WORKSHOP ON SURVEY SAMPLING THEORY AND METHODOLOGY

August 24–28, 2006
Ventspils, Latvia

Central Statistical Bureau of Latvia

Workshop on Survey Sampling Theory and Methodology, August 24–28, 2006, Ventspils

ORGANISERS

The Institute of Mathematics and Informatics, Lithuania
The Vilnius University, Lithuania
The University of Latvia
Ventspils University College
The University of Tartu, Estonia
The University of Umeå , Sweden
The Central Statistical Bureau of Latvia
The Statistics Lithuania
The Statistical Office of Estonia
Datorzinibu Centrs

ORGANIZING COMMITTEE

Daniel Thorburn, Sweden
Gunnar Kulldorff, Sweden
Imbi Traat, Estonia
Risto Lehtonen, Finnland
Danute Krapavickaite, Lithuania
Aleksandras Plikusas, Lithuania
Signe Balina, Latvia
Janis Lapins, Latvia

SPONSORS

Baltic-Nordic Conference on Survey Sampling
Stockholm University
The Central Statistical Bureau of Latvia
The Statistics Lithuania
Ventspils University College

PREFACE

Dear participants and readers of this book,

The book consists of the lectures and contributed papers presented at the Workshop on Survey Sampling Theory and Methodology, Ventspils, 2006. The workshop is already 10^{th} in a series of yearly Baltic-Nordic meetings on survey sampling within the co-operation program between Sweden, Finland and three Baltic countries that started in the first half of the 90s.

On behalf of the Organizing Committee, we wish you fruitful work at workshop and pleasant stay in Ventspils.

Signe Bāliņa

Jānis Lapiņš

Mārtiņš Liberts

CONTENTS

PROGRAMME

August 23, Wednesday

Arrival of participants

August 24, Thursday

Morning Session

Chair: Gunnar Kulldorff

9.00-9.10	Opening of the Workshop
9.15-10.00	**Danutė Krapavickaitė, Vilma Nekrasaite** (lecture): Some model-based estimator
10.10-10.55	**Daniel Thorburn** (lecture): Optimal inclusion probabilities and estimators when sampling with varying probabilities
11.25-11.50	**Martins Liberts**: Variance estimation in EU-SILC survey *Discussant: Vilma Nekrasaite*
12.00-12.25	**Karolin Toompere**: Strength of Auxiliary Information for Compensating Nonresponse *Discussant: Natalia Budkina*

Afternoon session

Chair: Danutė Krapavickaitė

14.00-14.45	**Imbi Traat** (lecture): Variance of quantile estimators in household surveys
15.00-15.25	**Anton Grafström**: Comparisons of methods for generating conditional Poisson samples *Discussant: Virgi Puusepp*
15.55-16.20	**Signe Bāliņa**: Usage of administrative data in EU-SILC survey *Discussant: Inga Masiulaityte*
16.35-17.00	**Genovaite Saluckiene**: Estimation of the number of non-official emigrants from the Labour Force Survey *Discussant: Janis Lapins*
19.00-20.30	***Welcome party***

August 25, Friday

Morning Session

Chair: Signe Bāliņa

9.00-9.45 **Risto Lehtonen** (lecture): The role of models in model-assisted and model-dependent estimation for domains and small areas, part I

10.00-10.45 **Risto Lehtonen** (lecture): The role of models in model-assisted and model-dependent estimation for domains and small areas, part II

11.15-11.40 **Pasi Piela**: Introduction to Emerging Methods for Imputation in Official Statistics
Discussant: Anton Grafström

11.55-12.20 **Andris Fisenko, Vita Kozirkova**: Imputation in EU-SILC survey
Discussant: Olga Vasylyk

Afternoon session

Chair: Risto Lehtonen

14.00-14.25 **Olga Vasylyk, Oksana Honchar**: Statistical analysis of a sample of small-scale enterprises
Discussant: Jelena Novika

14.55-15.20 **Inga Masiulaityte**: Imputed rent in Household Budget Survey
Discussant: Seppo Laaksonen

15.35-16.00 **Oksana Honchar**: Detection and Considering of Extremal Elements for Business Surveys
Discussant: Pasi Piela

17.00-18.00 ***City sightseeing***

August 26, Saturday

9.00-19.00 ***Excursion***

August 27, Sunday

Morning Session

Chair: Daniel Thorburn

9.00-9.45 **Seppo Laaksonen** (lecture): Sampling Design of Multi-National Surveys with Applications to the European Social Survey (ESS)

10.00-10.45 **Aleksandras Plikusas** (lecture): Nonlinear calibration

11.15-11.40 **Nataliya Skachek**: EU Farm Accountancy Data Network (EU FADN) and Ukraine
Discussant: Gunnar Kulldorff

11.55-12.20 **Olga Grakoviča**: Analysis of GREG Estimator in Farm Survey
Discussant: Viktoras Chadysas

Afternoon session

Chair: Pauli Ollila

14.00-14.45 **Martins Liberts**: Training on R-language

15.15-17.00 Individual work of participants with R program and sampling package

August 28, Monday

Morning Session

Chair: Imbi Traat

9.00-9.45	**Lennart Bondesson** (lecture): Is there a best πps sampling design?
10.00-10.45	**Pauli Ollila** (lecture): Variance estimation with imputed data
11.15-11.40	**Dalius Pumputis, Aleksandras Plikusas**: Calibrated estimators of finite population covariance *Discussant: Pauli Ollila*
11.55-12.20	**Milda Slickute-Sestokiene**: The use of administrative data sources for Lithuanian annual data of earnings *Discussant: Andris Fisenko*

Afternoon session

Chair: Janis Lapins

14.00-14.25	**Andrius Ciginas**: On Variance minimization for unequal probability sampling *Discussant: Imbi Traat*
14.35-15.00	**Juris Breidaks**: Quality Analysis in a Survey on Transportation of Goods by Road *Discussant: Vaiva Virketyte*
15.30-16.50	Round Table dicussion
16.50-17.00	Closing of the Workshop
19.00-21.30	***Farewell party***

August 29, Tuesday

Departure

IS THERE A BEST FIXED SIZE πPS SAMPLING DESIGN?

Lennart Bondesson
Umeå university, Sweden
e-mail: Lennart.Bondesson@math.umu.se

Abstract

Fixed size πps sampling with prescribed inclusion probabilities is considered. It is discussed whether there is a best πps design. Several candidates are presented, as the Sampford design, the adjusted conditional Poisson design, the adjusted Pareto design, and some other designs. No definite conclusion is presented.

1. Introduction

A population of units $1, 2, \ldots, N$ is considered. We want to take a sample without replacement of size n according to given inclusion probabilities $\pi_1, \pi_2, \ldots, \pi_N$ with sum $\sum_{i=1}^N \pi_i = n$. The inclusion probabilities are assumed to be roughly proportional to the y_i-values of an interesting y-variable. We intend to estimate the population total by the Horvitz-Thompson estimator

$$\hat{Y}_{HT} = \sum_{i=1}^{N} \frac{y_i}{\pi_i} I_i,$$

where I_i is 1 if unit i is sampled and otherwise 0. We set $a_i = \breve{y}_i = y_i/\pi_i$ and then $\hat{Y}_{HT} = \sum_{i=1}^N a_i I_i$. The variance of the HT-estimator can be written as

$$Var(\hat{Y}_{HT}) = \mathbf{a}^T \mathbf{\Sigma} \mathbf{a},$$

where $\mathbf{\Sigma} = (c_{ij})$ is the matrix of covariances $c_{ij} = Cov(I_i, I_j)$ with $c_{ii} = d_i = \pi_i(1 - \pi_i)$. We also have, the Sen-Yates-Grundy form of the variance:

$$Var(\hat{Y}_{HT}) = \frac{1}{2} \sum_{i,j} \tilde{c}_{ij} (a_i - a_j)^2, \quad \text{where} \quad \tilde{c}_{ij} = -c_{ij}.$$

Dividing here $\tilde{c}_{ij}$ by $\pi_{ij} = E(I_i I_j)$ and then summing instead over i and j in the sample, we get the SYG variance estimator.

There is no sampling design with smallest variance uniformly in $\mathbf{a}$. In fact, if such a design with covariance matrix $\mathbf{\Sigma}_0$ existed, we would have $\mathbf{a}^T \mathbf{\Sigma}_0 \mathbf{a} \leq \mathbf{a}^T \mathbf{\Sigma} \mathbf{a}$ for all $\mathbf{a}$ and all other $\mathbf{\Sigma}$ with diagonal elements $d_i = \pi_i(1 - \pi_i)$, $i = 1, \ldots, N$. Then $\mathbf{D} = \mathbf{\Sigma} - \mathbf{\Sigma}_0 \geq 0$ (in matrix sense) and hence the eigenvalues of $\mathbf{D}$ are nonnegative. But they sum to 0 because the diagonal elements of $\mathbf{D}$ are 0 and hence trace$(\mathbf{D}) = 0$. Thus all the eigenvalues are 0 and hence $\mathbf{\Sigma} = \mathbf{\Sigma}_0$, which is a contradiction.

We have not any superpopulation in mind except possibly the simplest one: $a_i = \alpha + \epsilon_i$, where the ϵ_is are uncorrelated with mean zero and variance σ^2. For that model, with E here denoting expected value w.r.t. the superpopulation, we have, since $\sum_{j;\, j\neq i} c_{ij} \equiv -\pi_i(1-\pi_i)$,

$$\mathrm{E}(Var(\mathbf{a}^T\mathbf{I})) = \frac{1}{2}\sum_{i,j;\, i\neq j} \tilde{c}_{ij}\mathrm{E}((a_i - a_j)^2) = \frac{1}{2}\sum_{i,j;\, i\neq j} \tilde{c}_{ij}2\sigma^2 = \sigma^2\sum_{i=1}^{N}\pi_i(1-\pi_i).$$

Hence all designs are equally efficient with respect to this superpopulation.

So to single out a 'best' design further considerations are needed. In the literature, e.g. Brewer *et al.* (1983), many πps designs are discussed. Some of them are approximate in the sense that $E(I_i) \approx \pi_i$ only. Such designs are not considered here. The following three designs deserve a lot of attention as candidates for being at least very good πps designs.

1. The Sampford design
2. The adjusted conditional Poisson design
3. The adjusted Pareto design.

These are discussed in sections 2 and 3. We give motivations for them and present advantages and drawbacks of them. In section 4 we derive some slightly more theoretical designs related to the conditional Poisson design. In section 5 and in an appendix we derive and discuss some further designs, which are of 2nd order type, i.e. are only given by their 2nd order inclusion probabilities. They are also more theoretical than practical. We illustrate and compare the methods in section 6 by looking at a small but not trivial population for which $N = 6, n = 3$, and $\pi_1 = \pi_2 = \pi_3 = 1/3$ and $\pi_4 = \pi_5 = \pi_6 = 2/3$. This population, the TBM-population, has earlier been considered in Traat *et al.* (2004). The paper ends with a brief discussion in section 7.

2. The Sampford, the adjusted conditional Poisson, and the adjusted Pareto designs

Here we look at the three designs mentioned in the introduction. We present the designs mainly by their probability functions (pf) $p(\mathbf{x}) = \Pr(\mathbf{I} = \mathbf{x})$, where $\mathbf{x} = (x_1, x_2, \ldots, x_N)$ with $x_i = 0$ or 1.

The **Sampford design** was introduced by Sampford (1967). Its pf is given by

$$p_S(\mathbf{x}) = C_S\prod_{i=1}^{N}\pi_i^{x_i}(1-\pi_i)^{1-x_i} \times \sum_{k=1}^{N}(1-\pi_k)x_k, \quad |\mathbf{x}| = \sum_{i=1}^{N} x_i = n.$$

It is a profound result that the true inclusion probabilities really equal π_i. The constant C_S is inexplicit but otherwise the pf is very explicit. It is possible to sample from this pf by

first sampling one unit with replacement according to the probabilities π_i/n, $i = 1, \ldots, N$, and then with replacement $n-1$ units according to the probabilities $p_i' \propto \pi_i/(1-\pi_i)$. If all these n units are distinct, the sample is accepted, otherwise the whole procedure is repeated. This is often a slow procedure but there are also other methods to sample from the pf (e.g. Grafström, 2005).

The **adjusted conditional Poisson design** was introduced by Hajek (1964, 1981). Tillé (2005) gives it a careful treatment. The pf is

$$p_{CP}(\mathbf{x}) = C_{CP} \times \prod_{i=1}^{N} p_i^{x_i}(1-p_i)^{1-x_i}, \quad |\mathbf{x}| = n,$$

where p_i with $\sum_{i=1}^{N} p_i = n$ must be chosen so that the desired inclusion probabilities π_i are obtained. Hajek presented various approximations but nowadays it is also possible to calculate the desired p_is numerically by a computer program (e.g. Tillé, 2005). A simple recent good approximation is, with $d = \sum_{k=1}^{N} \pi_k(1-\pi_k)$,

$$\frac{p_i}{1-p_i} \propto \frac{\pi_i}{1-\pi_i} \exp(\frac{1-\pi_i}{d}).$$

This approximation was derived in Bondesson *et al.* (2006) via the assumption that $p_{CP}(\cdot)$ is close to the Sampford pf. It turns out to yield a very good approximation. To sample from the conditional Poisson design is easy, one samples from the Poisson design (independent I_is, with $I_i \sim Bin(1, p_i)$) but only samples of the desired size are accepted.

The **Pareto design** was introduced by Rosén (1997a,b). The main idea dates back to Ohlsson (1990) and Saavedra (1995). Target probabilities λ_i such that $\sum_{i=1}^{N} \lambda_i = n$ are used. Let $U_1, U_2, \ldots, U_N$ be random numbers from $U(0,1)$ and let

$$Q_i = \frac{U_i/(1-U_i)}{\lambda_i/(1-\lambda_i)}, \quad i = 1, \ldots, N,$$

be *ranking variables*. Now select the n units with the smallest Q_is. If we put $\lambda_i = \pi_i$, $i = 1, \ldots, N$, the true inclusion probabilities will approximately equal the π_is but not exactly. It is possible to make an adjustment so that the true inclusion probabilities will be π_i (Aires, 2000). A very good approximation in this direction is provided by, with $d = \sum_{k=1}^{N} \pi_k(1-\pi_k)$,

$$\frac{\lambda_i}{1-\lambda_i} \propto \frac{\pi_i}{1-\pi_i} \exp(-\frac{\pi_i(1-\pi_i)(\pi_i - \frac{1}{2})}{d^2}).$$

It is derived in Bondesson *et al.* (2006) from the assumption that the adjusted Pareto pf is close to the Sampford pf. Hence the Q_is above, with $\lambda_i = \pi_i$, only have to be multiplied by the factor $\exp(\pi_i(1-\pi_i)(\pi_i - \frac{1}{2})/d^2)$ to yield a sample with inclusion probabilities π_i.

The pf for the Pareto design is given by

$$p_{Par}(\mathbf{x}) = \prod_{i=1}^{N} \lambda_i^{x_i}(1-\lambda_i)^{1-x_i} \times \sum_{k=1}^{N} c_k x_k,$$

where the constants c_k are given by certain integrals (Traat *et al.*, 2004, Bondesson *et al.*, 2006). Approximately, $c_k \propto 1-\lambda_k$, which shows that the Pareto and the adjusted Pareto pfs are close to the Sampford pf.

3. Advantages and drawbacks of the designs

The sampling designs in section 2 have pfs that are very close to each other. Should one of these designs be preferred? We look here at advantages and drawbacks of each of them, in the order: Sampford sampling, Pareto sampling and conditional Poisson sampling.

Sampford sampling. The main advantage of this design is that the pf is very explicit. A main drawback has been that the original methods to get a Sampford sample are slow. However, since the pf is explicit except for the normalizing constant, one can easily use MCMC methods as Gibbs sampling to sample from it. There are now also very rapid methods that use Pareto sampling in a first step and then acceptance/rejection technique (Bondesson *et al.*, 2006, Grafström, 2005). Another small drawback is that there is no known optimality property of Sampford sampling.

Pareto sampling. The big advantage of this method is that it is very easy to get a sample. Without adjustment the method gives a slight bias of the estimators. Although the bias is small, it is slightly disturbing and therefore one may advocate at least simple adjustment; cf. section 2. Another advantage of the method is that it permits the use of permanent random numbers. A drawback of the method is that there is no simple pf. It is also complicated to calculate the true inclusion probabilities. The method has no known optimum property except that it is asymptotically the best method among all order sampling procedures.

Adjusted conditional Poisson sampling. A main advantage of this design is that the entropy $-\sum p(\mathbf{x})\log(p(\mathbf{x}))$ is maximized under the given restrictions (Hajek, 1981). Thus the probabilities are spread over the possible samples as much as possible in some sense. The probability function belongs to an exponential family. A drawback is that the pf is not very explicit since the p_is must be calculated. Another drawback is that the standard rejective procedure for sampling takes some time for large populations and samples but there are list-sequential methods also (Chen & Liu 1997, Öhlund, 1999, Traat *et al.*, 2004, Tillé, 2005). There are also rapid methods based on preliminary Pareto samples which are accepted or rejected. (Bondesson *et al.*, 2006, Grafström, 2005).

4. Other designs related to the adjusted conditional Poisson design

Using two different starting points, we derive some further designs related to the conditional Poisson design.

The maximum entropy property for the adjusted conditional Poisson design can be expressed in another way too. Let $p_I(\mathbf{x})$ denote the pf for Poisson sampling with probabilities $\pi_i, i = 1, 2, \ldots, N$. For a fixed size design with pf $p(\mathbf{x})$ and inclusion probabilities π_i, let us minimize the Kullback-Leibler divergence $KL = \sum_{\mathbf{x};\,|\mathbf{x}|=n} p(\mathbf{x}) \log(p(\mathbf{x})/p_I(\mathbf{x}))$. We have

$$KL = \sum_{\mathbf{x};\,|\mathbf{x}|=n} p(\mathbf{x}) \log p(\mathbf{x}) \; - \sum_{\mathbf{x};\,|\mathbf{x}|=n} \left[p(\mathbf{x}) \sum_{i=1}^{N} (x_i \log \pi_i + (1 - x_i) \log(1 - \pi_i)) \right]$$

$$= -\text{Entropy} \; + \sum_{i=1}^{N} \left[\log \pi_i \sum_{\mathbf{x};\,|\mathbf{x}|=n} x_i p(\mathbf{x}) \; + \; \log(1 - \pi_i) \sum_{\mathbf{x};\,|\mathbf{x}|=n} (1 - x_i) p(\mathbf{x}) \right]$$

$$= -\text{Entropy} \; + \sum_{i=1}^{N} (\pi_i \log \pi_i + (1 - \pi_i) \log(1 - \pi_i)).$$

Now since the entropy is maximized for the adjusted conditional Poisson pf, which can be proved by a use of Lagrange multipliers, it follows that KL is minimized for that pf. Of course, one could then also try to minimize another distance measure, the squared Hellinger metric

$$d_H^2 = \sum_{\mathbf{x}} (\sqrt{p(\mathbf{x})} - \sqrt{p_I(\mathbf{x})})^2 = 2 \; - \; 2 \times E_I(\sqrt{\frac{p(\mathbf{x})}{p_I(\mathbf{x})}})$$

given that $\sum_{\mathbf{x}} x_i p(\mathbf{x}) = \pi_i$, $i = 1, 2, \ldots, N$, and $p(\mathbf{x}) = 0$ for $|\mathbf{x}| \neq n$. It is more difficult to minimize d_H^2 but it is possible for small populations for which the number of different samples is limited. It would have been some extra support for the adjusted conditional Poisson design if the 'Hellinger design' had been equal to that design. As will be seen in section 6 it is not the case.

The maximum entropy for the adjusted conditional Poisson design ought to guarantee that the variance of the HT-estimator is small though not in a very direct way. Since $p \log p$ is a limit of $p(p^\epsilon - 1)/\epsilon$ as $\epsilon \downarrow 0$, we see that maximum entropy corresponds to minimization of $\sum_{\mathbf{x};\,|\mathbf{x}|=n} (p(\mathbf{x}))^{1+\epsilon}$ for an ϵ close to 0. In this connection, Hölder's inequality may give some additional insight. We have, with $x_i = I_i$ and $\hat{Y}(\mathbf{x}) = \hat{Y}_{HT}$,

$$Var(\hat{Y}) = \sum_{\mathbf{x};|\mathbf{x}|=n} (\hat{Y}(\mathbf{x}) - Y)^2 p(\mathbf{x}) \leq \left(\sum_{\mathbf{x};\,|\mathbf{x}|=n} (p(\mathbf{x}))^{1+\epsilon} \right)^{\frac{1}{1+\epsilon}} \left(\sum_{\mathbf{x};\,|\mathbf{x}|=n} (\hat{Y}(\mathbf{x}) - Y)^{2\frac{1+\epsilon}{\epsilon}} \right)^{\frac{\epsilon}{1+\epsilon}}.$$

But of course here we could also use $\epsilon = 1$ (Cauchy's inequality) or $\epsilon = \infty$. This would lead to designs where we minimize

$$\sum_{\mathbf{x};\,|\mathbf{x}|=n} (p(\mathbf{x}))^2 \quad \text{or} \quad \max_{\mathbf{x};\,|\mathbf{x}|=n} p(\mathbf{x})$$

given the restrictions. At least for symmetry reasons, it may seem more natural to use $\epsilon = 1$ than ϵ very close to 0. We call these designs the minsum$-p^2$ and the minmax$-p$ designs, respectively. They are more difficult to manage than the conditional Poisson design but at least for small populations and samples they can be handled.

5. Some second order designs

In this section we look at designs defined by 2nd order inclusion probabilities only. Although it is not completely true that in sampling higher order inclusion probabilities are irrelevant, we focus on the 2nd order ones here.

Hajek (1981) thought that it would be desirable to have a design with $c_{ij} = Cov(I_i, I_j)$ of the simple product form $c_{ij} = -c_i c_j,\ i \neq j$. Then there is a simple expression for the variance of the HT-estimator. Moreover a good approximation of the covariances of the adjusted conditional Poisson design is obtained. It is possible to solve these equations by iterative methods but the solution is inexplicit. The solution is of the form

$$c_{ij}^H = -\frac{\pi_i(1-\lambda_i)\pi_j(1-\lambda_j)}{\sum \pi_k(1-\lambda_k)},$$

where $\lambda_i \approx \pi_i$. Hajek was not able to show that there really is a sampling design with the derived covariances and the 2nd order inclusion probabilities $\pi_{ij} = c_{ij}^H + \pi_i\pi_j$. Nowadays at least for small populations one can use linear programming to find such designs: we should solve the linear equations $\sum_{\mathbf{x};\ |\mathbf{x}|=n} x_i x_j p(\mathbf{x}) = \pi_{ij}$ for $p(\mathbf{x})$ under nonnegativity restrictions. It is also possible to use a pf of the form $p(\mathbf{x}) = \prod_{i=1}^N \pi_i^{x_i}(1-\pi_i)^{1-x_i}\, Q(\mathbf{x})$, where Q is a quadratic form that has to be calculated (Lundqvist & Bondesson, 2005).

Hajek also derived his product form by maximizing the 'entropy' $\sum_{i,j;\,i\neq j} c_{ij}\log(-c_{ij})$. Bondesson *et al.* (2006) instead minimized the measure

$$\text{SSCorr} = \sum_{i,j;\,i\neq j} \rho_{ij}^2,$$

where $\rho_{ij} = Corr(I_i, I_j)$. The restrictions $\sum_{j=1}^N c_{ij} = 0$ together with Lagrange multiplier technique show that there is an explicit solution:

$$c_{ij}^{BTL} = -\pi_i(1-\pi_i)\pi_j(1-\pi_j)(\ _i + \ _j), \quad i \neq j,$$

where

$$_i = \frac{\frac{1}{d-2\pi_i(1-\pi_i)}}{1+\sum \frac{\pi_k(1-\pi_k)}{d-2\pi_k(1-\pi_k)}} \quad \text{with} \quad d = \sum \pi_k(1-\pi_k).$$

Most often these covariances are very close to c_{ij}^H. Of course, to minimize SSCorr gives in some sense as much pairwise independence as possible to the inclusion variables I_i. We also have, with $a_i = y_i/\pi_i$, $\tilde{\rho}_{ij} = -\rho_{ij}$, and $d_i = \pi_i(1-\pi_i)$, by Cauchy's inequality,

$$Var(\hat{Y}_{HT}) = Var(\sum a_i I_i) = \frac{1}{2} \sum_{i,j;\, i \neq j} \tilde{c}_{ij}(a_i - a_j)^2$$

$$= \frac{1}{2} \sum_{i,j;\, i \neq j} \tilde{\rho}_{ij} \sqrt{d_i d_j}\,(a_i - a_j)^2 \;\leq\; \frac{1}{2}\sqrt{\text{SSCorr}} \times \sqrt{\sum_{i,j;\, i \neq j} d_i d_j (a_i - a_j)^4}.$$

For a fixed population and fixed inclusion probabilities, the last factor above is constant. However, we can affect SSCorr and by minimizing it we get in a direct way some guarantee that $Var(\hat{Y}_{HT})$ becomes small.

There are several simple variants of the approach above. Instead of focusing on the correlation, we may focus on the covariance. Minimizing the sum of squared covariances under the appropriate restrictions, we get the solution:

$$c_{ij} = \frac{1}{N-2}\left(\frac{d}{N-1} - \pi_i(1-\pi_i) - \pi_j(1-\pi_j)\right), \quad i \neq j.$$

Often this is not a useful solution since the signs of the covariances may vary. They should preferably be nonpositive to give a stable Sen-Yates-Grundy variance estimator. Instead of Cauchy's inequality, we may use Hölder's inequality. In particular, we may use Hölder's inequality with the exponents $p = \infty$ and $q = 1$. This leads to second order minimax-designs.

We now turn to such designs and use first the covariance. Let $a_i = y_i/\pi_i$ and $\mathbf{a} = (a_1, a_2, \ldots, a_N)^T$. Set $\mathbf{1} = (1, 1, \ldots, 1)^T$. Then $\mathbf{a} = \bar{a}\,\mathbf{1} + \mathbf{b}$, where $\mathbf{b} = (a_1 - \bar{a}, \ldots, a_N - \bar{a})$ is orthogonal to $\mathbf{1}$. Now, since $\sum I_i = n$,

$$Var(\hat{Y}_{HT}) = Var(\mathbf{a}^T\mathbf{I}) = Var(\mathbf{b}^T\mathbf{I}) = \mathbf{b}^T\mathbf{\Sigma}\mathbf{b} \;\leq\; \lambda_{\max}||\mathbf{b}||^2 = \lambda_{\max}\sum_{i=1}^{N}(a_i - \bar{a})^2,$$

where $\lambda_{\max}$ is the maximal eigenvalue of $\mathbf{\Sigma}$. Now we could try to choose a covariance matrix $\mathbf{\Sigma}$ with given diagonal elements $d_i = \pi_i(1-\pi_i)$ and eigenvector $\mathbf{1}$ with eigenvalue $\lambda_1 = 0$ and such that its maximal eigenvalue is minimal. We should add the condition that $\tilde{c}_{ij} \geq 0$, $i \neq j$, where $\tilde{c}_{ij} = -c_{ij}$. This is a problem that in some cases can be solved.

We can alternatively describe the problem as follows. We have

$$Var(\mathbf{a}^T\mathbf{I}) = \frac{1}{2}\sum_{i,j;i\neq j}\tilde{c}_{ij}(a_i-a_j)^2 \leq \max_{i,j;\,i\neq j}\tilde{c}_{ij}\,\frac{1}{2}\sum_{i,j;\,i\neq j}(a_i-a_j)^2 = \max_{i,j;\,i\neq j}\tilde{c}_{ij}\times N\sum_{i=1}^{N}(a_i-\bar{a})^2.$$

Now we should try to find a covariance matrix with $\mathbf{1}$ as eigenvector with eigenvalue 0 and such that $\max_{i,j;\,i\neq j}\tilde{c}_{ij}$ is minimal. Additionally we should require that $\tilde{c}_{ij}\geq 0$, $i\neq j$, to get a stable variance estimator. This is a problem that in small cases can be solved by linear programming for determination of the appropriate $\tilde{c}_{ij}$. Often, but not always the solution is of the form that the matrix has all its elements in the row with the largest $d_i=\pi_i(1-\pi_i)$ equal and if that row is the first row then equal to $-d_1/(N-1)$ (since all rows sums are 0). We then have

$$Var(\mathbf{a}^T\mathbf{I}) \leq \frac{Nd_1}{N-1}\sum_{i=1}^{N}(a_i-\bar{a})^2.$$

In fact, in this case the inequality becomes an equality for $\mathbf{a}\propto(N-1,\,-1,\,-1,\ldots,\,-1)$ which is an eigenvector with eigenvalue $Nd_1/(N-1)$. Hence the inequality is sharp in some sense.

We can also use minimax designs w.r.t. the correlation. We have, with $d_i=\pi_i(1-\pi_i)$,

$$Var(\mathbf{a}^T\mathbf{I}) \leq \max_{i,j;\,i\neq j}\tilde{\rho}_{ij}\,\frac{1}{2}\sum_{i,j;\,i\neq j}(a_i-a_j)^2\sqrt{d_id_j} = \max_{i,j;\,i\neq j}\tilde{\rho}_{ij}\,(\sum_{i=1}^{N}\sqrt{d_i})^2\sum_{i=1}^{N}(a_i-\bar{\bar{a}})^2p_i',$$

where $p_i'\propto\sqrt{d_i}$ with $\sum p_i'=1$ and $\bar{\bar{a}}$ is a weighted mean. We may therefore try to minimize $\max\tilde{\rho}_{ij}$ under simple restrictions (see below). An alternative, but not equivalent approach, is to set $b_i=a_i\sqrt{d_i}$ and then use the inequality

$$Var(\mathbf{a}^T\mathbf{I}) = \mathbf{b}^T\mathbf{R}\mathbf{b} \;\leq\; \lambda_{\max}||\mathbf{b}||^2$$

for the correlation matrix $\mathbf{R}$. The maximal eigenvalue of $\mathbf{R}$ should then be minimized under the restrictions that $(\sqrt{d_1},\sqrt{d_2},\ldots,\sqrt{d_N})^T$ is an eigenvector of $\mathbf{R}$ with eigenvalue 0 and $\tilde{\rho}_{ij}\geq 0$, $i\neq j$.

6. Example: The TBM population

Here we return to the TBM-population in section 1 with $N=6, n=3$, and $\pi_1=\pi_2=\pi_3=\frac{1}{3}, \pi_4=\pi_5=\pi_6=\frac{2}{3}$. The population is simple but it illustrates many things in a good way. There are 20 possible samples of size $n=3$ but only 4 with distinct probabilities: $\{1,2,3\},\{1,2,4\},\{1,4,5\},\{4,5,6\}$. Each of the samples 2 and 3 has 8 variants. We set

$p_1 = \pi_{123}$, $p_2 = \pi_{124}$, $p_3 = \pi_{145}$, $p_4 = \pi_{456}$. Then $p_1 + 6p_2 + 3p_3 = 1/3$ and $3p_2 + 6p_3 + p_4 = 2/3$ implying that $p_1 + 9p_2 + 9p_3 + p_4 = 1$. It is easy to experiment with this population since there are only a few parameters to vary, e.g. p_1 and p_2. In Table 1 different characteristics have been calculated for all the designs considered.

Table 1: The TBM-population; 2nd order inclusion probabilities, sample probabilities, and some other characteristics for seven different sampling designs.

	MinSSCorr	Sampford	Pareto(adj)	CP(adj)	Hellinger	minsum-p^2	minmax-p
π_{12}	0.06667	0.06918	0.06973	0.07081	0.07170	0.04762	0.03030
π_{14}	0.17778	0.17610	0.17574	0.17501	0.17442	0.19048	0.20202
π_{45}	0.400	0.40252	0.40306	0.40415	0.40503	0.38095	0.36364
p_1	0	0.00629	0.00647	0.00686	0.00730	0	0
p_2	0.02222	0.02096	0.02109	0.02132	0.02147	0.01587	0.01010
p_3	0.06667	0.06709	0.06678	0.06619	0.06574	0.07937	0.09091
p_4	0.20	0.20126	0.20272	0.20556	0.20780	0.14286	0.09091
SSCorr	**1.2**	1.2025	1.2037	1.2071	1.2103	1.3469	1.7355
$\max \tilde{\rho}_{ij}$	**0.20**	0.2075	0.2092	0.2125	0.2151	0.2857	0.3636
Entropy	2.7080	2.7150	2.7151	**2.7152**	2.7151	2.6796	2.5976

There are several solutions, $\mathbf{p} = (p_1, p_2, p_3, p_4)$ in the MinSSCorr case. Above an extreme solution is given. All the different designs considered in section 5 lead for this simple population to the MinSSCorr solution in the second column. Since $d_i = \pi_i(1 - \pi_i) \equiv 2/9$, it does not even matter whether we consider the correlation or the covariance. Because of the very symmetric character of this design, $\rho_{ij} \equiv -0.2$ for $i \neq j$, one may think that this is a very good solution in this simple case. Its entropy can be increased to 2.7142 by the choice of a less extreme MinSSCorr design among the possible variants. The Sampford, the Pareto-adjusted, the CP-adjusted, and the Hellinger designs do not agree with the MinSSCorr design although they are very close to it. These latter four designs are pairwise very equal in this example with {Sampford, Pareto(adj)} and {CP(adj), Hellinger} as the pairs. This is a relation that is true in general as found by Lundqvist (2006). The minsum-p^2 and the minmax-p designs seem a bit extreme compared to the other designs. They have much higher SSCorr and lower entropy than the other five designs. They have also very low values of π_{12} which leads to less stable variance estimators.

It was mentioned in the introduction that for the simple superpopulation model $a_i = \alpha + \epsilon_i$ with i.i.d. ϵ_is, the expected (design) variance of the HT-estimator is the same for all πps designs with the given inclusion probabilities. We may then also look at the expected design variance of the Sen-Yates-Grundy estimator of the variance of the HT-estimator. Assuming that the fourth central moment of ϵ_i equals $3Var(\epsilon_i) = 3\sigma^4$, we got the following expected values.

Table 2: Expected values of the design variance of the SYG variance estimator under a simple superpopulation model.

MinSSCorr	Sampford	Pareto(adj)	CP(adj)	'Hellinger'
$1.844\sigma^4$	$1.816\sigma^4$	$1.806\sigma^4$	$1.789\sigma^4$	$1.778\sigma^4$

Thus the Hellinger design gives a slightly more stable variance estimator than the other included designs. On the other hand, by setting $p_1 = 0.0813$, $p_2 = 0$, $p_3 = 0.0840$, and $p_4 = 0.1626$, we get the smallest possible expected value: $1.650\sigma^4$. Of course, it is well known that the requirement of a small variance for an estimator is in conict with the requirement of a stable variance estimator.

7. Discussion

It is a bit annoying that it is not really possible to single out a best fixed size πps sampling design. There is no doubt about that at present the (adjusted) Pareto design is the best one to select a sample easily. On the other hand the adjusted conditional Poisson design has a very attractive maximum entropy property. The Sampford design has a simple and nice pf and is very attractive from that point of view. Fortunately these three designs are close to each other. Brewer (2002) classifies the Pareto design as a high entropy design. The minimax designs considered in section 5 are at present not very practical but focus more directly on making the variance small than the other designs. Gabler (1990) presents many results on strongly related minimax designs.

Finally, it is appropriate to add that if there is relevant auxiliary information, many other designs are possible, as e.g. systematic πps sampling designs, and may be better than the ones considered here.

References

Aires, N. (2000). Comparisons between conditional Poisson sampling and Pareto πps sampling designs. *J. Statist. Plann. Inference* **88**, 133-147.

Bondesson, L., Traat, I., and Lundqvist, A. (2006). Pareto sampling versus Sampford and conditional Poisson sampling. *Scand. J. Statist.* **33**, to appear.

Brewer, K.R.W. & Hanif, M. (1983). *Sampling with unequal probabilities.* Lecture Notes in Statistics, No. 15. Springer-Verlag, New York.

Brewer, K.R.W (2002). Combined survey sampling inference: weighing of Basu's elephant. Hodder Arnold, London.

Chen, S.X. & Liu, J.S. (1997). Statistical applications of the Poisson-Binomial and conditional Bernoulli distributions. *Statistica Sinica* **7**, 875-892.

Gabler, S (1990). *Minimax solutions in sampling from finite populations.* Lecture Notes in Statistics **64**. Springer-Verlag, Berlin, Heidelberg.

Grafström, A. (2005). Comparisons of methods to generate conditional Poisson samples and Sampford samples. Master's thesis, Department of mathematics and mathematical statistics, Umeå university.

Hajek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Ann. Math. Statist.* **35**, 1491-1523.

Hajek, J. (1981). *Sampling from a finite population.* Marcel Dekker, New York.

Lundqvist, A. (2006). Comparing distances between some distributions that occur in sampling. Manuscript.

Lundqvist, A. & Bondesson, L. (2005). On sampling with desired inclusion probabilities of first and second order. Research report 3 in mathematical statistics, Umeå university, Sweden.

Ohlsson, E. (1990). Sequential Poisson sampling from a business register and its application to the Swedish consumer price index. Statistics Sweden R&D Reports 1990:6.

Öhlund, A. (1999). Comparisons of different methods to generate Bernoulli distributed random numbers given their sum. Master's thesis, Department of mathematical statistics, Umeå university, Sweden (in Swedish).

Rosén, B. (1997a). Asymptotic theory for order sampling. *J. Statist. Plann. Inference* **62**, 135-158.

Rosén, B. (1997b). On sampling with probability proportional to size. *J. Statist. Plann. Inference* **62**, 159-191.

Saavedra, P. (1995). Fixed sample size PPS approximations with a permanent random number. 1995 Joint Statistical Meetings American Statistical Association, Orlando, Florida.

Sampford, M.R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika* **54**, 499-513.

Tillé, Y. (2005). *Sampling algorithms.* Technical Report, Neuchâtel, Switzerland.

Traat, I., Bondesson, L. & Meister, K. (2004). Sampling design and sample selection through distribution theory. *J. Statist. Plann. Inference* **123**, 395-413.

SOME MODEL-BASED ESTIMATOR

Danutė Krapavickaitė[1] and Vilma Nekrašaitė[2]

[1] Institute of Mathematics and Informatics, Lithuania; Statistics Lithuania, Lithuania
e-mail: krapav@ktl.mii.lt

[2] Vilnius Gediminas Technical University, Lithuania; Statistics Lithuania, Lithuania
e-mail: vilma.nekrasaite@stat.gov.lt

Abstract

A design-based and model-based estimator of a total in a finite population (Valliant et. al 2000) and estimation of the variance of the model-based estimator (Valliant 1985) is discussed. The study variable having zero and positive values is considered. Some econometric models for this variable are being suggested to use. Some simulation results are given.

1 Design-based and model-based approaches in survey sampling

In the main branches of statistics data is considered as realisation of some random variables. The aim of a statistician is to make inference about the probability laws of these random variables. The survey sampling is historically isolated from the mainstream of statistics. The data in this field is considered as fixed and randomness is introduced by statistician when selecting data for observation from the whole amount of data. The aim of the survey statistician is to get a statistical estimate of some fixed parameter describing the whole amount of data. The main branches of statistics would investigate a probability law of the random parameter of the randomly generated data.

The main statistical approach to the survey sampling is taken in the book of Valliant et al. (2000) and prediction of the population parameters is investigated. The values of a study variable y : $y_1, y_2, ..., y_N$ in the finite population $U = \{1,2,..., N\}$ are considered to be random, generated by some statistical model. The population total

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

is also random. Given a probability sample s from U, $s \subset U$, the value of the total t_y can be predicted (denoted by $\hat{t}_y$) after the individual values of y_k, $k \in U \setminus s$ are predicted (let us denote them by $\hat{y}_k$):

$$\hat{t}_y = \sum_{k \in s} y_k + \sum_{k \in U \setminus s} \hat{y}_k \qquad (1)$$

Properties of linear (Valliant et al. 2000) and nonlinear (Valliant, 1985) models of y are being studied, accuracy of the predictor $\hat{t}_y$ for a total t_y of study variable y is investigated.

2 Models for some skew distributed study variable

We will discuss prediction of a finite population total for a special case of a study variable y, which obtains zero values in some cases and positive values in other cases. Such situation arises when investigating plots under the crops, which are being grown up only in some parts of the country; when investigating expenditure of the enterprises to the protection of the environment.

Let us denote two vectors of the auxiliary variables $\mathbf{x}^{(1)}$, $\mathbf{x}^{(2)}$ with the values $\mathbf{x}_1^{(1)},...,\mathbf{x}_N^{(1)}$ and $\mathbf{x}_1^{(2)},...,\mathbf{x}_N^{(2)}$, and two unobserved variables $y^{(1)}$, $y^{(2)}$ with the values $y_1^{(1)},...,y_N^{(1)}$ and $y_1^{(2)},...,y_N^{(2)}$ satisfying the models

$$y^{(1)} = \beta_1' \mathbf{x}^{(1)} + u_1,$$
$$y^{(2)} = \beta_2' \mathbf{x}^{(2)} + u_2 \qquad (2)$$

with the vectors of constants β_1, β_2 and random errors u_1, u_2 distributed according to a normal law with zero mean (the indices denoting the values of the variables are omitted here). Let us suppose that a study variable y satisfies condition

$$y = \begin{cases} y^{(1)} & \text{if } y^{(1)} \geq y^{(2)}, \\ 0 & \text{otherwise.} \end{cases} \qquad (3)$$

This is a censored regression model with unobserved stochastic threshold - a case of Heckman model (Maddala, 1983). In the case when $y^{(2)} \equiv 0$ instead of (2) the variable y in (3) satisfies the conditions of the censored regression (tobit) model (Greene (2002), Maddala (1983)).

The estimator (1) for the model (3) will be investigated in the lecture. The results of simulations show that the accuracy of the tobit-model based estimator of total in some cases can be much better than that of design-based estimator under a simple random sampling.

References

Greene W.H. (2002) *Econometric Analysis*. Prentice Hall, Upper Saddle River.

Maddala G.S. (1983) *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge University Press, Cambridge.

Valliant, R. (1985) Nonlinear Prediction Theory and the Estimation of Proportion in a Finite Population. *Journal of American Statistical Association*, **80**, 631-641.

Valliant R.,, Dorfman A.H. and Royall M. (2000) *Finite Population Sampling and Inference: a Prediction Approach*. John Wiley & Sons, New York.

Sampling and Estimation of Multi-National Surveys with Examples from the European Social Survey

Seppo Laaksonen
University of Helsinki and Statistics Finland
Seppo.Laaksonen@Helsinki.Fi [1]

Key words: Design effects, effective sample size, intra-cluster correlation, sampling frames, cross-country analysis of happiness

1. Introduction

This paper first discusses the objectives of sample design for cross-national surveys (section 2). Then we describe the principles and requirements for sample design that were developed for the European Social Survey (ESS) in order to meet these objectives (section 3). In particular, these include a requirement to predict design effects and to use these predictions in determining national sample sizes. The procedures used on the ESS are described in section 4, and some of the strengths and weaknesses are pointed out. In section 5, some cross-country analysis have been presented, just as examples how to do it and to motivate to exploit these data files that are freely available for everyone. Section 6 concludes and presents some ideas for improving the survey.

2. Sample Design for Cross-National Surveys

To enable comparisons between nations, the ESS sampling group suggests that national sample designs for cross-national surveys must meet two fundamental criteria:

- The study population must be equivalent in each nation. In practice, this will usually mean that the same population definition is applied in each nation and that no or minimal under-coverage can be permitted;
- Sample-based estimates must have known and appropriate precision in each nation. In practice, "known" precision means that a strict probability sample design must be used, and those aspects of sample design that affect precision (selection probabilities, stratum membership, primary sampling unit (psu) membership) must be available on the microdata to permit estimation of standard errors; "appropriate" precision may mean, a) meeting some minimum precision requirement in order for the estimates to be useful and, b) aiming for similar precision in each nation.

To best meet these criteria, it is likely that details of the sample design will vary between nations (Le and Verma, 1997). The goal is functional equivalence, not replication of parameters of the sample

design. As Kish (1994, 173) writes, "Sample designs may be chosen flexibly and there is no need for similarity of sample designs. Flexibility of choice is particularly advisable for multinational comparisons, because the sampling resources differ greatly between countries. All this flexibility assumes probability selection methods: known probabilities of selection for all population elements." Therefore, an optimal sample design for a cross-national survey should consist of the best probability sample design possible in each nation, where "best" can be interpreted as an optimum trade-off between cost and precision. The choice of a specific national design depends on the available frames, experiences, other constraints such as those that may be imposed by the national legal infrastructure and, of course, costs of sample selection and data collection (Häder and Gabler, 2003). If adequate estimators are chosen, the resulting estimates can be compared using appropriate statistical tests.

3. Requirements of Sample Design for the European Social Survey

3.1 The European Social Survey

The ESS is an academically-driven social survey designed to chart and explain the attitudes, beliefs and behaviour patterns of Europe´s diverse populations. In parallel with its substantive aims, it aims also to provide a model of best practice in methodology and to contribute towards improvement in methodological standards (further details: www.europeansocialsurvey.org). The ESS is funded via the European Commission's Framework Programmes, with supplementary funds from the European Science Foundation. In each participating nation, the cost of data collection and the appointment of a national co-ordinator (NC) is funded by the national research council or equivalent body. An important principal of the survey is that the data are made freely available: no-one involved in the survey has advance access and there are no restrictions on access. Data can be downloaded from http://ess.nsd.uib.no.

There is a core questionnaire that is administered in every round, along with modules of questions that will change from round to round. Nations are not asked to commit themselves to more than one round at a time, though of course continued participation is encouraged. All interviews are carried out face-to-face. However, after the interviewing a respondent is asked also to fill-in a self-completed supplementary questionnaire (the big part of this questionnaire includes Schwartz' life values) that will be submitted by mail to the country survey organisation (This was not done in Luxembourg and Italy). It should be noted that there is some second phase unit nonresponse since all first-phase respondents have not answered these supplementary questions.

The ESS consists of regular "rounds" of data collection, with each round involving an independent cross-sectional sample in each nation (it is a repeated survey, not a panel). The first

[1] The points relating to the sampling guidelines and conclusions have been made together with the other ESS sampling experts, that is, Sabine Häder (Zuma, Mannheim), Siegfried Gabler (Zuma. Mannheim) and Peter Lynn (Univ. of Essex).

round of field work took place in September-December 2002 (in a few nations fieldwork was not completed until 2003.). Consequently, the interviews for the second round were performed two years later. The third round is now in August 2006 approaching. Table 1 shows which countries have been participated in this survey. Until now, the 30 countries have contributed at least for one round. The 31th country in the list is Turkey for which the sampling design was accepted for round 2 but the data are still missing.

Table 1. Participation of countries in the ESS 2002-2007

Country	*Round 1*	*Round 2*	*Round 3*	*Country*	*Round 1*	*Round 2*	*Round 3*
Austria	Yes	Yes	Yes	Italy	Yes	No	?
Belgium (Flemish)	Yes	Yes	Yes	Luxembourg	Yes	Yes	No
Belgium (Francophone)	Yes	Yes	Yes	Netherlands	Yes	Yes	Yes
Bulgaria	No	No	Yes	Norway	Yes	Yes	Yes
Cyprus	No	No	Yes	Poland	Yes	Yes	Yes
Czech Republic	Yes	Yes	?	Portugal	Yes	Yes	Yes
Denmark	Yes	Yes	Yes	Romania	No	No	Yes
Estonia	No	Yes	Yes	Russia	No	No	Yes
Finland	Yes	Yes	Yes	Slovak Republic	Yes	No	Yes
France	Yes	No	Yes	Slovenia	Yes	Yes	Yes
Germany	Yes	Yes	Yes	Spain	Yes	Yes	Yes
Greece	Yes	Yes	Yes	Sweden	Yes	Yes	yes
Hungary	Yes	Yes	Yes	Switzerland	Yes	Yes	Yes
Iceland	No	Yes	?	Turkey	No	?	?
Ireland	Yes	Yes	Yes	Ukraine	No	Yes	?
Israel	Yes	No	No	United Kingdom	Yes	Yes	Yes

The ESS sampling design group developed the requirements for participating nations – which will be described in the remainder of this section under five broad headings – and then co-operated with participating nations in developing acceptable sample designs.

3.2 Population definition and coverage

The target population for each participating nation is defined as all persons 15 years or older resident in private households within the borders of the nation, regardless of nationality, citizenship, language or legal status. (In countries in which any minority language is spoken as a first language by 5 % or more of the population, the questionnaire must be translated into that language.) It is worth noting in passing that this definition was subject to considerable discussion by a 21-country steering group prior to agreement.

The requirement for sample design was that every person with the defined characteristics should have a non-zero chance of selection. In practice, the quality of available frames – e.g. coverage, updating and access - differs between nations, so careful evaluation of frames was necessary to assess the likely extent of under-coverage and ensure that any coverage bias was likely to be minimal.

Among others, we found the following kinds of frames:

a) nations with reliable lists of *residents* that are available for social research such as the Danish Central Person Register that has approximately 99 % coverage of persons resident in Denmark;
b) nations with reliable lists of *households* that are available for social research such as the "SIPO" database in the Czech Republic, that is estimated to cover 98% of households;
c) nations with reliable lists of *addresses* that are available for social research such as the postal delivery points in the Netherlands and in the UK;
d) nations without reliable and/or available lists such as Portugal, Russia and Greece.

Drawing a sample is most complicated if no lists are available (group d). In this instance area sample designs were usually applied, in which the selection of a probability sample of small geographical areas (e.g. Census enumeration areas within municipalities) preceded a complete field enumeration of households or dwellings within the sampled areas, from which a sample was selected. In nations where this approach was used (e.g. Greece), the sampling panel insisted that the selection stage should be separated from the enumeration and carried out by office staff or supervisors who had not been present for the enumeration. An alternative to area sampling in this situation is the application of random route sampling about which some survey organisations were enthusiastic. The basic idea of random route sampling is that within each sampled psu one address is selected by a random method to serve as a starting point and the interviewer then follows rules that specify the route he or she should take from there, sampling systematically using a pre-specified interval (Häder and Gabler, 2003). The question here, however, is the extent to which random routes can be judged to be "strictly random". That depends on both the rules for the random walk and the control of the interviewers by the fieldwork organisation in order to minimise interviewer influence on selection. A rigorous version of random route sampling was permitted in one country (Austria).

Even in countries where reliable lists exist, some problems had to be solved. For example, in Italy there is an electoral register available. But it contains, of course, only persons 18 years or older (and

only those who are eligible to vote). Therefore, it had to be used as a frame of addresses. This had not been attempted before and there were practical problems to be overcome, not least the fact that persons at the same address do not necessarily appear together on the list, making it difficult to ascertain the selection probabilities of addresses. Thus, under-coverage, while not zero, was restricted to persons at addresses with no registered electors. In countries with population registers, people with illegal status will be excluded because they are not registered. The practical task for the sampling panel was to ensure that levels of under-coverage were kept to an absolute minimum by considering all possible frames and evaluating the properties of each with respect to the ESS population definition.

3.3 Response rates

Non-response is the next problem for achievement of the objective to represent the target population. A carefully drawn sample from a perfect frame can be devalued if non-response leads to systematic bias. Therefore, it is essential to plan and implement adequate field work strategies to minimise non-contacts and refusals. For the ESS a target response rate of 70% was fixed although it was known that this would be challenging for the countries where response rates between 40 and 55 percent are common. Nevertheless, it was felt that a realistic but challenging target should encourage maximum efforts. Additionally, the ESS required that non-contacts should not exceed 3% of eligible sample units, that at least four personal visits must be made to a sample unit before non-contact was accepted as an outcome, and that the field period must last at least 30 days.

As expected, the target response rate was hard to achieve. Table 2 illustrates this from round 2 that also shows the net sample sizes by country. However, some success from round 1 was happened in Czech Republic and Switzerland, in particular.

Table 2. Response rates and realised interviews from round 2 based on the data from April 2006.

	Number of realised interviews	Rate of ineligibles (%)	Response rate (%)	Non-contact rate (%)	Refusal rate (%)
Austria	2256	1.7	62.5	7.8	28.6
Belgium	1778	4.9	61.5	7.1	22.7
Czech Republic	3026	1.3	55.5	n/a	n/a
Denmark	1487	6.4	65.1	5.6	23.9
Estonia	1989	12.7	79.5	5.1	11.4
Finland	2022	1.5	70.8	2.8	21.2

France	1806	7.1	44.2	12.1	39.5
Germany	2870	7.2	52.7	6.2	27.4
Greece	2406	0.1	78.8	3.7	16.4
Hungary	1498	13.5	70.3	6.0	16.0
Iceland	579	5.9	51.3	4.6	39.1
Ireland	2286	8.1	62.5	9.5	22.3
Luxembourg	1635	10.2	52.1	7.7	40.2
Netherlands	1881	3.0	64.5	2.7	28.0
Norway	1760	3.4	66.2	2.1	25.5
Poland	1716	3.8	74.4	2.3	18.2
Portugal	2052	6.4	70.9	2.8	20.0
Slovenia	1442	6.7	70.2	10.2	15.3
Spain	1663	7.8	56.1	13.6	18.6
Sweden	1948	2.3	66.5	4.3	22.6
Switzerland	2141	6.5	47.1	2.9	39.7
United Kingdom	1897	7.9	51.1	8.0	34.0

3.4 Sample selection methods

We have already argued that strict probability sampling is a necessary pre-requisite for cross-national comparability. However, partly as a measure to overcome the fear of non-response bias, many survey organisations habitually implement substitution of non-cooperative or not reachable primary sampling units, households or target persons by others. There are many varieties of substitution (Vehovar, 2003; Lynn, 2004), but none of them meet the requirement for probability sampling. Another important disadvantage of substitution in the field is that it can reduce the effort made by interviewers to gain a response at the original addresses/households.

For the ESS, substitution of non-responding households or individuals (whether 'refusals' or 'non-contacts') was not permitted in any circumstances. However, in exceptional circumstances substitution was permitted at the first stage of sampling. Administrative considerations may mean that addresses cannot be obtained for specific sampled areas (e.g. a particular municipality may refuse to grant access to the list, or be unable to co-operate within the available time).

3.5 Effective sample size

The ESS requirement was for a minimum estimated effective sample size of 1,500 completed interviews and a minimum of 2,000 actual interviews. (An exception was made for nations with a total population of less than 2 million persons, recognising that resources for funding surveys are considerably constrained in such nations. For such nations, the minimum requirement was an effective sample size of 800 and an actual sample size of 1,000.) Explanation was provided as to what was meant by *effective sample* size and how it should be predicted. This involved predicting, under certain simplifying assumptions, the design effect due to unequal selection probabilities (*DEFFp*) and the design effect due to clustering ($DEFF_C$). Additionally, realistic estimates of response rate and eligibility rate were required in order to calculate the sample size to select in order to produce the target number of completed interviews.

A reasonable approach to sample size determination is to predict the determinants of design effects within reasonable bounds. The aspects of the survey that make this possible are, 1) relatively low – and relatively stable over time - expected correlation between survey variables and psu's; 2) relatively small variation in selection probabilities; 3) prior estimates in several countries for similar variables on surveys with similar designs. Additionally, a repeating survey like ESS offers the opportunity to revise predictions at each round based on estimates from previous rounds.

3.5.1 Design effect due to unequal selection probabilities (DEFFp)

The ESS guidelines suggested that *DEFFp* should be predicted as follows:

$$\tilde{DEFF}_P = \frac{m\sum_{i=1}^{I} m_i \left(w_i^2\right)}{\left(\sum_{i=1}^{I} m_i w_i\right)^2} \qquad (1)$$

where m_i and w_i denote respectively the number of interviews and the design weight associated with the i^{th} weighting class. (This can be expressed equivalently as $1+cv_w^2$, where cv_w is the coefficient of variation of the weights)

In some nations, it is necessary to select the sample in stages, with the penultimate stage being addresses or households. In this case, each person's selection probability depends on the respective household size. The guidelines illustrated estimation of (1) with a hypothetical example of an address-based design of this sort, where the weighting classes were defined by the possible values of number of persons aged 15 or over resident at an address. Several nations use such an address-based design (e.g. Czech Republic, Greece, Ireland, Israel, Netherlands, Portugal, Russia, Spain, Switzerland, UK).

Another reason for unequal selection probabilities is that minority groups are over-sampled for substantive reasons. A third reason is that certain strata (typically, the largest cities) may be over-

sampled in anticipation of lower response rates, though in principle this should not affect variance of estimates as it will lead to equal inclusion probabilities if the response rate predictions turn out to be accurate.

A fourth source of variation in selection probabilities occurs in countries where the psu's are selected with probability proportional to a proxy size measure which does not correlate perfectly with the units sampled at the subsequent stage.

3.5.2 Design effect due to clustering ($DEFF_C$)

The cluster sample size and the intra-class correlation also influence the design effect. Following Kish (1987), the ESS guidelines suggested that $DEFF_C$ should be predicted as follows:

$$\tilde{DEFF}_C = 1 + \left(\bar{b} - 1\right)\rho \qquad (2)$$

where $\bar{b}$ is the mean number of interviews per cluster and ρ is the intra-cluster correlation. Expression (2) implies that, were cost not a consideration, the cluster sample size should be chosen as small as possible. The larger the average cluster size, the more interviews have to be conducted to reach the minimum effective sample size. The challenge, therefore, is to find the combination of $\bar{b}$ and *n* that delivers the desired effective sample size for the lowest cost. Participating nations were encouraged to seek estimates of ρ from other surveys in their country if possible, or alternatively to assume $\rho = 0.02$. In practice, ρ will take different values for different statistics and can also vary between subgroups for any particular statistic, but the ESS sample design requirements were stated only in terms of the total sample and only in terms of a "typical" ρ. Considerable variation in $\tilde{DEFF}_C$ was observed, primarily because of the variation in proposed cluster sample size.

3.5.3 Combined design effect

The ESS guidelines suggested that the total design effect should be predicted as:

$$\tilde{DEFF} = \tilde{DEFF}_P \times \tilde{DEFF}_C \qquad (3)$$

This ignores any design effect due to stratification of the sampling frame, but as this is generally modest in magnitude and beneficial in direction (i.e. less than one), ignoring this effect was felt to both simplify the calculation and build in a little conservatism to the required sampled size. Expression (3) also assumes no association between the weights and the clusters – see Lynn and Gabler (2005). Predictions of total design effect vary greatly between nations.

3.6 Summary of sampling procedure

When taking into account all the effects the gross sample size can be anticipated as Table 3 illustrates.

Table 3. Illustrative example of all factors related to anticipate an ideal gross sample size.

Operation	**Size calculation**
1. Target effective sample size - *neff* (size that can be received with *srs* without missingness).	**1500**
2. Anticipated missingness due to nonresponse (on average , may vary by strata, e.g.)	30% eli 1500/.7 = 2143
3. Anticipated missingness due to overcoverage (on average)	5% eli 2143/.95 = 2256
4. Anticipated cluster effect so that the final cluster size has been anticipated too * and intra-cluster correlation based on earlier experience on similar surveys	$DEFF_c$ = 1+(5.3- 1)*.025 = 1.108 2256*1.108= 2499
5. Anticipated design effect due to unequal inclusion probabilities used in the design*	$DEFF_p$ = 1.25 2499*1.25 = 3125
6. Anticipated risk in fieldwork and then we have the gross sample size (here net sample size = 3150*.7*.95 = 2095)	**3150**

* *should be consistent with figures in points 2 and 3*

3.7. Documentation

Comprehensive and clear documentation of all relevant methodological aspects of the survey was demanded. At the level of sampling units, this meant that indicators of sampling stratum, primary sampling unit and the selection probability at each stage of sampling should be included on a micro-level data file that carried the same identifiers as the questionnaire and other data files. A detailed file specification was provided. Supply of these data would allow the application of weights and the use of appropriate methods for the analysis of data from a complex survey.

A problem of the current procedure is that nonresponse has not been well taken into account. So, any adjusted weights due to this reason are not available on the web. It would first require to include more auxiliary data for the sampling file, both of respondents and non-respondents. This is very realistic for many countries but not for all. It is obviously the main reason that the central co-ordinating team (CCT) of the ESS has not required these operations even although I have proposed it.

4. Evaluation of the ESS Procedures

4.1 Predictions of DEFF

As already mentioned the ESS sampling system has not yet taken completely into account nonresponse. However, this present system as illustrated in Table 3 uses numbers of respondents when calculating the basic weights that are called design weights (variable DWEIGHT in the freely available web data file that are scaled so that the average for each country is equal to one).

The basic weights vary in all other countries except in those which have applied simple random sampling (three countries in round 1 and seven in round 2). This thus assumes that nonresponse is non-informative that does not hold, of course. On the other hand, the total DEFF is equal to one in these countries. In the case of the other countries, we have tried to analyse the DEFF's in order to improve the sampling procedure for the subsequent rounds. Since the DEFF's are variable-dependent we created nine variables from the round 1 file so that the different characteristics of the questionnaire were taken into account. Most variables were constructed from several initial variables, being thus like indicators. Table 4 gives some results on the DEFF's.

Table 4: Estimation of design effects for countries participating in both rounds

Country	Median ρ	max b^*	*DEFFc*	*DEFFp*	*DEFF*
AT	0.11	6.49	1.61	1.24	2.01
BE	0.04	6.56	1.22	1	1.22
CH	0.03	8.83	1.27	1.21	1.54
CZ	0.15	2.94	1.28	1.25	1.61
DE	0.06	18.85	2.03	1.11	2.26
ES	0.15	4.96	1.60	1.22	1.95
FR	0.05	7.42	1.34	1.23	1.65
UK	0.03	12.06	1.40	1.22	1.69
HU	0.05	8.68	1.36	1	1.36
NL	-	-	1	1.19	1.19
NO	0.01	30.03	1.41	1.43	2.03
PL	0.05	10.07	1.32	1.02	1.35
PT	0.14	5.07	1.57	1.83	2.88
SI	0.03	10.76	1.33	1	1.33

We see for example that the median intra-class correlation was in most countries higher than the initial minimum recommendation = 0.02 (this was given based on the UK experience and used due to the fact that many countries has no idea how high this correlation could be). Table 4 also shows that the average cluster size varies quite much. It was in most countries set to be initially quite constant but because the response rates may be varied substantially between clusters (psu's) some variability was present.

In most cases, the achieved response rates were lower than the predictions, but in some they were higher. The greatest proportionate under-prediction (round 1) was in Greece ($\tilde{\bar{b}} = 4.8; \bar{b} = 5.9$), while the greatest over-prediction was in Italy ($\tilde{\bar{b}} = 18.0; \bar{b} = 11.0$), followed by France ($\tilde{\bar{b}} = 12.0; \bar{b} = 8.9$). Where the response rate was less than predicted, this was not necessarily due to a failure to meet the ESS minimum requirements regarding contact efforts or indeed due to lack of efforts generally (Philippens and Billiet 2006).

Differences between the predicted and estimated values of *DEFF* are non-existent in some cases, but considerable in others. There was some uncertainty regarding the parameters of clustering, design weights, or both. In five nations in round 1, the uncertainty only concerned $\bar{b}$. These were three nations (BE, HU, SI) with an equal-probability sample selected from population registers and two (DE, PL) where the weights were completely determined by the sample design. The prediction turned out accurate in Belgium. In Slovenia, $\bar{b}$ was under-estimated as both the eligibility rate and response rate turned out higher than predicted. These two rates were also both under-estimated in Hungary, but this was more than compensated for by an increase in the number of psu's (and associated reduction in the selected cluster sample size), subsequent to the prediction made on the sign-off form.

Due to too optimistic anticipated DEFFs and response rates many countries failed to achieve the minimum requirement for the effective sample size in both rounds. This thus means that the confidence intervals are not to be very close to each other, and a user has to be careful with cross-country comparisons. Some countries made an improvement in round 2 either increasing the number of psu's, or increasing gross sample size or fighting better against nonresponse. It is unclear whether this tendency will continue in round 3 since more and more countries have met budgetary problems. Recently, it was also discussed whether it is necessary to lower the ESS sampling requirements, in the terms of effective sample size.

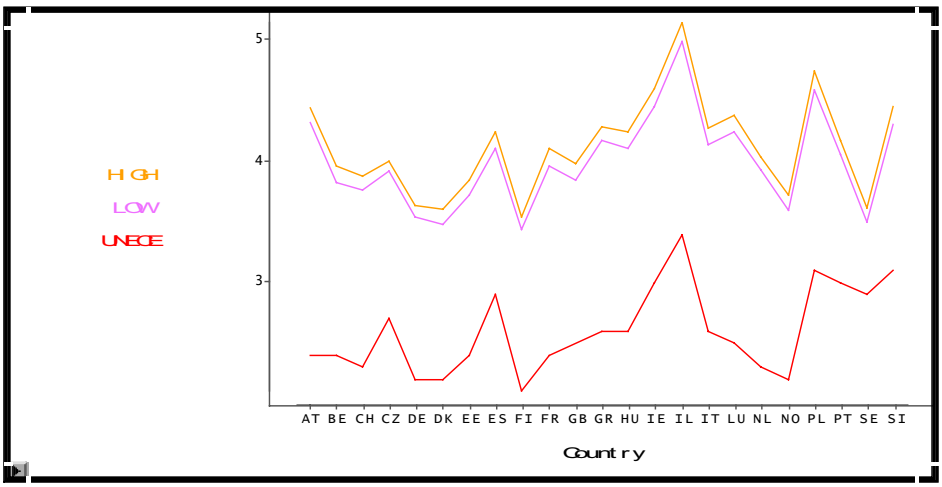
4.2 Need for nonresponse adjustments

Nonresponse has not taken into account enough well although its effect has been analysed in various reports. A problem is that there are no much individual-level data collected on nonrespondents in any country. It is possible to evaluate this problem indirectly, using outside-aggregate data, for example. I

performed one exercise which exploits the data from the United Nations web on one hand and the ESS micro data, on the other; see Figure 1.

Although it is not guaranteed that the UNECE data are complete, it is however rather correct. We see clearly that small households are much less represented in the ESS samples. It is not easy to see, why? Sampling procedure may be one reason but I think that the main reason is that single households have not contacted well and hence they have responded worsely than members of larger households. This is usual in most surveys, why not in the ESS. Such a bias could be reduced in field work to some extent but more using nonresponse adjustments by collecting more auxiliary data of non-respondents, and then constructing the adjusted weights (e.g. using the methodology presented in Laaksonen & Chambers 2006).

Figure 1. Average household size by country based on the UNECE data from early 2000 and from the ESS micro data from 2004-2005. HIGH = higher 95% confidence interval, LOW = lower 95% confidence interval.



5. Analysis of ESS data – and happiness example

The ESS micro files are easy to download and then to analyse. The web also includes rather good meta data and some para data derived from interviewing. Nevertheless, since the variables are coded so that all collected information is included in the files, a user has to be careful especially with such codes that give information about missingness which can be of different kinds. Hence a user always needs to make some refinements because starting real analysis.

The files also include the two types of weights, the ones based on the sampling design of each country and the others indicating the size of the target population of each country. A user thus has to select the first weights always and sometimes both weights. Nonresponse adjusted

weights are thus missing. When using the available weights in the analysis, the point estimates will be correct but it is not reasonable in the case of interval estimates. Some idea of the impact on interval estimates can be got from Table 4 but naturally it is best if a user can estimate the corresponding estimates him/herself since the DEFF's depend on each particular case. Currently, a big problem is that the information of psu's is not available on the web files. Obviously, this will be the case later. If you are using these files, ask for the data of psu's.

To illustrate a bit which types of analysis from the ESS files can be provided, I present some results. The variable of interest is called HAPPY and measured as follows:

Taking all things together, how happy would you say you are? Please use this card.

Extremely unhappy										Extremely happy	(Don't know)
00	01	02	03	04	05	06	07	08	09	10	88

I exclude the individuals with values '88' from the analysis and also those with other missing values (coded afterwards as '99'). The results in Figure 2 are direct averages for each country, however so that the reference country, Ukraine, has been set equal to 0. The people in the other countries are happier than Ukrainians, on average.

The standard errors are not in Figure 2 but these are rather even over countries without taking into account cluster effects. The standard errors are around 5.4-6.4 percent in most countries except Iceland (8.8 %). The cluster effect will increase this number to some extent. Since I had the German cluster data, I calculated the DEFF that was = 1.76.

I also looked the changes in happiness from round 1 to round 2. In most countries the change was not significant but it is interesting that in the following countries the average happiness was increased significantly: Hungary, Poland and Slovenia. All these joined the EU after round 1. Is it the explanation? On the other hand, some decrease was happened in Portugal, in particular, and also in Spain. I cannot explain these results although we have found that the quality of the Portuguese data seems to be weakened from round 1 to round 2.

Much further analysis can be performed. I briefly present some results in order to explain happiness/unhappiness. These are based on ordinary multivariate regression analysis using country weights. A number of explanatory variables were included in the model. I do here present only some results:

- Women are happier than men, and young people happier than older. On the other hand, the happiness of women decreases quite linearly by age but men are least happy at middle ages.

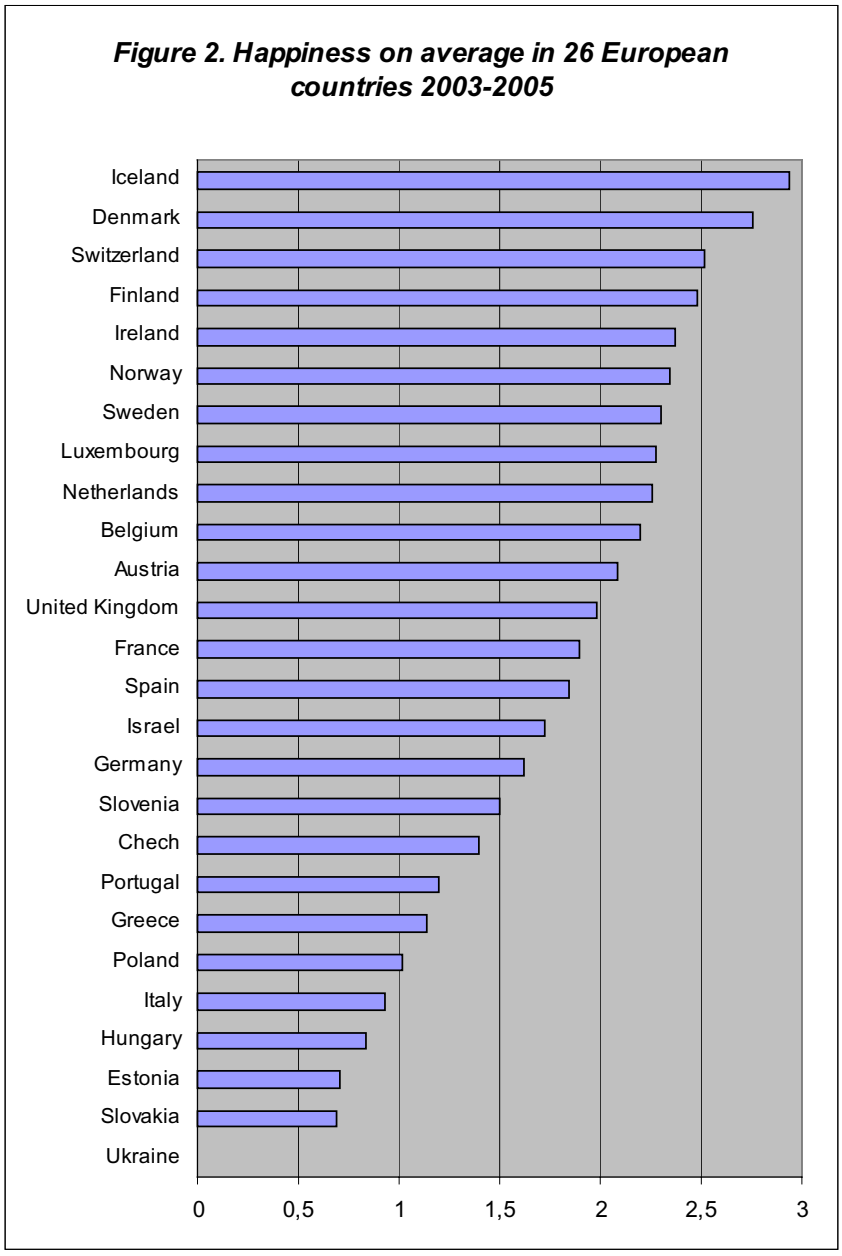


Figure 2. Happiness on average in 26 European countries 2003-2005

- Married are happier than never married who are happier than divorced and separated.
- Big household increases happiness to some extent.
- Trust on police and legal issues in the country are good for people's happiness. The same is concerned trust on administration including health organisations.
- People who feel to be discriminated by gender, race etc. are less happy.
- Very poor people seem to be least happy but there is not much difference if the income level is over some minimum.
- Bad health naturally decreases happiness.
- Active people are slightly happier than inactive.

- Foreigners are slightly less happy than native people.

The model includes also the country as one explanatory variable. Naturally, the differences between countries were reduced essentially after this modeling. The happiness order also changed to some extent. However, Iceland was still in the top, and Ireland the second before Denmark, but now Italy was the last, Slovakia the second last. Can a reader explain these?

6. Conclusion

The aims of the ESS, in terms of sample design standards and procedures for implementation of those standards, were ambitious. Though not realised in every detail, the ESS can be considered a great success. This is evidenced also so that the ESS (with subtitle "Innovations in comparative measurement") was one of the five winners of the 2005 EU Descartes Laureates (see http://www.sardinien.com/astronomie/pdf/pr02122005_annex_winners_dp_research2005_en.pdf). In particular, the process for developing and finalising sample designs can be considered successful both at a subjective level and in objective terms (guidelines used to estimate design parameters proved useful and estimates generally accurate; documentation is relatively complete).

In my opinion, the quality of the ESS is one of the best ones in the world if the demanding multinational surveys are concerned. This does not mean that the quality cannot be essentially improved. The evaluation of the estimation of design parameters presented here has provided several pointers to how such estimation might be improved on future cross-national surveys. The nature of uncertainty in the estimates has been described and the directions of errors documented.

In general, the ESS has provided advances in survey practice in a number of nations. Additionally, the procedures for sample design represent a useful advance in the methodology of cross-national surveys.

Oversampling has been used in some countries and also so that the anticipated differences in nonresponse/overcoverage between regions have been taken into account. But this could be exploited much more, also in *srs*-countries where it is well-known that response rates vary much by region and other domains. So, pre-stratification would be my recommendation for these countries too, and consequently leading to varying weights if the anticipation is not complete. Furthermore, I recommend to insert the new adjusted sampling weights into the ESS archive data files in addition to the current design weights (DWEIGHT).

In the first stage these weights should be required for the *srs* countries that have always the weights equal to one in the current integrated archive file. This is not even difficult since these countries have already created such weights, based on post-stratification or other calibration. Later, we should require all countries to add information for nonresponse analysis and adjustments. For example, all countries are able to add to a sampling file some variables of nonrespondents.

References

Gabler, S., Häder, S., Lahiri, P. (1999). A model based justification of Kish's formula for design effects for weighting and clustering. *Survey Methodology*, 25:1, 105-106.

Gabler, S., Häder, S., Lynn, P. (2005). Design effects for multiple design samples, ISER Working Paper No. 2005–12, Colchester: University of Essex. http://www.iser.essex.ac.uk/pubs/workpaps/pdf/2005-12.pdf

Häder, S. and Gabler, S. (2003). Sampling and estimation. In *Cross Cultural Survey Methods*, eds. J. Harkness, F. van de Vijver, P. Mohler, New York: John Wiley and Sons.

Kish, L. (1992). Weighting for unequal P_i. *Journal of Official Statistics*, 8:2, 183-200.

Kish, L. (1994). Multipopulation survey designs: five types with seven shared aspects. International Statistical Review, 62, 167-186.

Kish, L. (1995). Methods for design effects. *Journal of Official Statistics*, 11:1, 55-77.

Laaksonen, S. & Chambers, R. (2006). Survey estimation under informative non-response with follow-up. *Journal of Official Statistics, 81-95*.

Le, T. and Verma, V. (1997). An analysis of sampling designs and sampling errors of the Demographic and Health Surveys, DHS Analytic Report No.3, Calverton, Maryland: Macro International Ltd. http://www.measuredhs.com/pubs/details.cfm?ID=4

Lynn, P. (2003). Developing quality standards for cross-national survey research: five approaches. *International Journal of Social Research Methodology*, 6:4, 323-336.

Lynn, P. (2004). The use of substitution in surveys. *The Survey Statistician*, 49, 14-16.

Lynn, P. and Gabler, S. (2005). Approximations to b^* in the prediction of design effects due to clustering. *Survey Methodology*, 31:1, 101-104.

Lynn, P., Gabler, S. & Häder, S. & Laaksonen, S. (2006). Methods for achieving equivalence of samples in gross-national surveys. *Journal of Official Statistics*. (in print)

Philippens, M. and Billiet, J. (2006). Nonresponse in cross-national surveys. Results of the European Social Survey. *ESS Working Paper.*

Vehovar, V. (2003). Field substitutions redefined. *The Survey Statistician*, 48, 35-37.

THE ROLE OF MODELS IN MODEL-ASSISTED AND MODEL-DEPENDENT ESTIMATION FOR DOMAINS AND SMALL AREAS

Risto Lehtonen[1]

[1] University of Helsinki, Finland
e-mail: risto.lehtonen@helsinki.fi

Abstract

Estimation for population subgroups or domains is investigated for model-assisted generalized regression (GREG) and model-dependent EBLUP estimators, under different model choices and under unequal probability sampling. Two particular issues are addressed: (i) how to account for the domain differences in the model formulation, and (ii) how to account for the underlying unequal probability sampling design. Results on bias and accuracy of GREG and EBLUP are based on Monte Carlo experiments where PPS samples were drawn from an artificially generated population. The bias of GREG estimator remained negligible for all model formulations considered, and accuracy improved when including the PPS size variable in the assisting model. A "double-use" of the auxiliary data both in the sampling design and in the estimation design appeared favorable. In GREG, the mixed model formulation did not outperform the fixed-effects model formulation. For EBLUP, the model choice was critical and if not successful, large bias was introduced. For unweighted EBLUP, substantial bias reduction was attained with the inclusion of the PPS size variable in the model. We propose a new weighted EBLUP estimator for unequal probability sampling designs, as an alternative to the unweighted EBLUP. The results show that the weighted EBLUP behaves better that the unweighted EBLUP, but still the bias can be substantial and can dominate the MSE, which invalidates the construction of proper confidence intervals.

Acknowledgement. This working paper is joint work with Prof. Carl-Erik Särndal of University of Montreal, Dr Ari Veijanen of Statistics Finland and Mr Mikko Myrskylä of University of Helsinki.

1 Introduction

Estimation of reliable statistics for population subgroups or domains constitutes an area of increasing importance in the production of official statistics. A good example is the estimation of the number of unemployed and employed, and the accompanying standard errors, for regional areas in a country by using sample survey data from a Labour Force Survey and auxiliary data taken from the available register and census data sources. Typically, a LFS is planned to produce reliable statistics for the entire population and large or major areas. Standard design-based direct estimators, such as the Horvitz-Thompson estimator, are often used for such cases. The task can become challenging when the number of sample elements in a number of domains remains small or minor. In this case, more advanced methods that effectively use the available auxiliary information are needed.

Methods available for the estimation of totals for domains and small areas include model-assisted design-based estimators, referring to the family of generalized regression (GREG) estimators (Särndal, Swensson and Wretman 1992, Estevao and Särndal 1999, 2004), and model-dependent techniques, such as the EBLUP estimator (Empirical Best Linear Unbiased Predictor) and synthetic estimators (Ghosh 2001, Rao 2003). Properties of these estimator types are discussed for example in Lehtonen and Veijanen (1998, 1999) and Lehtonen, Veijanen and Särndal (2003, 2005). The documentation of the EURAREA project includes use-

ful comparative materials on properties of model-dependent estimators (EURAREA Consortium 2004, Heady and Ralphs 2005).

Known design-based properties related to bias, precision and accuracy of model-assisted estimators and model-dependent estimators are summarized in Table 1. Model-assisted estimators are approximately design-unbiased by definition, but their variance can become large in domains where the sample size is small. Model-dependent estimators are design-biased: the bias can be large for domains where the model does not fit well. The variance of a model-dependent estimator can be small even for small domains, but the accuracy tends to be poor because the squared bias often dominates the mean squared error (MSE), as shown for example by Lehtonen, Veijanen and Särndal (2003 and 2005). The dominance of the bias component together with a small variance can cause poor coverage rates and invalid confidence intervals for a model-dependent estimator. For model-assisted design-based estimators, on the other hand, valid confidence intervals can be constructed. Typically, model-assisted estimators are used for major or not-so-small domains and model-dependent estimators are used for small domains where model-assisted estimators can fail.

Table 1. Design-based properties of model-assisted and model-dependent estimators for domains and small areas.

	Design-based model-assisted methods - GREG family	**Model-dependent methods SYN and EBLUP**
Design bias	Design unbiased (approximately) by the construction principle	Design biased Bias does not necessarily approach zero with increasing domain sample size
Precision (Variance)	Variance may be large for small domains Variance tends to decrease with increasing domain sample size	Variance can be small even for small domains Variance tends to decrease with increasing domain sample size
Accuracy (Mean Squared Error, MSE)	MSE = Variance (or nearly so)	MSE = Variance + squared Bias Accuracy can be poor if the bias is substantial
Confidence intervals	Valid intervals can be constructed	Valid intervals not necessarily obtained

Survey statistician often faces challenging methodological choices when aiming at reliable estimation of population totals for domains and small areas. These choices include, for example, the inferential framework, model type (mathematical form, specification, parametrization, estimation of model parameters), and estimator type (point estimator, estimator of variance or MSE) for the unknown domain totals. Related to the problem of model choice, or the role of the model in model-assisted estimators and in model-dependent estimators, the two questions of special interest in this study are:

(i) How to account for the domain differences in the model formulation (relevant for model-assisted estimators in particular)?

(ii) How to account for the underlying unequal probability sampling design (relevant for model-dependent estimators in particular)?

We discuss points (i) and (ii) to some extent from a design-based perspective, under the fixed finite population approach. More specifically, we compare the relative performance (bias and accuracy) of the two estimator types of domain totals, GREG, and EBLUP, under different model choices. A continuous response variable is assumed. In the construction of models we

use both linear fixed-effects models and linear mixed models, where random effects are included in addition to the fixed effects. We fit the linear models with different parametrizations. In the estimation of the model parameters, we use both weighted and unweighted estimation procedures.

An underlying unequal probability sampling design is assumed. The case of unequal probability sampling is of importance for practical purposes in official statistics and many fields of empirical research. Without-replacement type fixed-size Probability Proportional to Size sampling (systematic PPS) was selected to represent an example of an unequal probability sampling design. This study extends the case of equal probability sampling investigated in Lehtonen, Särndal and Veijanen (2003, 2005) to unequal probability sampling designs.

The working paper is organized as follows. Chapter 2 introduces our notation and models and estimators used. Results for GREG and EBLUP estimators are given in Chapter 3. Conclusions are in Chapter 4.

2 Methods

2.1 Models and estimators of domain totals

We are interested in the estimation of totals of a continuous response variable *y* for the domains of interest. Availability of powerful auxiliary information is essential for the estimators of domain totals considered. We assume that we have access to unit-level data, which include domain membership indicators and vectors of auxiliary x-variables, for all units in the population. The auxiliary data vector also contains the size variable used in the PPS sampling procedure. The auxiliary data are incorporated in the estimation procedure by an appropriate model. Thus, the choice of the model that underlies the GREG, SYN and EBLUP estimators of domain totals is considered important.

Our question (i) was "How to account for the domain differences in the model formulation?". The domain differences can be accounted for by a proper model formulation. Basically, there are two options to facilitate the domain differences: (1) introduction of domain-specific fixed effects in the model, and (2) accounting for the domain differences by domain-specific random effects, such as random intercepts. It is obvious that these options are relevant for model-assisted estimators in particular. The reason is that in a standard GREG setting, a fixed-effects linear model is routinely used as the assisting model (Estevao and Särndal 1999, 2004), and a GREG estimator that uses a mixed model, the MGREG estimator, has been introduced only recently (Lehtonen and Veijanen 1999, Lehtonen et al. 2003, see also Goldstein 2003, p. 165). On the other hand, a mixed model formulation has a long tradition in the context of EBLUP estimation of small area totals (Fay and Herriot 1979, Rao 2003). The problem of model choice is discussed in a more general spirit in Firth and Bennett (1998).

To throw some light on question (ii) "How to account for the underlying unequal probability sampling design?", we study the different options to incorporate the information of the sampling design into the estimation procedure. In the modelling phase, there are two main options to account for the sampling design: (a) the incorporation of sampling weights in the estimation of model parameters, and (b) the inclusion of sampling design variables as additional covariates in the model. By default, sampling weights are incorporated in the estimation procedures for all assisting models of GREG estimators. As a rule, sampling weights are ignored in the estimation procedures for SYN estimators.

Typically, the underlying mixed model of a standard EBLUP estimator is fitted in an unweighted manner. Rao (2003) introduced a pseudo EBUP estimator, where sampling weights are included in the construction of the EBLUP estimator, but the parameters of the mixed model are estimated by unweighted techniques. As an alternative to the unweighted EBLUP

and pseudo EBLUP, we will introduce a new EBLUP estimator, where sampling weights are incorporated in the estimation of parameters of the underlying mixed model. We will also compare options (a) and (b) in their successfulness in accounting for the sampling design. It is obvious that these options are relevant for EBLUP estimators in particular.

We study the bias and accuracy properties of the estimators of domain totals by empirical methods. Our Monte Carlo simulation experiments consisted of repeated draws of systematic PPS samples from an artificially constructed fixed finite population.

Table 2 shows the model-dependent and model-assisted estimators to be discussed, in a two-way arrangement by estimator type and by model choice. Each of the rows corresponds to a different model choice. CC model (common intercepts, common slopes) is one whose only parameters are fixed effects defined at the population level; it contains no domain specific parameters. We obtain SYN-CC and GREG-CC estimators. SC model (separate intercepts, common slopes) is one having at least some of its parameters or effects defined at the domain level. These are fixed effects for SYN-SC and GREG-SC and random effects for EBLUP-SC, EBLUPW-SC and MGREG-SC. Table 2 also shows the estimation methods that are used in the estimation of model parameters.

To address points (i) and (ii) of Chapter 1, we discuss in more detail GREG-SC and MGREG-SC for GREG family estimators and EBLUP-SC and EBLUPW-SC for EBLUP family estimators.

Table 2. Schematic presentation of the model-dependent and model-assisted estimators of domain totals for a continuous response variable by model choice and estimator type, under unequal probability sampling.

Model choice				***Estimator type***	
Model abbreviation	**Model specification**	**Effect type**	**Estimation of model parameters**	***Model-dependent estimators***	***Model-assisted estimators***
CC	Common intercepts Common slopes	Fixed effects	OLS	SYN-CC	Not applicable(**)
			WLS	Not applicable(*)	GREG-CC
SC	Separate intercepts Common slopes	Fixed effects	OLS	SYN-SC	Not applicable (**)
			WLS	Not applicable(*)	GREG-SC
		Fixed and random	REML GLS	EBLUP-SC	Not applicable (**)
			Weighted REML GWLS	EBLUPW-SC	MGREG-SC

OLS Ordinary least squares
WLS Weighted least squares (sampling weights)
GLS Generalized least squares
GWLS Generalized weighted least squares (sampling weights)
REML Restricted (residual) maximum likelihood
Weighted REML Restricted pseudo maximum likelihood (sampling weights)

(*) In SYN, weights are ignored in the estimation procedure by default.
(**) In GREG, weights are incorporated in the estimation procedure by default.

We next introduce the notation used in this study.

Population and sampling design

$U = \{1, 2, ..., k, ..., N\}$ Population (fixed, finite)

$U_1, ..., U_d, ..., U_D$ Domains of interest (non-overlapping)

$Y_d = \sum_{U_d} y_k$, $d = 1, ..., D$ Target parameters (domain totals)

$\mathbf{x}_k = (x_{1k}, ..., x_{pk})'$ Auxiliary variable vector

$I_{dk} = 1$ if $k \in U_d$ Domain membership indicators,

$I_{dk} = 0$ otherwise $d = 1, ..., D$

Note that we assume the vector value $\mathbf{x}_k$ and domain membership to be known for every population unit $k \in U$.

Sampling design: Systematic PPS with sample size n

s Sample from U

$s_d = s \cap U_d$ Random part of s falling in domain d

$\pi_k = n \frac{x_{1k}}{\sum_{k \in U} x_{1k}}$ Inclusion probability for $k \in U$

$a_k = 1/\pi_k$ Sampling weight for $k \in s$

We observe y_k for $k \in s$. Note that for estimation purposes, sample data and auxiliary data are merged at the micro level by using unique ID keys that are available in both data sources.

Models for continuous response *y*

Linear fixed-effects models

CC models $y_k = \beta_0 + \beta_1 x_{1k} + ... + \beta_p x_{pk} + \varepsilon_k$

SC models $y_k = \beta_{0d} + \beta_1 x_{1k} + ... + \beta_p x_{pk} + \varepsilon_k$, $d = 1, ..., D$

Fitted values under fixed-effects models $\hat{y}_k = \mathbf{x}_k' \hat{\boldsymbol{\beta}}$

Linear mixed models

SC models $y_k = \beta_0 + u_d + \beta_1 x_{1k} + ... + \beta_p x_{pk} + \varepsilon_k$, $d = 1, ..., D$

where u_d are domain-specific random intercepts

Fitted values under mixed models $\hat{y}_k = \mathbf{x}_k' \hat{\boldsymbol{\beta}} + \hat{u}_d$, $d = 1, ..., D$

Note that fitted values $\hat{y}_k$ are calculated for every $k \in U$.

Estimators of domain totals

The predictions $\{\hat{y}_k ;\ k \in U\}$ differ from one model specification to another. For a given model specification, the estimator of the domain total $Y_d = \sum_{U_d} y_k$ has the following structure for the three estimator types (SYN, GREG, EBLUP):

Model-assisted GREG estimators

$$\hat{Y}_{dGREG} = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} a_k (y_k - \hat{y}_k)$$

Model-dependent SYN estimators

$$\hat{Y}_{dSYN} = \sum_{k \in U_d} \hat{y}_k$$

Model-dependent EBLUP estimators

$$\hat{Y}_{dEBLUP} = \sum_{k \in s_d} y_k + \sum_{k \in U_d - s_d} \hat{y}_k$$

where $d = 1,...,D$.

Note that $\hat{Y}_{dSYN}$ and $\hat{Y}_{dEBLUP}$ rely heavily on the truth of the model, and can be biased if the model is misspecified. On the other hand, $\hat{Y}_{dGREG}$ has a second term that protects against model misspecification.

We adopt the following conventions (Table 2). In SYN-CC, SYN-SC, GREG-CC and GREG-SC, a fixed-effects model formulation is assumed. A mixed model is assigned for EBLUP-SC, EBLUPW-SC and MGREG-SC estimators.

Measures used in Monte Carlo simulations

In Monte Carlo simulation experiments, by using estimates $\hat{Y}_d(s_v)$ from repeated samples s_v; $v = 1,2,...,K$, we computed for each domain $d = 1,...,D$ the following Monte Carlo summary measures of bias and accuracy.

(i) Absolute relative bias (ARB), defined as the ratio of the absolute value of bias to the true value:

$$\left| \frac{1}{K} \sum_{v=1}^{K} \hat{Y}_d(s_v) - Y_d \right| / Y_d$$

(ii) Relative root mean squared error (RRMSE), defined as the ratio of the root MSE to the true value:

$$\sqrt{\frac{1}{K} \sum_{v=1}^{K} (\hat{Y}_d(s_v) - Y_d)^2} \; / Y_d$$

Details of the simulations

There were 100 domains in the population. The size of domain d was proportional to $\exp(q_d)$, where q_d was simulated from U(0,2.9). Each observation was allocated to a domain by geometric probability: intervals of length $\exp(q_d)$ were concatenated and a random point was chosen in $(0, \sum_d \exp(q_d))$. The interval containing the point determined the domain of the observation.

There were 47 minor domains, 19 medium-sized domains and 34 major domains in the population. These three classes were defined on the basis of expected sample size $n(N_d / N)$: less than 70, 70-119 and 120 or more units, respectively. The smallest domain of the generated population had 1,711 units and the largest had 28,296.

The variable x_1 is the size variable used in PPS sampling. The variable was simulated from uniform distribution U(1,11). Another auxiliary variable x_2 was simulated from N(0,9). The random effects u_d were simulated independently from N(0,0.25). The error term ε followed N(0,1).

Responses were simulated as

$$y_k = 1 + 2x_{1k} + 1.5x_{2k} + u_d + \varepsilon_k \ (k \in d)$$

Correlations of the variables in the population were: $corr(y, x_1) = 0.779$, $corr(y, x_2) = 0.607$ and $corr(x_1, x_2) = -0.001$. Domain means of the response variable were approximately equal, but the totals differed considerably: The means of domain totals were 45,614 for minor domains, 117,308 for medium domains and 241,527 for major domains.

Our population size is N = 1,000,000 and sample size n = 10,000. In Monte Carlo experiments, K = 1000 independent systematic PPS samples were generated. The inclusion probabilities are $\pi_k = nx_{1k} / \sum_k x_{1k}$. The weights $a_k = 1/\pi_k$ varied between 54.6 and 596.5.

3 Results

3.1 GREG estimators

We first discuss results for GREG estimators. Our point (i) devoted to GREG was "How to account for the domain differences in the model formulation". This is demonstrated by the eight different model formulations in Table 3. In models A1, B1, C1 and D1, the domain differences are accounted for by domain-specific fixed effects β_{0d}. In models A2, B2, C2 and D2, we use random intercepts $\beta_0 + u_d$, where β_0 is the fixed intercept common for all domains, and the random term u_d is domain-specific. In addition, we have two explanatory variables at our disposal: the variable x_1, which was used in the PPS sampling design, and x_2, which is an auxiliary variable uncorrelated to x_1. Note that both variables correlate quite strongly with the response variable *y*. For x_1 and x_2, slope parameters β_1 and β_2 are common fixed effects for all domains.

For GREG, we incorporate the sampling weights in the estimation procedure of model parameters, including the mixed model underlying the MGREG-SC estimator. This facilitates the condition of "internal bias calibration" (a proper combination of model formulation and estimation procedure under a given sampling design) proposed by Firth and Bennett (1998).

Table 3 also shows our model building strategy. We start with simple models A1 and A2 and proceed step by step towards the population generating model D2. In all models considered, GREG family estimators are essentially unbiased, and a fixed-effects model formulation and a mixed model formulation yield similar accuracy. An explanation for this observation is that in the setting of this exercise, the average levels of the response variable did not vary much over the domains. Best accuracy (excluding the true model) is for models where the PPS size variable x_1 is included. This demonstrates the accuracy gains attained from the "double-use" of x_1 both in the sampling design and in the estimation design; see also Särndal (1996). We also note that accuracy differences between the different GREG estimators are substantial especially in minor and medium domains, and accuracy improves with increasing the domain sample size.

Table 3. Average absolute relative bias ARB (%) and average relative root mean squared error RRMSE (%) of model-assisted GREG estimators of domain totals for minor, medium-sized and major domains of the generated population.

Model and estimator	Average ARB (%)			Average RRMSE (%)		
	Domain size class			Domain size class		
	Minor (20-69)	Medium (70-119)	Major (120+)	Minor (20-69)	Medium (70-119)	Major (120+)
Model A1 $y_k = \beta_{0d} + \varepsilon_k$						
GREG-SC	1.4	0.5	0.3	13.7	8.1	5.7
Model A2 $y_k = \beta_0 + u_d + \varepsilon_k$						
MGREG-SC	0.2	0.2	0.1	13.7	8.1	5.6
Model B1 $y_k = \beta_{0d} + \beta_1 x_{1k} + \varepsilon_k$						
GREG-SC	0.2	0.1	0.0	7.8	4.6	3.2
Model B2 $y_k = \beta_0 + u_d + \beta_1 x_{1k} + \varepsilon_k$						
MGREG-SC	0.2	0.1	0.0	7.8	4.6	3.3
Model C1 $y_k = \beta_{0d} + \beta_2 x_{2k} + \varepsilon_k$						
GREG-SC	1.4	0.5	0.3	11.6	6.8	4.8
Model C2 $y_k = \beta_0 + u_d + \beta_2 x_{2k} + \varepsilon_k$						
MGREG-SC	0.2	0.1	0.1	11.6	6.8	4.7
Model D1 $y_k = \beta_{0d} + \beta_1 x_{1k} + \beta_2 x_{2k} + \varepsilon_k$						
GREG-SC	0.0	0.0	0.0	1.7	1.0	0.7
Model D2 $y_k = \beta_0 + u_d + \beta_1 x_{1k} + \beta_2 x_{2k} + \varepsilon_k$ (Population generating model)						
MGREG-SC	0.0	0.0	0.0	1.7	1.0	0.7
Variables	x_1 Size variable in PPS sampling, x_2 Auxiliary variable					

3.2 EBLUP estimators

For estimators of the EBLUP family, we asked "How to account for the underlying unequal probability sampling design?". We proposed two options for this purpose: (a) the incorporation of sampling weights in the estimation of model parameters, and (b) the inclusion of sampling design variables as additional covariates in the model.

We compare unweighted and weighted EBLUP estimators constructed with four mixed model formulations. Model A includes a random intercept, variable x_1 is included in Model B, variable x_2 is included in Model C and both variables appear in the population generating model D. Similarly as for GREG, domain differences are accounted for by random intercept terms, and slope parameters are common for all domains. For all models (except D), EBLUP estimators are calculated with unweighted and weighted estimation of model parameters.

For Models A and C, unweighted estimators EBLUP-SC are seriously biased. For these models, the PPS sampling design is not accounted for. The bias declines considerably when the sampling weights are incorporated in the estimation of the mixed model, as shown by the new EBLUPW-SC estimators for Models A and C. The unweighted estimator EBLUP-SC under Model B shows best bias behaviour, indicating that the inclusion of the PPS size variable in the model can offer a powerful tool for bias reduction for EBLUP family estimators. Use of both weighting and the inclusion of x_1 in the model appears to be less powerful.

Accuracy behaviour of all EBLUP estimators is infected by the dominance of the squared bias component in the MSE, as indicated by the RRMSE figures. This holds for all three do-

main size classes. Because of large bias and small variance, invalid confidence intervals can be obtained. This means that point estimates can be systematically far away from the true value, independently of the domain sample size. In addition, accuracy does not improve much with increasing the domain sample size.

Table 4. Average absolute relative bias ARB (%) and average relative root mean squared error RRMSE (%) of model-dependent EBLUP estimators of domain totals for minor, medium-sized and major domains of the generated population.

Model and estimator	Average ARB (%)			Average RRMSE (%)		
	Domain size class			Domain size class		
	Minor (20-69)	Medium (70-119)	Major (120+)	Minor (20-69)	Medium (70-119)	Major (120+)
Model A $y_k = \beta_0 + u_d + \varepsilon_k$						
EBLUP-SC	22.9	23.1	21.7	22.9	23.3	21.8
EBLUPW-SC	3.7	3.5	3.3	3.9	3.6	3.5
Model B $y_k = \beta_0 + u_d + \beta_1 x_{1k} + \varepsilon_k$						
EBLUP-SC	1.8	1.4	0.7	2.8	2.5	2.2
EBLUPW-SC	3.5	3.5	3.3	3.5	3.6	3.3
Model C $y_k = \beta_0 + u_d + \beta_2 x_{2k} + \varepsilon_k$						
EBLUP-SC	22.3	23.1	21.8	22.4	23.2	21.9
EBLUPW-SC	3.7	3.6	3.2	3.9	3.7	3.3
Model D $y_k = \beta_0 + u_d + \beta_1 x_{1k} + \beta_2 x_{2k} + \varepsilon_k$ (Population generating model)						
EBLUP-SC	0.3	0.1	0.0	1.3	0.8	0.6
Variables	x_1 Size variable in PPS sampling, x_2 Auxiliary variable					

4 Conclusions

Results indicate that under unequal probability sampling, model-assisted GREG family estimators are quite insensitive to the model choice, a property also shown in our previous research to hold under SRSWOR. Model formulation and the estimation strategy of the model are critical for model-dependent EBLUP family estimators. This is especially true when using EBLUP for unequal sampling designs.

Bias of GREG estimators remained negligible for all model choices. “Double-use” of the same auxiliary information, that is, the use of the size variable in the PPS sampling design and in the assisting model, appeared to be beneficial with respect to accuracy. The accuracy improved with increasing the domain sample size. In this case, the mixed model formulation did not outperform the fixed-effects model formulation.

For model-dependent EBLUP family estimators, the bias can be large for a misspecified model. The PPS sampling design could be accounted for with two options, by the inclusion of the PPS size variable in the mixed model, or by the use of the weighted version of the EBLUP estimator, where the sampling weights are incorporated in the estimation procedure of model parameters. Of these two options, the first one appeared to be more effective, producing an EBLUP estimator with small bias and good accuracy. However, for both options, the squared bias component can still dominate the MSE, even in minor domains, tending to invalidate the construction of proper confidence intervals. Dominance of the bias component also can cause that the accuracy does not show improvement, when increasing the domain sample size.

References

Estevao, V.M. and Särndal, C.-E. (1999) The use of auxiliary information in design-based estimation for domains. *Survey Methodology*, **25**, 213–221.

Estevao, V.M. and Särndal, C.-E. (2004) Borrowing strength is not the best technique within a wide class of design-consistent domain estimators. *Journal of Official Statistics*, **20**, 645–669.

EURAREA Consortium (2004) Project Reference Volume, Parts 1, 2 and 3 (PDF).
Website: www.statistics.gov.uk/eurarea/

Fay, R.E. and Herriot , R.A. (1979) Estimation of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269–277.

Firth, D. and Bennett, K.E. (1998) Robust models in probability sampling. (With discussion) *Journal of the Royal Statistical Society*, B, **60**, 3–56.

Ghosh, M. (2001) Model-dependent small area estimation: theory and practice. In: Lehtonen and Djerf K. (eds) *Lecture Notes on Estimation for Population Domains and Small Areas*. Helsinki: Statistics Finland, Reviews 2001/5, 51–108.

Goldstein, H. (2003) *Multilevel Statistical Models*. Third Edition. London: Arnold.

Heady, P. and Ralphs, M. (2005) EURAREA: an overview of the project and its findings. *Statistics in Transition,* 7, 557–570.

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

Lehtonen, R., Särndal, C.-E. and 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.

Lehtonen, R. and Veijanen, A. (1998). Logistic generalized regression estimators*. Survey Methodology*, 24, 51–55.

Lehtonen, R. and Veijanen, A. (1999) Domain estimation with logistic generalized regression and related estimators. *Proceedings, IASS Satellite Conference on Small Area Estimation*, Riga, August 1999. Riga: Latvian Council of Science, 121–128.

Rao, J.N.K. (2003) *Small Area Estimation*. Hoboken: Wiley.

Särndal, C.-E. (1996) Efficient estimators with simple variance in unequal probability sampling. *Journal of the American Statistical Association*, **91**, 1286–1300.

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

VARIANCE ESTIMATION WITH IMPUTED DATA

Pauli Ollila[1]

[1] Statistics Finland, Finland
e-mail: Pauli.Ollila@stat.fi

Abstract

The paper describes briefly the theoretical framework of variance of an estimator in the presence of imputation and the basis for estimating the variance.

1 Introduction

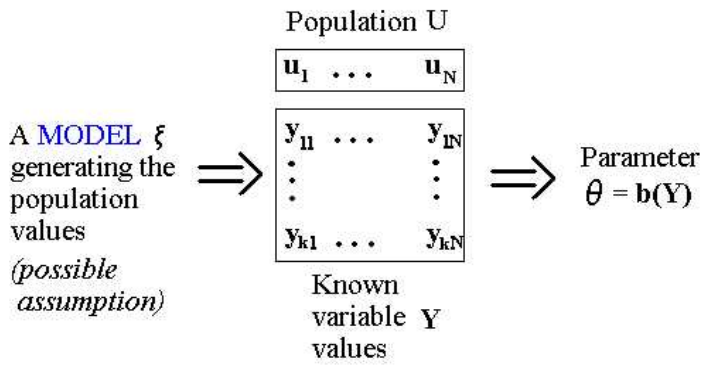
The aim of this paper is to give a brief overview on variance and its estimation with imputed data. In the presentation in Ventspils there will be examples of different variance estimation methods applied with a real data set. Most of the terminology here is based on Berger et al. (2004), although the graphical presentation and some expressions are by the author of this paper. Berger et al. (2004) is a deliverable of the DACSEIS project (2001 – 2004) concentrating on different variance estimation issues in survey sampling. This deliverable (and other DACSEIS papers as well) are available on the DACSEIS site www.dacseis.de. Also the imputation bulletins of Statistics Canada provide important information on the topic, see e.g. Rao (2003) and Kalton (2003). Rao's article also includes a somewhat comprehensive list of reference material.

2 Variance in the presence of non-response and imputation

2.1 Survey situation

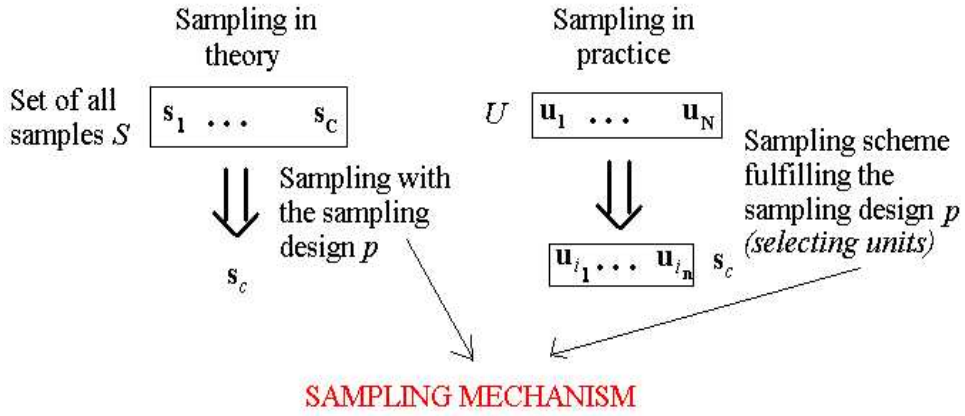
The population U of size N includes realised variable values in matrix $\mathbf{Y}$ ($k \times N$). One can make an assumption that there is an underlying superpopulation **model** ξ, which might have been behind the realised values of $\mathbf{Y}$ (Figure 1).

Figure 1. Population and parameter



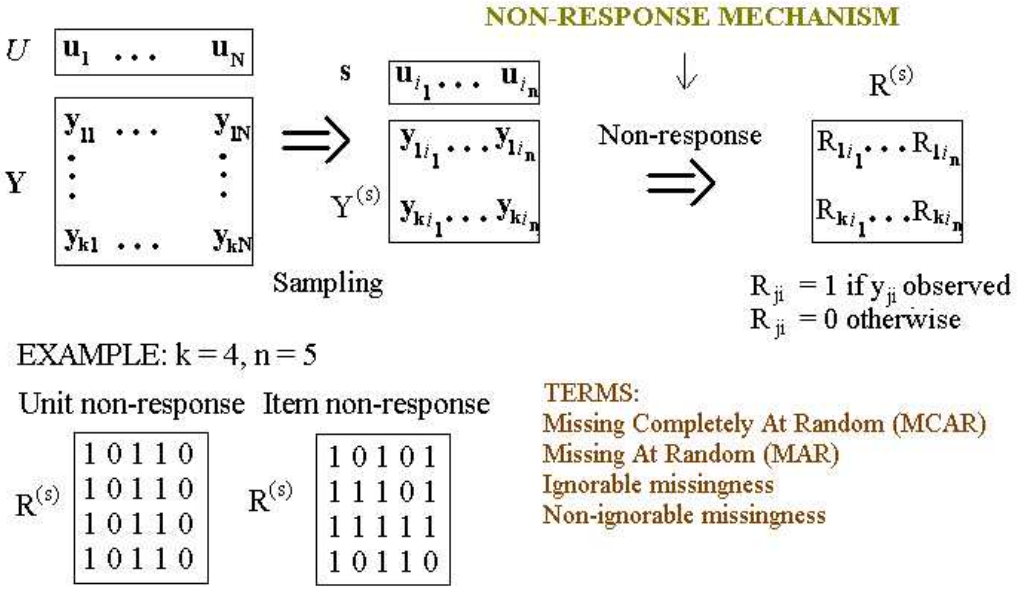
Theoretically sampling can be interpreted as a selection of a particular sample s_c size n_s with the probability $p(s)$ [sampling design] from the set of all samples $\boldsymbol{S}$. In practice we usually make the selection at the unit level following a scheme which fulfils the sampling design $p(s)$. In this context this process is called as a **sampling mechanism** (Figure 2).

Figure 2. Sampling in theory and practice



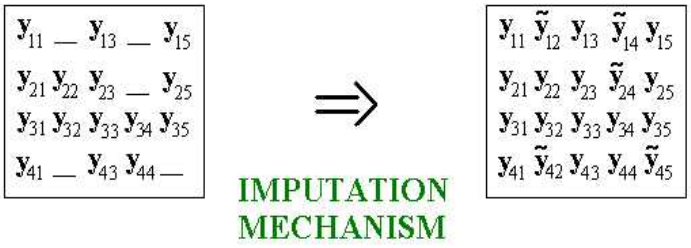
Almost always a survey includes unit non-response, and in many cases there is item non-response for certain variables as well. The matrix $\mathbf{R}^{(s)}$ with random variables R_{ji} ($i \in s$, $j=1, \ldots, k$) describes for each observation which variable values are observed. There might be underlying reasons and patterns why persons do not respond or some people do not answer all the questions asked. The sample s can be divided into two groups considering variable y: respondents s_r of size r and non-respondents s_m of size m. Different **non-response mechanisms** can be assumed as possible causes for non-response (Figure 3). The probabilities $p(R^{(s)})$ generating R_{ji} are denoted by q. Alternatively one can express the situation in terms of a **response mechanism**. For example in the case of *"Missing Completely At Random"* (MCAR) values of the variable (or a set of variables) for a point estimator are missing completely at random if missingness is independent of all these variables. Furthermore, in *"Missing At Random"* (MAR) values are "missing at random given an additional set of measured variables if missingness is independent of the values of the variables which are missing, conditional on the observed values of both sets of variables" (Berger et al. 2004). The mechanisms can also be further developed, for more information see e.g. Little and Rubin (2002). A specific application of the non-response mechanism is to expand it to the population level (see e.g. Rubin 1987), i.e. R_i are defined for all $i \in U$. The resulting matrix is denoted R_U. In most of the approaches using the population non-response mechanism this matrix can be constructed under the assumption of independence of sampling and non-response, i.e. s and R_U are independent. Correspondingly q^U denote the probability distribution $p(R_U)$.

Figure 3. Non-Response Mechanism



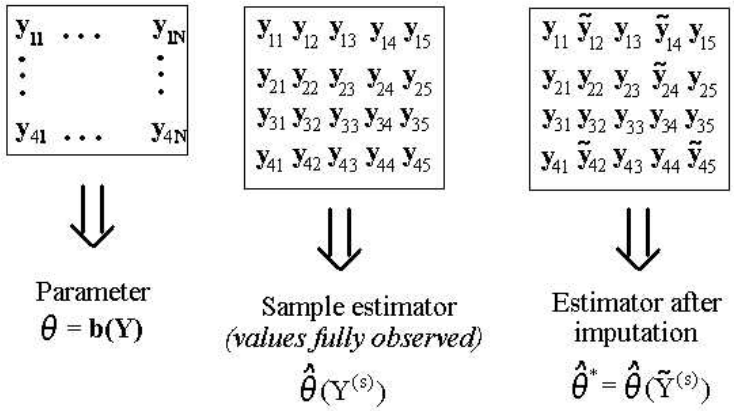
In most cases the unit non-response is dealt with weight adjustments. However, often in surveys there is a need to correct the item non-response by filling the missing information for variables according to a specific principle, i.e. under some **imputation mechanism** (Figure 4). The main reason behind imputation is the utilisation of the data for different analytic purposes, e.g. without imputation an analysis requiring several variables with item non-response concentrating on different observations in different variables may reduce the number of valid observations notably. The imputation based on marginal parameters (e.g. item totals, means, quantities [within groups]) is a simple way to solve the problem. A more sophisticated alternative is the modelling (e.g. regression or logistic regression based on auxiliary information, for the latter alternative see e.g. Ekholm and Laaksonen 1991) of the variable to be imputed. Auxiliary information is also utilised in the donor imputation, covering various techniques developed for the process (e.g. hot-deck imputation based on the order of another variable *x* in the data, which is in relation with the study variable y). Some imputation mechanisms are deterministic, i.e. the mechanisms do not produce any stochastic variation (e.g. simple mean imputation). On the other hand the mechanism can include variation (some donor methods) or it can be built in the mechanism (e.g. an error term in regression). Then the imputation mechanism is stochastic. In addition to these single imputation methods the multiple imputation methods (i.e. creating several imputed data sets for further operations) make a recent developing branch of imputation (Rubin 1987).

Figure 4. Imputation mechanism



In order to get results the parameters of interest must be estimated. If there was no unit non-response, the weights w_j (j = 1, …, n) for sample observations to be used in estimation would be created according to the sampling design. In the case of unit non-response weight adjustments would be carried out. For the variables with item non-response the use of these weights may provide biased estimates; most clearly this can be seen when estimating the total of a variable. In the case of imputation these weights can be used (Figure 5). However, if the variation is not introduced in the imputation mechanism (deterministic imputation), we end up usually too low standard errors when calculated from the imputed data.

Figure 5. Estimation

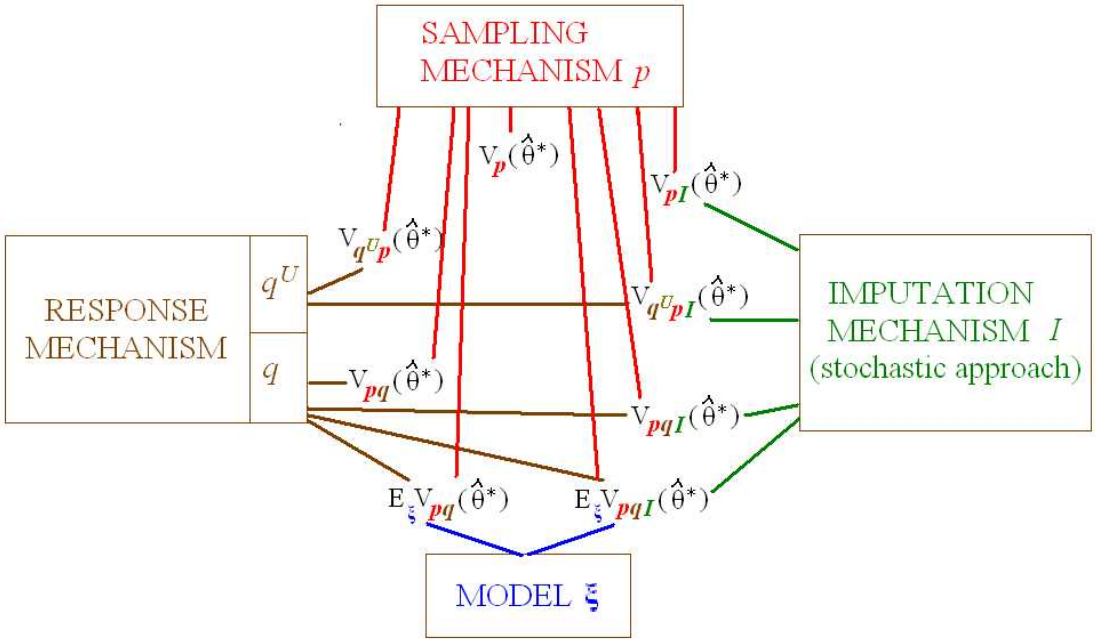


2.2 Sources of variation

What is the variance of the estimator including imputation, i.e. $V(\hat{\theta}^*)$? In the beginning one should determine what are the sources of variation in the whole process including the construction of the population, sampling, occurrence of non-response and conducting imputation (Figure 6). In the design-based approach the **effect of the sampling mechanism** cannot be avoided. Furthermore, there should always be an **assumption of the non-response mechanism** behind s_m in order to study the properties of the imputed estimator. The simplest assumption is that every observation has the same non-response probability for variable y. If there is **a stochastic element in the imputation mechanism**, that aspect is taken into account. An interesting alternative (see e.g. Deville and Särndal 1994) is to **introduce dependence upon a model** ξ (which is assumed to generate the population values) for the y_i

into the variance. This model-anticipated variance $E_\xi V_{pq}$ (or $E_\xi V_{pqI}$) applies the model-assisted regression theory familiar from Särndal, Swensson and Wretman (1992) as the case when non-response is treated as the second phase of selection incurred after the sample selection. The subscripts in the variances expressed in Figure 6 show the impact of different sources of variation.

Figure 6. Sources of variation



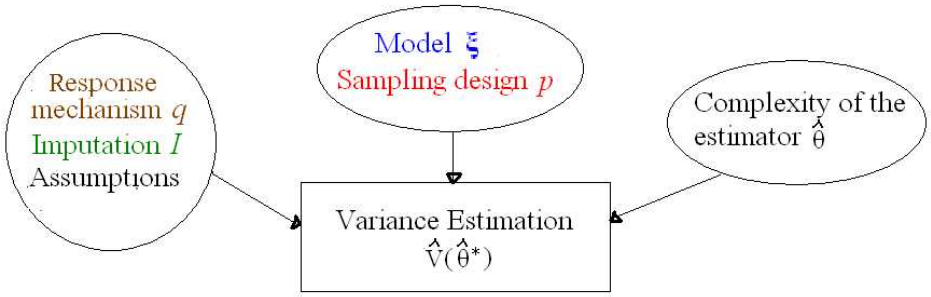
The principle of constructing the variance $V(\hat{\theta}^*)$ is theoretically rather simple. The sources of variation are decomposed into separate terms. For example, including the variation of the sampling mechanism and the response mechanisms provides the decomposition $V_{pq}(\hat{\theta}^*) = V_p E_q(\hat{\theta}^*) + E_p V_q(\hat{\theta}^*)$. When a third variation term is introduced, a further decomposition is conducted. These decomposition parts form the basis for variance estimation considering different sources of variation.

2.3 Variance Estimation

After defining the variance to be estimated, one should decide the method to be used in variance estimation (Figure 7). The complexity of the estimator is an important issue. Then one decision (partially connected to the sources of variation) is the same as in the case of non-imputed data: should we choose the analytic approach strictly based on the existing imputed data, possibly with an inflation factor (only for simple estimators); should we make a linearisation or other approximation of the estimator for variance estimation or should we concentrate on resampling methods (jackknife, bootstrap). A recent alternative is also the linearisation of the jackknife variance estimator (in the presence of imputation e.g. Sitter and Rao 1997, Berger et al. 2004). The analytic and linearisation methods usually apply the plug-

in data, which can be imputed data, but it may also include values adjusted for variance estimation purposes or it does not need to include all item non-response imputed (e.g. for variance estimation of regression imputation). The resampling methods may include the non-response mechanism and imputation mechanism more or less. An example of a construction of the population non-response mechanism is a modification of Sitter's bootstrap without replacement procedure (1992) with a pseudopopulation including a non-response structure (Berger et al. 2004).

Figure 7. Variance estimation



Solutions:
- Strict analytic formulae
- Approximations (e.g. linearisation)
- Resampling methods (e.g. jackknife, bootstrap)

Another decision is how to estimate the different sources of variation defined in the variance and to take the effect of the underlying assumptions into account. Then a crucial question is that from where we get the information for estimation of the different parts of variance. In suitable situations (e.g. small sampling fractions) some parts can be interpreted as negligible. Some expressions can be simplified with assumptions, e.g. estimating $V_p E_I(\hat{\theta}^*)$ simply with $\hat{V}_p(\hat{\theta}^*)$, which can be calculated in the absence of missing data with a non-response adjustment. Evidently one cannot always make such assumptions.

For the terms including the population response mechanism q^U one may develop e.g. linearisation functions of existing data values according to the structure of the mechanism and independence assumptions behind that (see Shao and Steel 1999). The usual way of introducing the non-response mechanism at the sample level (i.e. *q*) in the terms of variance estimation is the framework of two-phase sampling (e.g. Rao 2003), where we have the path population *U* --> complete sample *s* --> sample of respondents s_r . Methods for variance estimation in this case are developed utilising linearisation, resampling methods and linearised jackknife. See Rao (2003) and Berger et al. (2004) for more information.

If the imputation is stochastic, the variance estimator $\hat{V}_I(\hat{\theta}^*)$ should be somehow constructed, and this might be conducted e.g. by replication methods taking the stochastic nature of imputation into account or by theoretical adjustments in the case of regression with an additional error term. The terms including the model ξ can be constructed based on the structure of the model and the assumptions and auxiliary information behind it (Deville and Särndal 1994).

Examples of variance estimation methods with a real data set are shown (with some paper copies of presentation) in the workshop in Ventspils.

References

Berger, Y., Björnstad, J., Zhang, L-C., Skinner, C. (2004) Imputation and Non-response. *DACSEIS Deliverable 11.1*.

Deville, J. and Särndal, C-E. (1994) Variance Estimation for the Regression Imputed Horvitz-Thompson Estimator. *Journal of Official Statistics* **10**, 381-394.

Ekholm, A. and Laaksonen, S. (1991) Weighting via Response Modelling in the Finnish Household Budget Survey. *Journal of Official Statistics* **7**, 325-337.

Kalton, G. (2003) Imputation methods. The Imputation Bulletin

Little, R. and Rubin, D. (2002) *Statistical Analysis with Missing Data*, Second edition. New York: John Wiley & Sons.

Rao, J. N. K. (2003) Variance Estimation in the Presence of Imputation for Item Non-response. *The Imputation Bulletin*, **3**, *2-6*, Statistics Canada.

Shao, J. and Steel, P. (1999) Variance Estimation with Composite Imputation and Non-negligible Sampling Fractions, *Journal of the American Statistical Association* **94**, 254-265.

Sitter, R. R. (1992) Comparing Three Bootstrap Methods for Survey Data. *Canadian Journal of Statistics* **20**, 135-154.

Sitter, R. R. and Rao, J. N. K. (1997) Imputation for Missing Values and Corresponding Variance Estimation. *Canadian Journal of Statistics* **25**, 61-75.

Rubin, D. (1987) *Multiple Imputation for Non-response in Surveys*. New York: Wiley.

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

NONLINEAR CALIBRATION

[1]Aleksandras Plikusas

[1]Statistics Lithuania, Institute of Mathematics and Informatics, Lithuania
e-mail: Plikusas@ktl.mii.lt

Abstract

The definition of a calibrated estimator of the finite population parameter which may be not population total is discussed. Some estimators of the ratio of two population totals and population covariance is presented.

1 Introduction

Regression and calibrated estimators of the finite population totals are often met in the finite population statistics. These estimators are based on the use of auxiliary variables. The values of the auxiliary variables are known for all population elements. The definition and main properties of the calibrated estimator of the population total is given in the paper of Deville and Särndal (1992). The important subclass of the calibrated estimators are generalized regression estimators (GREG), which can be defined as calibrated estimator by choosing special loss function. The properties of GREG estimators of totals are considered in (Särndall, Swensson and Wretman, 1992). The estimation of the ratio of two population totals, population variance, population covariance as well as other population parameters is also topical. We will construct the calibrated (we may also call regression) estimators of the ratio of totals and the population covariance and provide the possible definition of a calibrated estimator of a more complicated parameters.

2 Calibrated estimator of total

Let us consider the finite population $\mathcal{U} = \{u_1, u_2, \ldots, u_N\}$. We can also assume $\mathcal{U} = \{1, 2, \ldots, N\}$. Denote the unknown population total of the variable y by

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

and Horvitz-Thompson estimator

$$\widehat{t}_y = \sum_{k \in s} \frac{y_k}{\pi_k} = \sum_{k \in s} d_k y_k.$$

Here $\pi_k = \mathbf{P}(k \in s)$, $k = 1, \ldots, N$ – inclusion probability of the element $k \in \mathcal{U}$ into the sample s, $d_k = 1/\pi_k$, $k \in \mathcal{U}$ – sample design weights.

Let us suppose that for every population element k the vector of auxiliary values $\mathbf{a}_k = (a_{k1}, \ldots, a_{kJ})'$ is known. It means we have J known auxiliary variables $a^{(1)}, \ldots, a^{(J)}$. In official statistics the auxiliary variables may be known from the previous census, administrative data, other sources. Denote the known total

$$\mathbf{t}_\mathbf{a} = \sum_{k=1}^{N} \mathbf{a}_k.$$

Calibrated estimator of the total t_y (Deville and Särndal 1992)

$$\widehat{t}_w = \sum_{k \in s} w_k \, y_k$$

is defined by the following conditions

a) using weights w_k the known total $\mathbf{t}_\mathbf{a}$ is estimated without error:

$$\hat{\mathbf{t}}_\mathbf{a} = \sum_{k \in s} w_k \, \mathbf{a}_k = \mathbf{t}_\mathbf{a};$$

b) the distance between the weights d_k and weights w_k is minimal according to some loss function L.

In most practical cases the loss function

$$L = L_1(w, d) = \sum_{k \in s} \frac{(w_k - d_k)^2}{d_k q_k}$$

is being used. Here q_k, $k = 1, \ldots, N$, – are free additional weights. We can also modify estimator by choosing weights q_k

Usually in survey practice we have many, say q, study variables $y^{(1)}, \ldots, y^{(q)}$. The notation can be summarized in the table below

Population element	study variables	auxiliary variable(s)
$u_1 \rightarrow$	$y_1^{(1)}, \ldots, y_1^{(q)}$	$\mathbf{a}_1 = (a_{11}, \ldots, a_{1J})'$
$u_2 \rightarrow$	$y_2^{(1)}, \ldots, y_2^{(q)}$	$\mathbf{a}_2 = (a_{21}, \ldots, a_{2J})'$
$\ldots$	$\ldots$	$\ldots$
$u_N \rightarrow$	$y_N^{(1)}, \ldots, y_N^{(q)}$	$\mathbf{a}_N = (a_{N1}, \ldots, a_{NJ})'$
totals	$t_y^{(i)} = \sum_{k=1}^N y_k^{(i)}$	$\mathbf{t}_\mathbf{a} = \sum_{k=1}^N \mathbf{a}_k$

It is known that in case auxiliary variables are well correlated with study variable, the mean square error of the calibrated estimator is lower compare to the Horvitz-Thompson estimator. It can be mentioned, that the problem of the selection of

auxiliary variables is not well studied. If J auxiliary variables are available one can choose from 2^J possible collections of auxiliary variables for the construction of calibrated estimators. In many practical applications the same auxiliary variables (it means the same weights) are being used for all study variables. Simulation study on the data of the Lithuanian Survey on Wages and Salaries show, that using different auxiliaries for different study variables we can reduce sampling error.

3 Calibrated estimator of the ratio

Let variables y and z be defined on $\mathcal{U}$ and take values $\{y_1, y_2, \ldots, y_N\}$ and $\{z_1, z_2, \ldots, z_N\}$, respectively. Let t_y and t_z be unknown population totals of y and z:

$$t_y = \sum_{k=1}^{N} y_k, \quad t_z = \sum_{k=1}^{N} z_k,$$

We are interested in the estimation of the ratio of two totals $R = t_y/t_z$. Suppose, the auxiliary variables a and b, having known population values $\{a_1, a_2, \ldots, a_N\}$ and $\{b_1, b_2, \ldots, b_N\}$ are available. We assign auxiliary variable a to the study variable y and and b to z. It means that a serves as auxiliary information for the study variable y and b – for the study variable z. So, we assume that the population totals

$$t_a = \sum_{k=1}^{N} a_k, \quad t_b = \sum_{k=1}^{N} b_k$$

and the ratio $R_0 = t_a/t_b$ are known.

One can take a straight estimator of the ratio R by taking the Horvitz-Thompson estimators of the totals t_y and t_z: $\widehat{R} = \widehat{t}_y/\widehat{t}_z$. Here

$$\widehat{t}_y = \sum_{k\in s} d_k y_k, \quad \widehat{t}_z = \sum_{k\in s} d_k z_k.$$

We shall construct a new estimator of the ratio R having the form

$$\widehat{R}_w = \frac{\sum_{k\in s} w_k^{(1)} y_k}{\sum_{k\in s} w_k^{(2)} z_k}. \tag{1}$$

Here the weights $w_k^{(i)}$, $i = 1, 2$, are defined under the two following conditions:

a) the weights $w_k^{(i)}$ satisfy the calibration equation

$$R_0 = \frac{\sum_{k\in s} w_k^{(1)} a_k}{\sum_{k\in s} w_k^{(2)} b_k}; \tag{2}$$

b) the weights $w_k^{(i)}$ are as close as possible to the initial design weights d_k according to the distance measure

$$L^2(w,d) = \alpha \sum_{k\in s} \frac{(w_k^{(1)} - d_k)^2}{d_k\, q_k} + (1-\alpha) \sum_{k\in s} \frac{(w_k^{(2)} - d_k)^2}{d_k\, q_k}\,. \tag{3}$$

Here q_k, $q_k > 0$, are free additional weights.

One can modify the calibrated estimator $\widehat{R}_w$ by choosing q_k or simply put $q_k = 1$ for all k.

The weights $w_k^{(i)}$, defining the estimator of the ratio $\widehat{R}_w$ can be found explicitly. Preliminary simulation results show that in some cases calibrated estimator of the ratio have lower variance than the ratio of two calibrated estimators of totals. It is not easy to compare the variances of these estimators analytically. Some special cases of the calibrated estimator of the ratio were considered by Plikusas (2003), and Krapavickaitė & Plikusas (2005).

4 Estimation of the population covariance

Suppose we are interested in the estimation of the population covariance

$$Cov(y,z) = \frac{1}{N-1} \sum_{k=1}^{N} \Big(y_k - \frac{1}{N}\sum_{k=1}^{N} y_k\Big)\Big(z_k - \frac{1}{N}\sum_{k=1}^{N} z_k\Big).$$

Consider the one of the standard estimators of the covariance

$$\widehat{Cov}(y,z) = \frac{1}{N-1} \sum_{k\in s} d_k\Big(y_k - \frac{1}{N}\sum_{k\in s} d_k y_k\Big)\Big(z_k - \frac{1}{N}\sum_{k\in s} d_k z_k\Big).$$

Let the variable a with the population values $\{a_1, a_2, \ldots, a_N\}$ and the variable b with the values $\{b_1, b_2, \ldots, b_N\}$ be known auxiliary variables. Denote their covariance by $Cov(a,b)$. We will construct a new calibrated estimator of the $Cov(y,z)$ using known auxiliary variables a and b. If the auxiliary variables are well correlated with the study variables, we can expect the variance of the calibrated estimator be smaller compare to the variance of estimator $\widehat{Cov}(y,z)$. The calibrated estimator

$$\widehat{Cov}_w(y,z) = \frac{1}{N-1} \sum_{k\in s} w_k\Big(y_k - \frac{1}{N}\sum_{k\in s} w_k y_k\Big)\Big(z_k - \frac{1}{N}\sum_{k\in s} w_k z_k\Big)$$

of the covariance $Cov(y,z)$ is defined under the following conditions:

a) the estimator $\widehat{Cov}_w$ estimates the known covariance $Cov(a,b)$ without error:

$$\widehat{Cov}_w(a,b) = \frac{1}{N-1} \sum_{k\in s} w_k\Big(a_k - \frac{1}{N}\sum_{k\in s} w_k a_k\Big)\Big(b_k - \frac{1}{N}\sum_{k\in s} w_k b_k\Big) = Cov(a,b); \tag{4}$$

b) the distance between the design weights d_k and calibrated weights w_k is minimal under the some loss function L.

It should be noted that in this case the explicit solution of the minimization problem does not exist even in the case of loss function (5) The iterative equations can be used to find the calibrated weights.

We can also use some other calibration equation instead of (4), for example,

$$\widehat{Cov}_w(a,b) = \frac{1}{N-1}\sum_{k\in s} w_k\big(a_k-\mu_a\big)\big(b_k-\mu_b\big) = Cov(a,b); \tag{5}$$

Here

$$\mu_a = \frac{1}{N}\sum_{k=1}^{N} a_k, \quad \mu_b = \frac{1}{N}\sum_{k=1}^{N} b_k.$$

The case when calibration equation (5) is used can be called linear calibration, because here we are calibrating the total of the variable $(a-\mu_a)(b-\mu_b)$.

5 Some general definition of nonlinear calibration

Taking into account the examples above we will define the (nonlinear) calibrated estimator, in case the parameter of interest θ is some function of the population totals: $\theta = f(t_y^{(1)},\dots,t_y^{(q)})$. Suppose we have selected q different collections of auxiliary variables $\mathbf{a}^{(1)},\dots,\mathbf{a}^{(q)}$ that are assigned to study variables $y^{(1)},\dots,y^{(q)}$. Denote the auxiliary totals by

$$\mathbf{t}_a^{(j)} = \sum_{k=1}^{N} \mathbf{a}_k^{(j)}, \quad j=1,\dots,q$$

and write formally

$$\hat{t}_{wy}^{(j)} = \sum_{k\in s} w_k^{(j)} y_k^{(j)}, \quad \hat{\mathbf{t}}_{wa}^{(j)} = \sum_{k\in s} w_k^{(j)} \mathbf{a}_k^{(j)}, \quad j=1,\dots,q.$$

The calibrated weights $w_k^{(j)}$ can be defined by the conditions

a) for some (it may be vector valued) functions g_1 and g_2

$$g_1(\hat{\mathbf{t}}_{wa}^{(1)},\dots,\hat{\mathbf{t}}_{wa}^{(q)}) = g_2(\mathbf{t}_a^{(1)},\dots,\mathbf{t}_a^{(q)})$$

b) the weight systems $w_k^{(j)}$ are as close as possible to the design weights d_k according to some loss function L.

The calibrated estimator of $\theta = f(t_y^{(1)},\dots,t_y^{(q)})$ be $\widehat{\theta} = f(\hat{t}_{wy}^{(1)},\dots,\hat{t}_{wy}^{(q)})$. Here we can take the loss function

$$L = \sum_{j=1}^{q} \alpha_j \sum_{k\in s} \frac{(w_k^{(j)}-d_k)^2}{d_k q_k}$$

with $\alpha_j \geq 0$ and $\sum_{j=1}^{q} \alpha_j = 1$. The loss function is minimized also by α_j, $j = 1, \ldots, q$. Of course, the existence of the solution of such calibration problem is under the question. The simulation examples of calibration of covariance show that for properly chosen iterative equations and loss functions the calibrated weights exist for almost all samples.

References

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

[2] D. Krapavickaitė, A. Plikusas. Estimation of a Ratio in the Finite Population. *Informatica*, 2005, **16**(3), p. 347-364.

[3] A. Plikusas, Calibrated weights for the estimators of the ratio, *Lith. Math. J.*,**43**, 543-547 (2003).

[4] C.-E. Särndal, B. Swensson, J.Wretman, *Model Assisted Survey Sampling*, Springer-Verlag, NewYork (1992).

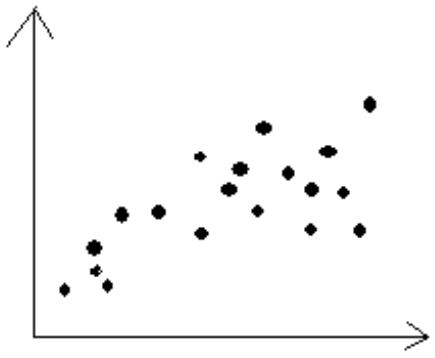
Optimal inclusion probabilities and estimators when sampling with varying probabilities

by
Daniel Thorburn,
Department of Statistics
Stockholm University
Ventspils August 2006

We discuss optimal allocation of inclusion probabilities in the presence of auxiliary information. In most situations one should use itboth when deciding the inclusion probabilities and in the estimator. The Horvitz-Thompson-(HT)-estimator is seldom optimal. We will only look at large sample theory and use a modelassisted designbased approach. The observations $y_i;\ i \in U$ are iid rv.

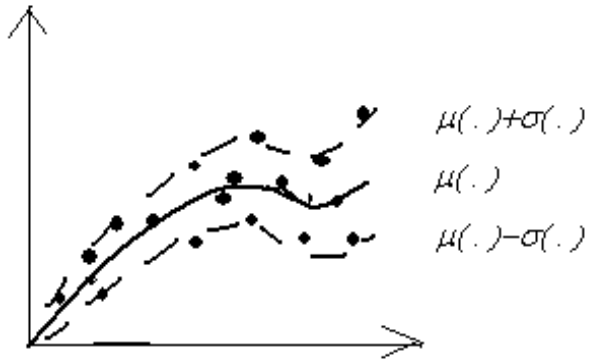
$$\mathrm{E}(Y_i|x_i) = \mu(x_i); \quad \mathrm{Var}(Y_i|x_i) = \sigma^2(x_i)$$

where μ is a nice function and x_i is the auxiliary information (perhaps multidimensional). In this talk we will assume that x is one-dimensional but generalisations to the multidimensional situation is straight-forward. If the population looks as follows (but with more data-points)$\mu(x_i)$ and $\sigma(x_i)$ may look as follows



We assume that μ varies slowly so that it can be estimated fairly well from the sample (e.g. with a moving average other kernel estimators or spline functions). Its estimator is denoted by $\mu^*(\bullet)$. A natural estimator of the total is then

$$\sum_U \mu^*(x_i) + \sum_s \frac{1}{\pi_i}(y_i - \mu^*(x_i))$$



The first part is a model-based estimator and the last part is an estimator of the design-bias. We call this estimator a generalised difference estimator. It is approximately unbiased for large samples and its variance can be estimated by the ordinary Sen-Yates-Grundy estimator

$$\sum_U \sum_U \frac{(\pi_{i,j} - \pi_i \pi_j)}{\pi_i \pi_j} (y_i - \mu(x_i))(y_j - \mu(x_j))$$

From the assumed independence the expected variance under the model this is approximately

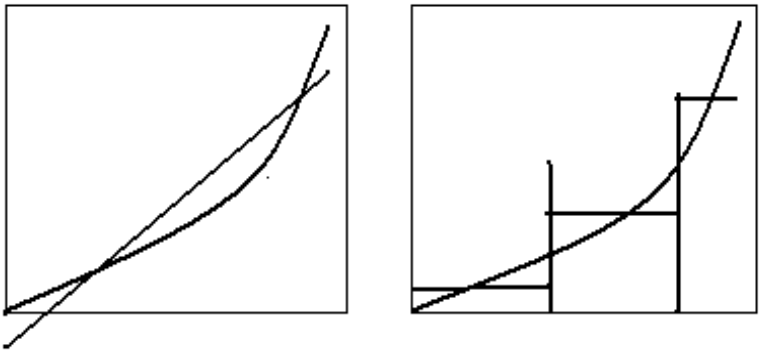
$$\sum_U \frac{1 - \pi_i}{\pi_i} \sigma^2(x_i)$$

Minimising this expression under a cost constraint gives that the inclusion probabilities (at least for large n and N) should be chosen

$$\pi_i \propto \frac{\sigma(x_i)}{c_i^{1/2}}$$

if the marginal costs are c_i. Those who have seen Neyman allocation recognises this expression.

We have not assumed anything about the procedure selecting the sample, i.e. the second order inclusion probabilities are unimportant, but $\mu(\bullet)$ must be estimated consistently and the above estimator used. With a more rigid model like polynomial regression with a bounded degree, a dependence may appear and the second order inclusion probabilities become important. In the next picture it is illustrated by a straight line regression and a curved mean value function. The residuals for two close x-values will mostly have the same sign. In that case one ought to choose a πps-design which spreads the observations so that $\pi_{i,j} < \pi_i * \pi_j$ if x_i and x_j are close. A similar effect occurs when stratifying with a limited number of strata or using splines with a bounded number of nodes.



With the above asymptotically optimal estimator the πps-method did not matter. The second order inclusion probabilities disappeared in the approximate variance. Then you can choose any sampe design like: systematic πps, Pareto-πps or Poisson-sampling. But if you intend to use a non-optimal method, like the ordinary HT-estimator, the design matters. Most commonly used πps-methods try to get a high independence i.e. they tries to mimic SRS, when $\pi(\bullet)$, is constant. In most cases this is a silly choice. It is e.g known that variants of systematic sampling and stratified sampling are better even when the inclusion probabilities are constant if the mean value function varies slowly. This holds here too. Systematic πps, ordering the observations after $\pi(x)$ or x or other sensible background variables is better. If you intend to use the HT-estimator and wants an asymptotically small variance you should avoid methods like Sampford, Pareto-πps or Poisson-sampling.

Systematic πps has, as we said, the advantage that you get a good and representative sample. But it has the disadvantage that the variance cannot be estimated exactly. But this is not a really a valid counterargument. Because everything you can estimate with e.g. SRS or Sampford, you can still estimate with a never larger variance. Thus you can give an upper bound on the variance-which is the variance with with Pareto-πps, say. There also exist list-sequential methods which have the same asymptotically optimal behaviour as systematic sampling and where the variance is possible to estimate. Another way to obtain fairly good sampling schemes is to use stratified sampling with decreasing strata widths.

If one uses a "silly" estimator like the HT-estimator or the regression estimator, the above inclusion probabilities are not optimal. Instead one must add a residual term to the variance getting

$$\pi_i \propto \frac{\left(\sigma^2(x_i) + (\mu(x_i) - \mathrm{E}(\mu^*(x_i)))^2\right)^{\frac{1}{2}}}{c_i^{1/2}}$$

where $\mathrm{E}(\mu^*(x_i))$ e.g. is 0 for the HT-estimator and the best regression line $\mathrm{E}(a^* + b^* x_i)$ for the regression estimator. If the variance function $\sigma^2(x_i)$ is

small compared to the size, $\mu(x_i)$, this formula says that for the HT-estimator, inclusion probabilities proportional to size are optimal. This is a well-known fact, which often is used to motivate πps.

Variance of quantile estimators in household surveys

Imbi Traat

University of Tartu, Estonia
e-mail: imbi.traat@ut.ee

Abstract

In this paper estimators for quantiles and for some of their functions are considered. These estimators are frequently used as various poverty measures. The corresponding variance estimators are derived for the design often used in household surveys (households are selected through people in the Population Register). A jackknife variance estimator for this design is also given. An illustration by simulation is presented.

1 Introduction

This research was made for the Estonian Statistical Office in 2005 (http://www.stat.ee/169971). Estonia had to run EU-SILC survey (Survey on Income and Living Conditions) and measure various poverty and income indicators called the Laeken Indicators. The survey design was the same as for the Household Budget survey – a stratified unequal probability design. But the required estimators were different. Instead of estimating ordinary means, totals and proportions the quantiles and their functions were needed (median, poverty threshold, at-risk-of-poverty rate, etc.). Variance estimators for Estonian EU-SILC survey were absent. The task was to derive them. The results will be briefly presented here. The variable used for quantile calculations was *equivalized disposable income.*

2 Design for household surveys

The survey design is stratified unequal probability sampling of households. Sampling is carried through among the records of population register, whereas the sampling frame consists of people 14 years old and older (14+). Strata are formed geographically by grouping Estonian counties (and the capital city Tallinn) into three strata. Within each stratum systematic sampling procedure of persons is used with different sampling fractions in the defined strata. Each selected person brings its household (hh) into the sample. All members 16+ of that hh are questioned.

Probabilistic description. Let I_{hi} be the sampling indicator of the hh i (shows how many times the hh is sampled) in stratum h. The expected sampling count of that hh is

$$E(I_{hi}) = np_i, \quad p_i = m_{14hi} / M_{14h},$$

where n is sample size in households, m_{14hi} is the number of 14+ people in hh i of stratum h, and $M_{14h} = \sum m_{14hi}$ is the total number of people in the frame (population register with 14+ persons). Note that here the index i refers to the hh. The expected sampling counts are proportional to the 14+ size of the households. The hh's of big size are more frequently

sampled causing over-representation of big-size hh's. This needs down weighting by sampling weights:

$$w_{hi} = k_{hi} / E(I_{hi}), \quad h = 1,2,3,$$

where k_{hi} is an outcome of the sampling indicator I_{hi}, usually equal to 1.

The joint distribution of sampling indicators I_{hi} in startum h is a multivariate hypergeometric distribution. This distribution is well studied and the variances and covariances of sampling indicators well known (Johnson et al 1997). However, this theoretical framework is not used in sampling literature. Sampling designs as multivariate distributions are considered in Traat et al. (2004).

3 Estimators for quantiles

Denote y_i as a study variable and w_i as a sampling weight (possibly adjusted). The index here refers to a person. The sampling weight is the same for the persons in the same household. Sample of persons is denoted by s and of households by s_{hh} .

Let y_i be sorted into ascending order. Then the estimated α-quantile of the variable y is

$$q_\alpha = \begin{cases} (y_j + y_{j+1})/2, & if \quad \sum_{i=1}^{j} w_i = \alpha \hat{M} \\ y_{j+1}, & if \quad \sum_{i=1}^{j} w_i < \alpha \hat{M} < \sum_{i=1}^{j+1} w_i \end{cases}, \qquad (1)$$

where $\hat{M} = \sum_s w_i$ is the estimated number of people

The estimated median is received for $\alpha = 0.5$ and estimated quintiles for $\alpha = 0.2, 0.4, 0.6, 0.8$.

Several indicators are calculated as functions of quantile estimators. For example, the indicator *at-risk-of-poverty-threshold* (I_{1e}) is defined as 60% of median, so its estimator is

$$\hat{I}_{1e} = 0.6\, q_{0.5}\,.$$

The *Income quintile share ratio* (I_2) is estimated as

$$\hat{I}_2 = q_{0.8} / q_{0.2}\,.$$

The quantiles in the formulae are based on the variable *eqinc*.

4 Variance estimators

There are several moments, which need special attention when developing variance formulae for Laeken indicators:

- sampling unit is household but the estimators are formed with person-level data;
- sampling design is complex – unequal probabilities for households and persons;
- estimators are non-linear;
- domain variable in estimators has a random threshold
- calibration and weight adjustments may have introduced unequal weights for persons of the same household.

We use inverse distribution function method for the variance of quantile estimators. The general results and some special cases are given in Särndal et al. (1992). The hypergeometric design is not considered there and will be done here. We skip the stratum index and present the following formulae for the hypergeometric design in one stratum.

First we estimate variance of the distribution function at a sample quantile:

$$\hat{V} = \hat{V}(\hat{F}) = \frac{M_{14} - n}{M_{14}} \frac{1}{n(n-1)} \frac{1}{\hat{M}^2} \sum_{s_{hh}} \left(\frac{m_i}{p_i}\right)^2 (\tilde{z}_i - \alpha)^2,$$

where

$$\tilde{z}_i = \begin{cases} 1, & if \quad y_{ij} \le q_\alpha, \ \forall j, \quad (j \text{ refers to the member of hh } i), \\ 0, & otherwise, \end{cases}$$

m_i is no. of eligible (questioned) members in hh *i*.

Now, if $\hat{F}$ at a sample quantile is approximately normally distributed around α (which is the case for big samples) we can say that (c_1, c_2) is an approximate 95% confidence interval for α, where

$$c_1 = \alpha - 1.96\sqrt{\hat{V}}, \ c_2 = \alpha + 1.96\sqrt{\hat{V}}.$$

Inverting the points c_1, c_2 with $\hat{F}^{-1}$ which means that we calculate q_{c_1}, q_{c_2} from (1), we get that (q_{c_1}, q_{c_2}) is the approximately 95% confidence interval for true quantile Q_α. From the last interval one can also estimate the variance of quantile estimator (assuming normality):

$$\hat{V}(q_\alpha) = [(q_{c1} - q_{c2})/(2 \cdot 1.96)]^2 \qquad (2)$$

The simplest function of quantiles is *the poverty threshold* $\hat{I}_{1e} = 0.6\, q_{0.5}$. Its variance in a straightforward way is

$$\hat{V}(\hat{I}_{1e}) = 0.6^2\, \hat{V}(q_{0.5}),$$

where $\hat{V}(q_{0.5})$ is calculated from (2) with $\alpha = 0.5$.

5 Jackknife variance estimator

The resampling methods are appealing due to their applicational simplicity. In this work we concentrate on the Jacknife method. The statistic considered is the design-weighted sample sum. The two-phase sampling framework is assumed with the hypergeometric design of sample size n_a in the first phase and SI-sampling of size *n* of already selected hh's (multiples included) in the second phase. The first phase estimator is $\hat{t}_a = \sum (I_{ai} y_i)/n_a p_i$ and the second phase one $\hat{t} = \sum (I_i \mid I_{ai}) y_i / n p_i$. The sampling indicators I_{ai} and $I_i \mid I_{ai}$ describe selections of hh *i* into the first phase and second phase samples, respectively. The following formula for the first phase variance can be derived:

$$V_a(\hat{t}_a) = c_a \frac{n}{n_a - n} V(\hat{t} \mid I_a),$$

where $c_a = (M_{14} - n_a)/(M_{14} - 1)$ comes from the hypergeometric sampling, given I_a means given the first phase sample. If to choose the subsampling size $n = n_a - 1$, and to use that $c_a \approx 1$, which is usually true in practice, the formula simplifies,

$$V_a(\hat{t}_a) = (n_a - 1) V(\hat{t} \mid I_a). \qquad (3)$$

The important thing here is the fact that $V(\hat{t} \mid I_a)$ can be estimated as variance of $\hat{t}$ over second phase samples. The result (3) wich is valid under EU-SILC design for sample sums, was applied to the Gini coefficient and it performed very well.

6 Illustrations

The population was formed with the Estonian HBS 2003 data and made similar to the true Estonian population by its hh characteristics (Leiten and Traat, 2005):

Total no. of people 2595; total no. of frame people $M_{14} = 2332$; total no. of hh's 1263;

Median of *eqinc* 3000; Poverty threshold $I_{1e} = 1800$;

At-risk-of-poverty rate $I_1 = 15.34\%$; Income quintile share ratio $I_2 = 2.58$;

100 persons were SI-selected in the frame and their hh's included into sample. The sample quantities were calculated, their means and variances over 4000 repetitions obtained. Some tables are below.

Table 1. *Estimated median and related quantities*

	med	$\sqrt{\hat{V}(med)}$	$\sqrt{\hat{V}(\hat{F})}$
Mean	3036	216.6	0.05
Std Dev	216.5	52.3	0.0006

The first column is for sample median, others for derived variance estimators over 4000 simulations. We see that median as calculated by (1) on average slightly overestimates the true median 3000. Overestimation is small: $36/3000 \approx 1\%$. The true sampling variability of the median is not big, $c.v. = 216.5/3036 \approx 7\%$. The second column says that the variance formula worked out by us performs very well, it produces almost unbiased variance estimator, and the estimator is also quite stable (with standard deviation 52.3). The third column characterizes variability of the estimated distribution function at the estimated median. This is basic component when finding confidence intervals and variance of the median with inversion method. Stability of $\hat{F}$ quarantees good performance of the method. It appeared that the confidence intervals of the median worked very well. The coverage rate was 95.2% instead of 95%.

The poverty threshold $\hat{I}_{1e}$ is median-based quantity and its performance is much defined by the median.

Table 5.3. *Poverty threshold* $\hat{I}_{1e}$ *and related quantities*

	$\hat{I}_{1e}$	$\sqrt{\hat{V}(\hat{I}_{1e})}$
Mean	1821.8	130.0
Std Dev	129.9	31.4

References

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

Leiten, E., Traat, I. (2005) *Variance of Laeken Indicators in complex surveys.* The statistical Office of Estonia: http://www.stat.ee/169971

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

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.

USAGE OF ADMINISTRATIVE DATA IN EU-SILC SURVEY

Signe Balina[1]

[1] University of Latvia, Latvia
e-mail: signe.balina@lu.lv

Abstract

The main purpose of this paper is to present the results of first analyses of the possible usage of Latvian administrative registers for collecting income data necessary for EU-SILC survey.

1 Introduction

The main purpose of this paper is to analyse the possibility of the usage of Latvian administrative registers for collecting income data necessary for EU-SILC survey.
The main tasks were as follows:

- Gathering of information on existing administrative income registers and the analysis of them;
- The analysis of possible links and integration between data from EU-SILC survey and administrative registers.

The first EU-SILC survey in Latvia was carried out in summer 2005, at the same time in March, 2004 Central Statistical Bureau of Latvia carried out the 1st wave of EU-SILC pilot survey and the 2nd wave was provided in October, 2004. Therefore it was possible to use data of the pilot survey for first analyses of the possible usage of Latvian administrative registers for collecting income data necessary for EU-SILC survey.

2 EU-SILC pilot survey in Latvia

The reference population of EU-SILC pilot survey in Latvia is all private households and their current members. Persons living in institutional households are excluded from the target population.

The method of sampling – a two-stage stratified random sample of households. Primary sampling units (PSU) are administrative territories. Primary sampling units are selected within each stratum by systematic probability proportional to size (the number of households

in PSU) sampling with a random starting point. At the 2nd stage households are selected by a simple random sampling procedure. As a sampling frame the dwelling data base was used.

The gross sample size for EU-SILC pilot survey in Latvia was 500 households. The financial resources of the pilot survey allowed to survey 200 responding households. In 200 responding households there were 505 persons from whom 408 persons belong to the EU-SILC target population (the EU-SILC target population is all persons aged 16 and over).

One of the main EU-SILC objectives is to produce comparable and timely cross-sectional and longitudinal data on income and on the level and composition of poverty and social exclusion.

The income data reference period in Latvian pilot survey was the preceding calendar year (year 2003), which for the respondents is a clear and unambiguous category. For regular social transfers paid by state the reference period is one month, and information is additionally obtained on the number of months during which transfers were received. However in the data file these transfers are also recalculated for the period of the full calendar year.

3 Identification of persons and related problems

All permanent residents of Latvia have a unique person identification code. The person ID code consists of 11 digits; the first six of them specify his/her birth date (*ddmmyy*). Since the person ID code is unique it can be effectively used as a key variable for merging different data bases.

During the interviews of the 1st wave of the EU-SILC pilot survey the person ID code was not registered. For identification purposes of respondents simply the name, surname and the birth date was used. It was suspected that registration of the person ID code during the survey may significantly decrease the survey response rate. It was the main reason why person ID code was not registered in any persons' survey (including the 1st wave of the EU-SILC pilot) carried out by CSB of Latvia.

The lack of person ID code in EU-SILC pilot data considerably complicated possibilities of merging survey data with the administrative data bases. As a key variables persons name, surname, birth date and address were used. Using these variables the person ID code was looking for in the Population register.

In the 1st wave the person ID code was identified for 311 persons (out of them 254 were respondents and 57 non-respondents). Thus person ID code was identified for slightly more than 60% of all sampled persons. In the 2nd wave respondents were asked also about their person ID code. In 35 cases respondents corrected the existing information about their person

ID code. Nevertheless, some respondents still did not give information on their person ID code. Looking once again for the person ID code in the Population register allowed finding 147 more codes. Altogether 485 person ID codes were identified, and for 20 persons identification of their person ID codes was impossible.

4 Income information of administrative registers

Questions about income are sensitive. Pilot survey showed that there is a high item non-response for these variables. It would be very useful to obtain data about income from other sources.

Analysis of income structure pointed out two main existing sources of income information – State Revenue Service (SRS) and State Social Insurance Agency (SSIA). Therefore in the current study the main attention was paid to the income variables of these two administrative registers, and analysis of building the links between registers' data and data from EU-SILC survey.

4.1 State Revenue Service information

The tax information about persons who are income tax payers in the administrative register of State Revenue Service is gathered in several ways:

- Every month employers statutory declare income and tax information about their employees,
- Self-employed persons statutory full-fill the annual income declarations till April of the next calendar year,
- Any tax payer can submit the annual tax declaration specifying all income (from different sources), calculated and withhold taxes, and expenses redeemed from taxes (including contributions to the private pension funds).

During the pilot study CSB of Latvia asked State Revenue Service for income data of 485 sampled persons having identified person ID code in EU-SILC. Complete income information was received. Altogether 242 records related to 201 people were received from the SRS. These records contained information about wages and salaries as well as other income information (sickness benefits, income from intellectual property, etc.).

Out of 201 persons 178 responded in the EU-SILC. The other 23 persons did not belong to the target population (were of age below 16 years) or did not respond.

The SRS did not deliver any information about 284 persons since these persons were not tax payers in the year 2003: the age of 79 persons is below 16 years and the SRS register did not

contain information about 97 pensioners because according to the law the annual pensions below Ls 1200 are not a subject of income tax.

There were 14 persons without SRS data declaring their status as “working full time” or “working part-time” in the EU-SILC pilot survey. It can be explained by several reasons. It may happen that a person is working at the time of interview and at the same time he/she was not working during the whole reference year. Another possibility is that a person was working without some official contract and his/her employer did not pay any taxes. Thus such person does not appear in the SRS register. It may happen also that the respondent gave false information during the pilot survey.

4.2. State Social Insurance Agency information

The State Social Insurance Agency (SSIA) is a state institution under supervision of the Ministry of Welfare, performing the public administration function in the area of social insurance and social services.

Thus SSIA administrative register contains information about different type of pensions, social benefits and allowances paid to Latvian residents. Thus in the SSIA register it is possible to find the income information of more than 40% of all respondents of the EU-SILC pilot survey. Unfortunately existing legislation did not allow obtaining person level income data from the SSIA. The negotiations between the CSB of Latvia and SSIA still continue. If the solution of the legal aspects will be found, the income data of social transfer compiling a significant income part will be available.

5 Analysis of Data from State Revenue Service

5.1 Available Information

From the State Revenue Service (SRS) we received income and tax information about 201 person. For 187 individuals we received information about wages and salaries paid in cash for time worked or work done in main and any secondary or casual job(s). Some persons have more than one type of income and 11 of them had submitted income declaration. Wages and salaries is the type of income for which the information is available for the biggest percentage of respondents.

In the EU-SILC pilot survey the respondents could report gross or net wages and salaries. One part of respondents reported the annual income (gross and/or net), another part – the monthly income (gross and/or net) as well as the number of months during which wages and salaries were received (however, in the data file monthly wages and salaries are also recalculated for the period of the full calendar year). It is some differences between wages and salaries information from the SRS registers and wages, salaries and other labour force

income information from the EU-SILC survey data, but this information is comparable in some way.

5.2 Comparison of gross wages and salaries

In EU-SILC pilot survey 79 respondents gave information about the annual gross wages and salaries paid in cash for time worked or work done in main and any secondary or casual job(s) (74 of them were with person identification number, 5 – without). For 70 of them information from the State Revenue Service is also available (for 69 available is information about wages and salaries, for one person – information about the income from intellectual property). Therefore comparable are data for 69 persons.

Table 1 summarises information about annual gross wages and salaries from two data sources – EU-SILC survey data and SRS registers information. Below we concentrate our attention to the differences between the SRS and survey data that are bigger than 10% (differences that do not exceed 10% we consider as statistically insignificant).

Table 1 Comparison of gross wages and salaries from SRS registers and EU-SILC survey data (annual data)

	Frequency	*Percent*	*Cumulative percent*
Gross income in EU-SILC > gross income in SRS	13	18.8	18.8
$\approx \pm$ 10%	29	42.0	60.9
Income in SRS > Income in EU-SILC	27	39.1	100.0
Total	69	100.0	

From the Table 1 one can see that gross income of 42.0% of 79 respondents is almost equal to the income shown in the SRS data. Income of 18.8% of respondents exceeds more than for 10% of the corresponding gross income of the SRS data. At the same time income of 39.1% of respondents is less (more than for 10%) than the corresponding gross income of the SRS data.

Mean annual gross wages and salaries estimated from the SRS registers for this group of respondents is Ls 2956.71, at the same time mean gross wages and salaries estimated from the EU-SILC data is lower – Ls 2663.91 (the corresponding medians are Ls 1824.86 and Ls 1509.00). The median estimated from the SRS registers is higher than median estimated from the EU-SILC because among 50% of respondents with lowest income the income data in the survey are not completely reported. Another reasons why mean and median of the gross income calculated from the SRS registers data is higher than mean and median gross

income estimated from the EU-SILC survey data is that very frequently respondents better understand and know what is net income rather than gross income.

Figure 1 and Table 2 show very high correlation between the SRS and EU-SILC data for persons, who gave information about their annual gross income. The estimated coefficient of correlation is equal to 0.983 and it is statistically significant at the significance level 0.01.

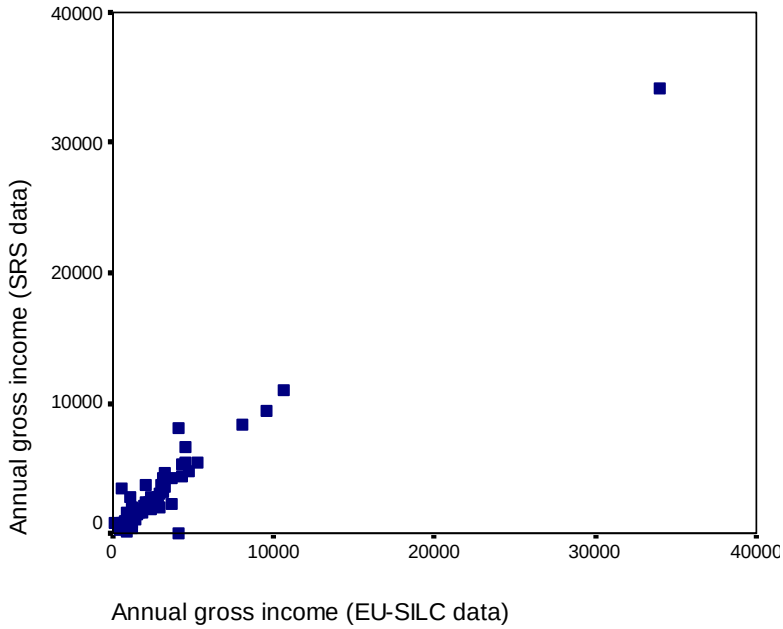


Figure 1 **Scatterplot of gross wages and salaries from SRS register and EU-SILC survey data (annual data)**

Table 2 **Correlation between gross wages and salaries from SRS register un EU-SILC survey data (annual data)**

	Gross wages and salaries, SRS	*Gross wages and salaries, EU-SILC*
Pearson Correlation	1	.983
Sig. (2-tailed)	.	.000
N	69	69
Pearson Correlation	.983	1
Sig. (2-tailed)	.000	.
N	69	69

** Correlation is significant at the 0.01 level (2-tailed).

The analysis shows that for group of respondents who reported their annual gross income from wages and salaries the quality of income information of data in both data sources is high. It means that in the next EU-SILC surveys we can use the SRS gross income

information and therefore we will reduce the response burden for persons participating in the EU-SILC survey. We have high item non-response in the EU-SILC survey therefore we could also use the SRS registers data at least for imputation purposes in the cases of item non-response.

Similar analyses were made also for monthly gross wages and salaries as well as for annual and monthly net wages and salaries.

6 Conclusions and Recommendations

Main conclusions that can be made from results of analysis of the possibility of the usage of Latvian administrative registers for collecting income data necessary for EU-SILC survey are as follows:

- A person identification code is the most appropriate key variable for merging different persons' registers or merging survey data with some administrative data sources. Searching for person ID in the copy of population register based on persons name, surname, birth date and living address allows identification of persons' ID codes, nevertheless, it is a time- and labour-consuming task that can be done if small number of person ID have to be found but that is difficult to realise if hundreds or even thousands of person ID code have to be found. Therefore, inclusion of a person identification code in the EU-SILC questionnaire (and in the EU-SILC survey data base) is a necessary condition for successful combining of the survey data with administrative data sources (SRS administrative register, SSIA administrative register, or some other data source).
- Comparison of EU-SILC pilot survey data with SRS administrative register data on gross/net wages and salaries for respondents reporting their annual gross/net wages and salaries shows that the quality of income data in both data sources for this group of employees is high.
- Comparison of EU-SILC pilot survey data with SRS administrative register data on net wages and salaries for respondents reporting their monthly net wages and salaries shows rather high quality of income data in both data sources also for this group of employees.
- Comparison of EU-SILC pilot survey data with SRS administrative register data on gross wages and salaries for respondents reporting their monthly gross wages and salaries shows significant differences between the two data sources. The analysis made so far does not indicate any serious reason, why the data quality of the SRS administrative register for this group of respondents (in contradistinction

to other groups of respondents) could be of a poor quality. It means, most likely there do exist some quality drawbacks in the EU-SILC pilot survey data (person's monthly income from wages and salaries has rather big seasonal fluctuation, or respondents reporting their monthly gross income in fact do not know precisely its amount, or probably some respondents even misunderstand the definition of the gross income, etc.).

- Usage of the SRS administrative register data as one of the main income data sources for wages and salaries:
  - allows significant reduction of the response burden for persons participating in the EU-SILC survey,
  - allows obtaining detailed and more complete income data on wages and salaries,
  - promotes the survey response rate, and thus,
  - improves the total quality of the survey.
- The list of variables of the SSIA register contains a wide range of important income components related to more than 40% of the target population of the EU-SILC survey. Successful solution of the existing legislation problems in usage of the person level income data from SSIA register would allow a further considerable simplification of the EU-SILC survey questionnaire and reduction of the response burden for persons participating in the EU-SILC survey. Due to a luck of SSIA register data in the current research it was impossible making an analysis of the completeness and other quality aspects of the SSIA register. Nevertheless, in common with the SRS administrative register, it is very likely that the usage of SSIA register and combining the data of this register with the EU-SILC survey will result in a considerable improvement of the total EU-SILC survey quality.

References

Balina S. (2005) The Usage of Administrative Registers for Collecting Income Data. EUROSTAT PHARE Multi – Country Programme organised by ICON-Institute GmbH/Germany on behalf of EUROSTAT. Pilot Project on Statistics on Income and Living Conditions.

Regulation (EC) No 1177/2003 of The European Parliament and of The Council (16 June 2003) Concerning Community Statistics on Income and Living Conditions (EU-SILC)

European Commission Eurostat (2003) Description of Target Variables: Cross-Sectional and Longitudinal, Directorate E: Social and Regional Statistics and Geographical Information System Unit E-2: Living Conditions

Quality Analysis in a Survey on Transportation of Goods by Road

Juris Breidaks[1]

[1] University of Latvia

e-mail: Juris.Breidaks@gmail.com

Central Statistical Bureau of Latvia

e-mail: Juris.Breidaks@csb.gov.lv

Abstract

The paper is devoted to the quality analysis of the ongoing survey "Transportation and Turnover of Goods by Road" organised by the Central Statistical Bureau of Latvia (CSB). This is a continuous survey. The stratified simple random sampling is used.

Starting from 2005 stratification is changed to improve the quality. The stratification from 2002 till 2004 is compared with the stratification from 2005 in the paper. These stratifications are compared by the sample error, bias and mean square error.

1 Introduction

The Survey on Transport of Goods by Road was initiated in January 1997 as a pilot project organized by *Eurostat* under the *Phare Programme*. It is a continuous survey where information about the vehicles in the sample is obtained through questionnaires mailed to respondents. The target of survey is to obtain the information about transportation of goods by road performed by transport vehicles registered in Latvia. The main variables of interest are tonnes transported, tonne-kilometres performed and kilometres travelled loaded for total goods road transport.

The survey covers transport vehicles that are owned by legal and natural persons and which at the moment of sample formation had undergone technical inspection and could be lawfully used. The data of the Road Traffic Safety Directorate about vehicle registrations and the number of vehicles that had undergone technical inspection reveal that only 43.2% of all the registered transport vehicles had passed the yearly technical inspection and could be legally used. Special vehicles such as fire-fighting engines, crane lorries, tower cranes, road repair vehicles and other special vehicles were not included in the survey.

Simple random stratified sampling is used. The weekly sample size is 120 vehicles.

2 Stratification

From 1st January of 2002 till 1st January of 2005 stratification has been made by capacity, place of registration of vehicles and year of release of vehicles.

Table 1 – Stratification in 2004

Stratum	Capacity and place of registration of vehicles	Year of release of the vehicles
1	cap. 1,5t, Riga(including the district of Riga)	All
2	cap. 1,5t, all Latvia without Riga and the district of Riga	All
3	1,5t<cap. ≤ 5t, Riga(including the district of Riga)	All
4	1,5t<cap. ≤ 5t, all Latvia without Riga and the district of Riga	All
5	5t<cap. ≤10t, Riga(including the district of Riga)	1998-2004
6	5t<cap. ≤ 10t, Riga(including the district of Riga)	1991-1997
7	5t<cap. ≤ 10t, Riga(including the district of Riga)	1990
8	5t<cap. ≤ 10t, all Latvia without Riga and the district of Riga	1998-2004
9	5t<cap. ≤ 10t, all Latvia without Riga and the district of Riga	1991-1997
10	5t<cap. ≤ 10t, all Latvia without Riga and the district of Riga	1990
11	cap.>10t, Riga(including the district of Riga)	1998-2004
12	cap.>10t, Riga(including the district of Riga)	1991-1997
13	cap.>10t, Riga(including the district of Riga)	1990
14	cap.>10t, all Latvia without Riga and the district of Riga	1998-2004
15	cap.>10t, all Latvia without Riga and the district of Riga	1991-1997
16	cap.>10t, all Latvia without Riga and the district of Riga	1990
17	the trucks, Riga(including the district of Riga)	1998-2004
18	the trucks, Riga(including the district of Riga)	1991-1997
19	the trucks, Riga(including the district of Riga)	1990
20	the trucks, all Latvia without Riga and the district of Riga	1998-2004
21	the trucks, all Latvia without Riga and the district of Riga	1991-1997
22	the trucks, all Latvia without Riga and the district of Riga	1990

After 1st January of 2005 stratification is made by capacity, place of registration of vehicles, year of release of the vehicles and status.

Table 2 – Stratification after 1st January of 2005

Stratum	Capacity and place of registration of vehicles	Year of release of the vehicles	Status of person
1	cap. 1,5t, Riga(including the district of Riga)	All	Legal
2	cap. 1,5t, all Latvia without Riga and the district of Riga	All	Legal
3	1,5t<cap. ≤ 5t, Riga(including the district of Riga)	All	Legal
4	1,5t<cap. ≤ 5t, all Latvia without Riga and the district of Riga	All	Legal
5	5t<cap. ≤10t, Riga(including the district of Riga)	1999-2005	Legal

Stratum	Capacity and place of registration of vehicles	Year of release of the vehicles	Status of person
6	5t<cap. ≤ 10t, Riga(including the district of Riga)	1992-1998	Legal
7	5t<cap. ≤ 10t, Riga(including the district of Riga)	1991	Legal
8	5t<cap. ≤ 10t, all Latvia without Riga and the district of Riga	1999-2005	Legal
9	5t<cap. ≤ 10t, all Latvia without Riga and the district of Riga	1992-1998	Legal
10	5t<cap. ≤ 10t, all Latvia without Riga and the district of Riga	1991	Legal
11	cap.>10t, Riga(including the district of Riga)	1999-2005	Legal
12	cap.>10t, Riga(including the district of Riga)	1992-1998	Legal
13	cap.>10t, Riga(including the district of Riga)	1991	Legal
14	cap.>10t, all Latvia without Riga and the district of Riga	1999-2005	Legal
15	cap.>10t, all Latvia without Riga and the district of Riga	1992-1998	Legal
16	cap.>10t, all Latvia without Riga and the district of Riga	1991	Legal
17	the trucks, Riga(including the district of Riga)	1999-2005	Legal
18	the trucks, Riga(including the district of Riga)	1992-1998	Legal
19	the trucks, Riga(including the district of Riga)	1991	Legal
20	the trucks, all Latvia without Riga and the district of Riga	1999-2005	Legal
21	the trucks, all Latvia without Riga and the district of Riga	1992-1998	Legal
22	the trucks, all Latvia without Riga and the district of Riga	1991	Legal
23	1,5t<cap. ≤ 5t, all Latvia	All	Private
24	5t<cap., all Latvia	All	Private
25	the trucks, all Latvia	All	Private

There are two types of vehicle. The status of the owner of a vehicle is legal or natural person.

Table 3 – Level of response by status in 2005

Status	Number of respondents	Sample size	Level of response (%)
Legal	5085	5772	88.10
Private	255	468	54.49
Total	5340	6240	85.58

Table 4 – Mean of indicators value in 2005 population

Capacity of vehicles	TONN		TKM		KML2	
	Legal	Private	Legal	Private	Legal	Private
1,5t<cap. 5t	13.4	1.0	971.0	78.8	1253.9	213.3
5t<cap.	330.5	81.5	24233.9	2500.5	1795.6	287.0
the trucks	231.1	244.9	77330.1	23328.2	4232.5	1157.0

Capacity of vehicles	TO_N		TK_N		KM_N	
	Legal	Private	Legal	Private	Legal	Private
1,5t<cap. 5t	13.2	1.0	913.0	78.8	1190.6	213.3
5t<cap.	321.7	81.5	16307.5	2500.5	1199.8	287.0
the trucks	179.3	223.1	12722.4	19027.1	612.7	937.1

Notations for this and further tables

TONN	Tonnes transported for total goods road transport
TKM	Tonne-kilometres performed for total goods road transport
KML2	Kilometres travelled loaded for total goods road transport
TO_N	Tonnes transported for national goods road transport
TK_N	Tonne-kilometres performed for national goods road transport
KM_N	Kilometres travelled loaded for total national road transport

Comparing the mean indicator of ton-kilometres performed and kilometres performed per vehicle in above mentioned vehicle groups, it is obvious that for vehicles owned by legal persons these indicators are higher. The non-response level of natural persons is higher compared to legal persons.

The status (legal or private) of the owner of a vehicle was not used as a stratification indication in the previous stratification. So the combination of these essential differences in these two vehicle groups can lead to biased estimates. The estimates of total indicators of private persons are underestimated if these differences are not taken into account. The estimates of total indicators of legal persons are overestimated because of the same reason. The estimates of total indicators are also overestimated because legal persons have higher response rate and the values of indicators for legal persons are also higher.

3 Comparing of the stratification

3.1 Sample errors of estimates

The variance of estimates was estimated for six indicators – total tonnes transported, total tonne-kilometres performed and total kilometres travelled loaded for total goods road transport and for national goods road transport.

Table 5 – The coefficients of variation for estimates of indicators in year 2004 and 2005

Year	TONN	TKM	KML2	TO_N	TK_N	KM_N
2004	4.210	2.086	2.688	4.575	3.500	4.175
2005	3.666	2.123	2.506	4.038	3.426	3.858

It is obvious from Table 5 that the quality of estimates of indicators concerning to tonnes and ton-kilometres, kilometres has improved, expect sample error of one indicator is worsen

In next capter sample error for the survey in 2005 will be estimated with stratification from 2002 till 2004, and this stratification will be compared with stratification from 2005. The comparison is inconsiderate in table 5, because are compared different population. In next chapter one population of survey was taken.

3.2 The estimation of the sample error with different stratification

It is possible to estimate the sample error if the different stratification is used. Assume that there are two stratification, new stratification with *G* strata and previous stratification with *H* strata. It is possible to compute population size M_g from sampling frame. The standard deviation of *g*-th stratum is

$$S_g^2 = \frac{1}{M_g - 1}\sum_{k=1}^{M_g}\left(y_{gk} - \overline{Y}_g\right)^2 = \frac{1}{M_g - 1}\sum_{k=1}^{M_g} y_{gk}^2 - \frac{M_g}{M_g - 1}\overline{Y}_g^2$$

$\sum_{k=1}^{M_g} y_{gk}^2$ and $\overline{Y}_g^2$ have to be estimated to estimate S_g^2. Estimate of $\sum_{k=1}^{M_g} y_{gk}^2$ is $\sum_{h=1}^{H}\frac{N_h}{n_h}\sum_{i=1}^{n_h} y_{hi}^2 z_{hi}$, where $z_{hi} = \begin{cases} 1, hi \in \theta_g \\ 0, hi \notin \theta_g \end{cases}$; θ_g is the index group of successfully surveyed vehicles belonging to *g*-th stratum. Estimate of $\overline{Y}_g^2$ is

$$\hat{\overline{Y}}_g^2 = \left(\hat{\overline{Y}}_g\right)^2 - \hat{Var}\left(\hat{\overline{Y}}_g\right)$$

$$\hat{\overline{Y}}_g = \frac{\hat{Y}_g}{M_g} = \frac{1}{M_g}\sum_{h=1}^{H}\frac{N_h}{n_h}\sum_{i=1}^{n_h} y_{hi} z_{hi}$$

$$\hat{Var}\left(\hat{\overline{Y}}_g\right) = \frac{1}{M_g^2}\sum_{h=1}^{H} N_h^2\left(\frac{1}{n_h} - \frac{1}{N_h}\right)\delta_h^2$$

$$\delta_h^2 = \frac{1}{n_h - 1}\sum_{i=1}^{n_h}\left(y_{hi} z_{hi} - \frac{1}{n_h}\sum_{t=1}^{n_h} y_{ht} z_{ht}\right)^2$$

So the estimate of S_g^2 is

$$s_g^2 = \frac{1}{M_g - 1}\sum_{h=1}^{H}\frac{N_h}{n_h}\sum_{i=1}^{n_h} y_{hi}^2 z_{hi} -$$
$$-\frac{M_g}{M_g - 1}\left(\left(\frac{1}{M_g}\sum_{h=1}^{H}\frac{N_h}{n_h}\sum_{i=1}^{n_h} y_{hi} z_{hi}\right)^2 - \frac{1}{M_g^2}\sum_{h=1}^{H} N_h^2\left(\frac{1}{n_h} - \frac{1}{N_h}\right)\frac{1}{n_h - 1}\sum_{i=1}^{n_h}\left(y_{hi} z_{hi} - \frac{1}{n_h}\sum_{t=1}^{n_h} y_{ht} z_{ht}\right)^2\right)$$

Two conditions have to realize to estimate S_g^2: $n_h > 1, \forall h$ and $\theta_g \neq \varnothing, \forall g$.

Variance of $\hat{Y}$ is

$$Var\left(\hat{Y}\right) = \sum_{g=1}^{G} M_g^2\left(\frac{1}{m_g} - \frac{1}{M_g}\right)S_g^2$$

Estimate of $Var\left(\hat{Y}\right)$ is

$$\hat{V}ar\left(\hat{Y}\right)=\sum_{g=1}^{G}M_g^2\left(\frac{1}{m_g}-\frac{1}{M_g}\right)s_g^2$$

Table 6 – The coefficients of variation of estimates of indicators using two stratifications

	TONN	TKM	KML2	TO_N	TK_N	KM_N
Stratification from 2002 till 2004	3.740	2.139	2.711	4.122	3.484	4.213
Stratification from 2005	3.666	2.123	2.506	4.038	3.426	3.858

It is obvious that if previous stratification would be used in 2005, than the sample error would be largercomparing with current situation.

3.3 Bias

In this survey there is bias of the estimates. The bias is caused by different reasons. One of the reasons (caused by two types of vehicles – legal and private) is discussed in the paper. Using stratification from 2002 till 2004, the estimation of those indicators are biased for these three indicators – total tonnes transported, total tonne-kilometres performed and total kilometres travelled loaded for total goods road transport.

An estimates of biases in percentage was taken from M. Liberts paper. I assume that the bias of estimates of indicators in percentage has not changed.

Table 7 – Estimates of biases (%) for totals (in 2003)

TONN	TKM	KML2
1.8	0.8	6

Table 8 – Bias of estimates of indicators in 2005 using stratification from 2002 till 2004

TONN	TKM	KML2
938 610.94	67 171 847.49	49 594 073.83

Using stratification from 2005, the estimation of those indicators are unbiased for these three indicators, because the status (legal or natural) of the owner of a vehicle is used as a stratification indication.

3.4 Mean square error (MSE)

Mean square error is

$$MSE\left(\hat{\Theta}\right) = Var\left(\hat{\Theta}\right) + \left(Bias\left(\hat{\Theta}\right)\right)^2$$

where $\mathrm{Var}\left(\hat{\Theta}\right)$ is variance of $\hat{\Theta}$ and $Bias\left(\hat{\Theta}\right)$ is bias of $\hat{\Theta}$.

Table 9 – MSE of estimates of indicators in 2005 using stratification from 2002 till 2004

	TONN	TKM	KML2
Variance	3 803 715 422 997	32 260 798 183 220 800	502 018 849 797 176
Bias	938 611	67 171 847	49 594 074
MSE	4 684 706 032 318	36 772 855 212 612 200	2 961 591 025 714 650

Table 10 – MSE of estimates of indicators in 2005 using stratification from 2005

	TONN	TKM	KML2
Variance	3 654 513 335 476	31 782 088 614 413 800	429 064 493 734 578
Bias	0	0	0
MSE	3 654 513 335 476	31 782 088 614 413 800	429 064 493 734 578

4 Conclusion

Comparing the mean square error of tons transported, ton-kilometres performed and kilometres performed per vehicle, it became obvious that stratification from 2005 is better than previous stratification. The sample error is improved for all indicator, expect sample error of one indicator is worsen. The bias is considerably dominished.

5 References

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

Lapins, J. (1997) Sampling surveys in Latvia: Current situation, problems and future developments. *Statistics in Transition: Journal of the Polish Statistical Association*, 3, 281-292.

Pandutang V, Sukhatme, Balkrishna V. Sukhatme , Shashikala Sukhatme, C. Asok., Sampling Theory of Surveys with Applications, IOWA State University press and Indian Society of Agricultural Statistics, 14-15

Lohr S. L. (1999) *Sampling: Design and Analysis.* Brooks/Cole Publishing Company, Pacific Grove, Calif.

SPSS Inc. (2002) *SPSS® Syntax Reference Guide*

Liberts. M. (2004) *Quality Analysis of a Sample Survey on Transportation of Goods by Road*. http://www.stat.jyu.fi/~knissine/nordstatabs/Sess07SurvSamp.pdf, 9.

On Variance minimization for unequal probability sampling

A. Čiginas

Vilnius University, Lithuania; Statistics Lithuania, Lithuania
e-mail: andrius.ciginas@maf.vu.lt

Abstract

We consider sampling designs, where inclusion (to sample) probabilities are mixtures of two components. The first component is proportional to the size of a population unit (described by means of an auxiliary information available). The second component is the same for every unit. We look for mixtures that minimize variances of various estimators of the population total and show how auxiliary information could help to find an approximate location of such mixtures.

We report theoretical and simulation results in the case of Poisson samples drawn from populations which are generated by a linear regression model.

1 Introduction

Consider the population $\mathcal{U} = \{u_1, \ldots, u_N\}$ and assume that we want to estimate the population parameter $t_y = \sum_{1 \le i \le N} y_i$, where $y_i = y(u_i)$ denotes a measurement of the population unit u_i. For this purpose we draw a sample s from $\mathcal{U}$. Assume that an auxiliary information is available in the form of the vector $x = (x_1, \ldots, x_N)$ with positive coordinates. We call x_i the size of the unit u_i. In the case where the variables y and x are highly correlated it is convenient to take into account the relative weights

$$p_i = x_i/t_x, \qquad t_x = \sum_{i=1}^{N} x_i, \tag{1}$$

when choosing the sampling design. For instance, one can define inclusion (to sample s) probabilities $\pi_i = P(u_i \in s)$ proportional to p_i,

$$\pi_i \approx cp_i, \qquad 1 \le i \le N. \tag{2}$$

Kröger, Särndal and Teikari (2003) give examples of skewed populations and sampling designs with inclusion probabilities close to (2) where the variances of several popular estimators $\hat{t}_y$ of the population total t_y are considerably larger then the variances of the same estimators, but with inclusion probabilities $\pi_i \approx cp_i(h)$, where

$$p_i(h) = (1-h)p_i + h/N, \qquad 1 \le i \le N. \tag{3}$$

Here $h \in [0, 1]$. They consider Horvitz-Thompson, regression and generalized regression (GREG) estimators and sampling designs, where sampling is without replacement and with a fixed sample size. The simulation study shows that the variances of these particular examples are minimized for $h \in (0.2; 0.5)$.

Let us call the value h^* of the parameter h optimal if it minimizes the variance. Generally, it is impossible to find h^* without complete knowledge of the population. Much easier question is whether $h^* > 0$ (i.e., whether inclusion probabilities with the uniform component are preferable) for various sampling designs, various populations and estimators $\hat{t}_y$. Another interesting question is how to make a decision about the location of the minimizer h^*, based on the auxiliary information available.

An attempt to answer these questions in some simple situations is made in the present article. Let us outline our approach. Assume that the population point scatter $\{(y_i, x_i) : 1 \le i \le N\}$ looks as if it had been generated according to a probabilistic model, where $y_1, \dots, y_N$ are assumed to be realized values of independent random variables $Y_1, \dots, Y_N$. Given an estimator $\hat{t}_y$ based on the sample s with the inclusion probabilities $\pi_i \approx cp_i(h)$, let $D_h^* = D_h^*(\hat{t}_y)$ denote the conditional variance of $\hat{t}_y$ given $Y_1 = y_1, \dots, Y_N = y_N$. Furthermore, let D_h denote the expected value of this variance, i.e., $D_h = \mathbf{E}D_h^*$. Assume, for the moment, that in the interval $0 \le h \le 1$ the function $h \to D_h$ has the unique minimizer

$$h_0 = argmin(D_h). \tag{4}$$

Then one may expect that, by the law of large numbers, for large N, the number h_0 is close to the minimizer of the function $h \to D_h^*(\hat{t}_y)$. Therefore, h_0 can be considered as an approximation to the unknown random variable h^*. In order to access the quality of the approximation one would like to evaluate the mean square error $\mathbf{E}(h^* - h_0)^2$ and to compare (expected) values of the target function: $D_{h^*}^*$, $D_{h_0}^*$, D_0^* and D_1^*.

In this article we study the simplest case of the Poisson sample drawn from a population which is generated by a linear regression model (see Särndal, Swensson and Wretman (1992), 226 p.). We have chosen the Poisson sample as a modelling example since here (unique) solutions to the corresponding minimization problems are available and the analysis is relatively simple and lucid.

The article is organized as follows. In Section 2 we introduce the population model and derive the inequality $h_0 > 0$ for two commonly used estimators: Horvitz-Thompson and regression estimator. The approximation h_0 to the random variable h^* can be find numerically, but we also propose explicit approximations to h^*. Examples of a simulation study are reported in Section 3. They demonstrate the empirical evidence of the accuracy of the approximation $h^* \approx h_0$.

2 Results

1. Population. We shall assume that $y_1, \dots y_N$ are realized values of independent random variables $Y_1, \dots, Y_N$ such that for every k,

$$\mathbf{E}(Y_k) = \beta_1 + \beta_2 x_k, \qquad \mathbf{V}(Y_k) = \sigma_k^2. \tag{5}$$

Here $\sigma_1, \dots \sigma_N$ and $x_1, \dots, x_N$ are non-random numbers and $x_k > 0$ for every k. We assume in what follows that $\beta_2 \neq 0$. Later we will assume that $\sigma_k^2 = \sigma^2 x_k^\gamma$, $1 \le k \le N$, $\gamma \in [0, 2]$.

2. Poisson sample includes the unit u_k in the sample s with probability π_k so that the inclusion events for different units are independent. In particular, the random variables $\mathbb{I}_k := \mathbb{I}_{\{u_k \in s\}}$ are independent. Given $n < N$ and $h \in [0,1]$ we choose probabilities

$$\pi_k = \pi_k(h) = np_k(h), \quad 1 \leq k \leq N. \tag{6}$$

Then the expected sample size

$$E(\mathbb{I}_1 + \cdots + \mathbb{I}_N) = \pi_1(h) + \cdots + \pi_N(h) = n.$$

For simplicity of notation we shall assume in what follows that

$$\pi_k(0) < 1, \qquad \text{for every} \qquad k = 1, \ldots, N. \tag{7}$$

Then $\pi_k(h) < 1$ for every $h \in [0,1]$ and $k = 1, \ldots, N$.
We shall show that in the case of the Poisson sample the functions $h \to D_h^*$ and $h \to D_h$ are convex for Horvitz-Thompson and regression estimator. Therefore, the numbers $h_0 = argminD_h$ and $h^* = argminD_h^*$ are well defined.

3. Horvitz-Thompson estimator (HT estimator for short)

$$\hat{t}_{yHT} = \sum_{i=1}^{N} \mathbb{I}_i y_i \pi_i^{-1}$$

is unbiased and its variance

$$D_h^* = \sum_{i=1}^{N} y_i^2 (1 - \pi_i) \pi_i^{-1}. \tag{8}$$

Proposition 1. *The functions $h \to D_h$ and $h \to D_h^*$ are convex. These functions are constants whenever $p_i = N^{-1}$ for every $i = 1, \ldots, N$.*

The next Proposition 2 shows that very often we have $h_0 > 0$. Therefore, the inclusion probabilities (6) with equal probability sampling component of size $h_0 > 0$ lead to a lower variance of HT estimator than the traditional choice of inclusion probabilities (2).

Proposition 2. *Assume that $\sigma_i^2 = \sigma^2 x_i^\gamma$, $\gamma \in [0,2]$. Assume that at least two of probabilities $\{p_i\}$ are distinct.*
(i) Assume that $\gamma \in [0,2)$. If $\beta_1 \beta_2 > 0$ then $0 < h_0 < 1$. If $\beta_1 = 0, \beta_2 \neq 0$ then we have $0 < h_0 < 1$ for $\sigma^2 > 0$ and $h_0 = 0$ for $\sigma^2 = 0$.
(ii) Assume that $\gamma = 2$. If $\beta_1 \beta_2 > 0$ then $0 < h_0 < 1$. If $\beta_1 = 0, \beta_2 \neq 0$ then $h_0 = 0$.

In our presentation at the conference we shall refer results of a simulation study where the values of variances D_h^* are compared for $h = h^*$, $h = h_0$, $h = 0$ and $h = 1$.

4. Regression estimator. It is convenient to treat the cases $\beta_1 = 0$ and $\beta_1 \neq 0$ separately.
4.1. Assume that $\beta_1 = 0$. In this case the regression estimator can be written in the form (see Särndal, Svensson, Wretman (1992))

$$\hat{t}_{yr} = \hat{t}_{yHT} + \hat{B}(t_x - \hat{t}_{xHT}),$$

where

$$\hat{t}_{xHT} = \sum_{k=1}^{N} \mathbb{I}_k x_k \pi_k^{-1}, \qquad \hat{B} = \Big(\sum_{k=1}^{N} \mathbb{I}_k \frac{x_k^2}{\sigma_k^2 \pi_k}\Big)^{-1} \sum_{k=1}^{N} \mathbb{I}_k \frac{x_k y_k}{\sigma_k^2 \pi_k}.$$

The variance formula is rather complex and, therefore, it is convenient to deal with the approximate variance (see ibidem),

$$D_h^* = \sum_{k=1}^{N} (y_k - B x_k)^2 (1 - \pi_k) \pi_k^{-1}, \quad \text{where} \quad B = D^{-1} \sum_{k=1}^{N} \frac{x_k y_k}{\sigma_k^2}$$

and $D = \sum_{k=1}^{N} \sigma_k^{-2} x_k^2$. A simple calculation shows that the expected value $D_h = \mathbf{E} D_h^*$ can be written in the form

$$D_h = \sum_{k=1}^{N} \Big(\frac{1}{n} \frac{1}{p_k(h)} - 1\Big)(\sigma_k^2 - D^{-1} x_k^2). \tag{9}$$

The same argument as above shows that the functions $h \to D_h^*$ and $h \to D_h$ are convex.

4.2. Assume that $\beta_1 \neq 0$. In this case the population size N can be considered as an auxiliary information and we have the regression estimator (see Särndal, Svensson, Wretman (1992))

$$\hat{t}_{yr} = \hat{t}_{yHT} + \hat{B}_1 (N - \hat{t}_{1HT}) + \hat{B}_2 (t_x - \hat{t}_{xHT}).$$

Here $\hat{t}_{1HT} = \sum_{i=1}^{N} \mathbb{I}_i \pi_i^{-1}$. The coefficients

$$\begin{pmatrix} \hat{B}_1 \\ \hat{B}_2 \end{pmatrix} = \Big(\sum_{i=1}^{N} \mathbb{I}_i X_i X_i' / \sigma_i^2 \pi_i\Big)^{-1} \sum_{i=1}^{N} \mathbb{I}_i X_i y_i / \sigma_i^2 \pi_i,$$

where $X_i = \binom{1}{x_i}$. The variance formula of this estimator is rather complex and we shall consider the approximate variance instead (see ibidem)

$$D_h^* = \sum_{k=1}^{N} (\pi_k^{-1} - 1)(y_k - B_1 - x_k B_2)^2, \tag{10}$$

where

$$\begin{pmatrix} B_1 \\ B_2 \end{pmatrix} = \Big(\sum_{i=1}^{N} X_i X_i' / \sigma_i^2\Big)^{-1} \sum_{i=1}^{N} X_i y_i / \sigma_i^2.$$

It is convenient to write the function $D_h = \mathbf{E} D_h^*$ in the form

$$D_h = \sum_{k=1}^{N} \Big(\frac{1}{n} \frac{1}{p_k(h)} - 1\Big)\Big(\sigma_k^2 - \frac{1}{W}(D - 2G x_k + H x_k^2)\Big), \tag{11}$$

where we denote

$$D = \sum_{k=1}^{N} \frac{x_k^2}{\sigma_k^2}, \qquad G = \sum_{k=1}^{N} \frac{x_k}{\sigma_k^2}, \qquad H = \sum_{k=1}^{N} \frac{1}{\sigma_k^2}, \qquad W = DH - G^2.$$

The same argument as above shows that functions $h \to D_h^*$ and $h \to D_h$ are convex.

In both cases expressions of the functions (9) and (11) are complicated for further theoretical analysis (the minimization problem of the functions (9) and (11) can be easily solved numerically), so we shall consider the approximation (see Särndal, Svensson, Wretman (1992))

$$D_h \simeq \sum_{k=1}^{N} \Big(\frac{1}{n} \frac{1}{p_k(h)} - 1 \Big) \sigma_k^2. \tag{12}$$

This approximation is convex function too.

Proposition 3. *Assume that $\sigma_i^2 = \sigma^2 x_i^\gamma$, $\gamma \in [0,2]$. Assume that at least two of probabilities $\{p_i\}$ are distinct. Assume that the functions (9) and (11) are changed by approximation (12). Let $\sigma^2 > 0$.*
(i) Assume that $\gamma = 0$. Then $h_0 = 1$.
(ii) Assume that $\gamma \in (0,2)$. Then $0 < h_0 < 1$.
(iii) Assume that $\gamma = 2$. Then $h_0 = 0$.

5. Explicit approximations to h^*. Assume that $\sigma_k^2 = \sigma^2 x_k^\gamma$, $1 \le k \le N$, $\gamma \in [0,2]$. For HT estimator (after some analytical and statistical assumptions) we have

$$h^* \approx h_{HT} = \frac{\beta_1 + \frac{cv(y)}{2}(1 - \frac{\gamma}{2})\sigma}{\beta_1 + \beta_2 \mu_x + \frac{cv(y)}{2}\mu_\sigma}, \tag{13}$$

where $\mu_x = t_x/N$, $\mu_\sigma = \frac{1}{N}\sum_{i=1}^{N} \sigma_i$ and $cv(y)$ is the coefficient of variation of y in the population $\mathcal{U}$.
For regression estimator can be similarly derived

$$h^* \approx h_R = \frac{(1 - \frac{\gamma}{2})\sigma}{\mu_\sigma}. \tag{14}$$

3 Simulation examples

We fix population size $N = 1000$, expected sample size $n = 100$. Consider auxiliary information vector $\tilde{x}_E$ with coordinates

$$x_i = \Big| \log(1 - \frac{i - 0.5}{N}) \Big|, \qquad 1 \le i \le N.$$

Note that this auxiliary information vector satisfy the condition (7).
Given an auxiliary information vector $\tilde{x}_E$ consider the population models $y_i = 2 + x_i + \sigma_i \eta_i$, where $\sigma_i^2 = \sigma^2 x_i^\gamma$, $\gamma \in \{0; 0.5; 1; 1.5; 2\}$, $1 \le i \le N$. Here $\eta_1, \eta_2, \dots$ denotes the sequence of independent standard normal random variables. For every γ we choose the value of σ so that the expectation of the coefficient of correlation between $\tilde{x}_E$ and y is near 0.9.
The first table report the simulation study of the HT estimator variance (8) and the second table report the simulation study of the regression estimator variance (10). Columns

Table 1 HT estimator

γ	h_0	$\frac{D_{h_0}}{D_0}$	$\frac{D_{h_0}}{D_1}$	E_1	E_2	E_3	E_4
0.0	0.675	0.2106	0.8940	0.2112	0.8937	0.9999	0.9999
0.5	0.668	0.2178	0.8905	0.2176	0.8895	0.9999	0.9999
1.0	0.663	0.2185	0.8877	0.2189	0.8889	0.9999	0.9999
1.5	0.660	0.2183	0.8856	0.2182	0.8865	0.9999	0.9999
2.0	0.658	0.2182	0.8836	0.2184	0.8836	0.9999	0.9999
	$\mathbf{E}(h^*-h_0)^2$	$cv(y)$	h_{HT}	V_1	V_2	V_3	V_4
0.0	3.37E-05	0.372	0.676	4.29E-04	8.33E-06	1.42E-09	1.60E-09
0.5	2.93E-05	0.361	0.671	2.73E-05	1.31E-05	1.27E-09	2.75E-09
1.0	3.07E-05	0.363	0.664	3.49E-06	2.02E-05	1.19E-09	1.10E-09
1.5	4.96E-05	0.356	0.658	2.91E-06	3.03E-05	3.56E-09	4.63E-09
2.0	6.22E-05	0.390	0.652	2.70E-06	4.76E-05	8.00E-09	1.61E-08

Table 2 Regression estimator

γ	h_0	$\frac{D_{h_0}}{D_0}$	$\frac{D_{h_0}}{D_1}$	E_1	E_2	E_3	E_4
0.0	1.000	0.1099	1.0000	0.1227	1.0000	0.9998	0.9998
0.5	0.716	0.4496	0.9362	0.4460	0.9345	0.9994	0.9903
1.0	0.419	0.7831	0.7832	0.7814	0.7857	0.9995	0.9866
1.5	0.161	0.9587	0.6045	0.9582	0.6048	0.9994	0.9895
2.0	0.000	1.0000	0.4458	1.0000	0.4434	0.9999	0.9999
	$\mathbf{E}(h^*-h_0)^2$	$cv(y)$	h_R	V_1	V_2	V_3	V_4
0.0	1.98E-04	0.372	1.000	1.18E-03	4.09E-31	2.21E-07	2.21E-07
0.5	8.33E-04	0.361	0.827	1.06E-03	2.05E-04	3.96E-07	2.36E-05
1.0	8.33E-04	0.363	0.564	4.47E-04	4.86E-04	5.44E-07	2.73E-05
1.5	6.41E-04	0.356	0.272	8.58E-05	1.01E-03	6.45E-07	2.47E-05
2.0	1.65E-05	0.390	0.000	2.09E-31	6.04E-04	2.65E-08	2.65E-08

E_1-E_4 shows the means of the ratios $\frac{D^*_{h_0}}{D^*_0}$, $\frac{D^*_{h_0}}{D^*_1}$, $\frac{D^*_{h^*}}{D^*_{h_0}}$, $\frac{D^*_{h^*}}{D^*_{h_{HT}}}$ (or $\frac{D^*_{h^*}}{D^*_{h_R}}$ for regression estimator) respectively and V_1-V_4 - their variances. Expected values given in the columns E_1-E_4, V_1-V_4 and the mean square error $\mathbf{E}(h^*-h_0)^2$ are evaluated using a tiny Monte Carlo study. We generate 50 independent copies of a given population and evaluate empirical mean values of the parameters of interest. Quantity $cv(y)$ is evaluated using first copy of a given population.

References

Särndal, C.E., Swensson, B., Wretman, J. (1992) *Model assisted survey sampling.* (Springer series in Statistics) Springer-Verlag Berlin, Heidelberg, New York.

Kröger, H., Särndal, C.E., Teikari, I. (2003). Poisson mixture sampling combined with order sampling, *Journal of Official Statistics*, 19, 59–70.

M. Bloznelis and A. Čiginas, On Variance minimization for unequal probability sampling, *Vilnius Univ. Preprint 05-11.*

IMPUTATION IN EU-SILC SURVEY

Andris Fisenko[1], Vita Kozirkova[2]

[1] Central Statistical Bureau of Latvia, Latvia
e-mail: andris.fisenko@csb.gov.lv

[2] Central Statistical Bureau of Latvia, Latvia
University of Latvia
e-mail: vita.besmenova@csb.gov.lv

Abstract

The goal of the research is to analyse imputation methods and make data imputation in EU-SILC survey. It is not yet possible to use administrative data sources in EU-SILC survey. Imputation is recommended by Eurostat. Eurostat has legislated regulation describing situations when it is necessary to use data imputation. Single and multiple imputation methods are analysed in this paper.

1. Introduction

EU-SILC survey is organized in EU countries. The main goal of survey is to achieve reliable statistics about income distribution and social exclusion. Central Statistical Bureau of Latvia (CSB) has started the EU-SILC survey in year 2005. EU-SILC is expected to become the EU reference source for comparative statistics on income distribution and social exclusion at European level. One of EU-SILC regulations states that for better quality of data it is necessary to carry out the imputation of missing data.

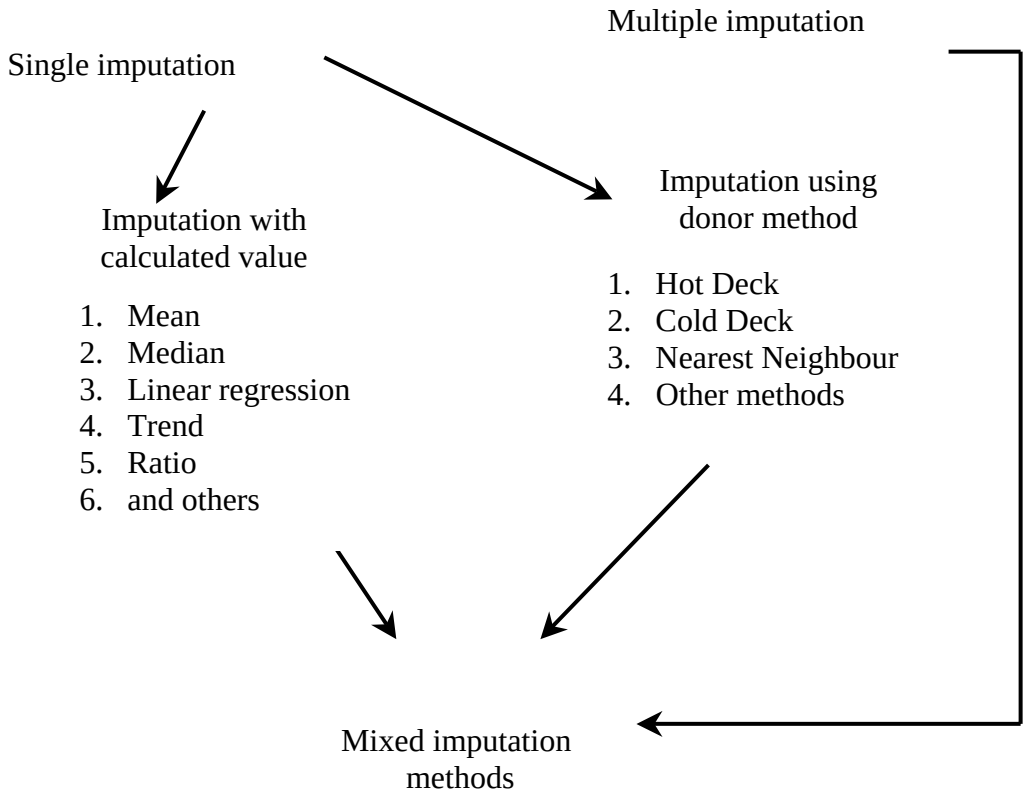
There are two possibilities for imputation: unit and item. Unit non-response usually is reduced by attaching appropriate weights to the responding cases. Item non-response arises when some data are collected for a unit but values of some items are missing, for example the respondent refuses to provide the answer to a sensitive question or is unable to provide the answer to a question requiring complex information. Items missing values are replaced using imputation methods.

2. Description

The traditional approach for imputation in official statistics is to produce just one imputed value for each missing item. It is called single imputation. However, single imputation could create a problem for variance estimation. An alternative approach is to design the imputation method in such a way that a simple variance estimator can be constructed. One such approach is multiple imputation. The basic idea of multiple imputations is to create *m* imputed values for each missing item.

In next chapter strong and weak sides of single imputation methods will be analysed. The general structure of imputations is given in figure 1.

Figure 1: Structure of imputations



3. Analysis

At first imputation analyse is made using in SPSS built-in procedure Replace Missing Value (RMV). These imputation methods are with calculated value. RMV calculates new variables using one of several methods. The estimated values are calculated from valid data in the existing variables.

There are 5 estimation methods for RMV: *Lint* (linear interpolation), *Mean* (mean of surrounding values), *Median* (median of surrounding values), *Smean* (variable mean), *Trend* (linear trend at that point).

Lint replaces missing values using a linear interpolation. The last valid value before the missing value and the first valid value after the missing value is used for the interpolation. If

the first or last case in the series has a missing value, the missing value is not replaced. *Lint* will not replace missing values at the endpoints of variables.

Mean/Median replaces missing values with the *mean/median* of valid surrounding values. The span of nearby points is the number of valid values above and below the missing value used to compute the *mean/median.*

Trend replaces missing values with the linear trend for that point. The existing series is regressed on an index variable scaled 1 to n. Missing values are replaced with their predicted values.

Smean replaces missing values in the new variable with the variable mean. This function is equivalent to the Mean function with a span specification of all.

There are several reasons to refuse RMV. From analyse we get that in 36% of observations RMV methods (*lint, mean, median*) are not imputed missing value. *Smean* method for all missing values gives all equal values. Finally, for income imputation other imputation methods are more recommended. (S.Laaksonen, U.Oetliker, S.Rssler, J.P.Renfer, C.Skinner, 2004)

3.1 Donor methods

Cold-deck imputation uses an external source to the current data collection to "fill-in" the missing item. Frequently a previous iteration of the same survey serves as the external source. This method is historical imputation. As EU-SILC survey is organised for the first time it is not possible to use historical information.

The main principle of the hot deck method is to use the current data (donors) to provide imputed values for records with missing values. In the nearest neighbourhood (NN) method, the missing value is replaced by a value of the donor, which is very close to the covariate of the missing case. NN method is the special case of hot deck method. Cold deck, hot deck and NN are single imputation methods.

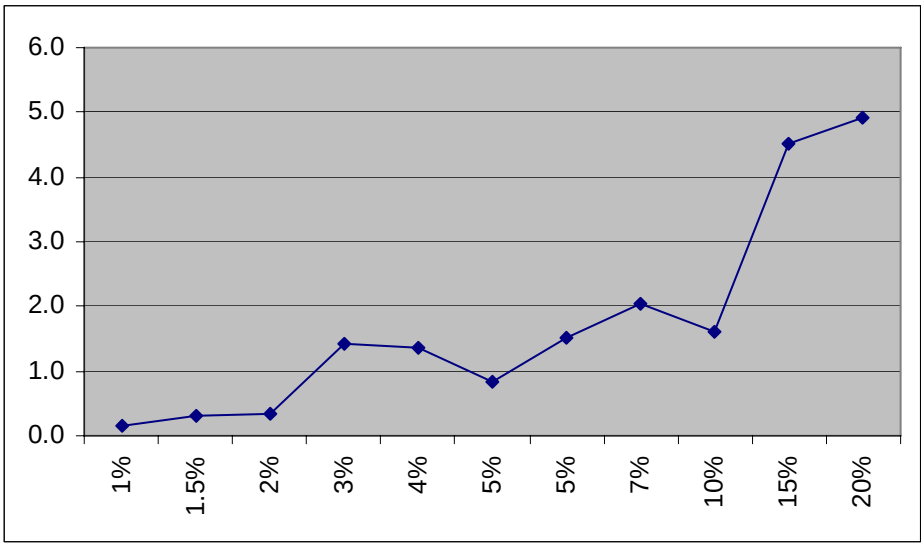
For EU-SILC data multiple imputation method is chosen for imputation. For multiple imputation it is necessary to choose one method from single imputation methods. For EU-SILC survey hot deck method is chosen and it is repeated a couple of times.

Figure 2 and 3 shows some results. There are analysed net amounts of additional payments by person. For this indicator missing data was 1.6 %. The main aim of imputation analyse was to make simulations and calculations of different situations and compare results.

Figure 2. Percentage of missing value compared to full data set.

Number of simulations	Percentage of missing data	Mean	Standar Deviation
	0	1673.3	1511.5
20	1	**1670.8**	**1511.2**
100	1.5	**1672.9**	**1515.7**
50	2	**1670.7**	**1508.7**
21	3	**1677.6**	1529.4
100	4	**1676.3**	1529.1
50	5	**1670.7**	1521.9
100	5	**1670.1**	1531.4
50	7	1682.1	1534.4
20	10	1659.4	1499.8
20	15	1657.5	1457.4
50	20	1693.4	1567.8

Figure 3. Differences in simulations with different value of missing data.



This small research shows how similar are data sets with imputed data compared to original data set. It is important that both indicators (Mean and standard deviation) are analysed together.

4. Conclusions

Analysing imputation methods on EU-SILC survey data there were some difficulties. Results from RMV methods are different as it was expected. Historical data is necessary for cold deck method. Using multiple imputation with hot deck imputation method, we get better results as using RMV imputation methods.

For future research it is necessary to try linear regression, especially if auxiliary information is available.

References

1. R.J.A. Little & D.B.Rubin (2002). Statistical Analysis with missing data. New York: Wiley
2. D.B.Rubin (1987). Statistical Analysis with missing data. New York: Wiley
3. S. Laaksonen, S. Rässler & C. Skinner (2004). DACSEIS research project Workpackage 11 Imputation and Non-Response. Eurostat
4. R.M. Grove, D.A. Dillman, J.L. Elting & R.J.A. Little (2002) Survey Nonresponse. New York: Wiley
5. S. Rassler DACSEIS research paper series No. 5 The Impact of Multiple Imputation for DACSEIS. (2004)
6. S.Laaksonen, U.Oetliker, S.Rssler, J.P.Renfer, C.Skinner DACSEIS research paper series Workpackage 11 Imputation and Non-response, Dliverable 11.2 (2004)
7. D.B. Rubin (1987) Multiply Imputation for Nonresponse in Surveys. US, John Wiley & Sons
8. SPSS® 11.5 Syntax Reference GuideCopyright © 2002 by SPSS Inc

http://www.stat.psu.edu/~jls/mifaq.html

Comparisons of methods for generating conditional Poisson samples

Anton Grafström
Umeå University, Sweden
e-mail: anton.grafstrom@math.umu.se

Abstract

Methods for conditional Poisson sampling (CP-sampling) are compared and the focus is on the efficiency of the methods. The time it takes to generate samples is investigated by simulation in the R-programming language. A new method introduced by Bondesson, Traat & Lundqvist in 2004 is found to be efficient. The new method is an acceptance rejection method that uses the efficient Pareto sampling method.

1 Introduction

Both conditional Poisson sampling (CP-sampling) and Sampford sampling are fixed size πps sampling designs. Thus, the methods can be used to get a sample of fixed size n from a population of size N with unequal inclusion probabilities. In 2004, Bondesson, Traat & Lundqvist introduced new methods for both CP-sampling and Sampford sampling. The new methods use Pareto sampling, which was introduced by Rosén (1997a,b). The methods are acceptance rejection (A-R) methods and they use the fact that the Pareto sampling design is very close to the design of both CP-sampling and Sampford sampling. A Pareto sample, which is rapidly generated, can be adjusted to become a true CP-sample or a Sampford sample by the use of an A-R filter.

In Grafström (2005), methods for both CP-sampling and Sampford sampling were compared. The methods were compared by simulation in the Matlab programming language and the new methods were found to be efficient. The focus in this text is on the methods for CP-sampling and we present some simulation results using the R-programming language. It is more appealing to use R since it is a free software which is specialised on statistical computing and it is widely used. Four methods for CP-sampling are compared and we wonder which method is the most efficient one.

CP-sampling is a modification of Poisson sampling. Let p_i be the given target inclusion probability for unit i, $i = 1, ..., N$. Each unit i in the population is included with probability p_i but only samples of size n are accepted. Usually it is assumed that $\sum_{i=1}^{N} p_i = n$ since it will maximize the probability to get samples

of size n. The assumption $\sum_{i=1}^{N} p_i = n$ is not restrictive. If it is not satisfied, the p_is can be transformed to satisfy that condition (Hajek, 1981, p. 66, Broström and Nilsson, 2000). When using CP-sampling, the true inclusion probabilities will only be approximately p_i. However, there is a possibility to adjust the p_is to obtain desired inclusion probabilities (Dupacova, 1979, Chen *et al.*, 1994, Aires, 2000, Tillé, 2005).

In section 2 there is a description of each of the sampling methods. Then in section 3, the methods are tested by simulation in some different sampling situations. The conclusions are presented in section 4.

2 The methods

The different sampling methods are described in this section.

2.1 CP-reject

The CP-reject method for CP-sampling can be found in Hajek (1981). Let the target inclusion probability for unit i be p_i with $\sum_{i=1}^{N} p_i = n$. Also, let I_i be independent and $Bin(1, p_i)$ distributed inclusion variables. Then unit i is included in the sample if $I_i = 1$. Simulate I_i for $i = 1, ..., N$ and accept the sample as a CP-sample if $\sum_{i=1}^{N} I_i = n$. Repeat the procedure until a sample is accepted.

2.2 CP-with replacement

CP-with replacement (Hajek, 1981) is another method for CP-sampling. Let the target inclusion probability for unit i be p_i with $\sum_{i=1}^{N} p_i = n$. Draw n units with replacement where unit i is drawn with probability $p_i' \propto p_i/(1 - p_i)$ and $\sum_{i=1}^{N} p_i' = 1$. If all n units are distinct, the sample is accepted as a CP-sample. Otherwise the procedure is repeated from the beginning.

2.3 CP-list sequential

The CP-list sequential method uses the definition of conditional probability and it was found to be efficient by Öhlund (1999). The method can also be found in Chen & Liu (1997), Traat *et al.* (2004) and in Tillé (2005). Let the target inclusion probability for unit i be p_i and $\sum_{i=1}^{N} p_i = n$. Also, let I_i be independent

$Bin(1, p_i)$ distributed random inclusion variables. Then the inclusion variables I_i can be successively generated from the conditional distributions

$$P\left(I_i = x \left| \sum_{j=i}^{N} I_j = n - n_{i-1} \right.\right), \quad x = 0, 1,$$

where $n_{i-1} = \sum_{j=0}^{i-1} I_j$ and $I_0 = 0$. We will always get a sample of size n. The conditional probabilities can be written as

$$P\left(I_i = 1 \left| \sum_{j=i}^{N} I_j = n - n_{i-1} \right.\right) = \frac{P\left(I_i = 1\right) P\left(\sum_{j=i+1}^{N} I_j = n - n_{i-1} - 1\right)}{P\left(\sum_{j=i}^{N} I_j = n - n_{i-1}\right)}.$$

To use this formula, one first has to calculate the probabilities $P\left(\sum_{j=i}^{N} I_j = k\right)$ for all i and k. That can be done recursively. Fortunately these probabilities need only to be calculated once. Then they can be used to generate as many samples as desired. The calculation may still be too time-consuming if N and n are large. Then it is possible to calculate only some of the probabilities exactly and use normal approximations for the rest of them.

2.4 Pareto sampling

Pareto sampling (Rosén, 1997a,b) is used to select a sample of fixed size n from a population of size N. Let λ_i be the given target inclusion probability for unit i and $\sum_{i=1}^{N} \lambda_i = n$. The method works as follows.

Generate $U_1, U_2, ..., U_N$, where the U_is are independent $U(0, 1)$ variables. Then calculate the Pareto ranking variables

$$Q_i = \frac{U_i/(1 - U_i)}{\lambda_i/(1 - \lambda_i)}$$

for each unit. Select the n units with the smallest Q-values as a Pareto sample of fixed size n. The true inclusion probabilities will be approximately λ_i.

2.5 CP-sampling via Pareto sampling

CP-sampling via Pareto sampling is the new method that was introduced by Bondesson, Traat & Lundqvist (2004). Let the target inclusion probability for unit i be p_i and $\sum_{i=1}^{N} p_i = n$. First a Pareto sample is generated with $\lambda_i = p_i,\ i = 1, ..., N$. Then the Pareto sample is either rejected or accepted as a CP-sample using the probability functions for the Pareto and CP designs. Let

$\mathbf{I} = (I_1, I_2, ...I_N)$ be the vector of random inclusion variables, i.e. $I_i \in \{0, 1\}$ and if $I_i = 1$ then unit i is sampled. Also, let $|\mathbf{I}| = \sum_{i=1}^{N} I_i = n$ be the sample size. The probability functions $p(\mathbf{x}) = P(\mathbf{I} = \mathbf{x})$ for the designs can then be written as

$$p_{CP}(\mathbf{x}) = C_{CP} \prod p_i^{x_i} (1 - p_i)^{1-x_i}, \quad |\mathbf{x}| = n,$$

and, for $\lambda_i = p_i$,

$$p_{Par}(\mathbf{x}) = \prod p_i^{x_i} (1 - p_i)^{1-x_i} \times \sum c_k x_k, \quad |\mathbf{x}| = n,$$

where

$$c_k = \int_0^\infty x^{n-1} \prod \frac{1 + \tau_i}{1 + \tau_i x} \cdot \frac{1}{1 + \tau_k x} dx \quad \text{and} \quad \tau_i = \frac{p_i}{1 - p_i}.$$

The sums and products are taken over the integers $1, 2, ..., N$. The constant C_{CP} is found from the normalizing condition $\sum_{\mathbf{x}:|\mathbf{x}|=n} p(\mathbf{x}) = 1$. We also have $C_{CP} \approx \sqrt{2\pi d}$ for large values of $d = \sum p_i(1 - p_i)$. The c_ks can be calculated exactly or approximated by Laplace approximations. One approximation is

$$c_k \approx c_k^* = (1 - p_k)\sqrt{2\pi}\sigma_k \exp\{\sigma_k^2 p_k^2/2\}, \text{ where } \sigma_k^2 = \frac{1}{d + p_k(1 - p_k)}.$$

This approximation can be improved by the following calibration

$$c_k^{*(cal)} = \frac{(N - n)c_k^*}{\sum_i c_i^*} c_0, \text{ where } c_0 = \int_0^\infty x^{n-1} \prod \frac{1 + \tau_i}{1 + \tau_i x} dx.$$

The constant c_0 can be calculated exactly or approximated by $c_0^* = \sqrt{2\pi/d}$. See Bondesson, Traat & Lundqvist (2004) for a full description of these approximations.

Now let us consider when we can accept a Pareto sample as a CP-sample. Let $p_1(\cdot)$ and $p_2(\cdot)$ be two probability functions. If there exists a constant B such that $p_1(\mathbf{x}) \leq Bp_2(\mathbf{x})$ for all $\mathbf{x}$, then a sample from $p_2(\cdot)$ can be generated and accepted as a sample from $p_1(\cdot)$ if $U \leq p_1(\mathbf{x})/(Bp_2(\mathbf{x}))$, where U is a random number from $U(0, 1)$. The procedure is repeated from the beginning until a sample is accepted.

If $p_1(\cdot) = p_{CP}(\cdot)$ without C_{CP} and $p_2(\cdot) = p_{Par}(\cdot)$, then the constant B must be chosen so that $1 \leq B \sum c_k x_k$ for all $\mathbf{x}$. If the probabilities $p_i,\ i = 1, ..., N$, are given in increasing order, then the $c_k s$ will decrease. The best choice of B will be $B^{-1} = \sum_{k=m}^{N} c_k$ where $m = N - n + 1$.

The conditional acceptance rate for accepting a Pareto sample as a CP-sample is

$$CAR(\mathbf{x}) = \frac{1}{B \sum c_k x_k}.$$

Thus a generated Pareto sample with $\lambda_i = p_i$ will be accepted as a CP-sample if $U \leq CAR(\mathbf{x})$, where $U \sim U(0, 1)$. See Bondesson, Traat & Lundqvist (2004) for more details.

3 Simulation and results

The sampling methods have been implemented in the R-programming language. For CP-sampling via Pareto we have used the calibrated Laplace approximation for calculation of the c_ks. In the CP-list sequential method all necessary probabilities for sums are calculated exactly.

The methods are first tested on a relatively small population and then a larger population is used, where the differences are more apparent.

Example 1. Sampling from the MU284 population. The population that consists of the 284 municipalities of Sweden is called the MU284 population and can be found in Särndal, Swensson & Wretman (1992, pp. 652-659). We use the variable P85, which is the population size in a municipal in the year 1985. Sampling is performed proportional to the size of the population (P85) in each municipal. We generated 1000 samples of size 50 and the results can be found in Table 1. The acceptance rate for CP-with replacement was too low for that method to be used in this example.

Table 1: Results for the MU284 population. We generated 1000 samples of size 50. The times are in seconds and $\hat{AR}_{Sim}$ is the acceptance rate for this simulation.

Method	n	Prel. calc.	Mean time	Total time	$\hat{AR}_{Sim}$
CP-reject	50	0	0.00232	2.32	0.069
CP-list sequential	50	0.66	0.01182	12.82	1
CP via Pareto	50	0	0.00336	3.36	0.791

We see from Table 1 that CP-reject has the lowest mean time. The simplicity of that method makes it efficient as long as the acceptance rate is not too low. CP-sampling via Pareto is also quite efficient and the high acceptance rate (0.791) implies that the probability functions for CP and Pareto are close. The CP-list sequential method is not as efficient as the other methods.

Example 2. Sampling from a large population. Let $N = 10000$ be the population size and $n = 2000$ be the sample size. Also let the target inclusion probabilities be

$$p_1 = 0.1,\ p_2 = 0.15,\ p_3 = 0.2,\ p_4 = 0.25,\ p_5 = 0.3,$$

where each p-value is used for 2000 units (thus we have $\sum_{i=1}^{N} p_i = n$). The acceptance rate for CP-with replacement was too low for that method to be used in this example. We generated 100 samples of size 2000 from this population and the results can be found in Table 2.

Table 2: Results for the large population. We generated 100 samples of size 2000. The times are in seconds and $\hat{AR}_{Sim}$ is the acceptance rate for this simulation.

Method	Prel. calc.	Mean time	Total time	$\hat{AR}_{Sim}$
CP-reject	0	0.373	37.31	0.009
CP-list sequential	616	0.3422	650.22	1
CP via Pareto	0	0.0501	5.01	0.901

In Table 2, we see that CP via Pareto has the lowest mean time. We also see that the acceptance rate (0.901) is even higher than in Example 1. If we look at the acceptance rate for CP-reject, we see that it is much lower now than in Example 1. The time for preliminary calculations in the list sequential method has increased a lot. After the preliminary calculations have been performed, the method is a little bit more efficient than CP-reject.

4 Conclusions

We found that the method CP-reject is efficient for sampling from a small population, but the acceptance rate decreases when the population size increases. We found that CP-with replacement is efficient only when n is much smaller than N. The method becomes inefficient very fast when the sample size n increases. The CP-list sequential method has preliminary calculations and the time for these calculations increases rapidly when the sample size n and the population size N increases. However, after the preliminary calculations have been performed the method is quite efficient. CP-sampling via Pareto seems to be very efficient in all situations. We have used Laplace approximation of the c_ks, but the approximation is very good and it makes this method faster than if the c_ks are calculated exactly. It is also easy to implement. The time it takes to generate a sample with this method is rather independent of the sample size n. The new method is the most efficient one in general, but not always. If the population and the sample size are not too big, then the list sequential method can be efficient and useful (Öhlund, 1999). The list sequential method might even be the most efficient one if many samples are to be generated, since the samples always are accepted.

References

Aires, N. (2000). *Techniques to calculate exact inclusion probabilities for conditional Poisson sampling and Pareto πps sampling designs.* Doctoral thesis, Chalmers University of technology and Göteborg Univerity, Göteborg, Sweden.

Bondesson, L., Traat, I., Lundqvist, A. (2004). Pareto Sampling versus Sampford and Conditional Poisson Sampling. Research Report No. 6 2004, Department of Mathematical Statistics, Umeå University. *To appear in Scand. J. Statist.*

Broström, G. & Nilsson, L. (2000). Acceptance-Rejection sampling from the conditional distribution of independent discrete random variables, given their sum. *Statistics* **34**, 247-257.

Chen, S.X., Dempster, A.P. & Liu, J.S. (1994). Weighted finite population sampling to maximize entropy. *Biometrica* **81**, 457-469.

Chen, S.X. and Liu, J.S. (1997). Statistical applications of the Poissonbinomial and conditional Bernoulli distributions. *Statistica Sinica* **7**, 875-892.

Dupacova, J. (1979). A note on Rejective Sampling, *Contributions to Statistics, Jaroslav Hajek Memorial Volume.* Reidel, Holland and Academia, Prague, 71-78.

Grafström, A. (2005). Comparisons of methods for generating conditional Poisson samples and Sampford samples. Master's thesis, Department of Mathematics and Mathematical Statistics, University of Umeå, Sweden.

Hajek, J. (1981). *Sampling from a finite population.* Marcel Dekker, New York.

Öhlund, A. (1999). Jämförelse av olika metoder att generera Bernoullifördelade slumptal givet deras summa (Comparisons of different methods to generate Bernoulli distributed random numbers given their sum). Master's thesis, Department of Mathematical Statistics, University of Umeå, Sweden.

Rosén, B. (1997a). Asymptotic theory for order sampling. *J. Statist. Plann. Inference* **62**, 135-158.

Rosén, B. (1997b). On sampling with probability proportional to size. *J. Statist. Plann. Inference* **62**, 159-191.

Särndal, C-E, Swensson, B. & Wretman, J. (1992). *Model assisted survey sampling.* Springer-Verlag, New York.

Tillé, Y. (2005). *Sampling algorithms.* Technical Report, Neuchâtel, Switzerland.

Traat, I., Bondesson, L. & Meister, K. (2004). Sampling design and sample selection through distribution theory. *J. Statist. Plann. Inference* **123**, 395-413.

ANALYSIS OF GREG ESTIMATOR IN FARM SURVEY

Olga Grakoviča

The Central Statistical Bureau of Latvia
University of Latvia
e-mail: olga.grakovica@csb.gov.lv

Abstract

The aim of the research is to compare Horvitz-Thompson (HT) estimator with GREG estimator in farm survey. It is also a good practice in R program, because I have used R only for calibration of weights before. The analysis is based on simulations. The object of research is the sample of Farm Structure Survey (FSS) 2005. Sample of FSS is considered as a population. This sample is big enough to make sub-sampling from it. A number of samples are made by simulation and totals are estimated using HT and GREG estimators for each sample. The sampling error is analysed for both estimators.

1 Sampling

The sample of FSS survey is made using stratified sampling. The farms are stratified in 120 strata. Stratification is made according to territorial location, group of specification and area of agricultural land. There are 58 429 responding farms in FSS. 45% of farms are selected in sub-sample for simulations. The sample allocation is determinate based on proportional allocation and Neyman optimal allocation in each stratum. Proportional allocation:

$$n_h = \frac{n \cdot N_h}{N} \tag{1}$$

Where n_h – the sample size in strata h;

n – total sample size.

Neyman optimal allocation:

$$n_h = n \cdot \frac{N_h S_h}{\sum_{h=1}^{H} N_h S_h} \tag{2}$$

Where n_h – the sample size in the strata h;

n – total sample size;

N_h – the number of farms in the strata h;

S_h^2 – the variance in the strata h;

N – total number of farms.

Where $$S_h^2 = \frac{1}{n_h - 1}\sum_{i=1}^{n}\left(y_{h_i} - \bar{y}_h\right)^2 . \quad (3)$$

2 Estimators

2.1 HT Estimator

Design weights d_k are used to calculate Horvitz – Thompson estimate (HT estimate).

$$\hat{Y}_{HT} = \sum_s d_k y_k \quad (4)$$

2.2 GREG Estimator

Definition: U={1,….,k,….,N} – target population; s – sample consisting of *n* elements. A wider and more efficient class of estimators are those that use auxiliary information explicitly at the estimation stage. Some information may already have been used at the design stage. Denote the auxiliary vector by x^*. It is constructed from one or more auxiliary variables. We assume that the population total, $\sum_U x_k^*$, is known. The total $\sum_U x_k^*$ represents information available about population *U*. When the value x_k^* is specified in the sampling frame for every element $k \in U$, we can simply sum the values x_k^* to obtain $\sum_U x_k^*$.

Given this setting for the auxiliary information, the theory of regression estimation forms a basis for constructing an estimator of $Y = \sum_U y_k$. An estimator that uses the information $\sum_U x_k^*$ is the generalized regression estimator (GREG estimator). It is given by

$$\hat{Y}_{GREG} = \hat{Y}_{HT} + (\sum_U x_k^* - \sum_s d_k x_k^*)' B_{s;d} \quad (5)$$

where

$$B_{s;d} = (\sum_s d_k c_k x_k^* (x_k^*)')^{-1} (\sum_s d_k c_k x_k^* y_k) \quad (6)$$

is a vector of regression coefficients, obtained by fitting the regression of y on x^*, using the data $(y_{k,} x_k^*)$ for the elements $k \in s$. In $B_{s;d}$, c_k are weights specified by the statisticians. The standard choice is $c_k = 1$ for all k.

We can also write $\hat{Y}_{GREG}$ as

$$\hat{Y}_{GREG} = \sum_U \hat{y}_k + \sum_s d_k (y_k - \hat{y}_k) \quad (7)$$

where $\hat{y}_k = (x_k^*)' B_{s;d}$. This form highlights the idea of prediction of the non-observed y_k values.

It is helpful to express the GREG estimator given by (5) and (6) as a sum of weighted observed values y_k . We have

$$\hat{Y}_{GREG} = \sum_s d_k g_k y_k \tag{8}$$

where the total weight given to the value y_k is the product of two weights, the design weights $d_k = 1/\pi_k$, and the weight g_k , which depends both on the element k and the whole sample s of which k is a member. It is given by

$$g_k = 1 + \lambda_s' c_k x_k^* . \tag{9}$$

where $\lambda_s' = (\sum_U x_k^* - \sum_s d_k x_k^*)'(\sum_s d_k c_k x_k^* (x_k^*)')^{-1}$. The value g_k is near unity for a majority of the elements $k \in s$, because λ_s' is near the zero vector.

When we apply the weights system $d_k g_k$ for $k \in s$ to the auxiliary vector x_k^* , and sum over s , we obtain an estimate of the population total of x_k^* . This estimate agrees exactly with the known values of total, that is,

$$\sum_s d_k g_k x_k^* = \sum_U X_k^* \tag{10}$$

The weight system is called calibrated. So we obtain for Y_q the estimator

$$\hat{Y}_{qGREG} = \sum_s d_k g_k y_{qk} . \tag{11}$$

3 Simulation and Results

By simulation number of samples is created and two totals (Y_1 – total number of cattle, Y_2 – total area of agricultural land) are estimated using HT and GREG estimator from each sample.

Matrix and two vectors are necessary for calibration – vector of auxiliary information (x^*), vector of inclusion probabilities *p=1/d* and matrix of auxiliary data for units in sample. The calibration is done for each stratum separately. Auxiliary vector x^* consists of information about total area of farms in each stratum. The variable containing information about area of each farm is restructured to matrix. It looks like this:

***x* in stratum 1**	**…**	***x* in stratum *n***
x_{1_1} . . x_{1a}	…	*0* . . *0*
. . .		. . .
0 . . *0*	…	x_{n_1} . . x_{n_z}

Standard error of total is estimated directly from totals estimated in simulation. Estimates of standard error of totals depending on number of simulations are shown in following table:

Number of simulations	200	300	400	500	600
$SE(Y_{1HT})$	267.0386	239.7611	265.8321	237.1940	261.8456
$SE(Y_{1GREG})$	256.1897	218.6111	255.2800	234.8421	238.6668
$SE(Y_{2HT})$	330.4524	354.7256	349.0017	335.9367	238.6668
$SE(Y_{2GREG})$	160.8756	163.4265	154.8191	162.2570	166.9982

In following tables we can see how agricultural land area (Y_2) correlates with total land area of farm (X^*) and number of cattle (Y_1) with total land area of farm (X^*):

Correlations

		Y_2	X^*
Y_2	Pearson Correlation	1	.960(**)
	Sig. (2-tailed)	.	.000
	N	55975	55975
X^*	Pearson Correlation	.960(**)	1
	Sig. (2-tailed)	.000	.
	N	55975	55975

		X^*	Y_1
X^*	Pearson Correlation	1	.567(**)
	Sig. (2-tailed)	.	.000
	N	55975	55975
Y_1	Pearson Correlation	.567(**)	1
	Sig. (2-tailed)	.000	.
	N	55975	55975

4 CONCLUSION

According to results it is easy to understand that GREG estimator depends of correlation among *X** and *Y*. If correlation is small, GREG estimator is not better compared to HT estimator.

REFERENCES

Särndal, C-E. Lundström, S (2005) Estimation in Surveys with Nonresponse. John Wiley and Sons, Inc.

Detection and Considering of Extremal Elements for Business Surveys

Oksana Honchar[1]

[1] Scientific and Technical Complex of Statistical Research, Ukraine;
e-mail: ohonchar@list.ru

Abstract

The problem of detection and considering of outliers, such as those encountered in many business sample surveys is discussed in this paper. Different aims of extremal elements detection are distinguished and methods that must be applied for these are given. Small-scale enterprises are described in this paper.

1. Introduction

For detection of extremal elements we use statistical tools (it allows to detect them in quantitative way) and logical analysis that consider qualitative factors of detection and considering of extremal elements. Decision about extremal elements of population must be based on complex approach to problem, in harmonious combination of statistical tools and logical analysis.

2. Definition

Extremal elements (outliers) are elements of population that differ from others. Except these elements in sample survey we also consider ones that have essential influence on the precision of the sample estimates. Certainly some characteristics of enterprises not only have significant impact on the precision of the sample estimates but also are greatly different from the other elements of this population.

The population elements that differ from others may be result of either introduced registration error or objective reasons. In case of data verification enterprises which characteristics greatly differ from others are called atypical enterprises.

Therefore it is the author's opinion that it is necessary to discern term extremal elements for atypical enterprises (units). However in practice the terms are considered synonyms and are distinguished one from another in case when it is necessary to show peculiarities of individual enterprises.

Also it is necessary to differ qualitative and quantitative extremes. If enterprise is typical but its characteristics are in some sense different or more significant than others then we deal with quantitative extreme. But if enterprise due to some objective reasons is different from the others then it is qualitative extreme.

Outliers which are registration errors usually are quantitative extremes. Enterprises that essentially influence on the precision of the sample estimates but are not atypical also are the quantitative extremes because they begin to worsen the precision of the sample estimates due to the fact that characteristics of enterprise exceed some bounds.

3. Two methods for detection of extremal elements

3.1. Method of precision control

To define method that may be used for detection extremal elements it is necessary to specify the purpose for which we find them. The method of precision control decrease the precision of the sample estimates on the planing stage. For population of small-scale enterprises that is stratified by the real economic activity and enterprise size this method consists of following.

Strata put in order by some criteria (for example, by stratum number). Suppose that mean $\bar{x}_h$ and standard deviation σ_h is needed for each stratum h. For the whole population coefficient of variation can be calculated as:

$$\overline{V^2} = \frac{\overline{\sigma^2}}{\bar{x}^2} \tag{1}$$

where $\overline{\sigma^2}$ is mean of group variances σ_h; overall mean $\bar{x}$ can be calculated as weighted mean of group means $\bar{x}_h$.

A sample size may be estimated from the $\overline{V^2}$ values as

$$n' = \frac{t^2 \overline{V^2} N}{\hat{h}^2 N + t^2 \overline{V^2}} \tag{2}$$

where t - quantile of normal distribution; $\hat{h}$ - relative limiting sampling error that is assigned by statistician on the planing sample stage; N - population size.

In the first stratum the largest element is extracted and the all mentioned characteristics are computed again. The total sample size is $n = n' + j_h$, where j_h is number of extremal elements in stratum. If total sample size is reduced then separating element is extremal. In this case we pick out the largest residuary element and calculate mentioned values. If total sample size rises then separating element is not extremal. Thus this element returns in it's stratum. The quantity extremal elements in stratum equals the number of separation steps in this stratum. After that these procedures are executed for next stratum. Extremal elements are placed in individual stratum that must be observed with 100% probability.

It is significant that the sample size and precision are functionally dependent. If we reduce sample size due to considering of extremal elements then for fixed sample size precision decreases. Hence influence on the sample size under fixed precision is similar to influence on precision under fixed sample size. Thus this method is used for precision control.

If precise data must be obtained for the economic activity then this method is applied but each economic activity is considered single population. Of course there are more extremal elements for

the economic activities than for whole population. In practice precision for the economic activities in formula of the sample size is usually smaller then for whole population. Thus there may be more extremal elements for whole population than for the economic activities.

3.2. Distance-based method (L)

If we are interested in elements that differ from others in some economic activity then we can apply one of methods that are based on distance. In part we suggest using method that rests on calculation of value:

$$L = \frac{l}{R} * 100\%, \tag{3}$$

where l is distance to previous element or

$$l = x_n - x_{n-1};$$

R is range of deviation or

$$R = x_n - x_1.$$

Thus

$$L = \frac{x_n - x_{n-1}}{x_n - x_1} * 100\%. \tag{4}$$

The value L shows the percentage of distance to previous element in deviation range. This method provides for the statistician the advantage of choosing its standard. For example, we are interested in investigating elements that differ from others more than 20%. Thus in this case the standard is 20%.

This method can be used both for economic activities and for strata. In economic activities it is important for analysis of atypical enterprises. In strata it is used for increasing homogeneity of these strata.

Note that in practise in strata there are more atypical enterprises then in population of economic activity. Usually in economic activity strata overlap absorbs extremal element of stratum. Within economic activity extremal elements can be largest elements of strata overlapping. They belong either to single most significant stratum or to different strata (these elements are contained in tails of these strata). Consequently we rather have atypical enterprises in strata than in economic activity population. Usually there is single standard for value L. Although frequently for economic activity population standard for value L must be lower than for stratum.

Conclusion

In practice we encounter with rather sizeable stratum of atypical enterprises given by fundamental methods of detection of extremal elements. In this case method of the precision control plays significant part because it gives moderate number of the most extremal elements. Moreover it defines criterion and scale of extreme.

L-method is only one of distance-based methods.

References

1. *Bartkowiak A., Szustalewicz A.* Outliers – finding and classifying which genuine and which spurious// Computational Statistics, 15, 3-12,, 2000.

2. *Hidiroglou M.A., Berthelot J.-M.* Statistical Editing and Imputation for Periodic Business Surveys// Survey Methodology, 12, pp. 73-84, 1986.

3. *Hyunshik L.* Outliers in Business Survey Methods, Ed. Wiley, 1995.

4. *Васєчко О.О.* Нові підходи до виявлення нетипових підприємств у структурній статистиці// Статистика України, 2004. - №3. – с. 7-12.

3. *Шикуленко А.В*. Методологічні підходи до виявлення нетипових підприємств у вибіркових обстеженнях// Статистика України, 2005. - №1.

VARIANCE ESTIMATION IN EU-SILC SURVEY

Mārtiņš Liberts[1]

[1] Central Statistical Bureau of Latvia
e-mail: Martins.Liberts@csb.gov.lv

Abstract

The task of the contribution is to develop methodology for estimation of sampling error for non-linear statistics. Estimation of sampling error for total, the ratio of two totals and Gini coefficient are considered. However developed methodology could be used also for other statistics. Re-sampling variance estimation methods are used – dependent random groups and jackknife. The result of the contribution is a developed program (in SPSS) for estimation of sampling error for arbitrary sample with broad possibilities of fine-tuning the parameters of methods applied. The data of two sample surveys organised by Central Statistical Bureau of Latvia is used – data of Household Budget Survey and Survey on European Union Statistics on Income and Living Conditions (EU-SILC).

Keywords: estimation of sampling error, re-sampling, dependent random groups, jackknife, Gini coefficient.

1 Introduction

Measurement of accuracy is important part in production of statistics based on survey sampling. The most common measure of accuracy is sampling error. The task of study is to develop methodology for estimation of sampling errors for complex (nonlinear) statistics and to apply it to Household Budget Survey (HBS) and EU-SILC (SILC).

2 Population Parameters

Three types of population parameters will be considered in the paper – total, the ratio of two totals and Gini coefficient.

Parameter	Population parameter	Estimate of parameter
Total	$X = \sum_{i=1}^{N} x_i$	$\hat{X} = \sum_{i=1}^{n} x_i w_i$
The ratio of two totals	$R = \frac{X}{Y}$	$\hat{R} = \frac{\hat{X}}{\hat{Y}}$
Gini index	$G = 100\left(\frac{2\sum_{i=1}^{N} x_i R_i - X}{NX} - 1 \right)$, $R_i = \sum_{j=1}^{i} 1$ – Rank of unit i if sorted ascending by x_i	$\hat{G} = 100\left(\frac{2\sum_{i=1}^{n} x_i w_i \hat{R}_i - \hat{X}}{N\hat{X}} - 1 \right)$, $\hat{R}_i = \sum_{j=1}^{i} w_i$ – Estimate of rank of unit i if sorted ascending by x_i

3 Design of Surveys

Both surveys considered in the study share similar design. Households and individuals are survey units. Two-stage sampling is used for households; two-stage cluster sample is used for individuals.

Stratified systematic *pps* (sampling with probability proportional to size) sample of population census (2000) areas is used at the first stage. Stratification is made by degree of urbanisation – Riga, 6 other largest cities, towns and rural areas (four strata). PSUs are selected by several starting points (6 or 3 for HBS, 4 for SILC).

Simple random sampling of households is used at second stage.

All individuals from selected households are sampled – so households form clusters of individuals.

4 Estimation of Sampling Errors

It is hard to find direct estimators of sampling errors for estimates of complex statistics – especially in case of sampling design described in previous section. The approximation methods are used as alternative. Re-sampling methods (dependent random groups and jackknife) and linearization methods are considered in the paper.

4.1 Dependent Random Groups

The sample s from population U is divided in A non-overlapping subgroups $s_1,\ldots,s_A$. The sample s should be divided so that all subgroups preserve the same sampling design as the sample s. The estimate of population parameter θ could be estimated as $\hat{\theta}_1,\ldots,\hat{\theta}_A$. It is possible to estimate a variance of $\hat{\theta}$ by

$$\hat{V}_{DRG2} = \frac{1}{A(A-1)}\sum_{a=1}^{A}\left(\hat{\theta}_a - \hat{\theta}\right)^2 \qquad (1)$$

4.2 Jackknife

Similarly to dependent random groups technique the sample is divided in non-overlapping A sub-samples. The parameter θ is estimated from the sample s by deleting one of sub-sample for each $a = 1,\ldots,A$. The resulting estimates $\hat{\theta}_{(a)}$ are used to estimate the variance of $\hat{\theta}$ by

$$\hat{V}\left(\hat{\theta}\right) = \frac{A-1}{A}\sum_{\alpha=1}^{A}\left(\hat{\theta}_{(\alpha)} - \hat{\theta}\right)^2 \qquad (2)$$

4.3 Linearization

The idea of linearization is to estimate a variance of complex statistics using the same estimator of variance as for totals. The goal of linearization is to find z_i for each unit in the sample so that variance of $\hat{\theta}$ could be approximated by

$$\hat{V}(\hat{\theta}) \approx \hat{V}\left(\sum_{i \in s} \frac{z_i}{\pi_i}\right) \tag{3}$$

Differentiable parameters can be linearized by expansion in Taylor series. For ratio of two totals $R = \frac{Y}{X}$ z_i can be expressed in form

$$z_i = \frac{1}{X}(y_k - Rx_k) \tag{4}$$

Broader class of parameters can be linearized using extended theory by J. C. Deville (1999). For example for Gini coefficient z_i can be expressed in form

$$z_i = \frac{2x_k\left(\sum_{i \in U} 1(x_i \le x_k)\right) + 2\sum_{i \in U} x_i 1(x_i \ge x_k) - x_k - (Gini+1)\left(\sum_{i \in U} x_i + Nx_k\right)}{N\sum_{i \in U} x_i} \tag{5}$$

5 Software for Estimation of Sampling Errors

To apply the theory described in previous section software in SPSS macro language has been developed.

5.1 Possibilities of the Software

It is possible to use the software for both single stage and multi-stage sampling. In case of multi-stage sampling errors are estimated at the level of PSUs. Stratification is allowed at the first stage.

Design weights should be available for software. For estimation design weights are increased proportionally to ratio of full sample size and sub-sample size. It is possible to apply non-response correction for user defined response homogeneity groups and post-stratification by one variable.

It is possible to estimate sampling errors for totals (SUM), ratio of two totals (RATIO) and Gini coefficient (GINI). Linearization of RATIO and GINI is possible to speedup the execution of software.

It is possible to use two re-sampling methods for estimating of sampling errors – jackknife and dependent random groups technique. Methods are applied at the level of PSUs. Correction of finite population is applied at level of PSUs.

User can freely choose the number of sub-samples and how sub-samples are created. PSUs could be sub-grouped in random or user defined order. The grouping of sub-groups and sub-sampling of these groups is possible.

Sample units can be divided in sub-units by applying parameters of sample unit to corresponding sub-units. For example Gini coefficient has to be estimated at individual level by applying to each individual equalised income. The income of household is divided by equalised household size (according to modified OECD scale) and the result is applied to all household members. Household is sample unit and individuals are sub-units.

5.2 Base of the Software

The software is written in SPSS® syntax using macro commands. Currently it is based on six macro commands:

- !linrat – linearization of ratio;
- !lingini – linearization of Gini coefficient;
- !estim – estimator of indicator;
- !weight – weighting of sub-sample;
- !e_tion – estimation of indicator using estimator and weights;
- !proc – estimation of sampling error;
- !proc_u – main procedure.

User can control the software using several parameters. For example:

- File – survey data file (in SPSS format);
- Strata – variable of stratifications;
- Psu – variable of PSUs;
- Diz_sv – variable of design weights;
- Meth – method of resampling – dependent random groups or jackknife;
- E_tor – estimator;
- Lin – linearization (Yes/No);
- Div – number of sub-samples;
- And other parameters.

Example of execution of the software:

```
!proc_u
 dir "C:\Darbs\Stockholm\DRG\files\SILC"
 file "C:\Darbs\Stockholm\DRG\Data\SILC\SILC2005_data_ver02.sav"
 p_file "C:\Darbs\Stockholm\DRG\Data\SILC\dem_info.sav"
 strata=prl /
 psu=atk iecirk /
 pop_psu=4263
 hh_id=db030 /
 per_sk=per_sk
 diz_sv=diz_sv
 resp=resp
 resp_gr=atk iecirk /
 p_gr=prl /
 p_var=per_sk
 p_tot=iedz_sk
 meth=DRG JACK /
 rorder=0 /
 repeat=1
 psu_gr=sel_nr /
 order=sel_nr /
 div=4 /
 e_tor=RATIO /
 lin=0 /
 level=H /
 eqscale=per_sk /
 var=hh07n hs13n /
 fast=1.
```

The software is good tool for research. It is possible to test different methods and parameters of methods for estimation of sampling error. The software has been used for estimation of sampling errors in EU-SILC and HBS surveys. It has been tested on different SPSS versions – SPSS 11.5, SPSS 12 and SPSS 14.

6 Results

The software has been used for estimation of sampling errors in EU-SILC 2005 survey. The next table shows results of sampling errors of two indicators – Lowest monthly income to make ends meet (X) and Total housing cost (Y).

Table 1 Estimates of sampling errors in EU-SILC survey

Method	Estimator	Estimation	Estimation of variance	Coefficient of Variation (%)
Dependent Random Groups	SUM(X)	39 351 774.67	821 235 601 716	2.30
	SUM(Y)	337 079 686.10	14 605 983 870 634	1.13
	SUM(X)/SUM(Y)	0.12	0.000004037	1.72
	SUM(Y)/SUM(X)	8.57	0.021500621	1.71
	GINI(X)	39.71	0.475	1.74
	GINI(Y)	30.25	0.673	2.71
Jackknife	SUM(X)	39 351 774.67	831 832 862 430	2.32
	SUM(Y)	337 079 686.10	14 743 756 770 632	1.14
	SUM(X)/SUM(Y)	0.12	0.000003679	1.64
	SUM(Y)/SUM(X)	8.57	0.019817048	1.64
	GINI(X)	39.71	0.530	1.83
	GINI(Y)	30.25	0.667	2.70

Study about the linearization shows that it could be used to gat faster estimates. In this case the estimates of sampling error are almost the same comparing estimates with and without linearization.

Table 2 Estimates of sampling errors using linearization for Gini coefficient

Method	Estimator	Number of sub-samples	Estimate of CV without linearization	Estimate of CV with linearization	Comparison
DRG	GINI	3	2.672	2.679	100.3%
DRG	GINI	4	2.284	2.277	99.7%
DRG	GINI	6	2.333	2.237	95.9%
DRG	GINI	12	1.923	1.849	96.2%
JACK	GINI	3	3.062	3.049	99.6%
JACK	GINI	4	1.999	1.999	100.0%
JACK	GINI	6	2.566	2.564	99.9%
JACK	GINI	12	2.011	2.010	100.0%

Estimates of sampling error are dependent on methodology of creating sub-samples (number of sub-samples, order of PSUs). The estimates of CV by different sub-sampling are varying. It can be seen in next table.

Table 3 Estimates of sampling errors by different sub-sampling

Nr	Method	Estimator	Number of sub-samples	Estimate of CV
1	DRG	GINI	52	2.224
2	JACK	GINI	52	2.680
3	JACK	GINI	26	2.649
4	DRG	GINI	34	2.449
5	DRG	GINI	42	2.312
1	JACK	RATIO	52	1.310
2	JACK	RATIO	42	1.284
3	DRG	RATIO	52	1.444
4	JACK	RATIO	20	1.350
5	DRG	RATIO	34	1.389

7 Conclusion

The software – created during the research is a good tool for using different methods of estimation of sampling errors. The software can be upgraded with additional methods or estimators of indicators. Analysis of linearization method shows that linearization is useful method in estimation of sampling errors. The analysis about the results of the survey will be continued.

References

Central Statistical Bureau of Latvia (2005) *Mājsaimniecības budžets 2004. gadā*, Rīga.

J. C. Deville (1999) Variance Estimation for Complex Statistics and Estimators: Linearization and Residual Techniques. *Survey Methodology*, **Vol. 25, No. 2**, 193-203, Statistics Canada.

European Commission, Eurostat, *The SAS macro for linearizing EU-SILC complex income indicators, User Guide*, Directorate F: Social Statistics and Information Society, Unit F-3: Living conditions and social protection statistics.

J. Lapiņš, E. Vaskis, Z. Priede, S. Bāliņa (2002) Household Sample Surveys in Latvia. *Statistics in Transition Journal of the Polish Statistical Association*, **Volume 5, Number 4**.

M. Liberts (2005) *Izlases apsekojumu teorija (Survey Sampling)*, LU, Rīga.

M. Liberts (2004) *Prakses darba atskaite*, LU, Rīga.

S. L. Lohr (1999) *Sampling: Design and Analysis*, Brooks/Cole Publishing Company, Pacific Grove, Calif.

A. Sandström, J. H. Wretman, B. Waldén (1988) Variance Estimators of the Gini Coefficient – Probability Sampling, *Journal of Business & Economic Statistics*, **Vol. 6, No. 1**, American Statistical Association.

SPSS® Syntax Reference Guide (2002) SPSS Inc.

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

Wikipedia, http://en.wikipedia.org/wiki/Gini_coefficient

K. M. Wolter (1985) *Introduction to Variance Estimation*, Springer-Verlag

IMPUTED RENT IN HOUSEHOLD BUDGET SURVEY

Inga Masiulaitytė

Faculty of Mathematics and Informatics, Vilnius University, Statistics Lithuania
e-mail: inga.masiulaityte@stat.gov.lt

Abstract

The imputation of rental expenditures is an important step in the estimation of a household's standard of living. The methods of imputation of the rent are discussed in this paper. Some simulation results are presented.

1 Introduction

The imputation of rental expenditures is an important step in the estimation of a household's standard of living. Rent imputation is especially important when one is wanting to make accurate welfare comparisons between households which own their housing ('owner-occupiers') and those who rent it. For example, an income comparison of two households having the same income but with one household renting and the other being an owner-occupier would, in absence of imputation, conclude that their position is the same; in reality the owner household is better-off because it enjoys housing services for free.

Definition.

The imputed rent refers to the value that shall be imputed for all households which do not report paying full rent, either because they are owner-occupiers or because they live in accommodation rented at a lower price than the market price, or because the accommodation is provided them for free.

2 Notation

Let us consider a population U consisting of *N* elements:

$$U = \{1,..., N\}.$$

Sample s from the population U is drawn according to the sampling design. Sample size is from *n* elements. The inclusion probability is π_k. Variable of interest is *y (imputed rent)*. We assume that only 10 per cent of the values of the variable *y* are known, and corresponding population elements belong to the subsample s_1 of size n_1. Subsample s_2 of size n_2 consists of the elements with unknown values of the variable *y*. Exactly, total sample is conjunction of two subsamples s_1 and s_2, i.e. $s = s_1 \cup s_2$.

We are interested in estimation of the population mean $\mu_y = \frac{1}{N}\sum_{k=1}^{N} y_k$:

$$\hat{\mu}_y = \frac{1}{\hat{N}} \sum_{k \in s} w_k y_k = \frac{1}{\hat{N}} \left(\sum_{k \in s_1} w_k y_k + \sum_{k \in s_2} w_k y_k \right). \quad (1)$$

Second term of the equality (1) is unknown. According to the selected methods unknown values of the variable *y* will be imputed. For imputation some vector of auxiliary variables $\mathbf{x}=(x_1,...,x_J)$ with the known values $\mathbf{e_{ke}}=(x_{1k},...,x_{Jk}), k \in s$ are needed.

2 Imputation methods for rent

Imputed rent is the main variable of interest. Also more variables like location of the dwelling (strata), number of rooms, type of dwelling and the amenities are analyzed and taken into account in to the rent imputation. Most of the auxiliary variables correlate with the rent and make influence on the rent price.

Some methods for imputation of the rent are analyzed. They will be introduced briefly.

2.1 Self – assessment method

Self – assessment method is based on the owner-occupiers answers about potential rent for their dwellings. This method is subjective method, and can not show real situation in the rental market.

2.2 Homogeneity groups method

Imputed rent is estimated using homogeneity groups method. It is based on actual rentals, and combines information on the housing stock, broken down by various groups, with information on actual rentals paid in each group. For all persons within each group, the average rent of tenants belonging to the same group is used for the imputation.

2.3 Heckman method

The sample selection model consists of two equations. The first equation indicates if target variable is observed, i.e. if there is a response or a nonresponse:

$$r_k^* = \alpha \mathbf{z}_k + v_k, \quad (2)$$

here $\mathbf{z}_k$ is a value of a vector of auxiliary variables, α is a vector of parameters, it has to be estimated using the sample data, and v_k is the error term, a random component. However, r_k^* is a latent variable that we do not observe. We only observe whether it exceeds a certain threshold, say 0, because that results in the response or nonresponse:

$$r_k = \begin{cases} 1, \ if \ r_k^* \geq 0, \\ 0, \ if \ r_k^* < 0. \end{cases} \quad (3)$$

Whether or not a person responds is the result of an underlying process for which we only observe the actual outcome.

The second equation assumes a linear relationship between the target variable *y* and a vector of auxiliary variables $\mathbf{x} = (x_1,...,x_J)$.

$$y_k^* = \beta \mathbf{x}_k + u_k, \tag{4}$$

here u_k is a random component, β is a vector of parameters that have to be estimated using the sample data. y_k^* is also a latent variable. The study variable *y* is defined by:

$$y_k = \begin{cases} y_k^*, \ if \ \ r_k = 1, \\ \cdot, \ if \ \ r_k = 0. \end{cases} \tag{5}$$

Contrary to equation (3), we do not observe values for the nonrespondents $(r_k = 0)$. These observations are simply missed.

Equation (2) and (4) are linked by a joint probability distribution of the error terms u_k and v_k. In practice it is usually assumed that they have bi-variate normal distribution:

$$\begin{pmatrix} u_k \\ v_k \end{pmatrix} \sim N\left(0, \begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix} \right), \tag{6}$$

here ρ is the correlation coefficient between the two error terms, and σ^2 is the variance of u_k. If there is no correlation between the two equations $(\rho = 0)$, then there is no selection bias and the target variables can be estimated using equation (4) only. If the error terms are correlated, however, the nonresponse is selective and restricting estimation (4) would result in a selection bias.

The first step of the Heckman method is implemented using probit model in PROC LOGISTIC of SAS. The second one is done by PROC REG.

2.4 Log-linear regression

Lets us introduce a variable *l* with the values:

$$l_k = \ln y_k, \quad k = 1,..., N. \tag{7}$$

Then the log-linear model can be defined as:

$$l_k = \beta \mathbf{x}_k + \varepsilon_k, \ k \in s_1, \tag{8}$$

here β – is vector of coefficients, $\mathbf{x}_k$ – are values of known auxiliary variables, ε_k – are random errors.

3 Simulation

3.1 Sample design

The HBS sample is chosen to be a population. Stratified random design is used. Population of individuals of Lithuania is divided into seven strata by place of living place: the 5 biggest towns, other cities and rural areas. Independent simple random sampling is used in each of the stratum. The total of the population is 19130 individuals; there are 1997 persons in Vilnius, 1854 – in Kaunas, 764 – in Klaipėda, 625 – in Šiauliai, 628 – in Panėvėžys, 6262 – in other cities and 7000 persons in the rural. The assumption that rent in population is known is made. 2000 individuals are selected from the population.

3.2 Simulation results

1000 stratified simple random samples have been selected. The sample consists of 200 elements, that are 10 per cent of tenants. Rent for owner-occupiers have been imputed using different methods. After, mean of imputed rent for each sample has been estimated and compare with "true" rent.

The comparison of average of estimates of mean rent estimated by different methods and "true" rent is presented in Fig.1.

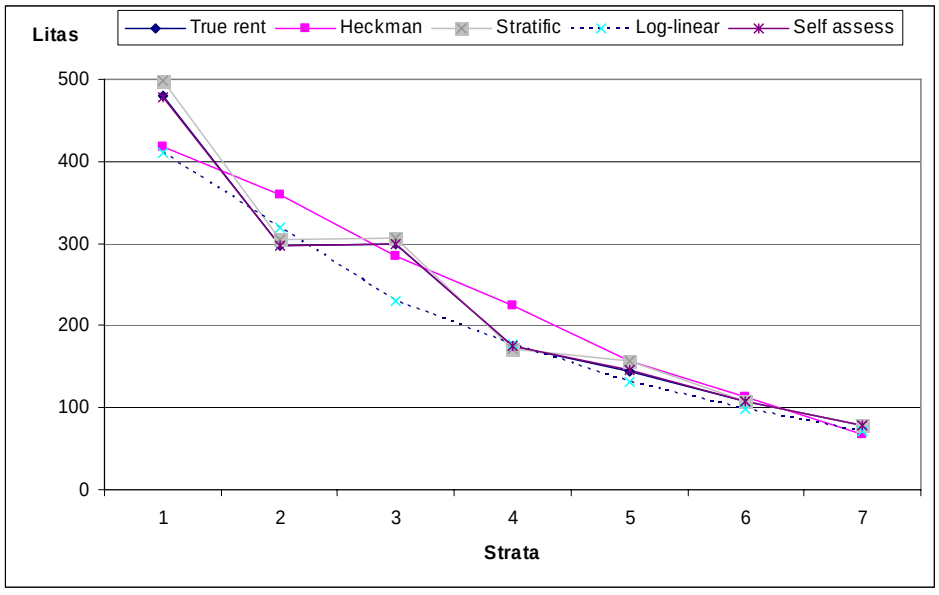


Figure 1. Empirical average of estimates of mean rent using different imputation methods by strata

4 Conclusion

The imputed rent estimated by Homogeneity groups method has a smaller bias than estimated by Heckman or Log-linear methods.

References

Jeroen Smits, (2003) *Estimating the Heckman two-step procedure to control for selection bias with SPSS*. http://home.planet.nl/~smits.jeroen/selbias/Heckman-SPSS.doc .

WEN Hai-zhen, JIA Sheng-hua. GUO Xiao-yu, (2004) *Hedonic price analysis of urban housing: An empirical research on Hangzhou*, China, Journal of Zhejiang University SCIENCE, 6A(8), p. 907-914.

EU–SILC DOC TFMC-12/06. *Third Meeting of the EU-SOLC task force on methodological issues. Imputed rent.*. Eurostat-Luxembourg, April 2006, EU-SILC DOC TFMC-12/06.

DOC HBS/161/2006/EN, DOC EU-SILC/162/06/EN. *Meeting of the working group on living conditions (HBS, EU-SILC and IPSE). HBS and EU-SILC Imputed rent.*. Eurostat-Luxembourg, May 2006.

Bethlehem J., Cobben F., Schouten B., May 2006, *Nonreponse in Household Surveys*. Version 1, Statistics Netherlands, Methods and Informatics Department, p. 141-142.

Bradley C. Martin, Rahuk Ganguly, *SAS program for Heckman 2-stage estimation*. University of Georgia, College of Pharmacy, Athens, GA 30602, http://www.rx.uga.edu/main/home/cas/faculty/heckman.pdf.

INTRODUCTION TO EMERGING METHODS FOR IMPUTATION IN OFFICIAL STATISTICS

Pasi Piela

Statistics Finland

e-mail: pasi.piela@stat.fi

Abstract

The quality of the data is one of the main issues in today's official statistics. Missingness of the data is a serious threat to the quality. Imputation, as one possibility to solve this problem, means statistical replacement of missing values. The family of imputation methods is wide and versatile. Here, re-clustering of the data by using auxiliary variables is an important approach for creating homogenous imputation classes in which imputations to the target variable with missing values could be made. Finally we also discuss about how complex data structures and clustering that *already exists* such as in a cluster sampling design could be taken into account. This paper presents my previous research on imputation techniques (Piela, 2005) and plans for future research.

Keywords: imputation, editing, statistical clustering, hierarchical clustering, cluster sampling, multilevel models.

1 Introduction to imputation in the quality framework of official statistics

The quality of the data is one of the main issues in the official statistics. The quality of the inferences that can be drawn from statistical data depends on the quality of the data. See the definitions for statistical quality in the quality quidelines of Statistics Canada 2003 and Eurostat 2002. Here, imputation is undertaken as the process of statistical replacement of missing values and *as part of a quality improvement strategy to improve accuracy, consistency and completenes*s (Charlton 2003).

The goal of imputation is to reduce non-response bias of a statistic or an estimator. Bias is reduced by using auxiliary information in imputation, that is, using the non-missing values of

other variables and other observations. Knowledge of the missingness mechanism (or *the missingness pattern*) could help a lot. Rubin (1987) used the division in the famous concept of missingness as follows: *missing at random* (the probability of an item being missing does not depend on the values of the missing items), *missing completely at random* (the probability of an item being missing does not depend on either the value of the missing item or the other characteristics of respondents) or *not missing at random* otherwise, abbreviatiated respectively as MAR, MCAR and NMAR. The alternative two-category divisions are *ingnorable* (MAR) or *non-ignorable missingness* (NMAR) and *informative* or *non-informative missingness*.

Methods of imputation vary considerably depending on the type of data set, the characteristics of the missingness and the extent of missingness in the data. There are at least three ways of classifying imputation methods, as presented in my licentiate thesis in statistics (Piela, 2005):

A1. Deterministic imputation, or

A2. Stochastic imputation

B1. Logical imputation,

B2. Real donor imputation, or

B3. Model donor imputation

C1. Single imputation, or

C2. Multiple imputation

A1. In deterministic methods the imputation procedure always gives exactly the same value when repeated. Thus, logical imputation, mean imputation, regression imputation and nearest neighbour belong to this class of methods.

A2. Stochastic imputation involves a random element. This means that when imputation is repeated imputed values are not exactly same. Random donor method, regression imputation with random element and nearest neighbour within imputation classes (clusters) created stochastically belong to this class of methods.

B1. Logical imputation is used when there is only one possible solution that is known given appropriate assumptions and restrictions. For example, one missing component when total and other components are known can be imputed by calculating the difference of the total and the sum of the known components. This method can be considered as a part of the editing process.

B2. Real donor method (*hot-deck* or donor imputation) gives a value that is "borrowed" from a real observed case, for example, a nearest neighbour. Hot-deck, however, could be used in a wider context referring to the donor imputation without historical information but in which the final chosen donor can also be a group of donors each giving *a fraction* as in chapter 4.

B3. Model donor method refers to methods where also non-observed values and even values out of the range of the variable are possible. Usually these methods use regression based imputations.

C1. Single imputation methods give a single value that replaces the missing value.

C2. Multiple imputation produces several imputed datasets which are then used to estimate imputation variance.

General five accuracy requirements for imputation procedure in official statistics (see Chambers, 2001) are *predictive accuracy* (the imputation procedure should maximise the preservation of true values), *distributional accuracy, estimation accuracy, imputation plausibility* (imputation procedure and imputed values must be plausible) and *ranking accuracy*. For official statistics purposes the methods should be automatized, fast and still suitable for the large data masses. The statistical production process has indeed many features similar to the industrial process. On the other hand, simple methods are easy to understand, to implement and to apply. It is often the case in official statistics that users of the imputation programs are not the ones who have planned and created these programs and there is not necessarily enough time to make a methodological analysis in the production process.

1.1 Processing imputation and editing

The current tendency in official statistics is to understand imputation and editing as their own *process* along the whole statistical survey process. The Banff system of Statistics Canada is a good example about computerized editing and imputation process. It is a collection of specialized SAS® procedures "each of which can be used independently or put together in order to satisfy the edit and imputation requirements of a survey" as stated in the Banff manual (Statistics Canada, 2006). Banff has been derived from possibly more well-known program named Generalized Edit and Imputation System (GEIS).

Thus, Banff includes appropriate imputation and editing algorithms but the special part is to see editing and imputation as its own process starting from the edit specification and preliminary data analysis and continuing to error localisation and outlier detection and ending to imputation model selections and prorating (and so called *mass imputation* when needed). Statistics Finland is starting to test and evaluate Banff in 2006 for its own use. Please contact the author of this paper for further details.

2 Past research work in imputation

The methodological goal of my past imputation research (Piela 2005) has relied on approaches that can be regarded as advanced imputation methods compared to the standard, traditional ones such as mean, nearest neighbour, random donor and regression imputation. The advanced imputation concept pertains to complex imputation problems and automated solutions. Naturally this is a very general definition, only giving an idea of more developed methods than the standard ones. One important approach that requires special attention and is emphasized here is *clustering* from the imputation point of view. Clustering as a statistical technique refers to re-grouping of the data by auxiliary variables so that clusters are internally homogenous and distinctive among each other.

Clustering makes it possible to use random techniques that give more variation to the imputed values. Subgrouping the data into homogenous groups creates a good base for any imputation method. The problem comes from too small groups and, of course, failures of grouping. By using standard measures of homogeneity like the Gini index, which does not refer to the true values of the missing ones, it is possible to check whether or not the grouping is satisfactory enough for forthcoming imputation.

Clustering techniques can be computationally heavy and they naturally have to be computerised. Typically statistical clustering is being done by iterative way in reducing the variance within clusters. In imputation this means that a unit with a missing item should belong to a cluster having similar kinds of observations with only very small variation. Missing items can then be replaced by mean, random donor or nearest neighbour or by a *local* regression model, e.g. by a model fitted within a corresponding cluster or within a group of similar clusters in case of small clusters.

2.1 *K*-means and hierarchical clustering

The basic and possibly the most well-known variance minimization technique is *K*-means. It starts by defining initial, often random, weights w_i. Then for every unit, e.g. the vectors with selected auxiliary variables, the "closest" w_i is selected. Closest is often defined by Euclidean distance metrics (the Euclidean norm). This creates an initial clustering which then continues by calculating the average points of the clusters. These points will now be updated positions of w_i. The updating process continues until there are no noticeable changes in the values of w_i. The resulting structure is called Voronoi tesselation or Thiessen diagram.

Another popular clustering method is hierarchical clustering. This far it has not been popular in imputation. Breiman et al. published a book about *classification and regression trees*,

abbreviated as CART, in 1984. Since then it has been regarded as a primary book on tree clustering methods in statistics. Tree-based methods are conceptually simple yet powerful.

A tree is usually created by splitting the data in a binary way. The data are divided into two sets and then each of the sets is divided again into two subsets in turn as long as the stopping criteria are reached. Binary splitting is partly due to simplicity, that is, a binary tree is easier to interpret than multiway trees.

In imputation it is relevant to keep the tree in appropriate size and especially keep the minimum size of the *terminal nodes* (the final nodes, clusters or leaves of the tree) decent for imputation, e.g. ad hoc value equal to 50. That is to say, the tree does not have to be explanatory in all the ways, so it is possible to create a large tree that is good for imputation. However, the minimum size of the nodes is naturally much more important. Piela & Laaksonen (2001) handled these issues by using the samples of the UK Census (SARs) 1991 and Finnish Household Expenditure Survey 1996 datasets.

2.2. Neural networks

Neural networks are nowadays regarded as interesting alternatives to the ordinary statistical methods. Especially very large data masses might need automated intelligent methods in finding homogenous groups and in visualisation. Neural networks or modern statistical pattern recognition can be seen very important from an imputation perspective.

Self-organizing map (Kohonen, 1997), SOM, is an iterative, "non-supervised", clustering algorithm related to the neural network methodology. Interestingly, neural networks can be seen just as generalizations of the well-known statistical methods. Specifically, the SOM algorithm can be interpreted as a discretized approximation procedure for the computation of principal curves or surfaces (Ritter et al., 1992). Koikkalainen, Horppu and Piela (2003) concentrated – among the other neural network methods – on its special modification, tree-structured self-organizing map or TS-SOM as abbreviated. It is a mixture of tree clustering, computational speed-up and SOM techniques. See Piela (2002) for successful results of the TS-SOM techinques from the view of official statistics.

The special problem, however, in neural network modelling is that even if we find a good neural method for imputation it can still be very hard to be implemented in practice. These models or systems especially require the user's good knowledge, e.g. on handling parameters that often differ from the transparent, standard, statistical model parameters. On the other hand, official statistics require fast imputation processes.

However, Piela (2004) concluded that editing and imputation as part of the data mining process can be the future of the neural networks in official statistics. Although not necessarily

in imputation but in editing the powerful potentiality of these methods is simply in their visualization possibilities. It is slightly harder to find any other areas where neural networks or statistical pattern recognition methods in general could be useful due to the confirmatory nature of the statistical analysis in national statistical offices.

2.4. Conclusive comments

What is then the imputation method that best meets the requirements in official statistics? According to my study, it is the nearest neighbour imputation in general – the worst one being the mean imputation. Of course, some clustering is useful before replacing missing values by nearest donors, but clustering methods are also based on the nearest neighbour idea. This method keeps imputed values within an appropriate range but still maintains enough variation. Often the standard Euclidean distance metrics seems to work fine. The nearest neighbour method is also a very natural way to look for appropriate values if the data are not too small.

Anyhow, *the best* method is actually a system that includes several competitive imputation methods (recall 1.1). Specifically, the development and evaluation of new forthcoming imputation methods, especially complex neural network models, is closely connected to the software development. SOM, for instance, can be used in clustering. After that the software should be able to produce several competitive model alternatives within those clusters. Depending on the type of missingness the imputation model is then chosen by using appropriate model statistics and the user's own experience in order to conclude *the imputation task*. However, finding the most adequate methods still requires further research.

Multiple imputation, MI, was not handled here because of the context of the research. But also detailed research in imputation variance and careful analysis of the datasets with hierarchical, multilevel nature (note the difference to the previously mentioned hierarchical clustering methods) containing cross-cassifications and missingness were also excluded. This will lead us to the forthcoming research that will be next outlined.

3 Multiple imputation, MI

Multiple imputation (Rubin, 1987), which works in the Bayesian framework, responds to the general problems of strongly dependent variables including missing cases in any of them, the problem of imputation variance estimation and the problem of inconsistency in the theoretical framework of imputation. In multiple imputation there are two conditions that must be met. First one is that the missing data should be missing at random, MAR, meaning that the probability of missing data on a target variable Y does not depend on missing part of Y. However, missing data is allowed to be dependent on observed part of Y (e.g. *planned*

missingness in surveys). Besides, there exist some MI applications for MCAR situations as pointed out by Schafer & Graham (2002). The second requirement is that one has to include all the covariates and interactions of the observed data as otherwise imputed observations will not have this structure. This is actually just what is expected in multilevel modeling.

Now, let the quantity that is target of interest be $\theta = \theta(Y_O, Y_M)$, where Y_M is the missing part of the data and Y_R is the observed part. We try to estimate the distribution of

$$f(\theta \mid Y_R) = \int \theta(Y_R, Y_M) f(Y_M \mid Y_R) dY_M .$$

MI assumes this distribution is approximately Normal described by its mean and variance. We note, that under MCAR or MAR regression models give valid parameter estimates. By using Bayesian framework or maximum likelihood statistics we are able to get a valid estimate of the distribution of $Y_R|Y_M$. After that we can impute the missing data *number of times* by drawing them from this estimated distribution. In practice the number of times is often between 3 and 10. The mean of the distribution of θ is approximated by the average of the estimates of the imputed datasets. Formulation of the variance combines both between and within imputation components of variance (see Rubin, 1987).

However, MI is not commonly used. Assumptions are strict and there also exist some difficulties with MI variance estimation as discussed by Rao (2005) and Kim et al. (2004). However, convenient estimation of variance under MI remains as an attractive feature.

4 Fractional imputation

Fractional imputation (Kalton & Kish, 1984) is a sort of mixed donor and model donor imputation method (one could still call this as a *hot deck* method), which involves using more than one donor for a recipient. For example, three imputed values might be assigned to each missing value, with each entry allocated a weight of one-third of the nonrespondent's original weight as the sum of the imputation fractions for each missing item is required to be one.

Kim and Fuller (2004) showed that fractional imputation and the suggested variance estimator are superior to multiple imputation estimators in general and for estimating the variance of a domain mean. Specifically, fractional imputation was designed to reduce the imputation variance while multiple imputation only gives a simplified way to estimate it. In the following we formulate "basic setup" to the fractional imputation methodology.

Consider a population of N elements identified by a set of indices $U = \{1, 2, \ldots, N\}$. Associated with each unit i there is a study variable y_i and a vector of auxiliary information; the use of auxiliary information is skipped here. Let A denote the indices of elements in a

sample selected by a set of probability rules called a sampling mechanism, $\hat{\theta}$ being a full sample and linear-in-*y* estimator of the population quantity of interest θ_N. Now, write

$$\hat{\theta} = \sum_{i \in A} w_i y_i .$$

Now, of course, $\hat{\theta}$ is unbiased for the population total if w_i is the inverse of the selection probability, corresponding to the Horwitz-Thomson estimator.

The essential assumption is that the *U* is divided into imputation cells *G* (e.g. clusters) in which the homogeneity is defined by using the response probability approach. We assume the within-cell uniform response model in which the responses in a cell are equivalent to a Bernoulli sample from the elements in a cell.

Let d_{ij} be the number of times that y_i is used as a donor for missing y_j and define $\mathbf{d} = \{d_{ij}; i \in A_R, j \in A_M\}$, where A_R and A_M denote the set of indices of the sample respondents and sample nonrespondents. The distribution of **d** is called the *imputation mechanism*, where as the distribution of a standard binary response indicator function **R** is called as the *response mechanism* (see Fuller and Kim, 2005). Now, let w_{ij}^* be the factor applied to the original weight for element *j* when y_i is used as a donor for element *j*. For element *j*, *j* belonging to A_M,

$$Y_{Ij} = \sum_{i \in A_R} w_{ij}^* y_i$$

is the weighted mean of the respondent values. The w_{ij}^* is called the *imputation fraction* (Fuller and Kim, 2005). Thus, it is the fraction that donor *i* is donating for the missing item y_j. Obviously, $w_{ij}^* = 0$ for $i \neq j$, $i, j \in A_R$ and $w_{ii}^* = 1$ for $i \in A_R$ and the sum of the fraction is restricted to equal 1.

The estimator with imputed values Y_{Ij} and some $w_{ij}^* < 1$ is called a *fractionally imputed* estimator. An imputation estimator that is linear in *y* can be written in the form

$$\hat{\theta}_I = \sum_{i \in A_R} \left(\sum_{j \in A} w_j w_{ij}^* \right) y_i .$$

Fuller and Kim (2005) presented also the fully efficient fractional imputation, FEFI. It requires uniform response probablities in an imputation cell and the use of every responding unit as a donor for every nonresepondent in the cell. However, FEFI has hardly any use in practice. Fuller and Kim (2005) responded this problem by giving approximations to FEFI. They outlined the procedure with fixed number of donors per recipient that is fully efficient for the grand total but not necessarily for subpopulations.

Kim and Fuller (2004) showed also how fractional imputation combined with the proposed *replication variance estimator* gives a set of replication weights that can be used to construct unbiased variance estimators for estimators based on imputed data (and for estimators based on the completely responding variables).

5 Multilevel modeling for imputation

Many kinds of data have a hierarchical or clustered structure. We refer to a hierarchy as consisting of units grouped at different levels. Thus children may be the level 1 units in a 2-level structure where the level 2 units are the families and students may be the level 1 units clustered within schools that are the level 2 units. Naturally children belonging to a same family have generally common characteristics both mentally and physically. But there is also a very well-known example how important it is to take into account schools as clusters when studying grades and success of the elementary school children. The study by Aitkin et al. (1981) is widely used in literature as an introduction to multilevel modelling (see Goldstein, 2003).

More recently there has been a growing awareness that many data structures are not purely hierarchical but contain cross-classifications of higher level units and multiple membership patterns (see the web pages of the Centre for Multilevel Modelling: http://www.mlwin.com/). An example of the former is where students in a longitudinal study "belong" to a combination of elementary school attended and secondary school. For a detailed discussion of such structures see the paper by Hill and Goldstein (1998).

Statistical multilevel models (Goldstein, 2003) take advantage of the correlation structure between different levels of hierarchy. Correlation structures and connections between the study variables can be a challenge in imputation tasks as well. Indeed, multilevel models for imputation can provide an interesting supplement to the imputation methods discussed in previous sections.

Currently in literature, there are some papers about the use of multiple imputation with multilevel models. Afterall, multiple imputation is popular among some social science research areas. Carpenter and Goldstein (2004) says that if a dataset is multilevel, then the imputation model should be multilevel too. MLwiN (Rasbash et al., 2004) is a commonly used software for multilevel modeling. It can fit a range of Bayesian models using Markov Chain Monte Carlo. Carpenter and Goldstein (2004) consider MLwiN as a natural tool for multiple imputation in multilevel modeling.

6 Plans for future research

In the forthcoming research I will study new imputation methods and the analysis of imputed and missing data when data structures are complex. This includes the use of multilevel imputation models in which clusters *in the cluster sampling design* can be incorporated in an imputation model as random effects. In addition, information on the complex sampling design can be incorporated, for example, by using strata indicators as fixed covariates. Multilevel imputation models do not necessarily refer to MI but also modified single and fractional imputation methods can be considered.

Another new avenue of research is the use of multilevel models and imputation in the context of small area estimation (Rao, 2003). Here I will concentrate on the analysis of missing data when estimating small area totals using imputation and reweighting methods. Model-assisted survey methods that are based on multilevel models (Lehtonen, et al., 2003 and 2005) will be addressed.

References

Aitkin, M., Anderson, D., and Hinde, J. (1981): Statistical Modelling of Data on Teaching Styles (with discussion). *Journal of the Royal Statistical Society*, **A 144**, 148-1461.

Breiman, L., Friedman, J.H., Olsen, R.A., and Stone, C.J. (1984): *Classification and Regression Trees.* Wadsworth, Belmont, CA.

Carpenter, J.R., and Goldstein, H. (2004): Multiple Imputation Using MLwiN, *Multilevel Modelling Newsletter*, **16**, 2. **

Charlton, J. (2003): Editing and Imputation Issues. *Towards Effective Statistical Editing and Imputation Strategies - Findings of the Euredit project.* *

Chambers, R. (2001): Evaluation Criteria for Statistical Editing and Imputation. *National Statistics Methodological Series*, United Kingdom, **28**, 1-41. *

Eurostat (2002): *Quality in the European Statistical System - the Way Forward.* European Comission.

Goldstein, H. (2003). *Multilevel Statistical Models (Third Edition).* Edward Arnold, London.

Hill, P.W., and Goldstein, H. (1998): Multilevel Modelling of Educational Data with Cross Classification and Missing Indentification of Units, *Journal of Educational and Behavioural Statistics*, **23**, 117-128.

Kalton, G., and Kish, L. (1984): Some Efficient Random Imputation Methods. *Communications in Statistics*, **A13**, 1919-1939.

Kim, J.K., and Fuller, W.A. (2004): Fractional hot deck imputation, *Biometrika*, **91**, 559-578.

Fuller, W.A., and Kim, J.K. (2005): Hot Deck Imputation for the Response Model, *Survey Methodology*, **31**, 2, 139-149.

Kim, J.K., Brick, J.M., Fuller, W.A. and Kalton, G. (2004): *On the Bias of the Multiple Imputation Variance Estimator in Survey Sampling*. Technical Report.

Kohonen, T. (1997): *Self-Organizing Maps.* Springer Verlag, New York.

Koikkalainen, P., Horppu, I., and Piela P. (2003): Evaluation of SOM based Editing and Imputation. *Towards Effective Statistical Editing and Imputation Strategies - Findings of the Euredit project.* *

Lehtonen, R., Särndal, C.-E., and Veijanen, A. (2003): The Effect of Model Choice in Estimation for Domains, Including Small Domains, *Survey Methodology*, **29**, 33-44.

Lehtonen, R., Särndal, C.-E., and 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.

Piela, P., and Laaksonen, S. (2001): Automatic Interaction Detection for Imputation – Tests with the WAID Software Package. In: *Proceedings of Federal Committee on Statistical Methodology Research Conference, Statistical Policy Working paper*, **34**, 2, 49-59, Washington, DC.

Piela, P. (2002): Introduction to Self-Organizing Maps Modelling for Imputation - Techniques and Technology. *Research in Official Statistics*, **2**, 5-19.

Piela, P. (2004): Neuroverkot ja virallinen tilastotoimi, *Suomen Tilastoseuran Vuosikirja 2004*. Helsinki.

Piela, P. (2005): *On Emerging Methods for Imputation in Official Statistics.* Licentiate Thesis. University of Jyväskylä.

Rasbash, J., Steele, F., Browne, W., and Prosser, B. (2004): *A User's Guide to MLwiN (version 2.0)*. London. **

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

Rao, J.N.K. (2005): Interplay Between Sample Survey Theory and Practice: An Appraisal, *Survey Methodology*, Statistics Canada, **31**, 2, 117-138.

Ritter, H., Martinez, T., and Schulten, K. (1992): *Neural Computation and Self-Organizing Maps: An Introduction*. Addison-Wesley, Reading, MA.

Schafer, J.L., and Graham, J.W. (2002): Missing Data: Our View of the State of the Art, *Psychological Methods* **7**, 2, 147-177.

Statistics Canada (2003): *Statistics Canada Quality Guidelines*. Catalogue no. 12-539-XIE.

Statistics Canada (2006): *Functional Description of the Banff System for Edit and Imputation.* Statistics Canada.

Rubin, D. (1987): *Multiple Imputation in Surveys*. John Wiley & Sons, New York.

*) The Euredit project: http://www.cs.york.ac.uk/euredit/

**) Centre for Multilevel Modelling: http://www.mlwin.com/

CALIBRATED ESTIMATORS OF FINITE POPULATION COVARIANCE

Aleksandras Plikusas[1] and Dalius Pumputis[2]

[1] Vilnius University, Institute of Mathematics and Informatics, Lithuania
e-mail: plikusas@ktl.mii.lt

[2] Institute of Mathematics and Informatics, Vilnius Pedagogical University, Lithuania
e-mail: dpumputis@yahoo.co.uk

Abstract

Calibrated estimators of the population covariance are constructed using auxiliary information in order to get more accurate estimates. Different distance measures are used to construct calibrated estimators. Four estimators of the covariance are presented in the paper. The experimental comparison of the considered estimators will be presented when correlation between study variable and known auxiliary variable be *0.8, 0.6, 0.4, 0.2*.

1 Introduction

Calibrated estimators are widely used in finite population statistics to improve the quality of estimators, using auxiliary information. The idea of calibration technique for estimating of population totals was presented in (J.-C. Deville, C.-E. Särndal *1992, 376–382* p.). Recently the calibration technique has been widely used in the presence of non-response (S. Lundström *1997*). The calibrated estimators of a ratio of two totals were introduced in (A. Plikusas *2003, 543–547* p.). The calibrated estimator of a ratio as well as of population variance and covariance can be defined in different ways. We can choose a different calibration equation and use different distance functions. Five distance functions have been presented in (J.-C. Deville, C.-E. Särndal *1992, 376–382* p.), but only one of them (L_1) is being used in practice. This distance function is the simplest one and there exists an explicit solution of calibration equations when calibrating the estimator of the total as well as of the ratio (A. Plikusas *2001, 457–462* p.). An undesirable property of this distance function is that for some populations calibrated weights can be negative.

2 Calibration problem

Consider a finite population $U = \{u_1, u_2, ..., u_N\}$ of N elements. Let *y* and *z* be two study variables defined on the population *U* and taking values $\{y_1, y_2, ..., y_N\}$ and $\{z_1, z_2, ..., z_N\}$ respectively. The values of the variables *y* and *z* are not known. We are interested in the estimation of the covariance S_{yz}. Let us consider the estimator of the covariance

$$\hat{S}_{yz} = \frac{1}{N-1}\sum_{k\in\mathbf{s}} d_k \left(y_k - \frac{1}{N}\sum_{i\in\mathbf{s}} d_i y_i \right)\left(z_k - \frac{1}{N}\sum_{i\in\mathbf{s}} d_i z_i \right).$$

Here ***s*** denotes a probability sample set, $d_k = 1/\pi_k$ are sample design weights, π_k is a probability of inclusion of the element *k* into the sample ***s***.

Suppose, that some auxiliary information is available. Let a variable x_y with the population values $\{x_{y1}, x_{y2}, ..., x_{yN}\}$ and a variable x_z with the values $\{x_{z1}, x_{z2}, ..., x_{zN}\}$ be auxiliary variables with the known covariance $S_{x_y x_z}$. The covariance estimator is constructed using the known auxiliary variables. It is known that in case the auxiliary variables are well correlated with study variables, the variance of the calibrated estimator of the covariance is lower. Using auxiliary variables the calibrated estimator $\hat{S}_{wyz} = \frac{1}{N-1}\sum_{k\in\mathbf{s}} w_k \left(y_k - \frac{1}{N}\sum_{i\in\mathbf{s}} w_i y_i \right)\left(z_k - \frac{1}{N}\sum_{i\in\mathbf{s}} w_i z_i \right)$ of the covariance S_{yz} is defined under the following conditions:

a) the weights w_k estimate the known covariance $S_{x_y x_z}$ without error:

$$\hat{S}_{wx_y x_z} = S_{x_y x_z} \tag{1}$$

b) the distance between the design weights d_k and calibrated weights w_k is minimal according to some loss function *L*.

3 Examples of distance measures

Let us introduce free additional weights $q_k, k = 1, ..., N$. One can modify calibrated estimators by choosing q_k. A number of a different estimators can be derived as a special case of the calibrated estimator by choosing weights q_k. We can also put $q_k = 1$ for all *k*. The following loss functions can be considered:

$$L_1 = \sum_{k\in\mathbf{s}} \frac{(w_k - d_k)^2}{d_k q_k},\ L_2 = \sum_{k\in\mathbf{s}} \frac{w_k}{q_k} \log\frac{w_k}{d_k} - \frac{1}{q_k}(w_k - d_k),\ L_3 = \sum_{k\in\mathbf{s}} 2\frac{(\sqrt{w_k} - \sqrt{d_k})^2}{q_k},$$

$$L_4 = \sum_{k\in\mathbf{s}} -\frac{d_k}{q_k} \log\frac{w_k}{d_k} + \frac{1}{q_k}(w_k - d_k),\ L_5 = \sum_{k\in\mathbf{s}} \frac{(w_k - d_k)^2}{w_k q_k},\ L_6 = \sum_{k\in\mathbf{s}} \frac{1}{q_k}\left(\frac{w_k}{d_k} - 1\right)^2,$$

$$L_7 = \sum_{k\in\mathbf{s}} \frac{1}{q_k}\left(\frac{\sqrt{w_k}}{\sqrt{d_k}} - 1\right)^2.$$

The functions $L_1 - L_5$ are mentioned in (see J.-C. Deville, C.-E. Särndal *1992, 376–382* p.). The distance measures L_6 and L_7 are introduced in (A. Plikusas *2003, 543–547* p.).

4 Results

Let us introduce some notations

$$\hat{\mu}_{w1} = \frac{1}{N}\sum_{k\in\mathbf{s}} w_k x_{yk},\quad \hat{\mu}_{w2} = \frac{1}{N}\sum_{k\in\mathbf{s}} w_k x_{zk},\quad \hat{N}_w = \sum_{k\in\mathbf{s}} w_k,$$

$$b_i = x_{yi} x_{zi} + \left(\frac{\hat{N}_w}{N} - 2\right)\left(x_{yi}\hat{\mu}_{w2} + x_{zi}\hat{\mu}_{w1}\right) + \hat{\mu}_{w1}\hat{\mu}_{w2},$$

$$B = (N-1)S_{x_y x_z} + (2N - \hat{N}_w)\hat{\mu}_{w1}\hat{\mu}_{w2} - \sum_{k\in\mathbf{s}} d_k x_{yk} x_{zk} \ .$$

Proposition 1. The weights w_i, which satisfy the calibration equation (1) and minimize the loss function $L_1 = \sum_{k\in\mathbf{s}} \frac{(w_k - d_k)^2}{d_k q_k}$ can be expressed as $w_i = d_i \nu_i$, with

$$\nu_i = 1 + B\left(\sum_{k\in\mathbf{s}} d_k b_k q_k x_{yk} x_{zk}\right)^{-1} b_i q_i \, .$$

Proposition 2. The weights w_i, which satisfy the calibration equation (1) and minimize the loss function $L_3 = \sum_{k\in\mathbf{s}} 2\frac{(\sqrt{w_k} - \sqrt{d_k})^2}{q_k}$ can be expressed as $w_i = d_i \nu_i$, with $\nu_i = 1 \Big/ \left(\frac{1}{2}\lambda b_i q_i - 1\right)^2$, here λ is a properly chosen root of the equation

$$\left(\frac{1}{4}\sum_{k\in\mathbf{s}} w_k q_k^2 b_k^2 x_{yk} x_{zk}\right)\lambda^2 - \left(\sum_{k\in\mathbf{s}} w_k q_k b_k x_{yk} x_{zk}\right)\lambda + B = 0 \, .$$

Proposition 3. The weights w_i, which satisfy the calibration equation (1) and minimize the loss function $L_6 = \sum_{k\in\mathbf{s}} \frac{1}{q_k}\left(\frac{w_k}{d_k} - 1\right)^2$ can be expressed as $w_i = d_i \nu_i$, with

$$\nu_i = 1 + B\left(\sum_{k\in\mathbf{s}} d_k^2 b_k q_k x_{yk} x_{zk}\right)^{-1} d_i b_i q_i$$

Proposition 4. The weights w_i, which satisfy the calibration equation (1) and minimize the loss function $L_7 = \sum_{k\in\mathbf{s}} \frac{1}{q_k}\left(\frac{\sqrt{w_k}}{\sqrt{d_k}} - 1\right)^2$ can be expressed as $w_i = d_i \nu_i$, with $\nu_i = 1 \Big/ \left(\lambda d_i b_i q_i - 1\right)^2$, here λ is a properly chosen root of the equation

$$\left(\sum_{k\in\mathbf{s}} w_k d_k^2 q_k^2 b_k^2 x_{yk} x_{zk}\right)\lambda^2 - \left(2\sum_{k\in\mathbf{s}} w_k d_k q_k b_k x_{yk} x_{zk}\right)\lambda + B = 0 \, .$$

References

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

S. Lundström, Calibration as Standard Method for Treatment of Nonresponse, *Doctoral dissertation*, Stockholm University (1997).

A. Plikusas, Calibrated weights for the estimators of the ratio, *Lith. Math. J.*, **43**, 543-547 (2003).

C.-E. Särndal, B. Swensson, J.Wretman, *Model Assisted Survey Sampling*, Springer-Verlag, NewYork.

ESTIMATION OF THE NUMBER OF NON-OFFICIAL EMIGRANTS FROM THE LABOUR FORCE SURVEY

Genovaitė Šalučkienė

Statistics Lithuania
E-mail: Genovaite.Saluckiene@stat.gov.lt

Abstract

In the first quarter of 2006, Statistics Lithuania carried out a survey, which allowed estimating the number of residents of Lithuania, who unofficially left the country in 2001–2005. This survey was carried out together with the Labour Force Survey (LFS).

1 Introduction

Labour market is open or partly open for Lithuanians almost in all Member States of the European Union. People can legally work in many countries, and many Lithuanians went abroad in search for a better life.

Statistics on international migration calculates all persons who leave Lithuania or arrive to Lithuania with the intention of permanent residence or for a period longer than 6 months (foreigners, who have arrived with temporary residence permits for 1 year and longer). This is official migration statistics. But not all Lithuanian residents leaving the country for more than 6 months declare their departure according to the Lithuanian legislation.

Statistics on persons who do not declare their emigration are very important and are needed to specify the emigration flow and demographic data. So, it is very important to have reliable statistics on emigration.

Statistics Lithuania decided to estimate the number of non-official emigrants using the Labour Force Survey (LFS). Some new variables were added to the LFS questionnaire for this reason.

2 Main definitions

Emigrant – person, leaving the Republic of Lithuania with the intention to take up the usual residence in another country permanently or for more than 6 months period (including

foreigners, whose temporary residence permit for 1 year or longer have expired, and a new permit has not been issued)

Non-official emigrant is an emigrant who has not declared his / her departure.

Employed population refers to the residents of a surveyed age, who have been working during the reference week for no less than 1 hour and for which they were compensated in cash or in kind (food products or other stuff) or had profit (income). They are the persons having the professional status of employers, owners, farmers, employees, contributing family workers, self-employed.

Employed persons refer also to those who were ill during the surveyed week, had vacations, did not work due to short-term or long-term idle time, took care of children less than 3 years of age and maintained official ties with the working place.

Unemployed by the ILO definitions are persons aged 15–74, who had no job during the reference week, were ready to start working if work was available during the coming fortnight and actively seeking for a job for a four-week period, i.e. applied to the public or private employment agencies, employers, friends, relatives, mass media, passed tests or had recruitment interviews, looked for premises, equipment for his / her own business, tried to obtain a business certificate, get a licence or financial resources

Inactive population are persons who cannot be ascribed neither to employed nor unemployed. Those are children, non-working pupils and students, housewives, non-working pensioners, disabled, renters, prisoners, discouraged persons.

3 Labour Force Survey

3.1 Sampling plan

The sampling plan is a one-stage simple random sample of 4 000 individuals aged 15 years and over, using the Population Register as a sampling frame. All persons in the household of the sampled individual are also included into the sample, so that the total sample is approximately 12 600 individuals per quarter. As a result of this sampling design, the inclusion probability of each household is proportional to the number of persons aged 15 years and over in the household. The cluster sample of persons is thus obtained. All the persons living at the address selected belong to the same cluster. The interviewer indicates the actual composition of the cluster when visiting the household.

Each household is surveyed for four quarters according to the rotation pattern 2-(1)-2.

3.2 Weighting Procedure

The weighting method for the Lithuanian LFS is based on the calibration method introduced by Deville and Särndal in 1992. The initial household design weights are adjusted by the use of auxiliary information relating to the population data on the intersection of 13 age groups, sex and urban-rural area as well as 58 municipality groups.

The calculation of the calibrated weights is carried out with SAS macro program CLAN, developed by Statistics Sweden. CLAN is also used for the estimation of the variances.

3.3 Data Collection

Data are collected by face-to-face interviews in the first wave and by face-to-face or telephone interviews in the next waves using paper questionnaires.

The interviewing is normally done during the week immediately following the reference week but never later than five weeks after the reference week.

4 Estimation of the number of non-official emigrants

Emigrants who have not declared their departure are included into in the Population Register. So, these persons can be selected to the LFS sample.

The Labour Force Sample Survey questionnaire is filled in by a direct interview mode, interviewing the person, who best knows the composition of the household and the reasons why the sampled person did not answer the questionnaire's questions. If there is no one living at the sampled address, the interviewer (if there are possibilities) asks the neighbours and indicates the reason for non-response to the questionnaire. One of the non-response reasons is "Left Lithuania (to work, live abroad)"; if this answer is marked, an annex to the questionnaire is to be filled in. The interviewer asks year and month of leaving the country, new country of residence, education and occupation before leaving.

Statistics Lithuania carried out this survey in the first quarter of 2006.

We consider that population data (LFS auxiliary information) include all Lithuanian residents and non-official emigrants. Number of employed, unemployed, inactive persons and non-official emigrants has to be estimated using auxiliary information from demographic data meeting the following condition:

Residents of Lithuania from population data = employed population + unemployed population + inactive population + non-official emigrants

Number of non-official emigrants who left Lithuania in 2001- 2005 has to be estimated.

4.1 Adjusted design weights. Let us introduce the notation:

M - population size according to the Population Register,

N - number of households in the population according to the Population Register,

n - number of responding households in the sample,

s – sample of households,

m_i - number of registered members of the i-th household, $i \in s$,

π_i - inclusion probability of the i-th household into the sample:

$$\pi_i = \frac{nm_i}{M}, \; i=1,2,...,N,$$

d_i - adjusted design weight of the i-th household:

$$d_i = \frac{1}{\pi_i}, \; i \in s .$$

4.2 Calibrated weights. Deville and Särndal (1992) have introduced the method of calibration of weights which can incorporate auxiliary information into the estimator.

To estimate the number of non-official emigrants, let us define values of the variable y: y_{ij}=1, if the j-th person of the i-th household is non-official emigrant and y_{ij}=0, if this person is not non-official emigrant. The total $T_y = \sum_{i=1}^{N} \sum_{j} y_{ij}$ has to be estimated.

Notation. Let us replace the variable y in the population of persons by the variable z in the population of households so that z_i be equal to the sum of values of the variable y in the i-th household, $z_i = \sum_j y_{ij}, i \in s$. The population of persons is divided into C groups by territory and intersection of age, sex, and urban/rural residence place. Sizes of these groups in the population are known from the demographic data.

Let us introduce a vector-column of auxiliary information for each household $\mathbf{x}_i = \left(x_{i1}, x_{i2},..., x_{iC}\right)'$, $i=1,2,...,N$, where x_{ic} means the number of individuals of the i-th household, belonging to the c-th group, $c=1,2,...,C$, the matrix $\mathbf{X} = \left(\mathbf{x}_1, \mathbf{x}_2,..., \mathbf{x}_n\right)$ and size $n \times n$ matrix $\mathbf{D}$ with design weights of the household d_i, $i \in s$ on the diagonal and other elements equal to zeros.

Denote by $\mathbf{T}_x = (T_{x1},..., T_{xC})'$ the population constants,

$T_{xc} = \sum_{i=1}^{N} x_{ic}, c=1,2,...,C,$

$\hat{\mathbf{T}}_{xw} = \left(\hat{T}_{x1},..., \hat{T}_{xC}\right)'$, $\hat{T}_{xc} = \sum_{i=1}^{n} d_i x_{ic}$, $c=1,2,...,C,$

$T_y = \sum_{i=1}^{N} \sum_{j} y_{ij} = \sum_{i=1}^{N} z_i = T_z .$

Definitions. The essence of the calibration method is to get new weights w_i, $i \in s$ for the households that would yield a new estimator of the total of the variable *z*:

$$\hat{T}_z^{(cal)} = \sum_{i \in s} w_i z_i \ ,$$

minimizing some distance between the design weights d_i and the new weights w_i:

$$\min \sum_{i \in s} d_i \left(\frac{w_i}{d_i} - 1 \right)^2 ,$$

and satisfying the calibration equation $\sum_{i \in s} w_i x_i = \mathbf{T}_x$.

New weights. The resulting weights $w_i = d_i \left(1 + \vec{\lambda}' \mathbf{x}_i\right), i \in s$ are called calibration weights, and $\vec{\lambda}$ is obtained as a solution of the system of linear equations $\left(\mathbf{X}'\mathbf{D}\mathbf{X}\right)\vec{\lambda} = \mathbf{T}_x - \hat{\mathbf{T}}_{xw}$. The estimator of $T_y = T_z$ defined by

$$\hat{T}_z^{(cal)} = \sum_{i \in s} d_i z_i \left(1 + \vec{\lambda}' \mathbf{x}_i\right) = \sum_{i \in s} \sum_j w_i y_{ij}$$

is the calibrated estimator of total and coincides with the regression estimator. The estimation of the number of the employed and unemployed persons is done in the same way.

5 Results

The survey results show that only each second-third resident of Lithuania declares his departure when emigrating, i.e. 1.8% of Lithuanian population are non-official emigrants.

Taking into consideration non-official migration, since 1990, over the past 16 years, 404 thousand persons emigrated from Lithuania.

After Lithuania's accession to the EU, in 2004 as compared with 2003, the number of non-official emigrants grew twice. Over the previous year, the number of non-official emigrants had stabilised.

Table 1. Non-official emigrants

	Non-official emigrants, thousand			Coefficient of variation, %		
	All	Males	Females	All	Males	Females
All	69.8	39.0	30.8	9.2	11.3	12.1
2001-2003	20.7	10.8	9.9	16.7	21.2	19.2
2004	24.7	11.4	13.3	16.7	22.4	20.8
2005	24.4	16.8	7.6	13.8	15.7	21.4

Value of the coefficient of variation is quite high (see Table 1), especially for the estimates of 2001 and 2002. It is not difficult to answer why. Interviewers cannot find persons at the addresses in case all family members emigrated, and neighbours do not know about it, or these persons have already declared their departure.

The results of official emigration shows similar tendency: emigration increased in 2003 (see Figure 1).

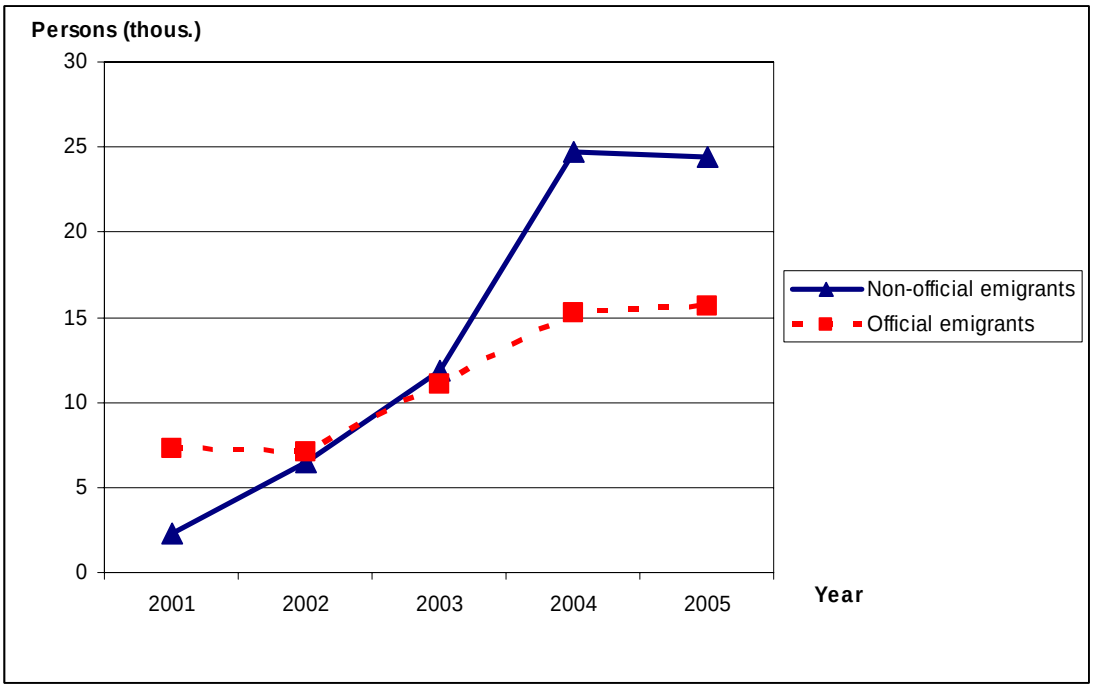


Figure 1. Official and non-official emigrants

6 Final remarks

Results on number of non-official emigrants are very interesting and useful. They are not very reliable (especially for 2001 and 2002) but give a general view of emigration in Lithuania.

Do we obtain the real number of emigrants if we sum non-official emigrants and official emigrants? It is difficult to answer.

Statistics Lithuania plans to repeat this survey next year.

References

SARNDAL C.-E., SWENSSON B., WRETMAN J. (1992) *Model Assisted Survey Sampling, Springer - Verlag.*

DEVILLE J.-C., SARNDAL C.-E. (1992), Calibration Estimators in Survey Sampling, *J. American Statistical Association*, vol. 87, No 418, 376-382.

EU Farm Accountancy Data Network (EU FADN) and Ukraine

Nataliya Skachek [1]

[1] Scientific and Technical Complex of Statistical Research, Ukraine;
e-mail: nskachek@mail.ru

Abstract

This paper deals with the approach of EU countries to creation and maintenance of Farm Accountancy Data Network. In comparison to it the situation in Ukrainian agricultural statistics is discussed. Main peculiarities determining the importance of Farm Accounting Data Network for the Ministry of Agricultural Policy of Ukraine and difficulties of its implementation are distinguished.

1. Introduction

Like many countries of the world Ukraine needs reliable sources for efficient administration process in agriculture and providing food security for it's citizens. Also data stream must be rather quick to provide timely administrative arrangements in the field of agriculture. It is known that sample survey basis provides less costly and less labour-intensive solution for decision-making. The experience of EU countries concerning this point may be considered as the example.

2. EU FADN

2.1. History

EU FADN was established in 1965 in accordance with the regulation of the European Commission. It provides physical and structural data like location, crop areas, livestock number, labour force and others. But the most important aim of EU FADN consists in providing economic and financial data on agriculture that are necessary for efficient management and quick interventions in this field. In European Union the network accounts for 60000 of agricultural enterprises of different forms of ownership, types of farming and economic sizes.

2.2. Methodology and Final Results

FADN Commission provides for each state certain number of enterprises to be included into sample. Than national authorities develop sampling plan for region institutions that collect data directly. Legal form, type of farming and economic size of the enterprise are main criteria for sampling technique. These indicators are also used for final presentation of results. For example, in Germany farm returns are represented by legal forms and types of farming with next sharing for strong, middle and weak enterprises. In member countries legal forms of the enterprises may vary but the types of farming are actually the same. Each country of the European Union represents its

final results by the following groups of farm return: income, sample and population, structures, subsidies, balance subsidies and taxes, balance sheet and financial situation. Indicators are calculated in accordance with the regulations of European Commission and than passed to FADN Committee of EC.

2.3. Liaison Agencies

It is important to note that each member country has its own institution responsible for collection and aggregation of indicators. These institutions vary from country to country. At national level National Statistical Offices and Ministries of Agriculture (with statistical services and associated bodies, like Agricultural Economic Institutes or other specialized institutions) are the main bodies. The distribution of tasks between National Statistical Offices and Ministries of Agriculture can vary considerably from country to country, so that, for example, in countries like France and Spain nearly all work within FADN is attributed to the statistical services of the Ministry, whereas in other countries, like Sweden, nearly all the work is done by National Statistical Office. Other countries have intermediate solutions. It is important to mention that in countries with more federal structure, like Germany, Spain and Italy, relatively independent regional statistical services are the basic units of data collection and thus coordination is necessary at national level. For example, in Germany such coordination is provided by the Federal Ministry of Consumers Protection, Food and Agriculture.

3. FADN in Ukraine

3.1. Agricultural Statistics in Ukraine

At present the Ministry of Agricultural Policy of Ukraine is main user of agricultural statistics in our country. It is interested in information for all its departments and divisions for the purposes of analysis, planning, efficient decision-making and agricultural policy maintenance. The State Statistics Committee of Ukraine is main generator of agricultural data. It provides aggregated agricultural data to the Ministry of Agricultural Policy in accordance with statistical forms that are collected from the enterprises. The process of data collection in the field of agriculture is conducted by Department of Agricultural and Environment Statistics of the State Statistics Committee of Ukraine. In close cooperation with the Ministry of Agricultural Policy of Ukraine Department of Agricultural and Environmental Statistics develops forms of agricultural statistical reporting. All legal enterprises are obliged to report on complete enumeration basis. Large farms represent annual forms of reporting and forms of reporting at certain date and private farms provide only annual reporting. Statistical forms conclude mostly natural indicators. Only form 50-agriculture contains such value indicators as costs of production, revenue from sales, operation costs, land rent and

others. Besides that legal enterprises compile financial accounting reporting due to National Accounting Standards that were agreed with the International Accounting Standards in 2000. These forms are also collected by statistical entities on complete enumeration basis. But they are not aggregated in departments of agricultural statistics at regional and national levels. They are compiled by Department of Financial Statistics which makes only general aggregations by economic activities including agriculture. But financial data together with all forms of agricultural statistics represent valuable source of information. Most of the indicators for the calculation of farm return may be obtained from these statistical forms. So the question of application of the farm accountancy data network first of all concerns the optimal maintenance of available sources.

3.2. Households in Agricultural Statistics

But there is one more problem. The matter is that agricultural enterprises (large farms and private farms) produce only about 40% of all agricultural output. And more than 60% of production is given by households. Households are not legal entities; they do not compile any forms of reporting. The only information about their activity is possible because of two sample surveys that are conducted by the State Statistics Committee of Ukraine. One of them is the Sample Survey of Expenses and Incomes of Households and the other is the Statistical Survey of the Agricultural Activity of the Households. The participation in these surveys is voluntary; heads of the households are not obliged to provide any documents except accordingly filled questioners so the point of reliability depends on the respondents.

3.3. Advisory Services

It is possible to improve the quality of information in mentioned surveys if the households within the sample survey work out accounting report either themselves or with the help of corresponding advisory services. Different advisory institutions function in more than 120 countries of the world. In Ukraine first steps of their creation goes back to 1993. But today quantity of these services is not enough throughout the country. That is why the Concept of State Program of Agricultural Advisory Activity was approved last year. It is intended to accelerate the process of creation of the efficient network of advisory services able to provide to agricultural producers its activity. With the help of these services it will be possible to enhance accounting procedures, consulting in tax assessment, insurance and legal implementation. Hence the process of data unification will be simplified.

Conclusion

FADN creation in Ukraine provides a lot of questions to be solved. It is necessary to ensure financial and legal support, to approve catalog of indicators, to confirm the methodology of data

converting to final results established by the regulations of European Commission and to come to an agreement about other important issues. It is worth to mention that obtained statistical data are of great interest not only for the purposes of collaboration with European Union but also as reliable source for decision-making in agricultural policy.

References

1. *Regulation* No 79/65/EEC of the Council of 15 June 1965 // CONSLEG 1965R0079 – 05/06/2003, Consolidated Text produced by Consleg system of the Office for Official Publications of the European Communnities – p.11.
2. *Regulation* No 2237/77/EEC of the Council of 23 September 1977 // CONSLEG 1977R2237 – 01/01/2002, Consolidated Text produced by Consleg system of the Office for Official Publications of the European Communnities – p.48.
3. *Ernährungs*- und agrarpolitischer Bericht des Bundesregierung 2003, Bundesministerium für Verbraucherschutz, Ernährung und Landwirtschaft, p.184.
4. *Віталій Дерлеменко.* Сільськогосподарські інформаційні консультаційні служби. – К.: Інститут аграрної економіки, 2001. – 452 с.
5. *Концепція* Державної цільової програми сільськогосподарської дорадчої діяльності на 2006-2009 роки, схвалена розпорядженням Кабінету Міністрів України від 20 червня 2005 р. № 210-р. – http://www.minagro.kiev.ua.

THE USE OF ADMINISTRATIVE DATA SOURCES FOR LITHUANIAN ANNUAL DATA OF EARNINGS

Milda Šličkutė-Šeštokienė

Statistics Lithuania, Lithuania

e-mail: milda.slickute@stat.gov.lt

1 Abstract

Statistics Lithuania has the full range of labour statistics that meet the timeliness and demands of Eurostat and national needs. The challenge is to keep this quality and timeliness and to publish even more detailed information and at the same time spare costs.

Users need more and more statistical information and at the same time respondents want to get less and less questionnaires. That enforce Statistics Lithuania to seek for new methods for estimation of statistical information required.

This presentation describes the introduction of administrative sources at estimation stage for data of earnings. Generalized Regression estimator of total and ratio is examined. Introduction of administrative sources at estimation stage significantly improved the quality of the statistical estimates and spared the burden and the costs.

2 History of Annual Survey of Earnings

Until 2003 Annual Survey of Earnings (ASE) used to be performed completely enumerating all enterprizes. According to the one of the goals of Statistics Lithuania, to diminish burden for enterprizes as much as possible using administrative sources, Labour statistics division decided to reject ASE and to calculate annual data of earnings for 2004 on the basis of Quarterly Survey of Earnings (QSE) and data of Social Insurance (SI).

It is supposed that usage of administrative sources will diminish the burden for enterprizes as well as for staff of Statistics Lithuania keeping quite good quality of statistical data.

The year 2003 were chosen for simulation and consideration of methods that could be used for estimation, because it is the only year when all three sources (ASE, QSE and SI) are available. The ASE where rejected since 2004 and data of SI become available

since 2003. All methods were analyzed for the year 2003 and it was compared with the real figures of Annual Survey of Earnings 2003.

3 Simulation and results

3.1 Sources available

As mentioned before annual data on earnings 2004 was estimated on the basis of two sources:

- Quarterly Survey of Earnings;
- Data of Social Insurance (administrative source).

Quarterly Survey of Earnings is conducted applying sampling methods. A simple random stratified sample is used. The Horvitz-Thompson estimator is applied to estimate the parameters of interest in each domain. The definitions of main variables of Quarterly Survey on Earnings and Annual Survey of Earnings is the same. The main reasons why two surveys duplicating variables used to be performed are following:

- Quarterly data are required every quarter for national needs;
- Detailed breakdown of annual data requires complete enumeration.

Data of Social Insurance that available for Statistics Lithuania are for the year 2003 and later. Definitions of statistical variables and variables of Social Insurance does not coincide but statistical variables are well correlated with the variables of SI. So it was decided to exploit variables of SI as auxiliary information at estimation stage. High correlation ensure improvement of quality. So it is expected to achieve the breakdown of annual data using the sample of quarterly survey.

Variables of SI analyzed:

- Number of insured persons: average on number of insured persons at the beginning of each quarter and end of each quarter;
- Taxable income per year;
- Days worked.

Coefficient of correlation for variables of Quarterly Survey of Earnings with variables of Social Insurance was calculated at NACE section level for each quarter of the years 2003 and 2004. The distribution of coefficients of correlation is presented in the table bellow.

Coefficients of correlation between variables of QSE and variables of SI, 2003 and 2004

Variable in SI	Coeff of corr	Variable in Quarterly Survey of Earnings				
		Number of employees	Number of full-time units	Gross remuneration	Hours worked	Hours paid
Number of insured	<0.8	0.0	0.0	3.1	0.0	0.0
	0.8-0.9	0.0	1.6	29.7	1.6	1.6
	0.9-1	100.0	98.4	67.2	98.4	98.4
Taxable income	<0.8	2.3	1.6	0.0	0.8	3.9
	0.8-0.9	26.6	19.5	0.0	21.9	19.5
	0.9-1	71.1	78.9	100.0	77.3	76.6
Days worked	<0.8	0.0	0.0	2.3	0.0	0.0
	0.8-0.9	0.0	0.0	28.1	0.0	0.8
	0.9-1	100.0	100.0	69.5	100.0	99.2

From the table above we can see that in most cases coefficient of correlation is higher than 0.9, some of them fall into interval [0.8; 0.9) and only few cases when it is less than 0.8. So it could be affirmed that there exist well-correlated auxiliary variables.

3.2 Notation

The purpose is to examine the estimation for domains using auxiliary information. It was decided to analyze possibility to introduce General Regression Estimator (GREG) for estimation of annual data.

Let us denote $U = \{1, 2, ..., k, ..., N\}$ - the *sample frame*. A probability sample s is drawn from U according to the specified sampling design. The sample size is denoted by n. The first order *inclusion probabilities* are denoted by $\pi_k = P(k \in s)$, the second order inclusion probabilities are denoted $\pi_{kl} = P(k\&l \in s)$. The corresponding *sampling weights* denoted $d_k = 1/\pi_k$ and $d_{kl} = 1/\pi_{kl}$.

Lets denote y - *the variable of interest*. The value of the *auxiliary variable vector* for the k-th element is denoted by $\mathbf{x_k} = (x_{1k}, ..., x_{Jk})$, J - the number of auxiliary variables.

The objective is to estimate the unknown y total:

$$t_y = \sum_{k \in U} y_k \tag{1}$$

when we have observed $(y_k, \mathbf{x_k})$ for $k \in s$ and when $\mathbf{x_k}$ is also known for $k \in U \backslash s$.

Generally, when J auxiliary variables are present, the General Regression Estimator is given by

$$\hat{t}_y^{greg} = \hat{t}_y + \sum_{j=1}^{J} \hat{B}_j (t_{x_j} - \hat{t}_{x_j}) = \hat{t}_y + \hat{\mathbf{B}}'(\mathbf{t_x} - \hat{\mathbf{t}}_\mathbf{x}) \tag{2}$$

where $\hat{t}_y$ is Horvitz-Thompson estimator of t_y, $\mathbf{t_x} = (t_{x_1}, ..., t_{x_J})'$ is the vector of known population total of the J auxiliary variables, and similar for $\hat{\mathbf{t}}_\mathbf{x}$, the vector of estimated population totals of the auxiliary variables. The $\hat{B}_1, ..., \hat{B}_J$ are components of the vector

$$\hat{\mathbf{B}} = \Big(\sum_{i \in s} d_i \mathbf{x_i} \mathbf{x_i'} \Big)^{-1} \Big(\sum_{i \in s} d_i \mathbf{x_i} y_i \Big) \tag{3}$$

The Generalized Regression Estimator can be alternatively written as

$$\hat{t}_y^{greg} = \sum_{i \in s} d_i g_i y_i, \tag{4}$$

where

$$g_i = 1 + (\mathbf{t_x} - \hat{\mathbf{t}}_\mathbf{x})' \Big(\sum_{i \in s} d_i \mathbf{x_i} \mathbf{x_i'} \Big)^{-1} \mathbf{x_i} \tag{5}$$

3.3 Simulation accomplished

The sampling frame and the sample of QSE is the same whole year. That is why it is possible to use the sample of QSE for annual data. The annual number of employees in the sampled enterprizes was calculated as average of four quarters, the gross earnings and hours - as the sum of the quarterly values.

The breakdowns required for Annual Survey of Earnings:

- NACE (two digits or sometimes even more detailed) & economic sectors (49 economic activities and 2 economic sectors), it is also the breakdown of QSE;

- NACE (section level) & size of enterprize & economic sector (15 economic activities, 6 sizes of enterprize and 2 economic sectors);;
- NACE (section level) & county (15 economic activities and 10 counties);
- Municipality (60 municipalities);

Total by $49 * 2 + 15 * 6 * 2 + 15 * 10 + 60 = 488$ partly overlapping domains are required. The sample size in 2004 is 6111 units. It is evident that it is impossible to get reliable data for such detailed breakdown using only the sample of QSE. As the data of SI became available for the statistical purposes it was decided to use this data as auxiliary information in order to calculate estimates by more detailed breakdown.

There are no problems with the first breakdown because it is also the breakdown of QSE and at the moment of sample was foreseen to get the results by this breakdown. The main problem is breakdown by regions because the quarterly survey does not aim to get data for the estimates for regions. If we want to get reliable results by detailed NACE and by regions using only the data from survey we need almost a complete enumeration of enterprizes. That used to be done till 2003.

The high correlation of variables of Social Insurance and variables of Annual Survey of Earnings let us expect that the usage of variables of Social Insurance will allow to switch from census of enterprizes to the sample survey.

The main task is to identify the most reliable vector of auxiliary information. As mentioned in 3.1 three auxiliary variables were analyzed. Also two levels of auxiliary information were examined:

- NACE at section level;
- NACE at section level & region at county level (10 counties).

Combining different auxiliary variables 14 GREG estimators were calculated: 7 possible combinations of three auxiliary variables multiplied by 2 levels of auxiliary information.

Notation of different GREG estimators

Notation	Auxiliary information used
G1, G8	Number of employees
G2, G9	Taxable income
G3, G10	Days worked
G4, G11	Number of employees and taxable income
G5, G12	Number of employees and days worked
G6, G13	Taxable income and days worked
G7, G14	All variables

G1 - G7 refer to auxiliary information at NACE section level and G8 - G14 refer to auxiliary information at NACE section level & county.

The main criteria for choosing the most suitable estimator from the list above was variance and distribution of weights g_i presented in formula (5). The variance should be as small as possible and the weights g_i should not be scattered too mush. But unfortunately, as presented in the tables bellow, the smaller the variance the weights g_i are more scattered.

Distribution of weights g_k for different GREG estimators 2004, in per cent

g_i	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12	G13	G14
< 0.4	0	0	0	0	0	0	0	2	2	67	3	4	65	5
$[0.4; 0.8)$	1	0	1	1	1	1	1	7	7	11	8	8	11	9
$[0.8; 1.2)$	97	97	97	96	95	96	94	59	59	7	56	54	7	52
$[1.2; 1.6)$	2	2	2	3	3	3	4	26	26	7	26	25	8	25
≥ 1.6	0	0	0	0	0	0	0	6	6	8	7	9	9	10

Distribution of the coefficients of variation for different GREG estimators 2004, in per cent

CV	**HT**	**G1**	**G2**	**G3**	**G4**	**G5**	**G6**	**G7**	**G8**	**G9**	**G10**	**G11**	**G12**	**G13**	**G14**
[0; 5)	44	44	44	44	44	43	44	44	70	63	83	82	75	88	86
[5; 10)	16	15	14	15	14	15	13	14	14	18	8	7	11	6	5
[10; 30)	26	26	26	26	27	27	26	26	11	13	6	7	10	3	5
≥ 30	15	16	16	16	16	15	16	16	5	6	3	4	4	4	4

The GREG estimator which was chosen for estimation of Annual Data of Earnings 2004 is G8, it uses only the number of insured persons as auxiliary information and level of auxiliary information is NACE at section level & county. This estimator is a compromise between small variance and scatter of weights g_i: 70% of coefficients of variation are less than 5% and 59% of weights g_i fall into interval [0.8; 1.2).

3.4 Precision gained

The chosen GREG estimator G8 was compared with the HT estimator. Some variables were improved very significantly but some only a bit. In the table bellow it is presented the distribution of coefficients of variation for the variable "Average number of employees" and for all variables altogether.

Distribution of statistical estimates (G8) by size of the coefficient of correlation (CV) 2004

CV	All variables		Average number of employees	
	G8	HT	G8	HT
Mean	9.8	14.4	7.0	18.5
Median	2.5	6.8	1.9	11.9
[0; 5)	70.1	43.7	76.7	32.4
[5; 10)	13.5	15.5	10.0	14.8
[10; 30)	11.4	25.7	8.1	31.4
[30; 50)	1.2	9.5	1.0	13.3
[50; 100)	3.7	5.6	4.3	8.1

We can see from the table above that median value of CV for chosen GREG estimator declines almost three times compare to HT estimator for all variables altogether and more than 6 times for average number of employees. Also the number of estimates with CV less than 5% is significantly higher for GREG estimator compare to HT.

As mentioned above all estimates for 2003 were compared with the real figures of ASE 2003. In fact the frame for ASE is not the same as for QSE but most enterprizes belongs to both frames. So G8 estimates may not coincide with the respective figures from ASE but should be close.

Distribution of statistical estimates for Average Number of Employees by deviation from ASE 2003, in %

Interval of deviation, in %	**Number of statistical estimates, in %**	
	G8	**HT**
[0; 5]	54.1	35.5
(5; 10]	16.9	19.8
(10; 20]	11.6	15.1
(20; 50]	12.8	21.5
50 and	4.7	8.1

From the table above it can be noticed that the G8 estimates for Average Number of Employees are closer to corresponding figures from Annual Survey of Earnings than the HT estimates. Similar situation are found calculating deviation for all other variables.

3.5 Improvements foreseen

Analyzing the results it was noticed some improvements that should be introduced for estimation of further annual data:

- More levels of auxiliary information should be analyzed and fist of all size of enterprize should be included;
- Maybe different auxiliary information should be used for estimation of different variables;
- Because of too detailed breakdown of data regression imputation should be implemented (variables should be imputed for whole population).

4 Conclusions

Introduction of administrative sources for estimation of data of Labour Statistics is undoubted a useful experience. Burden for enterprizes as well as for staff of Labour Statistics was significantly diminished. Approximately 40000 of enterprizes do not need to fill in annual questionnaire on earnings, staff of Statistics Lithuania do not need to enter and check those questionnaires and users are able to get information sooner than they used to. Useful experience enforced to start analysis of possibility to introduce administrative sources for other surveys on earnings.

5 References

[1] Deville, J; and Särndal, C.-E. Calibrated estimators in survey sampling. Journal of American Statistical Association, (1992, 87 p.376-382).

[2] Lundström, S. and Särndal, C.-E. Calibration as a standard method for treatment of nonresponse. Journal of Official Statistics, (1999, 15(2), p. 305-327).

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

Strength of Auxiliary Information for Compensating Non-response

Karolin Toompere

University of Tartu, Estonia
e-mail: ktoompere@hot.ee

Abstract

In this paper the calibration approach is introduced. Calibration approach is a simple and effective method to weighing in the presence of the non-response. It uses the auxiliary information. Since it depends on the selection of the auxiliary information how well the calibration estimate can reduce the bias, the main focus is on finding the best auxiliary information available. Some computable indicators, non-response indexes, are constructed to guide this search. In the practical example the non-response and samples are generated and the indexes and estimates are computed to see how well the non-response indexes work in practice.

1 Auxiliary information

The auxiliary information is given by an auxiliary vector. The value of an auxiliary vector is known for every responding object, $k \in r$ and also for a larger set than r.
Denote the auxiliary vector by $\mathbf{x}$. Its value for object k is denoted $\mathbf{x}_k$. The information for the larger set is provided by information input $\mathbf{X}$.

2 Calibration estimators

The calibration estimator of total $Y = \sum_U y_k$ is

$$\hat{Y}_W = \sum_U w_k y_k , \qquad (1)$$

where w_k is the calibrated weight for the object k. It depends on the auxiliary information how effective the calibrated weights are.

We are searching for a system of calibrated weights w_k , k∈r, that satisfy the calibration equation

$$\sum_r w_k \mathbf{x}_k = \mathbf{X}. \qquad (2)$$

When there is non-response then the sampling weights $d_k = 1/\pi_k$ are too small for all or most of the sampled objects. We are looking for the weights that are larger than the sampling weights. Therefore we multiply d_k with some factor v_k. That is $w_k = d_k v_k$. We construct the factors v_k so that they depend linearly on the known value x_k. One simple form is

$$v_k = 1 + \lambda' x_k . \qquad (3)$$

The vector λ should satisfy the calibration requirement. Insert (3) to equation (2) and solve it for λ. Then $\lambda = \lambda_r$, where (Särndal, Lindstöm, 2005)

$$\lambda_r' = (X - \sum_r d_k x_k)'(\sum_r d_k x_k x_k')^{-1}, \quad (4)$$

if the inverse matrix of $\sum_r d_k x_k x_k'$ exists. We have found the weights that account for non-response and are calibrated to the given information.

$$w_k = d_k(1 + \lambda_r' \mathbf{x}_k), \quad (5)$$

where λ_r' is given by (4).

3 Properties of calibrated weights

Lets take a look of the calibrated weights in (5). They can be written

$$w_k = w_{Mk} + w_{Rk},$$

where

$$w_{Mk} = d_k\{\mathbf{X}'(\sum_r d_k \mathbf{x}_k \mathbf{x}_k')^{-1}\mathbf{x}_k\}$$

and

$$w_{Rk} = d_k\{1 - (\sum_r d_k \mathbf{x}_k)'(\sum_r d_k \mathbf{x}_k \mathbf{x}_k')^{-1}\mathbf{x}_k\}.$$

The component w_{Mk} is called the main component and w_{Rk} is remainder component.
It can be seen that

$$\sum_r w_{Mk}\mathbf{x}_k' = \mathbf{X}'(\sum_r d_k \mathbf{x}_k \mathbf{x}_k')^{-1}(\sum_r d_k \mathbf{x}_k \mathbf{x}_k') = \mathbf{X}'$$

and

$$\sum_r w_{Jk}\mathbf{x}_k' = \sum_r d_k\{1 - (\sum_r d_k \mathbf{x}_k)'(\sum_r d_k \mathbf{x}_k \mathbf{x}_k')^{-1}\mathbf{x}_k\}\mathbf{x}_k' =$$
$$= \sum_r d_k \mathbf{x}_k' - \sum_r d_k \mathbf{x}_k \mathbf{x}_k'(\sum_r d_k \mathbf{x}_k \mathbf{x}_k')^{-1}(\sum_r d_k \mathbf{x}_k)' = 0.$$

When we choose the vector $\mathbf{x}_k$ so that $\mu'\mathbf{x}_k = 1$ for some constant vector μ, $k \in U$ then $w_{Rk} = 0$ for every $k \in U$ and $w_k = w_{Mk}$, because then $\sum_r d_k \mathbf{x}_k = \mu' \sum_r d_k \mathbf{x}_k \mathbf{x}_k'$ and

$$w_{Rk} = 1 - (\mu' \sum_r d_k \mathbf{x}_k \mathbf{x}_k')(\sum_r d_k \mathbf{x}_k \mathbf{x}_k')^{-1}\mathbf{x}_k = 1 - \mu'\mathbf{x}_k = 0$$

for every k. The calibrated weights are

$$w_k = w_{Mk} = \mathbf{X}'(\sum_r d_k \mathbf{x}_k \mathbf{x}_k')^{-1} d_k \mathbf{x}_k \ . \quad (6)$$

4 Classification

Consider one simple case of the calibration estimator where the auxiliary information is about a classification of elements. We have P groups U_P. The group identifier for object k is defined as

$$\gamma_k = (\gamma_{1k}, \gamma_{2k} \ldots, \gamma_{Pk})',$$

where $\gamma_{pk} = 1$ if k belongs to the group U_P and $\gamma_{pk} = 0$ if not.

Let r_p be the respond set in the group U_P. From the equation (5) or (6) we get

$w_k = d_k F_p^*$ for every $k \in r_p$, where $F_p^* = N_p / \sum_{r_p} d_k$. The calibration estimator (1) is

$$\hat{Y}_W = \sum_{p=1}^{P} F_p^* \sum_{r_p} d_k y_k = \sum_{p=1}^{P} N_P \bar{y}_{r_{p;d}} = \hat{Y}_{PWA},$$

where

$$\bar{y}_{r_{p;d}} = \sum_{r_p} d_k y_k / \sum_{r_p} d_k$$

is the design-weighted group mean for respondents.

5 Non-response indexes

5.1 IND1

Denote the response probability of object k by θ_k. In the presence of non-response the response influences $\phi_k = 1/\theta_k$ are unknown. Therefore we can't find the unbiased estimator $\hat{Y} = \sum_r d_k \phi_k y_k$ for the total $Y = \sum_U y_k$. But we can find a calibration estimator $\hat{Y}_W$. The estimator $\hat{Y}_W$ is nearly unbiased if ϕ_k is linearly related to the auxiliary vector $\mathbf{x}_k$. As we don't know the response influences the vector can't be determined. But we are using the proxies $\hat{\phi}_k$. We take $\hat{\phi}_k = v_k$.

It is important that the proxies reflected the differences between the sampled elements. The more the proxies vary, the better. Therefore use the variance of $\hat{\phi}_k = v_k$ to indicate the strength of $\mathbf{x}_k$.

$$IND1 = \frac{1}{\sum_r d_k} \sum_r d_k (v_k - \bar{v}_{s:r;d})^2,$$

where

$$\bar{v}_{s:r;d} = \sum_r d_k v_k / \sum_r d_k$$

is the mean of v_k.

If $\mu' \mathbf{x_k} = 1$ for every k and some constant μ then

$$IND1 = \frac{(\sum_s d_k x_k)' (\sum_r d_k x_k x_k')^{-1} (\sum_s d_k x_k)}{\sum_r d_k} - \frac{(\sum_s d_k)^2}{\sum_r d_k}.$$

Consider the classification, where $\mathbf{x_k} = \gamma_k = (\gamma_{1k}, \gamma_{2k} \ldots, \gamma_{Pk})'$ and x_k is known for every $k \in s$. Then

$$IND1 = \frac{1}{\sum_r d_k} \left(\sum_{p=1}^{P} \frac{(\sum_{s_p} d_k)^2}{\sum_{r_p} d_k} - \frac{(\sum_s d_k)^2}{\sum_r d_k} \right). \quad (7)$$

For SI the weights $d_k = N/n$ and the index (7) takes the form:

$$IND1 = \frac{1}{m}\left(\sum_{p=1}^{P}\frac{n_p^2}{m_p} - \frac{n^2}{m}\right),$$

where m_p is the number of respondents in the group U_p, n_p is the number of sampled elements in the group U_p and m is the total number of respondents from the sample n.

5.2 IND2

The bias is small when the auxiliary vector explains the study variable, that is the residuals $e_k = y_k - x_k\mathbf{B}_U$ are zero for every k, where $\mathbf{B}_U = (\sum_U \mathbf{x}_k\mathbf{x}_k')^{-1}\sum_U \mathbf{x}_k y_k$ (Särndal, Lundström, 2005). In practice this condition doesn't hold but if the residuals are small, the bias and variance of $\hat{Y}_W$ are also small. Therefore the second index measures how close are the residuals to zero.

The variability of y_k is given by

$$SST_Y = \sum_U (y_k - \bar{y}_U)^2,$$

where $\bar{y}_U = (1/N)\sum_U y_k$ is the population mean.

$$SSR_Y = \sum_U (y_k - \tilde{y}_k)^2$$

measures residual variance where $\tilde{y}_k$ is a value obtained for k by a fitting procedure that delivers $\tilde{y}_k$ as a function of the individual auxiliary vector value $\mathbf{x}_k$. Then $(SST_Y - SSR_Y)/SST_Y$ measures the variance explained. In our situation SST_Y and SSR_Y are respectively

$$\hat{SST}_Y = \sum_r d_k v_k (y_k - \bar{y}_{r;dv})^2$$

and

$$\hat{SSR}_Y = \sum_r d_k v_k (y_k - \hat{y}_k)^2,$$

where

$$\bar{y}_{r;dv} = \frac{\sum_r d_k v_k y_k}{\sum_r d_k v_k},$$

$$\hat{y}_k = \mathbf{x}_k'(\sum_r d_k v_k \mathbf{x}_k\mathbf{x}_k')^{-1}\sum_r d_k v_k \mathbf{x}_k y_{k.}.$$

The index for y is

$$IND2 = 1 - \frac{\sum_r d_k v_k (y_k - \hat{y}_k)}{\sum_r d_k v_k (y_k - \bar{y}_{r;dv})}.$$

This indicator shows how well the auxiliary vector explains the study variable y.

Consider the classification discribed before and SI sample.

Then the IND2 takes a form

$$IND2 = 1 - \frac{\sum_{p=1}^{P} \frac{n_p}{m_p} \sum_{r_p} (y_k - \hat{y}_k)^2}{\sum_{p=1}^{P} \frac{n_p}{m_p} \sum_{r_p} (y_k - \bar{y}_{r;dv})^2},$$

where

$$\bar{y}_{r;dv} = \frac{1}{n} \sum_{p=1}^{P} n_p \bar{y}_{r_p}$$

and $\hat{y}_k = \bar{y}_{r_p}$ when the object k belongs to the group U_p.

The indexes IND1 and IND2 are constructed to guide the search of auxiliary information to use for computing the calibrated weights. Before choosing the auxiliary vector both indexes should be computed for different auxiliary vectors. If both indexes are larger for some auxiliary vector than others then there is high probability that the auxiliary information is the best for reducing the non-response bias. The two indexes should be used together. We see one simple example of how the indexes work in practice.

6 Practical example

In a practical example we study the simple case of classification where the auxiliary information is about a classification of elements. The data used is the data of StatVillage (Schwartz,1997) that is based on the census on Canada. In the StatVillage there are 480 households that are divided into 60 blocks, in every block there are 8 houses. For every household different variables are measured.
Since IND1 and IND2 are random we are studying the indeces by generating repeatedly samples and non-response. Using different auxiliary information the bias and standard deviation of the calibration estimator is computed, also the indeces IND1 and IND2. We want to know how well the indeces indicate the strength of the auxiliary information – if the values of IND1 and IND2 are larger when the bias of calibration estimator is smaller. Also we want to know if the results are different if the non-response rate is increased. In this example the cases when the non-response is 30% and 50% are concerned.
Consider the SI sample. The sample size is 100. We are studying the total income of the household. To every object the probability of the non-response is generated:

$$(1 - \theta_k) = \frac{(n-m) y_k}{\sum_U y_k}$$

where n is the sample size, m is the number of people responding and y_k is the total income of object k. The non-response depends on the income of the household. Therefore there is considerable non-response error.
We use the probabilities of the non-response and generate the set of respondents.

- generate $u_k \sim U(0,1)/\theta_k$ for every objekt.
- order the object in the increasing order of u_k
- choose m first objects.

We are using three classifications:

1. The households are divided into groups using the location of dwelling (south/north). The first group consists of households who live in the blocks 1-30, the second group of households who live in the blocks 31-60. It is known that in the northern blocks (with smaller block numbers) live the wealthier households. Therefore the auxiliary information should describe well the total income of the household.
2. The household are divided into groups according to the location of the house inside the blocks. In the first group there are households whose house number is 1-4, the in the second group households whose house number is 5-8/.
3. The households are divided into four groups according to the location of dwelling(south/north) and the value of the dwelling.

The results can be seen in Table 1.

Auxiliary information	North/south		Location of the house in the block		North/south+value of the dwelling	
Nonresponse (%)	30	50	30	50	30	50
$B(\hat{Y}_{PWA})\times 10^6$	-1.94	-3.43	-2.8	-5.18	-1,67	-3,6
$\left[D(\hat{Y}_{PWA})\right]^{1/2}\times 10^6$	1,13	1,03	1,43	1,36	5,1	3,9
$E(IND1)$	0,03	0,27	0,002	0,03	0,06	0,52
$D(IND1)\times 10^{-16}$	0,19	2,94	0,07	0,18	0,002	2,8
$E(IND2)$	0,51	0,58	0,01	0,02	0,33	0,37
$D(IND2)\times 10^{-16}$	3,51	0,12	0,009	0,47	0,12	0

Table 1: The estimates and non-response indeces

The bias of the calibration estimator is in all cases negative. So the real total income is larger. The bias is largest if we use the location of the house inside the block. The same result can be seen of indicators IND1, IND2.

If we consider the value of the dwelling in addition to the geographical location we see that IND1 is really incresing but IND2 is decreasing. In both cases, when the non-response is 30% and 50%. Therefore we can't tell that the estimate is getting better when we add the information of the value of the house to the auxiliary vector. When we take a look of the bias of the estimator we can see that the bias is similar whether we use the informaton about the value of the dwelling or not.

The indicator IND2 helps us to find the information that decreases both – the bias and standard deviation. That can be seen also from the results. The variance is smallest when we are using the house's location whether in the north or south. The IND2 is also largest in that case.

References

Schwarz, C-J. (1997) Stat-Village: An on-Line, WWW-Accessible, Hypothetical city Based on Real Data for Use in an Introductory Class in Survey Sampling. *Journal of Statistics Education v.5, n.2.*.

Särndal, C-E. Lundström, S (2005) Estimation in Surveys with Nonresponse. John Wiley and Sons, Inc.

STATISTICAL ANALYSIS OF A SAMPLE OF SMALL-SCALE ENTERPRISES

Oksana Honchar [1], Nataliya Kravtsova[2] and Olga Vasylyk[3]

[1]Scientific and Technical Complex of Statistical Research, Ukraine
e-mail: ohonchar@list.ru

[2]Kyiv National Taras Shevchenko University, Ukraine
(graduate student)

[3] Kyiv National Taras Shevchenko University, Ukraine
e-mail: vasylyk@univ.kiev.ua

Abstract

In the paper we present some results of joint work performed by the authors within the framework of pilot sample survey of small-scale non-financial service enterprises in Ukraine.

1 Introduction

In 2005, with the purpose of application of sample survey methodology in the field of non-financial service the State Statistics Committee of Ukraine carried out pilot sample survey in some regions of Ukraine.

Population of small-scale non-financial service enterprises was stratified by real economic activity and by size of the enterprises (that is, the kind of basic activity and the number of employees were selected as criteria for stratification). Stratification by size resulted in three groups of enterprises:

1) enterprise with the number of employees from 0 till 9;
2) enterprise with the number of employees within the limits from 10 till 19;
3) 20 and more employees.

Also the following five arrays of enterprises were specified:

array 1: inactive enterprises – the enterprises with zero volume of realized service and no employees for two years;

array 2: zero enterprises - the enterprises with zero volume of realized service;

array 3: strata with less than 10 enterprises;

array 4: atypical enterprises;

array 5 (main array): other enterprises.

Each of these arrays of enterprises had to be analysed separately.

As a result of primary stratification 150 strata were obtained. Then the conditions and criteria for unification of strata were determined, so the number of strata decreased.

The size of the planned stratified sample was determined using Neyman optimum allocation (designed size of the sample was equal to 20,15%).

The volume of realized service was chosen as a parameter, which had to be estimated.

2 Analysis of data

In order to take into account the changes in strata during the survey, in particular due to nonresponses, and to obtain more precise results the weighting with subsequent calibration was used (Sarioglo, 2005).

An example of calculations made for the main array of data is presented below.

Base weight in the h-th stratum of the main array is

$$w_B = w_{Bh} = \frac{N_h}{n_h},$$

where N_h denotes the number of elements in the h-th stratum; n_h is designed size of a sample from the h-th stratum.

Weighting coefficient, which characterizes response in the h-th stratum, is denoted by

$$k_h = \frac{n_h - n'_h}{n_1},$$

where n_1 denotes real size of a sample from h-th stratum, n'_h is a number of elements out of survey among nonresponses in the h-th stratum.

Resulting weight in the h-th stratum: $w_{rh} = w_{Bh}\ k_h$.

The total volume of realized service for the main array of enterprises is estimated by the following expression:

$$X = \sum_{i=1}^{1457} x_i w_{ri}.$$

The mean error of the total for stratified sample under Neyman optimum allocation has the following form:

$$\mu = \sqrt{\sum_h \frac{\sigma_h^2}{n_h} N_h^2 \left(1 - \frac{n_h}{N_h}\right)},$$

where σ_h^2 denotes variance in the h-th stratum.

Since $\frac{N_h}{n_h} = w_{Bh}$, then one can easily derive the following expression:

$$\mu = \sqrt{\sum_h \sigma_h^2 w_{Bh} n_h \left(w_{Bh} - 1\right)}.$$

However, it was not possible to obtain data for all units planed for the survey, so the number of population units presented by one sample unit has changed. Therefore we use resulting weights w_{rh} instead of the base weights w_{Bh} in the expression for mean error:

$$\mu = \sqrt{\sum_h \sigma_h^2 w_{rh} n_h \left(w_{rh} - 1 \right)} .$$

Also we calculate limiting error Δ and relative limiting error Δ_r for the estimated total:

$$\Delta = t\mu \qquad \text{and} \qquad \Delta_r = \frac{\Delta}{X} \cdot 100\% ,$$

where *t* is a quantile of normal distribution.

Thus, for the main array of small-scale non-financial service enterprises the following results were obtained:

1) total volume of realized service: X = 169 450,1 × 10^3 UAH;

2) mean error of the total: μ = 6724.41 × 10^3 UAH;

3) limiting error: Δ = 13448.83 × 10^3 UAH (here we use quantile of normal distribution at the level 0,95);

4) relative limiting error: Δ_r = 8.65 %.

Thus, estimated total volume of realized service for the main array of non-financial service enterprises belongs to the interval (156001.2; 182898.9) with probability 95%.

Conclusion.

Here we presented some results of statistical analysis of the data, obtained from the pilot sample survey of small-scale non-financial service enterprises. The analysis was performed within the framework of cooperation between Kyiv National Taras Shevchenko University and Scientific and Technical Complex of Statistical Research. During this joint work we faced many interesting problems arising in processing of sample survey data and gained good experience of collaboration between theorists and practitioners. We plan to continue our cooperation, in particular to consider different approaches to the problem of nonresponses.

References

Chernyak, A. (2001) *Survey Sampling Technique*. Kyiv, Ukraine (in Ukrainian).

Honchar, O.V. (2004) Methodological aspects of processing nonresponses in sample surveys of small-scale enterprises. *Problems of Statistics*, no. 6, 109-114 (in Ukrainian).

Parkhomenko, V. (2001) *Survey Sampling Methods*. Kyiv, Ukraine (in Ukrainian).

Sarioglo, V.G. (2005) *Problems of statistical weighting of sample data*. Kyiv, Ukraine (in Ukrainian).

LIST OF PARTICIPANTS

Signe Bāliņa	signe@dzc.lv	Latvia	University of Latvia
Lennart Bondesson	Lennart.Bondesson@math.umu.se	Sweden	Umeå University, Sweden
Juris Breidaks	juris.breidaks@csb.gov.lv	Latvia	Central Statistical Bureau of Latvia
Natalja Budkina	budkinanat@gmail.com	Latvia	Riga Technical University, University of Latvia
Viktoras Chadyšas	viktorasch@gmail.com	Lithuania	VGTU
Andrius Ciginas	andrius.ciginas@maf.vu.lt	Lithuania	Statistics Lithuania
Andris Fisenko	Andris.fisenko@csb.gov.lv	Latvia	Central Statistic Bureau of Latvia
Anton Grafström	anton.grafstrom@math.umu.se	Sweden	Umeå University
Olga Grakoviča	Olga.grakovicha@inbox.lv	Latvia	Central Statistical Bureau of Latvia
Oksana Honchar	ohonchar@list.ru	Ukraine	Scientific and Technical Complex of Statistical Research
Vita Kozirkova	Vita.Besmenova@csb.gov.lv	Latvia	Central Statistical Bureau of Latvia
Danutė Krapavickaitė	krapav@ktl.mii.lt	Lithuania	Institute of mathematics and informatics, Lithuania; Statistics Lithuania
Gunnar Kulldorff	gunnar@matstat.umu.se	Sweden	University of Umeå
Seppo Laaksonen	Seppo.Laaksonen@Helsinki.Fi	Finland	University of Helsinki
Janis Lapins	Janis.Lapins@bank.lv	Latvia	Bank of Latvia
Risto Lehtonen	risto.lehtonen@helsinki.fi	Finland	University of Helsinki
Mārtiņš Liberts	Martins.Liberts@gmail.com	Latvia	Central Statistical Bureau of Latvia
Inga Masiulaityte	inmagik@yahoo.com, inga.masiulaityte@stat.gov.lt	Lithuania	Statistics Lithuania, Vilniaus University
Vilma Nekrasaite	nekrasaite.vilma@gmail.com	Lithuania	Statistics Lithuania, Vilnius Gediminas Technical University
Jelena Novika	Jelena.Novika@csb.gov.lv	Latvia	Central Statistical Bureau of Latvia
Pauli Ollila	Pauli.Ollila@stat.fi	Finland	Statistics Finland
Pasi Piela	pasi.piela@stat.fi	Finland	Statistics Finland
Aleksandras Plikusas	plikusas@ktl.mii.lt	Lithuania	Statistics Lithuania, Institute of Mathematics and Informatics
Dalius Pumputis	dpumputis@yahoo.co.uk	Lithuania	Institute of Mathematics and Informatics, Vilnius Pedagogical University
Virgi Puusepp	virgi.puusepp@stat.ee	Estonia	Statistics Estonia
Genovaite Saluckiene	genovaite.saluckiene@stat.gov.lt	Lithuania	Statistics Lithuania
Nataliya Skachek	nskachek@mail.ru	Ukraine	Scientific and Technical Complex of Statistical Research
Milda Slickute-Sestokiene	milda.slickute@stat.gov.lt	Lithuania	Statistics Lithuania

Kaja Sõstra	kaja.sostra@stat.ee	Estonia	Statistics Estonia
Daniel Thorburn	Daniel.Thorburn@stat.su.se	Sweden	University of Stockholm
Karolin Toompere	ktoompere@hot.ee	Estonia	University of Tartu
Imbi Traat	imbi.traat@ut.ee	Estonia	University of Tartu
Olga Vasylyk	ovasylyk@univ.kiev.ua	Ukraine	Taras Shevchenko Kyiv National University
Vaiva Virketyte	vaiva.virketyte@stat.gov.lt	Lithuania	Statistics Lithuania