{dd}$.

Remark 3.6. The situation considered in Corollary 3.11 holds in a survey where ratio estimator is used both for domain and population totals.

Our result that for ratio estimators, $AV(\hat{Y}_d) = ACov(\hat{Y}_d, \hat{Y}_0)$ (see Corollaries 2.6, 2.13 and Remark 2.4), gives importance to the following corollary.

Corollary 3.12. Under assumptions of Corollary 3.11 and under an additional assumption $V_{dd} = V_{d0}$, $d \in \mathfrak{D}$, GR-estimators satisfying restriction (3.37) for domain U_d and for the population total are:

$$\begin{aligned} \hat{Y}_d^{GR} &= \hat{Y}_d, \quad \hat{Y}_0^{GR} = \sum\nolimits_{\mathfrak{D}} \hat{Y}_i, \\ V_{dd}^{GR} &= V_{dd}, \; V_{00}^{GR} = \sum\nolimits_{\mathfrak{D}} V_{ii}, \\ V_{dg}^{GR} &= 0, \quad V_{d0}^{GR} = V_{d0}. \end{aligned}$$

Proof: The proof follows from Corollary 3.11 under assumption $V_{dd} = V_{d0}$.

□

Remark 3.7. The Corollary 3.12 gives interesting results for a survey where ratio estimator is used for domains as well for the population total, sample design is SI- or HG-design and sample size is not so small. The optimal case which satisfies the restrictions is achieved by keeping initial domain estimators unchanged and by changing the estimated population total. The restriction estimator for the population total is just the sum of the initial domain estimators.

Remark 3.8. To express the results of this paragraph for special sampling designs, one needs to know the expressions of variances and covariances of domain estimators under these designs. In this thesis these expressions are readily available for SI-, HG- and multinomial designs.

3.5 Conditional GR-estimator for domains

Conditional restriction estimator for domains is based on the general form of the conditional GR-estimator given in Theorem 3.2. We assume that some population parameters $\hat{\boldsymbol{\theta}}_1$ (e.g. population total) are previously estimated and desired to keep unchanged. We have initial unbiased estimators $\hat{\boldsymbol{\theta}}_2 = (\hat{Y}_1, \hat{Y}_2, ..., \hat{Y}_D)'$ for domain totals $\boldsymbol{\theta}_2 = (Y_1, Y_2, ..., Y_D)'$. For solving consistency problem for $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$ with conditional GR-estimator we define restriction

matrix $\mathbf{R}$. In case $\hat{\boldsymbol{\theta}}_1 = \hat{Y}$ is an estimator of the population total, the restriction (3.8) takes the form:

$$\mathbf{R}\hat{\boldsymbol{\theta}}_2^{GR} = \begin{pmatrix} 1 & \cdots & 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} \hat{Y}_1^{GR} \\ \cdots \\ \hat{Y}_d^{GR} \\ \cdots \\ \hat{Y}_D^{GR} \end{pmatrix} = \hat{Y}, \tag{3.66}$$

Variances/covariances of initial estimators are denoted by $\mathbf{V_1} = V(\hat{Y})$,
$\mathbf{V_2} = Cov(\hat{\boldsymbol{\theta}}_2)$ and
$\mathbf{V_{21}} = Cov(\hat{\boldsymbol{\theta}}_2, \hat{\boldsymbol{\theta}}_1) = [Cov(\hat{Y}_1, \hat{Y}), Cov(\hat{Y}_2, \hat{Y}), ..., Cov(\hat{Y}_D, \hat{Y})]'$.

Let the considered estimators $\hat{Y}$ and $\hat{Y}_1$, $\hat{Y}_2$, ..., $\hat{Y}_D$ be not additive, i.e. $\sum_{\mathfrak{D}} \hat{Y}_i = \hat{Y}_* \neq \hat{Y}$. Assume uncorrelated domain estimators with covariance matrix $\mathbf{V_2} = diag(V_{ii})$, the variance of the estimated population total is denoted by $\mathbf{V_1} = V_{00}$, the covariance between estimators of domain and of population totals is denoted by $\mathbf{V_{21}} = (V_{10}, V_{20}, ..., V_{D0})'$.

Corollary 3.13. In case of uncorrelated unbiased domain estimators $\hat{Y}_i$, $i \in \mathfrak{D}$, the conditional GR-estimator is

$$\hat{Y}_d^{GRc} = \hat{Y}_d + (\hat{Y} - \hat{Y}_*) \cdot \frac{V_{dd}}{v} \tag{3.67}$$

and its variance is

$$\begin{aligned} V_{dd}^{GRc} &= V_{dd} - \frac{V_{dd}^2}{\sum_{\mathfrak{D}} V_{ii}} + \frac{V_{00} V_{dd}^2}{v^2} \\ &+ 2\,\frac{V_{d0} V_{dd} \cdot v - V_{dd}^2 \sum_{\mathfrak{D}} V_{i0}}{v^2}, \end{aligned} \tag{3.68}$$

where

$$v = \sum_{\mathfrak{D}} V_{ii}.$$

The covariance of two GR-estimators is

$$\begin{aligned} Cov(\hat{Y}_d^{GRc}, \hat{Y}_g^{GRc}) &= -\frac{V_{dd} V_{gg}}{v} + \frac{V_{00} V_{dd} V_{gg}}{v^2} \\ &+ 2\,\frac{V_{d0} V_{gg} \cdot v - V_{dd} V_{gg} \sum_{\mathfrak{D}} V_{i0}}{v^2}. \end{aligned} \tag{3.69}$$

Proof: Since $\hat{Y}$ is considered fixed, the form of conditional GR-estimator $\hat{Y}_d^{GRc}$ comes from the form of the GR-estimator in (3.24) that assumes known population total. As a result, (3.67) follows where the estimated population total $\hat{Y}$ is used instead of the known population total Y.

The general covariance matrix formula (3.10) is used to develop the variance and covariance expressions for our case. The first term of (3.10) reduces in our case to the formulae (3.25) − (3.26) which will be included into V_{dd}^{GRc} and $Cov(\hat{Y}_d^{GRc}, \hat{Y}_g^{GRc})$ respectively. In our case the matrix $\mathbf{K}$ takes the form:

$$\mathbf{K} = \frac{1}{\sum_{\mathfrak{D}} V_{ii}} \begin{pmatrix} V_{11} \\ V_{22} \\ ... \\ V_{DD} \end{pmatrix}.$$

Therefore the second term of (3.10) is:

$$\mathbf{K}\mathbf{V_1}\mathbf{K}' = \frac{V_{00}}{\left(\sum_{\mathfrak{D}} V_{ii}\right)^2} \begin{pmatrix} V_{11} \\ V_{22} \\ ... \\ V_{DD} \end{pmatrix} \begin{pmatrix} V_{11} & V_{22} & ... & V_{DD} \end{pmatrix},$$

giving the subsequent terms in (3.68) − (3.69).

The third and fourth term of (3.10) are equal in our case and give the last term of (3.68) and (3.69).

□

Corollary 3.14. In case $V_{dd} = V_{d0}$, $d \in \mathfrak{D}$, the formulae of variance and covariance in Corollary 3.13 reduce to

$$\begin{aligned} V_{dd}^{GRc} &= V_{dd} - \frac{V_{dd}^2}{v} + \frac{V_{00}V_{dd}^2}{v^2}, \\ Cov(\hat{Y}_d^{GRc}, \hat{Y}_g^{GRc}) &= -\frac{V_{dd}V_{gg}}{v} + \frac{V_{00}V_{dd}V_{gg}}{v^2}. \end{aligned} \tag{3.70}$$

Proof: Replacing V_{d0} by V_{dd} in the numerator of the last term of (3.69) we get:

$$V_{dd}V_{gg}\sum\nolimits_{\mathfrak{D}} V_{ii} - V_{dd}V_{gg}\sum\nolimits_{\mathfrak{D}} V_{ii} = 0.$$

Analogically, the numerator of the last term of (3.68) is equal to zero.

□

Remark 3.9. The Corollary 3.14 can be used in not so small samples for estimating variance and covariance of the conditional GR-estimators based on the initial ratio estimators under SI and HG sampling designs. The approximate variance of domain ratio estimator, AV_{dd}, is equal to the approximate

covariance, AV_{d0}, between ratio estimators for a domain and population total (see Corollaries 2.6, 2.13 and Remark 2.4).

Remark 3.10. Note the opposite situation for building restriction estimators in the conditional and unconditional cases for initial ratio estimators. In the conditional case the estimated population total $\hat{Y}$ is kept fixed and domain estimators change. In the unconditional case (Remark 3.7), the domain estimators remain fixed but $\hat{Y}$ changes.

In general the conditional GR-estimators are not more effective than the initial estimators. The next corollary shows in which case they are.

Corollary 3.15. Under assumptions of Corollaries 3.13 and 3.14, the conditional GR-estimator for a domain is more effective than the initial estimator $\hat{Y}_d$ if

$$\frac{V_{00}}{\sum_{\mathfrak{D}} V_{ii}} < 1,$$

where $V_{00} = V(\hat{Y})$ and $V_{ii} = V(\hat{Y}_i)$.

Proof: The corollary is proved when writing (3.70) in the form:

$$V_{dd}^{GRc} = V_{dd} - \frac{V_{dd}^2}{v}\left(1 - \frac{V_{00}}{v}\right),$$

Notice that $V_{dd}^{GRc} < V_{dd}$ if $1 - V_{00}/v > 0$.

□

Chapter 4

Simulation study

In this chapter the performed simulation study is described, the simulation results are presented and compared. One aim of the simulation study was to check the derived variance/covariance formulae of domain estimators. Some of these formulae were the asymptotic ones. The important issue here was to check how the asymptotic formulae perform in the practical situation where the sample size is not so big. Another aim was to illustrate the magnitude of dependence of different domain estimators under different sampling designs, and to study the performance of GR-estimators for domains in different situations.

An artificial population was generated for simulations with two variables (continuous and binary). Two sampling designs, the SI- and the HG-design with comparable sample sizes (in persons) were used. Simulations were performed in SAS environment. Necessary programs were written by the author in SAS-base and SAS-IML.

4.1 Population, sample and performance criteria

4.1.1 Population

An artificial population was created for the simulation study. The population consisted of 2000 persons comprising 1192 households (HH). The real data of the Estonian Labour Force Survey was used. More precisely, the following variables were included into the population database for each person.

The two study variables were:

– monthly salary (a continuous variable: in thousand kroons),

– higher education (a binary variable: 1 - person has higher education, 0 - otherwise).

The necessary auxiliary variables were:

– household ID (identifier of the HH involving that person),

– household member ID (identifier of the person in the household),

– household size (the number of persons in the household),

– domain indicator d ($d = 1, 2, 3$).

All members of the household belong to one and the same domain. Tables 4.1 and 4.2 present main characteristics of the created population.

Table 4.1: Population characteristics, domain sizes

Domain	no. of persons	%	no. of HHs	%	Average HH size
1	1 019	51.0	604	50.7	1.69
2	733	36.6	442	37.1	1.66
3	248	12.4	146	12.2	1.70
Population	2 000	100.0	1 192	100.0	1.68

It can be seen from Table 4.1 that domains have different sizes. The first domain is the largest and the third domain is the smallest. The average household size is approximately the same in each domain.

The study variables perform differently in domains (Table 4.2). The mean value and the standard deviation (S_{yy}) of both study variable are the largest in the second domain. For higher education, the total shows the number of persons with higher education and the mean shows their proportion.

4.1.2 Sample design and data issues

10,000 independent samples were drawn from the population by SI- and HG-sampling. The sample size for SI-design was 200 persons and for HG-design 101 households, resulting on the average in 200.9 persons. Sample sizes for

Table 4.2: Population characteristics, study variables

Domain	Total	Mean	Minimum	Maximum	S_{yy}
Monthly salary					
1	4 998.9	4.906	0.100	37.580	3.194
2	4 614.8	6.296	0.500	46.660	4.309
3	1 396.3	5.630	0.200	23.560	3.285
Population	11 010.1	5.505	0.100	46.660	3.707
Higher education					
1	129	0.127	0	1	0.333
2	209	0.285	0	1	0.452
3	36	0.145	0	1	0.353
Population	374	0.187	0	1	0.390

domains were random in both cases. Table 4.3 shows average, minimum and maximum sample sizes in domains.

Table 4.3: Sample sizes in domains and population

Domain	Average	Minimum	Maximum
SI-design, persons			
1	101.8	77	127
2	78.3	48	99
3	24.9	9	43
Population	200	200	200
HG-design, households			
1	51.4	33	72
2	37.1	20	46
3	12.5	3	28
Population	101	101	101
HG-design, persons			
1	102.7	57	150
2	73.6	37	114
3	24.6	4	53
Population	200.9	172	234

The variability of sample size is larger for the third (smallest) domain where the maximum sample size is much bigger than the minimum sample size (almost five times for SI-design and about ten times for HG-design). For the first and the second domains this ratio is about twice for both designs.

Two types of initial estimators were calculated on each sample:

– linear estimator of domain and population totals,

– ratio estimator of domain and population totals using domain size as an auxiliary information.

Based on the population data, the theoretical (known) covariance matrices of initial linear and ratio estimators were calculated for both designs. The estimated covariance matrices of initial estimators, separately for linear and ratio estimators, were calculated on each sample (see Sections 2.3 – 2.5).

The following GR-estimators were calculated on each sample (see Sections 3.3 – 3.5):

– GR-estimator for domains when the population total is known (both for known and estimated covariance matrix of initial estimators);

– GR-estimator for domains and for the population total when the latter is estimated (only ratio initial estimators, both for known and estimated covariance matrices of initial estimators);

– conditional GR-estimator for domains when the population total is estimated and fixed (only ratio initial estimators and known covariance matrix of initial estimators).

The covariance matrices of GR-estimators were calculated with formulae in Sections 3.3 – 3.5, as well empirically, over simulated values of estimators.

4.1.3 Performance criteria

The following measures were applied to compare the performance of the different estimators over M simulations: the relative bias

$$RB(\hat{Y}_d) = \frac{\frac{1}{M}\sum_{m=1}^{M}\hat{Y}_d^{(m)} - Y_d}{Y_d} \tag{4.1}$$

and the relative root mean square error

$$RRMSE(\hat{Y}_d) = \frac{\sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(\hat{Y}_d^{(m)} - Y_d\right)^2}}{Y_d}, \tag{4.2}$$

where $\hat{Y}_d^{(m)}$ is the estimated domain total of the study variable from the mth simulation and Y_d refers to the true domain total. $\hat{Y}_d^{(m)}$ denotes any estimator in the simulation study (both initial and GR-estimators).

Three covariance matrices (theoretical, its estimate and empirical) were calculated for initial and GR-estimators. The theoretical covariance matrices were compared to the respective empirical covariance matrices with the aim to check the derived formulae, especially the performance of the asymptotic ones. The estimated covariance matrices were compared to the theoretical ones with respect to the bias and mean square error.

The elements of empirical covariance matrix were calculated over M simulations as follows:

$$V_{dg}^{Emp} = \frac{1}{M-1} \sum_{m=1}^{M} (\hat{Y}_d^{(m)} - \hat{\bar{Y}}_d^{(m)})(\hat{Y}_g^{(m)} - \hat{\bar{Y}}_g^{(m)}),$$

where $\hat{\bar{Y}}_d^{(m)} = \frac{1}{M} \sum_{m=1}^{M} \hat{Y}_d^{(m)}$ is the mean value of the estimates of Y_d. Diagonal elements of the covariance matrix, i.e. empirical variances were obtained if $d = g$.

Elements of the theoretical (known) covariance matrix V_{dg} of initial estimators were calculated based on the formulae in Sections 2.3 – 2.5. The elements of theoretical covariance matrix of GR-estimators using known covariance matrix of initial estimators were calculated with formulae of Section 3.3. The variances, as more important, were compared.

The relative difference between the theoretical and of empirical variance was formed:

$$RD(V_{dd}) = \frac{V_{dd} - V_{dd}^{Emp}}{V_{dd}^{Emp}}. \tag{4.3}$$

Thirdly, the covariance matrix of initial estimators was estimated on each sample using results of Sections 2.3 – 2.5. Based on this, the estimated covariance matrix of GR-estimator was calculated on each sample. The formulae in Section 3.3 were used. The estimated variances $\hat{V}_{dd}^{(m)}$ were averaged over simulations,

$$\hat{\bar{V}}_{dd} = \frac{1}{M} \sum_{m=1}^{M} \hat{V}_{dd}^{(m)},$$

and compared to the theoretical variances similarly to (4.1):

$$RB(\hat{V}_{dd}) = \frac{\hat{\bar{V}}_{dd} - V_{dd}}{V_{dd}}.$$

To show the stability of variance estimators, the relative root mean square error $RRMSE(\hat{V}_{dd})$ was calculated as in (4.2) inserting $\hat{V}_{dd}$ and V_{dd} instead of $\hat{Y}_d$ and Y_d respectively.

4.2 Simulation results

4.2.1 Illustration of a consistency problem

It is desirable that the estimated domain totals sum up to the known or estimated population total. Then the estimates are called consistent. However, often in practice this consistency does not hold. The difference between the sum of initial domain estimators and population total (known or estimated) characterises the consistency problem:

$$\begin{aligned} Diff &= \sum\nolimits_D \hat{Y}_d - Y, \\ RDiff &= Diff/Y, \end{aligned}$$

where $\hat{Y}_d$ is linear or ratio estimator of domain d, Y is known population total and *RDiff* is relative difference. For ratio estimator also the difference from estimated population total was analysed, i.e. when Y is replaced by $\hat{Y}$ in the above formulae. Tables 4.4 and 4.5 show the mean, minimum and maximum difference over simulations. Also the proportion of samples for which $|RDiff| < 0.05$ was calculated.

Table 4.4: Difference *Diff* if population total is known

Design, estimator	Mean	Minimum	Maximum	$\lvert RDiff\rvert < 0.05$
Continuous variable ($Y = 11010.1$)				
SI, linear estimator	1.1	-1536.4	1886.2	0.74
SI, ratio estimator	0.9	-1525.8	1982.9	0.74
HG, linear estimator	13.6	-1856.8	3105.1	0.64
HG, ratio estimator	11.8	-1784.7	2950.3	0.65
Binary variable ($Y = 374$)				
SI, linear estimator	3.4	-182.0	228.0	0.30
SI, ratio estimator	3.4	-182.1	221.9	0.28
HG, linear estimator	3.1	-226.8	301.3	0.22
HG, ratio estimator	2.8	-211.8	297.8	0.23

The consistency problem is stronger for the HG-design and binary variable. For binary variable the relative difference *RDiff* is larger than 0.05 for more

than 70% of samples. In case of continuous variable one quarter of samples for SI-design and one third for HG-design have relative difference larger than 0.05.

Table 4.5: Difference *Diff* if population total is estimated

Design, estimator	Mean	Minimum	Maximum	$\lvert RDiff \rvert < 0.05$
		Continuous variable		
SI, ratio estimator	-0.2	-550.1	611.7	0.9999
HG, ratio estimator	-1.8	-963.8	1104.6	0.996
		Binary variable		
SI, ratio estimator	0.0	-49.7	52.1	0.90
HG, ratio estimator	-0.3	-90.6	79.4	0.76

Difference from estimated population total is much smaller compared to the known population total. In most of the cases the relative difference is below 5%. Only for HG-design and binary variable the consistency problem is more severe. The figures 4.1 and 4.2 show the distribution of relative difference.

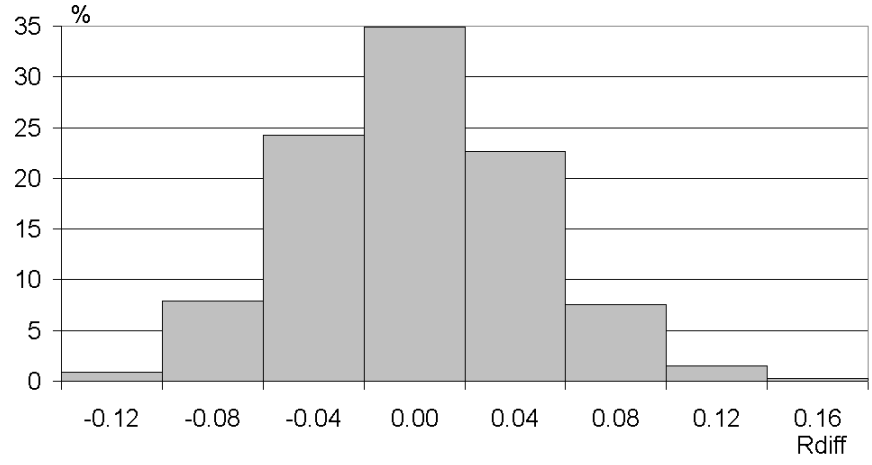

Figure 4.1: Distribution of relative difference (SI-design, known population total, continuous variable)

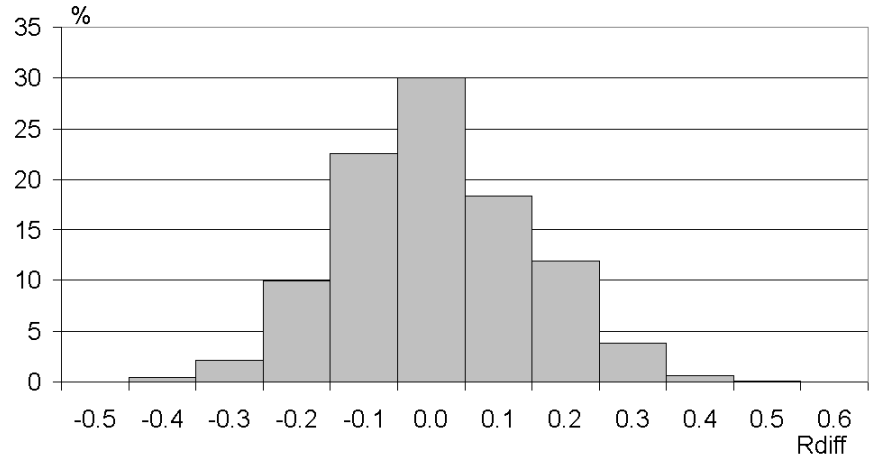

Figure 4.2: Distribution of relative difference (SI-design, known population total, binary variable)

4.2.2 Initial estimators

Two different initial estimators were considered, the linear and the ratio estimator, calculated respectively by (2.4) and (2.6). Under SI-design theoretical covariance matrix of initial linear estimators was calculated by (2.20) and estimated covariance matrix by (2.21). Under HG-design theoretical covariance matrix was calculated by (2.26) and estimated covariance matrix by (2.27).

For ratio estimators covariance matrix and its estimator were calculated by (2.16) and (2.18) using Remark 2.2 and Corollary 2.4. Ratio of known and estimated domain size, $N_d/\hat{N}_d$, was used.

In the following tables linear estimator is denoted by L, ratio estimator by R, continuous variable by C, binary variable by B and population total by Pop.

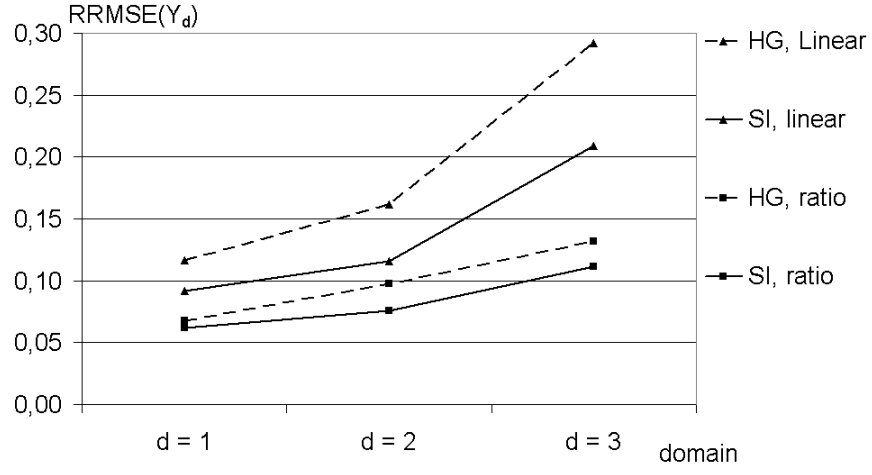

Figure 4.3: Relative root mean square error of initial estimators of continuous variable

Table 4.6 confirms that initial estimators are unbiased, though theoretically, the ratio estimator is only asymptotically unbiased. The relative bias is less than 0.2% for continuous and less than 0.5% for binary variable. The relative root mean square error is much lower for ratio estimator of the continuous variable (Figure 4.3) compared to the linear estimator of the same design. The same pattern, though not so strong, can be seen for binary variable (Table 4.6). Table 4.7 includes variances of initial domain estimators. Column *RD* shows that variance formulae of initial estimators are correctly derived and in spite of asymptotic nature of V_{dd} for ratio estimators, it can be used as a

true variance in most of the cases. Only for the smallest domain and the HG-design was the relative difference as big as 10%. We know that the estimated covariance matrix of initial estimators is unbiased or asymptotically unbiased (for ratio estimator). Comparison of V_{dd} and $\hat{\hat{V}}_{dd}$ shows that this holds, this is expressed also in the column *RB*. According to relative root mean square error the estimated variances of domain estimators are quite stable especially for linear estimators of continuous variable (*RRMSE* in Table 4.7).

Table 4.6: Simulation results for initial estimators

Design	Estim.	Domain	$\hat{\bar{Y}}_d^{(m)}$	$min[\hat{Y}_d^{(m)}]$	$max[\hat{Y}_d^{(m)}]$	RB	$RRMSE$
			Continuous variable				
SI	L	1	4998.54	3152.17	6841.82	0.000	0.092
		2	4613.53	2638.68	6890.35	0.000	0.116
		3	1399.13	466.87	2504.03	0.002	0.209
	R	1	5001.51	4101.01	6429.51	0.001	0.062
		2	4614.48	3453.84	6179.35	0.000	0.075
		3	1394.98	910.08	2210.81	-0.001	0.112
HG	L	1	5007.05	3106.40	7248.11	0.002	0.117
		2	4625.77	2238.04	7767.64	0.002	0.162
		3	1390.89	204.29	3206.52	-0.004	0.292
	R	1	5008.46	3858.43	6449.14	0.002	0.068
		2	4618.10	3257.52	7063.12	0.001	0.097
		3	1395.36	639.63	2354.07	-0.001	0.132
			Binary variable				
SI	L	1	129.57	20.00	270.00	0.004	0.255
		2	209.65	60.00	380.00	0.003	0.197
		3	36.15	0.00	120.00	0.004	0.497
	R	1	129.65	21.01	286.59	0.005	0.247
		2	209.70	73.30	354.28	0.003	0.177
		3	36.04	0.00	109.12	0.001	0.471
HG	L	1	129.48	9.90	304.95	0.004	0.316
		2	209.74	36.30	511.55	0.004	0.265
		3	35.88	0.00	155.12	-0.003	0.639
	R	1	129.50	13.41	302.30	0.004	0.303
		2	209.36	41.99	451.70	0.002	0.232
		3	35.93	0.00	140.53	-0.002	0.602

Table 4.7: Simulation results for variances of initial estimators

Design	Estim.	Domain	V_{dd}^{Emp}	V_{dd}	RD	$\hat{\hat{V}}_{dd}$	RB	$RRMSE$
Continuous variable								
			$\times 10^3$	$\times 10^3$	$\times 1$	$\times 10^3$	$\times 1$	$\times 1$
SI	L	1	209.3	201.8	-0.036	201.6	-0.001	0.271
		2	287.1	288.1	0.004	287.5	-0.002	0.258
		3	85.2	86.0	0.009	86.1	0.001	0.338
	R	1	95.9	93.5	-0.025	92.7	-0.008	0.472
		2	121.2	122.4	0.010	120.9	-0.012	0.446
		3	24.3	24.0	-0.011	23.3	-0.030	0.620
HG	L	1	341.0	338.3	-0.008	340.1	0.005	0.223
		2	555.8	542.2	-0.024	546.0	0.007	0.384
		3	166.3	160.2	-0.037	159.4	-0.005	0.372
	R	1	114.5	112.0	-0.022	112.0	0.000	0.455
		2	202.0	196.0	-0.030	195.8	-0.001	0.783
		3	34.0	30.6	-0.100	28.3	-0.077	0.654
Binary variable								
			$\times 1$	$\times 1$	$\times 1$	$\times 1$	$\times 1$	$\times 1$
SI	L	1	1079.2	1086.7	0.007	1091.2	0.004	0.236
		2	1691.9	1685.3	-0.004	1689.9	0.003	0.173
		3	320.1	318.3	-0.006	319.6	0.004	0.486
	R	1	1011.8	1014.5	0.003	1013.9	-0.001	0.220
		2	1364.5	1345.3	-0.014	1337.2	-0.006	0.140
		3	287.2	277.1	-0.035	269.4	-0.028	0.431
HG	L	1	1660.8	1666.0	0.003	1671.3	0.003	0.353
		2	3061.5	2939.1	-0.040	2949.2	0.003	0.263
		3	528.5	522.4	-0.012	519.7	-0.005	0.725
	R	1	1526.0	1515.3	-0.007	1505.7	-0.006	0.335
		2	2348.2	2229.0	-0.051	2198.1	-0.014	0.226
		3	469.7	436.3	-0.071	404.9	-0.072	0.672

4.2.3 GR-estimator when population total is known

Simulations with known population total include both the linear and the ratio estimator as initial estimators under SI- and HG-designs. GR-estimators GR(1) and GR(2) were calculated, the first one using the theoretical covariance matrix of initial estimators, and the second one, the covariance matrix estimated on each sample. Theoretical covariance matrix of GR-estimators was calculated by (3.30) and (3.31). Table 4.8 includes the values of the first three simulations to show the change of domain estimators from the initial estimators to the GR-estimators. Both GR-estimators guarantee the equality of the sum of domain estimators and the known population total.

Table 4.8: The values of estimators of selected samples (SI-design, continuous variable, linear estimator)

Sample	Estimator	Domain d			$\sum_{\mathfrak{D}}$
		d=1	d=2	d=3	
1	Initial $\hat{Y}_d$	5 089.6	5 759.7	1 501.7	12 351.0
	GR(1) $\hat{Y}_d^{GR}$	4 729.0	4 918.1	1 363.0	11 010.1
	GR(2) $\hat{\hat{Y}}_d^{GR}$	4 747.9	4 844.5	1 417.7	11 010.1
2	Initial $\hat{Y}_d$	4 828.9	4 146.3	1 339.3	10 314.5
	GR(1) $\hat{Y}_d^{GR}$	5 015.9	4 582.9	1 411.3	11 010.1
	GR(2) $\hat{\hat{Y}}_d^{GR}$	4 979.2	4 565.4	1 465.5	11 010.1
3	Initial $\hat{Y}_d$	4 762.6	5 103.2	1 385.6	11 251.4
	GR(1) $\hat{Y}_d^{GR}$	4 697.8	4 951.7	1 360.6	11 010.1
	GR(2) $\hat{\hat{Y}}_d^{GR}$	4 757.1	4 875.9	1 377.1	11 010.1

Table 4.9 confirms that the GR-estimator is unbiased for initial linear estimator, and already for our modest sample size, it is unbiased for initial ratio estimator too. Using estimated covariance matrix of initial estimators in the GR-estimator, makes its expression probabilistically complex, and it is not so easy to find the theoretical bias in this case. However, we mentioned earlier that the bias vanishes asymptotically. Table 4.10 confirms that the bias is negligible also in this more complex case.

The variability of the GR-estimator compared to the variability of the initial estimator decreases for all domains, the largest decrease being for the second domain (see Figure 4.4). The change of variability can be seen in more detail when comparing the minimal and maximal values of estimators in Tables 4.6, 4.9 and 4.10. According to the relative root mean square error in Tables 4.9 and 4.10 the variability of two different GR-estimators is practically the same;

the use of the estimated covariance matrix of initial estimators increases the variance only very slightly.

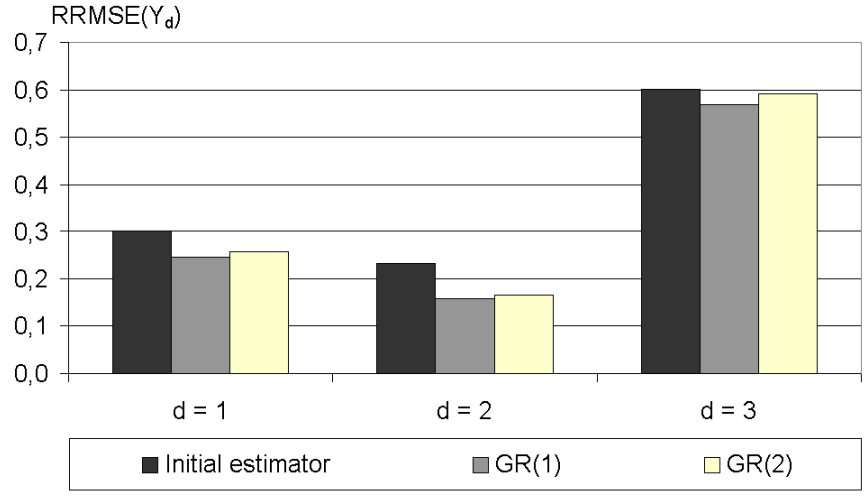

Figure 4.4: Relative root mean square error of initial ratio and GR-estimators (HG-design, binary variable)

Table 4.9: Simulation results of GR(1), the GR-estimator with known covariance matrix

Design	Initial	Domain	$\hat{\bar{Y}}_d^{(m)}$	$min[\hat{Y}_d^{(m)}]$	$max[\hat{Y}_d^{(m)}]$	RB	$RRMSE$
			Continuous variable				
SI	L	1	4998.24	3190.68	6746.67	0.000	0.087
		2	4612.84	3022.09	6401.37	0.000	0.095
		3	1399.02	442.52	2489.95	0.002	0.206
	R	1	5001.17	4104.12	6029.96	0.000	0.048
		2	4614.04	3598.44	5466.10	0.000	0.053
		3	1394.89	897.15	2119.84	-0.001	0.106
HG	L	1	5004.91	3044.05	7236.42	0.001	0.116
		2	4615.60	2524.02	6918.24	0.000	0.129
		3	1389.59	156.33	3126.27	-0.005	0.289
	R	1	5004.55	3838.50	6016.40	0.001	0.056
		2	4611.26	3549.50	5765.78	-0.001	0.063
		3	1394.29	664.91	2284.47	-0.001	0.126
			Binary variable				
SI	L	1	129.09	31.38	234.83	0.001	0.213
		2	208.89	110.18	314.29	-0.001	0.138
		3	36.02	-13.09	125.20	0.001	0.476
	R	1	129.11	43.14	229.83	0.001	0.193
		2	208.99	113.50	305.03	0.000	0.124
		3	35.89	-12.01	100.65	-0.003	0.444
HG	L	1	129.14	-2.52	270.81	0.001	0.276
		2	209.08	66.73	367.94	0.000	0.180
		3	35.78	-15.54	157.81	-0.006	0.617
	R	1	129.22	21.08	251.99	0.002	0.245
		2	208.93	76.35	323.76	0.000	0.158
		3	35.85	-16.90	127.89	-0.004	0.570

Table 4.10: Simulation results of GR(2), the GR-estimator with estimated covariance matrix

Design	Initial	Domain	$\hat{\bar{Y}}_d^{(m)}$	$min[\hat{Y}_d^{(m)}]$	$max[\hat{Y}_d^{(m)}]$	RB	$RRMSE$
			Continuous variable				
SI	L	1	4993.49	3145.21	6593.36	-0.001	0.088
		2	4612.52	2760.85	6320.36	-0.001	0.096
		3	1404.09	470.10	2824.11	0.006	0.209
	R	1	4997.81	4219.53	6029.14	0.000	0.049
		2	4614.33	3550.75	5518.04	0.000	0.054
		3	1397.96	916.83	2110.72	0.001	0.108
HG	L	1	5016.90	3076.71	7191.22	0.004	0.117
		2	4590.44	2321.24	6760.14	-0.005	0.131
		3	1402.76	197.41	3281.29	0.005	0.295
	R	1	5018.53	3947.25	6053.36	0.004	0.056
		2	4589.71	3360.06	5639.65	-0.005	0.064
		3	1401.86	640.17	2313.56	0.004	0.131
			Binary variable				
SI	L	1	129.10	28.77	238.94	0.001	0.215
		2	208.88	93.50	318.59	-0.001	0.139
		3	36.02	0.00	140.25	0.000	0.481
	R	1	128.53	32.97	228.67	-0.004	0.197
		2	209.70	106.85	303.76	0.003	0.127
		3	35.78	0.00	103.84	-0.006	0.453
HG	L	1	129.65	9.54	261.84	0.005	0.281
		2	208.51	58.19	329.07	-0.002	0.184
		3	35.84	0.00	206.81	-0.004	0.630
	R	1	128.47	14.45	257.80	-0.004	0.256
		2	209.90	68.05	321.38	0.004	0.166
		3	35.64	0.00	151.20	-0.010	0.591

Empirical variances of all the estimators are gathered together into the Table 4.11. The decrease of the variance compared to the one of the initial estimator is noticeable, especially for larger domains and for ratio initial estimator. Table 4.12 shows the correlations of the domain estimators. First of all we see that no effect on the correlations occurs whether to use the estimated covariance matrix of initial estimators or the theoretical ones. The empirical correlations of initial ratio estimators were close to zero, as was expected by the theoretical results. Correlations of linear initial domain estimators and all domain GR-estimators were negative. Correlations of the GR-estimator are stronger than correlations of the initial estimators, with an enormous increase of the negative dependence for ratio estimators.

Comparing the relative root mean square error of the estimated variances of initial and GR-estimators we can see slight increase in stability for GR-estimators (Figure 4.5). This is another positive feature of the GR-estimator. In addition to the smaller variability compared to the initial estimator, its variance estimator is also more stable.

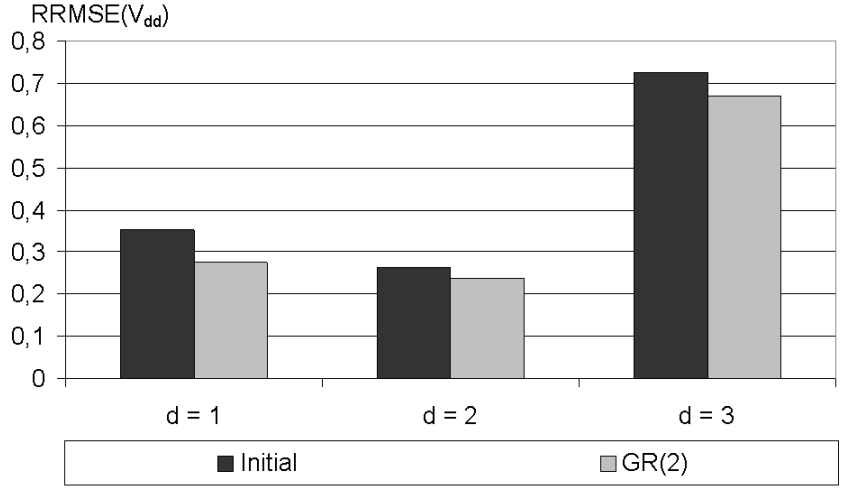

Figure 4.5: Relative root mean square error of estimated variances of the initial and GR (2), GR-estimator with estimated variance (HG-design, linear estimator, binary variable)

Table 4.11: Empirical variances of domain estimators: initial estimator, GR(1) and GR(2).

Design	Initial estimator	Domain	Estimator		
			Initial	GR(1)	GR(2)
Continuous variable, $V_{dd}^{Emp} \times 10^3$					
SI	L	1	209.3	190.8	193.6
		2	287.1	193.4	196.5
		3	85.2	82.5	84.8
	R	1	95.9	58.2	59.6
		2	121.2	60.2	62.0
		3	24.3	21.8	22.8
HG	L	1	341.0	333.4	343.3
		2	555.8	352.3	363.1
		3	166.3	162.6	170.1
	R	1	114.5	77.1	79.0
		2	202.0	85.3	86.3
		3	34.0	30.8	33.3
Binary variable, V_{dd}^{Emp}					
SI	L	1	1079.2	758.0	770.5
		2	1691.9	833.3	848.5
		3	320.1	293.7	300.0
	R	1	1011.8	621.0	647.3
		2	1364.5	669.7	701.3
		3	287.2	256.0	266.1
HG	L	1	1660.8	1269.1	1317.8
		2	3061.5	1422.5	1484.8
		3	528.5	493.3	515.0
	R	1	1526.0	1002.3	1089.3
		2	2348.2	1095.7	1206.3
		3	469.7	420.6	452.9

Table 4.12: Empirical correlations between domain estimators: initial estimator, GR(1) and GR(2).

Design	Initial estim.	Domain d	Domain g	Estimator Initial	Estimator GR(1)	Estimator GR(2)
Continuous variable, Cor_{dg}^{Emp}						
SI	L	1	2	-0.445	-0.785	-0.783
		2	3	-0.169	-0.337	-0.340
		1	3	-0.246	-0.318	-0.320
	R	1	2	-0.003	-0.816	-0.812
		2	3	0.006	-0.329	-0.335
		1	3	-0.006	-0.278	-0.277
HG	L	1	2	-0.510	-0.763	-0.759
		2	3	-0.208	-0.379	-0.382
		1	3	-0.279	-0.308	-0.311
	R	1	2	0.002	-0.811	-0.800
		2	3	-0.001	-0.380	-0.379
		1	3	-0.006	-0.233	-0.253
Binary variable, Cor_{dg}^{Emp}						
SI	L	1	2	-0.091	-0.816	-0.816
		2	3	-0.046	-0.373	-0.375
		1	3	-0.029	-0.231	-0.231
	R	1	2	0.003	-0.802	-0.803
		2	3	-0.008	-0.368	-0.370
		1	3	0.013	-0.260	-0.255
HG	L	1	2	-0.131	-0.818	-0.818
		2	3	-0.065	-0.386	-0.390
		1	3	-0.056	-0.215	-0.211
	R	1	2	-0.015	-0.800	-0.804
		2	3	-0.008	-0.378	-0.386
		1	3	-0.004	-0.252	-0.239

4.2.4 GR-estimator when population total is estimated from another survey

The initial ratio estimators of domain and population totals were considered. Estimator of the population total was assumed to be uncorrelated with domain estimators. Therefore, additional 10,000 samples with the same size were drawn where the population total was estimated.

Consequently, covariance matrix of initial estimators could be considered as a

diagonal matrix and results of Corollary 3.10 could be used for GR-estimator. Since, in theory, the domain ratio estimators are only approximately uncorrelated, the main issue here was to study, how will the use of diagonal covariance matrix influence the properties of GR-estimators, the bias and variability. It was also interesting to compare the GR-estimators with known and estimated population totals. The GR-estimators were calculated by (3.54) and (3.55), the first using a theoretical covariance matrix, and the second, a covariance matrix estimated on each sample. Theoretical covariance matrix of GR-estimators was calculated by (3.56) – (3.59).

Simulation results (Table 4.13) by domains are similar to the results presented in Section 4.2.3. According to the relative bias the obtained GR-estimators are unbiased. Decrease of the variance is smaller compared to the GR-estimator with known population total (Figure 4.6).

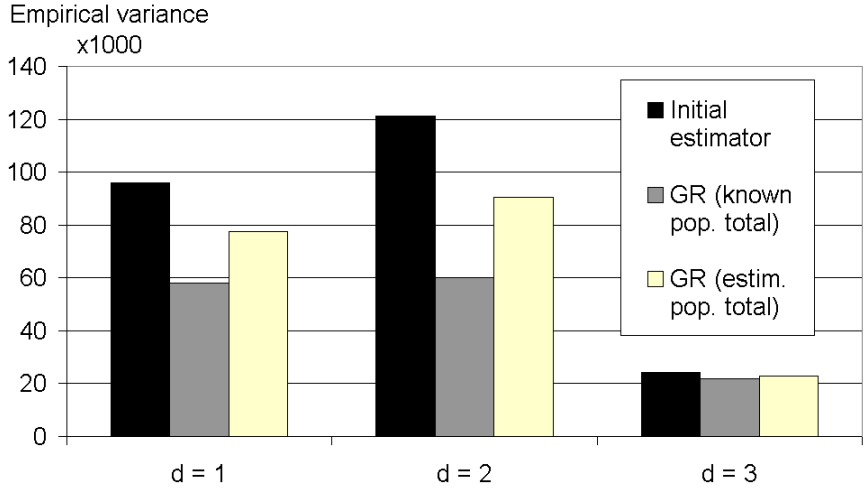

Figure 4.6: Variance of initial and GR-estimators assuming known and estimated population totals (HG-design, ratio estimator, continuous variable)

The theoretical variances differ from the empirical ones up to 5%, i.e. they describe well the true variances. The bias of the estimated covariance matrix is up to minus 5%.

Table 4.13: Simulation results of the GR-estimator

Variable	Design	Domain	$RB(\hat{Y}_d)$	$RRMSE(\hat{Y}_d)$	$RD(V_{dd})$	$RB(\hat{V}_{dd})$
Known covariance matrix of the initial estimator						
C	SI	Pop.	0.000	0.032	0.007	
		1	0.001	0.056	-0.026	
		2	0.000	0.065	0.013	
		3	-0.001	0.109	-0.007	
	HG	Pop.	0.001	0.039	-0.050	
		1	0.002	0.062	-0.025	
		2	0.001	0.083	-0.035	
		3	-0.001	0.129	-0.103	
B	SI	Pop.	0.002	0.099	-0.021	
		1	0.004	0.222	0.005	
		2	0.002	0.154	-0.025	
		3	0.000	0.457	-0.030	
	HG	Pop.	0.003	0.124	0.023	
		1	0.004	0.275	0.014	
		2	0.002	0.199	0.005	
		3	-0.001	0.587	-0.033	
Estimated covariance matrix of the initial estimator						
C	SI	Pop.	-0.004	0.032	-0.019	-0.046
		1	-0.003	0.055	-0.014	-0.048
		2	-0.005	0.065	0.013	-0.053
		3	-0.003	0.109	-0.010	-0.048
	HG	Pop.	0.001	0.039	-0.052	0.001
		1	0.003	0.063	-0.046	-0.044
		2	-0.001	0.082	-0.027	-0.034
		3	0.002	0.132	-0.138	-0.096
B	SI	Pop.	-0.005	0.100	-0.040	-0.009
		1	-0.007	0.223	-0.005	-0.010
		2	-0.003	0.155	-0.041	-0.011
		3	-0.010	0.459	-0.037	-0.036
	HG	Pop.	0.003	0.124	0.021	-0.033
		1	0.002	0.279	-0.016	-0.047
		2	0.005	0.203	-0.031	-0.072
		3	-0.004	0.598	-0.068	-0.126

4.2.5 Conditional GR-estimator

Initial estimators for conditional GR-estimator were ratio estimators under SI- and HG-designs. Population totals of both study variables and domain totals were estimated. The GR-estimates of domains were calculated so that the estimated population totals were kept fixed. This was done on each sample. The formulae of Section 3.5 were used. The main aim of the simulations here was to see the performance of the theoretical variance/covariance formulae of conditional GR-estimator. Under special interest was to display the increase of the variance due to the estimated population total. We recall that the estimated population total was fixed in the given sample but since it was estimated, it had a random nature which had to be taken into account in the variance formulae of domain GR-estimators.

For initial estimator the variance of the estimated population total was larger than the sum of variances of estimated domain totals (see Table 4.14). Consequently, the variance of GR-estimators is expected to increase compared to the initial estimator (Corollary 3.15). The results in Table 4.14 confirm this.

Table 4.14: Variances of initial and conditional GR-estimators

Design	Domain	Continuous variable, $V_{dd} \times 10^3$		Binary variable, V_{dd}	
		Initial	GR(1)	Initial	GR(1)
SI	1	93.5	94.6	1014.5	1029.5
	2	122.4	124.3	1345.3	1371.6
	3	24.0	24.1	277.1	278.2
	$\sum_{\mathfrak{D}}$	239.9		2637.0	
	Pop.	247.4		2737.9	
HG	1	112.0	113.7	1515.3	1543.0
	2	196.0	201.2	2229.0	2288.9
	3	30.6	30.8	436.3	438.5
	$\sum_{\mathfrak{D}}$	338.6		4180.5	
	Pop.	354.2		4391.4	

The variance of the conditional GR-estimator is composed by the variance of initial domain estimators and by variance of the initial population total estimator (3.68). In Figure 4.7 the lower part of the column of theoretical variance denotes the variance caused by initial domain estimators and the upper part of that column denotes the variance caused by estimated population total. We see that for larger domains these parts are of equal size. Thus, considering population total fixed in variance calculations leads to serious underestimation of the variance. Comparing the overall length of the columns of the empirical

and theoretical variances we see that our variance formula describes the true variance well.

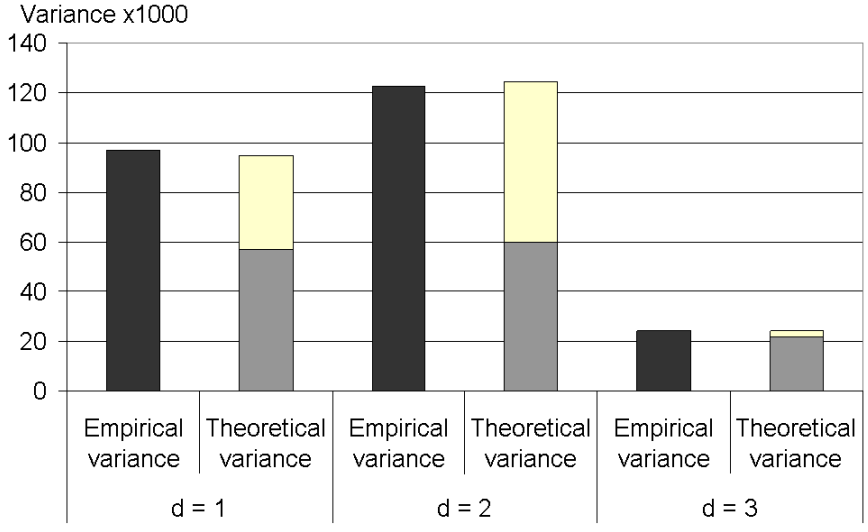

Figure 4.7: Empirical and theoretical variance of the conditional GR-estimator (initial ratio estimator, SI-design, continuous variable)

Table 4.15: Simulation results of the conditional GR-estimator

Variable	Design	Domain	$RB(\hat{Y}_d)$	$RRMSE(\hat{Y}_d)$	$RD(V_{dd})$
C	SI	1	0.001	0.062	-0.022
		2	0.000	0.076	0.015
		3	-0.001	0.111	-0.004
	HG	1	0.002	0.068	-0.024
		2	0.001	0.099	-0.029
		3	-0.001	0.131	-0.083
B	SI	1	0.005	0.248	0.004
		2	0.003	0.178	-0.007
		3	0.001	0.471	-0.031
	HG	1	0.005	0.305	-0.004
		2	0.003	0.235	-0.050
		3	-0.001	0.599	-0.056

4.2.6 Conclusions from simulations

The simulation study confirmed the theoretical results of this thesis. Even the asymptotic results worked well in our example with modest sample size.

The domain GR-estimators were considered in three cases: the population total was known, it was estimated and the conditional form of GR-estimator. In the first case, the effect of replacing known covariance matrix of initial estimators by the estimated one was also studied.

The main simulation results are summarized below.

- All domain GR-estimators were unbiased or the bias was so small that could not be detected from the simulation study. The issue of bias of GR-estimators was not theoretically clear when the estimated covariance matrix of initial estimators or the theoretical but approximate covariance matrix of ratio initial estimators was used in the GR-estimator.
- The variance of domain GR-estimators (except the conditional case) was smaller compared to the variance of initial estimators, expressing the optimality property. The largest relative decrease was for the domain with largest variance of the initial estimator.
- Using estimated covariance matrix of initial estimators instead of the theoretical one has increased the variance of GR-estimators, but only very little.
- The conditional GR-estimator for a domain was not more effective than the initial estimator for that domain. The increase in variance was quite small.
- The variance estimators of domain GR-estimators with known and estimated population total were unbiased or the bias was relatively small. The variance estimators were quite stable both for the initial and GR-estimators. The RRMSE of the estimated variance decreased slightly for GR-estimator.
- The dependence structure between domain estimators was illustrated for the case with known population total. The domain GR-estimators were stronger correlated than the initial domain estimators. The correlations were strong even when the initial estimators were uncorrelated. The correlations were negative. The use of estimated covariance matrix in GR-estimators did not affect the correlations.
- For the conditional domain GR-estimator where the estimated population total was kept fixed, its random nature has to be taken into account in the variance formula, otherwise the variance will be seriously underestimated.

Chapter 5

General conclusions

The main goals of this thesis were achieved. As a contribution of the author

- the domain GR-estimators satisfying known restrictions were derived; the restriction considered here was the summation restriction; three different cases were covered − the population total is known, it is estimated, it is estimated but conditionally fixed;
- the variance/covariance expressions of the GR-estimators were derived;
- it was shown that, excluding the conditional case, the GR-estimator is never less precise than the initial estimator; the domain GR-estimators were optimal among the linear estimators satisfying the same restrictions and using the same initial estimators;
- the important practical cases using linear and ratio initial estimators were elaborated in more detail;
- the results were expressed for two special sampling designs − the simple random sampling and the hypergeometric sampling design;
- the general variance/covariance expressions of domain linear and ratio estimators were derived, valid for both the WOR and WR designs; a remarkable result was the fact that domain ratio estimators are uncorrelated for some sampling designs;

– the theoretical results were tested and illustrated in a simulation study which confirmed good properties of the GR-estimators, even the asymptotic results worked well in our practical example with finite sample size;

A favorable feature of the derived GR-estimators is their practical applicability in domain estimation.

Bibliography

[1] Casella, G., Berger, R.L. (1990) *Statistical Inference.* Belmont: Duxbury Press

[2] Cochran W.G. (1977) *Sampling Techniques. Third Edition.* New York: Wiley

[3] Deming, W.E., Stephan, F.F. (1940) On a least square adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 11, 427−444

[4] Deville, J. C., Särndal, C.E. (1992) Calibration estimators in survey sampling, *Journal of the American Statistical Association*, 87, 376−382

[5] Van Duin, C., Snijders, V. (2003) *Simulation studies of Repeated weighting.* Voorburg/Heerlen: Statistics Netherlands

[6] EURAREA Consortium with K. Sõstra among the members (2004) *Enhancing Small Area Estimation Techniques to meet European Needs (EURAREA project) Final Reference Volume: Vol.*1−3
https://www.statistics.gov.uk/eurarea/default.asp

[7] European Commission (2005) *European Statistics Code of Practice.* http://epp.eurostat.ec.europa.eu/

[8] Hald, A. (1998). *A History of Mathematical Statistics.* New York: Wiley

[9] Hansen, M.H., Hurvitz, W.N., Madow, W.G. (1953) *Sample Survey Methods and Theory.* New York. Wiley

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

[11] Houbiers, M., Knottnerus, P., Kroese, A.H., Renssen, R.H., Snijders, V. (2003) *Estimating consistent table sets: position paper on repeated weighting.* Voorburg/Heerlen: Statistics Netherlands

[12] Johnson, N. L., Kotz, S., Balakrishnan, N. (1997) *Discrete Multivariate Distributions.* New York: Wiley

[13] Knottnerus P. (2003) *Sample Survey Theory. Some Pythagorean Perspectives.* New York: Springer

[14] Lehtonen, R., Pahkinen, E. (1995) *Practical Methods for Design and Analysis of Complex Surveys.* Chichester: Wiley

[15] Lehtonen, R., Särndal, C.-E., Veijanen, A. (2003) The effect of model choice in estimation for domains, including small domains. *Survey Methodology*, 29, 33−44

[16] Lehtonen, R., Särndal, C.-E., Veijanen, A. (2005) Does the model matter? Comparing model-assisted and model-dependent estimators of class frequencies for domains. *Statistics in Transition*, 7, 649−673

[17] Lepik, N., Sõstra, K., Traat, I. (2007) Conditional restriction estimator for domains. *The 8th Tartu Conference on Multivariate Statistics / The 6th Conference on Multivariate Distributions*
http://www.ms.ut.ee/tartu07/presentations/presentations.html

[18] Lütkepohl, H. (1996) *Handbook of Matrices.* New York: Wiley

[19] Meister, K. (2004) *On Methods for Real Time Sampling and Distributions in Sampling. Doctoral Dissertation.* Umea, 2004

[20] Montanari, C.E. (1987) Postsampling efficient prediction in large-scale surveys. *International Statistical Review*, 55, 191−202

[21] Rajaleid, K. (2004) Multivariate finite population inference under the assumption of linear pattern in the population. *Acta et Commentationes Universitatis Tartuensis de Mathematica*, 8, 235−242

[22] Rao, J. N. K. (2003) *Small Area Estimation.* New York: Wiley

[23] Sen, A. R. (1953). On the estimator of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics,* 5, 119-127

[24] Särndal C-E., Swensson, B., Wretman, J. (1989) The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76, 527-537

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

[26] Srivastava, M. S. (2002) *Methods of Multivariate Statistics.* New York: Wiley

[27] Tillé, Y. (2006) *Sampling Algorithms.* New-York: Springer-Verlag

[28] Traat I., Ilves M. (2007) The hypergeometric sampling design, theory and practice. *Acta Applicandae Mathematicae* vol. 97, pp. 311−321

[29] Traat, I., Bondesson, L., Meister, K. (2004) Sampling design and sample selection through distribution theory. *Journal of Statistical Planning and Inference*, vol. 123, 395−413

[30] Traat, I., Meister, K., Sõstra, K. (2001) Statistical inference in sampling theory. *Theory of Stochastic Processes*, vol. 7(23), 301−316

[31] Traat, I. (2000) Sampling design as a multivariate distribution. *New trends in Probability and Statistics 5, Multivariate Statistics.* Vilnius, Utrecht: VSP/TEV, 195−208

[32] Yates, F., Grundy, P. M. (1953) Selection without replacement from within strata with probability proportional to size, *Journal of the Royal Statistical Society*, Series B, 15, 253−261

Osakogumite kitsendustega hinnang

Kokkuvõte

Valikuuringute üheks oluliseks ülesandeks on osakogumite hindamine. Kasvanud on nõudmine usaldusväärsete ja kooskõlaliste hinnangute järele, mis omakorda on andnud tõuke vastava teooria arengule. Senistes teoreetilistes käsitlustes on suurt tähelepanu pööratud hinnangute täpsuse tõstmisele, seda eriti väikese valimimahuga osakogumite korral. Sellele teemale on pühendatud raamatuid (Rao, 2003), kirjutatud teadusartikleid (Lehtonen jt, 2003, 2005) ja läbi viidud suuri uurimisprojekte (EURAREA konsortsium, 2005).

Käesoleva dissertatsiooni põhiteemaks on teine tähtis probleem osakogumite hindamise valdkonnas, see on hinnangute kooskõlalisuse ehk ühilduvuse probleem. Osakogumite ja üldkogumi parameetrid on sageli omavahel seotud. Näiteks peavad osakogumite kogusummad summeeruma üldkogumi kogusummaks. Hinnangute jaoks pole need seosed aga sageli täidetud, st hinnangud pole ühilduvad. Statistika tarbijad ei aktsepteeri mitteühilduvaid hinnanguid. Statistiliste andmete ühilduvus on samuti oluline Euroopa statistikasüsteemi kvaliteedikomponent (Euroopa Komisjon, 2005, põhimõte 14). Mitteühilduvuse põhjusteks on hinnangute juhuslikkus, mitteaditiivsus, erinevate hinnangumeetodite kasutamine, hinnangute pärinemine erinevatest uuringutest jm. Kirjeldatud probleem tekib nii suurte kui ka väikeste osakogumite korral. Tavaliselt kasutatakse lihtsaid ad hoc meetodeid ühilduvuse probleemi lahendamiseks. Sageli pole teada selliselt saadud hinnangute dispersiooni valem. Kui soovitakse saavutada ühilduvus erinevatel tasemetel (väikesed osakogumid, suuremad osakogumid, üldkogum), siis muutub probleem matemaatiliselt keeruliseks.

Dissertatsiooni põhieesmärgid on:

– tuletada osakogumite hinnangud, mis rahuldaksid teadaolevaid kitsendusi mõnede lihtsate, kuid praktikas oluliste olukordade jaoks ja oleksid teatud hinnangute klassi jaoks optimaalsed;

– tuletada dispersiooni ja kovariatsiooni valemid hinnangute jaoks;

– testida ja illustreerida teoreetilisi tulemusi simuleerimiseksperimentidega.

Põhieesmärkide lahendamiseks kasutatakse raamatus Knottnerus (2003) välja pakutud üldist kitsendustega hinnangut (General Restriction estimator), lühidalt GR-hinnang. Dissertatsioonis rakendatakse Knottneruse üldisi ideid osakogumite hindamise ülesandele. Tuletatakse osakogumite jaoks GR-hinnangud ja nende dispersioonide ning kovariatsioonide avaldised. Tuletatud hinnangutele kandub üle GR-hinnangu optimaalsuse omadus. Seega on saadud osakogumite hinnangud dispersiooni mõttes parimad võimalikud kõigi osakogumite hinnangute hulgas, mis baseeruvad samadel alghinnangutel ja rahuldavad samu kitsendusi. Töös vaadeldakse summeeruvuskitsendust, st et osakogumite kogusummad peavad summeeruma üldkogumi kogusummaks. Valemid tuletatakse kolme juhu jaoks: üldkogumi kogusumma on teada, on hinnatud, on hinnatud aga tinglikult fikseeritud.

Käesolevas töös kasutatakse disainipõhist lähenemist, kus hinnangute omadused (keskväärtus, dispersioon, kovariatsioon) on määratud valikudisainiga. Vaadeldud on kahte valikudisaini: lihtne juhuslik valik (SI) ja hüpergeomeetriline valik (HG). Nendest SI-valik on võrdsete kaasamistõenäosustega tagasipanekuta valikudisain ja HG-valik on ebavõrdsete kaasamistõenäosustega tagasipanekuga valikudisain. Mõlemad valikudisainid on laialdaselt kasutuses valikuuringutes. Avaldiste esitamisel on kasutatud valikuvektori meetodit (Traat, 2000), mis võimaldab samaaegselt käsitleda tagasipanekuta ja tagasipanekuga valikudisaine. Seega on paljud töö tulemused üldisemad kui seni kirjanduses käsitletud, nad kehtivad mõlema disainitüübi jaoks.

Antud töös eeldatakse, et osakogumite valimimahud pole liiga väikesed, st spetsiaalseid väikese osakogumi hinnanguid pole siin vaadeldud. GR-hinnang vajab esialgseid hinnanguid oma konstruktsioonis. On vaadeldud kahte esialgset hinnangut osakogumite jaoks: lineaarset ja suhtehinnangut. Lineaarne hinnang on kirjanduses tuntud ka Horvitz-Thompsoni nime all (seda tagasipanekuta disainide korral). Mõlemad hinnangud on otsesed osakogumi hinnangud selles mõttes, et kasutavad ainult neid uuritava tunnuse väärtusi, mis on mõõdetud vaadeldavas osakogumis. Suhtehinnangus kasutatakse lisaks osakogumi kohta teadaolevat abitunnuse kogusummat.

Esimene peatükk annab ülevaate käesoleva töö jaoks vajalikest tulemustest. Valikuvektori jaotuse abil defineeritakse valikudisain. Erijuhtudena kirjeldatakse kolme valikudisaini: SI-valik, HG-valik ja multinomiaalne valik. Neid kasutatakse töös tulemuste rakendamisel. Defineeritakse esialgsed hinnangud, st. lineaarne ja suhtehinnang. Tuuakse nende dispersioonide ja kovariatsioonide üldised valemid, kusjuures suhtehinnangu korral on need Taylori reaksarendusele baseeruvad ligikaudsed avaldised. Üldistest valemitest on tuletatud dispersiooni ja kovariatsiooni valemid SI- ja HG-valikute jaoks. HG-valiku kovariatsiooni valem on uudne tulemus.

Teises peatükis käsitletakse osakogumeid ja osakogumi kogusumma hinda- mist, kasutades lineaarset ja suhtehinnangut. Antud töö seisukohalt on oluline teada nende hinnangute kovariatsioonimaatriksit. Seepärast tuletatakse üldised dispersiooni ja kovariatsioonivalemid osakogumi hinnangute jaoks ja ka valemid SI- ja HG-valikudisainide korral. Osakogumite hinnangute kovariatsiooni pole siiani kirjanduses väga põhjalikult käsitletud. Suhtehinnangu ja tagasipanekuta disainide korral on kovariatsiooniavaldis toodud Särndal (1992), teatud erijuhtudel on kovariatsioonimaatriksi hinnangut vaadeldud raamatus Lehtonen (1995). Antud töös saadud valemid võimaldavad teha mitmeid huvitavaid järeldusi osakogumite hinnangute sõltuvuse kohta. Selgus, et teatud valikudisainide ja suhtehinnangu korral on osakogumite hinnangud mittekorreleeritud. Samade valikudisainide jaoks võrdub osakogumi ja üldkogumi suhtehinnangute kovariatsioon osakogumi hinnangu dispersiooniga. Peatükis on esitatud ka kaks näidet väikese üldkogumi ja kahe osakogumi korral. Esitatud tulemuste uudsus seisneb nende üldisemas iseloomus. Need on rakendatavad nii tagasipanekuga kui ka tagasipanekuta valikudisainide korral.

Kolmandas peatükis teoreemides 3.2-3.4 ja arvukates järeldustes on esitatud väitekirja põhitulemused, mis on autoripoolne panus antud valdkonna teoreetilisse arengusse. Esiteks on antud ülevaade kitsendustega hinnangust ja selle põhiomadustest. Antud metoodika annab võimaluse hinnangute arvutamiseks, kui parameetrite vahel peavad kehtima etteantud seosed. Töös on vaadatud summeeruvuskitsendust, st. et osakogumite kogusummad peavad summeeruma üldkogumi kogusummaks. Tõestatakse teoreemid üldkogumi ja osakogumite GR-hinnangute kujust ja vastavatest dispersioonidest ja kovariatsioonidest. Tulemused on esitatud kolme tähtsa juhu jaoks: teadaolev üldkogumi kogusumma, samast või mõnest teisest uuringust hinnatud üldkogumi kogusumma, hinnatud ja tinglikult fikseeritud üldkogumi kogusumma. Esialgsete hinnangutena on eeldatud lineaarset ja suhtehinnangut. Tuletatud on ka valemid SI- ja HG-valikudisainide jaoks. Tuletatud GR-hin- nangutel on oluline omadus, nad on minimaalse dispersiooniga kõigi teiste samu kitsendusi rahuldavate ja samu esialgseid hinnanguid kasutavate hinnangute hul-

gas. Osutus, et GR-hinnangu dispersioon (v.a. tingliku GR-hinnangu juht) ei ole suurem esialgse hinnangu dispersioonist. Seega üldjuhul on GR-hinnangul mitu head omadust: rahuldab kitsendusi ja on täpsem kui esialgne hinnang. Töös arendatakse diskussiooni, kuidas GR-hinnangu analüütilise kuju uurimine võimaldab konstrueerida teisi lihtsamaid (pole vaja teada alghinnangu kovariatsioonimaatriksit), kuid siiski optimaalsele lähedasi hinnanguid.

Neljandas peatükis esitatakse simuleerimiseksperimendi tulemusi. Eesmärgiks oli uurida kitsendustega hinnangu käitumist praktikas ja seda, kuidas asümptootilised tulemused töötavad lõpliku valimimahu korral. Selleks moodustati 2000 isikust üldkogum ja võeti 10000 valimit suurusega ligikaudu 200 isikut, kasutades SI- ja HG-valikudisaine. Eksperimentides vaadeldakse kitsendustega hinnangut kolmel erijuhul: teadaolev üldkogumi kogusumma, hinnatud üldkogumi kogusumma ja tinglik kitsendustega hinnang. Simuleerimiseksperiment kinnitas töös saadud teoreetilisi tulemusi. Isegi asümptootilised valemid töötasid hoolimata küllaltki väikesest valimimahust hästi. Eksperiment illustreeris järgmisi aspekte:

- GR-hinnangud olid nihketa või oli nihe väga väike.
- GR-hinnangu dispersioon oli väiksem esialgse hinnangu dispersioonist. Suurim suhteline vähenemine oli osakogumis, mille esialgse hinnangu dispersioon oli suurim.
- Hinnatud kovariatsioonimaatriksi kasutamine GR-hinnangus suurendas vähesel määral hinnangu dispersiooni.
- Tinglik GR-hinnang oli vähem efektiivne kui sama osakogumi esialgne hinnang.
- Osakogumite GR-hinnangute korrelatsioon oli tugevam kui vastavatel esialgsetel hinnangutel. Tugev sõltuvus ilmnes isegi juhul, kui esialgsed hinnangud ei olnud korreleeritud. GR-hinnangute korrelatsioon oli negatiivne.
- Tingliku GR-hinnangu korral, kui üldkogumi kogusumma hinnang fikseeritakse, tuleb selle juhuslikkust kindlasti arvestada osakogumi hinnangu dispersiooni valemis, et mitte alahinnata dispersiooni.

Lõpetuseks võib öelda, et töös tuletatud summeeruvuskitsendust rahuldavad osakogumite hinnangud on uudsed. Neil on mitmeid häid omadusi, nad on ühilduvad ja üldiselt täpsemad kui esialgsed hinnangud. Nad on praktikas rakendatavad.

Curriculum Vitae

Kaja Sõstra

Citizenship: Estonian Republic
Born: April, 20, 1965, Viljandi Estonia
Marital Status: married, two daughters
Address: Kolde 67-48, Tallinn 10321, Estonia
Contacts: e-mail: kaja.sostra@stat.ee

Education

1983-1988 Faculty of Automatics, Tallinn University of Technology, diploma system engineer
1999-2005 Faculty of Mathematics and Computer Science, University of Tartu, PhD studies in Mathematical Statistics

Professional employment

1988-1990 Accountant, Viljandi Collective Farm
1990-1994 Chief Accountant, Nõges Ltd
1995-1999 Specialist, Statistical Bureau of Central Estonia
2000-2001 Specialist, Statistics Estonia
2002-2003 Researcher, Statistics Finland
Since 2004 Head of Methodology Department, Statistics Estonia

Curriculum Vitae

Kaja Sõstra

Kodakondsus: Eesti Vabariik
Sünniaeg ja -koht: 20. aprill, 1965, Viljandi, Eesti
Perekonnaseis: abielus, kaks tütart
Aadress: Kolde pst 67-48, Tallinn 10321, Eesti
Kontaktandmed: e-mail: kaja.sostra@stat.ee

Hariduskäik

1983-1988 Tallinna Tehnikaülikooli automaatikateaduskond, diplomeeritud süsteemiinsener
1999-2005 Tartu Ülikooli matemaatika-informaatikateaduskond, doktoriõpingud matemaatilise statistika erialal

Erialane teenistuskäik

1988-1990 raamatupidaja, Viljandi Kolhoos
1990-1994 pearaamatupidaja, AS Nõges
1995-1999 spetsialist, Kesk-Eesti Statistikabüroo
2000-2001 spetsialist, Statistikaamet
2002-2003 spetsialist, Soome Statistikakeskus
Alates 2004 metoodika osakonna juhataja, Statistikaamet

DISSERTATIONES MATHEMATICAE UNIVERSITATIS TARTUENSIS

1. **Mati Heinloo.** The design of nonhomogeneous spherical vessels, cylindrical tubes and circular discs. Tartu, 1991, 23 p.
2. **Boris Komrakov.** Primitive actions and the Sophus Lie problem. Tartu, 1991, 14 p.
3. **Jaak Heinloo.** Phenomenological (continuum) theory of turbulence. Tartu, 1992, 47 p.
4. **Ants Tauts.** Infinite formulae in intuitionistic logic of higher order. Tartu, 1992, 15 p.
5. **Tarmo Soomere.** Kinetic theory of Rossby waves. Tartu, 1992, 32 p.
6. **Jüri Majak.** Optimization of plastic axisymmetric plates and shells in the case of Von Mises yield condition. Tartu, 1992, 32 p.
7. **Ants Aasma.** Matrix transformations of summability and absolute summability fields of matrix methods. Tartu, 1993, 32 p.
8. **Helle Hein.** Optimization of plastic axisymmetric plates and shells with piece-wise constant thickness. Tartu, 1993, 28 p.
9. **Toomas Kiho.** Study of optimality of iterated Lavrentiev method and its generalizations. Tartu, 1994, 23 p.
10. **Arne Kokk.** Joint spectral theory and extension of non-trivial multiplicative linear functionals. Tartu, 1995, 165 p.
11. **Toomas Lepikult.** Automated calculation of dynamically loaded rigidplastic structures. Tartu, 1995, 93 p. (in Russian).
12. **Sander Hannus.** Parametrical optimization of the plastic cylindrical shells by taking into account geometrical and physical nonlinearities. Tartu, 1995, 74 p.
13. **Sergei Tupailo.** Hilbert's epsilon-symbol in predicative subsystems of analysis. Tartu, 1996, 134 p.
14. **Enno Saks.** Analysis and optimization of elastic-plastic shafts in torsion. Tartu, 1996, 96 p.
15. **Valdis Laan.** Pullbacks and flatness properties of acts. Tartu, 1999, 90 p.
16. **Märt Põldvere.** Subspaces of Banach spaces having Phelps' uniqueness property. Tartu, 1999, 74 p.
17. **Jelena Ausekle.** Compactness of operators in Lorentz and Orlicz sequence spaces. Tartu, 1999, 72 p.
18. **Krista Fischer.** Structural mean models for analyzing the effect of compliance in clinical trials. Tartu, 1999, 124 p.

19. **Helger Lipmaa.** Secure and efficient time-stamping systems. Tartu, 1999, 56 p.
20. **Jüri Lember.** Consistency of empirical k-centres. Tartu, 1999, 148 p.
21. **Ella Puman.** Optimization of plastic conical shells. Tartu, 2000, 102 p.
22. **Kaili Müürisep.** Eesti keele arvutigrammatika: süntaks. Tartu, 2000, 107 lk.
23. **Varmo Vene.** Categorical programming with inductive and coinductive types. Tartu, 2000, 116 p.
24. **Olga Sokratova.** Ω-rings, their flat and projective acts with some applications. Tartu, 2000, 120 p.
25. **Maria Zeltser.** Investigation of double sequence spaces by soft and hard analitical methods. Tartu, 2001, 154 p.
26. **Ernst Tungel.** Optimization of plastic spherical shells. Tartu, 2001, 90 p.
27. **Tiina Puolakainen.** Eesti keele arvutigrammatika: morfoloogiline ühestamine. Tartu, 2001, 138 p.
28. **Rainis Haller.** *M(r,s)*-inequalities. Tartu, 2002, 78 p.
29. **Jan Villemson.** Size-efficient interval time stamps. Tartu, 2002, 82 p.
30. **Eno Tõnisson.** Solving of expession manipulation exercises in computer algebra systems. Tartu, 2002, 92 p.
31. **Mart Abel.** Structure of Gelfand-Mazur algebras. Tartu, 2003. 94 p.
32. **Vladimir Kuchmei.** Affine completeness of some ockham algebras. Tartu, 2003. 100 p.
33. **Olga Dunajeva.** Asymptotic matrix methods in statistical inference problems. Tartu 2003. 78 p.
34. **Mare Tarang.** Stability of the spline collocation method for volterra integro-differential equations. Tartu 2004. 90 p.
35. **Tatjana Nahtman.** Permutation invariance and reparameterizations in linear models. Tartu 2004. 91 p.
36. **Märt Möls.** Linear mixed models with equivalent predictors. Tartu 2004. 70 p.
37. **Kristiina Hakk.** Approximation methods for weakly singular integral equations with discontinuous coefficients. Tartu 2004, 137 p.
38. **Meelis Käärik.** Fitting sets to probability distributions. Tartu 2005, 90 p.
39. **Inga Parts.** Piecewise polynomial collocation methods for solving weakly singular integro-differential equations. Tartu 2005, 140 p.
40. **Natalia Saealle.** Convergence and summability with speed of functional series. Tartu 2005, 91 p.
41. **Tanel Kaart.** The reliability of linear mixed models in genetic studies. Tartu 2006, 124 p.

42. **Kadre Torn.** Shear and bending response of inelastic structures to dynamic load. Tartu 2006, 142 p.
43. **Kristel Mikkor.** Uniform factorisation for compact subsets of Banach spaces of operators. Tartu 2006, 72 p.
44. **Darja Saveljeva.** Quadratic and cubic spline collocation for Volterra integral equations. Tartu 2006, 117 p.
45. **Kristo Heero.** Path planning and learning strategies for mobile robots in dynamic partially unknown environments. Tartu 2006, 123 p.
46. **Annely Mürk.** Optimization of inelastic plates with cracks. Tartu 2006. 137 p.
47. **Annemai Raidjõe.** Sequence spaces defined by modulus functions and superposition operators. Tartu 2006, 97 p.
48. **Olga Panova.** Real Gelfand-Mazur algebras. Tartu 2006, 82 p.
49. **Härmel Nestra.** Iteratively defined transfinite trace semantics and program slicing with respect to them. Tartu 2006, 116 p.
50. **Margus Pihlak.** Approximation of multivariate distribution functions. Tartu 2007, 82 p.
51. **Ene Käärik.** Handling dropouts in repeated measurements using copulas. Tartu 2007, 99 p.
52. **Artur Sepp.** Affine models in mathematical finance: an analytical approach. Tartu 2007, 147 p.
53. **Marina Issakova.** Solving of linear equations, linear inequalities and systems of linear equations in interactive learning environment. Tartu 2007, 170 p.