Baltic-Nordic-Ukrainian Summer School on Survey Statistics

August 23-27, 2009
Kyiv, Ukraine

Kyiv, 2009

Proceedings of the Baltic-Nordic-Ukrainian Summer School on Survey Statistics. – Kyiv, "TBiMC", 2009. – 177 p.

Organizing institutions

University of Tartu, Estonia
University of Helsinki, Finland
University of Latvia
Vilnius University, Lithuania
Stockholm University, Sweden
University of Umeå, Sweden
Taras Shevchenko National University of Kyiv, Ukraine
Institute of Mathematics and Informatics, Lithuania
Institute for Demography and Social Research, Ukraine
Scientific and Technical Complex of Statistical Research, Ukraine
Statistics Estonia
Central Statistical Bureau of Latvia
Statistics Lithuania
State Statistics Committee of Ukraine

Organizing committee

Yuliya Mishura (chair)
Danutė Krapavickaitė
Gunnar Kulldorff
Risto Lehtonen
Volodymyr Sarioglo
Olga Vasylyk
Tetyana Yakovenko

Sponsors

The Visby Programme of the Swedish Institute
The organizing institutions

June 2009

ISBN 966-8725-02-6

Preface

The Summer School on Survey Statistics in Kyiv 2009 is the first event within the Baltic-Nordic-Ukrainian Network on Survey Statistics that takes place in Ukraine. The main objectives of the School are to provide an opportunity for university teachers, research students and survey practitioners to discuss their problems and to learn from the experiences in other countries.

The School starts with an opening session which includes presentations by a founder of the Network Gunnar Kulldorff (University of Umeå, Sweden), Nataliya Vlasenko (Deputy Chair of the State Statistics Committee of Ukraine) and Yuliya Mishura (Head of the Department of Probability Theory, Statistics and Actuarial Mathematics, Taras Shevchenko National University of Kyiv).

The Programme Committee invited three main speakers: Imbi Traat (Estonia), Risto Lehtonen (Finland) and Susanne Rässler (Germany) to give series of lectures of teaching nature. There are seven more invited speakers: Daniel Thorburn (Sweden), Danutė Krapavickaitė and Aleksandras Plikusas (Lithuania), Olga Vasylyk (Ukraine), Mārtiņš Liberts and Jānis Lapiņš (Latvia), Volodymyr Sarioglo (Ukraine). They will deliver special lectures covering different topics of the theory and application of survey statistics.

There are 51 registered participants at the School. Most of them will present contributed papers included in this book. All presentations will be followed by discussions.

We wish you have a good time in Ukraine and hope that participation at the Summer School will be fruitful and enjoyable for everyone.

On behalf of the Organizing Committee,

Yuliya Mishura

Olga Vasylyk

Tetyana Yakovenko

Contents

Abstracts of lectures

Opening session

Series of invited lectures

Special invited lectures

Contributed papers

Networking — from Umeå to Minsk and Dnipropetrovsk : A historical, pictorial and statistical review

Gunnar Kulldorff

University of Umeå, Sweden

e-mail: gunnar@matstat.umu.se

Abstract

The Baltic-Nordic-Ukrainian Network on Survey Statistics has developed sequentially: Umeå – Tartu 1992 – Riga 1994 – Kyiv 1994 – Vilnius 1997 – Helsinki 1997 – Jyväskylä 1997 – Stockholm 1998 – Örebro 1999 – Oslo 2002 – Copenhagen 2002 – Minsk 2007 – Dnipropetrovsk 2008 – Uzhgorod 2009.

The Network has nine official partners in six countries. It is linked to many universities, research institutes and national statistical agencies in the Baltic and Nordic countries, Ukraine and Belarus.

13 annual events (Summer Schools, Workshops and Conferences) have been arranged in Tartu 1997, Jurmala 1998, Palanga 1999, Pärnu 2000, Jurmala 2001, Ammarnäs 2002, Palanga 2003, Tartu 2004, Vilnius 2005, Ventspils 2006, Kuusamo 2007, Kuressaare 2008 and Kyiv 2009.

Many exchange visits have been made since 1992 by university teachers, research students and survey practitioners in Estonia, Latvia, Lithuania and Ukraine to the Universities of Helsinki, Stockholm and Umeå.

Many doctoral, master and bachelor theses on survey statistics have been written by Network participants.

Prospects of future developments.

Reference

Website of the *Baltic-Nordic-Ukrainian Network on Survey Statistics*: http://wiki.helsinki.fi/display/BNU/Home

Umeå 1994

Taevaskoja 1997

Ammarnäs 2002

Zhirovichi 2007

Dnipropetrovsk 2008

THE STATISTICAL SCIENCE IN UKRAINIAN UNIVERSITIES

Yuliya Mishura[1]

[1] Taras Shevchenko National University of Kyiv, Ukraine
e-mail: myus@univ.kiev.ua

Abstract

The direction Statistics oriented for students specializing in Mathematics is necessary and very important for the national economy and public institutions in Ukraine. Graduates of mathematical faculties specializing in Statistics are employed in different governmental and private institutions such as State Statistical Committee of Ukraine, National Bank, research organizations, insurance and investment companies, pension funds, marketing and advertising departments of large companies.

There exists a stable contingent of those willing to study subjects related to the direction Statistics in the National University of Kyiv and other universities in Ukraine: in Lviv, Uzhgorod, Donetsk, Odesa, Chernivtsi, Kherson, Dnipropetrovsk and Nizhyn. The number of those who want to obtain this speciality is the largest comparing to the other mathematical specialities.

The speciality Statistics was opened for the first time in Ukraine in National University of Kyiv in 1996. The methodological and financial support for its implementation and development has been received from grants of International Development Agency (USA), TACIS program of European Commission, Soros International Renascence Foundation and from five International grants within activity of European Commission. During the running of these projects friendly relations have been established with Stockholm University, Malardalen University, Helsinki University, Cologne University and Aegean University. Teachers of Mechanics and Mathematics Faculty of the University of Kyiv training specialists in Statistics have prepared many teaching materials: programs, tutorials, about 20 textbooks and books with collection of problems, etc. A course of Survey Sampling was introduced within the first Tempus grant "Statistical Aspects of Economics: Financial mathematics, Insurance and Sample Survey in Industry and Agriculture" in 1996.

In order to share the experience obtained by the University of Kyiv in this direction with other Ukrainian universities it was created a national network of universities and practitioners in economic-statistical area directed on dissemination and implementation of the best outcomes of TEMPUS TACIS JEP-10353-97 Statistical Applications of Economics at the national level.

Estimation for domains and small areas with design-based and model-based methods

Risto Lehtonen, University of Helsinki, Finland
e-mail: risto.lehtonen@helsinki.fi

Abstract

The series of lectures on the estimation for domains and small areas will cover to some extent the following topics:

1. Estimator type, model choice and auxiliary information
2. Design-based methods of estimation
   Horvitz-Thompson, GREG and model calibration estimators
3. Model-based methods of estimation
   Synthetic and EBLUP estimators
4. Composite estimators
5. Computation, Software
   SAS Macro EBLUPGREG
   Program Domest
6. Further examples

The treatment of the materials is practically oriented. The necessary statistical theory will be introduced. Applications will be presented, including the estimation of domain totals and means and their variances. The estimation of selected poverty indicators for domains will be discussed. Properties of the estimators are compared and discussed based on Monte Carlo simulation results. An option will be offered for university students to gain some credit points with a completed practical homework.

Materials

Rao J.N.K. (2003). *Small Area Estimation*. New York: John Wiley & Sons.

Lehtonen R. and Pahkinen E. (2004). *Practical Methods for Design and Analysis of Complex Surveys.* Second Edition. Chichester: John Wiley & Sons. Chapter 6 (copies of the chapter will be distributed to participants).

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.
Free access PDF: http://www.statcan.gc.ca/bsolc/olc-cel/olc-cel?catno=12-001-X20030016605&lang=eng

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.
Free access PDF issue: http://www.stat.gov.pl/cps/rde/xbcr/gus/POZ_Stat_in_Trans_8_3.pdf

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

Lehtonen R. and Veijanen A. (2009). Design-based methods of estimation for domains and small areas. Chapter 31 in Rao C.R. and Pfeffermann D. (Eds.). *Handbook of Statistics. Sample Surveys: Inference and Analysis. Vol. 29B.* New York: Elsevier. (Forthcoming). (Copies of the chapter will be distributed to participants).

Myrskylä M. (2007). Generalized regression estimation for domain class frequencies. Helsinki: Statistics Finland, *Research Reports* 247.
Free access PDF: https://oa.doria.fi/bitstream/handle/10024/11985/generali.pdf?sequence=1

Web materials (free access)

VLISS-virtual laboratory in survey sampling http://mathstat.helsinki.fi/VLISS/
EURAREA Project, http://www.statistics.gov.uk/eurarea/
EWORSAE – the European Working Group on Small Area Estimation http://sae.wzr.pl/

OVERVIEW OF THE STATE-OF-THE-ART OF SURVEY SAMPLING

Imbi Traat

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

Abstract

This series of four lectures gives an introduction to the basic issues in survey sampling. The design-based approach is in focus. The explanation of the very basic concepts is tied with the research interests of the lecturer.

The lectures will cover the following topics:

- Population, parameters, sample, sampling design, sampling procedure.
- Unbiased estimation and variance estimation, the effect of the second order inclusion probabilities.
- More on specific sampling designs and on estimation.
- Calibration; for increasing precision, for compensating non-response, for achieving consistency between estimates.

References

Bondesson, L., Traat, I. (2007), A nonsymmetric matrix with integer eigenvalues, *Linear Multilinear A*, **55**, 239-247.

Bondesson, L., Traat, I., Lundqvist, A. (2006), Pareto sampling versus Sampford and Conditional Poisson sampling. *Scand. J. Statist.*, **33**, 699-720.

Cochran, W.G. (1977), *Sampling Techniques*, 3rd ed., Wiley, New York.

Lohr, S. (1999), *Sampling: Design and Analysis*. Duxbury Press, Pacific Grove.

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

Särndal, C.E., Traat, I. (2009), Domain Estimators Calibrated on Information from Other Surveys. Submitted.

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

Traat, I., Ilves, M. (2007), The hypergeometric sampling design, theory and practice. *Acta Appl. Math.*, **97**, 311-321.

TEACHING THE SURVEY SAMPLING THEORY AND METHODOLOGY IN LITHUANIA

Danutė Krapavickaitė[1] and Aleksandras Plikusas[2]

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

[2] Institute of Mathematics and Informatics; Vilnius University, Lithuania
e-mail: plikusas@ktl.mii.lt

Abstract

The overview of courses on the sampling theory and methodology provided at Lithuanian universities is given here. Several teaching programs are presented and discussed.

1 Introduction

Statistical surveys by sampling methods started to be carried out with the restitution of independence. People get interested in the results of public opinion surveys in 1989. The methodology of the Household Budget Survey – the first sampling survey at Statistics Lithuania – originated in 1992. Statisticians of Statistics Sweden shared their knowledge in survey sampling giving a course for the employees of Statistics Lithuania in 1993 and little by little the sampling surveys started to be carried. The knowledge of survey sampling theory and methodology was necessary for statisticians, therefore courses on survey sampling started to be delivered at Vilnius University (VU) by Aleksandras Plikusas in 1995.

2 University courses

2.1 Teaching programs

The studies in this field of statistics are extended also now to other universities. The sampling courses are given at several universities by several university instructors:

Šiauliai University (ŠU) by Marijus Radavičius,

Vilnius Gediminas Technical University (VGTU) by Danutė Krapavickaitė,

Vytautas Magnus University (VMU) by Algimantas Bikelis,

Vilnius Pedagogical University (VPU) by Dalius Pumputis,

Vilnius University (VU) by Aleksandras Plikusas.

There are two main types of courses: the basic course and the advanced one. The scope of the courses in words and in the European Credit Transfer and Accumulation System (ECTS) is presented in Table 1.

Table 1. Courses on survey sampling at the universities of Lithuania

Kind of course	Course title	Place	Timing			Self-sustaining work	ECTS Credits
			Lectures	Practicals	Total		
Basic	Finite Population Statistics	VU	32	16	48	Practical work	4
Basic	Sampling Methods	VU	32	32	64	Practical work	4,5
Basic	Sampling Methods	VGTU	26	26	52	Course w., 3 contr. w.	4,5+1,5
Advanced	Statistical Surveys by Sampling Methods	VGTU	32	16	48	Course w., 2 contr. w.	4,5+1,5
Basic	Sampling methods	ŠU	32	0	32	1 contr. w.	3
Basic	Experiment design	VMU	45	15	60	1 colloquium	6
Basic	Basics of sample theory	VPU	36	24	60	2 contr. w.	4,5

2.2 Main topics included into the basic course

1. The object of survey sampling. The main concepts and definitions. The main paraneters of estimation, accuracy measures of estimators.

2. Simple random sampling. Sampling schemes, estimators of a total, mean, proportion in the population and domains. Determination of the sample size.

3. Sampling with replacement, estimators of a total. Normal approximation of the estimator distribution.

4. Unequal probability sampling with replacement.

5. Estimator of the ratio of two totals and ratio estimator of a total.

6. Stratified sampling design. Allocation of the sample size and estimators.

7. One-stage and two-stage cluster sampling. Systematic sample, analysis of variance.

8. Dealing with nonresponse.

9. Examples of the real surveys.

The course at VU for the students with a strong mathematical background is supplemented by the following topics:

Unequal probability sampling, Horvitz-Thompson estimator,

Bernoulli, Poisson sampling,

Regression and calibrated estimators,

Resampling methods for variance estimation in complex surveys.

For the students of VGTU with the a low mathematical background the basic topics are supplemented with the topics:

Repetition of the main facts of the probability theory,

Data quality,

Computer software for estimating in the case of complex sampling design.

The basic course is being delivered according to the textbook of Krapavickaitė and Plikusas (2005).
Students of VMU are studying the sampling theory independently and afterwards deliver it in the classroom one by one. Bernoulli sampling, results of Hájek (1981), normal as well as self-decomposable approximations of the distribution of the estimator in finite population sampling are also included into their course.

2.3 Main topics included into the advanced course

At VGTU, the advanced course is delivered for master students who have already attended the basic course. This course includes the following topics:

1. The object of survey sampling. The main concepts and definitions. The main parameters of estimation, accuracy measures of estimators.

2. Repetition. Sampling deigns and design-based estimators in the case of simple random sampling with and without replacement, unequal probability sampling with replacement, stratified sampling.

3. Representative sampling in the direct and generalized sense. Horvitz-Thompson estimator of a total.

4. Use of auxiliary information at the estimation stage: poststratification, a separate and common ratio estimator for stratified sampling, regression estimator.

5. Estimation of variance of complex estimators by the methods of

   Taylor linearization,

   random groups,

   jackknife,

   bootstrap.

6. Two-phase sampling. Its application to ratio estimation, stratification, and dealing with nonresponse.

7. Unequal probability sampling: systematic sampling with probability proportional to size, Poisson, Pareto sampling.

8. Estimation of the median and quantile.

9. Small area estimation. Synthetic and composite estimators.

10. Application of sampling methods in nature surveys:

    detectability and sampling,

    line transects,

    capture-recapture sampling.

The advanced course is delivered according to the textbook of Krapavickaitė and Plikusas (2005), other sources are also used. Some of the topics of this source are changed from time to time.

2.4 Practicals

Short numeric problems that follow the lecture topic are solved without using a computer. They corroborate the theoretical results, comparing the accuracy of estimators. They are especially useful to the students with a week mathematical background, because they show how the theoretical results should be applied. The sources of numerical problems are the textbook in Lithuanian (2005), Ardilly and Tillè (2006), and originally created problems.
Control works consist of short theoretical questions, definitions, proofs of the properties and solutions of the numerical problems.

2.5 Self-sustaining work

Practical work The course on survey sampling includes the practical problem to be solved by students at VU. The main idea of this practical problem is to simulate, to some extent, the work of a survey statistician. Approximately at the beginning of the last month of the semester, the data of some population are presented. The variables are described and divided in two categories: survey variables and auxiliary variables. Some possible topics of the practical work are also discussed. For example, "Regression estimator and its comparison with Horvitz-Thompson estimator" or "Stratified sampling, optimal allocation of the sample size". The students have a week to choose a topic for themselves. The topics must be different if the number of students is not larger than 12-15.
The students are encouraged to combine the practical problem with some theoretical study. This can be evaluated by some extra points. For example: "Inclusion probabilities for Poisson sampling when minimizing the variance of the ratio estimator". Sometimes some "not traditional" estimators are considered, such as a linear combination of the ratio and regression estimator.
The main part of practical work is simulation: one thousand samples should be selected from the given population and average characteristics (coefficient of variation, average estimate of the variance, average bias, etc.) must be calculated. These requirements are formulated in the stage of problem selection.
The students are free to choose software they need. Nevertheless, the R package is the main tool at the moment.
At the end of the semester, the students present their results to colleagues and to the teacher. The results are discussed and compared with that of other students. Because of that the same "survey" variable is considered and the same sample size is fixed for almost all the problems.
The weight of the practical problem varies between 20 and 40 percent.

Course work This work allows the students to acquire the primary practical experience in carrying out the sample survey.

The first source for this is the data and collection of the Survey exercises, presented in Lohr (1999) and translated into Lithuanian by Krapavickaitė (2002). The TV company provides for introduction of the cable TV in the Stephens County, and before starting the work, the company is carrying out the sample survey of the citizens in order to find out which programs should be translated and what payment from the citizens may be received. Students are asked to draw samples, to estimate the parameters and to find the strategy for the best accuracy of the estimators.
The Statvillage data of Schwarz (1997) are the real household data obtained in Canada during the population census. A description of the variables is translated into Lithuanian. The problems are being formulated for students in order they could feel the dependence of accuracy of the estimates on the sampling design, estimators, and the properties of the study variables.
The SAS and Excel software is used mainly at VGTU.
Students evaluate the course work as the main tool for understanding the survey sampling theory. They often use the course work examples when explaining the subject details often during the exam.
The topic of the course work has to be changed each year in order to avoid copying of the previous semester works.

Theses There are many students at all universities writing bachelor and master theses on survey sampling. Their topics are usually simple theoretical problems with the following simulation study. These problems are quite often connected with the problems of the real surveys at Statistics Lithuania. The real data are also often used. Sometimes the essence and topics of theses are similar to the practical work ones described above.
The VPU teacher, Dalius Pumputis, is the first statistician in Lithuania, who has defended his PhD thesis in Survey Sampling. Viktoras Chadyas at VGTU is by the end of his PhD studies.

3 Short courses for employees of Statistics Lithuania

The first textbook on survey sampling in Lithuanian has been written by Plikusas (1997) for statisticians of Statistics Lithuania. It includes a description of the main sampling designs, estimators and their variances. The theory described there seemed very complicated for the employees of Statistics Lithuania at that time. Many other courses in Europe have been attended by the statisticians since then, and it turned out that the results delivered there were even more complicated. People have got used to this field of statistics.
Two types of courses are sometimes given at Statistics Lithuania: the basic course for beginners and the advanced course for that who had already attended the basic

course depending on the needs of statisticians. In 2009, the advanced course has included the following topics:

determination of stratification boundaries,

dealing with nonresponse,

calibration,

estimation of the regression model parameters using SAS procedure `surveyreg`.

4 Experience of international teaching

Certain materials are prepared and some experience acquired in delivering the basic survey sampling course in Russian.

5 Conclusion

More practical work is needed for students. They don't like to study the pure theory. Therefore more attractive tools using computers and real data have to be developed with the view of successful studies.

References

Ardilly P., Tillè Y. (2006). *Sampling Methods, Excercises and Solutions.* Springer Science+Business Media, New York, 2006. 404 p.

Hájek, J. (1981). *Sampling from a Finite Populations.* New York and Basel: Marcel Dekker, inc. 247 p.

Krapavickaitė D. (2002) *Imčių metodai statistiniuose tyrimuose. Kompiuterinė programa APKLAUSA ir jos duomenis nagrinėjantys pratimai* (*SURVEY excercises, Lohr* (1999)). Technika, Vilnius. 53 p.

Krapavickaitė, D., Plikusas, A. (2005) *Imčių teorijos pagrindai* (*Basics of Sample Theory*). Technika, Vilnius. 312 p.

Lohr, S. L. (1999) *Sampling: Design and Analysis.* Duxbury Press, Pacific Grove. 512 p.

Schwarz C. J. (1997) StatVillage: An On-Line, WWW-Accessible, Hypothetical City Based on Real Data for Use in an Introductory Class in Survey Sampling. *Journal of Statistics Education*, **5**(2).
http://www/amstat.org/publications/jse/v5n2/schwarz.html

Plikusas, A. (1997) *Imčių metodai ir jų taikymai* (*Sampling Methods and their Applications*). Vilnius: Statistikos departamentas prie LRV. 51 p.

SURVEY SAMPLING METHODOLOGY IN OFFICIAL STATISTICS IN LATVIA

Mārtiņš Liberts[1] and Jānis Lapiņš[2]

[1] Central Statistical Bureau of Latvia
e-mail: Martins.Liberts@csb.gov.lv

[2] Bank of Latvia
e-mail: Janis.Lapins@bank.lv

Abstract

The paper briefly describes the development of survey sampling methodology in official statistics in Latvia. Methodology of household and business surveys will be covered separately. The paper will focus on such survey sampling methodology elements as sampling frame, sampling design, weighting and precision estimation. The paper will cover methodology used by Central Statistical Bureau of Latvia and Bank of Latvia.

1 Background

Central Statistical Bureau of Latvia (CSB) is the main producer of the official statistics in Latvia. There are several other producers of the official statistics. Not all of them are using survey sampling methodology in the production of statistics. The main producers of official statistics using survey sampling methodology are CSB, Bank of Latvia and State Employment Agency.

Survey sampling methodology is a relatively new field in Latvia. Survey sampling as it is known today in CSB was started in mid 90'ties. A probabilistic sampling in the production of the official statistics in Latvia for the first time was used in 1994. It was NORBALT I project organised by Fafo and funded by the Norwegian Ministry of Foreign Affairs, the Nordic Council of Ministers and the Norwegian Research Council. CSB was co-operating partner in the project. It was a household survey on living conditions.

The first sample surveys organised by CSB was Household Budget Survey (HBS) and Labour Force Survey (LFS). The surveys were lunched for the first time in 1995. Both surveys are household surveys.

2 Survey Typology

There are a lot of sample surveys organised by CSB. Lot of them are sharing similar methodology. By methodology we mean construction of sampling frame, sample selection, weighting and precision estimation techniques. Sharing of similar methodology gives several benefits:

- Cost efficiency. It takes less efforts to apply previously used methodology;
- Comparability over different surveys. It is easier to compare results produced by similar methodology;
- Higher quality of the results in some aspects. It is easier to maintain certain level of quality for survey if known methodology is used.

It is natural to classify all survey into three groups by methodology used:

- **Business surveys** cover all probabilistic surveys where the sampling unit is enterprise, organisation or farm. There is the Survey of Transportation of Goods by Road where a vehicle is a sampling unit. This survey is similar to business surveys;
- **Household and person surveys** cover all probabilistic surveys where private address, private household or person is a sampling unit;
- **Other surveys** are non-probabilistic surveys like survey of travellers on borders, survey of prices of products and services and other surveys.

3 Business Surveys

3.1 Statistical Registers

CSB has created several statistical registers. Statistical registers are used to build sampling frames, produce statistics and to support other statistical work. There are lot of auxiliary data sources for statistical registers. Most of them are different state registers and administrative data sources. Auxiliary data are combined with survey data and stored in statistical registers.

Statistical Register of Enterprises and Organisations – this is one of the oldest statistical registers in CSB. The sources for Statistical Register of Enterprises and Organisations are:

- The Register of Enterprises of Latvia;
- State Revenue Service;
- Statistical surveys;
- Other sources.

The Statistical Register of Farms was created before the Farm Census 2001. The sources of the Statistical Register of Farms are:

- The Land Register;
- The Population Register;
- Statistical Register of Enterprises and Organisations;
- Other sources.

The Register of Vehicles is a state register used to build a sampling frame for the Survey of Transportation of Goods by Road.

The development of the statistical registers in CSB has been done step by step. It has taken quite a long time to combine so many data sources in statistical registers. There are several obstacles to overcome until a data source can be used in statistical registers:

- Administrative obstacles. The process of developing cooperation between state institutions about data sharing is quite difficult. Often it requires changes in administrative documentation about new tasks. It requires extra resources for data sharing in some cases;
- Data quality. Often data quality in not good enough for usage in statistics:
  - Precision (the Population Register – the information about living place of person is not always reliable. It is because there is not motivation for persons to declare the living place in the Population Register. Even more – there will be a new rule that person has to pay tax for the declaration of the living place);
  - Timeliness is quite often problem for administrative information;
  - Coverage;
  - Other quality aspects.

3.2 Sampling Frames

The first step in building a sampling frame is to define sampling unit. Sampling unit could be defined by type, economic activity, size, location, activity period and other parameters.

There could be several types of units in business surveys:
- Enterprise;
- Local unit[1];
- Kind-of-activity unit (KAU)[2];
- Local kind-of-activity unit (local KAU)[3];
- Farm.

There is quite a lot of auxiliary information usually available in sampling frames for business surveys. The auxiliary information that could be available in sampling frame:
- The main economical activity coded by NACE classification[4];
- The location of enterprise coded by national territory classification;
- Number of employees;
- The amount of transaction imposed by value added tax.

The main difficulty for sampler is to choose the information to be used in sampling designing according to the targets of the survey. We have to remember that there is not 100% correct information available. Evaluation of available information is very important. The imperfections of sampling frames have to be noted and documented as much as possible. It is because the quality of the sampling frame will influence the overall quality of the survey.

We will demonstrate two typical situations in case of sampling frames for business surveys:
- The information units (variables and values) in registers are not updated with the same frequency. There are units with updated information and units with outdated information. We can split all units from a register in two parts by the definition of the target population. We can build our sampling frame as a set of units corresponding to the target population. We have to remember that sampling frame will not precisely cover the target population. There are over-coverage errors and under-cover errors. Sometimes the way how to deal with under-cover errors is to select a small sample from units that do not correspond to the definition of the target population;
- The parameters of the sampling units in business surveys are rapidly changing. So it is very important to keep the copy of sampling frame used for sampling. Otherwise it will be not possible to “recreate” it afterwards.

---

[1] The local unit is an enterprise or part thereof (e.g. a workshop, factory, warehouse, office, mine or depot) situated in a geographically identified place. At or from this place economic activity is carried out for which – save for certain exceptions – one or more persons work (even if only part-time) for one and the same enterprise.

[2] The kind of activity unit (KAU) groups all the parts of an enterprise contributing to the performance of an activity at class level (4-digits) of NACE Rev. 1 and corresponds to one or more operational subdivisions of the enterprise. The enterprise's information system must be capable of indicating or calculating for each KAU at least the production value, intermediate consumption, manpower costs, the operating surplus and employment and gross fixed capital formation.

[3] The local kind-of activity unit (local KAU) is the part of a KAU that corresponds to a local unit.

[4] Statistical Classification of Economic Activities in the European Community.

3.3 Sampling Design

Usually stratified simple random sampling is used for business surveys. Very rarely other designs are used. A motivation for stratified sampling in business surveys is:

- Populations of enterprises or farms are very skewed by different indicators. There are relatively low number of big enterprises and high number of small enterprises in the population of study, where the size of the units can be defined by different indicators;
- The main results of business surveys are required to be split by many domains (small areas). Stratification by these domains is a way to control the precision for the domain estimates.

In case of stratified sampling there are two main tasks to do for a statistician:

- Designing stratification:
  - The number of strata;
  - Definition of strata;
- Sample size allocation over strata.

The basic concept of stratification is to split units in homogenous groups by size variables and to improve precision of estimates for main domains of interest. There could be different size variables for different surveys. Often used size variables are total turnover in a period (year, quarter, month) and number of employees (average in period or at fixed time point). Domains usually are defined by economic activity, number of employees, location. The number of strata usually is high for business surveys.

Several sample allocations can be used. Traditionally Neyman sample allocation is used to minimise variance for total of one variable. All theoretical sample allocations have to be adjusted to meet practical needs. For example, Neyman allocation can give strata sample size equal to 0 or higher then strata population size.

3.4 Weighting and Precision Estimation

Usually number-raised estimation[5] is used for business surveys. The weights are computed for each stratum independently. Calibration technique has been used in some cases. The weights from number-raised estimator are calibrated to known population totals.

One of the problems in weighting is rapidly changing population in case of business surveys. It is quite typical that the information about the target population is very different during sampling and during weighting. It is good practice to take these differences into account to get more precise estimates of the population.

In case of number-raised estimator classical variance estimator for stratified simple random sampling is used.

[5] The application of weights to the individual survey records. Number-raised weights are given by N/n (where N is the total number of units in the population for the stratum, and n is the number of responding units in the sample for that stratum). The weight assigned to each survey unit indicates the number of units in the target population that the survey unit is meant to represent.

4 Social Surveys

4.1 Sampling Frames

There are several administrative registers with information available for statistics:
- The Population Register;
- The Address Register;
- The Land Register;
- Other registers.

The information from administrative registers and fieldwork operations are gathered in statistical registers. Mainly there is one statistical register used for building sampling frames of social surveys. It is the statistical register of households and dwellings.

There could be different ultimate sampling units in social surveys – households and persons. There are several definitions of household:
- Population census definition – household is defined by dwelling;
- Household is a single person or group of persons sharing a budget in common.

The latter definition is usually used in social surveys. It is not always easy to define households in the statistical register. So information in the statistical register does not always match with real information.

The information available in the sampling frames for households:
- Address;
- The population census counting area;
- The list of persons living in a household. For each person the following information is available:
  - Name of person;
  - Sex;
  - Date of birth, age;
  - Other information.

4.2 Sampling Design

Usually multi-stage sampling is used for social surveys. The main reason for multi-stage sampling is the minimisation of survey costs. Traditionally social surveys have been organised as face-to-face interviews. Recently telephone interviews have been started. There are surveys done partly by face-to-face interviews and partly by telephone interviews. So there is always necessity for face-to-face interviewers for all social surveys. Doing surveys with multi-stage design allows lowering the cost of surveys.

The population census counting areas are sampling units at the first stage. Census counting areas are small geographic areas defined for all territory of Latvia. The purpose of the census counting areas was the population census organised in 2000. There are several benefits for using these areas for sampling:
- Areas are similar by size where the size of area is defined by the number of households or persons living in area;
- The areas were planned for the purpose of the population census interviewers. They are optimised to cover geographically compact areas – so the distances inside areas are not long – they are feasible for interviewers;
- Areas are sorted in serpentine order covering all territory of Latvia. It provides implicit stratification if systematic sampling is used;
- Areas are stratified by the degree of urbanisation. Usually there are four strata defined:

    - The capital city – Riga (approximately 1/3 of the population);
    - Six republic cities – six biggest cities after the capital city (approximately 1/6 of the population);
    - Other cities and towns (approximately 1/6 of the population);
    - Rural areas (approximately 1/3 of the population).
- The information about areas can be updated by the information from the statistical register of households and dwellings.

Stratified systematic sampling is used to sample areas. It differs if survey is one-off or continuous survey. Classical systematic sampling with probabilities proportional to size can be used for one-off surveys. Different design is used for continuous surveys. The following targets are set to achieve for the design of continuous surveys:

- Probabilistic sample;
- Rotation of areas according to the selected rotation scheme;
- Uniformly distributed areas over space;
- Uniformly distributed areas over time;
- Easy management of areas in sample;
- Coordination of different surveys;
- Variance estimation.

These targets are achieved with stratified double systematic sampling with probabilities proportional to size. The design is used for the core sample of areas. The core sample is used for several continuous surveys (LFS, HBS, The Survey of Domestic Travellers (SDT)). The core sample can be used also for other one-off surveys.

Double systematic sampling comes from two sampling steps used in scheme. There is a brief description of the sampling scheme:

- Assumptions:
    - This is a continuous survey were weekly samples are selected;
    - A is a number of areas to be selected in a weekly sample;
    - Rotation scheme – units are sampled m times repeatedly with time interval T weeks;
- The first area is selected randomly by probability proportional to the size of area;
- The first sampling step is computed and $A-1$ extra areas are sampled by consecutively adding the sampling step. The step has been selected so that areas are almost uniformly distributed over whole population. A small derivation from the uniform distribution allows keeping selected areas in the sample m times and moving to new areas afterwards. At the end we have A areas selected. This is a sample for one week;
- The second sampling step is computed. The second step is added to all A previously selected areas. So we have selected extra A areas. This is a sample for the second week. Repeating the process each time we select A areas for the next weekly sample. The second sampling step is computed so that $A(1-1/m)$ areas selected for the week $T+1$ are matching the same $A(1-1/m)$ areas selected for the first week. A/m areas will be rotated out of sample and replaced by new areas;
- The second sampling step is used continuously to select areas for each week;
- Similar sampling scheme can be applied for other rotation schemes and survey designs.

Simple random sampling of households or dwellings is used at the second stage. There could be the third stage of sampling if persons are ultimate sampling units. Usually all persons living in selected household are sampled.

At the end we have achieved self-weighting sampling design.

4.3 Weighting

Different weight can be computed for social surveys:
- Weights for individuals and households;
- Cross-sectional and longitudinal weights;
- Other specific weights.

The design weights are computed according to the selection probabilities. The correction of non-response is applied splitting sample in response homogeneity groups. Calibration is applied to get final weights. Historically post stratification was used instead of calibration.

4.4 Precision Estimation

Because of sampling design and complexity of statistics computed in social surveys approximate variance estimators are used. There are several techniques how variance estimation can be done:
- Dependent random group method can be applied. The design of the survey is very feasible to divide the sample in dependent random groups;
- Jackknife method can be applied in similar way as dependent random group method by dropping out one of sub-sample. Jackknife delete one can be applied dropping out one primary sampling unit (PSU) to compute each Jackknife estimate;
- Linearization technique can be applied for complex statistics.

5 The Coordination of Samples

Historically positive coordination was used in business surveys. Usually the yearly and quarterly samples were designed as core samples. Different subsets of these core samples were used as samples for different surveys. Often these subsets were overlapping. The reasons for this coordination were:
- Easy management of respondents over all surveys;
- Better comparability of estimates over different surveys;
- Micro data sharing or transfer between surveys.

Quarterly and yearly core samples were also positively coordinated. Of course it resulted with high burden for respondents who were sampled in the core samples.

Sampling for business surveys were reorganised during 2006-2007. Independent sampling was used for each survey. The targets for reorganisation were:
- To reduce the burden of respondents – to share a burden over more enterprises;
- To optimise the designs for each survey. The targets of surveys differ – for example main indicators to estimate, domains of interest. Target populations can differ also.

Both positive and negative coordination is used for social surveys. There is a positive coordination over time because several social surveys (LFS, Survey on Income and Living Conditions) are panel surveys where respondents have to be contacted repeatedly. Another aspect of positive coordination is between different surveys. Often surveys are sharing the same primary sampling units – areas. It is done to reduce the cost and time of interviewing. For example LFS, HBS and SDT are coordinated in such way. There are areas where interviewers are doing LFS and HBS and there are areas where interviewers are doing LFS and SDT.

Negative coordination in social surveys is done in all stages of sampling. There is a negative coordination at the PSU level. One area cannot be used for a long time, because the number of households is not very high in areas. So the areas used for surveys are excluded from sample

for some time – this is achieved by the sampling design used. Of course there could be some independent samples that are not coordinated with core sample.

There is a negative coordination at the level of households. There is a rule that a household who has taken part in any social survey by CSB has to be left out of social surveys for at least two years (after the last time of interviewing). This is managed by the listing of households used for surveys. It is not always possible to achieve this rule because households are units changing over time and there is not unique identifier for a household.

6 Other Aspects

Other aspects like dealing with non-response and imputation will be discussed during the presentation. There will be also the presentation about survey sampling methodology in Bank of Latvia.

References

Fafo (2001) NORBALT I and II, Living Conditions in the Baltic Countries, http://www.fafo.no/norbalt/.

Lapins, J. (1997) Sampling Surveys in Latvia: Current Situation, Problems and Future Development. *Statistics in Transition*, Vol. **3**, No. 2, 281-292.

Lapins, J., Vaskis, E., Priede, Z. and Balina, S. (2002) Household Surveys in Latvia. *Statistics in Transition*, Vol. **5**, No. 4, 617-641.

OECD, Glossary of Statistical Terms, http://stats.oecd.org/glossary/.

BAYESIAN METHODS IN SURVEY SAMPLING

Daniel Thorburn[1]

[1] Stockholm University, Sweden
e-mail: daniel.thorburn@stat.su.se

Abstract

In this talk I will sketch a Bayesian approach to survey sampling. There are different variants of Bayesians. This will be a quite orthodox presentation. Due to the short time it will also be a superficial treatment.

For a true Bayesian the whole analysis should be completely Bayesian. Thus I will start by describing the Bayesian philosophy and its consequences for survey sampling, estimation and reporting. I will then give a full Bayesian treatment in a few simple standard situations, Dichotomous data under simple and stratified sampling. The Normal-gamma and the Dirichlet process approaches to continuous data. I will also mention the use of auxiliary information. Many papers on Bayesian methods in survey sampling are not quite orthodox and suggest that Bayesian methods should be used only for those purposes, when the standard design-based approach does not work, like handling of non-response and small area estimation. I will illustrate some of these cases too. A true Bayesian, on the other hand, says that Bayesian methods should only be used when they are better than classical methods, i.e. always.

The following plan is only a preliminary sketch. I will probably have to omit some parts of it due to time restrictions

1 Bayesian philosophy I

With a special view on Survey Sampling and Official Statistics.
Subjective probability and multiple users.
Updating, Bayes'formula.
Predictive inference.
Decisions, the of the statistics.

2 Standard situations.

2.1 Polya urn scheme for dichotomous variables

Exchangeability, de Finetti's theorem.
Do parameters exist?

2.2 Stratified sampling for dichotomous variables

Effects of too informative priors.

2.3 Dirichlet-multinomial and the Dirichlet process

Discrete and continuous variables.
Bayesian non-parametrics.

2.4 The Linear and Normal-gamma approaches to sampling for continuous variables

Superpopulations.
Conjugate priors.
BLUE and restriction to linearity.

2.5 Auxiliary variables

Ratio and regression type situations.

3. Bayesian philosophy II

3.1 The role of randomisation

Why do we have to randomise?
In a model based approach the design should not matter. But it does!

3.2 Preposterior analysis

How to chose the sample size and the design af a survey.
Balanced designs.

4. Special situations, where the Bayesian approach can be seen as a complement.

4.1 Missing data and non-response

Rubin's classification, MAR, MCAR, NMAR

4.2 Multiple Imputation

Obtaining the full posterior through Gibbs' sampling and other MCMC-methods

4.3 Small area estimation

The use and estimation of covariance structures

4.4 Outlier detection

How to take the possibility of outliers into account, when there are no outliers in your sample?

4.5 Editing

Estimating the probability and impact of response errors. Effects of data checking

5. Conclusion

This is a very superficial talk. It is impossible to give an elegant and thorough treatment of survey sampling in one hour. But it is easier to do so with the Bayesian approach than in an old-fashioned design-based setting.

TEACHING SURVEY SAMPLING THEORY AND METHODOLOGY – WITH A UKRAINIAN PERSPECTIVE

Olga Vasylyk[1] and Tetyana Yakovenko[2]

[1] Taras Shevchenko National University of Kyiv, Ukraine
e-mail: ovasylyk@univ.kiev.ua

[2] Taras Shevchenko National University of Kyiv, Ukraine
e-mail: yata452@univ.kiev.ua

Abstract

An overview of teaching programs for the course on Survey Sampling Theory and Methodology at some Ukrainian universities will be presented.

The timing of this course varies between 14 and 36 hours of lectures and between 8 and 36 hours of practical lessons at different universities. A typical teaching program is based on the books by O.Chernyak and V.Parkhomenko and includes the following topics:

1. General scheme of a survey, main concepts and definitions
2. Simple random sampling with and without replacement
3. Sampling with unequal probability
4. Systematic sampling
5. Stratified sampling
6. Cluster sampling
7. Multistage sampling
8. Linear regression models
9. Variance estimation
10. Errors in surveys, their sources and methods of reduction

Possible ways for improvement of teaching of survey sampling theory and methodology will be suggested.

Also current situation with employment of students specialized in Statistics, in particular in Survey Sampling, will be discussed.

References

Chernyak, O. (2001) *Survey Sampling Technique.* Kyiv (in Ukrainian).

Parkhomenko, V. (2001) *Survey Sampling Methods.* Kyiv (in Ukrainian).

SAMPLING IN VITAL STATISTICS

Nastassia Bobrova [1]

1 Economic Institute of National Academy of Science, Minsk, Belarus
e-mail: nastassiabobrova@mail.ru

Abstract

The methodology of sample survey of household is given. The number of demographic sampling in Belarus is provided..

1 Introduction

Sampling in vital statistics can be regular and irregular. Surveyed object can be household or certain population group (a person) for special purposes.

The sample survey of households (Income and Expenditures of Households Survey) is an important part of the regular operations of the National Statistical Committee of the Republic of Belarus. It represents the major source of information on the socio-economic status of the Belorussian households. The results of the survey are widely used in the various economic estimations made by the Committee (for instance, in the calculation of the Gross Domestic Product (GDP) and its distribution, in the computation of the consumer price index, in the System of National Accounts and the balance of agricultural production, etc.).

An Act of the Cabinet of Ministers of the Republic of Belarus #259 from the 23rd December, 1994 ("About the conduction of the sample survey of households") provides the legal basis for the carrying out this survey. The Act assumes confidentiality of received information as well as a compensation given to the households for their participation in the survey (10% of the monthly minimum salary).

The elaboration and implementation of the survey were conducted under the supervision of the World Bank. Advisers from the Bank consulted and coordinated the staff of the household department on the issues of the overall design of the survey, methodological aspects of the sampling, the content of questionnaires and data processing.

The primary aims of the survey are:

- To obtain reliable information for the evaluation of the quality of life of population;
- To provide the Belorussian Government and other interested organizations with reliable information;
- To create a computerized system for the collection and processing of the survey results.

The survey is conducted annually since January 1995. Each year a sample of about 6000 households is selected. Every household has an equal probability to be selected, and this allows extending the obtained results over all resident households. The survey covers all types of households with the exclusion of the institutionalized population (persons living in nursing homes, prisons, convents, military barracks, etc.).

The household survey is a continuous one and assumes participation of a household on the voluntary basis. Selected households are free to refuse their participation in the survey. Besides, a household may leave the survey due to death of its member, a change of address and so on; no household might be replaced by a new one..

2 Contents and structure of the survey

The main principles of the survey are:

to collect detailed income and expenditure data for a representative sample of households;

to use a relatively short recall period (due to the unstable economic situation and particularly, high inflation households might not be able to recall the exact amounts of income or expenses).

The survey is restricted to one calendar year and basically designed as a sequence of four quarterly interviews that cover an entire year for the same sample of households.

Main components of the survey are:

1. The baseline interview. It is designed to establish a first contact with a household, to collect the information on a household and its members' characteristics, and make an appointment for the next visits. Prior to the interview, each household receives an official letter from the Committee with a tentative date and time for the interviewer to visit the household.

2. Four quarterly interviews which are conducted in April, July, October of the current year and January of the following year. Each interview covers the household expenses and income for the previous three months. At the beginning of each quarter, a household receives so-called a "memory aid" used to record the major expenses made during the quarter. The memory aid is used during the subsequent quarterly interview.

3. Four two-week diaries which a household receives every quarter. Each diary is used for the recording of the every day expenses on food and other non-food items made by a household during 14 days.

The following principles were considered during the development of questionnaires:

The content of the questionnaires should correspond to the general structure of the survey;

The content of the questionnaires should be aimed at the collection of data necessary for the development of the social policies;

The questionnaires should be simple and understandable from the one hand but detailed and concrete from the other.

Regarding the occasional sample surveys in demographic statistics, there is a number of studies on population migration. One of the examples is the project "Analysis of the youth labour migration 2005-2006", financed by the Belorussian State Fund of the Fundamental Research.

As a database for the analysis the legal documents on the problem of youth and migration in Belarus, statistical reports and the results of sample surveys were used.

The following types of questionnaire were used in the project:

A questionnaire of a potential migrant. Interviews were conducted at working place or place of study.

An individual questionnaire of a labour migrant. In addition to the general questions on migration, the questionnaire consisted of three separated parts:

A set of questions for those working (hired) on the constant basis;

A set of questions for those engaged in reconstruction work, construction of a house or summer residence;

A set of questions for those engaged in trade.

The selection was conducted on the "snowball" basis.

3. The questionnaire of the labour out-migrant (youth summer labour migration). The interviews were conducted by the companies organizing this migration.

4. The questionnaire of the labour emigrant. The interviews were conducted by the companies organizing this migration.

5. In addition to that, there is an on-line questionnaire for those who left the country for work and have not returned on time.

In order to study the specifics of non-return youth migration, the joint group of researchers conducted a survey of the young people who moved to the USA for the summer vacation and did not returned to Belarus after the expiration of the issued visa. The survey was conducted within the framework of international project "Migration: theory, methods and the practice of the regulation of migration". The method of "snowball" was chosen in the sample design.

In addition, the Scientific Institute of the Ministry of Labor of the Republic of Belarus conducted a study in 1998 on the potential of external labor migration, migration flows and 'push' and 'pull' factors.

The special sampling of border area was carried out by professor Shakhotska within the bound of International CIS project in 2002.

The author is researching special population group (migrants). Statistic registration doesn't provide reliability of migration data. So now it is necessary to carry out sample survey of migrants because of their benefit. For example, the author's sampling of the primary data sources on migrants at Statistics Department in Minsk was allowed to reveal a lot of problems with registration of migrants. As a result of this problems the database mobility of migrants isn't qualitative. Nowadays only sampling can help to correct the number of migrant and to research the subjective data. The main purpose of our demographers today is sampling promotion to vital statistics.

Dual to ratio-cum-product estimator for general sample design and some simulation results

Ignas Bartkus[1]

[1] Vilnius Pedagogical University, Lithuania
e-mail: ignas.bartkus@gmail.com

Abstract

Ratio-cum-product and dual to ratio-cum-product estimators of the finite population total are presented for general sample design. An empirical study is carried out to observe the performance of these estimators in the case of stratified simple random sample design.

1 Introduction

The ratio method for estimation of population parameters are generally used when auxiliary information about study variable is available. The ratio estimators can be effective when auxiliary variable is positively correlated with study variable. The product estimators behaves similarly as ratio, however, they can be used in case negatively correlated auxiliary variable is available. In order to improve the efficiency of estimation, combination of ratio and product estimators can be used. Some of the important work in this direction are made. The ratio-cum-product estimator is presented in Singh (1969), dual variables for estimation of population parameters were introduced by Bandyopadhyay (1980) and Srivenkataramana (1980), dual to ratio-cum-product estimator is presented in Singh et all (2005). In all these papers estimators are defined for simple random sample case.
In this paper ratio-cum-product and dual of ratio-cum-product estimators are examined for arbitrary sample design. The dual variables for the construction of this estimator are defined in Plikusas (2008).

2 Ratio-cum-product estimator

2.1 General sample design case

Consider a finite population $\mathcal{U} = (u_1, u_2, ..., u_N)$ of N units. Let a sample s be drawn from this finite population and π_k determines the inclusion probability of kth element. Let y_k and (x_k, z_k) represent the values of a response variable y and two auxiliary variables (x, z) respectively available on the kth unit $(k = 1, 2, \ldots, N)$. If no auxiliary information is available, then Horvitz-Thompson estimator

$$\hat{t}_\pi = \sum_{k \in s} \frac{y_k}{\pi_k}$$

is a very good estimator of the population total $t = \sum_{k=1}^{N} y_k$. However if an auxiliary variable x (z), that is positively (negatively) correlated with a response variable y, is available, then the ratio (product) estimators performs better than the Horvitz-Thompson estimator.

The ratio and product estimators are defined as

$$\hat{t}_{R\pi} = \frac{\hat{t}_{y\pi}}{\hat{t}_{x\pi}} t_x, \quad \hat{t}_{P\pi} = \frac{\hat{t}_{y\pi}\hat{t}_{z\pi}}{t_z},$$

where

$$\hat{t}_{x\pi} = \sum_{k\in s} \frac{x_k}{\pi_k}, \quad \hat{t}_{z\pi} = \sum_{k\in s} \frac{z_k}{\pi_k}.$$

If both a positively and negatively correlated auxiliary variables are available, then a combination of the ratio and product estimators can be used for further improve of estimation. Singh [3] suggested the following combination, ratio-cum-product estimator

$$\hat{t}_{RP\pi} = \hat{t}_{y\pi} \frac{t_x}{\hat{t}_{x\pi}} \frac{\hat{t}_{z\pi}}{t_z}.$$

The approximate variances of the estimators $\hat{t}_{R\pi}$, $\hat{t}_{P\pi}$ and $\hat{t}_{RP\pi}$ are

$$AVar(\hat{t}_{R\pi}) = \sum_{k,l\in\mathcal{U}} (\pi_{kl} - \pi_k\pi_l) \frac{y_k - R_x x_k}{\pi_k} \frac{y_l - R_x x_l}{\pi_l},$$

$$AVar(\hat{t}_{P\pi}) = \sum_{k,l\in\mathcal{U}} (\pi_{kl} - \pi_k\pi_l) \frac{y_k + R_z z_k}{\pi_k} \frac{y_l + R_z z_l}{\pi_l},$$

$$AVar(\hat{t}_{RP\pi}) = \sum_{k,l\in\mathcal{U}} (\pi_{kl} - \pi_k\pi_l) \frac{y_k - R_x x_k + R_z z_k}{\pi_k} \frac{y_l - R_x x_l + R_z z_l}{\pi_l},$$

where

$$R_x = \frac{t_y}{t_x}, \quad R_z = \frac{t_y}{t_z}.$$

π_{kl} - the probability that both of the elements k and l will be included in a sample s.

2.2 Stratified simple random sample case

Assume the population $\mathcal{U}$ consists of H strata: $\mathcal{U} = \mathcal{U}_1 \cup \ldots \cup \mathcal{U}_H$. The size of stratum $\mathcal{U}_h$ is N_h, and the size of simple random sample s_h in stratum $\mathcal{U}_h$ is n_h, $h = 1, \ldots, H$. The ratio and product estimators are

$$\hat{t}_{Rst} = \frac{\hat{t}_{yxt}}{\hat{t}_{xst}} t_x, \quad \hat{t}_{Pst} = \frac{\hat{t}_{yst}\,\hat{t}_{zst}}{t_z},$$

where

$$\hat{t}_{yst} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{k\in s_h} y_k, \quad \hat{t}_{xst} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{k\in s_h} x_k, \quad \hat{t}_{zst} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{k\in s_h} z_k,$$

The ratio-cum-product estimator is defined as

$$\hat{t}_{RPst} = \hat{t}_{yst} \frac{\hat{t}_{zst}}{\hat{t}_{xst}} \frac{t_x}{t_z}.$$

And the approximate variances of these estimators are

$$AVar(\hat{t}_{Rst}) = \sum_{h=1}^{H} N_h^2 \frac{1 - f_h}{n_h} \bar{Y}^2 \big(C_{yh}^2 + C_{xh}^2(1 - 2K_{yxh})\big),$$

$$AVar(\hat{t}_{Pst}) = \sum_{h=1}^{H} N_h^2 \frac{1 - f_h}{n_h} \bar{Y}^2 \big(C_{yh}^2 + C_{zh}^2(1 + 2K_{zxh})\big),$$

$$AVar(\hat{t}_{RPst}) = \sum_{h=1}^{H} N_h^2 \frac{1-f_h}{n_h} \bar{Y}^2 \big(C_{yh}^2 + C_{zh}^2(1+2K_{yzh}) + C_{xh}^2(1-2K_{yxh}-2K_{zxh}) \big),$$

where

$$f = \frac{n_h}{N_h}, \quad K_{yxh} = \rho_{yxh} C_{yh}/C_{xh}, \quad K_{zxh} = \rho_{zxh} C_{zh}/C_{xh}, \quad K_{yzh} = \rho_{yzh} C_{yh}/C_{zh},$$

$$C_{yh} = \frac{s_{yh}}{\bar{Y}}, \quad s_{yh}^2 = \frac{1}{N_h - 1} \sum_{k \in U_h} (y_k - \bar{Y}_h)^2, \quad \bar{Y}_h = \frac{1}{N_h} \sum_{k \in U_h} y_k, \quad \bar{Y} = \frac{1}{N} \sum_{k=1}^{N} y_k,$$

$$\rho_{xyh} = \frac{s_{xyh}}{s_{xh} s_{yh}}, \quad s_{xyh} = \frac{1}{N_h - 1} \sum_{k \in U_h} (y_k - \bar{Y}_h)(x_k - \bar{X}_h).$$

3 Dual to ratio-cum-product estimator

3.1 General sample design case

The dual variable for the general unequal probability sampling design with the inclusion probability π_k of the element k was introduced by Plikusas in 2008.

Denote $g_k = \pi_k/(1-\pi_k)$ for $k \in s$, and define transformations of the auxiliary variables x and z:

$$x_k^* = \left(\sum_{k \in s} \frac{1}{\pi_k} \right)^{-1} \sum_{k=1}^{N} (1+g_k) x_k - g_k x_k,$$

$$z_k^* = \left(\sum_{k \in s} \frac{1}{\pi_k} \right)^{-1} \sum_{k=1}^{N} (1+g_k) z_k - g_k z_k.$$

Using the notation

$$\hat{t}_{x\pi}^* = \sum_{k \in s} \frac{x_k^*}{\pi_k}, \quad \hat{t}_{z\pi}^* = \sum_{k \in s} \frac{z_k^*}{\pi_k},$$

the dual to ratio-cum-product estimator can be defined as

$$\hat{t}_{RP\pi}^* = \hat{t}_y \frac{\hat{t}_{x\pi}^*}{t_x} \frac{t_z}{\hat{t}_{z\pi}^*}.$$

In previous formula, if auxiliary variable z is not used then the dual to ratio-cum product estimator reduces to dual to ratio estimator

$$\hat{t}_{R\pi}^* = \hat{t}_y \frac{\hat{t}_{x\pi}^*}{t_x}$$

and when x is not used then it reduces to dual to product estimator

$$\hat{t}_{P\pi}^* = \hat{t}_y \frac{t_z}{\hat{t}_{z\pi}^*}.$$

The approximate variances of the estimators $\hat{t}_{RP\pi}^*$, $\hat{t}_{R\pi}^*$ and $\hat{t}_{P\pi}^*$ are

$$AVar(\hat{t}_{RP\pi}^*) = \sum_{k,l \in U} (\pi_{kl} - \pi_k \pi_l) \frac{y_k - R_x g_k x_k + R_z g_k z_k}{\pi_k} \frac{y_l - R_x g_l x_l + R_z g_l z_l}{\pi_l},$$

$$AVar(\hat{t}_{R\pi}^*) = \sum_{k,l \in U} (\pi_{kl} - \pi_k \pi_l) \frac{y_k - R_x g_k x_k}{\pi_k} \frac{y_l - R_x g_l x_l}{\pi_l},$$

$$AVar(\hat{t}_{P\pi}^*) = \sum_{k,l \in U} (\pi_{kl} - \pi_k \pi_l) \frac{y_k + R_z g_k z_k}{\pi_k} \frac{y_l + R_z g_l z_l}{\pi_l}.$$

Estimators of the approximate variances are

$$\widehat{Var}\hat{t}^*_{RP\pi} = \sum_{k,l\in s} \frac{\pi_{kl}-\pi_k\pi_l}{\pi_{kl}} \frac{y_k - \hat{R}_x g_k x_k + \hat{R}_z g_k z_k}{\pi_k} \frac{y_l - \hat{R}_x g_l x_l + \hat{R}_z g_l z_l}{\pi_l},$$

$$\widehat{Var}\hat{t}^*_{R\pi} = \sum_{k,l\in s} \frac{\pi_{kl}-\pi_k\pi_l}{\pi_{kl}} \frac{y_k - \hat{R}_x g_k x_k}{\pi_k} \frac{y_l - \hat{R}_x g_l x_l}{\pi_l},$$

$$\widehat{Var}\hat{t}^*_{P\pi} = \sum_{k,l\in s} \frac{\pi_{kl}-\pi_k\pi_l}{\pi_{kl}} \frac{y_k + \hat{R}_z g_k z_k}{\pi_k} \frac{y_l + \hat{R}_z g_l z_l}{\pi_l}.$$

3.2 Stratified simple random sample case

Consider stratified simple random sampling of size n_h in stratum $\mathcal{U}_h$, denote $g_h = n_h/(N_h - n_h)$ for $h = 1, \ldots, H$. The dual transformation for stratified and arbitrary sampling design is defined in Plikusas (2008).

Here we use the direct generalization of dual transformation, and define transformation of the auxiliary variable x:

$$x^*_k = (1+g_h)\bar{X}_h - g_h x_k, \quad \text{for} \quad k \in \mathcal{U}_h,$$

where $\bar{X}_h = \frac{1}{N_h}\sum_{k\in\mathcal{U}_h} x_k$. The transformation for the variable z are defined analogously. Note that $\sum_{k=1}^N x^*_k = \sum_{k=1}^N x_k = t_x$. The dual to ratio and dual to product estimators are

$$\hat{t}^*_{Rst} = \hat{t}_y \frac{\hat{t}^*_{xst}}{t_x}, \quad \hat{t}^*_{Pst} = \hat{t}_y \frac{t_z}{\hat{t}^*_{zst}}.$$

and the dual to ratio-cum-product estimator is defined as

$$\hat{t}^*_{RPst} = \hat{t}_{yst} \frac{\hat{t}^*_{xst}}{t_x} \frac{t_z}{\hat{t}^*_{zst}},$$

where

$$\hat{t}_{yst} = \sum_{h=1}^H \frac{N_h}{n_h} \sum_{k\in s_h} y_k, \quad \hat{t}^*_{xst} = \sum_{h=1}^H \frac{N_h}{n_h} \sum_{k\in s_h} x^*_k, \quad \hat{t}^*_{zst} = \sum_{h=1}^H \frac{N_h}{n_h} \sum_{k\in s_h} z^*_k.$$

The approximate variances of the estimators $\hat{t}_{Rst}$, $\hat{t}_{Pst}$ and $\hat{t}_{RPst}$ are

$$AVar(\hat{t}^*_{Rst}) = \sum_{h=1}^H N_h^2 \frac{1-f_h}{n_h} \bar{Y}^2 \big(C^2_{yh} + g_h C^2_{xh}(g_h - 2K_{yxh})\big),$$

$$AVar(\hat{t}^*_{Pst}) = \sum_{h=1}^H N_h^2 \frac{1-f_h}{n_h} \bar{Y}^2 \big(C^2_{yh} + g_h C^2_{zh}(g_h + 2K_{yzh})\big),$$

$$AVar(\hat{t}^*_{RPst}) = \sum_{h=1}^H N_h^2 \frac{1-f_h}{n_h} \bar{Y}^2 \big(C^2_{yh} + g_h C^2_{zh}(g_h + 2K_{yzh}) + g_h C^2_{xh}(g_h - 2g_h K_{zxh} - 2K_{yxh})\big).$$

And the estimators of the approximate variances

$$\widehat{Var}(\hat{t}^*_{Rst}) = \sum_{h=1}^H N_h^2 \frac{1-f_h}{n_h} \bar{Y}^2 \big(\hat{C}^2_{yh} + g_h \hat{C}^2_{xh}(g_h - 2\hat{K}_{yxh})\big),$$

$$\widehat{Var}(\hat{t}^*_{Pst}) = \sum_{h=1}^H N_h^2 \frac{1-f_h}{n_h} \bar{Y}^2 \big(\hat{C}^2_{yh} + g_h \hat{C}^2_{zh}(g_h + 2\hat{K}_{yzh})\big),$$

$$\widehat{Var}(\hat{t}^*_{RPst}) = \sum_{h=1}^H N_h^2 \frac{1-f_h}{n_h} \bar{Y}^2 \big(\hat{C}^2_{yh} + g_h \hat{C}^2_{zh}(g_h + 2\hat{K}_{yzh}) + g_h \hat{C}^2_{xh}(g_h - 2g_h \hat{K}_{zxh} - 2\hat{K}_{yxh})\big).$$

4 Simulation study

In this section some empirical study is presented to observe the behavior of the estimators in the case of stratified simple random sample design. A real populations from some Lithuanian Enterprise survey were used for the simulation. During the simulation study several populations were examined.
It should be noted that both auxiliary variables initially are positively correlated with the study variable. So, first of all we transform the variable z to dual, and consider the transformed variable as given negatively correlated auxiliary variable. Below some results when dual to ratio-cum-product estimator performs efficiently are presented.

Population **I**

y - An income of enterprise, x - Number of employees, z - Wages-fund (dual variable).

$N = 636$, $t_y = 119060206$, $t_x = 43785$, $t_z = 1827869$, $\rho_{yx} = 0.7795$, $\rho_{yz} = -0.9496$, $\rho_{zx} = -0.7964$.

Population **II**

y - Gross wage, x - Average earnings, z - Number of employees (dual variable).

$N = 150$, $t_y = 6249836$, $t_x = 23199$, $t_z = 11719$, $\rho_{yx} = 0.8462$, $\rho_{yz} = -0.7001$, $\rho_{zx} = -0.5738$.

These populations are stratified into three and two strata respectively by the size of the variable x. In the first population 4000 samples and in the second population 2000 samples were drawn.

Table 1. **Simulation results for the Population I**

Estimator	Sample size n	Average estimate	Estimated bias	Average estimate of variance $\times 10^{13}$	Approximate variance $\times 10^{13}$	CV
$\hat{t}_{HT}$	100	118929894	-130312	5.1596	5.1330	0.0602
	200	119116419	56213	2.0387	2.0598	0.0381
	300	119031379	-28827	1.0494	1.0639	0.0274
	400	119050873	-9333	0.5517	0.5554	0.0198
$\hat{t}_R$	100	118945710	-114496	4.1413	4.0891	0.0538
	200	119057349	-2857	1.6558	1.6437	0.0341
	300	119062762	2556	0.8313	0.8484	0.0245
	400	119070111	9905	0.4370	0.4435	0.0177
$\hat{t}_{RP}$	100	118683511	-376695	2.9601	2.8936	0.0453
	200	118832365	-227841	1.1656	1.1637	0.0287
	300	118978467	-81739	0.5994	0.6005	0.0206
	400	119067962	7756	0.3101	0.3140	0.0149
$\hat{t}^*_{RP}$	100	118828780	-231426	4.0056	3.9739	0.0531
	200	118958156	-102050	1.1574	1.1568	0.0286
	300	119002767	-57439	0.5488	0.5522	0.0197
	400	119193156	132950	0.8245	0.8075	0.0238

Table 2. **Simulation results for the Population II**

Estimator	Sample size n	Average estimate	Estimated bias	Average estimate of variance $\times 10^{13}$	Approximate variance $\times 10^{13}$	CV
$\hat{t}_{HT}$	40	6249261	-575	1.4846	1.4625	0.0612
	50	6261114	11278	1.0362	1.0741	0.0523
	60	6257137	7301	0.8113	0.7815	0.0447
$\hat{t}_R$	40	6253267	3431	1.0378	0.9861	0.0502
	50	6261150	11314	0.7065	0.7304	0.0432
	60	6257925	8089	0.5284	0.5167	0.0363
$\hat{t}_{RP}$	40	6248636	-1200	0.9007	0.8497	0.0466
	50	6245913	-3923	0.6022	0.6281	0.0401
	60	6242498	-7338	0.4420	0.4474	0.0339
$\hat{t}^*_{RP}$	40	6239766	-10070	0.8725	0.8482	0.0467
	50	6245259	-4577	0.5193	0.5409	0.0372
	60	6240266	-9570	0.3711	0.3732	0.0310

Tables 1 and 2 show that for stratified simple random sampling design the dual to ratio-cum-product estimator can be more efficient than other estimators considered. For samples of size $n > N/2$, the coefficient of variation of dual to ratio-cum-product estimator increases with the size of the sample (Table 1).

References

Bandyopadhyay, S. (1980) Improved ratio and product estimators. *Sankhyā Series C*, **42**(2), 45-49.

Plikusas, A. (2008) Some overview of the ratio type estimators In: *Workshop on survey sampling theory and methodology*, Statistics Estonia.

Singh, M. P. (1969) Comparison of some ratio-cum-product estimators. *S*ankhyā Series B, **31**, 375-382.

Singh, H. P., Singh, R., Espejo, M. R., Pineda, M. D., Nadarajan, S. (2005) On the efficiency of the dual to ratio-cum-product estimator. *Mathematical Proceedings of the Royal Irish Academy*, **105A**(2), 51-56.

Srivenkataramana, T. (1980) A dual to ratio estimator in sample surveys. *Biometrica*, **67**, 199-204.

Upadhyaya, L. N., Singh, H. P. (1999) Use of transformed auxiliary variable in estimating the finite population mean. *Biometrical Journal*, **41**(5), 627-636.

Upadhyaya, L. N., Singh, H. P., Vos, J. V. E. (1985) On the estimation of population means and ratios using supplementary information. *Statistica Neerlandica*, **39**(3), 309-318.

Sampling in Belarus: history, state and prospects

Natalia Bokun[1]

[1] Belarusian State Economic University (BSEU), Minsk, Belarus

e-mail: nataliabokun@rambler.ru

Abstract

The history and problems of introduction of sample surveys in practice of official byelorussian statistics are considered. Features of formation of sample survey of households and the enterprises are shown. Using of a combination univariate and multivariate sampling is offered.

1 History of development of sample surveys

In the conditions of command economy in the national statistics of Belarus, as well as in other countries of FSU Region, a priority it was given to methods of continuous survey with the exception of 3,5 thousand family budgets survey of workers, employees and collective farmers. Then the two-level stratified sample was used: at the first step the enterprises was selected within branches, than hired workers was selected. Such principle of selection ensured wages data representativeness.

In consequence of disintegration of the USSR and occurrence of market relations the economic situation has changed. Notably restrictions on individual labor activity have been removed, the structure of sources of revenue has changed, the number of small state and private enterprises has sharply increased in all economic branches. So, the total number of the small enterprises (SE) in republic has came to 28 310 in 2000, 33 094 in 2005, 51 240 in 2007. From each of them was inexpedient to demand of statistic registration. Full coverage of population has become economically unjustified and almost unrealizable. Thereby only sampling was left a reliable method of an estimation minute, small and even the average enterprises. In addition, changed statistic registration of family incomes also demanded change of survey character. As a result process stage-by-stage introduction of sampling in the practical statistics has begun:

1. 1995-1996 Building of new model of sample surveys of household (HH), based on the international standards in the field of construction of the sampling plan, workings out of toolkit, data processing;
2. 1997-2005 Theoretical workings out and pilot sample surveys of the enterprise (retail trade, services, small business);
3. Since 2006 until now. Theoretical workings out and selection of the enterprises on a regular basis (retail trade, small business, labor statistics).

The methodology of sample surveys of household was developed by Ministry of statistics and the analysis of Republic of Belarus (since 2008 – National Statistical Committee, or Belstat) by means of experts from IMF.

At the second stage of introduction sampling in statistical practice (1997-2005) Statistics research institute provided with methodology and software of branch survey of the enterprises, based on using of group of methods of univariate selection: systematic sampling,

random selection without allocation, simple random sample, stratified sample with proportional and optimal allocation. Pilot surveys of SE in retail trade were carried out in 1998-1999, survey of enterprises in services – in 2002, survey of small enterprises in economic branches – in 2003. In 2005-2006 problems of building of multivariate sample are investigated, the first version of the program is developed, trial multivariate samples of SE are spent.

At the third stage (since the end of 2006) researches of multivariate sampling and improvement data extrapolation are hold on. State statistics began to carry out quarter samples of SE in area of labor statistics on a regular basis. Since 2008 has added sample surveys in retail trade and in catering. Special quarter sample surveys of SE concerning employment and unemployment, and also personal subsidiary plots is predicted.

Despite such advantages of sample, as enough low expenses, efficiency of and high reliability, statisticians was confronted with a number of problems: non-responses; atypical units, samples in small domains, splitting of survey population into small groups and sample fraction, inadequate data extrapolation, problems of compromise, complexity of a choice of an optimal way of the multivariate selection, complexity of a choice a leading indicator, technical impossibility of construction of multivariate for large population (over 400-500 units), absence of the standard methods of calculation of errors and estimation.

Changing of structure of base population, allocation in separate files of the atypical enterprises (HH), using of procedures of reweighing or replacement help overcome the problems non-responses and «atypical units of population». Other problems can be resolved by building of stratified sample or multivariate sample which ensure representativeness of sample of small population and can adequate extrapolate the sample data on survey population. The author is offered to apply a combination of univariate and multivariate methods. This approach is realized in carrying out of sample survey of SE, survey of level and structure of wages and survey of enterprises in retail trade.

2 Sample surveys of household

The survey is conducted annually since January 1995. Their main purposes are welfare estimation of all population and particular groups. Survey object is household. Survey is carried out at all country regions and separately in Minsk. It's annually is covered 0,2 % or 6000 HH. In this kind of sampling is used territorial probabilistic three-stage sample. At the first step sampling units are cities and rural soviets; on the second step sampling units are local-polling districts in city and data of the soviet account in rural soviets, on the third – HH. Procedure of selection of administrative and territorial units repeats 1 time in 10 years. Selection of polling districts and HH carries out annually. The methodology of weighing and raising of the selective data on a large population is based on assignment of each finite unit (HH) the corresponding weight (Bi):

$$B_i = \frac{1}{\rho_1 \cdot \rho_2 \cdot \rho_3},$$

here ρ_1 – probability of selection of each city and rural soviet;

ρ_2 – probability of selection of each polling district in cities, zones in rural soviet;

ρ_3 – probability of selection of everyone HH within polling district or a zone.

Base HH weights are corrected on uninhabited apartments and non-responses by using mathematics methods.

The sample program assumes filling of some questionnaires, which are containing the information on living conditions, personal subsidiary plots, education, health, employment. Daily and quarterly questionnaires include such parts as expenses on food and unfood, payment of services etc.

3 Sample surveys of small enterprises

Since 2005 year they spent quarterly. Survey objects are artificial persons of small business, i.e. SE. According to the legislation this is the organization with number of the working from 100 persons in industry and transport branches till 25-30 persons in services. Sample frame is the file of SE. Sampling carries out at each regions and Minsk by branches. The used sampling model provides possibility of a choice to use a method of sample building depending on population, number and character of survey variables (the program «Multivariate sampling»). It should be done several steps for searching optimal sample size for i-th branch and j-th region.

1. The population of survey variables is allocated (for example, the wages fund, average number working). Average rates, total rates and variability rates are calculated .

2. It should be executed one of third conditions for applying multivariate sampling:

1). coefficient of variation more than 100 %;

2). survey objects are non-uniform on many variables;

3). the small size large population (top limit 400–500, bottom limit – 30–40 units). Otherwise it should be used univariate sampling: systematic sampling, random selection without allocation, simple random sample, stratified sample with proportional and optimal allocation.

3. If it is expediently to use of multivariate sample, selection is carried out by the cluster analysis:

– large population is broken on homogeneous groups to k-variables:, i.e. clustering;

– stratification within clusters to the allocated leading sign (volume of output);

– optimal sample population is choose for each cluster, where standard sample error of k-variables is a criteria of productivity.

–

If the error exceeds admissible bounds, three methods of its reduction may be applied: 1) increasing sample population in cluster; 2) additional stratification of the enterprises in cluster to a leading variable; 3) repetition of clustering, but with larger number of steps, or using of an iterative method with the preliminary number of clusters r> l.

Extrapolation of total value of variables on all population is carried out by traditional group raising factors (ratio of number of units in i-cluster of total population and corresponding cluster of sample) and simple errors.

Sampling frame is 20–30 % from all number of small enterprises. As to branch sampling fraction is depending on number of SE and the degree of accuracy on a leading variable: a relative sampling error on regions less than 2 %, on branches less than 5 %, and on small branches less than 8–9 %.

4 Sample survey in labor statistics

This sample population is a base of special quarter sampling.

Sampling includes three types of survey: 1) survey of level of average wages by category; 2) survey of distribution of workers on wages; 3) survey of number of workers of SE on an educational level, studying, age.

Survey of average wages by category is carried out once in two years. The large and average enterprises of the basic industries are surveyed. Two-level sample is used: at the first step as a multivariate sampling the enterprises are selected, on the second workers in each enterprise are selected by mechanical sampling. The first frame is an array of the enterprises reporting on report № 1-work (monthly), the second frame is the list of workers, including heads, experts, other employees and the workers who completely have worked surveyed month (October). For small enterprises an approximate sample fraction of workers is 20-30 %, for average and large enterprises is 5-10 %. The frame for all Belarus fluctuates within 30-35 % from all organizations.

Raising of the sample data on the total population is carried out with application of the aggregative weight () a finite unit (worker):

$$\kappa_{ai} = \kappa_n \cdot \kappa_{ui}$$

$$x_{,эij} = \kappa_n \cdot \kappa_{ui} \cdot x_i$$

$$\kappa_n = \frac{N_m}{n_m}; \quad \kappa_{ui} = \frac{T_i}{t_i},$$

here Kn – enterprise weight;
Kui – individual worker weight for i-th category;
xi – value of x for worker i-th category;
xэij – the extrapolated value of x for i-th category of workers on j-th enterprise;
Nm, nm – number of the enterprises of m-th group accordingly in general and sample population;
Ti – an aggregate number of workers of the enterprise of i-th category who completely have worked surveyed month;
ti – number of workers of the enterprise of i-th category who in sampling.

Information files of database with registration variables of the surveyed persons and the weight rates are formed at regional levels. This information is systematized, extrapolated and generalized at republic level.

Survey of distribution of workers by wages

Sampling is spent annually. The surveyed period is calendar month (May), the purpose is the information about territorial differentiation of a branch wages. Survey object is the enterprises of all branches of economy, except for subjects of small business. Combination univariate and multivariate stratified sampling is used. The units by the form their activity and by region are stratified (selective variables are a number of workers and a wages fund). Selection should be optimal if limiting and standard errors are minimum with fixed sample size. Sampling fraction is about 20-35 % from all enterprises.

Survey of number of workers of SE by educational level, studying, age.

Sampling is spent annually. The purpose is the territorial analysis of sex, age and educational structure of SE staff in economic branches. The surveyed parameters are dependent on number workers variables such as structure of employed by sex, age, the professional studying. It could be unreasonable apply multivariate sampling. So univariate stratified sampling is used (simple random, proportional, optimum stratification). Sampling fraction is about 20-25 % from all enterprises 2005 according to their reporting in 2006 (form 6-t-staff). Raising sample data is made by of weight of organization and sample structure of workers. The weight of organization is calculated as reciprocal group fraction of the enterprises for i-th branch and j-th region. Sample errors are within the range from 0,02 to 5 %

5 Sample survey in retail trade

Sampling is covered the organizations without departmental governance. The purpose is the analysis of retail goods turnover and public catering in retail trade. The sample frame is represented by an array of the organizations reporting the form № the 1-trade (monthly). The enterprises code, the name, territorial code, a code of a kind of activity, goods turnover indicators, sales of products, sales food and unfood are considered. Sample population is formed by each region and Minsk separately in retail trade and catering. The multivariate sampling is used. The surveyed variables are retail goods turnover in accounting and previous years. Sampling fraction is about 15-20 % from all organization. Sampling fraction in retail trade is less than 10-15%, in catering is more than 20-25 %.The relative error should be less than 1 % by the country and less than 2-3 % by the regions. The weight of organization is calculated as reciprocal group fraction of the organization.

The experience of building of samples of household in Belarus has shown:

- the main problems of sample of household are non-responses (20-30%), necessary localization small sample, building of regional and demographic subsampling, small statistic estimation area;
- the main problems of branch sampling are necessary estimation of the whole variables, splitting of sample population on the smaller groups, little subsampling, using difficult estimations, adequacy of the expert judgements;
- the most applied model of branch sampling is based on a combination of univariate and multivariate samplings;
- recommended sampling fraction of the enterprises in the survey of small business, for wages – 20-30 %, for survey in retail trade a sampling fraction – 10-15 %, for sampling of standard of living – 0,2 %, at employment and unemployment – 0,6 %, private farming – 0,3-0,4 %.
- optimal and simple random stratification is most efficient among the methods of univariate sampling; methods of cluster analysis are the most comprehensible on degree of reliability and availability to the user among multivariate methods.

TEACHING SURVEY SAMPLING THEORY AND METHODOLOGY AT THE UNIVERSITY OF LATVIA

Natalja Budkina

University of Latvia, Latvia
e-mail: budkinanat@gmail.com

Abstract

The paper deals with teaching the theory and methodology of survey sampling at the University of Latvia. The overview of the programme of professional studies in Mathematical Statistics is given. The short program of the course of survey sampling in this programme is presented. The problems that arose during teaching the course and the perspectives of development are mentioned.

1 Introduction

Sample surveys are essential tools in a modern society to provide accurate information to politicians, businesses and the general public about living conditions and opinions. Unfortunately, the Baltic countries started to develop survey statistics only after gaining their independence. Now the course Survey Sampling is given in the programme of professional studies in Mathematical Statistics of the University of Latvia.

2 Programme of professional studies in mathematical statistics

The programme of professional studies in Mathematical Statistics has been running at the Department of Mathematics, University of Latvia since 1997/1998 academic year. In 2001 it was accredited for the period till December 2007 and reaccredited in 2007 for the period till 2013. During whole period 2001-2008 the programme remained the most popular among all mathematical programmes at the University of Latvia. The reaccredited programme is 4,5 years long, to compare with the previous 5 years long programme. Naturally, this transition required a serious rework and modernization of the programme. The modernization of the programme was necessary in order to keep the educational process up to day in spite of reduction of number of hours and courses. Besides, it was essential to introduce new actual courses in some branches of modern mathematical statistics and to apply new software. Now this programme is valued at 180 national credit points (70 of them are common with Bachelor Programme in Mathematics).

Now the programme foresees the following modules of the courses and requested activities of our students:

- a module of Pure Mathematics;
- a module of Mathematical Statistics and Probability Theory;
- a module of Analysis of Social and Economic Processes;
- a module of Software Provision for Statistical and Numerical Problems;

- a module of General Subjects;
- probation work;
- a Diploma Thesis.

The module of Pure Mathematics aims at two mutually connected goals and tasks. First, to guarantee the students, after successfully mastering the module, to have knowledge and professional skills in different areas of Pure Mathematics which are necessary in order to acquire speciality courses in probability, statistics, econometrics, decision making and others. Second, to give the students an opportunity to continue studies at the second, master level in one of the four offered directions in Mathematics at our university (Probability Theory and Mathematical Statistics, Mathematical Simulation, Mathematical Didactics and Modern Elementary Mathematics, Pure Mathematics, orientated towards Algebra, Topology and Analysis).

The aim of the module of Mathematical Statistics and Probability Theory is to educate specialists of high level of proficiency in Mathematical Statistics, whose knowledge and professional skills would be sufficient in order to make statistical analysis in all areas of modern economics, social and scientific activities.

The module of Analysis of Social and Economic Processes provides students with principles and methods in simulation and decision making.

There are courses in the software provision for statistical and numerical problems in the fourth module.

The module of General Subjects is relatively small and foresees a course in professional English and a course in Natural Sciences.

The probation period is increased from 4 months in the previous programme to 26 weeks in the modernized programme. The problem to organize fruitfully the probation work for students is quite difficult. The staff of the Department tries to develop the collaboration with enterprises where the probation for the students could be successfully organized.

Diploma work constitutes an essential part of the programme. By defending the Diploma a student demonstrates that his/her theoretical knowledge and developed skills are sufficient to start professional activities. The subjects of Diploma work for graduates of our programme are usually related to Probability Theory, Mathematical Statistics, Applied Statistics and Actuary Sciences as well as to Mathematical Simulation of Social and Financial Processes. One of the principal criteria to accept the work for defence is its mathematical correctness.

3 The teaching programme of survey sampling

3.1 Course in Survey Sampling

The special course in Survey Sampling was introduced at the Department of Mathematics in 1996 by Dr. Math. J. Lapiņš. It was intended for the students of the programme of the professional studies in Mathematical Statistics in the fourth-fifth year of studies. For the reaccreditation of this programme in 2007 the course was reworked. The changes were not very essential, but they were important taking into account the development of Survey Sampling Theory and Methodology during the last years. Now this course consists of 54 hours of lectures and 10 hours of practical lessons during the autumn semester in the fourth year of studies.

The lectures on Survey Sampling are based mostly on the book by W. G. Cochran (1) and S. L. Lohr (), which are available in the library of the Faculty of Physics and Mathematics of the University of Latvia. Unfortunately, we do not have a good literature on survey sampling in Latvian. For example, the overview of some sampling designs is given in Z. Goša (2003),

O.Krastiņš (1998), O.Krastiņš, I.Krūmiņa (1993). Most of the students can use the methodological materials by N. Budkina.

3.2 The programme of the course "Survey Sampling"

Course plan is as follows:

1. Introduction to sampling theory and methodology (the main concepts and definitions of Survey Sampling Theory, the main stages of modernization and realization of survey, types of sampling).
2. Simple random sampling with and without replacement (sampling scheme, definitions and notions, estimators of total, mean, proportion in the population and in domain, estimators of variance, sample size).
3. Sampling with Unequal Probability (description of the techniques, estimators).
4. Stratified sampling (sampling scheme, allocation of the sample size, estimators of total, mean, proportion, poststratifcation).
5. Ratio and regression estimators (description, estimators).
6. Systematic sampling (description, estimators, population with linear trend, population in random order, population with periodic variation).
7. Single-stage and multistage cluster sampling (description, estimators for the case of sampling with equal and unequal probabilities wit and without replacement, optimum selection of sampling rates, Jackknife method).
8. Double sampling (description, estimators).
9. Errors in surveys, their sources and Methods of their reduction. (nonresponse).

Practical works

During the course each student must work out 3 home works and control work. The practical works, the home works and the control work give 40% of the final mark.

Control work consists of short theoretical question and the numerical problems. Numerical problems are solved without using a computer during the practical lectures and the control work. Two home works consist of two common tasks and the one individual. The third home work is individual for each student. The Statvillage data of C. J. Schwarz (1997) and the exercises of S. L. Lohr (1999) are used for the preparation of the variants for these home works. Therefore a computer is needed for the solution of some numerical problems from home works. Most of the students use SPSS and Excel. The results of home works are discussed during the practical lectures.

The main problem that arises in teaching Survey Sampling is the lack of practical training. More practical work is needed for students and a possibility to provide the students with real data is needed for teachers.

3.3. Probation work and Diploma Thesis

During the whole period of teaching the course “Survey Sampling” it has been popular among the students. Every year there is at least one student choosing a theme on Survey Sampling for his/her probation work or Diploma paper.

Probation work

As it was mentioned above, Mathematical statistics programme foresees an essential increase in the role of probation work. The problem to successfully organize a half year long probation work is difficult. The tasks which the students have to fulfil are essentially different but a lot of students choose themes connected with survey sampling. A long period of probation work gives the possibility for the students to participate in real sampling projecting, carrying it out and its analysis.

There have been cases when the students themselves offered the companies to carry out a survey and got the support. The aim of these surveys was to find out the opinion of the clients and partners of the companies about different aspects of cooperation, to find out the opinion of the workers of the companies about working conditions, to estimate some parameter. In carrying out these surveys different types of sampling have been applied: simple random sampling, stratified random sampling, single-stage cluster sampling and two-stage cluster sampling.

Unfortunately, there have been some cases when the probation work was connected with statistical data processing, when these data were restricted and the problem to let the students use these data was aroused.

Diploma Thesis

Since the course "Survey Sampling" is intended for the students of the programme of professional studies, most of the Diploma Papers have practical purpose. It is common (however not obligatory) that the subject of the Diploma is closely related to the work which a student has fulfilled in the result of the probation work.

Since 1996 22 Diploma Theses on Survey Sampling have been defended by students of the speciality "Mathematician-Statistician" of the University of Latvia. There have been theoretical, practical and methodological works, which have been written under the supervision of J. Lapiņš, N. Budkina, S. Bāliņa, K. Lece and M. Liberts. 9 Master Theses have been defended under the supervision of J. Lapiņš, I. Priedola, S. Bāliņa.

4 Conclusions

The course „Survey Sampling" is very important and interesting for the students of Programme of Professional Studies in Mathematical Statistics at the University of Latvia. It could be developed by introducing new methods and modern techniques, the number of hours for practical works should be increased.

References

Cochran, W. G. (1977) *Sampling techniques*. Wiley and Sons.

Goša, Z. (2003) *Statistka*, LU, Rīga

Krastiņš, O. (1998) *Statistika un ekonometrija: mācību grāmata augstskolām*. Latvijas Republikas Centrālā statistikas pārvalde, Rīga.

Krastiņš, O. and Krūmiņa I. (1993) *Izlases metode*. LU, Rīga.

Lohr, S. L. (1999) *Sampling: Design and Analysis*. Duxbury Press, Pacific Grove.

Schwarz C. J. (1997) StatVillage: An On-Line, WWW-Accessible, Hypothetical City Based on Real Data for Use in an Introductory Class in Survey Sampling. *Journal of Statistics Education*, 5(2). http://www/amstat.org/publications/jse/v5n2/schwarz.html

Use and Theory of Random Digit Dialing in Sweden

Erik Bülow[1]

[1] Department of Mathematical Sciences, University of Gothenburg, Sweden
e-mail: bulow@student.chalmers.se

Abstract

This article is a meta study which describes the use and theory of *Random Digit Dialing (RDD)* in Sweden with special emphasis on the choice of respondents.

1 Introduction

There has been a long history of telephone surveys in Sweden. The Public Switched Telephone Network (PSTN) covered almost 99 percent of the Swedish population during the 1980s (Groves et. al. 1988). There have also been very good registers available through the state provider of telecommunication solutions, Televerket[1]. This has come to change however. Sweden is now one of the leading countries in terms of using cell phones instead of landlines[2]. Among people living in households consisting of members aged below 26, as many as one third have no longer any connection to the landline network (Thörn 2006).
However, due to integrity aspects of the respondents (among other things), the leading companies in the field of survey statistics in Sweden (Synovate Sweden[3] and Sifo Research International[4]) still limit their surveys to the landline networks. This report will therefore focus on such surveys.
Section 2 is a brief description of the construction of a Swedish telephone number. Section 3 will focus on some different ways to choose such a number to call. Section 4 describes the choice of a respondent within a given household. Section 5 gives some information about post stratification and section 6 gives some thoughts about the future.

2 Sample Frame

2.1 The Construction of a Swedish Telephone Number

In this section we describe the different components of a Swedish telephone number. The system follows the international public telecommunication numbering plan E.164. According to this plan, a telephone number that can be internationally reached, consists of maximum 16 digits.[5]
Here is an example of a Swedish telephone number written in full international format:

+46 3 1 713 7938

[1] Televerket is now a private company named TeliaSonera. It is the leading provider of telecommunication solutions in the Nordic and Baltic region.

[2] Telephone by landline will herein include telephones using PSTN, ISDN (*Integrated Services Digital Network*), xDSL (where the "x" refers to a number of different techniques and where DSL stands for *Digital Subscriber Line*) and cable TV-network.

[3] Synovate Sweden was founded in 2006 by merging Temo and LUI Marknadspartner. RDD is performed by a subsidiary called Fieldwork International.

[4] Sifo is an old acronym for "Svenska Institutet för Opinionsundersökningar". The name was later used as a brand name for Research International in Sweden. Research International and TNS-Gallup merged to TNS in February 2009.

[5] The exceptions are, for instance, the alarm number 112 that can not be reached outside the specific country.

where + marks the international call prefix (00 is used by most countries and is also recommended by the International Telecommunication Union (ITU). **46** is the country code for Sweden. **3** leads to a geographically bound area of Sweden consisting of the provinces of Halland (except the city of Laholm), Southern Bohuslän, Western Västra Götaland, and Western Småland. **1** goes to Gothenburg, the major city within this area. **713** is a local serial number connected to a specific node used by the telephone operator (the geographical connection is however not that strong anymore). The length of this number can very between one and four digits. **7938** is at last the individual number specifying the end user. The length of this number is always four digits.
Calling a Swedish telephone number inside Sweden is done by replacing the international code prefix and the country code by "0". The example number is called by:

0 3 1 713 7938[6]

Hence, the total length of a Swedish telephone number can vary. This is important since most literature in the field describe Random Digit Dialing (RDD) in USA where telephone numbers have a fixed length of ten digits.

2.2 Telephone Directories

There have always been good registers available in Sweden, connecting individuals to telephone numbers. In the USA as a counterpart, there are approximately 5 000 different registers that have to be taken under consideration when performing a telephone survey. Televerket was for a long time the main provider of such registers in Sweden. This register is now owned by a private company, Teleadress AB, which is in turn owned byowned by the daily evening paper Aftonbladet. PAR is an acronym for "Postens Adressregister" and was originally founded by the Swedish state messaging and logistics operator Posten. PAR is now a part of Bisnode, owned by Ratos and Bonniers.[7] Synovate Sweden uses a list of telephone numbers from PAR. Sifo buys lists from Teleadress.
These registers are good but not perfect. Approximately 6.77 percent of the Swedish landline numbers were not listed in 1997 (Forsman and Danielsson 1997).
Telephone numbers used for bigger companies and organisations are clustered in seperate number series. Private numbers, numbers not in use, secret numbers and numbers used by smaller companies are mixed in the remaining number series. The distribution among these number types is not clear. Some operators give subscribers number randomly, some operators follow different patterns. It is also possible for a subscriber to switch operator while keeping the old telephone number (number portability). This distribution was however investigated by Forsman and Danielsson (1997). They concluded (using 20 000 numbers from the Swedish city Linköping and Wald-Wolfowitz runs test) that the number types had a discrete uniform distribution within the number series used for private subscribers.

3 Randon Digit Dialing

3.1 Theoretical Suggestions

Several methods for the choice of respondents have been suggested. We describe some of them here. See for example Groves et.al. (1988) and Lepkowski et.al. (2008) for additional suggestions and details.

At random from register. The quality then depends on the updating frequency in the register.[8] The inclusion probability for subscribers with a secret or unlisted number equals zero.

Totally at random. This would lead to a large amount of misdirected calls. An alternative is to use lists of active number series (*List Assisted Random Digit Dialing*, LA-RDD). In 2006, 78 percent of the Swedish telephone numbers were however assigned to private subscribers (Thörn 2006).

[6] This is usually written as 031-71 37 938.

[7] Bonnier is also the owner of Aftonbladets biggest competitor Expressen.

[8] The register might be regularly updated by the register company but the survey company often buys extract every third month or so. In the summer of 2006 a Swedish survey institute also discovered that the register they bought had a lack of the 500 000 latest registered subscribers. This might lead to severe bias. (Forsman 2007)

In addition to that, all numbers in the active series are not in use. In USA for example only 12 percent of the numbers in the active series are actually in use (Lepkowski et.el. 2008, p. 6).

Sudman's Method. Choose m telephone numbers by the "at random from register"-method described above. Remove the last $\log N$ digits from each number. The resulting k ($k \leq m$) numbers can now be used for generating kn new numbers by appending $[x_{ij}] \sim \text{Unif}(0, N)$ to number $i = 1, 2, \ldots, k$ where $j = 1, 2, \ldots, n$.

A version of this method is used by Synovate Sweden ($N = 10$) and was used also by TNS-Gallup ($N = 100$) before 2003.

Mitofsky–Waksberg's Method (MW). This was the most well used method in the USA in the 1980s. Unlike in Sweden, there was a huge number of number series registered as active but with no active subscriber numbers within it.[9] Several modifications have been suggested (see f.eg. Potthoff's Method below). The purpose of the method is to first investigate if the number series are active and if so, choose respondents within them. Series in the USA, with at least one active subscriber tend to be active to at least 60 percent (Casady and Lepkowski 1999).

Here we describe a modification of the MW method which takes advantage of the Swedish telecom system. Choose m active number series at random, available from the Swedish Post and Telecom Agency (PTS). These series might have different lengths (100, 1 000 or 10 000 numbers in each). Denote the number in the smallest series by N. Limit the longer series by adding digits in the end randomly. Choose one full number from each of the m series at random. Investigate (by register or by calling) if these numbers leads to a private subscriber or not. We now limit our sampling frame to the m_p number series with positive results from this test. We finally choose k numbers from each of these series to call. It has been suggested in the USA to choose $mk \geq xn$ where n is the number of respondents we wish to have and where x varies between 6 (Bourque and Fiedler 2003) and 40 depending on which source to rely on. In Sweden it might be enough with $x = 2.19$. This number relies however on estimates from Synovate Sweden which do not use the MW method.

Potthoff's Method. This is a generalization of the MW method. We investigate c numbers from each originally selected m series (or "cluster" to use Potthoff's own terminology). Let c_i denote the number of active subscribers among the c in each number series $1 = 1, 2, \ldots, m$. Discard the series where $c_i = 0$. If $c_i \geq 2$ then choose ck more numbers at random from this series to use. If $c_i = 1$ then choose $k(c-1)$ numbers at random and then call as many numbers you need in order to find k additional subscribers within the series.

Plus Digit Sampling. Choose n numbers at random from a register. Add to each number a digit or number x. $x = 1$ is a common choice but the inclusion probability will then equal zero for approximately 15.5 percent of the Swedish subscribers (Forsman and Danielsson 1997). It is possible in Sweden to add numbers up 20 and still have a resulting subscribing number in 94 cases out of 100 (ibid.). An alternative is of course to assign x values at random.

All the new numbers are used for calling while the original ones are discarded.

3.2 Methods Used in Sweden

All methods described above were originally developed to suite the needs in the USA. The methods used in Sweden rely, to some extent, on these methods but are often modified in order to use the Swedish registers and the Swedish implementation of the international public telecommunication numbering plan E.164.
We will describe three methods used by the biggest companies in the Swedish telephone survey sector. Each company might also use other methods but here we concentrate on the so called omnibus surveys.

3.2.1 TNS–Gallup

TNS–Gallup used RDD until 2003. They regularly bought geographically stratified numbers from Teleadress. The stratification was done with respect to the 70 Swedish A-regions. Sudman's method

[9] This is not possible in Sweden. The Swedish Post and Telecom Agency (PTS) only sells number series to the operator if their intention is to activate them in the nearest future.

(with $N = 100$) were then used to construct the numbers to call. Computer Assisted Telephone Interviewing (CATI) was used. If there was no answer on the call, a new time for calling back was assigned randomly. Each number was called up to 32 times.
The company changed their method in 2003. They now use the Swedish population registry, SPAR (Statliga Personadressregistret) for choice of respondents.[10] If possible, each person is then connected to his or her telephone number by using an additional register. If no number is available, a postal questionnaire is instead sent by mail. It is often possible to find telephone numbers for 87 to 90 percent in the population (Petersson and Holmberg 2006). This method has the big advantage that the survey includes most Swedish citizens (homeless and institutionalised people excluded), not only those with a landline telephone.

3.2.2 Synovate Sweden

The actual calling is performed by Fieldwork International (FI). The company is situated in Karlskrona. The call centre has room for 92 telephone interviewers. RDD is used for omnibus surveys. Too many questions in one survey might have a negative influence on the response rate. It is therefore possible to have several smaller surveys at the same time.
FI buys a selection of active telephone numbers every month from PAR. It consist of one third of all registered landline telephone numbers in Sweden except bigger clusters assigned to organisations.[11]
At first, 2 500, telephone numbers are randomly selected from the selection frame bought from PAR. This selection of telephone numbers is used during two subsequent weeks, each week consisting of workdays Monday to Thursday with working ours 5 to 9 pm. It is a standard to perform 250 interviews per shift. In order to achieve this, 250 numbers is randomly selected out of the 2 500. Sudman's method is applied to each number with $N = 10$. Assume that *031-71 37 938* was chosen. We then construct new number series by substituting the last digit, 8, to $0, 1, \ldots, 9$. Then we have 250 number series consisting of 10 numbers each. The goal is to perform one (and only one) interview in every one of these smaller number series during the ongoing shift. Let's say we choose to call number x in one of these series. If a private subscriber answers the call, we perform an interview after selecting respondent within the household using the Troldahl–Carter method (see section **??**). No more number is then called from that series. If the number does not lead to any private subscriber, the number is substituted within the same number series.[12] If the line is busy or if there is no answer within 25 seconds, the number will be called again later. This event is scheduled to take place $k_s + y$ minutes later where k_s is a constant where $k_{occupied} \leq k_{no\ answer}$ and where y is chosen by a random variable $Y \sim \text{Unif}(0, y_s)$ where y_s is again a constant depending on the reason for the new call. After 10 unfavourable calls to the same number, a new one is taken as a substitute from the same number series. If it is not possible to perform any interview in a specific number series during the shift, the series will be reused for up to ten days.

3.2.3 Sifo Research International

The actual calling is performed by a subsidiary to Research International ("Sifo" is just a brand name). The company is situated in Ronneby (approximately 30 kilometres from FI in Karlskrona). The call centre has room for 90 telephone interviewers. RDD is used for omnibus surveys every second week.[13] Numbers are bought from Teleadress every month. The amount of numbers is not known to the author. The selection is said to be performed using "advanced computer algorithms". Geographical stratification due to Sweden's 70 A-regions is used. 250 numbers out of the bigger selection are randomly chosen to be used during each shift (Mondays to Thursdays at 5 to 9 pm). Plus digit sampling is then

[10] SPAR is administrated by Infodata AB, a company owned by Bisnode AB, which is also the owner of PAR.

[11] PAR had a register of 4.7 million subscribers in 2007. One third of this equals 1.56 million numbers. All numbers in the register are ordered in some way. Every third number is taken from this order. The selection is therefore not totally random.

[12] The procedure is a little different depending on the actual number status (not in use, company subscriber, fax machine, computer modem etc).

[13] There are some obvious similarities between the procedures used by Sifo and FI. The Sifo call centre was organised in 1990. The same organiser was later also involved in the building of FI.

performed on each number by adding digits from 0 to 9. Suppose the initially chosen number was *031-71 37 938*. We will out of this number construct a number series containing the numbers *031-71 37 9x* where $x \in \{38, 39, \ldots, 47\}$. Then we have 250 number series consisting of 10 numbers each. The goal is to perform one (and only one) interview in every one of these smaller number series during the ongoing shift.

Let's say we choose to call number k in one of these series. If a private subscriber answers the call, we perform an interview after selecting respondent within the household using the Troldahl–Carter method (see section **??**). No more number is then called from that series. If the number does not lead to any private subscriber, the number is substituted with a new number from the same number series.[14]

If the line is busy or if there is no answer within 6 signals, the number will be called again later. This event is scheduled to take place 30 minutes later. This is repeated during the evening. If there is still no answer, the number will be reused in the upcoming omnibus survey (two weeks and one day later)[15]. Substitution is made within the number series if it is not possible to recruit any respondent within a answering household.

The consequence of the substitution at Sifo has been rigorously investigated in a series of articles and papers by Gösta Forsman et.al (f.eg. 1992, 1997 and 2002). We will here only mention that households with fever members tend to be under represented in favour of bigger households.

4 Selection of Respondent Within Households

The proportion of single person households in Sweden is among the highest in the world. In 2005, 46.5 percent of all households consisted on only one person. The following section will however account for the remaining 53.7 percent where the respondent has to be chosen among the household members. It is a widely accepted fact that senior citizens are over represented in favour of younger people. It is also more common that the person answering the call is a female.[16]

Kish Tables (K). The method was suggested by Kish 1949 (see Kish 1965 for a rewritten version). It uses 12 tables in 8 different versions. The interviewer makes a list of all members in the household including their gender and age (this is done by asking questions to the person answering the call). Then the interviewer uses one of the tables to choose the respondent. Assume we have 4 adults in the household and we use for the moment table C. Hence adult number 2 (from the prepared list) is chosen as the respondent.

C	Number of adults in household:	1	2	3	4	5	6+
	Choose adult number:	1	1	2	2	3	3

Sifo used a modified version of K at least until 1993. They did however not ask for the members age and gender. They later changed to use the Troldahl–Carter method.

The intention of the method is to find a good stratification with the emphasis on age. Note however that most non-single households in Sweden consist of two persons (53.2 percent in 1992). A very strong positive age correlation has been found in these households (as well as a very strong negative gender correlation). The choice of a household is therefore much more important than the choice of respondent within the households (to obtain a good mix with respect to age).

The Troldahl–Carter Method (TC). [17] The method was suggested by Troldahl and Carter (1964) as a modification of the Kish method.[18] The purpose was to find a faster method with the same inclusion probabilities. A meta study investigating the use of different methods was done in 1993. Among 18 survey companies in Sweden, Denmark, Germany, France, USA, England and

[14] The procedure is a little different depending on the actual number status (not in use, company subscriber, fax machine, computer modem etc).

[15] If the number was originally called on Monday, it will be called again the next survey on Tuesday. This is repeated four times.

[16] This result can be found in reports from Sweden, Austria, UK, USA, Canada and Australia.

[17] The method is sometimes called Troldahl–Carter–Bryant's Method.

[18] Note however that the Kish method was then slightly modified the year after, in 1965.

the Netherlands, none used the TC method (Forsman 1993). Synovate for example used the birthday method at least until April 2006 (Petersson and Holmberg 2006). In 2007 both Sifo and Synovate Sweden did however use this method.

The telephone interviewer asks two questions: 1) How many adults live in your household? 2) How many of these are male/female?[19]

A table is used for the choice of respondent. Suppose f.eg. that the household has three members and two of these are male, then the oldest male is selected as a respondent.

Hagan–Colliers Method (HC). This is a simplification of the TC method (and therefore a simplification also of the K method). The interviewer directly asks for the "youngest female" or the "oldest male" etc according to a predefined plan. The purpose is to increase the speed of the selection. However the person answering the phone might be confused if such a person does not exist in the household. The method has not been used in Sweden to any known extent.

The Birthday Method (B). This is the simplest method. The interviewer asks for the member of the household that most recently had its birthday. It requires no further knowledge of the household. The method was widely used in the USA during the 1970s and 1980s. It does however have some disadvantages. The person answering the phone is more likely to be a woman than a man (this is a widely accepted fact referred to in many articles). The person answering might also be curious by the question and therefore reply that he or she is the one with the most recent birthday (hoping for a present). This might lead to over representation of females to males. The Swedish statistics for birthdays also shows that a survey performed during spring tends to include younger people than one performed during the fall (Forsman 1993).

4.1 Comparison Between Methods

The choice of a method might not be obvious. At least three components must be taken into consideration. 1) *Time:* How time consuming is the method? 2) *Non-response rates:* Some questions needed for the respondent choice might lead to a higher amount of drop outs. 3) *Representative sample:* Is the distribution regarding age, gender etc the same in the sample as in the population?
This has been investigated by Forsman (1993) together with Sifo. The K method was, as expected, the slowest. The TC method was however faster than both B and K (HC was not used for comparison). There were no significant differences regarding drop outs but the TC method is slightly better than its competitors (3.6 % compared to 4.3 % and 4.4 % for B and K respectively). Significant differences have been found in the USA. The K method gives the best proportions between males and females but TC is also quite close. The B method leads to over representation of females.

4.2 Substitution Within Households

Substitution within a given household has not been described much in the literature. It is however used by many survey companies.

Synovate Sweden. FI does not use any substitution within households. If the selected respondent is not at home at the moment, the interviewer calls back later.

Sifo. Only one answered call per household is made. If the first selected respondent is not at home, the youngest male is chosen instead. This is, of course, due to the fact that younger males is the most under represented group in age and gender. The third choice is to choose the oldest woman. The last choice is to choose the person answering the call as the respondent. Note that the high number of single and two person households in Sweden makes this choice the most common. Substitution was used in 15 to 20 percent of all households in 1993 (ibid.). 80 percent of all substitutions were made in households consisting of two persons. The method led to under representation of male to female, clerks to labour workers, private employers to public employers and older to younger people. All differences were proven by a significance level of $\alpha = 0.05$.

[19] It is most common to ask for males in USA but both Sifo and Synovate Sweden ask for females.

5 Post Stratification and Weighting

Suppose we are interested in some variable Y. We have conducted n telephone interviews and have recorded values $S = \{y_1, y_2, \ldots, y_n\}$. Suppose we can post stratify this sample into different subgroups S_i:s with $i = 1, 2, \ldots, k$ and where $S = \bigcup_{i=1}^{k} S_i$ and $\emptyset = S_{i_1} \bigcap S_{i_2}$ for any $i_1, i_2 = 1, 2, \ldots, k$. Suppose, for example, that S_{i_1} is the group of working men 15-17 years old. It is now very likely that $\#(S_{i_1})/\#(S) \leq \#(P_{i_1})/\#(P)$ where P is the total population and P_{i_1} is the subset of people described by the characteristics in S_{i_1} in the whole population. This leads to under representation for strata S_{i_1} and over representation for example for S_{i_2}, the group of non working women aged 65-74 years old. This can be fixed by multiplying the estimators $\bar{y}_{s_i} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$, where $n_i = \#(S_i)$ and where $S_i = \{y_{ij}\}_{j=1}^{n_i}$, by weights w_i:s before estimating the total value of $\bar{y} = \sum_{i=1}^{k} w_i \bar{y}_{s_1}$.

5.1 Individual Weights

One possibility is to stratify even further and to choose individual weights for every respondent. This has been suggested for the MW method (section **??**). Then $\bar{y}_{s_i} = \sum_{j=1}^{n_i} w_{ij} y_{ij}$. This weighting can take the following aspects into account: 1) The size of the strata compared to the sample and the population in total. 2) The number of telephone landlines, T_{ij} connected to the household, often truncated as $\min(2, T_{ij})$. 3) Some geographic information connected to socioeconomic groups etc. 4) The number of members in the respondent's household.
Individual weights are not used in Sweden to any known extent. There is often no information available on T_{ij}. Synovate Sweden for example never asks if the household has more than one landline. If the same household is called twice, it is up to the household to report the mistake.

5.2 Methods Used in Sweden

The stratum, S_i:s, are constructed as follows:

Sifo. Sifo uses the same weights in all omnibus surveys. The stratum must therefore be very general to apply to all different kinds of questions. They use nine different age groups (15-17, 18-20, 21-23, 24-29, 30-39, 40-49, 50-64, 65-74 and 74-), gender (male/female) and working/not working (where non working includes students, unemployed and senior citizens). This implies a total of 36 different stratum. With 1 000 respondents and 36 stratum there is, in general, 28 respondents in each strata. It is, however, recommended not to have fewer than 20 persons in each strata. Therefore merging should be quite common.

Synovate Sweden. FI has a similar method but instead uses gender, age and education level.

The weights, w_i:s are constructed in such a way that the proportions among the S_i:s to S (using weights) will be the same as the proportions among the P_i:s to P.

6 The Future

The telecommunication industry is under constant development. A reasonable development is that the number of landlines will continue to decrease in favour of cell phones. The use of Voice over Internet Protocol (VoIP) such as Skype etc will probably also increase in the future. There have been discussions among the Swedish survey institute to start calling cell phones. This method is already used together with non random respondent selection using telephones.
It is also quite common to use Internet surveys instead of telephones but the recruitment to the Internet panels is sometimes made by first calling the possible respondent, whereby the selection procedure is still the same. This is done, for example, by Sifo. The total number of telephone surveys has decreased in Sweden during the 21^{st} century in favour of Internet surveys.
Note that there was a big conference held in Miami 2006, "Second International Conference on Telephone Survey Methodology (TSM II)". The first conference was held 15 years earlier. The development since then has been tremendous. The conference book (Lepkowski et.al. 2008) is highly recommended!

References

Bourque, L. B, and Fielder, E. P. (2003). *How to conduct telephone surveys*, vol. 4 of *The Survey Kit*, 2nd ed. Sage Publications, London. ISBN 0-7619-2591-0.

Casady, R. J and Lepkowski, J. M. (1999). *Telephone Sampling.* Wiley Series in probability and statistics, 3 ed, Wiley, New York, chapter 15, pp. 455–479. ISBN 0-471-15576-6.

Forsman, G. (1993). Sampling individuals within households in telephone surveys, American Statistical Association, 1429 Duke Street, Alexandria. Proceedings of the Section on Survey Research Methods, volume II.

Forsman, G. (2007). Correspondence with the author using telephone (+46 70 66 14 288) and e-mail (gosta.forsman@swipnet.se) during spring 2007.

Forsman, G. and Berg, S. (1992).Telephone interviewing and data qualiity, an overwiew and empirical study, *Technical Report LiU-MAT-R-92-2,* Matematiska institutionen, Linköongs universitet.

Forsman, G and Berg, S. (2002). Sifo's telefonbussar, *premomoria*, SIFO.

Forsman, G and Danielsson, F. (1997). Inference from Plus Digit Sampling – a Model Based Approach, University of Linköping. Unpublished.

Groves, Biemar, Lyberg, Massey, Nicholls and Waksberg. (1988.) *Telephone Survey Methodology.* Wiley series in probability and mathematical statistics, New York. ISBN 0-471-62218-4.

Lepkowski, J M, Tucker, C, Brick, J M, Leeuw, E, Japec, L, Paul, J L, Link, M W and Sangster, R L. (2008). *Advances in telephone survey methodology*, Wiley-Interscience. ISBN 0471745316, 9780471745310.

Kish, L. (1965). *Survey Sampling.* John Wiley, New York.

Petersson, O and Holmberg S. (2006). Svenska partibarometrar: En dokumentation, *Technical Report*, Studieförbundet Näringsliv och samhälle. Rapport till SNS demokratiråd.

Thörn, L (2006) Televerksamhet 2005, *SIKA Statistik Tele 2006:18*, Statens Institut fr Kommunikationsanalys. ISBN 91-89586-63-8, ISSN 1404-854X. ISSN 1650-3465.

ESTIMATION OF TOTAL USING AUXILIARY INFORMATION

Viktoras Chadyšas[1]

[1] Vilnius Gediminas Technical University, Lithuania
e-mail: viktoras.chadysas@fm.vgtu.lt

Abstract

In this paper we focus on constructions of the total estimator for rotated sampling design. Successive sampling procedure using multi-stage sampling design have been developed. The composite estimator of the total using auxiliary information and its approximate variance is constructed.

1 Introduction

Some surveys are continuous surveys using repeated samling from the finte populations. For example labour-force surveys are conducted monthly or quartely. When information from the previuos surveys is available, the estimation procedure can be combined with multi-phase sampling. The previous-phase sample data can be used as auxiliary information for estimation of the population total.
In this paper the composite estimator of the population total with use of auxiliary information using rotated sampling design is proposed.

2 Sampling rotation scheme

Now we briefly, step by step, describe the sampling rotation scheme. Suppose we have a finite population $\mathcal{U} = \{1, \ldots, N\}$ with values y_i, $i = 1, \ldots, N$.

- **Step 1**. Firstly, the sample $s_1^{(1)}$ of size $n^{(1)}$ is drawn from the population $\mathcal{U}$ according to a certain sampling design with probability $p(s_1^{(1)})$. The corresponding first and second order inclusion probabilities are denoted by $\pi_{1i}^{(1)}$, $\pi_{1ij}^{(1)}$ for $i, j \in \mathcal{U}$.

- **Step 2**. From the set $s_2^{(1)} = \mathcal{U} \setminus s_1^{(1)}$ with corresponding first and second order inclusion probabilities $\pi_{2i}^{(1)}$ and $\pi_{2ij}^{(1)}$ respectively, sample $s_2^{(2)}$ of size $m^{(2)}$ is drawn in accordance with a certain sampling design, such that $p(s_2^{(2)}|s_2^{(1)})$ is the conditional probability of selection the sample $s_2^{(2)}$. The corresponding first and second order inclusion probabilities are denoted by $\pi^{(2)}_{2i|s_2^{(1)}}$, $\pi^{(2)}_{2ij|s_2^{(1)}}$ for $i, j \in \mathcal{U}$.

- **Step 3a**. The sample $s_1^{(2)}$ of size $n^{(2)}$ is drawn from $s_1^{(1)}$ to a certain sampling design with probability $p(s_1^{(2)}|s_1^{(1)})$. The corresponding first and second order inclusion probabilities are $\pi^{(2)}_{1i|s_1^{(1)}}$, $\pi^{(2)}_{1ij|s_1^{(1)}}$ for $i, j \in \mathcal{U}$.

- **Step 3b**. The sample $s_2^{(3)}$ of size $m^{(3)}$ is drawn from $s_2^{(2)}$ accrding to a certain sampling design with probability $p(s_2^{(3)}|s_2^{(2)})$. The corresponding first and second order inclusion probabilities are $\pi^{(3)}_{2i|s_2^{(2)}}$, $\pi^{(3)}_{2ij|s_2^{(2)}}$ for $i, j \in \mathcal{U}$.

- **Step 3c**. The set $s_3^{(1)} = \mathcal{U} \setminus (s_1^{(1)} \cup s_2^{(2)})$ is considered being a sample drawn from the finite population $\mathcal{U}$ and first and second order inclusion probabilities are $\pi_{3i}^{(1)}$, $\pi_{3ij}^{(1)}$ respectively. Finally sample $s_3^{(2)}$ of size $u^{(2)}$ is drawn from $s_3^{(1)}$ accordance with a certain sampling design, so that $p(s_3^{(2)}|s_3^{(1)})$ is the conditional probability of selecting $s_3^{(2)}$. The corresponding first and second order inclusion probabilities are $\pi^{(2)}_{3i|s_3^{(1)}}$, $\pi^{(2)}_{3ij|s_3^{(1)}}$ for $i, j \in \mathcal{U}$.

Sampling rotation scheme is shown in Figure 1. It is seen in presented scheme that the whole sample s consists of the union of three samples: $s_1^{(2)}$, $s_2^{(3)}$ and $s_3^{(2)}$. Firstly, we will construct three separate estimators of the total using data of samples $s_1^{(2)}$, $s_2^{(3)}$ and $s_3^{(2)}$ respectively. Secondly we will propose composite estimator of the total using that sample rotation scheme.

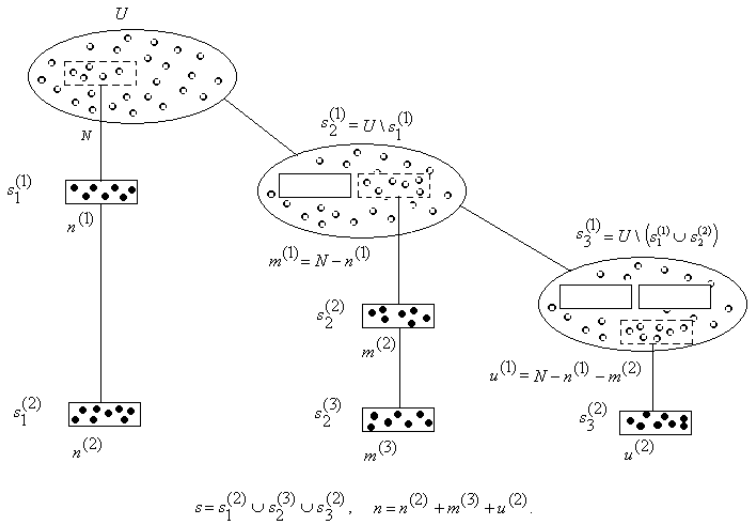


Figure 1: Sampling rotation scheme

3 Estimator of total

In sample survey, auxiliary information can be used at the estimation stage to increase the accuracy of estimators. Using proposed sampling rotation scheme we can construct estimators of total with values obtained by observing the elements in the previous phases as auxiliary information.

Case 1. The sample $s_1^{(2)}$ is obtained by a two-phase sampling:

$$\mathcal{U} \longrightarrow s_1^{(1)} \longrightarrow s_1^{(2)}.$$

In this two-phase sample design, on the first phase the sample $s_1^{(1)}$ is drawn from the population $\mathcal{U}$ and on the second-phase a matched sample $s_1^{(2)}$ is drawn from $s_1^{(1)}$. The values of the study variable on the first-phase can be used us auxiliary information. Let us denote the study variable on the first-phase by x with values x_i, $i \in s_1^{(1)}$, and the same variable on the second-phase by y with the values y_i, $i \in s_1^{(2)}$.

Using the first-phase sample $s_1^{(1)}$ and the second-phase sample $s_1^{(2)}$ we form a well known ratio estimator of the total

$$\hat{t}_{1y}^{(2)r} = \hat{t}_{1x}^{(1)} \frac{\hat{t}_{1y}^{(2)}}{\hat{t}_{1x}^{(2)}} = \hat{t}_{1x}^{(1)} \hat{r}, \tag{1}$$

where

$$\hat{t}_{1x}^{(1)} = \sum_{i \in s_1^{(1)}} \frac{x_i}{\pi_{1i}^{(1)}}, \quad \hat{t}_{1y}^{(2)} = \sum_{i \in s_1^{(2)}} \frac{y_i}{\pi_{1i}^{(1)} \pi_{1i|s_1^{(1)}}^{(2)}},$$

$$\hat{t}_{1x}^{(2)} = \sum_{i \in s_1^{(2)}} \frac{x_i}{\pi_{1i}^{(1)} \pi_{1i|s_1^{(1)}}^{(2)}}, \quad \hat{r} = \frac{\sum_{i \in s_1^{(2)}} x_i}{\sum_{i \in s_1^{(2)}} y_i}.$$

The variance of the estimator (1) has a form (Krapavickaite, Plikusas 2005, p. 248)

$$\mathrm{D}\hat{t}_{1y}^{(2)r} = \mathrm{DE}(\hat{t}_{1y}^{(2)r} | s_1^{(1)}) + \mathrm{ED}(\hat{t}_{1y}^{(2)r} | s_1^{(1)}). \tag{2}$$

The approximate variance of $\mathrm{D}(\hat{t}_{1y}^{(2)r} | s_1^{(1)})$ is given by

$$\begin{aligned} \mathrm{AD}\hat{t}_{1y}^{(2)r} &= \sum_{i=1}^{N} \sum_{j=1}^{N} (\pi_{1ij}^{(1)} - \pi_{1i}^{(1)} \pi_{1j}^{(1)}) \frac{y_i}{\pi_{1i}^{(1)}} \frac{y_j}{\pi_{1j}^{(1)}} \\ &+ \mathrm{E} \sum_{i \in s_1^{(1)}} \sum_{j \in s_1^{(1)}} (\pi_{1ij|s_1^{(1)}}^{(2)} - \pi_{1i|s_1^{(1)}}^{(2)} \pi_{1j|s_1^{(1)}}^{(2)}) \frac{R_i}{\pi_{1i}^{(1)} \pi_{1i|s_1^{(1)}}^{(2)}} \frac{R_j}{\pi_{1j}^{(1)} \pi_{1j|s_1^{(1)}}^{(2)}}, \end{aligned} \tag{3}$$

with $R_i = y_i - r x_i$, $r = \frac{\sum_{i \in U} x_i}{\sum_{i \in U} y_i}$.
The variance $\mathrm{D}\hat{t}_{1y}^{(2)r}$ is recommended to estimate by

$$\begin{aligned} \widehat{\mathrm{D}}\hat{t}_{1y}^{(2)r} &= \sum_{i \in s_1^{(2)}} \sum_{j \in s_1^{(2)}} \frac{\pi_{1ij}^{(1)} - \pi_{1i}^{(1)} \pi_{1j}^{(1)}}{\pi_{1ij}^{(1)} \pi_{1ij|s_1^{(1)}}^{(2)}} \frac{y_i}{\pi_{1i}^{(1)}} \frac{y_j}{\pi_{1j}^{(1)}} \\ &+ \sum_{i \in s_1^{(2)}} \sum_{j \in s_1^{(2)}} \frac{\pi_{1ij|s_1^{(1)}}^{(2)} - \pi_{1i|s_1^{(1)}}^{(2)} \pi_{1j|s_1^{(1)}}^{(2)}}{\pi_{1ij|s_1^{(1)}}^{(2)}} \frac{\hat{R}_i}{\pi_{1i}^{(1)} \pi_{1i|s_1^{(1)}}^{(2)}} \frac{\hat{R}_j}{\pi_{1j}^{(1)} \pi_{1j|s_1^{(1)}}^{(2)}}, \end{aligned} \tag{4}$$

with $\widehat{R}_i = y_i - \hat{r} x_i$, $\hat{r}$ given in (5).
Case 2. The sample $s_2^{(3)}$ is considered as a three-phase sample:

$$\mathcal{U} \longrightarrow s_2^{(1)} = \mathcal{U} \setminus s_1^{(1)} \longrightarrow s_2^{(2)} \longrightarrow s_2^{(3)}.$$

To extend the results to three-phase estimation, assume a third-phase sample $s_2^{(3)}$ of size $m^{(3)}$ is selected from a second-phase sample $s_2^{(2)}$ of size $m^{(2)}$, which is itself a sample of a first-phase sample $s_2^{(1)}$ of size $m^{(1)}$.
Under three-phase sampling, using idea of π^* estimators (Särndal et al. 1992, p. 347) the population total $t = \sum_{i=1}^{N} y_i$ can be unbiasedly estimated by

$$\hat{t}_{2y}^{(3)} = \sum_{i \in s_2^{(3)}} \frac{y_i}{\pi_{2i}^{(1)} \pi_{2i|s_2^{(1)}}^{(2)} \pi_{2i|s_2^{(2)}}^{(3)}}. \tag{5}$$

The expression for variance of estimator of the total in the three-phase sampling design (5) is given in (Fuller 2003, p. 311):

$$\begin{aligned}
\mathrm{D}\hat{t}_{2y}^{(3)} &= \mathrm{DE}(\hat{t}_{2y}^{(3)}|s_2^{(1)}) + \mathrm{ED}(\hat{t}_{2y}^{(3)}|s_2^{(1)}) \\
&= \mathrm{D}\hat{t}_{2y}^{(1)} + \mathrm{ED}\Big(\mathrm{E}(\hat{t}_{2y}^{(3)}|s_2^{(2)})|s_2^{(1)}\Big) + \mathrm{ED}\Big((\hat{t}_{2y}^{(3)}|s_2^{(2)})|s_2^{(1)}\Big) \\
&= \mathrm{D}\hat{t}_{2y}^{(1)} + \mathrm{ED}(\hat{t}_{2y}^{(2)}|s_2^{(1)}) + \mathrm{ED}(\hat{t}_{2y}^{(3)}|s_2^{(2)}). \qquad (6)
\end{aligned}$$

For three-phase sampling we can also form an estimator of the population total using values of the study variable on the second-phase as auxiliary information. Let us denote the study variable on the second-phase by x with values x_i, $i \in s_2^{(2)}$, and the same variable on the third-phase by y with the values y_i, $i \in s_2^{(3)}$. Using the second-phase sample $s_2^{(2)}$ and the third-phase sample $s_2^{(3)}$ we can form ratio estimator of the total

$$\hat{t}_{2y}^{(3)r} = \hat{t}_{2x}^{(2)} \frac{\hat{t}_{2y}^{(3)}}{\hat{t}_{2x}^{(3)}} = \hat{t}_{2x}^{(2)} \hat{r}, \qquad (7)$$

where

$$\hat{t}_{2x}^{(2)} = \sum_{i \in s_2^{(2)}} \frac{x_i}{\pi_{2i}^{(1)} \pi_{2i|s_2^{(1)}}^{(2)}}, \quad \hat{t}_{2y}^{(3)} = \sum_{i \in s_2^{(3)}} \frac{y_i}{\pi_{2i}^{(1)} \pi_{2i|s_2^{(1)}}^{(2)} \pi_{2i|s_2^{(2)}}^{(3)}},$$

$$\hat{t}_{2x}^{(3)} = \sum_{i \in s_1^{(2)}} \frac{x_i}{\pi_{2i}^{(1)} \pi_{2i|s_2^{(1)}}^{(2)} \pi_{2i|s_2^{(2)}}^{(3)}}, \quad \hat{r} = \frac{\sum_{i \in s_2^{(3)}} x_i}{\sum_{i \in s_2^{(3)}} y_i}.$$

The approximate variance is given by

$$\begin{aligned}
\mathrm{AD}\hat{t}_{2y}^{(3)r} &= \sum_{i=1}^{N} \sum_{j=1}^{N} (\pi_{2ij}^{(1)} - \pi_{2i}^{(1)} \pi_{2j}^{(1)}) \frac{y_i}{\pi_{2i}^{(1)}} \frac{y_j}{\pi_{2j}^{(1)}} \\
&+ \mathrm{E} \sum_{i \in s_2^{(1)}} \sum_{j \in s_2^{(1)}} (\pi_{2ij|s_2^{(1)}}^{(2)} - \pi_{2i|s_2^{(1)}}^{(2)} \pi_{2j|s_2^{(1)}}^{(2)}) \frac{y_i}{\pi_{2i}^{(1)} \pi_{2i|s_2^{(1)}}^{(2)}} \frac{y_j}{\pi_{2j}^{(1)} \pi_{2j|s_2^{(1)}}^{(2)}} \\
&+ \mathrm{E} \sum_{i \in s_2^{(2)}} \sum_{j \in s_2^{(2)}} (\pi_{2ij|s_2^{(2)}}^{(3)} - \pi_{2i|s_2^{(2)}}^{(3)} \pi_{2j|s_2^{(2)}}^{(3)}) \frac{R_i}{\pi_{2i}^{(1)} \pi_{2i|s_2^{(1)}}^{(2)} \pi_{2i|s_2^{(2)}}^{(3)}} \frac{R_j}{\pi_{2j}^{(1)} \pi_{2j|s_2^{(1)}}^{(2)} \pi_{2j|s_2^{(2)}}^{(3)}}. \qquad (8)
\end{aligned}$$

The variance is recommended to estimate by

$$\begin{aligned}
\widehat{\mathrm{D}}\hat{t}_{2y}^{(3)r} &= \sum_{i \in s_2^{(3)}} \sum_{j \in s_2^{(3)}} \frac{\pi_{2ij}^{(1)} - \pi_{2i}^{(1)} \pi_{2j}^{(1)}}{\pi_{2ij}^{(1)} \pi_{2ij|s_2^{(1)}}^{(2)} \pi_{2ij|s_2^{(2)}}^{(3)}} \frac{y_i}{\pi_{2i}^{(1)}} \frac{y_j}{\pi_{2j}^{(1)}} \\
&+ \sum_{i \in s_2^{(3)}} \sum_{j \in s_2^{(3)}} \frac{\pi_{2ij|s_2^{(1)}}^{(2)} - \pi_{2i|s_2^{(1)}}^{(2)} \pi_{2j|s_2^{(1)}}^{(2)}}{\pi_{2ij|s_2^{(1)}}^{(2)} \pi_{2ij|s_2^{(2)}}^{(3)}} \frac{y_i}{\pi_{2i}^{(1)} \pi_{2i|s_2^{(1)}}^{(2)}} \frac{y_j}{\pi_{2j}^{(1)} \pi_{2j|s_2^{(1)}}^{(2)}} \\
&+ \sum_{i \in s_2^{(3)}} \sum_{j \in s_2^{(3)}} \frac{\pi_{2ij|s_2^{(2)}}^{(3)} - \pi_{2i|s_2^{(2)}}^{(3)} \pi_{2j|s_2^{(2)}}^{(3)}}{\pi_{2ij|s_2^{(2)}}^{(3)}} \frac{\widehat{R_i}}{\pi_{2i}^{(1)} \pi_{2i|s_2^{(1)}}^{(2)} \pi_{2i|s_2^{(2)}}^{(3)}} \frac{\widehat{R_j}}{\pi_{2j}^{(1)} \pi_{2j|s_2^{(1)}}^{(2)} \pi_{2j|s_2^{(2)}}^{(3)}}, \qquad (9)
\end{aligned}$$

with $\widehat{R_i} = y_i - \hat{r} x_i$, $\hat{r}$ given in (7).

Case 3. The sample $s_3^{(2)}$ is considered as a two-phase sample:

$$\mathcal{U} \longrightarrow s_3^{(1)} = \mathcal{U} \setminus (s_1^{(1)} \cup s_2^{(2)}) \longrightarrow s_3^{(2)}.$$

In this case we have not auxiliary information for elements in the first-phase sample $s_3^{(1)}$. In two-phase sampling, the population total $t = \sum_{i=1}^{N} y_i$ is estimated unbiasedly (Särndal et al. 1992, p. 348) by the estimator

$$\hat{t}_{3y}^{(2)} = \sum_{i \in s_3^{(2)}} \frac{y_i}{\pi_{3i}^{(1)c} \pi_{3i|s_3^{(1)}}^{(2)}}. \tag{10}$$

Turning to the variance, we have

$$\mathrm{D}\hat{t}_{3y}^{(2)} = \mathrm{DE}(\hat{t}_{3y}^{(2)}|s_3^{(1)}) + \mathrm{ED}(\hat{t}_{3y}^{(2)}|s_3^{(1)}), \tag{11}$$

or

$$\begin{aligned} \mathrm{D}\hat{t}_{3y}^{(2)} &= \sum_{i=1}^{N}\sum_{j=1}^{N}(\pi_{3ij}^{(1)} - \pi_{3i}^{(1)}\pi_{3j}^{(1)})\frac{y_i}{\pi_{3i}^{(1)}}\frac{y_j}{\pi_{3j}^{(1)}} \\ &+ \mathrm{E}\sum_{i \in s_3^{(1)}}\sum_{j \in s_3^{(1)}}(\pi_{3ij|s_3^{(1)}}^{(2)} - \pi_{3i|s_3^{(1)}}^{(2)}\pi_{3j|s_3^{(1)}}^{(2)})\frac{y_i}{\pi_{3i}^{(1)}\pi_{3i|s_3^{(1)}}^{(2)}}\frac{y_j}{\pi_{3j}^{(1)}\pi_{3j|s_3^{(1)}}^{(2)}}. \end{aligned} \tag{12}$$

The variance is estimated unbiasedly by

$$\begin{aligned} \widehat{\mathrm{D}}\hat{t}_{3y}^{(2)} &= \sum_{i \in s_3^{(2)}}\sum_{j \in s_3^{(2)}}\frac{\pi_{3ij}^{(1)} - \pi_{3i}^{(1)}\pi_{3j}^{(1)}}{\pi_{3ij}^{(1)}\pi_{3ij|s_3^{(1)}}^{(2)}}\frac{y_i}{\pi_{3i}^{(1)}}\frac{y_j}{\pi_{3j}^{(1)}} \\ &+ \sum_{i \in s_3^{(2)}}\sum_{j \in s_3^{(2)}}\frac{\pi_{3ij|s_3^{(1)}}^{(2)} - \pi_{3i|s_3^{(1)}}^{(2)}\pi_{3j|s_3^{(1)}}^{(2)}}{\pi_{3ij|s_3^{(1)}}^{(2)}}\frac{y_i}{\pi_{3i}^{(1)}\pi_{3i|s_3^{(1)}}^{(2)}}\frac{y_j}{\pi_{3j}^{(1)}\pi_{3j|s_3^{(1)}}^{(2)}}. \end{aligned} \tag{13}$$

By linear combination of estimators (1), (7) and (10) we obtain a new composite estimator of total of study variable y

$$\hat{t}_y = \alpha\hat{t}_{1y}^{(2)r} + \beta\hat{t}_{2y}^{(3)r} + \gamma\hat{t}_{3y}^{(2)}, \tag{14}$$

where α, β and γ are constants ($0 < \alpha < 1, 0 < \beta < 1, 0 < \gamma < 1$), satisfying $\alpha + \beta + \gamma = 1$. An approximate variance of the composite estimator (14) is expressed by

$$\begin{aligned} \mathrm{AD}\hat{t}_y &= \alpha^2\mathrm{AD}\hat{t}_{1y}^{(2)r} + \beta^2\mathrm{AD}\hat{t}_{2y}^{(3)r} + \gamma^2\mathrm{D}\hat{t}_{3y}^{(2)} \\ &+ 2\alpha\beta\mathrm{cov}(\hat{t}_{1y}^{(2)r}, \hat{t}_{2y}^{(3)r}) + 2\alpha\gamma\mathrm{cov}(\hat{t}_{1y}^{(2)r}, \hat{t}_{3y}^{(2)}) + 2\beta\gamma\mathrm{cov}(\hat{t}_{2y}^{(3)r}, \hat{t}_{3y}^{(2)}). \end{aligned} \tag{15}$$

$\mathrm{AD}\hat{t}_{1y}^{(2)r}$, $\mathrm{AD}\hat{t}_{2y}^{(3)r}$ and $\mathrm{D}\hat{t}_{3y}^{(2)}$ given in (3), (8) and (12) respectively.

4 Conclusions

Simulation is still in progress. The results will be presented at the summer school.

References

Fuller, A. W. (2003) *Estimation for Multiple Phase Samples. In: Analysis of Survey Data, eds. Chambers R. L. and Skinner J. L.* John Wiley & Sons, Chichester.

Krapavickaite, D., Plikusas A. (2005) *Imciu teorijos pgrindai (Basics of sample theory).* Technika, Vilnius.

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

APPLICATION OF SURVEY SAMPLING METHODS TO MARKET RESEARCH

Oleksandr Chernyak[1] and Valentyn Nebukin[2]

[1] Kyiv National Taras Shevchenko University, Ukraine
e-mail: chernyak@univ.kiev.ua

[2] Kyiv National Taras Shevchenko University, Ukraine
e-mail: v_nebukin@univ.kiev.ua

Abstract

Main characteristics of different kinds of market research are described in detail. The area of sampling methods usage in market research is defined. Based on 3 real researches of Ukrainian markets, the effectiveness of the proposed technique of sampling methods usage in market research is confirmed.

1 Introduction

Under present conditions of accelerating globalization processes and strengthening world economic crisis appearances, the competitive activity has sharpened much in all industries of Ukrainian economy. It forces companies' management to make more efforts to increase the level of customer demands satisfaction by their goods and services. In order to provide the required effectiveness of managerial decisions companies need more accurate and comprehensive information on tastes and incomes of their consumers and also information on characteristics of their economic environment, on the basis on which marketing management decisions are made. Such information can be obtained through marketing research, the crucial part of which is made up by research of markets. Nowadays most companies have to carry out or order more research of markets, which they are going to entry or on which they are working, because the price of mistake in the marketing plan is too high when markets are falling down.

It's obvious that the main criteria for choosing the method of research are its accuracy and its costs. Survey sampling methods in contrast to other market research methods provide the possibility to measure their accuracy and to assess the needed expenses, which makes the choice of research method rather rational. And, as survey sampling methods of research can significantly decrease the cost obtaining the information on customers, competitors and other components of a company's environment, they are becoming especially relevant under such conditions.

Theory of market research is presently rather mature (see Churchill (1995) and Malhotra (2002)) as well as survey sampling theory is (see Lohr (1999), Särndal, Swensson, and Wretman (1992), Thompson (1997), and Thompson (1992)); but different survey designs and different methods of estimation show different effectiveness when applied to the research of different markets. That is why the issues of effective application of one or another sampling method to market research of one or another kind on one or another market require separate consideration. Thus applied to research of transition economy markets, which are notable for a high unpredictability and among which Ukraine is, survey sampling methods require practical testing and need working out the recommendations on certain cases of one or another method of application. All these issues are to be addressed in the paper.

2 Market research classification

Market research, as defined by Malhotra (2004), is a set of systematic and objective identification, collection, analysis, and dissemination of information for the purpose of assisting management in decision making related to the identification and solution of problems and opportunities in marketing. Another definition of market reasearch, in which emphasis is placed on its informational and communicating role in the business run, is given by Churchill (1995): **market research** is a function, which connects a company with a customer through the information, that is used in order to define marketing problems and opportunities; to develop, specify, evaluate and control marketing activities; to improve understanding the marketing management process.
As follows from the definitions the **process of market research** consists of 4 steps: definition of a problem and research goals, research plan design, research plan implementation, data analysis and results presentation.
In order to understand the area of survey sampling methods application in research of markets we are going to give various classifications of market research in this section.
Firstly, subject to the goal of research, market research is classified into:

- problem identification research,
- problem solving research.

Problem identification market research is more widespread (see Malhotra (2004)). Research of this type provides information about the marketing environment and helps to diagnose a problem. The recognition of economic, social, or cultural trends, such as changes in consumer behavior, may point to underlying problems or opportunities. For example, if the market potential is increasing, but the firm is losing market share, it means that the firm has its own specific problems. Otherwise, a declining market potential indicates that all the firms are likely to have a problem achieving their growth targets. Examples of problem identification research include market potential, market share, brand or company image, market characteristics, sales analysis, forecasting, and business trends research.
Once a problem or opportunity has been identified **problem solving market research** is undertaken by most companies to arrive at a solution. The findings of problem solving research are used in making decisions which are to solve specific marketing problems. Examples of problem solving research include market segmentation, goods, competitors' pricing policy, advertisement, and production distribution research.
Secondly, all market research projects can be divided into 2 basic types:

- exploratory research,
- conclusive research.

Exploratory research is a type of market research, which is intended for preliminary data collection, main ideas formulation, obtaining deep insight into the essence of the problem to solve, and hypothesis proposition. Exploratory research is to be done when a company's management recognizes the existence of a problem, but has no idea about its sources. As the informational need doesn't have some certain form at this stage, exploratory research should be flexible and shouldn't be bound by any structure. This type of market research usually fulfills the following tasks:

- to formulate the problem and define its sources,
- to define alternative actions,
- to propose hypotheses,
- to choose key variables and main relations for the following research,
- to obtain some basic idea about the solution of a problem.

Exploratory research is usually conducted with a small size sample at the preliminary stages of research project, so its results shouldn't be generalized; they only constitute one of possbile variants, which can be confirmed or disproved by the following conclusive research.
Conclusive research is intended to help the manager in defining, evaluating, and choosing optimal action variants to solve the problem. Conclusive market research can be used to verify the results of exploratory research. The aim of research of this type is to verify the

proposed hypotheses and to test some certain correlations. It is usually more formal and structured than exploratory research. Conclusive research draws conclusions: the results of the study can be generalized to the whole population.

Conclusive research includes descriptive, causal and descriptive-causal types of research. **Descriptive market research** is a type of conclusive research, which is aimed at description of marketing problems and opportunities, situations and markets through numerical values of their characteristics. Most market research is descriptive by nature. This type of market research usually focuses on product performance, market size, trends, competitive strategies and market share. Descriptive research presumes the existence of some a-priori knowledge about the problem. Descriptive research is based on clear understanding of a problem, certain hypotheses already formulated, knowledge about the nature of information that is to be collected during the research, and also on large representative samples. The results of descriptive market research are used in order to obtain general conclusions about some customer group or the whole market.

Subject to its design descriptive market research can be classified into cross-sectional research and longitudinal research. **Cross-sectional research** is a type of descriptive research, with which data is collected at one point in time on several variables. This type of research should be considered as a market configuration snapshot that characterizes a certain moment of time. In cross-sectional research a group of respondents is studied only once. Cross-sectional research is the most widely used type of market research (see Malhotra (2004)).

Longitudinal research is a type of descriptive research, with which data on one or several variables is collected over certain period of time. This type of research represents a set of market configuration snapshots, which enables analysis of changes that occurred in time. In longitudinal research a fixed group of respondents is studied regularly. When changes of market characteristic are revealed from a set of cross-sectional researches, conducted several times, they could be caused by occasional nature of a sample. That is why if research is aimed at exploring the dynamics of demand and other market characteristics, longitudinal design of research is preferable.

Causal market research is a type of conclusive research, the main aim of which is to find cause-effect relations between marketing variables. The essence of this type of research consists in testing the existence of some certain functional relation between a dependent marketing variable and a set of independent marketing variables. Causal market research can be conducted in the form of cross-sectional research and in the form of longitudinal research as well. When cross-sectional design is used, causal research helps to find out the relation between marketing variables that exists at a certain moment of time. Causal research with longitudinal design allows to figure out how the relations between marketing variables are changing over some period of time.

Descriptive-causal market research is a combination of descriptive research and causal research and, thus, is aimed at quantitative description of certain market characteristics and the same time serves for defining the existence of functional relations between them. Descriptive-causal research is always based on data collected via survey or observation. As well as descriptive research descriptive-causal research can have cross-sectional and longitudinal design. By means of **cross-sectional descriptive-causal research** current values of market characteristics as well as cause-effect links between them are studied. And with the use of **longitudinal descriptive-causal research** it is possible to investigate and model changes of market characteristics values in time and also time changes in functional relations between these variables.

Exploratory research, descriptive research, and causal research often complement each other in a single research project (see Churchill (1995)):

- exploratory research serves to formalize research aims and propose the hypotheses;
- with help of descriptive research market parameters are provided with quantitative characteristics;

– causal research gives an opportunity to reveal cause-effect relations between characteristics.

3 Efficiency of sampling methods application to market research

Taking into consideration the essence of **survey sampling theory** (see Lohr (1999), Särndal, Swensson and Wretman (1992)):

– methods of element sampling design for collecting the data on values of element characteristics;
– methods for constructing the estimates of these characteristics values over the whole population (total value, mean value, portion value, etc.) and confidence intervals of these values on the of basis data, collected on sampled elements

– we can infer that the main area of their application to market research is **descriptive research**. When it is needed to measure the current size of a market or any of its segment, sales volume of competitors, company's market share, etc., usage of sampling methods is very reasonable. Sampling methods can also be used for short-term forecasting of the market size, when under the research respondents are asked not about how many items of a certain good they usually purchase during a certain time period, but about how many items they are going to buy during the next time period. Data needed for defining the population, the sampling frame, and the sampling design, when doing descriptive research with help sampling methods, is often available from the previously published secondary data. These secondary data usually can also help to make effective stratification of the population in order mark out potentially homogeneous groups. Otherwise the needed information could be obtained by an exploratory research. Anyway, investigation of a small part of market elements, sampled from the whole population of market elements, costs much lower than investigation of the whole population. At the same time the obtained estimates of the studied market characteristic have a known level of accuracy and a comparable value for the company, conducting a research.

Survey sampling methods can be also widely-used in **causal research** and **descriptive-causal research**, when under the research project data is collected via survey or observation. Application of survey sampling methods to causal market research is bound to methods of element sampling design: the data, based on which a search for cause-effect relations is carried out, is collected by studying sampled elements. As far as descriptive-causal market research is concerned both sampling design methods and methods of estimation can be used for conducting this type of market research.

As descriptive, causal and descriptive-causal research can have as cross-sectional as longitudinal design from one side and can be aimed as at problem identification as problem solving, sampling methods theory could be applied to:

– problem identification research and problem solving research;
– cross-sectional research and longitudinal research;

– subject to the research project is not exploratory.

Further, in this section we will consider some research projects of Ukrainian markets that have been conducted with help of survey sampling methods – in order to prove the power of survey sampling methods application to market research. First attempt to apply sampling methods to research of Ukrainian markets was made by Chernyak (2001): in the research of Ukrainian banking market the total capital of Ukrainian banking system was successfully measured using survey sampling methods. These attempt was continued by Zatonatska and Nebukin (2002), Vasechko, Chernyak, Zhuykova, and Nebukin (2003), and Nebukin (2005), whose papers we will consider in more detail.

3.1 Estimation of Ukrainian potential market of tobacco products

In the paper by Zatonatska and Nebukin (2002) **survey sampling methods** were applied to **descriptive cross-sectional research** of Ukrainian **market** of tobacco products, the aim of

which was to estimate a monthly potential market size. The research was based on data, obtained as a result of stratified random sampling potential tobacco products consumers from the population of Ukraine: stratification criteria consisted of potential consumer's sex and home region. Using classical estimators of stratified random sampling (see Lohr (1999)) a monthly potential market size (for whole market and for each brand) was estimated for the next month (September) of 2001 and confidence intervals (p = 0.99) were calculated. The success of the research was confirmed by actual observations of tobacco products sales volume in Ukraine for two previous (up to September of 2001) months: actual sales volumes for July of 2001 and for August of 2001 were approximately 10% and 5% respectively lower than September volume estimate, obtained by this market research, and its confidence interval contained both of these actually observed values.

3.2 Estimation of output and sales volume of Ukrainian small businesses

In the paper by Vasechko, Chernyak, Zhuykova, and Nebukin (2003) a set of **survey sampling methods** was applied to **descriptive cross-sectional research** of Ukrainian organizations **market segment** – small enterprises, the subject of which was their mean and total values of yearly sales volume and output volume for 2001.
Using simple random sampling and stratified random sampling, samples of different sizes (1%, 10%, and 20%) were obtained form the population. Under stratified sampling the criterion of stratification was represented by a company's size indicator, which was measured by the number of employees. On the basis of sampled data from samples of different sizes, mean and total values of small enterprises yearly sales volume and output volume for 2001 were estimated, using 8 types of estimators:

- classical estimators of simple and stratified random sampling (see Thompson (1992)),
- ratio estimator and regression estimator of simple random sampling (see Thompson (1992)),
- separate and combined ratio estimators and regression estimators of stratified random sampling (see Särndal, Swensson, and Wretman (1992)).

An independent variable used in ratio estimators and regression estimators was a company's size indicator, measured by the number of employees.
As at the moment of paper publication the information on true values of the investigated characteristics had become available, it has given the authors an opportunity to evaluate each combination of sampling design method and method of estimation by accuracy of estimation criterion (relative estimation error) and to formulate recommendations for choosing sampling design methods and methods of estimation, when conducting similar research in the future:

- With samples of small size (1%) estimators of stratified random sampling are preferable: classical estimator or separate ratio estimator – for output volume estimation and combined or separate ratio estimators – for sales volume estimation; as the same data is necessary for calculating the estimators of both types, it is possible to use these estimators of different types in one research.
- With samples of middle size (10%) classical estimator of stratified random sampling is preferable for output volume estimation as well as for sales volume estimation; irrespective of the fact, that combined and separate regression estimators of stratified random sampling show a better accuracy of estimation, it's not recommended to use them, because R^2 in the relevant regression models is too low and high accuracy of estimation is more likely to be occasional.
- With samples of large size (20%) classical estimators of simple random sampling are preferable: they have relative estimation error at the level of 0.54% for output volume estimation and 0.79% for sales volume estimation and the same time they don't require data collection on supplementary characteristic (company's number of employees).

– We should mention that classical estimators of stratified random sampling show high accuracy of estimation with samples of any size when estimating output volume as well as sales volume (the maximum value of relative estimation error is 3.25%).

3.3 Estimation of Ukrainian pharmaceutical market size

In the paper by Nebukin (2005) **survey sampling methods** were applied to **descriptive cross-sectional research** of Ukrainian pharmaceutical **market**, as a result of which on the basis of sampled data on yearly sales volume of medicines in Ukraine for 2003 and 2004 estimates of yearly pharmaceutical market size of Ukraine for 2004 were obtained.

By means of simple random sampling and stratified random sampling samples of different sizes (5%, 10%, and 20%) were obtained from the population. When doing stratified sampling, yearly sales volume of medicines in Ukraine for 2003 was used as stratification criterion: the results of sampled data variance analysis of sales volume of medicines in Ukraine for 2004 confirmed the correctness choosing the stratification criterion.

Based on sampled data on sales volume of medicines in Ukraine for 2003 and 2004 from samples of different sizes, 8 types of estimates of Ukrainian pharmaceutical market yearly size for 2004 were calculated:

- classical estimates of simple and stratified random sampling (see Thompson (1992)),
- ratio based estimate and regression based estimate of simple random sampling (see Thompson (1992)),
- separate and combined ratio based estimates and regression based estimates of stratified random sampling (see Särndal, Swensson, and Wretman (1992)).

An independent variable, used to calculate ratio and regression based estimates, was sales volume of medicines in Ukraine for 2003.

As at the moment of paper publication the information on true values of Ukrainian pharmaceutical market yearly size for 2004 were already available, it has made possible to evaluate each combination of sampling design method and method of estimation by accuracy of estimation criterion (relative estimation error):

- with samples of any size the best performance is demonstrated by estimates of stratified random sampling, among which ratio and regression based estimates show the highest level of accuracy;
- with samples of small size (5%) and middle size (10%) the highest level of accuracy is demonstrated by combined ratio based estimate of stratified random sampling;
- with samples of large size (20%) combined regression based estimate of stratified random sampling is the best-performing.

This also allowed to formulate recommendations for choosing sampling design methods and methods of estimation, when conducting similar research in the future:

- with samples of small size (5%) and middle size (10%) combined ratio based estimate of stratified random sampling is preferable;
- increase of sample size up to 20% is excessive, because it doesn't improve the accuracy of estimation; so if it is cost-effective to increase sample size from 5%, it's optimal to stop at sample size of 10%.

4 Conclusion

Whereas from theoretical standpoint the main area of application **survey sampling methods** to **market research** turns out to be **descriptive market research**, success of descriptive market research projects, discussed in Section 3, proves its effectiveness on practice. However, to make this application really efficient it is necessary to choose an optimal combination of sample size, survey design method and method of estimation, which is expected to be unique for every single market during some period of time. That is why periodical conduction of research projects, structured like the ones, discussed in the Subsections 3.2 and 3.3, which make this choice rational, is very useful on the target markets.

References

Chernyak O. I. (2001) The sampling strategy for banking survey in Ukraine. *Theory of Stochastic Processes*, **7 (23)**, **1-2**, 45-52.

Churchill, G. A. (1995) *Marketing research: methodological foundations, 6th edition*. South-Western/Thomson Learning, Mason, Ohio.

Lohr S.L. (1999) *Sampling: Design and analysis.* Duxbury Press, New York.

Malhotra, N. K. (2004) *Marketing research: an applied orientation, 4th edition,* Prentice-Hall International, London.

Nebukin V.O. (2005) Estimation of Ukrainian Pharmaceutical Market Size with Survey Sampling Methods, *Bulletin of Lviv State Finance Academy. Economics*, **10**, 324-332. (in Ukrainian).

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

Thompson M.E. (1997) *Theory of sample surveys.* Chapman & Hall, London.

Thompson S.K. (1992) *Sampling.* John Wiley & Sons, New York.

Vasechko O.O., Chernyak O.I., Zhuykova E.M., Nebukin V.O. (2003) Application of Sample Survey Technique for Estimation of Small Enterprises Output and Sales Volume. In: *Statistics of Ukraine*, **3**. 4-9. (in Ukrainian).

Zatonatska T.G., Nebukin V.O. (2002) Application of Statistical Methods to Potential Market Estimation, *Bulletin of Kyiv National Taras Shevchenko University. Economics,* **58-59**, 26-30. (in Ukrainian).

Sampling of the enterprises for the purpose of the wages analysis

Chystsenka K. [1]

[1] National statistical committee of Belarus
e-mail: chistenko@gmail.com

Abstract

The mechanism and results of sampling in Belarus, in particular survey of distribution of number working on wages size are considered.

Development of various patterns of ownership, increasing of a fraction of a private sector of economy, expansion of the rights of the enterprises in an establishment of systems of wages and questions of allocation of additional expenses on the labor maintenance have led to considerable changes in conditions of a payment working in various economic branches. First of all these changes were showed in increasing of differentiation of number working by wages and redistribution of economic branches by the monthly average wages.

Observable tendencies in payment sphere have caused necessity of the organization of new methods of statistical wage research, in particular carrying out of some samples of wages of working and expenses for a labor, allowing to receive a detailed information about expenses of employers on a labor, for example, such as distribution of number working by wages size and structure of expenses of the enterprises on a labor.

Sampling of distribution of number working by the monthly average wages takes a prime place and allows to estimate differentiation of a payment in branches and regions on the basis of set of following indicators: structures of number occupied on the sizes of wages, coefficients of a parity of the sizes of average wages working in 10%-s' groups of the enterprises with the greatest and least level of wages, indicators of distribution of a total sum of the means directed on a payment on 10%-s' groups of working, for which wages below living wage size and below the minimum wages etc.

Annually since 2007 sampling of wages in Belarus, in particular survey of distribution of number working by the monthly average wages for May is carried out. The purpose of the given supervision is reception of the information characterizing differentiation of a payment in branches of economy and regions of Belarus. Objects of statistical supervision are legal enterprises and their isolated divisions with average number working more than 16 persons of all patterns of ownership of all economic branches, except for small enterprises of not state pattern of ownership. Formation of sample is made on the basis of a file of the enterprises representing the state statistical report «The Report on work and movement of the working».

By means of the software of multivariate sample the author in March 2009 on the basis of a database for February of current year has executed sample of distribution of number working by the monthly average wages working in seven regions and 60 economic branches.

Sampling of the enterprises was carried out within the outlined general aggregate in the chosen branch and region simultaneously for two indicators:

– average number working (G012);

– a wage fund of workers (G022), it is basic variable.

The mechanism of sampling of the enterprises for i-th branch and j-th area by means of the program «Multivariate sampling» includes some stages:

1. Installation of parameters on which sample of enterprises will be made. Such as:

1.1. Region and branch;

1.2. Sampling variables including the basic indicator on which the general aggregate if necessary breaks into groups for decreasing of sampling error;

1.3. A sampling fraction;

1.4. An admissible sampling error (10%);

1.5. Kind and method of sampling. Simple or optimal stratification and casual or mechanical method is used at unvariate sample. Cluster analysis is used at multivariate sample;

1.6. Number and borders of groups (for the chosen basic indicator; for multivariate sample they are established inside clusters).

2. Formation of a selective circle of the enterprises and an estimation of a divergence of indicators of sample and general aggregate on the established parameters and each of the chosen signs. Calculations are carried out under classical formulas of the theory of sample and the multidimensional statistics. Formation of variants of the sample repeats with the specified step until demanded degree of accuracy will be reached or there will be a situation at which the volume of the sample will be above its established limit, but the actual sampling error will exceed still admissible error. In this case last result of calculations can be used for realization of the subsequent procedure of sampling.

Result of sampling is independent sample enterprises lists on each region and branch which then unite in one general base for republic.

However the problem of extrapolation of sample data on a general aggregate remains even at construction of the univariate stratified sample with comprehensible by a standard error and a sampling fraction. On the basis of the received sample enterprises list distribution coefficients are calculated. Distribution coefficients (kij) of wages for extrapolation from univariate and multivariate sample on general aggregate of the enterprises are counted up as follows:

$$k_{ij} = \frac{N_{ij}}{n_{ij}}$$

here Nij – number of the enterprises in a general aggregate for i-th branch of j-th region,

nij – number of the enterprises in a sample for i-th branch of j-th region.

They were calculated individually for each organization, but are identical within group. Distribution coefficients allow to extrapolate precisely enough values of the basic variable, but deform other varying indicators of observe. It can avoid by means of formation of the multivariate sample: sampling errors in group of indicators will be much lower, than in the univariate sample, except the basic indicator. So, if at univariate sample on the basic indicator it is possible to achieve an sampling error less than 1%, and on another it can reach 50–100%, at multivariate sample an errors on all set of indicators will be in admissible limits (to 10%), but will appear considerably above, than on the basic indicator in univariate sample (approximately 4–6%).

Results of sampling testify to a coordination of indicators univariate and multivariate samples with data of general aggregate.

In the field of univariate sample an example of a coordination of data is results of sample for branch "Culture" in Minsk City: sampling coefficients of a variation in groups (34–55%) were close or more low, than in general aggregate (42–88%). Sampling errors in groups reached 10%, as a whole on average number working – 0,7% and on a wage fund of workers – 1,4%, the sampling fraction has made 17,5%.

Similar processes were observed and at multivariate sample. So, coefficients of a variation of indicators in a general aggregate were close or is considerable above, than in a sampling group in casual multivariate sample for branch "Education" in Minsk City. If in a general aggregate in cluster №1 group coefficients of a variation on average number working (G012) has made 35; 43; 41; 70% and on a wage fund (G022) – 36; 33; 24; 59%, in a sample accordingly 24; 40; 35; 50% and 41; 30; 23 and 36%. In cluster №2 group coefficients of a variation on G012 in general aggregate and a sample were equal accordingly 20; 30 and 22; 30%, on G022 – 17; 0,2 and 13; 0,2%. Sampling errors in cluster №1 has made 4,5% (G012) and 3,4% (G022), the sampling fraction – 19,6%; sampling errors in cluster №2 were equal 6,1% (G012) and 0,5% (G022), the sampling fraction – 44,4%. The general sampling error in branch "Culture" in Minsk City on average number working it was equal 5,4% and on a wage fund 1,5%.

In the course of formation of samples univariate and multivariate simple casual stratification was used because of the best sampling results.

Group errors of sample in a cut of branches and regions in the bulk fluctuated on average number working in limits from 0,1 to 6% and on a wage fund from 0,01 to 4%. In separate branches of regions sharp emissions of group errors 10 were observed, 12 and even 15 % on average number working and maximum 8% on a wage fund. However for a whole branch it has affected the general sampling error slightly, that is connected with small volume of such groups. In each of branches in seven regions the received sampling errors were in admissible borders: on a wage fund – from 0,00 to 4% and on average number working – from 0,002 to 7%. As a whole in regions owing to the big number of independent samples in branches an errors did not exceed 2%. So, on a wage fund the error was equal for the Brest region 0,37%, Vitebsk region 0,6%, Gomel region 0,19%, Grodno region 0,09%, Minsk region 0,31%, Mogilyov region 0,36%, Minsk City 0,31%, on average number working accordingly 1,75; 0,5; 0,94; 0,25; 0,17; 0,09 and 0,29%. The sampling fraction as a whole for republic has made 21,2%.

The estimated value of a wage fund as a whole for Belarus for January-February, 2009 has made 3433642,4 million rubles (actual a wage fund – 3435314,6 million rubles) and on average number working 3876617,0 persons (actual average number working – 3873377,0 persons). It accordingly on 0,05 and 0,08 % differs from data of total survey, i.e. the received samples are characterised by high degree of accuracy.

At sample carrying out there were some problems:

– in connection with complexity of computing procedures multivariate sample applied if coefficients of a variation exceed 100%, the enterprises are non-uniform on many indicators, the small size of a general aggregate (the top limit is 400–500 units, the bottom limit is 30–40 units);

– the enterprises are grouped on basic indicator, and often errors to an additional sign appear much more, than for the basic indicator. Therefore borders of groups are established taking into account the additional sign;

– stratification of small aggregate on the big number of groups (7−8) for reception of a low sampling error conducts or to general survey, or to a high sampling fraction in these groups, and, as consequence, to the general high sampling fraction in investigated aggregate;

– liquidation of the enterprises, their transition in other branch involves necessity of updating of a sample;

– for some enterprises the identical code for parent organization and its branches is incorrectly appropriated. The program for sample perceives all records in a database with identical codes within branch of economy as the uniform enterprise and summarizes on them all values on an indicator. Therefore there is a necessity of additional informing the enterprises about representation of the summary report taking into account data of branches.

The analysis of the results received for the period in 2007–2009, allows to draw following conclusions:

– the comprehensible sampling fraction of the enterprises at survey of distribution of number of workers on size of wages for region makes 20–30%, a relative sampling error for republic and regions – no more than 2%, for branches of economy – no more than 5–6%;

– simple casual stratification has the priority by efficiency at carrying out univariate sample;

– it is expedient to apply a combination of univariate and multivariate methods of sampling to leveling of lacks and use of advantages of multivariate sample;

– higher error was observed on average number working. As at stratification into groups a borders are established in more degree taking into account coefficients of a variation of a wage fund and the variation of average number working in these groups exceeds an optimum level, it is attracts increase of an sampling error for an additional indicator.

ORTHOGONAL DECOMPOSITION OF FINITE POPULATION L STATISTICS

Andrius Čiginas

Vilnius University, Lithuania
e-mail: andrius.ciginas@mif.vu.lt

Abstract

In this paper we study orthogonal decomposition of finite population L statistics. We propose quite simple form of first two terms of such decomposition.

1 Introduction

Consider the population $\mathcal{X} = \{x_1, \dots, x_N\}$ of size N and assume that $x_1 < \cdots < x_N$. Let $X_1, \dots, X_n$ is simple random sample of size $n < N$ drawn without replacement from $\mathcal{X}$ and let $X_{(1)} < \cdots < X_{(n)}$ denote the order statistics of $X_1, \dots, X_n$. Then for arbitrary real numbers $c_1, \dots, c_n$ define L-statistic $L = c_1 X_{(1)} + \cdots + c_n X_{(n)}$.
This statistic can be decomposed into the sum

$$L = \mathbf{E}L + U_1 + \cdots + U_n, \tag{1}$$

$$U_m = \sum_{1 \le i_1 < \cdots < i_m \le n} g_m(X_{i_1}, \dots, X_{i_m}), \qquad m = 1, \dots, n.$$

Decomposition (1) is called orthogonal (called also Hoeffding) decomposition and U_m are called U-statistics. The symmetric kernels $g_m, m = 1, \dots, n$ are linear combinations of conditional expectations

$$h_j(x_{k_1}, \dots, x_{k_j}) = \mathbf{E}(L - \mathbf{E}L | X_1 = x_{k_1}, \dots, X_j = x_{k_j}), \quad 1 \le j \le m. \tag{2}$$

The decompositon (1) and its applications for finite population symmetric statistics were studied by Bloznelis and Götze (2001). In that paper we can find coefficients of linear combinations of (2).
The main interest of present work is conditional expectations (2) for $j = 1, \dots, n$ and functions

$$g_1(x) = \frac{N-1}{N-n} h_1(x), \tag{3}$$

$$g_2(x,y) = \frac{N-2}{N-n} \frac{N-3}{N-n-1} \left(h_2(x,y) - \frac{N-1}{N-2} (h_1(x) + h_1(y)) \right), \tag{4}$$

which can be useful for various applications.
In the case where random variables $X_1, \dots, X_n$ are independent and identically distributed the orthogonal decomposition of L-statistics was studied by Putter and van Zwet (1998). We shall adopt some ideas from that paper to get a similar form of orthogonal decomposition. Note that for samples without replacement from finite population variables $X_1, \dots, X_n$ are identically distributed, but they are not independent.
Acknowledgement. I'm grateful to M. Bloznelis for introduction to the problem and valuable discussions.

2 Results

For the given n let fix $0 \le m \le n$ and define a set of conditions $A_m = \{X_1 = x_{k_1}, \dots, X_m = x_{k_m}\}$, where $1 \le k_1 < \cdots < k_m \le N$. Let $k_0 = 0$, $k_{m+1} = N + 1$ and define $X_{(0)} = x_0$, $X_{(n+1)} = x_{N+1}$ where $x_0 = x_1$, $x_{N+1} = x_N$. Consider statistics $X_{(r+1)} - X_{(r)}$, $r = 0, \dots, n$.

Lemma 1. *For any $m = 0, \dots, n$ and $r = 0, \dots, n$ we have*

$$\mathbf{E}(X_{(r+1)} - X_{(r)}|A_m) = \sum_{s=1}^{m+1} \sum_{i=k_{s-1}}^{k_s-1} \Delta_{m,s,i}(r)(x_{i+1} - x_i), \tag{5}$$

where we denote

$$\Delta_{m,s,i}(r) = \binom{N-m}{n-m}^{-1} \binom{i-s+1}{r-s+1} \binom{N-i-m+s-1}{n-r-m+s-1}.$$

We shall use differences $X_{(r+1)} - X_{(r)}$, $r = 0, \dots, n$ to get convenient expression of (2).

Proposition 2. *For chosen $m = 1, \dots, n$ we have*

$$\mathbf{E}(L - \mathbf{E}L|A_m) = \sum_{j=1}^{n} c_j \sum_{r=0}^{j-1} \{\mathbf{E}(X_{(r+1)} - X_{(r)}|A_m) - \mathbf{E}(X_{(r+1)} - X_{(r)})\}. \tag{6}$$

Next we shall propose simple form of kernels (3) and (4).

Theorem 3. *(i) For $1 \le k \le N$*

$$g_1(x_k) = -\sum_{j=1}^{n} c_j \sum_{i=1}^{N-1} \varphi_k(i) \frac{\binom{i-1}{j-1}\binom{N-i-1}{n-j}}{\binom{N-2}{n-1}} (x_{i+1} - x_i), \tag{7}$$

where

$$\varphi_k(i) = \begin{cases} -\frac{i}{N}, & \text{if } 1 \le i < k \\ 1 - \frac{i}{N}, & \text{if } k \le i < N. \end{cases}$$

(ii) For $1 \le k < l \le N$

$$g_2(x_k, x_l) = -\sum_{j=2}^{n} (c_j - c_{j-1}) \sum_{i=1}^{N-1} \phi_{k,l}(i) \frac{\binom{i-2}{j-2}\binom{N-i-2}{n-j}}{\binom{N-4}{n-2}} (x_{i+1} - x_i), \tag{8}$$

where

$$\phi_{k,l}(i) = \begin{cases} \frac{i(i-1)}{(N-1)(N-2)}, & \text{if } 1 \le i < k \\ -\frac{(i-1)(N-i-1)}{(N-1)(N-2)}, & \text{if } k \le i < l \\ \frac{(N-i)(N-i-1)}{(N-1)(N-2)}, & \text{if } l \le i < N. \end{cases}$$

3 Proofs

Proof of lemma 1. For any $m = 0, \dots, n$ and $r = 0, \dots, n+1$ straightforward calculations give

$$\begin{aligned} \mathbf{E}(X_{(r)}|A_m) &= \binom{N-m}{n-m}^{-1} \Big[\sum_{s=1}^{m+1} \sum_{i=k_{s-1}+1}^{k_s-1} \binom{i-s}{r-s} \binom{N-i-m+s-1}{n-r-m+s-1} x_i \\ &+ \sum_{s=0}^{m+1} \binom{k_s-s}{r-s} \binom{N-k_s-m+s}{n-r-m+s} x_{k_s} \Big]. \end{aligned}$$

For $r = 0, \dots, n$ write

$$\begin{aligned} \mathbf{E}(X_{(r+1)}|A_m) &= \binom{N-m}{n-m}^{-1} \Big[\sum_{s=1}^{m+1} \sum_{i=k_{s-1}+1}^{k_s-1} \binom{i-s}{r-s+1} \delta'_{m,s,i}(r) x_i \\ &+ \sum_{s=0}^{m+1} \binom{k_s-s}{r-s+1} \binom{N-k_s-m+s}{n-r-m+s-1} x_{k_s} \Big], \end{aligned}$$

where

$$\delta^{'}_{m,s,i}(r) = \binom{N-i-m+s}{n-r-m+s-1} - \binom{N-i-m+s-1}{n-r-m+s-1}$$

and

$$\begin{aligned}\mathbf{E}(X_{(r)}|A_m) &= \binom{N-m}{n-m}^{-1}\Big[\sum_{s=1}^{m+1}\sum_{i=k_{s-1}+1}^{k_s-1}\delta^{''}_{m,s,i}(r)\binom{N-i-m+s-1}{n-r-m+s-1}x_i \\ &+ \sum_{s=0}^{m+1}\binom{k_s-s}{r-s}\binom{N-k_s-m+s}{n-r-m+s}x_{k_s}\Big],\end{aligned}$$

where

$$\delta^{''}_{m,s,i}(r) = \binom{i-s+1}{r-s+1} - \binom{i-s}{r-s+1}.$$

Then it is easy to see, that for $r = 0, \ldots, n$, $\mathbf{E}(X_{(r+1)} - X_{(r)}|A_m)$ is the same as in lemma's statement.

□

Proof of proposition 2. Applying summation by parts we can write

$$L = \sum_{r=1}^{n-1}\alpha_r(X_{(r+1)} - X_{(r)}) + \bar{c}\sum_{j=1}^{n}X_j,$$

where $\alpha_r = -\sum_{j=1}^{r}(c_j - \bar{c})$ for $r = 1, \ldots, n-1$ and $\bar{c} = \frac{1}{n}\sum_{j=1}^{n}c_j$.
Then for $m = 1, \ldots, n$ we have

$$\begin{aligned}\mathbf{E}(L - \mathbf{E}L|A_m) &= -\sum_{j=1}^{n}c_j\sum_{r=j}^{n}\{\mathbf{E}(X_{(r+1)} - X_{(r)}|A_m) - \mathbf{E}(X_{(r+1)} - X_{(r)})\} \\ &+ \bar{c}\Big[\sum_{r=0}^{n}r\{\mathbf{E}(X_{(r+1)} - X_{(r)}|A_m) - \mathbf{E}(X_{(r+1)} - X_{(r)})\} \\ &+ \frac{N-n}{N(N-m)}\sum_{s=1}^{m}\Big(\sum_{i=0}^{k_s-1}i(x_{i+1} - x_i) - \sum_{i=k_s}^{N}(N-i)(x_{i+1} - x_i)\Big)\Big].\end{aligned}$$

Note that the term in brackets vanishes, because using lemma 1 and changing order of summation, for fixed $m = 0, \ldots, n$, $s = 1, \ldots, m+1$, $i = k_{s-1}, \ldots k_s - 1$

$$\begin{aligned}\sum_{r=0}^{n}r\Delta_{m,s,i}(r) &= \sum_{r=s-1}^{n-m+s-1}(r-s+1)\Delta_{m,s,i}(r) + (s-1)\sum_{r=s-1}^{n-m+s-1}\Delta_{m,s,i}(r) \\ &= \frac{n-m}{N-m}(i-s+1) + s - 1,\end{aligned}$$

where

$$\Delta_{m,s,i}(r) = \binom{N-m}{n-m}^{-1}\binom{i-s+1}{r-s+1}\binom{N-m-(i-s+1)}{n-m-(r-s+1)}$$

and the remaining verifying is quite simple.
Applying of Vandermonde's identity completes the proof.

□

Proof of theorem 3. (i) For chosen $1 \le k \le N$ using proposition 2 for $m = 1$ and lemma 1 for $m = 0; 1$ from (3) we have

$$\begin{aligned} g_1(x_k) &= \binom{N-2}{n-1}^{-1} \sum_{j=1}^{n} c_j \sum_{r=0}^{j-1} \Big\{ \sum_{i=1}^{k-1} \frac{i}{N} \theta_{21}(i,r)(x_{i+1} - x_i) \\ &- \sum_{i=k}^{N-1} \left(1 - \frac{i}{N}\right) \theta_{22}(i,r)(x_{i+1} - x_i) \Big\}, \end{aligned}$$

where

$$\theta_{21}(i,r) = \frac{N}{i} \binom{i}{r} \left\{ \binom{N-i-1}{n-r-1} - \frac{n}{N} \binom{N-i}{n-r} \right\},$$

$$\theta_{22}(i,r) = -\frac{N}{N-i} \binom{N-i}{n-r} \left\{ \binom{i-1}{r-1} - \frac{n}{N} \binom{i}{r} \right\}.$$

It is easy to verify that $\theta_{21}(i,r) = \theta_{22}(i,r)$. Next using principle of mathematical induction it is easy to show that for every $j = 1, \ldots, n$

$$\sum_{r=0}^{j-1} \theta_{22}(i,r) = \binom{i-1}{j-1} \binom{N-i-1}{n-j}$$

and the proof of the part (i) follows.
(ii) For chosen $1 \le k < l \le N$ using proposition 2 for $m = 1; 2$ and lemma 1 for $m = 0; 1; 2$ from (4) we have

$$\begin{aligned} g_2(x_k, x_l) &= \binom{N-4}{n-2}^{-1} \sum_{j=1}^{n} c_j \sum_{r=0}^{j-1} \Big\{ \sum_{i=1}^{k-1} \frac{i(i-1)}{(N-1)(N-2)} \theta_{31}(i,r)(x_{i+1} - x_i) \\ &- \sum_{i=k}^{l-1} \frac{(i-1)(N-i-1)}{(N-1)(N-2)} \theta_{32}(i,r)(x_{i+1} - x_i) \\ &+ \sum_{i=l}^{N-1} \frac{(N-i)(N-i-1)}{(N-1)(N-2)} \theta_{33}(i,r)(x_{i+1} - x_i) \Big\}, \end{aligned}$$

where

$$\begin{aligned} \theta_{31}(i,r) &= \frac{(N-1)(N-2)}{i(i-1)} \binom{i}{r} \Big\{ \binom{N-i-2}{n-r-2} - 2\frac{n-1}{N-2} \binom{N-i-1}{n-r-1} \\ &+ \frac{n(n-1)}{(N-1)(N-2)} \binom{N-i}{n-r} \Big\}, \end{aligned}$$

$$\begin{aligned} \theta_{32}(i,r) &= -\frac{(N-1)(N-2)}{(i-1)(N-i-1)} \Big[\binom{i-1}{r-1} \Big\{ \binom{N-i-1}{n-r-1} - \frac{n-1}{N-2} \binom{N-i}{n-r} \Big\} \\ &- \frac{n-1}{N-2} \binom{i}{r} \Big\{ \binom{N-i-1}{n-r-1} - \frac{n}{N-1} \binom{N-i}{n-r} \Big\} \Big], \end{aligned}$$

$$\begin{aligned} \theta_{33}(i,r) &= \frac{(N-1)(N-2)}{(N-i)(N-i-1)} \binom{N-i}{n-r} \Big\{ \binom{i-2}{r-2} - 2\frac{n-1}{N-2} \binom{i-1}{r-1} \\ &+ \frac{n(n-1)}{(N-1)(N-2)} \binom{i}{r} \Big\}. \end{aligned}$$

Similarly $\theta_{31}(i,r) = \theta_{32}(i,r) = \theta_{33}(i,r)$. Next using principle of mathematical induction we can show that for every $j = 1, \ldots, n$

$$\sum_{r=0}^{j-1} \theta_{33}(i,r) = \binom{i-2}{j-1}\binom{N-i-2}{n-j-1} - \binom{i-2}{j-2}\binom{N-i-2}{n-j}.$$

Then summation by parts completes the proof of the part (ii).

□

References

Bloznelis, M., Götze, F. (2001) Orthogonal decomposition of finite population statistics and its applications to distributional asymptotics. *Ann. Statist.*, **29**, 899-917.

Putter, H., van Zwet, W. R. (1998) Empirical Edgeworth expansions for symmetric statistics. *Ann. Statist.*, **26(4)**, 1540-1569.

SURVEY SAMPLING AT UZHHOROD NATIONAL UNIVERSITY

Tetiana Fedorianych [1]

[1] Uzhhorod National University, Ukraine
e-mail: fedoryanicht@gmail.com

Abstract

The structure of studies on educational direction "Statistics" is considered. A short programs, a basic teaching material and theses on "Survey Sampling" at Uzhhorod National University are presented. Some plans and activities for the future are discussed.

1 General Information about Speciality "Statistics"

Uzhhorod National University is the main high educational institution of Transcarpathia and one of the well-known university of Ukraine. It was opened in 1945. Nowadays university contains 15 faculties. The Faculty of Mathematics is one of the oldest faculty of the University. It was opened in 1950. The Faculty of Mathematics consists of 5 departments:

- *Department of Probability Theory and Mathematical Analysis;*
- *Department of Algebra;*
- *Department of Differential Equations and Mathematical Physics;*
- *Department of Computational Mathematics;*
- *Department of Cybernetics and Applied Mathematics.*

Approximately 120 graduated students of school enter to the university every year. There are 3 directions of training within the Faculty of Mathematics.

- Students specializing in *"Mathematics"* are trained to be teachers of mathematics in the secondary school.
- Students specializing in *"Applied Mathematics"* are trained to be specialists in the field of programming.
- Students specializing in *"Statistics"* are trained to be statisticians.

"Statistics" is the newest specialization. It started to function in 2002-2003 on the base of the Department of Probability Theory and Mathematical Analysis. Tempus Tacis Joint European Project *"Improvement of Education in Statistical Applications in Economics"* coordinated by D.S.Silvestrov (Mälardalen University, Västerås, Sweden) played a very important role in development of this new specialization. Thanks to the

Project, the Mathematics Faculty of Uzhhorod National University had the opportunity to improve its material and technical basis, to computerize studies, to establish scientific contacts between lecturers of our University and those from the European Union.

Educational direction "Statistics" includes one specialization - Applied Statistics. There are 2 educational levels of this training.

- The first educational level is **Bachelor of Mathematics** (4 years of training). On this educational level students follow the educational program for Bachelor of Mathematics with some additional courses of professional choice for the speciality "Statistics".
- The second educational level is **Specialist of Statistics** with one year of training.

The students of the level "Specialist of Statistics have eight weeks of practical training in statistical institutions and banks (four weeks in the beginning and four weeks before completing their studies). They also have to take a state exam and defend a diploma thesis.
The graduated students can be employed as statisticians in different economic and financial organizations, including those of insurance, in the administration, in the banks and so on.

2 Teaching Survey Sampling

Special course "Methods of Survey Sample Theory is an additional courses of professional choice for the speciality "Statistics on the first educational level. It consists of 30 hours of lectures and 16 hours of practical lessons.
The following two books are used as a basic teaching material:

- Chernyak, A. (2001) Survey Sampling Technique, Kyiv (Ukraine).
- Parkhomenko, V. (2001) Survey Sampling Methods, Kyiv (Ukraine).

Experience of teaching this course in Kyiv National University was also taken into account.
A short program of "Methods of Survey Sampling" course are:

- Goals and methods of surveys
- General scheme of survey
- Simple random sampling with and without replacement
- Sampling with unequal probabilities
- Systematic sampling
- Stratified random samples
- Simple cluster samples
- Errors in surveys, their sources and methods of reduction.

The main objective of this course is to acquaint students with the main concepts of Survey Sample Theory and the basic types of probability sampling. On practical lessons students solve different exercises using different software.

In 2008-2009 the new courses on Survey Sampling Methods for the specialist of statistics and Survey Sampling Methods for bachelor of mathematics within the Probability Theory" specialization was introduced.

The first one consists of 14 hours of lectures and 6 hours of practical lessons. The main objective of this course is to deepen the earlier obtained knowledge. Namely, students of speciality "Statistics" study in details different sampling designs and properties of unbiased estimator for this designs.
Considerable attention is allocated to the following designs:

- Bernoulli sampling;
- Simple random sampling (with and without replacement);
- Systematic sampling;
- Poisson sampling;
- Probability proportional-to-size sampling (with and without replacement).

The second one consists of 20 hours of lectures and 8 hours of practical lessons. The main objective is to acquaint students of speciality "Mathematics" with basic concepts, terms and methods of Survey Sampling Theory.

There haven't been any theses on Survey Sampling in our university before, but this year 2 Bachelor's theses have been prepared. Students are supposed to make some exercises, to estimate population parameters and to compare obtained results using different software (Statistica, Mathematica, Excel) for some of the main probability samplings. Data published in specialized statistical collections were proposed to students for processing. But they can also use data found via the Internet. The purpose of theses is to create an interest among students to survey sampling in order to continue their research over the next year of study.

For the best results of study, we also recommend to our students to pay special attention on the organization of statistical surveys along with methods of statistical estimation during their practical training in statistical institutions.

3 Perspectives

During the period March 17 - April 16, 2009 I was invited to visit the University of Umea for studies and research of survey sampling theory and methodology. This visit was supported by the Swedish Institute and its Visby Programme through the project *"Extension of the Baltic-Nordic network on survey statistics to include Ukraine and Belarus"* coordinated by Gunnar Kulldorff.

I was able to use research facilities, including libraries with a good collection of journals and books in survey sampling, and computer network of Department of Mathematical Statistics University of Umea.

The conducted work allowed to renew educational program, to form the detailed list of the recommended literature for the theoretical and worked out tasks, for the self-practical studies and the tests for checking the level of knowledge for students of Uzhhorod National University. The detailed plan for methodical recommendations on Survey Sampling for the students specializing in Statistics at the Mathematics Faculty of Uzhhorod University was composed. As a basic teaching material the following books were used:

- Carl - Erik Sarndal, Bengt Swensson, Jan Wretman "Model Assisted Survey Sampling";
- Pascal Ardilly, Yves Tille "Sampling Methods: Exersises and Solutions".

The participation in the Workshop on Survey Sampling Theory and Methodology in Kuressaare (2008) and in Kiev (2009) provides the opportunities to study the main trends and actual directions in survey sampling. This experience will be used for developing and introducing new methods and modern techniques in the teaching of survey sampling in our university and for sharing investigation on Survey Sampling at our department.

Studying actual directions on Survey Sampling enabled to draw in the titles of the Bachelor theses on "Survey Sampling", which our 4th-year students will prepar next year.

References

Chernyak, A. (2001) *Survey Sampling Technique.* Kyiv (in Ukrainian).

Fedorianych, T and Motsa, A. (2008) Speciality "Statistics" at Uzhhorod National University. *Baltic-Nordic Workshop on Survey Sampling Theory And Methodology*, Kuressaare, Estonia, August 25-29, 73-77.

Motsa, A. (2003) Theory of Probability and Mathematical Statistics at Uzhhorod National University. *Theory of Stochastic Processes*, **3-4**, 132-144.

Parkhomenko, V. (2001) *Survey Sampling Methods.* Kyiv (in Ukrainian).

Generalized Regression and Calibration Estimators for the Domain Study

Merike Hindrikson [1]

[1] University of Tartu, Estonia
e-mail: mercs@ut.ee

Abstract

Estimation for domains is an important objective in most surveys. In this paper generalized regression and calibration estimators for the domains is presented.

1. Introduction

It is common to provide estimates for the finite population of interest as well as for a number of subpopulations, called domains. When the realized domain sample size is sufficient and auxiliary information is available, the design-based domain estimation is used. In this paper two approaches for constructing a design-based domain estimators from a fixed set of auxiliary information is presented. These approaches are generalized regression and calibration estimation.

2. Use of auxiliary information in domain estimation

Let the finite population $U = \{1,...,i...,N\}$ be divided into D non-overlapping domains $U_1,...,U_d,...U_D$. The variable of interest is the domain specific y-variable, $\boldsymbol{y}_d$, whose value for unit i is equal to y_i if $i \in U_d$ and 0 otherwise. The main parameter of interest in this paper is the domain total of the variable $\boldsymbol{y}$, $t_y^{\,d} = \boldsymbol{y}_d' \boldsymbol{1}$, where $\boldsymbol{1}$ denotes the vector of ones.
Sampling is performed by a random vector $\boldsymbol{I} = (I_1, I_2, ..., I_N)'$ called sampling vector. Behaviour of $\boldsymbol{I}$ depends on the used sampling design. Elements of $\boldsymbol{I}$ - inclusion indicators –show the number of times a population element is sampled. For without-replacement (WOR) sampling schemes the possible values of inclusion indicators are 0 and 1, for with-replacement (WR) schemes inclusion indicators may take values from 0 to n, where n is the sample size. At the estimation stage the expanded sampling vector $\breve{\boldsymbol{I}} = (\breve{I}_1, \breve{I}_2, ..., \breve{I}_N)'$, where $\breve{I}_i = I_i / E(I_i)$ is used. Note that $E(\breve{\boldsymbol{I}}) = \boldsymbol{1}$.

An estimator of the domain total $t_y^{\,d} = \boldsymbol{y}_d' \boldsymbol{1}$ that does not use auxiliary data is given by

$$\hat{t}_y^{\,d} = \boldsymbol{y}_d' \breve{\boldsymbol{I}} = \breve{\boldsymbol{y}}_d' \boldsymbol{I}, \qquad \textbf{(1)}$$

where $\breve{\boldsymbol{y}}_d$ is the expanded study vector of domain with elements $y_i / E(I_i)$ if $i \in U_d$ and 0 otherwise.
Under WOR design (1) is called Horwitz-Thompson (HT) estimator and under WR designs it is called the Hansen-Hurwitz estimator.

The design-based variance of $\hat{t}_y^{\,d}$ is

$$V(\hat{t}_y^{\,d}) = \breve{\boldsymbol{y}}_d{}'\Delta\breve{\boldsymbol{y}}_d, \qquad \textbf{(2)}$$

where Δ is the $N \times N$ covariance matrix of $\boldsymbol{I}$.

The use of auxiliary information is essential for efficient estimation. It consists of information on the variables that make up the matrix $\boldsymbol{X}$ of the dimension $N \times J$, $J \geq 1$. Its value for unit i is denoted by $\boldsymbol{x}_i$, $\boldsymbol{x}_i = (x_{i1}, x_{i2}, ..., x_{iJ})$. Like Estevao and Särndal (2004) viewed in their article, there are one or more population groups called control groups, calibration groups or C-groups. The C-groups define the C-level. The general notation for a calibration group is U_C.

The auxiliary information about U_C may consists of the following two components:

i) the auxiliary total $\boldsymbol{t}_x^C = \boldsymbol{X}_C'\boldsymbol{1}$, where matrix $\boldsymbol{X}_C$ have elements $[\boldsymbol{x}_C]_i = \begin{cases} \boldsymbol{x}_i, \ if \ \ i \in U_C, \\ \boldsymbol{0}, \ otherwise, \end{cases}$ is known.

ii) for every sample element i, the vector value $\boldsymbol{x}_i$ and membership or not of i in U_C are known.

The target of interest in this paper is the y-total, $t_y^d = \boldsymbol{y}_d'\boldsymbol{1}$ of the domain U_d ($U_d \subseteq U_C \subseteq U$). Available for this purpose are the data $\{(\boldsymbol{x}_i, y_i) : i \in s\}$, the C-total $\boldsymbol{t}_x^C$ and its unbiased estimator $\hat{\boldsymbol{t}}_x^C = \boldsymbol{X}_C'\breve{\boldsymbol{I}} = \breve{\boldsymbol{X}}_C'\boldsymbol{I}$, where matrix $\breve{\boldsymbol{X}}_C$ elements are $x_i / E(I_i)$ if $i \in U_C$ and 0 otherwise. To solve this problem, generalized regression and calibration estimation are studied.

3. Generalized regression estimation

An estimator of $\boldsymbol{y}$ that uses auxiliary information $\boldsymbol{X}$: $N \times J$ is based on a linear regression model, called ζ. The regression model ζ (Särndal *et al*., 1992, p. 226) has the following features that can be written in a matrix form:

(i) $y_1, y_2, ..., y_N$ are assumed to be realized values of independent random variables; if considered as random variables, the following holds:

(ii) $E_\zeta(\boldsymbol{y}) = \boldsymbol{X\beta}$,

(iii) $V_\zeta(\boldsymbol{y}) = \boldsymbol{\Sigma} = diag(\boldsymbol{\sigma}^2)$,

where E_ζ and V_ζ denote expected value and variance with respect to the model ζ, and where $\boldsymbol{\beta} = (\beta_1, ..., \beta_J)'$ and $\boldsymbol{\sigma}^2 = (\sigma_1^2, \sigma_2^2, ..., \sigma_N^2)'$ are model parameters.

The generalized regression estimator (the GREG estimator) constructed under these assumptions has the following form (Särndal *et al*., 1992, p. 232):

$$\hat{t}_{GREG} = \hat{t}_{HT} + (\boldsymbol{t}_x - \hat{\boldsymbol{t}}_x)'\hat{\boldsymbol{B}}, \qquad \textbf{(3)}$$

where the involved totals and their design-unbiased estimators in matrix form are:

$$\boldsymbol{t}_x = \boldsymbol{X'1}, \qquad \boldsymbol{B} = (\boldsymbol{X'\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X'\Sigma}^{-1}\boldsymbol{y},$$

$$\hat{\boldsymbol{t}}_x = \boldsymbol{X}'\breve{\boldsymbol{I}}, \qquad \hat{\boldsymbol{B}} = (\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\breve{\boldsymbol{I}}_{diag}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\breve{\boldsymbol{I}}_{diag}\boldsymbol{y},$$
$$\breve{\boldsymbol{I}}_{diag} = diag(\breve{\boldsymbol{I}}).$$

Auxiliary information helps to reduce the variance of the estimator. For example, Särndal *et. al.* (1992, p. 239) claim that the GREG estimator is more accurate than (1) in sense of variability of the estimator. But the GREG estimator has a very small bias, which is of order n^{-1} (Särndal *et al*., 1992, p. 238). It is also known, that the bias ratio (the bias divided by the standard error of the estimator) tends to zero with increasing sample size (Estevao & Särndal, 2004).

The variance of the (3) cannot be obtained exactly, because of the complex nature of the GREG estimator. The linearization techniques (Särndal *et al.*, 1992, p. 246) give the approximate variance

$$Cov(\hat{t}_{GREG}) = (\boldsymbol{y} - \boldsymbol{XB})'\breve{\Delta}(\boldsymbol{y} - \boldsymbol{XB}), \qquad \textbf{(4)}$$

where $\breve{\Delta} = Cov(\breve{\boldsymbol{I}})$ is the covariance matrix of the expanded sampling vector.

3.1 The GREG estimator of domain total

The regression estimator can be carried out at different levels, leading to different $\hat{\boldsymbol{B}}$. The estimator of domain total t_y^d is built by the principle

$$\hat{t}^d_{GREG} = \hat{t}^d_{HT} + (\boldsymbol{t}_x^C - \hat{\boldsymbol{t}}_x^C)'\hat{\boldsymbol{B}}. \qquad \textbf{(5)}$$

This leads to a reduction in variance, compared to the simple HT estimator, if there exists a negative correlation between the HT term $\hat{t}^d_{HT}$ and the regression adjustment term $(\boldsymbol{t}_x^C - \hat{\boldsymbol{t}}_x^C)'\hat{\boldsymbol{B}}$, which is a very nearly unbiased estimate of zero (Estevao and Särndal, 2004).

3.1.1 The GREG estimator at the domain level

A general regression estimator at the domain level is

$$\hat{t}^d_{GREG/DOM} = \hat{t}^d_{HT} + (\boldsymbol{t}_x^C - \hat{\boldsymbol{t}}_x^C)'\hat{\boldsymbol{B}}_{s_d}. \qquad \textbf{(6)}$$

where the coefficient $\hat{\boldsymbol{B}}_{s_d}$ is

$$\hat{\boldsymbol{B}}_{s_d} = (\boldsymbol{X}_d'\boldsymbol{\Sigma}^{-1}\breve{\boldsymbol{I}}_{diag}\boldsymbol{X}_d)^{-1}\boldsymbol{X}_d'\boldsymbol{\Sigma}^{-1}\breve{\boldsymbol{I}}_{diag}\boldsymbol{y}_d. \qquad \textbf{(7)}$$

Units outside the domain do not contribute to the sum, so this estimator is a direct estimator. This estimator recognizes differences between domains.

3.1.2 The GREG estimator at the full sample level

A general regression estimator at the full sample level is

$$\hat{t}^d_{GREG/SAMPLE} = \hat{t}^d_{HT} + (\boldsymbol{t}_x^C - \hat{\boldsymbol{t}}_x^C)'\hat{\boldsymbol{B}}_s, \qquad \textbf{(8)}$$

where the coefficient $\hat{\boldsymbol{B}}_s$ is

$$\hat{\boldsymbol{B}}_s = (\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\breve{\boldsymbol{I}}_{diag}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\breve{\boldsymbol{I}}_{diag}\boldsymbol{y}. \qquad \textbf{(9)}$$

This estimator is an indirect estimator that attempts to borrow strength by using y-data from the entire sample to strength a potentially weak regression estimator.

Other options exist. For example C-level estimator is

$$\hat{t}^{d}_{GREG/C-level} = \hat{t}^{d}_{HT} + (\boldsymbol{t}^{C}_{x} - \hat{\boldsymbol{t}}^{C}_{x})' \hat{\boldsymbol{B}}_{s_C}, \quad \textbf{(10)}$$

where $\hat{\boldsymbol{B}}_{s_C}$ is given by (7) if we replace $\boldsymbol{y}_d$ and $\boldsymbol{X}_d$ by $\boldsymbol{y}_C$ and $\boldsymbol{X}_C$, respectively.

4. Calibration estimation

The calibration approach to estimation for finite populations consists of (Särndal, 2007)

(a) a computation of weights that incorporate specified auxiliary information and are restrained by calibration equation(s),
(b) the use of these weights to compute linearly weighted estimates of totals and other finite population parameters,
(c) an objective to obtain nearly design unbiased estimates as long as nonresponse and other nonsampling errors are absent.

The objective is to determine weights $\boldsymbol{w}$, $\boldsymbol{w} = \breve{\boldsymbol{I}} \circ \boldsymbol{g}$ to satisfy the calibration equation

$$\boldsymbol{t}_x = \boldsymbol{X}'\boldsymbol{1} = \boldsymbol{X}'(\breve{\boldsymbol{I}} \circ \boldsymbol{g}), \quad \textbf{(11)}$$

where

$$\boldsymbol{g} = \boldsymbol{1} + \boldsymbol{\Sigma}^{-1}\boldsymbol{X}\hat{\boldsymbol{T}}^{-1}(\boldsymbol{t}_x - \hat{\boldsymbol{t}}_x),$$
$$\hat{\boldsymbol{T}} = \boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\breve{\boldsymbol{I}}_{diag}\boldsymbol{X},$$

and then use them to form the calibration estimator of t_y as

$$\hat{t}_{CAL} = \boldsymbol{y}'\boldsymbol{w}, \quad \textbf{(12)}$$

which we can confront with the unbiased HT estimator by writing

$$\hat{t}_{CAL} = \hat{t}_{HT} + (\boldsymbol{t}_x - \hat{\boldsymbol{t}}_x)'\hat{\boldsymbol{B}}. \quad \textbf{(13)}$$

4.1 Calibration estimator at the domain level

The calibrated weights will be computed with the given auxiliary information (i) and (ii) in section 2. We apply these weights to the domain variable $\boldsymbol{y}_d$. The resulting estimator is

$$\hat{t}^{d}_{CAL} = \boldsymbol{y}'_d\boldsymbol{w}, \quad \textbf{(14)}$$

where $\boldsymbol{w} = \breve{\boldsymbol{I}} \circ \boldsymbol{g}$ and

$$\boldsymbol{g} = \boldsymbol{1} + \boldsymbol{\Sigma}^{-1}\boldsymbol{X}_C\hat{\boldsymbol{T}}^{-1}(\boldsymbol{t}^{C}_{x} - \hat{\boldsymbol{t}}^{C}_{x}),$$
$$\hat{\boldsymbol{T}} = \boldsymbol{X}'_C\boldsymbol{\Sigma}^{-1}\breve{\boldsymbol{I}}_{diag}\boldsymbol{X}_C.$$

The weights are calibrated to the C-level, the calibration equation is

$$\boldsymbol{t}^{C}_{x} = \boldsymbol{X}'_C\boldsymbol{1} = \boldsymbol{X}'_C\boldsymbol{w}.$$

We can write the estimator (14) as

$$\hat{t}_{CAL}^{d} = \hat{t}_{HT}^{d} + (\boldsymbol{t}_x^C - \hat{\boldsymbol{t}}_x^C)' \hat{\boldsymbol{R}}, \qquad \textbf{(15)}$$

where

$$\hat{\boldsymbol{R}} = (\boldsymbol{X}_C' \breve{\boldsymbol{I}}_{diag} \boldsymbol{X}_C)^{-1} \boldsymbol{X}_C' \breve{\boldsymbol{I}}_{diag} \boldsymbol{y}_d .$$

Its properties are (Estevao and Särndal, 2004):

i) In (15) we have $E(\hat{t}_{HT}^{d}) = t_y^d$ and the expectation of the other term tends to zero, making $\hat{t}_{CAL}^{d}$ design-consistent and very nearly design unbiased for t_y^d.

ii) In practice, the same weight system, $\boldsymbol{w} = \breve{\boldsymbol{I}} \circ \boldsymbol{g}$, is often used to produce estimates for any domain $U_d \subseteq U_C$.

iii) Estimator (14) is direct because the only $\boldsymbol{y}$ values used are for units inside the domain.

5. Simulations and comments

Investigation is in progress.

6. References

Estevao, V.M., Särndal, C-E. (2004) Borrowing strenght is not the best technique within a wide class of desing-consistent domain estimators. *Journal of Official Statistics* vol. 20, No.4, pp. 645-669

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

Särndal, C.-E. (2007) The calibration approach in survey theory and practice. *Survey Methodology* vol. 33, No. 2, pp. 99-119

EFFICIENCY OF DOUBLE SAMPLING IN THE LABOUR FORCE SURVEY

Edita Kemzūraitė[1]

[1] Vilnius Gediminas Technical University, Lithuania
e-mail: edita.kemzuraite@gmail.com

Abstract

The accuracy of the estimator of a total in the Lithuanian Labour Force Survey is investigated. Two sample designs and approximately design-unbiased estimators of a total are used. The simulation with the real data is studied. The estimates are calculated and the estimates of their accuracy measures are compared.

1 Introduction

The Labour Force Survey (LFS) is one of the most important surveys in official statistics of any country. The data are collected from the members of the households selected according to some probabilistic sample design. The total number of employed and unemployed individuals, as well as the unemployment level in the population and in its domains are estimated. The estimates in some domains are not accurate enough under the current sample design. It is important to find an efficient sample design and improve the accuracy of estimators in the Lithuanian LFS. An alternative sample design, described in Ilves (2004), is used. The dependency of the main estimates in the LFS upon the sample design is studied in this paper. The main purpose is to describe a simulation study with the real data using two – currently used in the Lithuanian LFS and alternative – sample designs, and to compare the accuracy of the estimates obtained.

2 Sampling designs and estimators of the parameters

Let $U = (1,2,...,N)$ be the finite population of the persons of working age. They are grouped into M households of size m_k, k=1,2,…,M, so that

$$\sum_{i=1}^{M} m_i = N .$$

The main parameters of interest of the population are the number of employed individuals, number of unemployed individuals, and the unemployment rate. Let us introduce a study variable y with the values

$$y_k = \begin{cases} 1, & \text{if } k \text{ has an attribute,} \\ 0, & \text{otherwise,} \end{cases} \qquad (1)$$

k=1,2,…,N. The attribute is considered as an employed or unemployed for a person. The parameters of interest of the LFS – the number of the employed t_{emp} and that of unemployed individuals t_{unemp} – can be expressed as totals of the variable y:

$$t_y = \sum_{k=1}^{N} y_k \ , \tag{2}$$

While the unemployment rate as $r = t_{unemp} / (t_{emp} + t_{unemp})$.

2.1 The currently used LFS sample design

A simple random sample $\mathbf{s}'$, $\mathbf{s}' \subset U$, of size *n* is drawn from the population of individuals, and the members of households of the selected individuals are included into the sample. The final sample is denoted by $\mathbf{s}$, its size is variable and equal to $\sum_{k \in \mathbf{s}'} m_k$.

The sample of individuals $\mathbf{s}$ is drawn according to the with replacement sampling design with the probabilities proportional to the household size $p_k = m_k / N$. Then the inclusion probability of the individual to the sample $\mathbf{s}$ is

$$\pi_k = P(\mathbf{s}, k \in \mathbf{s}) = 1 - (1 - p_k)^n \approx \frac{n m_k}{N} . \tag{3}$$

The population total t_y (2) is estimated by the Hansen-Hurwitz estimator (Krapavickaitė D. and Plikusas A., 2005) (replacements include into the sum):

$$\hat{t}_y = \sum_{k \in \mathbf{s}} d_k y_k \ , \qquad d_k = \frac{N}{m_k n} . \tag{4}$$

In such a way we obtain estimators for the number of the employed individuals $\hat{t}_{emp}$, number of the unemployed individuals $\hat{t}_{unemp}$, and the unemployment rate

$$\hat{r} = \hat{t}_{unemp} / (\hat{t}_{emp} + \hat{t}_{unemp}) . \tag{5}$$

2.2 The alternative sample design

The inclusion probabilities of individuals into the sample $\mathbf{s}$ expressed by (3) are unequal, and it may be one of the sources of high variance of the estimator of a total. The alternative LFS sample design is constructed in Ilves M., 2004. It is a two-phase sample design for stratification, and it is also used in this study.

The average household size in the population is equal to

$$\overline{m} = \frac{1}{M} \sum_{k=1}^{M} m_k \ .$$

The alternative sample design consists of the following steps:

1) The initial simple random sample $\mathbf{s}'$ of individuals of the size $\widetilde{n} = n\overline{m}$ is drawn from the population *U*, and the sample $\mathbf{s}^{(1)}$ of all the members of their households is considered as being the first phase sample. It is selected according to the LFS sample design as described in Section **2.1**, and its size is $\sum_{k \in \mathbf{s}'} m_k \approx \widetilde{n}\overline{m}$.

2) The obtained sample $\mathbf{s}^{(1)}$ is divided into the strata $\mathbf{s}^{(1)} = \mathbf{s}_1^{(1)} \cup \mathbf{s}_2^{(1)} \cup \ldots \cup \mathbf{s}_H^{(1)}$ by the size of the household *h*, *h*=1,2,…, *H*.

3) A simple random sub-sample $\mathbf{s}_h^{(2)}$ of $n_h^{(2)} \approx n_h^{(1)} / h$ households is selected from the stratum $\mathbf{s}_h^{(1)}$, *h*=1,2,…*H*, and the individuals of all sub-samples constitute the second phase sample $\mathbf{s}^{(2)}$. Its size is $\widetilde{\widetilde{n}} = n_1^{(2)} + \ldots + n_H^{(2)} \approx \widetilde{n}$.

The inclusion probability of the element *k*, $k \in U$, into the second phase sample $\mathbf{s}^{(2)}$, $\mathbf{s}^{(2)} \subset U$, is

$$\widetilde{\pi}_k = P(\mathbf{s}^{(2)} : k \in \mathbf{s}^{(2)}) = P(k \in \mathbf{s}^{(1)}, k \in \mathbf{s}^{(2)}) =$$

$$= P\left(k \in \mathbf{s}^{(2)} \middle| k \in \mathbf{s}^{(1)}\right) P\left(k \in \mathbf{s}^{(1)}\right) = \frac{1}{m_k} \cdot \pi_k \approx \frac{1}{m_k} \cdot \frac{m_k \tilde{n}}{N} = \frac{\tilde{n}}{N}.$$

The estimator $\tilde{t}_y$ for two phase sampling design for stratification is used for estimating totals:

$$\hat{\tilde{t}}_y = \frac{N}{\tilde{n}} \sum_{h=1}^{H} \frac{1}{h} \frac{n_h^{(1)}}{n_h^{(2)}} \sum_{k \in \mathbf{s}_h^{(2)}} y_k \approx \sum_{k \in \mathbf{s}^{(2)}} \tilde{d}_k y_k = \tilde{t}_y \tag{6}$$

with the weights $\tilde{d}_k = \frac{N}{\tilde{n}} \approx \frac{1}{\tilde{\pi}_k}$, and replacements included. Using formula (6), the number of the employed and that of unemployed individuals is estimated by $\tilde{t}_{emp}$ and $\tilde{t}_{unemp}$, and the unemployment rate by (5) with the estimators $\tilde{t}$ instead of $\hat{t}$.

3 Simulation plan

The population for the simulation study consists of N=13 383 individuals, grouped into the M=5 851 households of the size h=1,2,…,6. The data of the real Lithuanian LFS survey in 2005 are used. The real values of the parameters t_{emp}, t_{unemp} and r are calculated.

G=1 000 samples are selected according to each sample design, and the estimates $\hat{\theta}_j$, j=1,2,…,G of the estimator $\hat{\theta}$ of the parameters θ (t_{emp}, t_{unemp} and r) are obtained. Their average, variance, bias

$$\bar{\theta} = \frac{1}{G} \sum_{j=1}^{G} \hat{\theta}_j, \quad Var\left(\hat{\theta}\right) = \frac{1}{G} \sum_{j=1}^{G} \left(\hat{\theta}_j - \bar{\theta}\right)^2, \quad Bias\left(\hat{\theta}\right) = \bar{\theta} - \theta$$

and a relative mean squared error

$$MSE\left(\hat{\theta}\right) = \frac{\sqrt{Var\left(\hat{\theta}\right) + Bias\left(\hat{\theta}\right)}}{\theta}$$

are calculated for each of the estimators. The computer program SAS is used for simulations.

It is planning to test the hypothesis on the normality of the distribution by the Shapiro-Wilk test with the significance level $\alpha = 0.05$ (Čekanavičius V. and Murauskas G., 2000) and the hypothesis on the equality of variances of the estimators for both sampling designs will be testing by the F statistics with the same significance level.

The simulation is already in progress and its results will be presented at the Summer School.

References

Čekanavičius V. and Murauskas G. (2000) Statistika ir jos taikymai (*Statistics and its Applications*). Vilnius: TEV.

Ilves M. (2004) Variance and its estimator for one two-phase design. In: *Workshop on Survey Sampling Theory and Methodology* [*Tartu, Estonia, June* 18-22]. University of Tartu, Estonia. p.37-41.

Krapavickaitė D. and Plikusas A. (2005) *Imčių teorijos pagrindai* (*Basics of sample theory*). Vilnius: Technika.

ON USING DATA MINING METHODS IN ECONOMIC MODELING UNDER SMALL SAMPLE

Alexander Kolosov[1]

[1] Donetsk National University, Ukraine
e-mail: kolosov@dongu.donetsk.ua

Abstract

Difficulties of application of statistical analysis of data in applied investigations and example of using cluster analysis for generating hypothesis about structural changes in regional economics on example of Donetsk region of Ukraine are considered.

1 Introduction

Statistical analysis is essential of applied investigations. These are approbation of investigation results, basing of suppositions in making and using models of investigating phenomena. For this purposes methods of statistical analysis are used traditionally. They came classical methods: for example, regression, dispersion, and factor analysis. One of base conditions of its using is either enough sample size or gaussian distribution of data. However, these conditions are fulfilled not always. And causes of this are different. In medicine they are natural limitations on numbers of patients (for example, expansion of illness), in sociology they are limitations of financing, in economics they are peculiarities in collection of information and not enough continuance of processes that are investigated.
At the same time the number of methods of data investigation exists, which not need a priori assumptions. Traditionally in marketing analysis this methods are strongly used. However, it is possible essentially to expand the area of using these methods. Of course, we can't use confirmatory analysis of data in this case, but hypothesis making is quite possible. In other words, in this case at first data analysis takes place in applied investigations, so-called exploring analysis, and then discovered data peculiarities are explained on general theory of investigation direction base.

2 Data Mining Conception

Recently in world number of new conceptions of saving and analysis corporative data are formed:

1. on-line analytical processing (OLAP);
2. intellectual data analysis (Data Mining).

If we were considering these conceptions, we can get to know that Data Mining and OLAP consists statistical methods. In this methods the main accent is devoted to classical methods: factor, correlation, cluster analysis.
The data multidimensional presentation is in base of conception of OLAP. In 1993 E.F. Codd introduced the term OLAP. He defined general requirements to OLAP systems that expands functionality of control databases systems and includes multidimensional analysis as a one of its characteristics.

The term OLAP (or fast distributed multidimensional information analysis) denotes methods, which give possibility to receive answers on different analytical queries to users of multidimensional databases, and so give process of multidimensional databases analysis by way of compilation of effective multidimensional queries to data of different types. Analysis, which is realized by OLAP methods, can be both simple (for example, frequency tables, descriptive statistics, simple tables) and enough complex (for example, it can include different methods of data cleaning).

Concept “data mining” is defined as process of analytical investigation of information arrays (usually of economic character) with aim of bringing to light of certain regularities and systematical interdependences between variables, which then can be used to new data aggregate. This process includes three base stages: investigation, modeling and its testing.

In ideal case under enough number of data it can to organize iterative procedure for constructing of stable model. At the same time, it is practically impossible to test economic model in real situation on analysis stage and therefore initial results have heuristic character that can be used in making decision process.

Data mining methods obtain greater and greater popularity as instrument for economic information analysis, especially in those cases, when it is supposed, that it is possible to extract knowledge for making decision in uncertainty conditions from having data. Though at last time interest to development of new data analysis methods increases, that is specially designed for business sphere (for example, classification tree), as a whole as usual data mining systems base on classical principles of exploring analysis and making models and use same approaches and methods. Unlike traditional hypothesis testing that designed for a priori assumptions testing that is concerned relations between variables exploring data analysis applies for detection relations between variables in situations, when is lacking (or not enough) a prior idea about nature of this relations. As a rule, under exploring analysis a big number of variables are compared and took into account and for finding regularities very different methods are used. Computing methods of exploring data analysis consist base statistic methods as well as more complicated, specially developed methods of multidimensional analysis that are designed for detection of regularities in multidimensional data.

In base of modern technology of data mining it is put conception of patterns that reflect fragments of multifold relations in data. These patterns represent regularities that are peculiar to data subsamples, which may be compact expressed in understandable form. The finding of patterns is made by methods that aren’t limited by frames of prior assumptions about sample structure in distribution form of analyzing activities values.

Data mining isn’t one method, but it is aggregate of big number of different methods of knowledge detection. The choice of method often depend from data type that is had and from which information your try to obtain. Some methods are enumerated below:

- association, this is allocation of structures that is repeated in time consecution;
- sequence-based analysis or sequential association, it allow to find regularities between transactions in time;
- clustering, it is grouping of records that have same characteristics, for example by nearness of fields values;
- classification, it is put of record to one of beforehand defined classes;
- estimation;
- neural networks, data is passed over layers of nodes that is “trained” for recognition of those or other patterns.

3 Example of Clustering in Economic Modelling

Clustering it is grouping of records that have same characteristics, for example by nearness of fields values. It is possible to use statistical methods or neural networks. Clustering is often considered as first necessary step for subsequent data analysis.

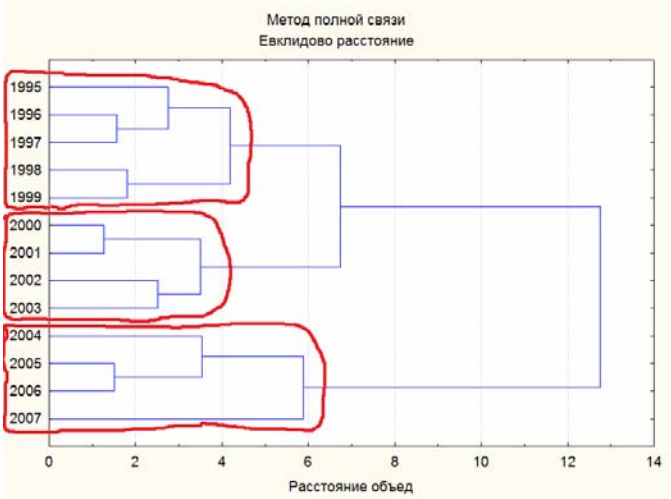
“Cluster analysis” is common name of set of computing procedures. “Clusters” or groups of similar objects are made as result of work with procedures. Different applications of cluster analysis can be reduced to four base problems:

1) development of typology or classification;
2) investigation of useful conceptual schemas of objects grouping;
3) generating of hypothesis on data investigation base;
4) hypothesis testing or investigation for determination of presence of types or groups in using data that is chose by some way.

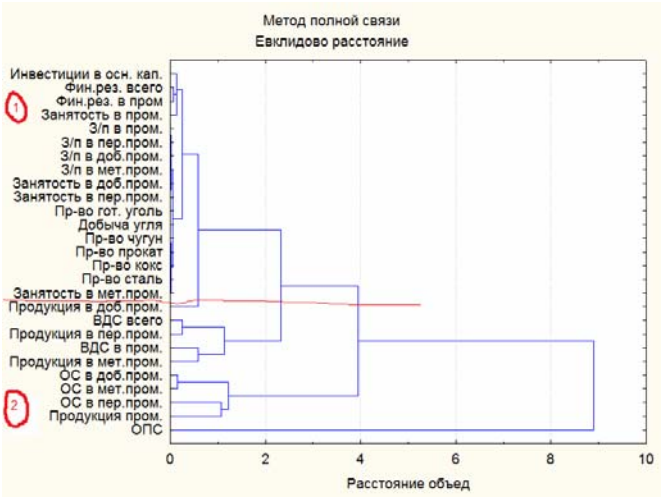
Many methods of cluster analysis are enough simple procedures, which as a rule haven’t sufficient statistical grounds. In other words, many methods of cluster analysis are heuristic. It is sharp difference, for example, from methods of factor analysis, which is well statistical grounded.

We consider the example of cluster analysis using for generating hypothesis about structural changes in regional economics on example of Donetsk region of Ukraine. Period of investigation was cover 1995-2007 years. Processes in economics that are related with breakdown of USSR are finished to 1995 year. Actions for economics of region recovery are carried out in period from 1995 to 2007. We can choose two most significant of its: “experiment in coal-mining- metallurgical industry” (1999-2002) and “special economical zone” (2000-2004). Displaying of structural changes in regional economics in discovered period is the aim of investigation. Economical development is described by set of activities that economists join in such groups: “natural activities”, “investments”, “financial result”, “industrial production”, “employment”, “salary”. In each group more significant for region activities was chose. For example, it is considered in “natural activities” such activities: coal mining, production of iron, steel, rolled metal, finished coal, coke. Thus, it was chose of 32 activities. The main peculiarity in collection of statistical data of these activities is that it’s accessible only on annual base. Thus, for investigation sample of 13 cases and 32 variables were taken. Classical methods of statistical analysis in those conditions are ineffective. At the same time cluster analysis (treelike clustering) gave information for generating economic hypothesis. In the first place similarity of variables in years was investigated. On results we can suppose presence of three periods in economic evolution of region (picture 1). Further, in each of supposed periods separately similarity of 32 variables is investigated (pictures 2-4) and then on each similar fragment similarity is investigated (example on pictures 5-7).

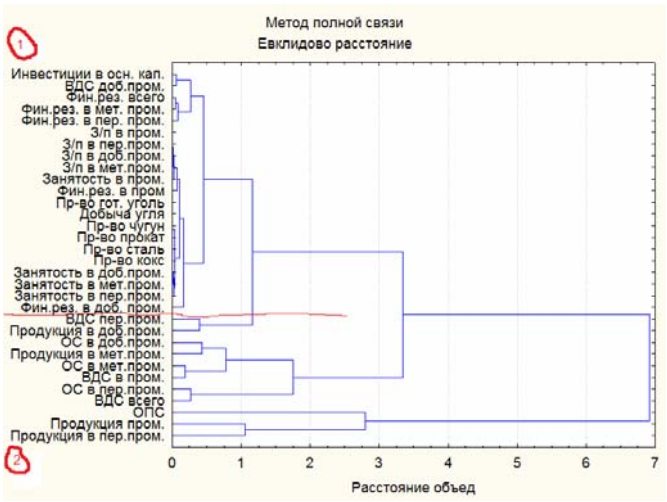
Thus, changes in similarities of some variables in dependence from time period were displayed. Further study and explanation of displayed changes in point of view of economic theory can give answer on question about structural changes in regional economics that are concerned with actions for economics of region recovery.



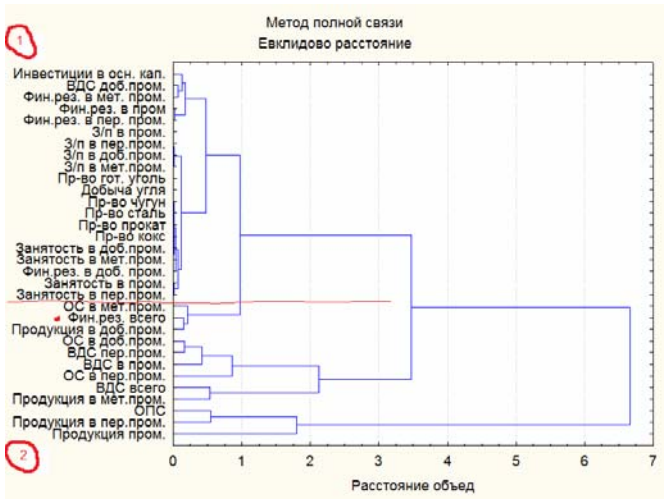
Picture 1



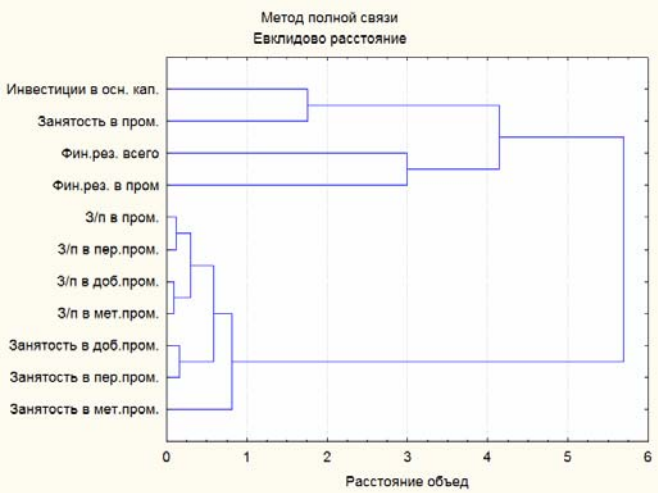
Picture 2



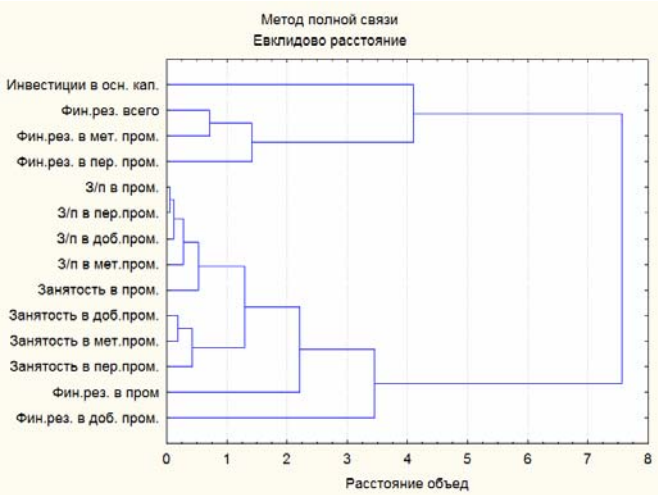
Picture 3



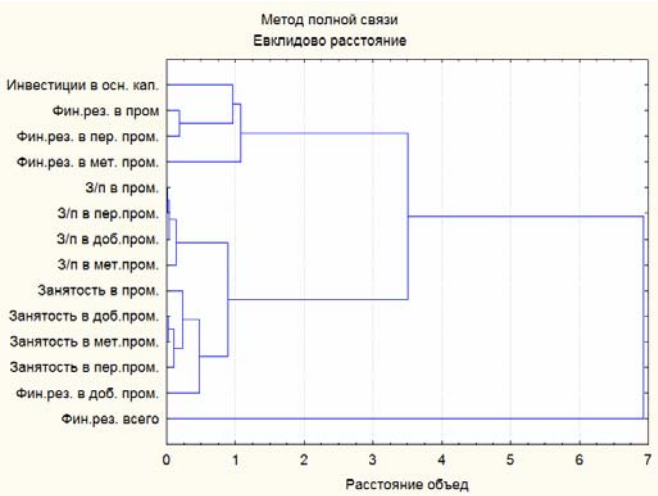
Picture 4



Picture 5



Picture 6



Picture 7

References

Duran B.S., Odell P.L. (1974) *Cluster analysis*. Springer Verlag. Berlin-Heidelberg-New York.

Ledermann W., Lloyd E. (1989) *Handbook of Applicable Mathematics. Volume IV: Statistics*. A wiley interscience publication. New York.

Synthetic estimator for domains

Natalja Lepik[1]

[1] University of Tartu, Estonia
e-mail: natalja.lepik@ut.ee

Abstract

In this paper overview of the synthetic estimator for the domain total is given considering the design based approach. In this case it is known that synthetic estimator is not unbiased for the domain total. The expression of the bias is given for the population and the domain case.

1 Introduction

1.1 Design based approach in matrix notation

Consider a finite population $U = (1, 2, ..., N)$ that consist of N units. A probability sample is drawn from U according to some sampling design, characterized by a random sampling vector $\mathbf{I} = (I_1, I_2, ..., I_N)'$ with I_i indicating the number of selections of unit i from U:

$$\mathbf{I} \sim p(\boldsymbol{k}) = Pr(\mathbf{I} = \boldsymbol{k}),$$

where $\boldsymbol{k} = (k_1, k_2, ..., k_N)'$ is an outcome of $\mathbf{I}$ (Traat et al. 2004, Tillé, 2006).

For without-replacement (WOR) designs $I_i \in \{0, 1\}$ and for the with-replacement designs $I_i \in \{0, 1, 2, ...\}$. Note that the sample size n can be expressed as $n = \mathbf{I}'\mathbf{1}$, where $\mathbf{1}$ is the N-dimensional vector of ones.

Depending on the sampling design, n can be random or fixed. A sampling design with fixed n is called a fixed size sampling design.

The parameter of interest is the population total t_y,

$$t_y = \boldsymbol{y}'\mathbf{1}, \tag{1}$$

where $\boldsymbol{y} = (y_1, y_2, ..., y_N)'$ is the study variable. In general, surveys have many study variables. For our purpose, as well in sampling literature elsewhere, it has been enough to develop formulas for the one-dimensional case.

In this paper we consider the design-based approach, i.e. properties of the estimators (expectation and variance/covariance) are determined only by the sampling design.

Let $\breve{\mathbf{I}} = (\breve{I}_1, \breve{I}_2, ..., \breve{I}_N)'$ be the expanded sampling vector, where

$$\breve{I}_i = \frac{I_i}{\mathbb{E}(I_i)} \tag{2}$$

and $\mathbb{E}(I_i)$ is the design-based expectation. Note that $\mathbb{E}(\breve{\mathbf{I}}) = \mathbf{1}$.

The *linear estimator* of t_y that uses only study variable values is

$$\hat{t}_{y,lin} = \boldsymbol{y}'\breve{\mathbf{I}} = \breve{\boldsymbol{y}}'\mathbf{I}, \tag{3}$$

where $\breve{\boldsymbol{y}}$ is the expanded study vector with elements $\breve{y}_i = y_i/\mathbb{E}(I_i)$. The estimator (3) is design-unbiased, $\mathbb{E}(\hat{t}_{y,lin}) = t_y$.

Since $\hat{t}_{y,lin} = \boldsymbol{y}'\breve{\mathbf{I}} = \sum_U \frac{I_i}{\mathbb{E}(I_i)} y_i$, where the summation goes over the all elements $i \in U$, the qualities $\breve{I}_i$ (given by formula (2)) are usually called design weights. They weights up sampled y_i values to produce estimated population total. From this prospective we call sometimes the vector $\breve{\mathbf{I}}$ the weight vector.

The design-based variance of $\hat{t}_{y,lin}$ can be written in a matrix form as

$$\mathbb{V}(\hat{t}_{y,lin}) = \breve{\boldsymbol{y}}'\Delta\breve{\boldsymbol{y}}, \tag{4}$$

where Δ is the $N \times N$ covariance matrix of $\mathbf{I}$.

Under WOR design (3) is usually called Horvitz-Thompson (HT) estimator and under WR designs it is called the Hansen-Hurwitz estimator. Note that the unified consideration of WOR and WR designs is not the usual one in sampling literature. It has been forcefully developed in Traat (2000), Traat, Meister, Sõstra (2001), Tillé (2006). Since in addition to WOR-designs, also some WR designs are very implemental in practice (Traat, Ilves, 2007) we continue to use the unified consideration in this thesis.

1.2 Linear estimator for domain

Nowadays the estimation of the domain parameters became an undividable part of the estimation in a whole. As it is defined in Srndal et al. (1992, p. 386) we use the term *domain* for the subpopulation for which separate point estimates and confidence intervals are required.

Let U be divided into D non-overlapping domains U_d, $d \in \mathcal{D} = \{1, 2, ..., D\}$ with N_d being the size of the domain U_d. We are interested in the domain totals of study

variable y:

$$t_y^d = \sum_{i \in U_d} y_i = \boldsymbol{y}_d' \mathbf{1}, \tag{5}$$

where $\boldsymbol{y}_d$ is vector of elements $y_i \in U_d$ and the 1 is the vector consisting of N_d ones. The sampling is carried on the whole population, therefore the sample sizes in the domains are random. We also assume that we can identify for the object $i \in U$ if it belongs to the domain or not. In the traditional approach the domain indicator-vector $\boldsymbol{\delta}_d = (\delta_{d1}, \delta_{d2}, ..., \delta_{dN})'$, $d \in \mathcal{D}$, where $\delta_{di} = 1$ if $y_i \in U_d$ and 0 otherwise, is usually used as possibility to apply the estimation theory of the population total for the domain estimation. According to this approach the domain total (5) can be rewritten as

$$t_y^d = \sum_{i \in U} \delta_{di} y_i = \boldsymbol{\delta}_d' \boldsymbol{y}, \tag{6}$$

and the linear estimator (3) can be applied to t_y^d.

2 Synthetic estimator

An estimator of the total y that uses auxiliary information $\mathbf{X} : N \times p$ is based on an assisting linear regression model, called ξ. The regression model ξ (Särndal *et al.*, 1992, p. 226) has the following features that can be written in a matrix form:

(i) The components of $\boldsymbol{y} = (y_1, y_2, ..., y_N)'$ are assumed to be realized values of independent random variables;
if considered as random variables, the following holds:

(ii) $\mathbb{E}_\xi(\boldsymbol{y}) = \mathbf{X}\boldsymbol{\beta}$,

(iii) $\mathbb{V}_\xi(\boldsymbol{y}) = \boldsymbol{\Sigma} = \mathrm{diag}\,(\boldsymbol{\sigma}^2)$,

where $\mathbb{E}_\xi$ and $\mathbb{V}_\xi$ denote expected value and variance with respect to the model ξ, and where $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)'$ and $\boldsymbol{\sigma}^2 = (\sigma_1^2, \sigma_2^2, ..., \sigma_N^2)'$ are model parameters.

2.1 Synthetic estimator for the population total

Synthetic estimator is defined as a sum of the predicted values of some assisting regression model ξ (Srndal et al., 1992, p. 399),

$$\hat{t}_{y,syn} = \boldsymbol{t}_x' \hat{\boldsymbol{b}}, \tag{7}$$

where $\hat{\boldsymbol{b}}$ is a the sample-estimated regression coefficient-vector and $\boldsymbol{t}_x = \mathbf{X}'\mathbf{1}$. The quantity $\hat{\boldsymbol{b}}$ can be found as

$$\hat{\boldsymbol{b}} = \hat{\mathbf{T}}^{-1}\hat{\boldsymbol{t}}_{xy}, \tag{8}$$

where $\hat{\boldsymbol{t}}_{xy} = \mathbf{X}'\boldsymbol{\Sigma}^{-1}\breve{\mathbf{I}}_{diag}\boldsymbol{y}$ and $\hat{\mathbf{T}} = \mathbf{X}'\boldsymbol{\Sigma}^{-1}\breve{\mathbf{I}}_{diag}\mathbf{X}$ are design-unbiased estimators of totals $\boldsymbol{t}_{xy} = \mathbf{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{y}$ and $\mathbf{T} = \mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}$.

The estimator (7) is non-linear. To study its properties it need to be expand into the Taylor series.

Proposition 1. The Taylor expansion of the synthetic estimator (7) up to the second order term is

$$\begin{aligned}\hat{t}_{y,syn} \approx\ & \boldsymbol{t}_x'\mathbf{T}^{-1}\hat{\boldsymbol{t}}_{xy} - (\boldsymbol{t}_{xy}'\mathbf{T}^{-1} \otimes \boldsymbol{t}_x'\mathbf{T}^{-1})\mathrm{vec}(\hat{\mathbf{T}} - \mathbf{T}) \\ & -(\hat{\boldsymbol{t}}_{xy} - \boldsymbol{t}_{xy})'(\mathbf{T}^{-1} \otimes \boldsymbol{t}_x'\mathbf{T}^{-1})\mathrm{vec}(\hat{\mathbf{T}} - \mathbf{T}) \\ & +\mathrm{vec}'(\hat{\mathbf{T}} - \mathbf{T})(\mathbf{T}^{-1}\boldsymbol{t}_{xy} \otimes \mathbf{T}^{-1} \otimes \boldsymbol{t}_x'\mathbf{T}^{-1})\mathrm{vec}(\hat{\mathbf{T}} - \mathbf{T}),\end{aligned} \tag{9}$$

where $\otimes$ is the Kronecker product, $\mathrm{vec}(\cdot)$ is the operation of the vectorization.

Proposition 2. Theoretical expression of the bias of $\hat{t}_{y,syn}$, obtained from the Taylor expansion (9), is

$$\begin{aligned}b(\hat{t}_{y,syn}) \approx\ & -\mathrm{vec}'\left[\mathbb{C}ov\left(\hat{\boldsymbol{t}}_{xy}, \mathrm{vec}(\hat{\mathbf{T}})\right)\right]\mathrm{vec}(\mathbf{T}^{-1} \otimes \boldsymbol{t}_x'\mathbf{T}^{-1}) \\ & +\mathrm{vec}'\left[\mathbb{V}\left(\mathrm{vec}(\hat{\mathbf{T}})\right)\right]\mathrm{vec}(\mathbf{T}^{-1}\boldsymbol{t}_{xy} \otimes \mathbf{T}^{-1} \otimes \boldsymbol{t}_x\mathbf{T}^{-1}).\end{aligned} \tag{10}$$

The bias expression is obtained from the second-order term of (9) that makes its very small, almost negligible. So, in the case of the estimation of the population total t_y the synthetic estimator is almost unbiased.

2.2 Synthetic estimator for the domain total

The idea of the synthetic estimator of the domain total is that all domains are similar, so we could "borrow strength" from all domains in order to construct an estimate for any single domain,

$$\hat{t}^d_{y,syn} = (\boldsymbol{t}^d_x)'\hat{\boldsymbol{b}}, \tag{11}$$

where $\boldsymbol{t}^d_x$ the vector of auxiliary sums in the domain U_d.

From (11) it can be easily seen that synthetic estimator for domain has the property of additivity,

$$\sum_{d\in\mathcal{D}} \hat{t}^d_{y,syn} = \left[\sum_{d\in\mathcal{D}} (\boldsymbol{t}^d_x)'\right] \hat{\boldsymbol{b}} = (\boldsymbol{t}_x)'\hat{\boldsymbol{b}} = \hat{t}_{y,syn}. \tag{12}$$

Analogically to the synthetic estimator for the population total, estimator (11) can be expanded into the Taylor series in order to study its properties. We use only the linear term of the Taylor series because it describes the majority of the estimator.

Proposition 3. The linear term of the Taylor expansion of the synthetic estimator (11) for the domain d is

$$\hat{t}^d_{y,syn} \approx (\boldsymbol{t}^d_x)'\mathbf{T}^{-1}\hat{\boldsymbol{t}}_{xy} - (\boldsymbol{t}'_{xy}\mathbf{T}^{-1} \otimes (\boldsymbol{t}^d_x)\mathbf{T}^{-1})\text{vec}(\hat{\mathbf{T}} - \mathbf{T}). \tag{13}$$

Proposition 4. Theoretical expression of the bias of $\hat{t}^d_{y,syn}$, obtained from the Taylor expansion (9), is

$$b(\hat{t}^d_{y,syn}) \approx (\boldsymbol{t}^d_x)'\mathbf{T}^{-1}\boldsymbol{t}_{xy} - t^d_y. \tag{14}$$

2.3 Simulation study

For the simulation an artificial population was created with to independent auxiliary variables, $X_1 \sim Bin(1, 0.7)$ and $X_2 \sim N(20, 3^2)$, so that the first auxiliary variable is binary variable. For the study variable the following rule was used:

$$Y_i = 2 * X_{1i} + 0.7 * X_{2i} + \epsilon_i, \quad i = 1, ..., 1000,$$

where $\epsilon \sim N(0, 1)$. The relationships between study and the auxiliary variables is strong enough, $Corr(X_1, Y) = 0.41$ and $Corr(X_2, Y) = 0.84$.

This population was divided into three domains of different sizes. Some basic characteristics of the domains are given below.

Table 1: Population characteristics, domain sizes

Domain	N_d	Mean, t^d_y/N_d	s^d_y
1	99	11.64	3.51
2	259	13.91	6.14
3	642	16.75	8.19
Population	1000	15.50	2.55

Domains are formed so that domain characteristics were different from the population ones. The simple random sampling procedure was used to produce estimates with $R = 1000$ repetitions and the sample size $n = 100$. The sampling was carried on the whole population, which gave the random sample sizes in the domains. In the average, sample sizes in the domains were $n_1 = 10$, $n_2 = 26$ and $n_3 = 64$. Two cases were observed in simulations, one with binary and another with continuous auxiliary variable.

The following characteristics were calculated:

the mean syn. estimator in the domain U_d: $\overline{\hat{t}^d_{syn}} = \frac{1}{R}\sum_{i=1}^{R} \hat{t}^d_{syn,(i)}$,

the mean bias of the syn. estimator: $b(\hat{t}^d_{y,syn}) = \overline{\hat{t}^d_{syn}} - t^d_y$,

the mean linearly estimated bias (from (14)): $\overline{\hat{b}^d} = \frac{1}{R}\sum_{i=1}^{R} \hat{b}(\hat{t}^d_{syn,(i)})$,

the mean linear estimator in the domain U_d: $\overline{\hat{t}^d_{lin}} = \frac{1}{R}\sum_{i=1}^{R} \hat{t}^d_{lin,(i)}$,

the st. deviation of the syn. estimator: $D(\hat{t}^d_{syn}) = \sqrt{\frac{1}{R}\sum_{i=1}^{R}(\hat{t}^d_{syn,(i)} - \overline{\hat{t}^d_{syn}})^2}$,

the st. deviation of the lin. estimator: $D(\hat{t}^d_{lin}) = \sqrt{\frac{1}{R}\sum_{i=1}^{R}(\hat{t}^d_{lin,(i)} - \overline{\hat{t}^d_{lin}})^2}$,

the root mean square error of (11): $RM(\hat{t}^d_{syn}) = \sqrt{\frac{1}{R}\sum_{i=1}^{R}(\hat{t}^d_{syn,(i)} - t^d)^2}$.

Results are given in Table 2 separately for the binary auxiliary variable X_1 and for the continuous auxiliary variable X_2.

Table 2: Simulation results

X_1-binary variable is used								
	t^d_y	$\overline{\hat{t}^d_{syn}}$	$b(\hat{t}^d_{y,syn})$	$\overline{\hat{b}^d}$	$\overline{\hat{t}^d_{lin}}$	$D(\hat{t}^d_{syn})$	$D(\hat{t}^d_{lin})$	$RM(\hat{t}^d_{syn})$
Dom. 1	1 153	1 053	-100	-84.6	1 122	17.5	334	102.1
Dom. 2	3 602	2 834	-768	-747	3 616	47.2	587	770.8
Dom. 3	10 754	7 450	-3 304	-3 364	10 780	124	803	3309.2
X_2-continuous variable is used								
	t^d_y	$\overline{\hat{t}^d_{syn}}$	$b(\hat{t}^d_{y,syn})$	$\overline{\hat{b}^d}$	$\overline{\hat{t}^d_{lin}}$	$D(\hat{t}^d_{syn})$	$D(\hat{t}^d_{lin})$	$RM(\hat{t}^d_{syn})$
Dom. 1	1 153	1 119	-34.2	-24.0	1 122	9.58	334	35.56
Dom. 2	3 602	3551	-51.5	-23.5	3 616	30.41	587	59.83
Dom. 3	10 754	10811	57.2	31.4	10 780	92.59	803	108.85

From the Table 2 it can be seen that results are more accurate (in a sense of bias and variance) for the continuous auxiliary variable. For both auxiliary variables, the synthetic estimator has a smaller variance than the linear estimator. In a case of the continuous auxiliary variable, where assisted linear regression model is well fitted for the sampled units, the root mean square error of the synthetic estimator is better that variance of the linear estimator. It is not so in the case of the binary auxiliary variable, where linear regression model is not the better one.

The estimated bias (14), where linear estimators are used for the unknown totals, is able to describe the majority of the bias. But due to the large variance of the linear estimator, it is not reasonable to use it in a practise in order to correct the synthetic estimator.

References

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

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

[3] Traat, I. (2000) Sampling design as a multivariate distribution. *New trends in Probability and Statistics*, vol. 5, pp. 195208

[4] 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

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

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

Restriction Estimator for Unidentified Domains

Kaur Lumiste[1]
[1] Tartu University, Estonia
e-mail: klumiste@ut.ee

Abstract

Domain estimation has become an important area in survey sampling, but a lot of problems are associated with it. For example small sample sizes in domains produce inaccurate estimates. Another problem with domain estimation is the lack of consistency between estimates. It is known that domain totals have to sum up to the population total. But this relationship does not always hold for respective estimates.

The inconsistency problem is studied in the dissertation Sõstra (2007) and restriction estimators for domains are derived which satisfy the summation restriction.

The goal of this paper is to give short overview of Kaur Lumiste's bachelor thesis (Lumiste, 2008) which was based on the dissertation and elaborates a specific case not considered in Sõstra (2007). We assume that there is an auxiliary variable in the sampling frame which helps us to construct a ratio estimator for population total. Also we assume that domains are unidentified in the sampling frame, meaning that we cannot find the true domain totals of the auxiliary variable. Therefore we cannot construct ratio estimators for domain totals, so linear estimator is considered for domains.

At first, the basic concepts and theorems are introduced and then step by step components of restriction estimator for domains are derived. The novel results are design-based covariance formulae of a ratio and of a linear estimator; their elaboration for domains, and the subsequent elaboration for simple random sampling design.

Later the concept of general restriction (GR) estimator (Knottnerus, 2003) is introduced. Restriction estimators for domains and population totals are constructed. They are based on the specific initial estimators studied in this thesis. The resulting restriction estimators for domains are consistent with the estimated population total.

Finally the results are tested in a simulation study, which shows that we can solve the inconsistency problem with the GR-estimators. Also by importing restrictions the variance of the estimators is reduced.

1 Preliminaries

1.1 Estimation of population parameters

Let $U = (1, 2, \ldots, N)$ denote a finite population of N units. Let a random vector (design vector) $\mathbf{I} = (I_1, I_2, \ldots, I_N)$ describe the sampling process on U. Elements I_i show the number of possible selections of the unit $i \in U$. For without-replacement

(WOR) designs $I_i \in \{1, 0\}$ and $I_i \in \{1, 2, \ldots\}$ for with-replacement (WR) designs. The distribution of $\mathbf{I}$ is the sampling design, $p(\mathbf{k}) = Pr(\mathbf{I} = \mathbf{k})$, where $\mathbf{k} = (k_1, k_2, \ldots, k_N)$ is an outcome of $\mathbf{I}$. The moments of $\mathbf{I}$, such as $E(I_i)$, $V(I_i)$ and $Cov(I_i, I_j)$ play a crucial role in finite population estimation theory. It is assumed that $E(I_i) > 0$, $\forall i$ for any sampling design. In the case of WOR design, the inclusion indicator I_i is a random variable with a Bernoulli distribution, $I_i \sim B(1, \pi_i)$. In this case

$$E(I_i) = \pi_i,\ V(I_i) = \pi_i(1 - \pi_i),$$
$$Cov(I_i, I_j) = \pi_{ij} - \pi_i\pi_j,$$

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

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

In this paper we use the unbiased estimator for the population total $Y = \sum_U y_i$. For any sampling design the unbiased estimator $\hat{Y}$ of Y and its variance are:

$$\hat{Y} = \sum\nolimits_U I_i \breve{y}_i = \sum\nolimits_U \omega_i y_i, \tag{1}$$

$$V(\hat{Y}) = \sum\sum\nolimits_U \Delta_{ij}\breve{y}_i\breve{y}_j, \tag{2}$$

where $\breve{y}_i = y_i/E(I_i)$ and $\Delta_{ij} = Cov(I_i, I_j)$. Provided that $E(I_i, I_j) > 0$, $\forall i \neq j$, the unbiased estimator of variance (2) is

$$\hat{V}(\hat{Y}) = \sum\sum\nolimits_U \breve{\Delta}_{ij}\omega_i y_i \omega_j y_j, \tag{3}$$

where $\breve{\Delta}_{ij} = \Delta_{ij}/E(I_iI_j)$ and ω_i is the design weight $\omega_i = I_i/E(I_i)$.

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

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

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

$$\hat{R} \approx R + \frac{1}{H}(\hat{Y} - R\hat{H}). \tag{4}$$

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

$$\hat{Y}^r \approx Y + (\hat{Y} - R\hat{H}). \tag{5}$$

The expansions (4) - (5) are used to derive approximate variance formulae.

1.2 Estimation of domain parameters

Let us assume that population U consists of D domains ($d = 1, 2, \ldots, D$). Let U_d be a domain with size N_d, $\sum_{d=1}^{D} N_d = N$. Let the sampling design in the population be $\mathbf{I} \sim p(\mathbf{k})$. Components I_i are a part of the design vector $\mathbf{I}$ where index $i \in U_d$ describes sampling in domain U_d. Sample size in U_d is $n_d = \sum_{U_d} I_i$, which is usually random even if the overall sample size $n = \sum_{d=1}^{D} n_d$ is fixed.

Expected sample size in U_d is

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

Let us define a domain indicator z_i^d:

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

and create a new variable y_i^d:

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

Using them we can directly apply all general formula brought earlier for estimation of population parameters. It is important to note that the domain total,

$$Y_d = \sum_{U_d} y_i$$

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

$$Y_d = \sum_{U} y_i^d = \sum_{U} z_i^d y_i.$$

An unbiased domain total estimator Y_d derives directly from (1):

$$\hat{Y}_d = \sum\nolimits_{U} \omega_i y_i^d = \sum\nolimits_{U} \omega_i z_i^d y_i = \sum\nolimits_{U_d} \omega_i y_i. \tag{6}$$

Domain size N_d is the total of a special variable $y_i \equiv 1$ in (6):

$$\hat{N}_d = \sum_{U} \omega_i z_i^d y_i = \sum_{U_d} \omega_i.$$

2 Simple random sampling without replacement

Under simple random sampling (SI) design all samples with fixed size are equally probable. Characteristics of the SI-design with population size N, sample size n and sampling fraction $f = n/N$, are for all $i, j \in U$

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

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

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

We now present the covariation formulas needed later.

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

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

where

$$\begin{aligned} S_{y^d y^g} &= \frac{1}{N(N-1)}\left[N\sum\nolimits_U y_i^d y_i^g - Y_d Y_g\right], \\ s_{y^d y^g} &= \frac{1}{N(N-1)}\left[N\sum\nolimits_U I_i y_i^d y_i^g - f^2 \hat{Y}_d \hat{Y}_g\right], \end{aligned}$$

$Y_d = \sum_U y_i^d$, $Y_g = \sum_U y_i^g$, $\hat{Y}_d = \sum_U \omega_i y_i^d$ and $\hat{Y}_g = \sum_U \omega_i y_i^g$

Corollary Under SI-design the approximate covariance between a ratio estimator of population total $\hat{Y}_0^r$ and a linear estimator of domain total $\hat{Y}_d$ and its unbiased estimator are:

$$\begin{aligned} ACov(\hat{Y}_0^r, \hat{Y}_d) &= N^2(1-f)S_{uy^d}/n, \\ \widehat{Cov}(\hat{Y}_0^r, \hat{Y}_d) &= N^2(1-f)s_{uy^d}/n, \end{aligned}$$

where

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

where $u = y_i - R_y h_i$ and $\tilde{u} = y_i - \hat{R}_y h_i$.

3 General restriction estimator

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

A longer introduction of GR-estimators can be seen in Knottnerus (2003) and Sõstra

(2007), here we only present the final formulas for SI-design and for our special case where we estimate the population total with a ratio estimator and domain totals are linear estimates.

Theorem. Under SI-design the GR-estimators for domain total Y_d and population total Y are:

$$\hat{Y}_d^{GR} = \hat{Y}_d + (\hat{Y}_0^r - \hat{Y}_*)(\sum_{i\in\mathcal{D}} S_{y^d y^i} - S_{uy^d})/v', \quad (7)$$

$$\hat{Y}_0^{GR} = \hat{Y}_0^r + (\hat{Y}_0^r - \hat{Y}_*)(\sum_{i\in\mathcal{D}} S_{uy^i} - S_{uu})/v', \quad (8)$$

where

$$v' = \sum\sum_{\mathcal{D}} S_{y^i y^j} + S_{uu} - 2\sum_{i\in\mathcal{D}} S_{uy^i},$$

$\hat{Y}_* = \sum_{\mathcal{D}} \hat{Y}_i$ and $\mathcal{D} = \{1, 2, \ldots, D\}$.
GR-estimators variance and covariances under SI-design are:

$$\begin{aligned}
V_{dd}^{GR} &= c[S_{y^d y^d} - (\sum_{i\in\mathcal{D}} S_{y^d y^i} - S_{uy^d})^2/v'] \\
V_{00}^{GR} &= c[S_{uu} - (\sum_{i\in\mathcal{D}} S_{uy^i} - S_{uu})^2/v'] \\
V_{dg}^{GR} &= c[S_{y^d y^g} - (\sum_{i\in\mathcal{D}} S_{y^d y^i} - S_{uy^d})(\sum_{i\in\mathcal{D}} S_{y^g y^i} - S_{uy^g})/v'] \\
V_{d0}^{GR} &= c[S_{uy^d} - (\sum_{i\in\mathcal{D}} S_{uy^i} - S_{uu})(\sum_{i\in\mathcal{D}} S_{y^d y^i} - S_{uy^d})/v']
\end{aligned}$$

where

$$\begin{aligned}
c &= N^2(1-f)/n, \\
S_{y^d y^g} &= \frac{1}{N(N-1)}\left[N\sum\nolimits_U y_i^d y_i^g - Y_d Y_g\right], \\
S_{uy^d} &= \frac{1}{N-1}\sum\nolimits_U u_i y_i^d, \\
S_{uu} &= \frac{1}{N-1}\sum\nolimits_U u_i^2,
\end{aligned}$$

$u_i = y_i - R_y h_i$, $Y_d = \sum_U y_i^d$ and $Y_g = \sum_U y_i^d$.

We see from (7)-(8) that the summation restriction is satisfied,

$$\sum_{d\in\mathcal{D}} \hat{Y}_d^{GR} = \hat{Y}_0^{GR}.$$

4 Simulations

We now present the results of the simulation study that was done to test the derived GR-estimators for SI-design.

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

- monthly salary (a continuous variable: in thousand kroons);
- household size (number of persons in the household) - auxiliary variable;
- domain indicator d ($d = 1, 2, 3$) - auxiliary variable.

Tables 1 and 2 show the main characteristics of the population. It can be seen that domains have different sizes. The first domain is the largest and the third domain is the smallest. The average household size is approximately the same in each domain.

Table 1: Domain sizes

Domain	nr. of persons	%	Average HH size
1	1019	51.0	1.69
2	733	36.6	1.66
3	248	12.4	1.70
Population	2000	100.0	1.68

Table 2: Monthly salary characteristics ($\times 10^3$)

Domain	Total	Mean	Minimum	Maximum	Std
1	4998.9	4.906	0.100	37.580	3.194
2	4614.8	6.296	0.500	46.660	4.309
3	1396.3	5.630	0.200	23.560	3.285
Population	11010.1	5.505	0.100	46.660	3.707

From Table 2 we can see, that mean and variance of monthly salary are highest in the second domain.

Table 3: Sample sizes

Domain	Mean	Minimum	Maximum
1	101.8	75	124
2	73.4	49	95
3	24.8	12	41
Population	200	200	200

Table 4: Covariance matrix of initial estimations over simulations

Estimates	$\hat{Y}_0^r$	$\hat{Y}_1$	$\hat{Y}_2$	$\hat{Y}_3$
$\hat{Y}_0^r$	350056	75766	162860	16657
$\hat{Y}_1$	75766	216989	-116263	-26787
$\hat{Y}_2$	162860	-116263	311380	-36916
$\hat{Y}_3$	16657	-26787	-36916	80655

1000 independent samples were drawn from the population by SI-sampling. The sample size was 200 persons and the sample sizes for domains were random. Table 3 shows average, minimum and maximum sample sizes in domains.

With each sample we calculated the initial estimators. For domain totals we used the linear estimator and for population total a ratio estimator, where the household size was the auxiliary variable. Then we calculated the GR-estimators on each sample using formulas (7) - (8). Our interest is concentrated on two questions:
Do the initial estimations sum up to the population total and if not, then can the problem be solved using the GR-estimations?
Are GR-estimators more accurate than the primary estimators?

The sum of initial estimates of domain totals calculated from samples are not consistent with the estimate of population total. Assuming that the covariance matrix $\mathbf{V}$ is known we found the GR-estimates and they where consistent.

Table 5: GR-estimations covariance matrix over simulations

Estimates	$\hat{Y}_0^{GR}$	$\hat{Y}_1^{GR}$	$\hat{Y}_2^{GR}$	$\hat{Y}_3^{GR}$
$\hat{Y}_0^{GR}$	248849	74012	157919	16617
$\hat{Y}_1^{GR}$	74012	217162	-116312	-26837
$\hat{Y}_2^{GR}$	157919	-116312	311145	-36913
$\hat{Y}_3^{GR}$	16617	-26837	-36913	80668

Covariance matrices for initial estimators, $\mathbf{V}$, and GR-estimators, $\mathbf{V}^{GR}$, are shown in Tables 4 - 5. Comparing variances we can see, that the variance of GR-estimator for population total has significantly reduced. Although there is a slight increase in some variances of estimated domain totals variance (the difference is less than 0,08% from initial estimates) we can say that the variances remained nearly the same.

Table 6 contains the means, standard deviations and minimum and maximum of estimates of both initial estimates and GR- estimates over all samples. Comparing it to Table 2 we can see that means are almost the same with the true totals for both estimation methods. That is as expected since the linear estimates are unbiased and the ratio estimate is asymptotically unbiased. Comparing standard deviations we witness

Table 6: Initial and GR-estimated parameters over simulations

	Domain	$\hat{\bar{Y}}_d$	Std	$min(\hat{Y}_d)$	$max(\hat{Y}_d)$
Initial	1	4984	465.8	3683	6548
estimates	2	4625	558.0	2910	6303
	3	1395	284.0	656	2390
	Population	11009	591.7	9259	13188
GR-estimates	1	4984	466.0	3675	6541
	2	4625	557.8	2920	6314
	3	1395	284.0	651	2389
	Population	11003	498.8	9397	12773

the same effects as described earlier for variances. GR-estimates leave the standard deviations for domains unchanged, but significantly decrease standard deviation of the estimated population total.

References

Sõstra K. (2007) *Restriction Estimator for Domains.* Tartu likooli kirjastus, Tartu.

Lumiste K. (2008) *Restriction Estimator for Unidentified Domains* (in Estonian). Tartu University

Knottnerus P. (2003) *Sample Survey Theory. Some Pythagorean Perspectives.* Springer, New York.

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

IMPROVEMENT OF RELIABILITY OF LFS INDICATORS MONTHLY ESTIMATES

Olha Lysa

Institute for Demography and Social Research, Ukraine
e-mail: Olysa@ukr.net

Abstract

The paper presents main results of investigations on problem of reliable estimation of monthly unemployment indicators for regional level by the data of state household sample survey of population economic activity that fulfilled by State Statistics Committee of Ukraine. The approach to estimator construction for indicators estimation based on the method of composite estimation is considered. A composite estimator is constructed on the basis of direct estimator and indirect estimators, received as composite estimator of the previous month.

1 Introduction

Indicators of employment and unemployment are the basis for development of the well-grounded social and economic policy and estimation of its efficiency. Unemployment level side by side with a real internal product and inflation are widely used as the general indicators of the current condition of country economy. The information needs about the condition and tendencies on the labour market constantly grow and first of all about labour force characteristics. Thus information necessity exists both on international, and at the state, regional and local levels. According to modern international standards of statistical information quality (Helsinki, 2002), one should be characterized by maximal completeness and timeliness, should corresponds to users needs, should be reliable, accessible and clear, comparable in time and in the space, coordinated with the available comparable data from other sources. Also the important aspect is expediency, optimality of expenses financial and manpower resources on data obtaining.

Household sample surveys are most widespread and recognized in the world the way for reception of the information concerning labour force. Unconditional advantage of this method is integrated approach and completeness of the data, flexibility and ample opportunities for the analysis.

The state sample surveys, by results of which economic activity, employment and unemployment indicators are measured, mainly provide an opportunity of their reliable estimation on the nation-wide level. The estimates received for lower levels in many cases are insufficiently reliable and demand application of special approaches for more precise definition. It is typical also for indicators estimation of separate social and economic groups of the population. Therefore in the state statistics of Ukraine more and more attention is given to the problem of calculation of reliable estimates of these indicators for the regional level (regions, districts, separate cities).

2 Measurement of labour force indicators

In Ukraine labour force indicators are measured on the basis of household sample survey of population economic activity that fulfilled by State Statistics Committee of Ukraine (SSCU) on the constant basis from 1995. Since the survey program was repeatedly changed with the purpose of the account of ILO and EC recommendations as well as increase of survey

efficiency for the fullest satisfaction of users' needs. First it was carried out once an year, but it did not allow revealing and displaying dynamics of changes at national labour market. Therefore from 1999 quarterly LFS was implemented under the new advanced program and 2004 transition has been carried out to monthly survey *(Council Regulation, 1998)*.

State sample surveys of population economic activity in Ukraine are carried out on the basis of interview of household non-institutional set that is formed on the procedure of stratified multistage random selection. From 2004 for monthly survey 11,1 thousand households are selected that represent all regions of Ukraine. With the purpose of reliability increase of economic activity, employment and especially unemployment indicators in rural areas on the basis of sample of household agriculture activity survey 7,4 thousand households are selected additionally for interview on the program of LFS.

In every household that have taken part in survey demographic data on all household members and special information concerning economic activity of household members age 15 – 70 years were collected within a week which preceded surveyed week.

Sample set represents all population of Ukraine except for military men, persons who are in places of imprisonment, persons who constantly live in boarding houses of different type, and also marginal layers of the population (homeless, etc.). Besides that territories which concern to I and II zone of radiating pollution because of accident on the Chernobyl APS have not got into sample set.

Population stratification is realized with the purpose of adequate reflection the basic features of administrative and territorial division in the sample as well as for maintenance of selection from more homogeneous under the characteristics of households' sets. Accordingly inside each of 25 regions of Ukraine allocate two strata: urban settlements (cities and towns) and rural administrative districts; besides city Kyiv and city Sevastopol are surveyed. The sample size is distributed on strata proportionally number of population.

All regions are represented in the sample almost equally, part of sample by regions is varied from 0,06% to 0,08%. The variation is accounted for necessity of the greater representation of some regions due to others because of the small sizes of these territories that makes impossible calculation of reliable estimations of general indicators for ones.

Household sample set consists of six approximately equal parts – rotational groups, each from which is representative sample from population. Monthly two rotational groups (third of set) are replaced by two others groups – one new group and rotational one that has been surveyed in corresponding month of the last year.

In other words, the SEAP uses rotation scheme 3-9-3. Each household is surveyed at six periods: three months running, nine months isn't surveyed and after that it is surveyed three more months. At such rotation scheme two thirds of this sample passes to next month and in one year half of sample of the current month comes back (see fig. 1).

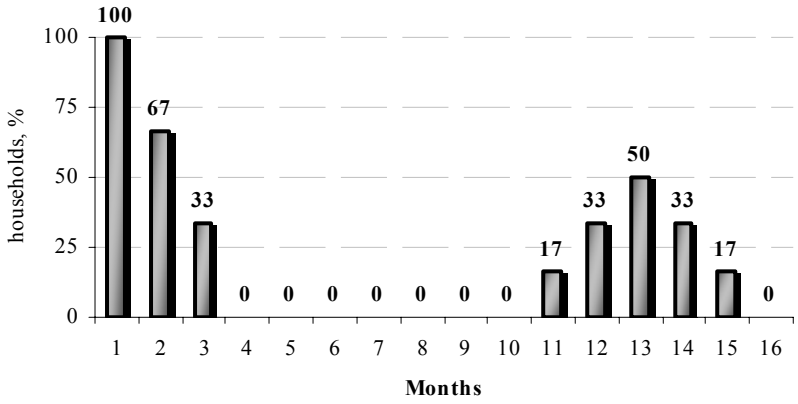


Figure 1. The scheme of household staying in the sample of SEAP 2004-2008

Realization of household rotation in the sample allows increasing the reliability level of indicators comparison in dynamics, not overloading any separate households group with too

long period of interview. On the one hand it provides renovation of sample set for each period of supervision, on the other – allows keeping constant its determined part. Besides use of such rotation scheme allows to measure changes of labour force indicators more reliably from year to year and to carry out comparisons of the corresponding periods of two next years.

3 Direct estimation

By results of LFS the estimates of labour force indicators that characterize all population in the age of 15–70 years are calculated. For receiving of labour force indicators estimates, which represent all population, multistage procedure of the statistical weights system calculation and calibration (poststratification) is realized that includes:

- the account of the general probabilities of households selection;
- the account of the actual level of households and persons refusals;
- association of the data received under different interview programs;
- harmonization of survey results with the data of demographic statistics concerning number and sex-age structure of the population.

Direct estimates of indicators which calculated directly by the survey results take into account statistical weights of respondents (households and persons). The estimator for calculation of indicators estimates at presence of additional information is defined as *(Ghosh M., Rao J.N.K., 1994)*:

$$\hat{Y}^{(D)} = \hat{Y}^{(HT)} + \sum_{d=1}^{D} \hat{\beta}_d \left(X_d - \hat{X}_d^{(HT)} \right), \qquad (1)$$

where $\hat{Y}^{(D)}$ – direct estimator; $\hat{Y}^{(HT)}$ – Horvitz- Thompson estimator; $\hat{\beta}_d$ – estimates of regression coefficients (d =1,2,…,D – number of auxiliary data); X_d – values of indicators estimates, known from external sources; $\hat{X}_d^{(HT)}$ – Horvitz- Thompson estimates for auxiliary data.

The Horvitz-Thompson estimators are:

1) for totals :

$$\hat{Y}^{(HT)} = \sum_{i=1}^{n} w_i \, \mathrm{I}_{i \in S} \, , \qquad (2)$$

where w_i – statistical weights; $\mathrm{I}_{i \in S}$ – indicator of the belonging of respondent to the certain group S (employed, unemployed, etc.); n – sample size.

2) for rates:

$$\hat{Y}^{(HT)} = \frac{\sum_{i=1}^{n} w_i \, \mathrm{I}_{i \in S_1}}{\sum_{j=1}^{n} w_j \mathrm{I}_{j \in S_2}} . \qquad (3)$$

Estimates of general indicators of labour force are calculated on the basis of estimator (1). They are: number of economically active, employed and unemployed population as well as rates of economic activity, employment and unemployment and so on. Data of last population census, current data of demographic statistics, and current data of social statistics as for number and placement of institutional population are used as external information. One is used by the calibration of statistical weights system for what special procedures are developed *(Sarioglo, 2005)*.

At the estimation of LFS indicators reliability method of balanced repeated replications is used. Taking into account recommendations of International Labour Organization, Convention and Resolutions of international conferences on labour statistics by bodies of official statistics in the SSCU next standards concerning sample estimates reliability have been adopted:

- if CV ≤ 10%, indicators estimates are reliable and can be used for quantitative analysis;
- if 10% ≤ CV ≤ 25%, estimates are suitable only for qualitative analysis.

Special research of estimates quality for labour force indicators has shown, that mid-annual, quarter and month estimates indicators of economic activity and employment are suitable for quantitative analysis both on nation-wide, and at the regional level except of month estimates for Sevastopol city, where $CV \geq 10\%$ (for national level $CV < 1{,}5\%$, for regional level $CV < 6\%$); mid-annual estimates of unemployment rates can be used for quantitative analysis on nation-wide level, $CV < 3\%$ (table 1). As to regional level, for 7 regions from 27 indicators estimates are reliable, for other 19 regions CV doesn't exceed 19% and only for Sevastopol city $CV \geq 25\%$; quarter estimates of unemployment indicators can be used for quantitative analysis on nation-wide level ($CV < 5\%$), at the regional level it is possible to use only estimates for separate regions; month estimates of unemployment indicators can be used for quantitative analysis on nation-wide level ($CV < 5\%$). At the regional level indicators estimates aren't suitable, in most cases $CV \geq 15\%$ (see, for example, fig. 2).

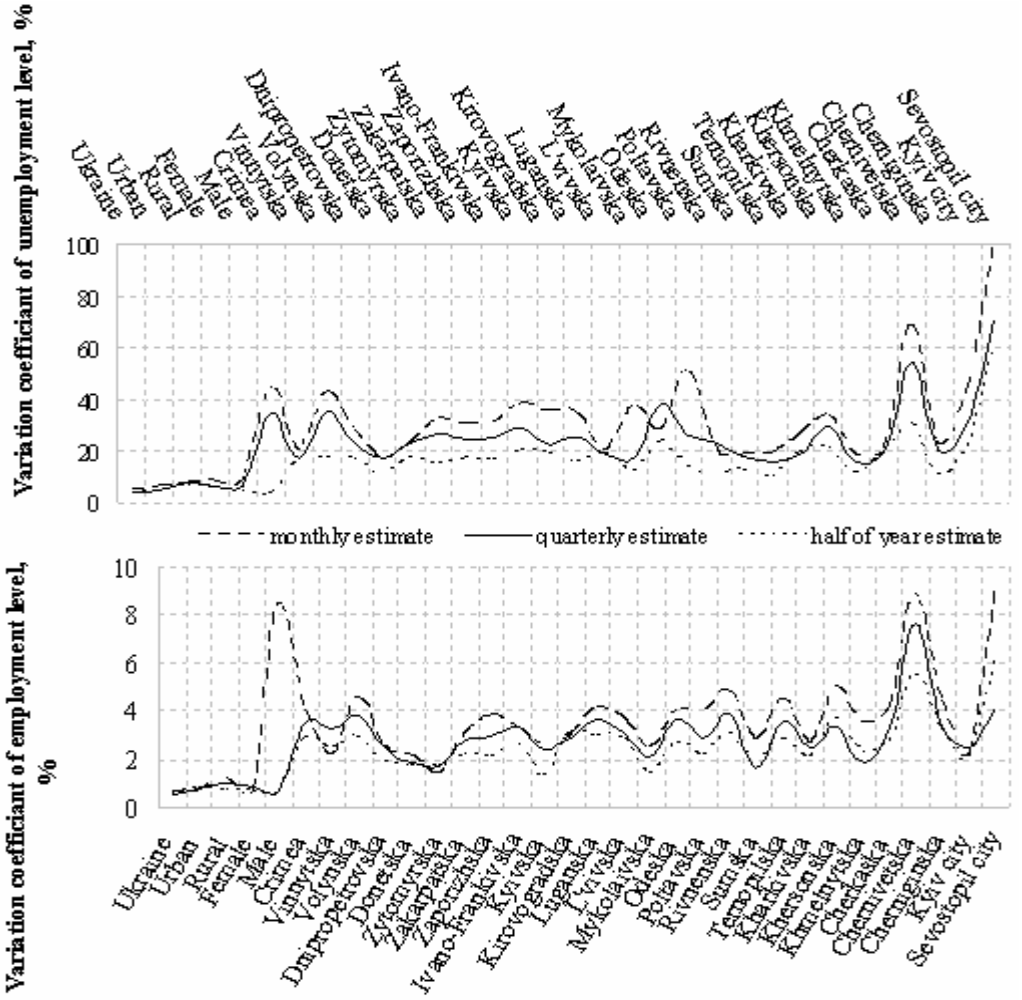


Figure 2. Variation coefficients of monthly, quarterly and half of year estimates of employment and unemployment rates (LFS, 2007)

This figure presents variation coefficient of monthly, quarterly and half of year estimates of employment and unemployment rates for national, sub national and regional levels for April

and accordantly quarter and half of year 2007. From the presented data it is evident that employment rates estimates have low variation of indicator values for all aggregation levels; unemployment rates estimates are reliable only for national level, but on the regional level all estimates are suitable only for qualitative analysis.
On the basis of reliability rate analysis of state LFS data it is possible to draw the conclusion: for regional level direct indicator estimates are less accurate than for Ukrainian level: it is caused first of all by considerably smaller sample size for each separate region. And for indicators of population economic activity and employment accuracy of received estimates, for example for regions, is satisfactory. Low reliability of population unemployment indicators is also caused by rather small values of the indicator.

4 Indirect estimation

On the basis of studying results of existing approaches to increase of data reliability level of household sample surveys it is necessary to draw a conclusion that now is expedient complex use of different ways:
- sample design optimization in view of the small area;
- use of statistical weight calibration procedure for account of the existing reliable external information;
- use of the small area methods at indicators estimation;
- use of coordination procedure for small area estimates and direct estimates for different levels of data aggregation.

During realization of the state surveys which are carried out on the constant basis, the design of sample changes rather seldom. For surveys of the labour force the design can be constant during five and more years. Thus, this approach isn't effective at the decision of the current problems with reliability of separate parameters.
Among methods of sample design development, used for improvement of indicators direct estimations for small areas without sample size increase, it is necessary to note the next (*Rao J.N.K, 2003*):
- increase of strata quantity at population stratification that provide more detail areas presentation in sample;
- decrease of clusters size for cluster sample;
- reallocation of sample volumes on the areas with the purpose of homogenization of indicators estimations quality and so on.

When forming the new territorial sample for LFS in 2004-2008 the strata quantity in rural area was increased from 25 (as it was in sample 1999-2003) to 490 that corresponds to quantity of administrative rural rayons all over Ukrainian regions. With the help of modeling methods on the basis of existing sets data of LFS the improvement degree of indicator estimates reliability with increase of strata quantity was estimated. According to the results of analysis, the main stratification effect has taken place for rural area; estimations reliability of unemployment level was increased on the average in 1.35 times here (*Sarioglo, 2003*).
The main feature of existing approaches for surveys that are carried out on the constant basis (monthly, quarterly, annually) is use of the information from additional data sources. One of approaches to estimator construction for indicators estimation is the method of composite estimation.
In this paper as composite estimator is considered estimator received on the basis of direct estimator and indirect estimators, received on the basis of a method synthetic estimation or a method of model base estimation. Also indirect estimator can be constructed based on the LFS information for previous months. The reliability improvement of monthly estimations is arrived at due to high correlation of economic activity indicators between consecutive months (see, fig. 3) on 67% of same sample units.

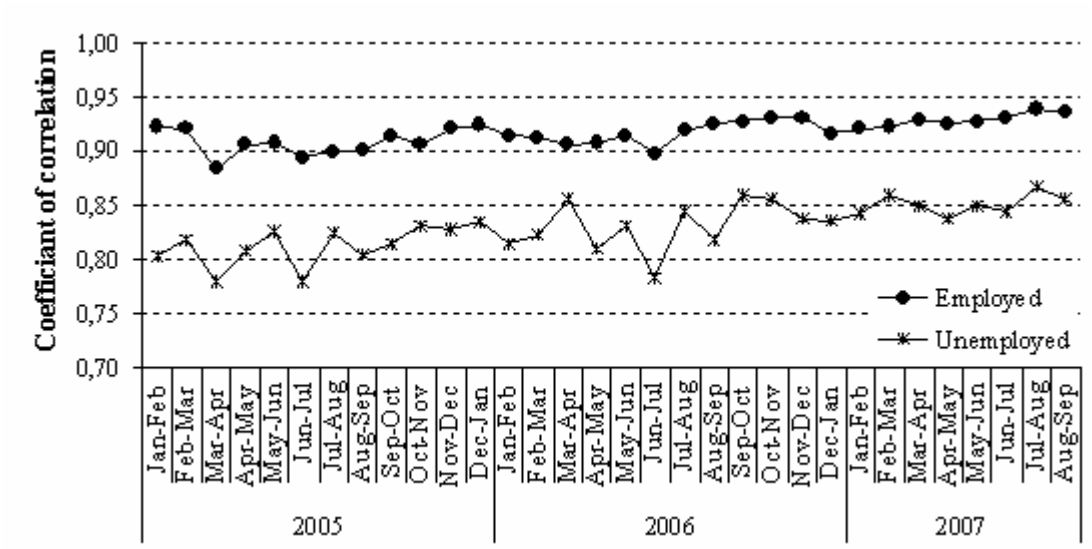


Figure 3. Coefficients of correlation between consecutive months (LFS, 2005-2007)

Composite estimator for unemployment level (number of unemployed) in view of monthly rotation for month *t* is *(Local Area Unemployment Statistics, 2001; Design and Methodology, 2000)*:

$$Y_t' = (1-K)\hat{Y}_t + K(Y_{t-1}' + \Delta_t), \tag{4}$$

where: Y_t' – indicator composite estimate for the current month; $\hat{Y}_t$ – direct estimate for the current month; Y_{t-1}' – composite estimate of the previous month; Δ_t – estimate of indicator changing concerning the last month that received on the basis of 4 rotation groups data, which are the common for months t and $t-1$; K – weight coefficient.

$$\hat{Y}_t = \sum_{i=1}^{6} x_{t,i}, \tag{5}$$

$$\Delta_t = \frac{4}{3}\sum_{i \in S}(x_{t,i} - x_{t-1,i-1}), \tag{6}$$

i = 1,2, …, 6 – number of rotation group for sample of current month; $x_{t,i}$ – the sum of the weights, for example unemployed, for month t and rotation group i; $S = \{2,3,5,6\}$ – rotation groups, which went from the last month.

The value of coefficient K is defined from the condition of variance minimization of indicator composite estimate:

$$V(Y_t') = (1-K)^2 \cdot V(\hat{Y}_t) + K^2 \cdot V(Y_{t-1}' + \Delta_t) \to \min. \tag{7}$$

The results of the special researches showed expedience of the use of constant weighing coefficients during a year. Thus, for every cluster it follows to use the separate weighing coefficient (see fig. 4).

This method of composite estimation of labour force indicators was realized on micro-level with the procedure of reweighing which account received composite estimates of employment, une4mployment and non-in-labour force levels. It give the opportunity to improve the reliability of estimates on the basis of LFS micro-data set. Calculations of are carried out for everyone domain separately: urban and rural area, female and male and region of Ukraine – in total there are 31 domains. The size of domains for unemployed changes from 2 to 47 persons on regional level and from 174 (female) to 326 (rural area) persons on national level; for employed changes from 612 to 1313 persons on regional level and from 6773 (urban area) to 326 (rural area) persons on national level. For most domains it is not

enough for reception of reliable estimates of employment levels, and furthermore unemployment.

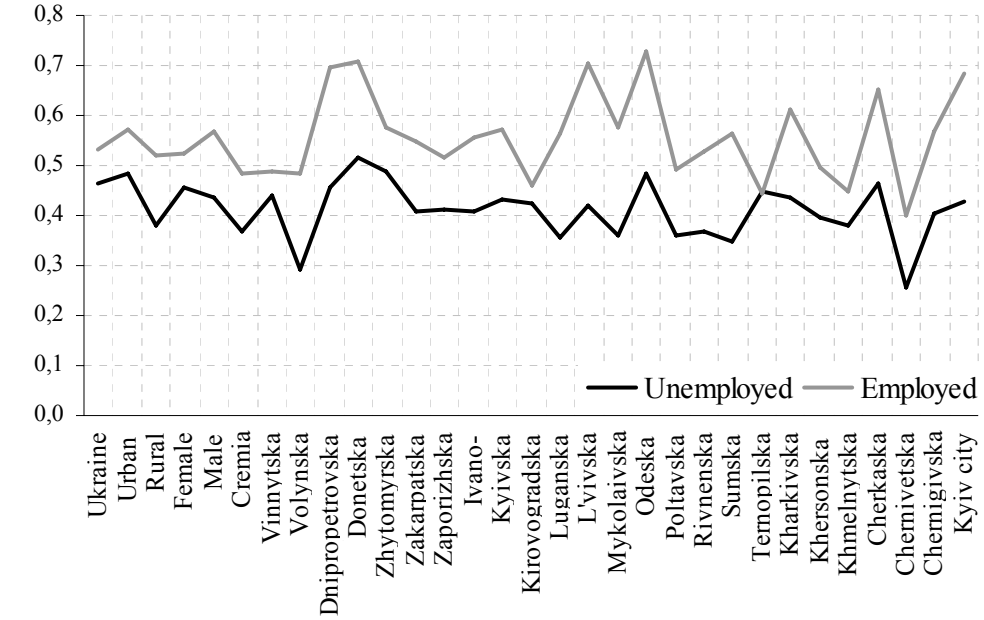


Figure 4. Distribution of weighting coefficients for employment and unemployment levels (LFS, 2007)

5 Results

Results of application of the method of composite estimations on the basis of the data for previous survey periods testify to efficiency of use of this approach to solving the problem of reliability improvement for results of LFS (see, fig. 5).

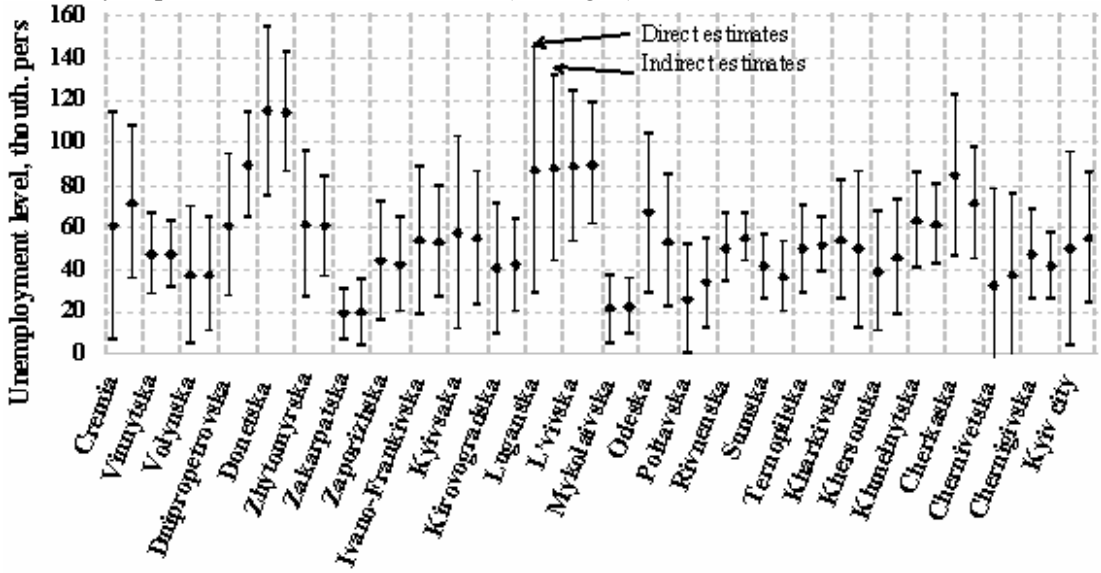


Figure 5. Comparison of direct and indirect estimates of employment and unemployment levels and their confidence intervals for regions of Ukraine (LFS, April, 2007)

Fig. 5 shows direct and indirect estimates and their confidence intervals for regional domains. As a whole estimates reliability improvement for employment and unemployment levels is observed for all regions of Ukraine on the average on 33.3 % and 19.6 % accordingly. As

research shows the estimates reliability is improving when the weighting coefficient not higher then 0.5. At that time, for separate regions the reliability level is not enough, it is needed the additional research and approaches.
At use of small areas methods the important problem can become the necessity of estimates coordination, because the estimates can be calculated on the base of survey data (direct estimates) and of some estimator specially constructed for small area.
As a whole the method of composite estimation allows to improve the reliability level of estimated indicators basically due to use additional data with sufficient reliability level. Therefore this method can be considered as one of the basic for solving small areas problems.

6. Conclusion

Research in direction of reliability improvement of regional estimations of labour force indicators in Ukraine are on the initial stage without regard to that the problems of development of the applied methods of statistical estimation of employment and unemployment on regional, subregional and local levels become more actual.
Application of the method of composite estimations allows increasing reliability of estimation of labour force indicators on average in 1.5 and 1.3 times accordingly on the basis of sample survey results for previous periods. At the same time completely to solve the problem of reliability without use of the external information and more effective models apparently is impossible. More effective are methods that developed on the basis of explicit statistical models, in particular models that take into account estimates of interterritorial variation of target indicators.

References

Council Regulation (EC) No 577/98 of March 1998 on the organization of a labour force sample survey in the Community. *Official Journal of the European Communities* 14.3.98.

Design and Methodology. (2000) Current Population Survey: *Technical Paper.* Washington: U.S. Department of Labor, BLS.

Ghosh M., Rao J. N. K. (1994) Small Area Estimation. *An Appraisal, Statistical Science.* **Vol. 9, № 1,** 55–93.

Kish L. (1995) *Survey sampling*. Wiley, New York.

Local Area Unemployment Statistics, *Program Manual*, U. S. Department of Labor, Bureau of Labor Statistics, Revised August 2001.

Quality Guidelines for Official Statistics. Helsinki: Statistics Finland, 2002.

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

Sarioglo V.G. (2005) *Problems of the sample data statistical weighting.* SSCU, Kyiv.

Sarioglo V. (2003) Methodological Approaches to Increase of Data Reliability Level of Sample Surveys of Population Economic Activity *Theory of Stochastic processes***. Vol. 9(25), № 3–4,** 176–183.

Small area estimation in victimization surveys

Måns Magnusson[1]

[1] Swedish National Council for Crime Prevention, Sweden
e-mail: mans.magnusson@bra.se

1 Introduction

This paper is intended to be the foundation of a research application this fall. Therefore all constructive criticism and ideas are very welcome.

1.1 Background

Today, both in Sweden and internationally, there is an increased demand for statistical estimates at regional levels. The use of estimates can shift between different users, but often estimates are demanded for public policy, decision making and benchmarking. When it comes to register based surveys or censuses, region based estimates are less of a problem since all elements are observed. In sample based surveys on the other hand, regional estimates become a problem when estimates are sought for smaller domains than the survey was originally intended for.

One of Sweden's largest victimization surveys, Nationella Trygghetsundersökningen (NTU), has a sample size of 20 000, but when it comes to estimates at lower regional levels as counties, many counties have a sample size of just 500 elements. One of the primary user of the regional estimates is the Swedish police who wants to use the survey for benchmarking between different counties since this is the level of organisation of the Swedish police. The estimates that are demanded is also often estimates concerning very rare events. For example only 0.8 percent of the respondents reported that they had experienced robbery and only 0.9 percent of the Swedish households have reported experience of burglary.

The increased demand of small area estimates has stimulated a new branch in sampling theory where new methods have been developed to produce small area estimates that have a better precision than direct estimates. In official statistics research is being made in labour force surveys and in demographic surveys, but in victimization surveys less research has been made.

Some studies have been made regarding small area estimation in victimization surveys. A study in Holland, Buelens (2009), used methods of small area estimation models to estimate the true victimization rate when the GREG-estimator failed. In this case the problem is similar to that of the Swedish victimization survey. The size of the sample in different police zones was approximately 750 but was still not large enough to enable estimates for year to year comparison. By using the number of violent crimes reported to the police new small area estimates were tested. Even though many violent crimes are never reported, Buelens showed that by using the number of reported crimes in an area as auxiliary information, it was possible to reduce the confidence intervals of police zone estimates by as much as 40 percent and thereby enable year to year comparisons.

In the area of small area estimation the two main components used to increase the precision of small area estimates are a linking model and good auxiliary information (Rao 2003). In criminological research there has been a tradition to model criminal events as Poisson

processes (Piquero et al. 2003). The Poisson distribution, used as a linking model, together with auxiliary information from administrative police registers in Sweden makes small area estimation in victimization surveys an interesting and demanded topic for research.

1.2 Purpose of study

The main purpose of the study should be to examine different model based estimators of small area estimation in victimization studies. Estimates of interest are totals and proportions and the main region level of interest are estimates of victimization rates for counties to enable comparisons between counties as well as for year to year comparisons. The main areas are estimates for different types of offences that often very rare events, such as battery and assault, robbery, burglary and sexual offences.

Even though small area estimates may have a better precision than direct estimates, the purpose is to evaluate whether the quality of small area estimates is of sufficient quality for the users of the statistics (i.e. enable year-to-year comparisons at county level, as well as benchmarking between different counties) and if the estimates are possible to include in production of official statistics.

1.3 Research questions

What linking models and what auxiliary information can be used for small area estimation of rare events in victimization surveys and what estimates have the better characteristics with regard to different user needs as well as precision?
How do small area estimates compete with direct estimates of rare event victimization?
Is it possible to produce small area estimates of sufficient quality to be used in official statistics?

References

Buelens, B. (2009) Estimating regional violent crime victim rates using police registers. abstract online: http://cio.umh.es/sae2009/AbstractList.asp#c24

Piquero, Alex R., David P. Farrington and Alfred Blumstein (2003). The criminal career paradigm. In: Tonry, M. (ed.), *Crime and Justice: A Review of Research*, **30**, 359–506.

Rao, J. N. K. (2003). *Small area estimation*. Wiley-Interscience, Hoboken, N.J.

COURSE OF SAMPLE SURVEYS FOR STUDENTS OF ECONOMIC SPESIALITIES

Tetyana Manzhos[1]

[1] Vadim Getman Kyiv National Economic University, Kyiv
e-mail: tmanzhos@gmail.com

Abstract

Historical data about Vadim Getman Kyiv National Economic University is presented. It is given information about different directions of training economists in all faculties. Mathematical subjects for students of the University are presented. Ways of introduction of special course on Sample Survey for the students of economical specialities are given.

1 History of Vadim Getman Kyiv National Economic University

The University known now as Vadim Getman Kyiv National Economic University started from commercial courses in 1906. In the end of the XIX century industry and agriculture started to develop quickly in Russia and also in Ukraine as a part of Russian Empire. It was appeared a question about the training of specialists in economics at regional level. Such specialists were supposed to know features of the economical development of region, commercial relations, traditions, and the way of life of population. The professor of st. Volodymyr Kyiv University Mytrofan Dovnar-Zapol'skiy (1867-1934) became the initiator of creation of such educational institution. It was needed three years to prove the need of creation of such educational establishment. And in February, 1906 Ministry of Commerce and Industry allowed him to open the quadrennial Higher Commercial Courses in Kyiv. History of our University began since this moment.

Soon became clear, that commercial courses could not satisfy the needs of Ukrainian economy and trade. In July 1908 The Kyiv Higher Commercial Courses were regenerate to the Kyiv Commercial Institute. There specialists in two faculties (economic and commercial) were trained not only for the needs of Ukraine but also for Caucasus. In June 1912 the institute got status of the Kyiv State Commercial Institute. An educational and financial base became better; students got the rights of the graduating students of state universities. The team of teachers was formed, changes were made to the educational process and practical studies of students, a number of educational auditoriums and laboratories were created, a library was organized. Student scientific activity got development. The First All-Ukrainian congress of economists and statisticians was initiated by the Kyiv Commercial Institute in October, 1918.

In 1920 the Kyiv commercial institute was reformed to the Kyiv Institute of Socio-Economic Sciences, and in September, 1920 the institute was renamed to the Kyiv Institute of National Economy. It included the economic, the co-operative, the industrial, the exploitations of roads faculties and faculty of law. In 1930 three independent institutes were created on the base of university. One of them was Kyiv Financial-Economic Institute.

In 1934 the Kyiv Financial-Economic Institute was removed in Kharkiv, where it was united with the Kharkiv Financial-Economic Institute. So, there was created the Ukrainian Financial-Economic Institute. It had functioned in Kharkiv till 1941.

Years of Great Patriotic War (1941-1945) were the most difficult for the Ukrainian Financial-Economic Institute, as well as for all people. The 1941-1942 studying year did not start. Almost 400 students and teachers (the day before war 600-700 students studied and 45 teachers worked in the institute) voluntarily went to war. During years of occupation of Kharkiv, fascists had destroyed the institute (educational corps, laboratories, students' dormitories); an institute library was burned out. After the liberation of Kharkiv in autumn in 1943 the Ukrainian Financial-Economic Institute continued to work in Kharkiv. In 1944 the government of Ukraine made decision on returning the Ukrainian Financial-Economic Institute to Kyiv. Then it returned the name Kyiv Financial-Economic Institute.

1945-1965 years have the special place in the history of the University. After the war 1941-1945 years it was needed to form economic potential, to prepare the new specialists and workers of all professions, to give the impulse to development economic and administrative structures. Development of University had begun since then. In 1955 there were over 3 thousand students, 110 teachers including 3 professors (doctors of sciences) and 61 associate professors (candidates of sciences). There had been already 5 thousand students in 1960 year, and in the end of 1985 – 13 thousand students. The institute grew up step by step in accordance with the needs of national economy into the institute of many profiles. The name of the institute didn't meet its real essence. At that moment there were 12 specialities, but they were not only of financial and economic profiles. In 1960 Ministry of Higher Education of Ukraine made a decision to rename the institute to the Kyiv Institute of National Economy composing five faculties: industrial economics, agricultural economics, financial-economical faculty, faculty of accounting and economical-statistic faculty. By the decision of Cabinet of Ministers of Ukraine in 1992 the institute was transformed into the Kyiv National Economic University.

Long-term experience of research work and relations of the University scientists with scientific institutions of Ukraine and foreign countries provided creation of university scientific school in record-keeping, audit and analysis; credit, banking and money circulation; state finances and taxes; statistics; economic theory; economics of enterprises and others. Wide creative relations were formed with the universities of Germany, Great Britain, Austria, Canada, France, USA and Holland. The university also became the active participant of the European programs TASIS, INTAS, TEMPUS.

On July, 11, 2005 the name of Vadym Get'man was given to the Kyiv National Economic University by the decision of Cabinet of Ministers of Ukraine. This prominent figure, financier of Ukraine also began his career as a student of this university. Vadym Get'man is one of authors of conceptions of creation the domestic pay system and principles of functioning of the system of electronic payments. He was also the organizer of creation of national currency – hryvnyas.

2 Directions of training

2.1 Faculties of University

Now Vadym Get'man Kyiv National Economic University consists of nine faculties:

- Faculty of Economics and Management;
- Faculty of International Economics and Management;
- Faculty of Law;
- Faculty of Personnel Management and Marketing;

- Faculty of Accounting;
- Faculty of Economy of Agro-industrial Complex;
- Faculty of Finance;
- Faculty of Crediting;
- Faculty of Information Systems and Technologies.

The Faculty of Economics and Management is one of the leading Faculties at the university. More than 3500 students study at the Bachelor's and Master's programs of the Faculty full time and part time. Specialists with the degree in economics and management can be employed as economists, managers of different business organizations, state bodies, enterprises, corporations, and as advisors for economic questions, and top-managers. Fundamental economic education with a broad perspective combined with professional specialization gives the graduates a considerable advantage within the labour market.

The Faculty of International Economics and Management is the youngest in the university. The Faculty prepares students for a variety of career opportunities. Graduates typically find employment in Ukrainian companies engaged in international trade; state bodies; companies with foreign investment, affiliates and subsidiaries of international corporations, foreign investment banks, affiliates of foreign banks, investment funds and companies. The graduates are also engaged in the work of international projects financed by the International Monetary Fund, the World Bank, and the European Union.

The degree programs at the Law Faculty serve to prepare students academically and to equip them with the necessary qualifications for professional careers in business and society for which an advanced knowledge of law is required. The degree awarded at the Law Faculty guarantees that the graduate is not only a qualified lawyer, but also possesses deep knowledge in economics.

The Faculty of Marketing offers two specializations of the direction "Economy and Entrepreneurship": "Marketing" and "Personnel Management and Economy of Labour". Full time training on the programs "Marketing Management" and "Advertising Management" enables the student to work as a manager and marketing directors in different business organizations. Graduates of the Faculty who specialized in "Personnel Management and Economy of Labour" (Master's program "Personnel Management") have well-developed skills and knowledge to carry out organizational, managerial, planning, analytical, project, scientific and methodological activity.

The Faculty of Accounting has a strategic mission of training highly qualified specialists in the field of accounting, economic analysis and auditing.

The Faculty of Economy of Agro-industrial Complexes prepares a new generation of specialists, who will be able to work not only as accountants and economists of agro-industrial complexes, but being universal specialists, with the sufficient knowledge they will be able to manage investments and capital at all the stages of making agro-industrial products starting with farming and finishing with the production of food products and their promotion in internal and external markets.

The Faculty of Finance prepares the student for employment in the Ministry of Finance of Ukraine, in the financial departments of regional and local state bodies, the State Treasury, the State Inspection, the State Tax Administration, insurance companies, financial departments of enterprises and different business organizations.

The Faculty of Crediting trains bachelors and masters in the specializations "Banking", "Economy and Entrepreneurship". The Faculty emphasizes the practical teaching of students and enables them to gain practical experience in the branches of the National Bank, banking

organizations of the 2-nd level, banks of Ukraine, and investment companies. Ukrainian banks have a great need for the graduates of the Faculty.

The experts in economic cybernetics, applied statistics and intellectual systems of decision making process are trained in the bachelor's and master's programs of Faculty of Information Systems and Technologies. The Faculty offers the following master's programs: Information Systems in Management, Data Management, Information Systems in Banks and Financial Organizations, Systems Analysis and Modeling of Economic Processes, Applied Statistics.

2.2 Theory of Probability and Mathematical Statistics at Economic University

Mathematics, as educational discipline, took the important place in all of the economic specialities. Wide application of mathematical methods is a feature of modern economy. The Department of Higher Mathematics conducts the fundamental mathematical training of students of all faculties, except the Faculty of Law. In the first year of studies, students study such courses, as "Mathematics for economists" and "Theory of probability and mathematical statistics". Course "Mathematics for economists" consists of such sections of mathematics, as linear algebra, analytical geometry, mathematical analysis, differential equations. This course also illustrates an application of these mathematical topics for solving applied economical tasks. The purpose of course "Theory of probability and mathematical statistics" is to acquaint students with basic concepts, methods, theorems and formulas of probability theory and mathematical statistics and help them to get primary skills in solving different problems. The University is successfully integrating into the European system of education based on the Bologna declaration, which defined the approaches to creating a single European educational environment. Therefore teachers of Department of Higher Mathematics need to create new interesting for students and up to date courses.

The Department of Statistics provides more teaching of statistics for the students of the second, the third and the fourth year of studies. According to the programs for bachelor and master's degrees the department of statistics provides the teaching of such courses as bank statistics, demographic statistics, legal statistics, statistics of state finances, labour statistics, financial statistics et al. The teachers of department give lecture on the series of the special courses for the students specializing in Statistics among which is "Sample surveys".

But this important course needs to be developed for the students of other specialities. As survey sampling is needed for marketing researches, this special course should be added to the education programs of the Faculty of Personnel Management and Marketing first of all.

A marketing concept is meant to study the market with a certain goal. In marketing, which should satisfy the needs of people, knowledge matters very much. In the conditions of market those firms and companies who better than other know these needs get advantages. But a market always changes, the needs of people under act of different factors change too. Therefore firms should always watch the state of market to get an income. So the task of marketing specialists is to conduct the most exact sample surveys with minimum expenses. Therefore the special course of Sample Surveys will be appropriate for training specialists of marketing. Now the teachers of Department of Higher Mathematics create such special course. The program for this course is composed according to the specific of student knowledge base.

SMALL AREA ESTIMATION IN EU-SILC SURVEY

Inga Masiulaitytė

Statistics Lithuania, Vilnius University

e-mail: inga.masiulaityte@stat.gov.lt

1 Introduction

The European Survey on Income and Living Conditions (EU-SILC) survey is the main survey which provides the poverty and income inequality statistics in Lithuania. A good measure of poverty and income could compare poverty over the time, make comparisons with over countries, assess the effects of the projects on poverty and so on. It is very important to measure poverty not only at the all country level, but also at smaller geographical level. At small areas (Vilniuas, Kauno counties and so on), there is lack of the survey data and the direct estimators have large variances, these data cannot be published.

The goal of the research to find small area estimation method to estimate poverty and income inequality. The target is the estimation of the poverty rate, mean of the income and Gini coefficient. The research is in progress now. The results will be presented later.

References

Giusti C., Pratesi# M., and Salvati N., Estimation of poverty inicators: a comparison of small area methods at LAU1-2 level in Tuscany. Department of Statistics and Mathematics Applied to Economics, University of Pisa. Research methodology paper, Eurostat webpage.

Khandker, S. (2004), *Poverty Manual,* World Bank.

Molina I. and Rao, J. N. K., Estimation of Poverty Measures in Small Areas. Research methodology paper, Eurostat webpage.

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

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

SAMPLING OF PRIVATE FARMING

Julia Orlova[1]

[1] Department of Sstatistics, Belarusian State Economic University, Belarus
e-mail: 55xx@mail.ru

Objective of this research is creation of methodical support intended for use by the Ministry of statistics and the analysis of Belarus at the organisation of selective inspections of personal part-time farms of citizens with a view of definition of total gathering and productivity of production made by personal part-time farms.

For formation of a sample of economy of the population application of methodology of three-level serial (nested) sample at which sample is made in three stages is offered: on the first sample of areas is taken, on the second – in each of the selected areas sample of the private farming is taken. Personal part-time farms of the rural soviets which have made a sample, are subject to continuous inspection.

For the organisation of specialised selective inspections in personal part-time farms of citizens it is necessary to generate at republic level a general formation on the basis of the size of an area under crops on the crop's eve of agricultural crops in personal part-time farms of the population. The general formation is formed on the basis of the data private farming books. In the subsequent construction of the general formation reflecting a real picture of manufacture of agricultural crops (animals) in personal part-time farms of the population, becomes possible after carrying out of agricultural census.

The algorithm of formation of a sample of specialised inspection of economy of the population includes following stages:

1. Calculation of characteristics of a general formation of areas, the Rural Soviets necessary for scoping of sample.
2. Scoping of sample of areas.
3. Formation of a sample of areas (selection on a streamer).
4. Scoping of sample of the Rural Soviets.
5. Formation of a sample of the Rural Soviets (on the basis of division of a general formation of the Rural Soviets on groups (striations) on the basis of the size of areas under crops).
6. Calculation of characteristics of a sample of the Rural Soviets necessary for an estimation of parameters of a general formation of the rural soviets, estimations of errors of sample.

The volume of sample of areas at Republic level has been calculated at the values of relative errors of sample making 0,5 %, 1 %, 3 %, 5 % and 7 % by data for 2007 and is presented in table 1.

Table 1

Volume of sample of areas of Belarus at an absolute error of the sample making 0,5 %, 1 %, 3 %, 5 %, 7 %

Number of areas	Average value of areas under crops on area	Sample Volume at the set absolute error of sample				
		0,50%	1%	3%	5%	7%
118	3955,79	118	117	108	95	79
Sample fraction		99,8%	99,0%	91,8%	80,2%	67,4%

To admissible and desirable 5 % level of accuracy it agree with the calculation resulted more low there corresponds sample volume in 95 areas.

Procedure of selection of areas included following steps:

1. All areas within areas of Belarus settled down as it should be defined by movement on a streamer from the east on the West;

2. The volume of sample of areas within each area proceeding from the total amount of sample making 80 % of all areas of Republic was defined.

3. On each of areas the selection step was defined.

4. All areas for which the cumulative sum calculated by consecutive addition of a step of selection, was less than the cumulative sum of areas under crops of agricultural crops of personal part-time farms of the population were subject to selection. Process of formation of a sample of areas of the Brest area is presented in table 2.

Table 2

Formation of a sample of areas of the Brest area
(Selection on a streamer on an area under crops)

Area	The Area under crops of the earth of personal part-time farms of area, hectare	Probability of selection	The Cumulative sum	The Cumulative sum of intervals	The Selected (Taken away) areas
Baranovichsky	5 012,12	0,067	5 012,12	2858	*
Ljahovichsky	4 057,28	0,055	9 069,40	8575	*
Pruzhansky	4 165,52	0,056	13 234,92	14 292	
Ivatsevichsky	4 309,30	0,058	17 544,22	14 292	*
Gantsevichsky	3 055,51	0,041	20 599,73	20 009	*
Kamenetsky	3 084,80	0,042	23 684,53	25 726	
Berezovsky	4 278,41	0,058	27 962,94	25 726	*
Zhabinkovsky	1 856,24	0,025	29 819,18	31 443	
Kobrinsky	6 302,28	0,085	36 121,46	31 443	*
Drogichinsky	6 183,14	0,083	42 304,60	37 160	*
Ivanovo	5 821,23	0,078	48 125,83	42 877	*
Pinsky	6 553,11	0,088	54 678,94	48 594	*
Luninetsky	6 500,57	0,087	61 179,51	54 311	*
Brest	3 120,72	0,042	64 300,23	60 028	*
Maloritsky	2 416,40	0,033	66 716,63	65 745	*
Stolinsky	7 604,59	0,102	74 321,22	71 462	*

Calculation of volume of a sample of quantity of the Rural Soviets includes following steps:

1. The general formation of the Rural Soviets is sorted in ascending order the size of areas under crops of personal part-time farms of the Rural Soviet.

2. The general formation of the Rural Soviets is distributed(allocated) on the basis of the size of areas under crops of personal part-time farms of the Rural Soviet on number of groups (страт), defined as private from division of scope of a variation into an interval defined under the formula:

$$i = \frac{\widetilde{S}_{ij\,\max} - \widetilde{S}_{ij\,\min}}{1 + 1{,}44 \ln M}, \tag{1}$$

Where - the maximum value of the size of areas under crops on a general formation of the Rural Soviets;

- The minimum value of the size of areas under crops on a general formation of the Rural Soviets;

- Number of the Rural Soviets in a general formation.

3. For each of groups all mathematical-statistical characteristics (number of the Rural Soviets which have got to group, the average size of areas under crops on each group, a dispersion, variation factor) pay off.

On the first step the general formation of the Rural Soviets has been sorted in ascending order the size of areas under crops of personal part-time farms of the Rural Soviet.

On the second step the general formation Rural Soviets have been grouped in qualitatively homogeneous five to a sign of the size of areas under crops of personal part-time farms of the Rural Soviet of groups (table 3).

Table 3

Scoping of sample of the Rural Soviets on the basis of division of a general formation of the Rural Soviets on qualitatively homogeneous groups

Interval	Number of the Rural Soviets	Average value of an area under crops of the Rural Soviet	Dispersion	Variation, %	Number of the Rural Soviets × Dispersion
1-150	159	105,6	1736,9	39,5	276 173,91
150-300	579	227,0	1892,6	19,2	1 095 817,34
300-450	390	370,1	1800,3	11,5	702 128,00
450-1 000	286	564,2	11637,5	19,1	3 328 338,41
1 000 and more	11	1163,8	28015,0	14,4	308 165,17
Total	1425	327,55	30735,1	53,5	5 710 622,82

On the third step on each of groups indicators of average value of the size of areas under crops of the Rural Soviet, an intragroup dispersion and variation factor have been calculated. The volume of sample of the Rural Soviets was defined under the formula

$$m = \frac{k \cdot \sum_{l=1}^{k} \sigma_l^2 \cdot \left(\frac{M_l}{M}\right)^2}{\Delta^2 + \sum \frac{M_l}{M^2} \cdot \sigma_l^2} \tag{2}$$

where k – number of the allocated groups of the Rural Soviets on the basis of the size of areas under crops of personal part-time farms of the Rural Soviet;

M_l – quantity(Amount) of the Rural Soviets which have got to group;

M – quantity(Amount) of the Rural Soviets of a general formation;

$\Delta 2$ – an absolute error of sample;

σ_l^2 – an intragroup dispersion of an indicator on the basis of the size of areas under crops of personal part-time farms of the Rural Soviet.

The volume of sample of the Rural Soviets at Republic level has been calculated at the values of relative errors of sample making 0,5 %, 1 %, 3 %, 5 % and 7 % by data for 2007 and is presented in table 4.

Table 4

Volume of sample of the Rural Soviets of Belarus at an absolute error of the sample making 0,5 %, 1 %, 3 %, 5 %, 7 %

Number of the Rural Soviets	Average value of areas under crops on the Rural Soviet	Sample Volume at the set absolute error(mistake) of sample				
		0,50%	1%	3%	5%	7%
1 425	327,55	855	347	47	17	9
Sample fraction		59,99%	24,34%	3,32%	1,22%	0,62%

From the offered five variants the comprehensible volume of a sample depending on necessary level of accuracy and admissible financial expenses is defined.

To admissible 1 %-s' absolute error there corresponds sample volume in 24 % of the Rural Soviets.

Proceeding from the general percent of sample making 24 %, each group the certain percent of selection (table 5) is given;

Table 5

Definition of a step of selection on each of groups the Rural Soviets

Interval	Number of the Rural Soviets in a general formation	Number of the Rural Soviets in a sample of areas (I stage of sample)	Selection Percent	Number of the Rural Soviets in a sample of the Rural Soviets (II stage of sample)	A selection Step
A	1	2	3	4	5
1-150	159	124	10	12	10
150-300	579	489	20	98	5
300-450	390	359	25	90	4
450-1 000	286	263	50	132	2
1 000 and more	11	10	100	10	1
Total	1425	1245	24	341	

The selection beginning is in a random way defined (the selection beginning, that is 1st Rural Soviet to sample, the first Rural Soviet or the Rural Soviet which is in the middle of an interval of selection) can be got;

Each subsequent Rural Soviet which has got to sample, is defined by addition of an interval of sample.

For maintenance(support) of rotation of sample at repeated selective inspection it is recommended to replace each of the Rural Soviets making a sample, on following under the list (considering that at carrying out of the second stage of sample in the allocated five groups the Rural Soviets settled down in alphabetic order).

Distribution on a general formation of the given specialised selective inspections of economy of the population can be spent(lead) in two variants: calculation of simple estimations of total and average values of the size of areas under crops of the Rural Soviets; calculation of simple estimations of total and average values of the size of areas under crops of personal part-time farms as a part of the Rural Soviets.

Distribution on a general formation of results of selective inspection according to the Rural Soviets is recommended to be spent as follows:

1. On a sample of the Rural Soviets grouped in five groups on the basis of the size of an area under crops of the Rural Soviet selective intragroup average value of an indicator pays off.

2. The number of the Rural Soviets in each of groups selective and a general formation of the Rural Soviets contains accordingly in columns 4 and 1 appendices 5.

3. The estimation of total value of an indicator is made by a finding of the sum of products of selective intragroup average values of an indicator on number of the Rural Soviets in each of general formation groups.

Distribution on a general formation of results of selective inspection of the Rural Soviets on areas of Belarus is presented in table 6.

Table 6

Estimation of divergences of indicators of the size of areas under crops of the Rural Soviets occupied(borrowed) under crops of a potato, on sample and a general formation on areas of Byelorussia

Area	The Area under crops, hectare	A potato Area under crops, hectare	The Estimation of a share of areas under crops of the Rural Soviets occupied(borrowed) under crops of a potato, %	The Estimation of total value of the size of areas under crops of the Rural Soviets occupied(borrowed) under crops of a potato, by results of selective supervision, hectare	The Divergence of total value of the size of areas under crops of the Rural Soviets occupied(borrowed) under crops of a potato, the received by results of selective and continuous supervision, %
Brest area	74 987,77	47 748,96	64,02	48 007,17	0,541
Vitebsk area	92 102,02	33 515,28	34,77	32 023,87	-4,450
Gomel area	78 146,38	41 354,06	48,97	38 268,28	-7,462
Grodno area	57 198,51	34 864,57	60,02	34 330,55	-1,532
Minsk area	95 922,68	45 287,65	48,17	46 205,95	2,028
Mogilyov area	66 731,84	25 012,50	37,36	24 931,02	-0,326
Belarus	461 024	227 783,02	48,28	222 582,39	-2,283

The estimation of divergences of indicators of sample and general formation has shown that sample is representative as the greatest divergence observed in the Grodno area and the made 3,86 % corresponds to admissible and desirable 5 % level of accuracy.

KERNEL IMPUTATION

Nicklas Pettersson[1]

[1] Stockholm University, Sweden

e-mail: nicklas.pettersson@stat.su.se

Abstract

This paper gives a brief introduction to hot deck imputation. Focus is on a certain type of hot deck method where the probability of selecting donors is calculated using a kernel density function. Some issues concerning kernel imputation are highlighted. The presentation will focus more on visualizing some of the methods and concepts that are described.

1 Introduction

Missing or incomplete data in sample surveys might bias the results and lead to invalid inference. Missing data in sample surveys is typically due to nonresponse, where either all values of a unit (unit nonresponse) or some of the values (item nonresponse) may be missing. Unit nonresponse may be caused by language problems or refusal to respond, while item nonresponse could occur with sensitive questions, or if the respondent does not have enough information to be able to answer.

Methods that adjust for nonresponse make use of observed information, such as auxiliary variables from registers and observed survey data. Unit nonresponse is often dealt with by adjusting the design weights (which are inversely proportional to selection probabilities) so that the nonresponse is treated as part of the design. Imputation is often used for item nonresponse, where the missing values are filled in with values estimated from the observed data.

The aim of an imputation process is generally to reduce nonresponse bias. Other aims are to preserve distributions of and associations between variables, and to recreate individual values of the true data. The importance of each of these aims depends on the purpose of the analyses. With a descriptive purpose, such as estimating a population total, focus will be on bias, while associations are of more importance in surveys with an analytic purpose. Recreating individual values will assist in fulfilling the other aims, but is also most difficult to achieve.

1.1 Model and donor imputation

Imputation methods can be categorized into model donor and real donor imputation. Model donor imputation replaces the missing values with predictions from a model, typically a linear regression model. Real donor imputation replaces the missing values with actual values from units similar to the recipient according to a distance measure. The distinction between model and donor methods is not always obvious or possible. One example is fractional imputation, which use a hot deck methodology but then impute a local mean based on very donors. Another example is when a parametric model is used to predict a value, but the imputed value is taken to be a true value that is closest to the predicted value.

Model donor methods are in general better at predicting individual values if the model is good, but might also bias the results if the model is bad. Real donor methods are in general more robust then model donor methods, because they assume less about the underlying distribution of the data. Imputed values from real donor imputation will also be

more plausible, since the imputed values already occur in the data, which is generally not the case with imputed values from model donor imputation. Another advantage of real donor methods is that they easily extend to imputation of several missing values on the same unit at once, which make them good at preserving associations between variables according to Kalton and Kasprzyk (1986).

1.2 Imputation uncertainty

Imputed values (almost always) differ from the true values. Rubin (1987) denote an imputation method that takes account of this imputation uncertainty as proper. According to Little and Rubin (2002), two imputation methods that may be proper are resampling and multiple imputations. Both methods are based on repeating the imputation procedure several times, producing B>1 different imputed datasets. Imputation uncertainty is then estimated from the variation between the datasets.

Multiple imputations are carried out as a repetition of a single imputation method on with the dataset consisting of the originally drawn sample. It therefore requires the imputation method to be random, since otherwise there is no variation between the imputed datasets.

Resampling methods create imputed datasets by drawing new samples with replacement from the original sample. Compared to multiple imputation methods, the methods are less dependent on that the assumed model is true. However, they also require more datasets than multiple imputations, and since they are based on large-sample theory, their properties might be questionable in small samples. Shao (2003) show that if interest is to estimate population totals or means, resampling with a deterministic model might be preferred to a random imputation model.

1.3 Disposition

Section 2 describe hot deck imputation, which is a real donor method based on duplication of the observed data. In section 3 the description is extended to kernel imputation, which allows varying probability distributions to be applied in the hot deck duplication process. Some comments on the coming presentation are given in section 4.

2 Hot deck imputation

Hot deck imputation is a very common real donor method, but according to Little and Rubin (2002) its theoretical properties are less explored than that of model donor methods. First assume a population U with N units, where the joint distribution of the auxiliary variables $X=(X_1,\ldots,X_d)$ and a study variable Y is determined by a parameter vector θ. The mean of Y is denoted as θ_Y. A random sample of size n is then drawn from the population

$$(X_i,Y_i,\delta_i)\ i=1,\ldots,n$$

where X_i is always observed, while for the first r units $\delta_i=1$ and Y_i is observed ($Y_{obs,i}$), while for the n-r last units $\delta_i=0$ and Y_i is missing ($Y_{mis,i}$). Further, assume that the conditional expectation of Y given $X=X_i$ may be described by

$$m(X_i)=E[Y|X=X_i] \quad \textbf{(1)}$$

Also assume that the mechanism that leads to nonresponse is strongly ignorable missing at random (MAR), as defined by Rosenbaum and Rubin (1983). It then follows that

$$P(\delta_i=1|X_i,Y_i)=P(\delta_i=1|X_i)=\pi(X_i), \quad \textbf{(2)}$$

so that Y_i and δ_i are conditionally independent given X_i, and $\pi(X_i)$ denotes the probability that Y_i is missing. A justification for hot deck imputation is the MAR assumption which states that the reason for missingness is assumed to be observed in the auxiliary data

and therefore don't need to be modelled outside of the data, which is also partly seen from the conditional independence in (2). Further (1) shows that the variable with missing values is related to the fully observed auxiliary variables. Thus, given that X_i is reasonably similar to X_j the relation may be utilized to replace a missing value $Y_{mis,i}$ with an observed value $Y_{obs,j}$,.

Duplications of the r observed values in Y_{obs} are thus used to replace the n-r missing values in Y_{mis}. For each of the i=1,...,r units where δ_i=0, a list is first made of the j=1,…,k units (where δ_i=1) that are closest to unit i according to some distance metric. One unit j is then drawn at random from the list, and $Y_{mis,i}$ is then replaced with a duplicated value of $Y_{obs,j}$. The n-r imputed values are denoted as Y_{imp}, and (X,Y_{obs},Y_{imp}) is thus an imputed dataset.

If all units are classified into adjustment cells based on X, then one type of distance metric for calculating nearness give all completely observed units j in the same cell as unit i the distance d(i,j)=0, while d(i,j)>0 for units in other cells. Another common choice of metric is Euclidian distance $d(i,j)=(X_i-X_j)^T S_{xx}^{-1}(X_i-X_j)$, where S_{xx} is the diagonal of the sample covariance matrix of X. The two types of metrics might also be combined into a single metric. The method of selecting the k nearest neighbour units as potential donors is sometimes denotes as kNN imputation.

With a single imputed dataset θ_Y can be estimated with

$$\hat{\theta}_Y = \frac{1}{n}\sum_{i=1}^{n} \delta_i Y_{obs,i} + (1-\delta_i) Y_{imp,i} . \quad \textbf{(3)}$$

An estimator of θ_Y based on b=1,…,B imputed datasets is

$$\tilde{\theta}_Y = \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}_Y^{(b)} , \quad \textbf{(4)}$$

where $\hat{\theta}_Y^{(b)}$ is the estimate in (3) from a single imputed dataset.

2.1 Sampling strategies and variance estimation

The choice of sampling strategy that is used to draw donor units will affect variance estimates. If sampling from Y_{obs} is done without replacement or by placing some other restriction on the number of times that each $Y_{obs,i}$ are allowed to be duplicated, estimates of $Var(\hat{\theta}_Y)$ are expected to be smaller compared to when sampling is made with replacement. This might be used when estimating $Var(\hat{\theta}_Y)$ from a single imputed dataset.

Another sampling strategy is the finite population Bayesian bootstrap (FPBB) as described by Lo (1988), where previously imputed units are allowed to serve as donors in subsequent imputations. Assume that data is sorted in increasingly order of missingness, starting with the r fully observed units and the n-r units where Y is missing at the end. Imputations are then carried out in order of i=r+1,…,n, and the k potential donors for each unit i are then found among the 1,…,i-1 units. If Y would contain several variables, the sampling scheme could maximize the conditioning on Y_{obs}, see Kong, Liu and Wong (1994).

Rubin (1987) show a basic multiple hot deck imputation with replacement that is not proper. However, they also give an example of a proper bootstrap technique denoted as approximate Bayesian bootstrap (ABB). For each dataset, a sample with replacement of size r is first drawn from the r units where δ_i=1. The sampled values from Y_{obs} are denoted as Y_{obs}*. Then Y_{imp} is created from a sample of size n-r drawn with replacement from Y_{obs}*. Another bootstrap suggested for hot deck imputation is suggested by Shao and Sitter (1996). Each imputed dataset is then created by drawing a random sample of size n-1 with replacement

from the original sample, and then randomly imputing the missing values using only donors in the drawn sample.

3 Kernel imputation

A generalization of hot deck imputation allows the kNN units to have unequal probabilities of donating. First assume that the probability of unit j becoming the donor for unit i is determined by a weight W_{ij}, which is related to d(i,j). Further assume that weights are calculated using a kernel function K(), which itself is controlled by a bandwidth (or smoothing) parameter h. If X is univariate, the kernel weights of the k potential donors to unit i are then given as

$$W_{ij}(X_i) = K\left(\frac{X_i - X_j}{h}\right) \Big/ \sum_{j=1}^{k} K\left(\frac{X_i - X_j}{h}\right), \text{ for j=1,...,k.}$$

Since $W_{ij}(X_i) \geq 0$ and $\sum_{j=1}^{k} W_{ij}(X_i) = 1$, the weights may also be treated as selection probabilities for the kNN units of becoming donors for unit i.

The kernel function is usually chosen to be a symmetric unimodal probability density function, which satisfy which integrates to 1. By setting K() to be a uniform density function

$$K\left(\frac{X_i - X_j}{h}\right) = \begin{cases} 1/2 & if \left|X_i - X_j\right| < h \\ 0 & else \end{cases},$$

all units j within a distance h from X_i are given the same weight, while all other units are given the weight zero. This is essentially the hot deck imputation described in section 2 with k determined by h. The Epanechnikov kernel function

$$K\left(\frac{X_i - X_j}{h}\right) = \begin{cases} 3\left(1 - \left[\left(X_i - X_j\right)^2 / h^2\right]\right)/4 & if \left|X_i - X_j\right| < h \\ 0 & else \end{cases}$$

will instead assign higher weights to units with smaller distances, while large distant units gets zero weights.

The relation between the number of potential donors, and the smoothing parameter is that if h is set to a fixed value, then k will be determined by h. On the other hand if k is fixed, then h will be variable as a function of k and set proportional to the largest distance among the k units, $h_{k(X_i - X_j)} \propto \max\left[d(i,j)\right]$. Whether h or k is fixed this reparameterization allows the same list of donors for the imputation of unit i. However, it is difficult to determine an optimal fixed value for each imputation, so all units for which δ_i=0 are usually imputed using the same fixed h or k. The lists will therefore usually differ.

3.1 Different studies on Kernel imputation

Cheng (1994) used kernel imputation with a modified type of equation (3) for estimating θ_Y

$$\hat{\theta}_Y = \frac{1}{n}\sum_{i=1}^{n} \delta_i Y_{obs,i} + (1 - \delta_i)\hat{m}(X_i),$$

where Y_{imp} was replaced with an estimate of the conditional expectation (1). The kernel weights are used when calculating $\hat{m}(X_i)$, which is also known as the Nadaraya-Watson estimator

$$\hat{m}(X_i) = \frac{\sum_{i=1}^{k} W_{ij} Y_{obs,j}}{\sum_{i=1}^{k} W_{ij}} = \frac{\sum_{i=1}^{k} K\left(\frac{X_i - X_j}{h}\right) Y_{obs,j}}{\sum_{i=1}^{k} K\left(\frac{X_i - X_j}{h}\right)}. \quad \textbf{(5)}$$

Silverman (1986) show that $Bias[\hat{m}(X_i)]$ increases with h and $Var[\hat{m}(X_i)]$ decreases with h. They also show that its mean squared error

$$MSE[\hat{m}(X_i)] = E[\hat{m}(X_i) - m(X_i)]^2 = Bias[\hat{m}(X_i)]^2 + Var[\hat{m}(X_i)]$$

can be minimized if K() is set to be an Epanechnikov kernel function. It is easily seen that (5) is the expectation of the imputed value. Now (5) is almost always biased since the X_j are unevenly dispersed around Xi, especially when Xi are at the boundary of the data.

A solution is then to change to linearly transformed kernel weights W'$_{ij}$,which may be found by minimizing $f(W_{ij} - W_{ij}')^2$ with restrictions that $W_{ij}' \geq 0$ for j=1,...,k, $\sum_{j=1}^{k} W_{ij}' = 1$ and $\sum_{j=1}^{k} W_{ij}' X_j = X_i$. Additional constraints might also be imposed, for example that $W_{ij}' < c$ for j=1,...,k, where $0 < c \leq 1$ is the max allowed selection probability that is given to one potential donor. More suggestion on how to deal with boundary bias is given in Simonoff (1996).

Aerts, Claeskens, Hens and Molenberghs (2002) use the kernel weights W'$_{ij}$ to estimate θ_Y by (4). Opposite to Cheng (1994) they use a proper kernel imputation method based on bootstrapping, which allows for additional parametric assumptions about Y_{imp}. Their findings are that the choice of kernel function is of less importance, since the precise choice of their kernel weights has a small effect on the final estimator. Wang and Rao (2002) extend the results of Cheng (1994) to an empirical likelihood method and make a simulation study on confidence intervals for θ_Y. Even though Wang and Rao (2002) show that the exact choice of smoothing was not so important in their estimation, the choice of whether h (or k) should be fixed or variable and what the size of h (or k) should be, is usually of more importance, see for example Simonoff (1996) for a further description.

A more recent study of kernel imputation is given in Conti, Marella and Scanu (2008). They make a simulation study comparing kernel imputation using a Gaussian kernel and different ways of selecting the smoothing to other imputation methods. Relations in the data are also varied. Their findings are that the performance of kernel imputation is close to that kNN hot deck in estimating marginal distributions, but that kernel imputation is better at estimating conditional expectations when the relationship in the data is more complex.

If X instead is multivariate, then the kernel function will also be multivariate, see for example Silverman (1986). The bandwidth parameter then becomes a matrix similar to S_{xx}, while k is still a single integer. However, when X is high dimensional, kernel imputation might lose some of its attractiveness compared to model donor imputation, see Aerts et al (2002). The main reason is that required sample increase drastically with the dimensionality of the data. This is also referred to as the "curse of dimensionality".

Finally, the most important factor to successful imputation is probably to have strong predictive (auxiliary) variables (and a reasonable model).

4 Presentation

The presentation will focus more on visualizing some of the methods and concepts that are described in this paper. Ideas for future research will also be discussed.

References

Aerts, M., Claeskens, G., Hens, N. and Molenberghs G. (2002). Local multiple imputation. *Biometrika*, **89**, 375-388.

Cheng, P. E. (1994). Nonparametric estimation of mean functionals with data missing at random. *Journal of the American statistical association*, **89**, 81-87.

Conti, P. L., Marella, D. and Scanu, M. (2008). Evaluation of matching noise for imputation techniques based on nonparametric local linear regression estimators. *Computational statistics and data analysis*, **53**, 354-365.

Kalton, G. and Kasprzyk, D. (1986). The treatment of missing survey data. *Survey Methodology*, **12**, 1-16.

Kong, A., Liu, J. S. and Wong, W. H. (1994) Sequential Imputations and Bayesian Missing Data Problems. *Journal of the American statistical association*, **89**, 278-288

Little, R. J. A. and Rubin, D. B. (2002). *Statistical analysis with missing data*. John Wiley & Sons, New York.

Lo, A. Y. (1988). A bayesian bootstrap for a finite population. *The annals of statistics*, **16**, 1684-1695.

Rosenbaum, P. R., and Rubin, D. B., (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, **70**, 41–55.

Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. John Wiley & Sons, Hoboken.

Shao, J. (2003). Impact of the bootstrap on sample surveys. *Statistical science*, **18**, 191-198.

SHAO, J. and SITTER, R. R. (1996). Bootstrap for imputed survey data. *Journal of the American statistical association,* **91** 1278–1288.

Silverman (1986). *Density estimation for statistics and data analysis*. Chapman & Hall, London.

Simonoff (1996) *Smoothing methods in statistics*. Springer-Verlag, New York.

Wang, Q. and Rao, J. N. K. (2002). Empirical likelihood-based inference under imputation for missing response data. *The annals of statistics*, **30**, 896-924.

ON THE USE OF SEVERAL WEIGHT SYSTEMS FOR ESTIMATION OF FINITE POPULATION COVARIANCE

Dalius Pumputis

Vilnius Pedagogical University, Lithuania
e-mail: dpumputis@vpu.lt, dpumputis@yahoo.co.uk

Abstract

We extend the theory of estimation with several systems of weights. The new calibrated estimators of the finite population covariance (variance) are derived using two and three weighting systems, which are defined by various calibration equations and loss function.

1 Introduction

Survey statisticians are always concerned in improvement of methods of estimation for finite population total, mean, proportion and others parameters. An auxiliary information may be used for that purpose. The estimators which are using auxiliary variables are often much more accurate than standard ones. The calibrated estimators belong to this class of estimators. The idea of calibration technique for estimating the population totals is presented in (Deville and Särndal, 1992).
The estimation of more complicated parameters using calibration methods is not widely studied in the literature. The calibrated estimator of the ratio of two totals is considered by Plikusas (2003), Krapavickaitė and Plikusas (2005). Calibration estimation for quantiles is studied by Harms and Duchesne (2006), Rueda et al. (2007). Sitter and Wu (2002) proposed model-calibrated method to estimate the quadratic finite population functions. Singh et al. (1999) applied the calibration technique for estimation of variance of Horvitz-Thompson estimator.
In the paper (Plikusas and Pumputis, 2007) we introduce some calibrated estimators of the finite population covariance. They use one weighting system, which is defined using various calibration equations and loss functions. In the following section we recall these estimators and extend the theory of estimation with several systems of weights.

2 Calibrated estimators of the finite population covariance

Consider a finite population $\mathcal{U} = \{u_1, u_2, \ldots, u_N\}$ of N elements. Let y and z be two study variables defined on the population $\mathcal{U}$ and taking real nonnegative values. The values of the variables y and z are not known. Suppose the known auxiliary variables a and b are available.

Let the covariance

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

be parameter of interest.
Denote by $s, s\subset\mathcal{U}$, probability sample set drawn from the population $\mathcal{U}$.

2.1 Estimators using one system of weights

In the case of no auxiliary information we can estimate the population covariance using well-known only design based estimator

$$\widehat{Cov}(y,z)=\frac{1}{N-1}\sum_{k\in s}d_k\left(y_k-\frac{1}{N}\sum_{k\in s}d_ky_k\right)\left(z_k-\frac{1}{N}\sum_{k\in s}d_kz_k\right), \tag{1}$$

here d_k- sample design weights. It is considered in the Särndal, Swensson and Wretman's book (1992, p. 187).
Our case is different, the auxiliary variables are available. In the paper (Plikusas and Pumputis, 2007) we apply calibration technique to modify the design weights d_k. Here we consider the calibrated estimator of the covariance of the following shape

$$\widehat{Cov}_w(y,z)=\frac{1}{N-1}\sum_{k\in s}w_k\left(y_k-\frac{1}{N}\sum_{k\in s}w_ky_k\right)\left(z_k-\frac{1}{N}\sum_{k\in s}w_kz_k\right). \tag{2}$$

The new (calibrated) weights w_k are defined under the following conditions:

a) the weights w_k satisfy some calibration equation;

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

The conditions a) and b) can be specified in different ways. The following calibration equations are used in our paper:

I)

$$\frac{1}{N-1}\sum_{k\in s}w_k(a_k-\widehat{\mu}_{aw})(b_k-\widehat{\mu}_{bw})=Cov(a,b), \tag{3}$$

$$\widehat{\mu}_{aw}=\frac{1}{N}\sum_{k\in s}w_ka_k,\quad \widehat{\mu}_{bw}=\frac{1}{N}\sum_{k\in s}w_kb_k.$$

II)

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

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

III)

$$\sum_{k\in s} w_k a_k = \sum_{k=1}^{N} a_k, \quad \sum_{k\in s} w_k b_k = \sum_{k=1}^{N} b_k. \tag{5}$$

The loss function

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

and some other ones are applied for the final specification of calibrated weights w_k. We extend the definition given in this subsection to the case of multiple weighting systems.

2.2 Estimators using several systems of weights

Let us consider some other, more general estimators of the finite population covariance, which are constructed using several weighting systems. The new calibrated estimators of the covariance are of the following shape:

$$\widehat{Cov}_{mw}(y,z) = \frac{1}{N-1}\sum_{k\in s} w_k^{[1]}\Big(y_k - \frac{1}{N}\sum_{l\in s} w_l^{[2]} y_l\Big)\Big(z_k - \frac{1}{N}\sum_{l\in s} w_l^{[3]} z_l\Big). \tag{6}$$

Several calibration equations may be used for definition of the calibrated weights $w_k^{[1]}$, $w_k^{[2]}$, $w_k^{[3]}$. Let us consider some of them.

Case 1. One can take a nonlinear equation

$$\widehat{Cov}_{mw}(a,b) = Cov(a,b). \tag{7}$$

Case 2. The systems of weights $w_k^{[1]}$, $w_k^{[2]}$, $w_k^{[3]}$ are defined by employing calibration equations (4) and (5):

$$\frac{1}{N-1}\sum_{k\in s} w_k^{[1]}\big(a_k - \mu_a\big)\big(b_k - \mu_b\big) = Cov(a,b), \tag{8}$$

$$\sum_{k\in s} w_k^{[2]} a_k = \sum_{k=1}^{N} a_k, \quad \sum_{k\in s} w_k^{[3]} b_k = \sum_{k=1}^{N} b_k. \tag{9}$$

Case 3. The first system of weights $w_k^{[1]}$ is defined by the nonlinear calibration equation (3). Calibration equations (9) define the other two systems of the weights $w_k^{[2]}$ and $w_k^{[3]}$.

A reasonable choice of the loss function in the first three cases may be as follows:

$$L(w,d) = \alpha_1 \sum_{k\in s} \frac{(w_k^{[1]} - d_k)^2}{d_k\, q_k} + \alpha_2 \sum_{k\in s} \frac{(w_k^{[2]} - d_k)^2}{d_k\, q_k} + \alpha_3 \sum_{k\in s} \frac{(w_k^{[3]} - d_k)^2}{d_k\, q_k}\,, \tag{10}$$

where $\alpha_1 + \alpha_2 + \alpha_3 = 1$, $0 < \alpha_i < 1$, $i = 1, 2, 3$.

Case 4. We can consider the estimator of covariance which uses two systems of weights:

$$\widehat{Cov}_{mw}(y,z)=\frac{1}{N-1}\sum_{k\in s}w_k^{[1]}\Big(y_k-\frac{1}{N}\sum_{l\in s}w_l^{[2]}y_l\Big)\Big(z_k-\frac{1}{N}\sum_{l\in s}w_l^{[2]}z_l\Big). \qquad (11)$$

The first system of weights $w_k^{[1]}$ is defined by equation (8), whereas the second system $w_k^{[2]}$ satisfies the following equations

$$\sum_{k\in s}w_k^{[2]}a_k=\sum_{k=1}^{N}a_k, \qquad \sum_{k\in s}w_k^{[2]}b_k=\sum_{k=1}^{N}b_k. \qquad (12)$$

Case 5. We can use another combination of two systems of calibrated weights: the first one $w_k^{[1]}$ satisfies nonlinear calibration equation (3), where the system $w_k^{[2]}$ is defined by (12).

Case 6. The system of weights $w_k^{[1]}$ satisfies equation (8), whereas the system $w_k^{[2]}$ is equal to $w_k^{[3]}$, and it is obtained using nonlinear calibration equation (3).
The following loss function may be used for the last three cases:

$$L'(w,d)=\beta\sum_{k\in s}\frac{(w_k^{[1]}-d_k)^2}{d_k\,q_k}+(1-\beta)\sum_{k\in s}\frac{(w_k^{[2]}-d_k)^2}{d_k\,q_k}, \quad 0<\beta<1. \qquad (13)$$

The first case is most complicated analytically, the expressions for the approximate iterative solutions of calibration equation (7) are cumbersome.

The following proposition defines the weights $w_k^{[1]}$, $w_k^{[2]}$, $w_k^{[3]}$ of estimator (6) for all six cases mentioned in this subsection. The values $\alpha_1=\alpha_2=\alpha_3=\frac{1}{3}$, $\beta=\frac{1}{2}$ are taken for minimization of loss functions (10) and (13).

Proposition. *The weights $w_k^{[i]}$, $k\in s$, $i=1,2,3$, which satisfy calibration equation* (7) *and minimize loss function* (10), *satisfy the equation $w_k^{[i]}=d_ku_k^{[i]}$. Here $u_k^{[i]}=1+\lambda q_k e_k^{[i]}$,*

$$\begin{aligned}
e_k^{[1]}&=a_kb_k-a_k\widehat{\mu}_{bw^{[3]}}-b_k\widehat{\mu}_{aw^{[2]}}+\widehat{\mu}_{aw^{[2]}}\widehat{\mu}_{bw^{[3]}},\\
e_k^{[2]}&=-a_k\Big(\widehat{\mu}_{bw^{[1]}}-\frac{\hat{N}_{w^{[1]}}}{N}\widehat{\mu}_{bw^{[3]}}\Big),\\
e_k^{[3]}&=-b_k\Big(\widehat{\mu}_{aw^{[1]}}-\frac{\hat{N}_{w^{[1]}}}{N}\widehat{\mu}_{aw^{[2]}}\Big),\\
\lambda&=\widehat{A}\Big(\sum_{k\in s}d_kq_k(a_kb_ke_k^{[1]}+e_k^{[2]}-e_k^{[3]})\Big)^{-1},
\end{aligned}$$

$$\begin{aligned}
\widehat{A}=(N-1)Cov(a,b)+N\widehat{\mu}_{bw^{[3]}}\Big(\widehat{\mu}_{aw^{[1]}}-\frac{\hat{N}_{w^{[1]}}}{N}\widehat{\mu}_{aw^{[2]}}\Big)+N\widehat{\mu}_{aw^{[2]}}\widehat{\mu}_{bw^{[1]}}+\\
+\hat{N}_{w^{[2]}}-\hat{N}_{w^{[3]}}-\sum_{k\in s}d_ka_kb_k,
\end{aligned}$$

$$\widehat{\mu}_{aw^{[1]}} = \frac{1}{N}\sum_{k\in s} w_k^{[1]} a_k, \qquad \widehat{\mu}_{aw^{[2]}} = \frac{1}{N}\sum_{k\in s} w_k^{[2]} a_k$$
$$\widehat{\mu}_{bw^{[1]}} = \frac{1}{N}\sum_{k\in s} w_k^{[1]} b_k, \qquad \widehat{\mu}_{bw^{[3]}} = \frac{1}{N}\sum_{k\in s} w_k^{[3]} b_k$$
$$\hat{N}_{w^{[1]}} = \sum_{k\in s} w_k^{[1]}, \qquad \hat{N}_{w^{[2]}} = \sum_{k\in s} w_k^{[2]}, \qquad \hat{N}_{w^{[3]}} = \sum_{k\in s} w_k^{[3]}.$$

In Cases 2, 4 *and* 6, *the first system of weights* $w_k^{[1]}$ *is defined by the equations:*

$$w_k^{[1]} = d_k\Big(1 + q_k\Big(\sum_{l=1}^{N} x_l^{[1]} - \sum_{l\in s} d_l x_l^{[1]}\Big)\Big(\sum_{l\in s} d_l q_l (x_l^{[1]})^2\Big)^{-1} x_k^{[1]}\Big),$$

where $x_k^{[1]} = (a_k - \mu_a)(b_k - \mu_b)$.
The equations

$$w_k^{[i]} = d_k\left(1 + \widehat{A}\Big(\sum_{l\in s} d_l q_l f_l a_l b_l\Big)^{-1} q_k f_k\right)$$

define the first system of weights $w_k^{[1]}$ *in Cases* 3 *and* 5, *and the systems* $w_k^{[2]}$ *and* $w_k^{[3]}$ *in Case* 6. *Here*

$$\widehat{A} = (N-1)Cov(a,b) + N\left(2 - \frac{\widehat{N}_{w^{[i]}}}{N}\right)\widehat{\mu}_{aw^{[i]}}\widehat{\mu}_{bw^{[i]}} - \sum_{k\in s} d_k a_k b_k,$$

$$f_k = (a_k - \widehat{\mu}_{aw^{[i]}})(b_k - \widehat{\mu}_{bw^{[i]}}) - \left(1 - \frac{\widehat{N}_{w^{[i]}}}{N}\right)\left(\frac{\widehat{\mu}_{aw^{[i]}}}{a_k} + \frac{\widehat{\mu}_{bw^{[i]}}}{b_k}\right)a_k b_k, i = 1,2,3.$$

In Cases 2 *and* 3, *the systems of weights* $w_k^{[2]}$ *and* $w_k^{[3]}$ *are defined by the following equations:*

$$w_k^{[i]} = d_k\Big(1 + q_k\Big(\sum_{l=1}^{N} x_l^{[i]} - \sum_{l\in s} d_l x_l^{[i]}\Big)\Big(\sum_{l\in s} d_l q_l (x_l^{[i]})^2\Big)^{-1} x_k^{[i]}\Big),$$

where $x_k^{[2]} = a_k$, $x_k^{[3]} = b_k$, $i = 2,3$.
In Cases 4 *and* 5, *the second system of weights* $w_k^{[2]}$ *satisfies these equations:*

$$w_k^{[2]} = d_k\Big(1 + q_k\Big(\sum_{l=1}^{N} \boldsymbol{x}_l^{[2]\prime} - \sum_{l\in s} d_l \boldsymbol{x}_l^{[2]\prime}\Big)\Big(\sum_{l\in s} d_l q_l \boldsymbol{x}_l^{[2]}\boldsymbol{x}_l^{[2]\prime}\Big)^{-1}\boldsymbol{x}_k^{[2]}\Big),$$

here $\boldsymbol{x}_k^{[2]} = (a_k, b_k)'$.

Let us denote the new estimators by $\widehat{Cov}_{mw}^{(i)}(y,z)$, $i = 1,2,\ldots,6$. Here the index i refers to the case of definition of calibrated weights. For example, $\widehat{Cov}_{mw}^{(1)}(y,z)$ denotes the estimator which uses three weighting systems defined by calibration equation (7) and loss function (10).

The presented calibrated estimators of the covariance are complicated enough and it is not easy to derive the variance estimators in the most cases considered above. So, a short simulation study is performed to check if the use of more than one weighting system reduce the variance of estimators. The simple only design based estimator (1) is also included into the simulation. We consider the real population of size 300 and having skewed distribution which is close to exponential. Stratified simple random sampling is used.

The main result is that if the correlation between study and auxiliary variables is high ($\rho(y,a) = 0.81$ and $\rho(z,b) = 0.90$), then the estimators $\widehat{Cov}_{mw}^{(1)}, \widehat{Cov}_{mw}^{(2)}, \widehat{Cov}_{mw}^{(4)}, \widehat{Cov}_{mw}^{(6)}$ outperform all the estimators (2) with one system of weights. They have smaller bias, variance, mean square error and coefficient of variation. In the case of one well correlated auxiliary variable ($\rho(y,a) = 0.21$ and $\rho(z,b) = 0.90$) situation is totally different. Only two estimators ($\widehat{Cov}_{mw}^{(3)}$ and $\widehat{Cov}_{mw}^{(5)}$) perform better than that which use one system of weights. In the case of low correlated auxiliary variables all the estimators are of similar quality.

References

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

Harms, T., Duchesne, P. (2006) On calibration estimation for quantiles. *Survey Methodology*, **52**, 37-52.

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

Plikusas, A. (2003) Calibrated weights for the estimators of the ratio. *Lithuanian Mathematical Journal*, **43**, 543-547.

Plikusas, A., Pumputis, D. (2007) Calibrated estimators of the population covariance. *Acta Applicandae Mathematicae*, **97**, 177-187.

Rueda, M., Martínez, S., Martínez, H., Arcos, A. (2007) Calibration methods for estimating quantiles. *Metrika*, **66**, 355-371.

Rueda, M., Martínez, S., Martínez, H., Arcos, A. (2007) Estimation of the distribution function with calibration methods. *Journal of Statistical Planning and Inference*, **137**, 435-448.

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

Singh, S., Horn, S., Chowdhury, S., Yu, F. (1999) Calibration of the estimators of variance. *Australian and New Zealand Journal of Statistics*, **41**, 199-212.

Sitter, R. R., Wu, C. (2002) Efficient Estimation of Quadratic Finite Population Functions in the Presence of Auxiliary Information. *Journal of the American Statistical Association*, **97**, 535-543.

APPLICATION OF IMPUTATION METHODS FOR SAMPLING ESTIMATION

Iryna Rozora[1] and Natalia Rozora[2]

[1] Kyiv National Taras Shevchenko University, Ukraine
e-mail: irozora@bigmir.net

[2] Nielsen Ukraine company
e-mail: rozora@ukr.net

Abstract

The focus of this paper is to provide the overview of statistical methods which used for estimation of missing data and which also can be applied to estimate sampled data, mentioning advantages and shortcoming of each method.

1 Introduction

Missing data is a frequent issue of statistical research studies, which imply extensional data collection and analysis. Though commonly it is carried out operational procedures to have a complete data, the real-fact situation is that some data may be not collected for different reasons like some of responders decline to provide the data or conceal it to protect confidentiality. Especially such facts is frequent for economic, political and social studies, when responders resist to provide information due to privacy connected with developing competition, etc. Another area, which defines the recognition of the missing data as a crucial factor for the analysis, is medical studies, for example studies connected with cancer diseases. Besides the operational reasons missingness could be provoked by the specific of sample design for the research study.

Technology development makes possible implementation in practice complicated theoretical approaches and evolving more sophisticated computations. Recently the development of statistical methods was concentrated to address the incomplete data problem, which causes bias or inefficient analysis, including such methods like the imputation, likelihood and weighting approaches. Overviews of different methods are given in Little and Rubin (1990), GSS (1996), Schafer and Graham (2002), Raghunathan (2004) and Ibrahim et al (2005), Little and Rubin (2002), Schaffer (1997), etc.

Generally speaking, one can consider sample data case as such with missing data from the population, thus extending missing data evaluation methods as an approach for sample data expansion. From this perspective it is important to understand the advantages and shortages of each of the methods, which allow better fit the reality of real data deficiency.

The following notation and classification is used. Missing data is defined as *univariate* if missing values only occur in a single response variable; and is defined as *multivariate* if missing values occur in more than one variable. When the observation is not done whole sampled unit (element of the sample) it is considered as *unit non-response,* while a failure of obtaining of part of information for observed unit is defined as *item non-response.* For

example, having a questionnaire list, one of the sampled responder did not answer a few questions, but answered most of other questions.

Let introduce some notation for the further definitions:
Let U be a finite universe of N units and s a sample of sample size n. Let H denote the complete data matrix with element h_{ik}, where $i = 1,...,n$, and $k = 1,...,K$. H_{obs} refers to the observed part of the matrix H and H_{mis} to the missing part. Let R denote a matrix with elements

$$r_{ik} = \begin{cases} 1 \text{ if } h_{ik} \text{ observed,} \\ 0 \text{ if } h_{ik} \text{ missing.} \end{cases}$$

Let y_i for unit i denote the sample value of the variable subject to missing data, r a binary indicator of whether y is observed and x a vector of fully observed auxiliary variables, $x_i = (x_{1i} ,x_{2i} ,...,x_{Pi})$. The notation Y denotes a vector and X a matrix of values respectively for multivariate cases. It is then $h = (x ,y)$.
Besides the known design of the sample done by researcher, it is important to understand how sample was selected and implemented and why more data are missing if such in order to make correct assumptions for the model of estimation. Informally missing data can be thought of being caused by one of three ways – random process, measured process, and process that not measured. Modern missing data methods work well with first two, but not the last. More formally, the major problem with missing data is that the distribution $f(R \mid H)$, referred to as the nonresponse mechanism, where f denotes the probability density function, is unknown, i.e. it is unknown how nonresponse for each variable is generated, so researcher have to make assumptions about this distribution of nonresponse, which often cannot be verified.
It is distinguished (Rubin, 1987) 3 types of assumptions about missing data distribution: missing completely at random (MCAR), missing at random (MAR), not missing at random (NMAR).
- Data are ***missing completely at random (MCAR),*** defined as $f(R \mid H_{obs} ,H_{mis}) = f(R)$, if the missingness depends neither on H_{obs} nor on H_{mis}. This means that the probability of response depends neither on the variable subject to nonresponse, nor on any other variable. When data are MCAR missing cases are not different from not missing ones in terms of the analysis being performed. MCAR assumption is a strong assumption, which is likely to be violated in many social science applications thus it is considered some weaker assumptions.
- If the distribution of the missing-data mechanism does not depend on the missing values H_{mis} it is said that the data are ***missing at random (MAR)***, i.e. $f(R \mid H_{obs} ,H_{mis}) = f(R \mid H_{obs})$, which implies that missingness does depend on the observed but not on the missing values. Under MAR it is possible that the missingness depends on y_i , however, when conditioning on other values x_i this dependency is wiped out and there is no residual relationship between y and r . The MAR assumption is a weaker assumption than MCAR. If either MCAR or MAR hold the missing-data structure is ignorable and produce unbiased results in the analysis.
- If the probability that an item is missing depends on the variable itself even when conditioning on the observed values, and therefore neither MAR nor MCAR hold, the missing-data mechanism ***is not missing at random (NMAR)*** or nonignorable (Schafer, 1997; Little and Rubin, 2002). Since the missing data depends on events, which researcher cannot measure, this is the most difficult situation for the analysis.
It is important to note that MCAR is testable given observed data but MAR is usually untestable since nonrespondent data are unobserved. Often researchers regard the MAR assumption as a workable approximation although MAR may not hold in reality. To analyze the effects of MAR-based methods under a departure from MAR is therefore desirable.
Let consider further each type of estimation of missing data.

2 Weighting Method

In this approach a model for the probability of missigness is fit and the inverse of these probabilities are used as weights for the completed cases.
In ideal study the sample assumed random with equal probability of selections and complete item data for the sampled unit. In ideal case the expantion factor used is equal to fraction of sample from the total population. The problem in practice that even is one gets data from the unit selected, data still have item non-response. In practice it is the most commonly used method, when incomplete observations are dropped from the sample study. Especially it is widely-used in case of large proportion of item non-responses for the specific unit . The variety of methods is related to list-wise or case deletion, pair wise deletion of observations, etc. Weighting methods are fit in the software allowing for the weighs, like SAS, SPSS, Stata, etc. Although these methods are applied by social scientists they have several shortcomings which are reported in Schafer and Graham (2002). Such methods are usually found not adequate to compensate for nonresponse bias in particular when estimating parameters other than means and may only be valid under strong assumptions about the mechanism that generated the missing values. In addition, variance estimation may not be straightforward and relationships between variables may be distorted. Exclusion of variables with missing data is not also attractive as may imply of exclusion of significant portion of data from the analysis.
Some adjustments to weighting methods are discussed in David et al. (1983), Little (1986) and Little and Rubin (2002).

3 Imputation

Imputation is a method to fill in missing data with plausible values to produce a complete data set. A distinction may be made between *deterministic and stochastic (or random) imputation methods*. Given a selected sample deterministic methods always produce the same imputed value for units with the same characteristics. Stochastic methods may produce different values.
The main reason for carrying out imputation is to reduce nonresponse bias, which occurs because the distribution of the missing values, assuming it was known, generally differs from the distribution of the observed items. When imputation is used, it is possible to recreate a balanced design such that procedures used for analyzing complete data can be applied in many situations. Rather than deleting cases that are subject to item-nonresponse the sample size is maintained resulting in a potentially higher efficiency than case deletion. Imputation usually makes use of observed auxiliary information for cases with nonresponse maintaining high precision (Schafer and Graham, 2002). In terms of sampling estimating, having know auxiliary variable fully observed for population, one can have more efficient estimation applying imputation if the variable of interest has strong relationship with auxiliary ones. However, it can have serious negative impacts if imputed values are treated as real values. To estimate the variance of an estimator subject to imputation adequately, often special adjustment methods are necessary to correct for the increase in variability due to nonresponse and imputation. It is also possible to increase the bias by using imputation if the relationship between known and unknown variables is poor (Kalton, 1983; Little and Rubin, 2002).

Under imputation let Y. denote the vector of imputed and observed values of Y in the univariate case, such that

$$y_{.i} = \begin{cases} y_i \text{ for } r_i = 1, \\ y_i^I \text{ for } r_i = 0, \end{cases}$$

where $i \in s$ and y_i^I denotes the imputed value for nonrespondent *i*. Let θ denote the parameter of interest in the population, e.g. a mean or a regression coefficient, which is a

function of the data in the population, and $\hat{\theta}$ an estimator of based on the sample in the case of full response, such that $c = \hat{\theta}\ (H)$.

Applying imputation in the case of nonresponse an estimator is obtained of the form $\hat{\theta} = \hat{\theta}\ (H_{obs}, H_{mis})$ called the ***imputed estimator***. The aim is to define an approximately unbiased and efficient estimator by choosing an appropriate imputation method. Another important aspect of an imputation method is its robustness under misspecification of underlying assumptions. When choosing among imputation procedures it is important to consider carefully the type of analysis that needs to be conducted. In particular, it should be distinguished if the goal is to produce efficient estimates of means, totals, proportions or other official aggregated statistics, or a complete micro-data file that can be used for a variety of different analyses. Other issues when choosing an imputation method are the availability of variance estimation formulae and practical questions concerning implementation and computing time.

3.1 Simple Imputation Methods

This is deductive methods to impute a missing value by using logical relations between variables and derive a value for the missing item with high probability. The method of (unconditional) mean imputation imputes the overall mean of a numeric variable for each missing item within that variable. A variation of this method is to impute a class mean, where the classes may be defined based on some explanatory variables. Disadvantages of such procedures are that distributions of survey variables are compressed and relationships between variables may be distorted (Kalton, 1983; Little and Rubin, 2002).

3.2 Regression Imputation

Predictive regression imputation, also called deterministic regression or conditional mean imputation, involves the use of one or more auxiliary variables, of which the values are known for complete units and units with missing values in the variable of interest. A regression model is fitted that relates y_i to auxiliary variables x_i , i.e. the imputation model. The predicted values are used for imputation of the missing values in Y . Usually, linear regression is used for numeric variables, whereas for categorical data logistic regression may be used. A potential disadvantage of predictive regression imputation is that it distorts the shape of the distribution of the variable Y and the correlation between variables, which are not used in the regression model. It might also artificially inflate the statistical association between Y and the auxiliary variables.

Under random regression imputation , sometimes referred to as imputing from a conditional distribution , the imputed value for the variable Y is a random draw from the conditional distribution of Y given X . If a linear model between Y and X is considered a residual term is added to the predicted value from the regression, which allows for randomisation and reflects uncertainty in the predicted value. This residual can be obtained in different ways, e.g. by drawing from a normal distribution, either overall or within subclasses, or by computing the regression residuals from the complete cases and selecting an observed residual at random for each nonrespondent. A random regression model maintains the distribution of the variables and allows for the estimation of distributional quantities (Kalton and Kasprzyk, 1982; Kalton, 1983).

An advantage of regression imputation is that it can make use of many categorical and numeric variables. The method performs well for numeric data, especially if the variable of interest is strongly related to auxiliary variables. The imputed value, however, is a predicted value either with or without an added on residual and not an actually observed value. Another potential disadvantage of such a parametric approach is that the method may be sensitive to model misspecification of the regression model. If the regression model is not a good fit the predictive power of the model might be poor (Little and Rubin, 2002).

3.3 Hot Deck Imputation Methods

Such approaches has been developed to assign the value from a record with an observed item, the donor, to a record with a missing value on that item, the recipient. Such imputation methods are referred to as donor or hot deck methods, setting $y_j^I = y_{i^*}$ for some donor respondent i^* (Kalton and Kasprzyk, 1982; Lessler and Kalsbeek, 1992). This involves consideration of how best to select the donor value. A simple way is to impute for each missing item the response of a randomly selected case for the variable of interest. Alternatively, imputation classes can be constructed, selecting donor values at random within classes. Such classes may be defined based on the crossclassification of fully observed auxiliary variables. An advantage of the method is that actually occurring values are used for imputation. Hot deck imputation is therefore common in practice, and is suitable when dealing with categorical data. Hot deck methods are usually non-parametric (or semi-parametric) and aim to avoid distributional assumptions. Under hot deck imputation the imputed values will have the same distributional shape as the observed data (Rubin, 1987). For a hot deck method to work well a reasonably large sample size may be required.

3.4 Nearest-Neighbour Imputation

Nearest-neighbour imputation, also called distance function matching, is a donor method where the donor is selected by minimizing a specified 'distance' (Kalton, 1983; Lessler and Kalsbeek, 1992). This method involves defining a suitable distance measure, where the distance is a function of the auxiliary variables. The observed unit with the smallest distance to the nonrespondent unit is identified and its value is substituted for the missing item according to the variable of concern. The easiest way is to consider just one continuous auxiliary variable X_1 and to compute the distance D from all respondents to the unit with the missing item, i.e. $D_{ji} = |x_{j1} - x_{i1}|$, where j denotes the unit with the missing item in Y. The missing item is replaced by the value y* , where the respondent i* is the donor for nonrespondent j if $D_{ji^*} = \min_i D_{ji}$.

An advantage of nearest neighbour imputation is that actually observed values are used for imputation. Another advantage may be that if the cases are ordered for example geographically it introduces geographical effects. However, the variance of $\hat{\theta}(y)$ under nearest neighbour imputation may be inflated if certain donors are used much more frequently than others. The multiple usage of donors can be penalised or restricted to a certain number of times a donor is selected for imputation. For example, the distance function can be defined as $D_{ji^*} = \min_i \{|x_{j1} - x_{i1}| \cdot (1 + \mu t_i)\}$, where $\mu \in R^+$ is the assigned penalty for each usage, t_i is the number of times the respondent i has already been used as a donor (Kalton, 1983).

3.5 Predictive Mean Matching Imputation

A hot-deck imputation approach that makes use of the regression or imputation model, is the method of predictive mean matching imputation. In its simplest form it is nearest neighbour imputation where the distance is defined based on the predicted values of y_i from the imputation model, denoted $\hat{y}_i$. Predictive mean matching is essentially a deterministic method. Randomisation can be introduced by defining a set of values that are closest to the predicted value and choosing one value out of that set at random for imputation (Little and Rubin, 2002). Another form of predictive mean matching imputation is hot deck imputation within classes where the classes are defined based on the range of the predicted values from the imputation model. This method achieves a more even spread of donor values for imputation within classes, which reduces the variance of the imputed estimator. The method of predictive mean matching is an example of a composite method, combining elements of

regression, nearest-neighbour and hot deck imputation. Since it is a semi-parametric method, which makes use of the imputation model but does not fully rely on it, it is also assumed to be less sensitive to misspecifications of the underlying model than for example regression.

3.6 Repeated Imputation: Multiple and Fractional Imputation

Apart from single value imputation discussed, where one value is imputed for each missing item, it is also possible to use repeated imputation, in the sense that M , M > 1, values are assigned for each missing item, by repeating a random imputation method several times.

There are two reasons for using repeated imputation. One reason is to reduce for example the random component of the variance of the estimator arising from imputation. Another reason for using repeated imputation is simplification of variance estimation of a point estimator which may be difficult in the presence of imputation as indicated earlier. The method of multiple imputation (MI), as proposed by Rubin (1987), is also a form of repeated imputation in the sense that several values are assigned for each missing item. The idea behind this approach is that the repeated imputed values themselves already reflect uncertainty about the true but non-observed values.

Rubin (1987) presented this method for combining results from a data analysis performed *m* times, once for each of *m* imputed data sets, to obtain a single set of results.
From each analysis, one must first calculate and save the estimates and standard errors. Suppose that $\hat{Q}_j$ is an estimate of a scalar quantity of interest (e.g. a regression coefficient) obtained from data set *j* (*j*=1,2,...,*m*) and U_j is the standard error associated with $\hat{Q}_j$. The overall estimate is the average of the individual estimates, $\bar{Q} = \frac{1}{m}\sum_{j=1}^{m}\hat{Q}_j$. For the overall standard error, one must first calculate the within-imputation variance, $\bar{U} = \frac{1}{m}\sum_{j=1}^{m}U_j$ and the between-imputation variance, $B = \frac{1}{m-1}\sum_{j=1}^{m}(\hat{Q}_j - \bar{Q})^2$.

The total variance is $T = \bar{U} + (1+\frac{1}{m})B$.

Confidence intervals and significance test are obtained by considering $t = \bar{Q}/\sqrt{T}$, that has Student's t-distribution with degrees of freedom $df = (m-1)\left(1+\frac{m\bar{U}}{(m+1)B}\right)^2$.

Additional methods for combining the results from multiply imputed data are reviewed by Schafer (1997, Ch. 4).

An advantage of MI (multiple imputation) is that it is possible to produce complete micro-data files that can be used for a variety of analyses. This is particularly useful when providing a public use dataset that may be analyzed by a wide range of researchers with different types of analyses in mind. Consideration needs to be given to the definition and choice of the imputation model and the relationship to the analysis model. Generally, the imputation model should be chosen such that it coincides approximately with subsequent analysis performed on observed and imputed data, e.g. regression analysis. The model should be rich enough in the sense that it should preserve associations and relationships among variables that are of importance to the subsequent analysis.

The number of multiple imputations is for many applications recommended to be between 3 and 10, which may make the computational burden feasible in particular when using modern

computer software. MI has the advantage to offer a relatively simple and flexible variance estimation formula, in the sense that it is in principle applicable to any type of imputed estimator. MI can also be used to fill in missing values in a multivariate missing data setting, and is suitable for numeric and categorical variables. It is currently probably the most practical and general approach, in particular for social scientists carrying out a large number of different analyses and missing values in several variables.

One approach to imputation when there are both continuous and discreet missing values is the **Conditional Gaussian approach**, popularized by Schafer (1997). A log-linear model is specified for discrete random variables and conditional on this distribution a multivariate normal distribution is assumed for continuous variables. This general location model can be fit as a saturated multinomial with separate means and shared covariance. This approach is implemented in MIX program.

Raghunathan et al. (2001) developed a sequential regression approach to MI, it is also called **Chained Equation approach**. The idea is to regard a multivariate missing data problem as a series of univariate missing data problems. The main procedure is as follows: First regress Y_1 on the set of fully observed variables X and impute the missing cases in Y_1 for example using a random regression imputation method, then regress Y_2 on Y_1 and X and so forth until Y_q . This procedure is repeated c times, however, now including all variables as predictors in the regression models apart from the variable being imputed. After c rounds the final imputations are used. Repeating the process M times results in M multiple imputations.

An advantage of the method may be that a specific form for the multivariate distribution can be avoided. The method, however, assumes that the multivariate posterior distribution exists, which may not always be the case, leading to non-convergence of the algorithm. There is thus a lack of a well established theoretical basis, and a note of caution needs to be applied, although the method is computationally attractive.

4 Software available for imputation

For multiple imputation a wide range of free and commercial software has been developed in recent years which makes MI more widely applicable to many researchers. The following gives a brief overview and update of programmes that are currently available. All of these procedures assume MAR and are not readily available for nonignorable nonresponse mechanisms.

The SPSS procedure Missing Value Analysis (MVA) includes a variety of techniques to analyse the missing data pattern, including a test of MCAR . It calculates some basic statistics of variables subject to nonresponse and handles missing data based on a listwise or pairwise method, regression imputation or the EM (estimation – maximization) method, a maximum-likelihood based method, etc.

Another computer package that is often used by social scientists is STATA. STATA includes options for various forms of hot deck imputation based on the approximate Bayesian bootstrap and regression imputation.

Package Splus used by statisticians and economists facilitates the implementation of hot deck procedures, regression and predictive mean matching imputation. Splus includes a missing data analysis library which enables parametric model-based procedures. It facilitates the implementation of the EM algorithm and MI based on data augmentation (MCMC) for numeric variables, assuming multivariate normality, and for categorical and mixed variables as described in Schafer (1997).

The programme PAN, NORM, Cat, Mix have been developed for panel data (Schafer, 2001). The missing data library includes methods for the analysis of convergence, the analysis of multiple complete datasets and options for the analysis of missing data patterns. Norm, Cat, Mix and Pan are available from http://www.stat.psu.edu/~jls/misoftwa.html.

The SAS-based programme SEVANI, System for Estimation of Variance due to Nonresponse and Imputation, developed at Statistics Canada, enables variance estimation for certain types of estimators under imputation methods such as regression and nearest neighbour imputation.

The IVEware, Imputation and Variance Estimation Software, is implemented in SAS and performs single and multiple imputations of missing values using the sequential regression imputation method (Raghunathan et al., 2001; http://www.isr.umich.edu/src/smp/ive/). It is also available as a stand-alone software.

SOLAS is a commercial programme to perform six imputation techniques including two techniques for MI and benefits from a well designed user interface. It incorporates mean imputation, hot deck imputation either overall or within imputation classes and regression imputation imputing predicted values. MI can be implemented either by using parametric regression imputation or by the propensity score method. (More information about SOLAS is available from http://www.statsol.ie/solas/solas.htm).

Other software programmes such as AMELIA, EMCOV, etc. are also available for imputation, more information about about which is available from http://www.multiple-imputation.com.

References

[1] Little, R.J.A. and Rubin, D.B. (2002): *Statistical Analysis with Missing Data*, New York.

[2] Schafer, J. L. (1997): *Analysis of Incomplete Multivariate Data*, London.

[3] Little, R.J.A. and Rubin, D.B. (1990): The Analysis of Social Science Data with Missing Values, *Sociological Methods and Research*, 18, 3, 292-326.

[4] www.missingdata.org.uk

[5] Schafer, J.L. and Graham, J.W. (2002): Missing Data: Our View of the State of the Art*, Psychological Methods,*7, 2, 147-177.

[6] Ibrahim, J.G., Chen, M.H. Lipsitz, S.R. and Herring, A.H. (2005): Missing-Data Methods for Generalised Linear Models: A Comparative Review, *Journal of the American Statistical Association*, 100, 469, 332-346.

[7] Raghunathan, T.E., Lepkowski, J.M. van Hoewyk M., Solenberger P.W. (2001): A Multivariate Technique for Multiply Imputing Missing Values using a Sequence of Regression Models, *Survey Methodology*, 27, 85-95

[8] Rubin, D.B. (1987): *Multiple Imputation for Nonresponse in Surveys*, New York, Chichester.

[9] Kalton, G. (1983): *Compensating for Missing Survey Data,* Michigan.

[10] Kalton, G. and Kasprzyk, D. (1982): Imputing for Missing Survey Responses*, Proceedings of the Section on Survey Research Methods*, American Statistical Association, 22-31

[11] Lessler, J.T. and Kalsbeek W.D. (1992): *Nonsampling Error in Surveys*, New York, Chichester.

[12] www.multiple-imputation.com

IMPLEMENTING OF SURVEY SAMPING STUDYING AT BOHDAN KHMELNITSKY NATIONAL UNIVERSITY OF CHERKASY

Svitlana Rychka[1]

[1] Bohdan Khmelnistky National University of Cherkasy, Ukraine
e-mail: rychka-svetlana@ukr.net

Survey sampling is an area of statistics that needs a great developing in Ukraine. Survey information is very important for a healthy development of the democratic life of Ukrainian society, for efficiency of the government, and for effective business operations in a market-oriented society [1].

The Bohdan Khmelnitsky National University of Cherkasy has actively begun to introduce the studying of survey sampling.

Beginnings from 2007 there have been defended several term papers on the mathematical faculty which are connected to statistics of sample surveys.

In 2009 the first final thesis to obtain the bachelor diploma on “Methods of the sample surveys” was defended. The basic methods of the sample surveys were considered in that works. Large attention was devoted to the methods of processing data with nonresponse.

The course on “Statistical data processing” is given at the National University of Cherkasy. Within it students specializing in “Mathematics” and “Applied mathematics” become acquainted with some methods of survey sampling. They processed their work in Statistica, SPSS, Matlab, Mathcad. The bachelor thesis on “Processing statistical data with Statistica and Mathcad” were defended in 2009. There was considered the practical realization of exercises connected to survey sampling in Statistica and Mathcad.

Students on the economic faculty of the National University of Cherkasy also use the methods of surveys sampling in their term and diploma theses.

References

[1] Parkhomenko, V. (2001) *Survey Sampling Methods*. Kyiv (in Ukrainian).

Creating Networks for Simulation on Network Sampling

Termeh Shafie[1]

[1] Stockholm University, Department of Statistics, Sweden
e-mail: termeh.shafie@stat.su.se

Abstract

One approach for sampling members of a rare or hidden population while still obtaining unbiased estimates of population characteristics is through the use network sampling. To do this, network data is needed. In this paper, a model is proposed for generating networks with a structure of a certain complexity and flexibility. Further, some network properties are analyzed and discussed for simulated networks.

1 Introduction

In an empirical research, the investigator must attend two issues: sampling and measurement. Using the extra information from the network structure allows for the design of a sampling and estimation scheme that may be both more feasible and cheaper to apply. Also, standard sampling and estimation techniques may not be adequate tools when sampling social networks since they require known sampling frames and inclusion probabilities. This may not always be the case when dealing with networks. For instance, when dealing with social networks, the population may be hidden or hard-to-reach (e.g. illegal immigrants, drug users, commercial sex workers etc). An alternative to statistical standard analysis techniques is using a network perspective when approaching the study of these hidden populations. There are various types of sampling methods developed for network data, where the actors of a system and their connectivity in that system plays an important part in the sampling process. Some examples are time-space sampling (Muhir et al. 2001), respondent-driven sampling (Heckathorn, 1997) and snowball sampling (Frank, 1978).
Real-life networks are generally very large, implying that it is a time-consuming task to collect data to delineate their structure in detail. This makes it desirable to develop models that capture essential features of real networks. In this paper a model is presented for generating flexible networks with a certain degree of complexity in structure. An algorithm for generating such networks is proposed and can be used for simulation and for theoretical calculations.

2 Network and Network Analysis

During recent decades, network analysis have attracted interest from the social and behavioral science community. This interest can be ascribed to the appealing focus of network analysis on connections among network entities. The pattern and the implications of these connections attract researchers since they produce leverage for answering social and behavioral science questions by giving precise formal definition to aspects of the political, economic, or social structural environment.

A network is a set of items, which can be called nodes, with connections between them, called edges. There are many kind of systems taking the form of networks abound the world. Different types of relations identify different networks, even when imposed on the identical set of elements. Measurements of connectivity are essential for the characterization, analysis, classification and modeling of networks. By forming networks of interesting people, organizations, places and events, connected with links that indicates connection, we can create conceptual images that facilitate understanding. Some examples of different kinds of networks are the Internet, science collaboration networks, cellular networks, ecological networks and social networks of acquaintances or other connections between individuals.
Network analysis emphasizes the relationships that connect the positions within a system. The organization of structures thus becomes a central concept in analyzing the structural properties of a network and understanding particular elements within the structure. These analysis take into the account both the relations that occur and those that do not exist between the actors. The formation of present and absent links among the network uncovers the network structure. A theoretical problem in network analysis is to explain the occurence of different structures and to account for the variation in linkage to other actors.

2.1 Graph Theory

There are a variety of procedures for describing and analyzing network data, one which is related to mathematical graph theory. In this section some elementary terms for visual and algebraic representations of network data is presented.
First some notations. A graph G consists of a set of nodes $\{1, 2, \ldots, N\}$ and an edge set E representing the relations between the nodes. The number k_i of edges incident with a given node i is called its degree. The classical algebraic representation of a graph is done using a $N \times N$ adjacency matrix $\mathbf{A}$. For instance, in a directed graph the actors arrayed in the matrix rows are the initiator of of the specified relation and the actors arrayed across the columns are the recipients of that relation. The elements of the matrix are N^2 values indicating the linkage between every pair of nodes in the network. These values are binary and defined as

$$z_{ij} = \begin{cases} 1 & \text{if there exists a link from node } i \text{ to node } j \\ 0 & \text{otherwise} \end{cases}$$

A general adjacency matrix is given as

$$\mathbf{A} = \begin{pmatrix} z_{11} & z_{12} & \cdots & z_{1N} \\ z_{21} & z_{22} & \cdots & z_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ z_{N1} & z_{N2} & \cdots & z_{NN} \end{pmatrix}. \tag{1}$$

By summing across matrix entries some useful statistics can be obtained. The outdegree of actor i is the sum of 1s whithin actor i's row

$$\text{outdegree}_i = \sum_{j=1}^{N} z_{ij}, \tag{2}$$

and actor j's indegrees is calculated as the sum of 1s within j's columns

$$\text{indegree}_j = \sum_{i=1}^{N} z_{ij}. \tag{3}$$

Note that when the graph is undirected, the in- and outdegrees are equal.

The order and size of the graph are defined to be the number of nodes and egdes, respectively. If all $N(N-1)/2$ possible lines between the set of N nodes are present, the graph is said to be complete (that is, the graph in which every pair of nodes is linked by an edge). Two nodes are adjacent if a line directly connects them. A directed graph (or digraph) consists of the N points linked by a set of directed lines. The density of a graph is a ratio of its actual size to the maximum possible size of N nodes, and is in the range $[0, 1]$.
Now, an alternative algebraic representation of a graph is given taking into account not only the link information, but the link information given different characteristics of nodes and links. To clarify an eaxmple is given. Assume that we are looking at four persons in a social network denoted A,B,C and D. These four nodes have different properties (for instance gender) and relations to one another. See Figure (1).

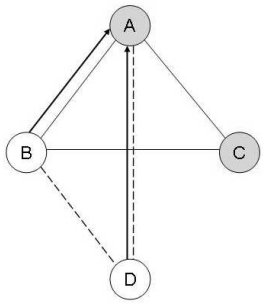


Figure 1: An example of a network consisting of four nodes. The solid lines represent work collegues, the dashed line represents neighbours and the directed arrows represent appreciation. The white and shaded circles may for instance be men and women.

The solid lines represent work collegues, the dashed line represents neighbours and the directed arrows represent appreciation by a node to another. As seen, person A, B and C work together and B appreciates A as a collegue. Person A has B and D as neighbors and D appreciates A as a neighbor. There is no link between person C and D.
In a two dimensional matrix representation we would only see the links between the nodes (as given in matrix (1)), not the characteristics of these relations. Thus, not only may the link information be of interest, but also information about the type of relations moving across the graph. This valuable information should be considered when performing network analysis. The general algebraic presentation of the example above is as given below.
We have a set of nodes $1, \ldots, N$ and properties for each node denoted by X_{ik} where $i = 1, \ldots, N$ and $l = 1, \ldots, p$. These properties may be out degrees or the number of edges from the nodes. Other may be ordinary variables like monthly income or political preference. A $N \times p$ matrix where the nodes are the matrix rows and the columns represents the node properties is

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Np} \end{pmatrix}. \tag{4}$$

Further, a set of edges or links between node pairs i and j is defined as

$$(i, j) \in E \subseteq \{(X, Y); \quad X, Y = 1, \ldots, N, \quad X \neq Y\}. \tag{5}$$

and properties for the node pair links are defined as

$$Z_{ijk}(i,j) \in E \quad k = 1, \ldots, q \tag{6}$$

Letting M denote the count of the subsets E, a $M \times q$ matrix can be created where the rows represent node pairs and the link properties are given in the columns.
Following the example given in Figure (1) we have X as a 4×2 gender matrix where the elements are 1 if woman and 0 if man,

$$\mathbf{X} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{7}$$

Further, we have that A is the three dimensional tensor where the binary elements z_{ijk} are the values of a relation k between individual i and j. Thus we have for property 'appreciation' and relation 'collegue'

$$\mathbf{A} = \begin{pmatrix} - & 1 & 1 & 0 \\ 2 & - & 1 & 0 \\ 1 & 1 & - & 0 \\ 0 & 0 & 0 & - \end{pmatrix}, \tag{8}$$

and for property 'appreciation' and relation 'neighbour'

$$\mathbf{A} = \begin{pmatrix} - & 1 & 0 & 1 \\ 1 & - & 0 & 0 \\ 0 & 0 & - & 0 \\ 2 & 0 & 0 & - \end{pmatrix}. \tag{9}$$

3 Generating a Network

There are $\{1, 2, \ldots, N\}$ nodes in the population. For each node a lot of properties are determined from a multivariate distribution F, one component of which is the degree, e.i. we first generate matrix $\mathbf{X}$ as defined in previous section. This is done independently but conditional on the sum of the degrees being even (the number of in- and outdegrees are equal). Given these links, properties for all links can be created. These properties should be independent for different links given the properties of the nodes. The degree counts can also be obtained from a given distribution and the graph is simulated thereafter. When the number k_i of node i is determined for each node, the network is constructed by the following algorithm. A simple graph is described here with only one type of undirected edges.

1. Let k_i denote the degree of node i where $i = 1, \ldots, N$.
2. The cumulative degree is given as

$$\sum_{j=1}^{i} k_j = r_i, \quad r_0 = 0. \tag{10}$$

3. Next, create a row vector of length r_N where

$$x_i = m, \quad \text{if} \quad r_m \geq i \geq r_{m-1} \tag{11}$$

It will look as follows

$$\left[\underbrace{1 \quad \cdots \quad 1}_{k_1 \text{ elements}} \quad \underbrace{2 \quad \cdots \quad 2}_{k_2 \text{ elements}} \quad \cdots \quad \underbrace{N-1 \quad \cdots \quad N-1}_{k_{N-1} \text{ elements}} \quad \underbrace{N \quad \cdots \quad N}_{k_N \text{ elements}}\right] \tag{12}$$

e.g.

$$[1 \quad 1 \quad 1 \quad 2 \quad 2 \quad 3 \quad 4 \quad 4 \quad 4 \quad 4 \quad 5 \quad 7 \quad 7 \ldots]$$

corresponds to $k_1 = 3$, $k_2 = 2$, $k_3 = 2$, $k_4 = 5$, $k_5 = 1$, etc.

4. Permute vector randomly. Note that there are $r_N!$ possible permutations available here.
5. Create pairs in the vector by dividing into sequences of length 2. These are the node pairs.
6. If no nodes have link loops to themselves and no link pairs appear twice or more, accept the generated network. Otherwise, restart from (4).

We need the condition that the proposed network may be formed, e.g. no degree may be larger than half the sum of degrees, or larger than $(N-1)$. In the cases discussed in this paper, only the first condition (that the degree sum is even) is essential from a practical point of view.
A similar routine may be created for digraphs but the condition will now be that the sum of the out degrees equals the sum of in degrees, which is a more restrictive condition from a simulation point of view.
It can be shown that the number of rejected networks is relatively small and that a graph with the following properties can be created within a reasonable amount of time, as long as the degree order k_i is kept low relative to the graph size.

4 Properties of a Network

Some properties of networks are presented and later applied on simulated networks.

4.1 Transitivity/Clustering

In many networks it is found that if node A is connected to node B and node B connected to node C, then there is a higher probability that node A also will be connected to node C. Considering social networks, this means that the friend of your friend is likely also to be your friend. A measure for this is the clustering coefficient. Watts and Strogatz (1998) introduced the clustering coefficient for the whole system as

$$C = \frac{1}{N}\sum_i C_i, \tag{13}$$

where C_i is the local cluster coefficient for a given node i

$$C_i = \frac{2E_i}{k_i(k_i-1)}, \tag{14}$$

and E_i is the number of links between the nearest neighbours of node i. The sum in equation (2) is restricted to nodes with degrees larger than 1. It can be said that the cluster coefficient quantifies how close the node and its neighbors are to being a clique or a complete graph, e.i. a graph in which every node is connected to every other node in the graph. This fraction can range from 0 to 1.

4.2 Average Path Lengths

The average distance between all pairs of nodes is a way to determine whether a network is compact- with shortest path linking most pairs- or spread out, with many long paths. Average path legnth is defined as the average number of steps along the shortest paths for all possible pairs of network nodes. It can be viewed as a measure of how quickly information can flow through the network. Most real networks have a very short average path length leading to the concept of a small world where everyone is connected to everyone else through a very short path.
The average path length l_G for an unweighted graph G with N nodes is defined as

$$l_G = \frac{1}{N(N-1)} \sum_{i,j} d(z_i, z_j), \tag{15}$$

where $d(z_i, z_j)$ denotes the number of edges (or distance), corresponding to the shortest path, required to get from node i to node j, where $i \neq j$.

5 Simulation Study

In this section, the strucutural properties of a simulated networks following the algorithm given in section (3) is considered. The hypothetical populations created here consist of two equally sized groups conditioned on having fixed number of degrees in each group.
As mentioned in section (4) of this paper, among the structural parameters, the average shortest path length (APL) and the cluster coefficient (CC) are important characteristics of a graph. APL is the average fewest number of steps it takes to get from each node to every other, and is thus an emergent property of a graph indicating how compactly its nodes are interconnected. CC is the average probability that any pair of nodes is linked to a third common node by a single edge, and thus describes the tendency of its nodes to form local clusters. High values of both APL and CC are found in regular graphs, in which neighboring nodes are always interconnected yet it takes many steps to get from one node to the majority of other nodes, which are not close neighbors. At the other extreme, if the nodes are instead interconnected completely at random, both APL and CC will be low.
The results from 16 different simulated networks with arbitrary links is presented. Each population consist of two equally sized groups in two population sizes, $N = 10$ and $N = 100$, where each group has been given fixed out-going links conditional on their in-going links. The structural properties of these networks are quantified by their clustering coefficient (CC) and average path lengths (APL). Table 1 and 2 present calculated 95% predicition intervals for these two measures. These intervals are of interest when simulating the distribution of a real network with the given degree counts.

6 Results and Discussion

Not all simulations are ready at this moment so the simulation results will be discussed during presentation.

Table 1: 95% *prediction intervals for 1000 simulations of the the cluster coefficient (CC) in a two group network consisting of size* $N = 10$ *and* $N = 100$*. Each group in the population is equally large and simulations are run for some arbitrary degree combinations,* k_1 *and* k_2 *(see Appendix B).*

	CC (N=10)		CC (N=100)	
k_1,k_2	lower	upper	lower	upper
2,4	0.0008	0.5090	0.0088	0.0466
3,5	0.2056	0.5268	0.0041	0.0503
4,6	0.4014	0.5915	0.0162	0.0562
5,5	0.3334	0.5201	0.0138	0.0509

Table 2: 95% *prediction intervals for 1000 simulations of the avergae path legnth (APL) in a two group network consisting of size* $N = 10$ *and* $N = 100$*. Each group in the population is equally proportioned and simulations are run for some arbitrary degree combinations,* k_1 *and* k_2 *(see Appendix B).*

	APL (N=10)		APL (N=100)	
k_1,k_2	lower	upper	lower	upper
2,4	1.7245	2.1021	0.0100	0.0477
3,5	1.5436	1.6661	3.4521	3.5710
4,6	1.4295	1.4666	2.9968	3.0608
5,5	1.4444	1.4444	3.0094	3.0754

References

Frank, O. 1978. *Sampling and Estimation is Large Social Networks.* Social Networks 1: 91-101.

Heckathorn, D. D. 1997. *Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations.* Social Problems, Vol. 44, No. 2.

Muhir, F. B., Lin, L. S., Stueve, A., Miller R. L., Ford, W. L, Johnson W. D., and Smith P. J. 2001. *A Venue-Based Method for Sampling Hard-to-Reach Populations.* Public Health Reports 116(Supp. 1):216-222.

Watts, D.J. and Strogatz, S.H. 1998. *Collective dynamics of 'small-world' networks.* Nature 393 (6684): 40910.

MERGING DATA FROM ANONYMOUS AND OPEN SURVEYS: TWO-POPULATION PROBLEMS

Artem Shcherbina[1] and Rostyslav Maiboroda[2]

[1] Kyiv University, Ukraine
e-mail: artshcherbina@gmail.com

[2] Kyiv University, Ukraine
e-mail: mre@univ.kiev.ua

1 Introduction

In the statistical analysis of survey data a problem of merging anonymous and personal surveys frequently occurs. In this paper it is considered that each element can belong to one of 2 sub–populations (classes). Also, all the elements are divided into several groups. Results of an anonymous survey provide information about proportions of sub–populations in different groups of elements. Personal survey is a sample of elements from different groups. During the personal survey some characteristic of interest Z is observed. An important remark is that classes of the observed elements are unknown. The aim of this two surveys is to determine some statistical properties of characteristic Z (such as distribution, mean value, variance) for each of two classes.
In this paper the problem of estimation of mean value for the elements of the first class is considered.

2 Model

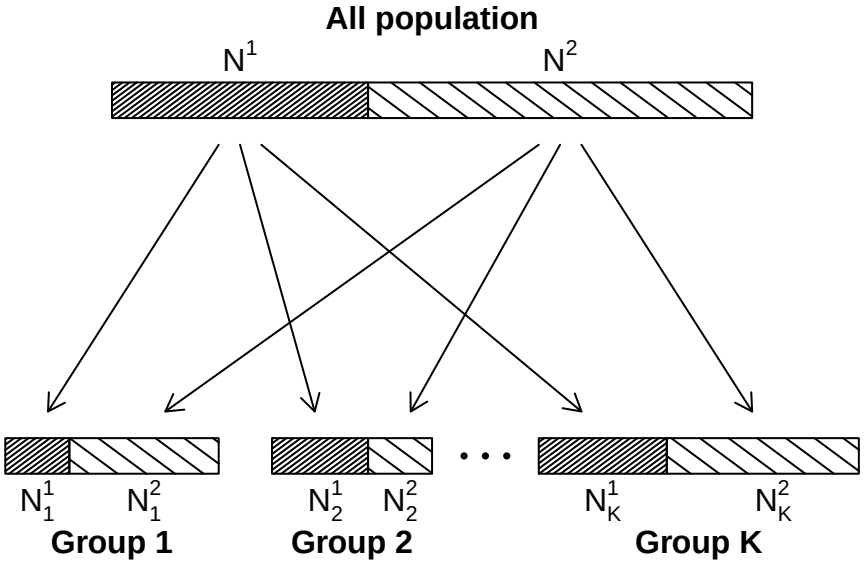
For analysis of such data finite mixture model will be used. In classical case the data is a set of independent random vectors (variables) $Z_1,\dots, Z_n$ and the distribution of Z_i is a mixture of M different probabilistic distributions:

$$\mathbf{P}\{Z_j \in A\} = p_1 H_1(A) + p_2 H_2(A) + \ldots + p_M H_M(A), \tag{1}$$

where H_m, $m = 1, \dots, M$ is the distribution of observed variables for the units from the m-th component of the mixture, p_i is the probability to observe a unit from the m-th component (the mixing probability), A is any measurable subset of the observations space.
Let us consider a case of two classes and suppose that each of them consists of a finite number of elements. Then distributions H_1 and H_2 will be discrete and, more important, different variables Z_i will be dependent.
Let the whole population consist of N elements $O_1, O_2, \dots, O_N$. Each of them can belong to classes $\mathcal{P}_1$ or $\mathcal{P}_2$. The sizes of the classes are N^1 and N^2 respectively. Also, all elements are divided into K groups $\mathcal{G}_1, \dots, \mathcal{G}_K$ with N_k units in the group $\mathcal{G}_k$. We suppose that exact proportions of the classes $\mathcal{P}_1$ and $\mathcal{P}_2$ in the groups are known. Let

the group $\mathcal{G}_i$ consists of N_i^1 elements of the first class and N_i^2 elements of the second class. This scheme is introduced in the next figure.

Figure 1. Scheme of population structure.



Let us Denote mean values for the elements from the i-th class by m_i, $i = 1, 2$. Similarly, let us denote variances for the elements from the i-th class by s_i^2, $i = 1, 2$. Consider a problem of estimation m_1 using a sample of the elements from the groups $\mathcal{G}_1, \ldots, \mathcal{G}_K$.

3 Estimation

Having data of an anonymous survey, we can obtain concentrations of the mixture model:

$$p_j^1 = \frac{N_j^1}{N_j},\ p_j^2 = \frac{N_j^2}{N_j}, \quad j = 1, 2, \ldots, K.$$

Thus, current model is

$$\mathbf{P}\{Z_j \in A\} = p_j^1 H_1(A) + p_j^2 H_2(A). \tag{2}$$

Let the personal survey consists of n_j elements from the group $\mathcal{G}_j$, $1 \le n_j \le N_j$. Obtained values of characteristic of interest are Z_j^i, $i = 1, 2, \ldots, K$, $j = 1, 2, \ldots, n_i$. Since the sample is taken without replacement, values Z_j^i are dependent. For estimation we will use the following weighted sum

$$\hat{m}_1 = \frac{1}{K} \sum_{i=1}^{K} a_i \frac{1}{n_i} \sum_{j=1}^{n_i} Z_j^i.$$

It is unbiased if and only if coefficients a_i satisfy equations

$$\frac{1}{K} \sum_{i=1}^{K} a_i p_i^1 = 1, \quad \frac{1}{K} \sum_{i=1}^{K} a_i p_i^2 = 0.$$

This equations define many possible sets of the coefficients a_i. In Maiboroda(1999) it is suggested to use minimax coefficients. They satisfy this equations and minimize sum of squares

$$\sum_{i=1}^{K} a_i^2. \tag{3}$$

It can be shown, that in our case minimax coefficients are

$$a_i = \frac{(1-q_1)p_i^1 + q_2 - q_1}{q_2 - q_1^2}, \tag{4}$$

where q_1 and q_2 are first two moments of r_i^1

$$q_1 = \frac{1}{K}\sum_{i=1}^{K} p_i^1, \quad q_2 = \frac{1}{K}\sum_{i=1}^{K} \left(p_i^1\right)^2.$$

It can be shown that estimate $\hat{m}_1$ with coefficients (4) has variance

$$\mathbf{D}\,\hat{m}_1 = \frac{1}{K^2}\sum_{i=1}^{K} a_i^2 \left[\frac{1}{n_i}\left(p_i^1 s_1^2 + p_i^2 s_2^2 + p_i^1 p_i^2 \frac{N_i}{N_i - 1}(m_1 - m_2)^2\right) - \frac{p_i^1 p_i^2}{N_i - 1}(m_1 - m_2)^2\right]$$

Under some assumptions estimate $\hat{m}_1$ will be consistent and asymptotically normal.

4 Discussion

Minimax coefficients, considered in this work, minimise the sum (3). However, last formula shows, that the variance of the estimate $\hat{m}_1$ has the form

$$\mathbf{D}\,\hat{m}_1 = \frac{1}{K^2}\sum_{i=1}^{K} a_i^2 \left(\frac{\alpha_i}{n_i} + \beta_i\right).$$

Thus, to obtain the most precise estimates, we need to minimize this expression. Unfortunately, parameters α_i and β_i contain unknown parameters s_1^2, s_2^2 and $(m_1 - m_2)^2$.

References

R. E. Maiboroda, Estimators of components of a mixture with varying concentrations, Ukrain. Mat. Zh. 48 (1996), no. 4, 562-566; English transl. in Ukrainian Math. J. 48 (1997), no. 4, 618-622

McLachlan, G.J., Peel, D.(2000).*FiniteMixtureModels.*Wiley,New York.

Titterington, D.M., Smith, A.F., Makov, O.E.(1985). *Analysis of Finite Mixture Distributions.* Wiley, New York.

THE EFFECT OF MODEL CHOICE IN ESTIMATION FOR DOMAINS

Milda Šličkutė-Šeštokienė[1]

[1] Statistics Lithuania, Lithuania
e-mail: milda.slickute-sestokiene@stat.gov.lt

Abstract

The effect of model choice on different types of estimators for domain totals is examined in this paper. For a given estimator type we derive different estimators, depending on the choice of model and compare them. In this paper Synthetic and General Regression estimators are discussed. We show how the choice of model effects different types of estimators for domains for Stratified Simple Random Sampling when domains do not correspond to stratum. The simulation study is accomplished on real data from Lithuanian Quarterly Survey on Earnings.

1 Introduction

During the last years in Statistics Lithuania as well as in most national statistical offices there is a contradictory tasks that need to be solved: on the one hand there is a huge need to diminish burden for respondents and to spare the costs for production of official statistics, but on another hand the volume of statistics required is constantly increasing. The only possibility to achieve those contradictory goals is to apply methods of domain estimation including small domains.

A number of estimators for domains, including small domains, have been proposed in recent literature. In this paper only two of them are analyzed: model-assisted Generalized Regression estimator (GREG) and model-dependent Synthetic estimator (SYN). The effect of model choice for those estimators for Stratified Simple Random Sampling are analyzed when domains do not correspond to strata.

For every specified model we derive one GREG estimator and one SYN estimator. We analyze how the accuracy of those estimators differ for the same model and also we analyze how the accuracy of each estimator differ applying different models. Table 1 shows the estimators to be discussed. Population model is one having its parameters defined at the population level, and domain model - at least some of its parameters defined at domain level.

Table 1

GREG and SYN Estimators by Model Choice and Estimator Type

Model Choice		Estimator type	
		Model-depended Synthetic	Model-assisted Generalized Regression
Population model	Linear	SYN-P	GREG-P
	Logistic	LSYN-P	LGREG-P
Domain model	Linear	SYN-D	GREG-D
	Logistic	LSYN-D	LGREG-D

2 Estimators and models

An objective of this paper is to analyze two main aspects of domain estimation: estimator type and model choice. The effect of model choice is analyzed for selected estimators: Generalized Regression (GREG) and Synthetic (SYN). GREG estimator is unbiased and SYN estimator is biased, but the variance of GREG estimator can be very large compared to SYN estimator, especially for small domains.

Consider a finite population $U = \{1, 2, ..., k, ...N\}$ from which a probability sample s ($s \subseteq U$) is drawn with a given sampling design, $p(\cdot)$. That is, $p(s)$ is the probability that s is selected. Consider a mutually exhaustive subgroups of the population U denoted $U_1, U_2, ..., U_d, ..., U_D$ - domains of interest, such that $\cup_{d=1}^{D} U_d = U$. The inclusion probabilities π_k=Pr($k \in s$) and π_{kl}=Pr(k & $l \in s$) are assumed to be strictly positive. Let y_k be the value of the variable of interest, y, for the kth population element. Let $\mathbf{x_k} = (x_{k1}, x_{k2}, ..., x_{kj}, ..., x_{kJ})'$ be the auxiliary variable vector of dimension $J \geq 1$, the value $\mathbf{x_k}$ is assumed to be known for every element $k \in U$. The target parameters are the set of domain totals, $Y_d = \sum_{U_d} y_k$, $d = 1, ..., D$

For a given model specification, the estimator of domain total $Y_d = \sum_U y_k$ has the following structure:

$$\hat{Y}_d^{GREG} = \sum_{U_d} \hat{y}_k + \sum_{s_d} a_k (y_k - \hat{y}_k) \tag{2.1}$$

$$\hat{Y}_d^{SYN} = \sum_{U_d} \hat{y}_k \tag{2.2}$$

where $a_k = 1/\pi_k$, $s_d = s \cap U_d$ is the part of sample s that falls in U_d, $d = 1, ..., D$ and the predictions $\{ \hat{y}_k; k \in U \}$ depends on the designated model and vector of auxiliary information $\mathbf{x_k}$.

Fixed-effect, linear and logistic, models are considered, such that $E_m(y_k) = f(\mathbf{x_k}; \beta)$, where β is an unknown parameter vector requiring estimation, and E_m refers to the expectation under the model. The model fit yields the estimate $\hat{\beta}$. The predicted values $\hat{y}_k = f(\mathbf{x_k}; \hat{\beta}_k)$ can be computed for $\forall k \in U$. Those predictions $\{\ \hat{y}_k; k \in U\}$ depends on the model and have the following structure:

$$\text{Population model:} \qquad \hat{\beta} = (\textstyle\sum_s a_k \mathbf{x_k} \mathbf{x_k'})^{-1} \sum_s a_k \mathbf{x_k} y_k \qquad (2.3)$$

$$\text{Domain model:} \qquad \hat{\beta}_d = (\textstyle\sum_{s_d} a_k \mathbf{x_k} \mathbf{x_k'})^{-1} \sum_{s_d} a_k \mathbf{x_k} y_k \qquad (2.4)$$

Our objective is to measure the effect of model for a given estimator in the case of Stratified Simple Random Sampling. For this purpose we compare coefficient of variation, bias and mean square error of each estimator using different models.

3 Simulation study

The data for simulation study came from the Lithuanian Quarterly Survey on Earnings. The population of enterprizes in this survey equals to about 50000 and sample size equals to about 7000. To generate realistic population of enterprizes, where target parameters are known, we combined total 8 quarters (2 years: 2007 and 2008) of responded enterprizes. In order to keep realistic distribution of enterprizes by economic activity, size and region we randomly dropped some enterprizes in some domains (where population size occurred too large compared to realistic), there were also some domains where population size was too small, we replicated the respondents in those domains by Simple Random Sampling With Replacement until a realistic population size was reached.

500 samples were selected from artificial population (this population based on real data) using Stratified Simple Random Sampling Without Replacement. In order to keep realistic sampling design stratification criterions were as following: form of ownership (public or private sector), economic activity (by NACE Rev. 1 at two digit level) and size of enterprize (determined by the number of employees: 1-9, 10-49, 50-99, 100-249, 250-499, 500-999, 999 and >). Domains of interest are regions based on NUTS4 classification (European Union's Nomenclature of Territorial Units for Statistics), domain sample size is random. In Lithuania we have 60 regions. Variables of interest are: number of employees converted in full-time units, number of part-time employees and gross earnings. For each sample we calculated the estimates, coefficients of variation, bias and mean square error.

Auxiliary variables are derived from register of Social Insurance, it is number of insured persons and taxable income. The coefficients of correlation between auxiliary variables and variables of interest are presented in Table 2.

Table 2

Coefficients of correlation between variables of interest and auxiliary variables

Auxiliary variables	**Variables of interest**		
	Gross Earnings	Total number of employees	Number of part-time employees
Number of insured persons	0.906	0.979	0.512
Taxable income	0.951	0.908	0.374

The results will be presented later.

4 Summary and discussions

Will be presented later.

References

EURAREA Consortium (2004). Project Reference Volume. Published by the EURAREA Consortium. URL: http://www.statistics.gov.uk/eurarea/.

Lehtonen R. and Pahkinen E. (2004). *Practical Methods for Design and Analysis of Complex Surveys*. John Wiley & Sons, Ltd.

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.

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

STATISTICAL ANALYSIS OF OBSERVATIONS WITH ADMIXTURE

Rostyslav Maiboroda[1] and Olena Sugakova[2]

[1] Kyiv University, Ukraine
e-mail: mre@univ.kiev.ua

[2] Kyiv University, Ukraine
e-mail: sugak@univ.kiev.ua

Abstract

Applications of mixture models in survey sampling are considered. Finite mixture model and model of mixtures with varying concentrations are discussed. Estimates based on minimax weighting scheme are proposed for estimation of means, variances and probability density functions of mixture components.

1 Introduction

Surveys data usually contain information on units belonging to different sub-populations. If the objective of a survey is to investigate some statistical characteristics of a prespecified sub-population then all units from the other sub-populations should be considered as a contamination (admixture). Ignoring contamination one obtains estimates which are, generally speaking, biased and inconsistent. The data with contamination can be considered as a mixture of different components. One of these components is of primary interest, while all others are admixtures.
In this paper we discuss applications of statistical mixture models to the analysis of such data.

2 Finite mixture model

In the classical finite mixture model (FMM) it is assumed that statistical data is a set of independent identically distributed random vectors (variables) $\xi_1,\ldots,\xi_n$ and the distribution of ξ_i is a mixture of M different probabilistic distributions:

$$\mathsf{P}\{\xi_j \in A\} = p_1 H_1(A) + p_2 H_2(A) + \ldots + p_M H_M(A), \tag{1}$$

where H_m, $m = 1,\ldots,M$ is the distribution of observed variables for units from the m-th component of the mixture, p_i is the probability to observe a unit from the m-th component (the mixing probability), A is any measurable subset of the observations space. If there exist pdfs h_m of the components distributions H_m the pdf f of ξ_j can be represented as

$$f(x) = p_1 h_1(x) + p_2 h_2(x) + \ldots + p_m h_m(x).$$

In statistical problems, where the model (1) is used, the distributions H_m and/or the mixing probabilities p_m are usually unknown. In fact, if no assumptions are made on

the distributions H_m then the FMM (1) is unidentifiable, i.e. it is impossible to construct a consistent estimate for H_m by the data even when the mixing probabilities are known. Therefore in the classical mixture analysis it is assumed that the distributions of the components belong to some parametric family of distributions. Then a parametric technique of estimation is used to fit obtained mixed distribution to the data. E.g. if all the components' distributions are Gaussian then the model (1) is identifiable and consistent estimates for the means, variances and mixing probabilities of the components can be constructed. See McLachlan and Peel (2000), Titterington et al. (1985).

This parametric modeling seems to be too restrictive for the purposes of surveys data analysis. Distributions of such data frequently have a complicated nature and can't be fitted by simple parametric models. So there is a need for nonparametric technique of mixture analysis applicable to surveys data.

Recently some nonparametric models were developed for different particular cases of FMM analysis. In Hall & Zhou (2003) a two-component mixture is considered, in which the observed data ξ_j are multivariate vectors with independent entries for both components. Consistent estimates for cdfs of the entries and mixing probabilities are constructed. Two-component mixtures of symmetric distributions were considered in Bordes et al. (2006), Hunter et al. (2007), Maiboroda (2007). Such models are more flexible than the parametric ones but they can be applied to mixtures with small number of components only and are based on some a priori assumptions as independence or symmetry.

3 Merging anonymous and non-anonymous surveys data

Let us consider a statistical problem which suggests some changes in the classical FMM model. Assume that two surveys were made: (1) anonymous and (2) non-anonymous. In the first survey there were N respondents (units) O_1, O_2,..., O_N. Each of them belongs to one of M sub-populations $\mathcal{P}_1$,...,$\mathcal{P}_M$. The units were divided into K groups $\mathcal{G}_1$,...,$\mathcal{G}_K$ with N_k units in the group $\mathcal{G}_k$ ($N_1 + N_2 + \ldots + N_K = N$). Anonymous surveys were conducted separately in each group. As a result we obtain the data on the numbers N_k^m of units belonging to the sub-population $\mathcal{P}_m$ in the group $\mathcal{G}_k$ for $k = 1, \ldots, K$, $m = 1, \ldots, M$.

The data of the second survey contain information on some sample O_{i_1}, O_{i_2},..., O_{i_n} from the respondents of the first survey. For each unit O in the sample some variable $\xi = \xi(O)$ is measured. Let $\xi_j = \xi(O_{i_j})$ be the value of this variable for the j-th unit in the sample. The sub-population to which O_{i_j} belongs is unknown but we know the group G_k in which O_{i_j} was surveyed anonymously. Let us denote the number of this group by k_j. The unobservable number of the sub-population $\mathcal{P}_m$ to which O_{i_j} belongs will be denoted by ν_j.

The aim of the analysis is to estimate such statistical characteristics as mean, variance, cdf, pdf for $\xi(O)$ over units O belonging to the sub-population $\mathcal{P}_m$. Let us denote the distribution of interest by

$$H_m(A) := \mathsf{P}\{\xi(O) \in A \mid O \in \mathcal{P}_m).$$

If O_{i_j} was chosen from $\mathcal{G}_{k_j}$ in random then

$$p_j^m := \mathsf{P}\{\nu_j = m\} = \frac{N_{k_j}^m}{N_{k_j}}$$

and the distribution of the observation ξ_j is

$$\mathsf{P}\{\xi_j \in A\} = p_j^1 H_1(A) + p_j^2 H_2(A) + \ldots + p_M H_M(A). \tag{2}$$

So we obtain a model analogical to (1), in which the mixing probabilities are different for different observations. This model is called the mixture with varying concentrations (MVC).

4 Example: Sociological analysis of elections results

As an example of problems discussed above we consider the analysis of sociologic characteristics distribution on sub-populations of different political parties adherents. Denote the variable of interest by ξ. It can be "satisfaction of life" or "importance of religion in the respondent's life" or any other variable which can be measured by answers of a respondent on some questionnaire. Suppose that ξ is measured in some scale, e.g. from 0 ("absolute dissatisfaction" or "unimportant") to 100 ("absolute satisfaction" or "highly important"). We are interested in distribution of ξ for voters which voted, say, "against all" at the recent parliament elections. Does this distribution differ from the distribution of ξ for persons who voted for the party which won the elections? Adherents of what parties have higher satisfaction of life or less interest to religion?
To answer these questions one can perform a survey on a sample of voters and obtain the values ξ_1, ξ_2,..., ξ_n of the variable ξ for the respondents. But it is not politically correct to ask a respondent for whom he/she voted at the resent elections. To estimate political preferences of respondents the results of elections averaged by election districts can be used. They can be considered as results of an anonymous survey.
So, all the population of voters is divided into sub-populations of different electorial choices adherents: $\mathcal{P}_1$ contains the persons who voted for the first party in the ballot, $\mathcal{P}_2$ contains the voters for the second party... $\mathcal{P}_{M-1}$ consists of the persons who voted against all and the sub-population of voters who didn't vote at this elections is denoted by $\mathcal{P}_M$. Let H_m be the distribution of ξ for voters from $\mathcal{P}_m$. Then the distribution of ξ_j is described by (2), where p_j^m is the frequency of m-th electoral behavior at the electoral district where the j-th respondent voted.
If the number of respondents taken from any electoral district is small in comparison to the number of voters at this district then ξ_j can be considered as independent random variables.

5 Estimates based on MVT model

Let ξ_1,...,ξ_n be independent random variables with distributions defined by (2). The mixing probabilities p_j^m are known, the distributions of the components H_m are unknown.

If all H_m were equal ($H_1 = H_2 = \ldots = H_M$), they could be estimated by the empirical distribution $\hat{F}(A) = \frac{1}{n}\sum_{j=1}^{n} 1_{\{\xi_j \in A\}}$, where $1_{\{\xi_j \in A\}}$ is an indicator of the event $\{\xi_j \in A\}$: $1_{\{\xi_j \in A\}} = \begin{cases} 1 & \text{if } \xi_j \in A, \\ 0 & \text{if } \xi_j \notin A. \end{cases}$ Since the distributions of different components are different, $\hat{F}_n$ can't be a consistent estimate to them. We propose to use a weighted empirical distribution function

$$\hat{H}_n(A, a) = \frac{1}{n}\sum_{j=1}^{n} a_j 1_{\{\xi_j \in A\}}$$

to estimate H_m. Here the weights a_j should be taken to suppress influence of all other components then H_m and to derive an unbiased estimate for H_m.
The unbiasedness means that $\mathsf{E}\hat{H}_n(A, a) = H_m(A)$ for all possible A, H_1,...,H_M. This implies the following condition on the weight vector $a = (a_1, \ldots, a_n)$:

$$\frac{1}{n}\sum_{j=1}^{n} a_j p_j^l = 1_{\{l=m\}}. \tag{3}$$

This condition doesn't define the weights unambiguously. To derive the best weights vector which satisfies (3) we adopt the minimax approach. Define the quadratic risk of the estimate $\hat{H}(\cdot, a)$ as

$$R(a; H_1, \ldots, H_M, A) = \mathsf{E}(\hat{H}_n(A, a) - H_m(A))^2.$$

Then the assured (minimax) risk of the estimate is

$$R^*(a) = \sup_{A, H_1, \ldots, H_M} R(a; H_1, \ldots, H_M, A),$$

where the sup is evaluated over all possible H_l and A. It is readily seen that for unbiased estimates

$$R^*(a) = \frac{1}{n}\sum_{j=1}^{n} (a_j)^2.$$

The assured risk attains its minimal value under the condition (3) on the weights $a^m = (a_1^m, \ldots, a_n^m)$, where

$$a_j^m = \sum_{l=1}^{M} \bar{\gamma}_{ml} p_j^l, \tag{4}$$

$\Gamma^{-1} = (\bar{\gamma}_{ml})_{m,l=1}^{M}$ is the matrix inverse to $\Gamma = (\gamma_{ml})_{m,l=1}^{M}$, $\gamma_{ml} = \frac{1}{n}\sum_{j=1}^{n} p_j^m p_j^l$. The weights a^m are called the minimax weights for the estimation of H_m. They can be applied to derive consistent estimates for means, variances and other probabilistic characteristics of distributions H_m. E.g. an unbiased estimate for the mean

$$\mu_m = \mathsf{E}(\xi(O) \mid O \in \mathcal{P}_m) = \int x H_m(dx)$$

is

$$\hat{\mu}_m = \frac{1}{n}\sum_{j=1}^{n} a_j^m \xi_j = \int x \hat{H}_n(dx, a^m).$$

As an estimate for the variance

$$\sigma_m^2 = \mathsf{E}((\xi(O) - \mu_m)^2 \mid O \in \mathcal{P}_m) = \int (x - \mu_m)^2 H_m(dx),$$

one can use

$$\hat{\sigma}_m^2 = \frac{1}{n} \sum_{j=1}^{n} a_j^m (\xi_j - \hat{\mu}_m)^2.$$

Note that this estimate is consistent but not unbiased.
An estimate for the pdf h_m of H_m can be constructed as a weighted version of the well known kernel density estimate:

$$\hat{h}_m(x) = \frac{1}{\lambda n} \sum_{j=1}^{n} a_j^m K\left(\frac{x - \xi_j}{\lambda}\right),$$

where K is a kernel (i.e. a pdf), λ is a bandwidth.

6 Discussion

Proposed estimates are consistent and asymptotically normal under mild conditions. But they also have some unwanted features. For example, due to (3) the weights a_j^m always are negative for some j. So the estimates $\hat{\sigma}_k^2$ and $\hat{h}_m$ also can be negative. Possible improvements and alternatives to these estimates should be discussed.

References

Bordes L., Mottelet S., Vandekerkhove P. (2006). Semiparametric estimation of a two-component mixture model. *Ann. Statist.* **34**, 1204-1232.

Hall P., Zhou X.-H. (2003). Nonparametric estimation of component distributions in a multivariate mixture. *Ann.Statist.*, **31**, 201-224.

Hunter D.R., Wang S., Hettmansperger T.R. (2007). Inference for mixtures of symmetric distributions. *Ann. Statist.* **35**, 224-251.

Maiboroda, R. (2008). Estimation of locations and mixing probabilities by observations from two-component mixture of symmetric distributions. *Teorija Imovirnosti ta Matematychna Statystyka* **78**, 133-141 (in Ukrainian, engl. transl. in *Theor. Probab. and Math. Statist.*).

McLachlan, G.J., Peel, D. (2000). *Finite Mixture Models.* Wiley, New York.

Sugakova O.V. (1999). Asymptotics of a kernel estimate for distribution density constructed from observations of a mixture with varying concentrations. *Theor. Probab. and Math. Statist.* **59** 161-171.

Titterington, D.M., Smith, A.F., Makov, O.E. (1985). *Analysis of Finite Mixture Distributions.* Wiley, New York.

APPROACHES TO STATISTICAL MATCHING OF STATE SAMPLE SURVEYS DATA IN UKRAINE

Ganna Tereshchenko

Institute for Demography and Social Research
of the National Academy of Sciences of Ukraine
e-mail: a_tereschenko@ukr.net

Abstract

In this work the basic results of approaches to statistical matching of state sample household surveys data in Ukraine are submitted. The methods of merging of the sample household surveys data obtained from samples with different design and method of statistical matching of different sample household surveys data with use of harmonized indicators system are proposed.

State sample population surveys are the main source of data as to demographic, social and economic characteristics of population in between the population censuses in Ukraine. They are necessary for macroeconomic calculations, calculations of major indicators in agriculture, various analyses of standards of living, monitoring poverty, evaluation of labour market conditions according to international standards and also for data users of different levels. The problems of development of methodological maintenance of the Statistical matching sample surveys data and estimation of matching data are important first of all for the state statistics.

State sample population (household) surveys are performed in all regions of Ukraine at the permanent residence of population. There are two surveys in the cities: survey of households living conditions and survey of population economic activity. In the rural area there are three surveys: survey of households living conditions, survey of population economic activity and survey of agricultural activity.

Indicators of labour force in Ukraine are measured by the results of households sample survey of population economic activity (LFS). Since 1995 LFS is carried out by bodies of the state statistics. In LFS the population of 15–70 years of age is surveyed. Indicators of economic activity, employment and unemployment are measured in survey according to the international standards by ILO methodology.

Such social and economic characteristics of the population of Ukraine as a standard of living, incomes and expenditures are defined by the results of the state sample survey of households living conditions (HLCS). HLCS data is used for information maintenance the measurement of poverty, for definition of households' social and demographic characteristics during the periods between population censuses, etc. But, not corresponding to the ILO methodology. Since 1999 the survey is carried out on the quarterly basis. Annually in HLCS approximately 10 thousand non-institutional households are interviewed.

Survey of agricultural activity of the rural population (AAS) was introduced by the State Statistics Committee in September 2000. Around 30 000 households that own land are surveyed every year on AARP. The data from this survey is used in calculated gross amount of major plant and animal production, amount and price of agricultural products for sale, production of agricultural products (labour, service).

Since 2004 by the special decree of State Statistics Committee of Ukraine harmonization of some indicators, which are measured in all population state surveys, is ratified. These indicators are: number of household members; family relations of household members; date of the birth of each household member; sex and educational level of each household member. Harmonization of survey programs in particular also provides the

foundation for approbation and implementation in the state statistics the modern methodological bases of matching data from different sources, including on the microlevel.

The statistical matching methods are becoming widely used in official (state) statistics of developed countries of the world. There are two major directions of statistical matching that should be pointed out:

Matching of the data obtained from different sources of information;

Matching of the data from one source of information in time (the most common example of such matching is the creation of cumulative data of information based on results of surveys that are done periodically – monthly, quarterly or annually).

Methods of statistical matching from different sources have started to develop only recently, though separate works have appeared in 1970s. It is the result of the need of official statistics. Methods of matching of the second direction have used since 1960s. They are considered traditional in such fields of statistics as labour monitoring, poverty monitoring, etc.

Current methodological approaches to matching the data obtained from different sources are based on the researches of such experts as Rubin, D.B., Rässler S, D'Orazio M., Di Zio M., Scanu M., Van der Laan P.

We think that it is prudent to differentiate between directions of matching the data obtained from different sources. We suggest accepting as basic directions the vertical and horizontal data matching (see fig. 1). As you can see on the fig. 1 vertical data matching is used in cases when the data from different levels of aggregation needs to be linkage, e.g. matching regional level data with microlevel or national level data. In horizontal matching the data from the same aggregation level is matched, e.g. data from several regions, districts, etc. It should be noted that data matching can be conducted in both directions at the same time.

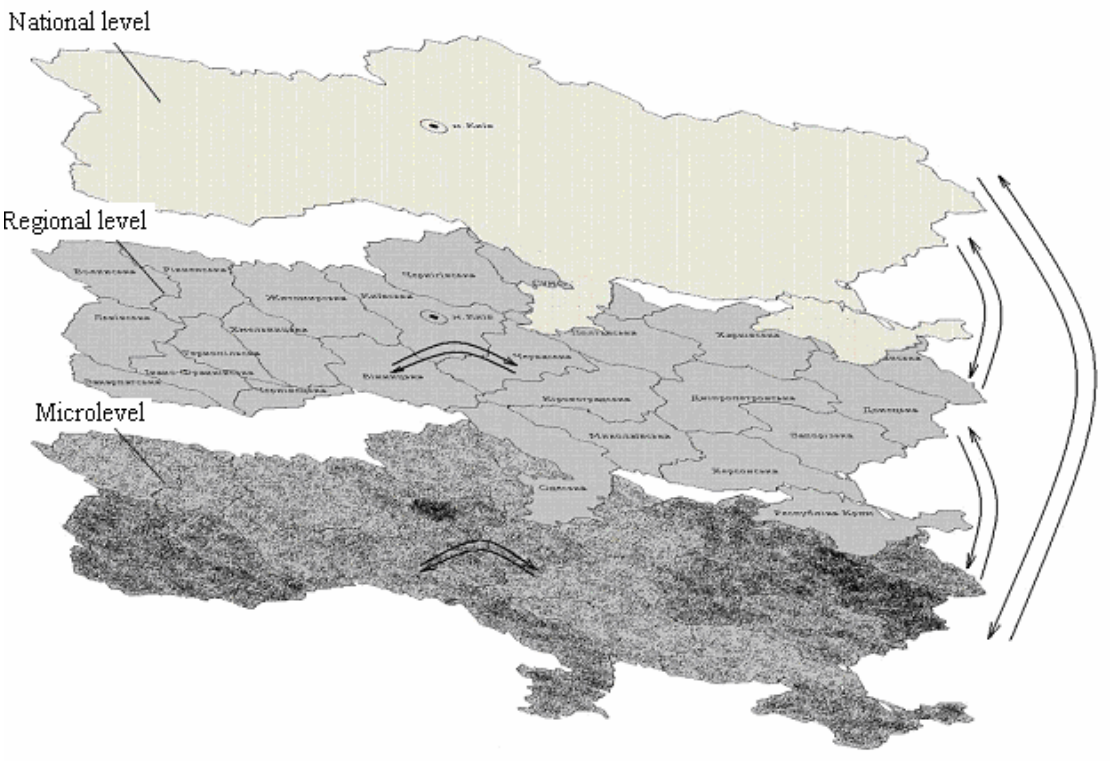


Fig. 1. Scheme of the use of information is from the different levels of aggregating

D'Orazio M., Di Zio M., Scanu M. in the works selects such basic methods of statistical matching of different sample surveys:

– data merging;
– record linkage;

– matching data (data fusion – name which is used in Europe).

Data merging is used in conditions when, from different sources the data is obtained in units of an identical level and by identical attributes. It is necessary to notice that at the use of method of data merging with the purpose of matching data of population sample surveys, as a rule, there is a serious methodological problem, related to the different levels of results reliability. It is predefined divergences in the designs of samples, by different quality of organization of surveys.

Record linkage is used, when for the same units get information from different sources that is to survey data it is necessary to add additional variables. For possibility of Record linkage it is necessary to foresee the special variables – keys which used for linkage records with units of survey.

These two methods are often used at households sample surveys.

Statistical matching is the procedure of combining information from different sample surveys that contain no common unit. Statistical matching was in detail learning by such scientists, as: Rässler S, D'Orazio M., Di Zio M., Scanu M. Statistical matching it is enough methodologically and technologically by a difficult method, bases an analysis and account of statistical properties of data.

Merging of LFS data, received on the basis of two samples with different design

To improve reliability of employment and unemployment estimates by regions of Ukraine in rural area since 2004 the survey is carried out on the of two probability stratified two stage samples: sample of LFS and sample of household agricultural activity survey (AAS). Interview of households on questions of economic activity in the AAS is carried out under the identical programmer, as in the LFS, but the sample design in AAS is differ from LFS. In AAS households are selected in the second stage with probability proportionally to their area of agricultural allotment, in LFS - on the basis of the procedure of systematic selection.

The size of monthly LFS sample in the rural area makes approximately 3,6 thousand households, and the size of AAS sample of households which have to be interviewed under LFS questionnaire is 7,4 thousand households. Total size of a monthly sample for interview under LFS questionnaire in the rural area due to AAS has increased three times and is equal to 11,1 thousand households.

Based on the results of the primary data analysis of LFS and AAS we can establish, that surveys data is mutually coordinated; in the February of 2007 in particular, the coefficient of correlation for employment rate is equal 0,79, and for unemployment rate – 0,77. The similar picture is observed throughout all months of 2007-2008.

Matching of LFS data, obtained on the basis of two different samples is carried out with the use method of data merging at the microlevel. This approach is used in conditions when, from different sources the data is obtained in units of an identical level (in this case – on household members aged 15–70) and by identical attributes (under identical questionnaires). The result of data merging is the file where we have observations from both - LFS file and AAS file.

The employed and unemployed population is calculated by formula for composite estimation:

$$\hat{Y}_{em} = \hat{\phi}_{em}\hat{Y}_{em}^{(LFS)} + (1-\hat{\phi}_{em})\hat{Y}_{em}^{(AAS)}$$
$$\hat{Y}_{un} = \hat{\phi}_{un}\hat{Y}_{un}^{(LFS)} + (1-\hat{\phi}_{un})\hat{Y}_{un}^{(AAS)} \qquad (1)$$

where – $\hat{Y}_{em}^{(LFS)}$ – estimate of employed population number on LFS sample; $\hat{Y}_{em}^{(AAS)}$ – estimate of employed population number on AAS sample; $\hat{Y}_{un}^{(LFS)}$ – estimate of unemployed

population number on LFS sample; $\hat{Y}_{un}^{(AAS)}$ – estimate of unemployed population number on AAS sample.

Based on the performed researches it has been established that direct estimates of the employed and unemployed population number, which are obtained on the basis of AAS sample are biased. The estimates of bias for the employed and unemployed population number for the current month in this work were defined as an average bias for current month and the previous two.

During file merging of AAS with LFS unit weights, calculated separately on each file are corrected for rural area on each region with the use of coefficients ϕ_{em} for employed population and ϕ_{un} – for unemployed (Sarioglo, 2005):

$$\hat{\phi}_{em} = \frac{SE^2(\hat{Y}_{em}^{(AAS)}) + \overline{B}^2(\hat{Y}_{em})}{SE^2(\hat{Y}_{em}^{(LFS)}) + SE^2(\hat{Y}_{em}^{(AAS)}) + \overline{B}^2(\hat{Y}_{em})} \quad for\ employed\ persons$$

$$\hat{\phi}_{un} = \frac{SE^2(\hat{Y}_{un}^{(AAS)}) + \overline{B}^2(\hat{Y}_{un})}{SE^2(\hat{Y}_{un}^{(LFS)}) + SE^2(\hat{Y}_{un}^{(AAS)}) + \overline{B}^2(\hat{Y}_{un})} \quad for\ unemployed\ persons \qquad (2)$$

where $SE(\hat{Y}_{em}^{(LFS)})$ – standard error of estimate of employed population number $\hat{Y}_{em}^{(LFS)}$ on LFS sample; $SE(\hat{Y}_{em}^{(AAS)})$ – standard error of estimate of employed population number $\hat{Y}_{em}^{(AAS)}$ on AAS sample; $SE(\hat{Y}_{un}^{(LFS)})$ – standard error of estimate of unemployed population number $\hat{Y}_{un}^{(LFS)}$ on LFS sample; $SE(\hat{Y}_{un}^{(AAS)})$ – standard error of estimate of unemployed population number $\hat{Y}_{un}^{(AAS)}$ on AAS sample; $\overline{B}(\hat{Y}_{em})$ – the bias of estimate of number of employed population; $\overline{B}(\hat{Y}_{un})$ – the bias of estimate of number of unemployed population.

The corrected statistical weights of employed and unemployed persons in rural area of each region are calculated by the formula:

$$w_i' = w_i \cdot k_i \qquad (3)$$

Correction coefficient of statistical weights system k_i is calculated for each region in two stages. On the first stage the value of k_i is calculated for employed and unemployed persons in rural area:

$$k_i = \begin{cases} \hat{\phi}_{em} & for\ employed\ person\ on\ LFS\ sample \\ \hat{\phi}_{un} & for\ unemployed\ person\ on\ LFS\ sample \\ (1-\hat{\phi}_{em}) & for\ employed\ person\ on\ AAS\ sample \\ (1-\hat{\phi}_{un}) & for\ unemployed\ person\ on\ AAS\ sample \end{cases} \qquad (4)$$

On the second stage the value of k_i is calculated for economically inactive persons in rural area for each regionIf to compare reliability characteristics of estimates of employment and unemployment indicators for regions of Ukraine on the matched data file and LFS data file, it is necessary to draw the conclusion that data matching allows to improve essentially the reliability characteristics of indicators estimates in rural area.

If we want to compare reliability characteristics of employment and unemployment rates estimates for regions of Ukraine on the matched data file and LFS data file, it is necessary to draw the conclusion that data matching allows essential improvement in the reliability characteristics of employment and unemployment rates estimates in rural area. It

should also be noted that for 8 regions CV of unemployment rate estimates has decreased twice and for 3 regions - three times (fig. 2).

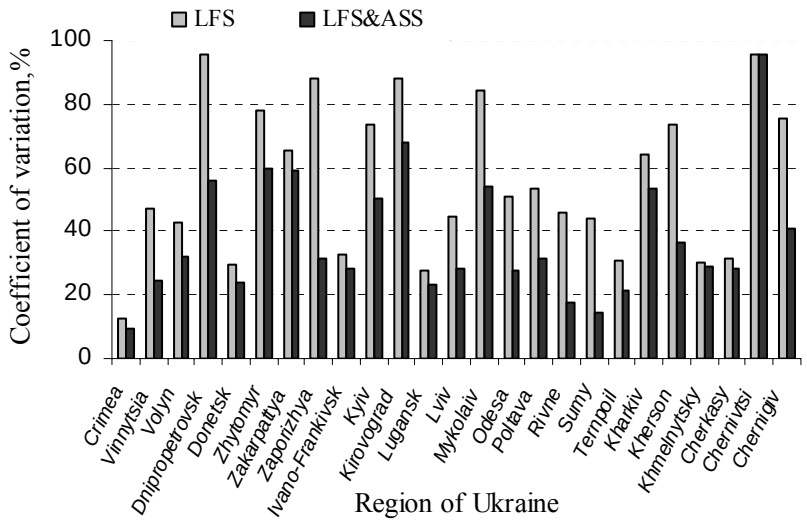


Fig. 2. Reliability of unemployment rate monthly estimates in rural area before and after statistical matching of the LFS data, February, 2007

Matching of HLCS and LFS data with use of harmonized indicators system, for the analysis of unemployment on ILO methodology and the level of expenditures of the unemployed population

In the HLCS social and economic status of persons is measured, and, particularly, the indicator of unemployment. But, not corresponding to the ILO methodology.

Model of interconnection of unemployment rate and certain person's characteristics was constructed on the basis of LFS data. Following variables of LFS microdata file were used for this model constructing:

"Unemployed population". This variable accepted value 1 if person is unemployed, 0 – if one isn't.

As factorial variables the following dummy variables were used:

WOMEN – sex of the person is female (1 – yes, 0 - no);

ED_HIGH – the person has higher or base higher education (1 – yes, 0 - no);

V15_19 - the age of the person is within 15-19 years (1 – yes, 0 - no);

V20_24 - the age of the person is within 20-24 years (1 – yes, 0 - no);

V25_29 - the age of the person is within 25-29 years (1 – yes, 0 - no);

V30_34 - the age of the person is within 30-34 years (1 – yes, 0 - no);

V35_39 - the age of the person is within 35-39 years (1 – yes, 0 - no);

V40_44 - the age of the person is within 40-44 years (1 – yes, 0 - no);

V45_60 - the age of the person is within 45-49 years (1 – yes, 0 - no);

V61_70 - the age of the person is within 61-70 years (1 – yes, 0 - no).

Besides these variables, which have come into final models by results of executed research, many other binary variables were considered, such as family condition, citizenship and so forth, but their influence on unemployed share appeared rather insignificant.

The constructed model has following view:

$$UN = 0.04 - 0.008 \cdot WOMEN - 0.014 \cdot ED\_HIGH + 0.011 \cdot V15\_19 + 0.063 \cdot V20\_24 +$$
$$+ 0.044 \cdot V25\_29 + 0.038 \cdot V30\_34 + 0.041 \cdot V35\_39 + 0.037 \cdot V40\_44$$

Quality characteristics of model:

$R^2 = 0.821;\ F = 26.786;(F_{(0.95)} = 2.24).$

On the basis of given model it was drown the conclusion, that share of unemployed is interconnected with such variables as person's sex, age and educational level.

Model of interconnection of unemployed population shares in the LFS and in the HLCS was also constructed:

$UN_{LFS} = 4.08 + 0.23 \cdot UN_{HLCS}$

Quality characteristics of model:

$R^2 = 0.562;\ F = 32.069;(F_{кр(0.95)} = 3.35).$

On the basis of given model it is possible to drown the conclusion, that shares of unemployed in both surveys are mutually statistically connected.

For reproduction of variable "unemployed by ILO" in HLCS data file the procedure of statistical data matching on microlevel is used (Marcello D'Orazio, Marco Di Zio, Mauro Scanu, 2006).

Reproduction of variable "unemployed by ILO" in HLCS data file was carried out with use of the imputation procedure by "hot deck" method in view of unemployed from HLCS. As in survey the population is interviewed only in the age of 15 – 70 years and in the HLCS variable "unemployed by ILO" is reproduced not for all persons who have taken part in survey but only for persons in the age of 15 – 70 years. The essence of this method consists in selection of the person as the donor from LFS file and use of corresponding attribute of this person for reproduction of "unemployed" value in HLCS file. The person-donor was selected from the condition of similarity with that person for whom values are selected, to such attributes: economic rayon, sex, educational level, age group. At indicators estimation in Ukraine in some cases it is expedient to consider territories bigger, than regions. As such large areas can be considered, for example, economic rayons – groups of regions, which integrated by proximity of such characteristics as production of industrial output, concentration of industrial potential and labor-power.

For use of "hot deck" method cells system is constructed in HLCS and LFS files to such classification attributes:

- economic rayon, 8 discrete values (East, Donetsky, Pridnyprovsky, Prichornomorsky, Podilsky, Central, Karpatsky, Polisky economic rayons);
- sex of the person, 2 discrete values (female, male);
- educational level of person, 2 discrete values (higher, no higher education);
- age group of person, 8 discrete values (15-19 years, 20-24 years, 25-29 years, 30-34 years, 35-39 years, 40-44 years, 45-60 years, 61-70 years).

Cod of cell is calculated by the formula:

Cod = attribute of economic rayon ·1000 + attribute of sex·100 + attribute of educational level·10 + attribute of age group

During construction of system of cells some collapsing of cells is needed where cells not contain unemployed persons.

For realization of imputation procedure the analysis of various indicators distribution on the generated cells has been carried out. On fig. 3 the data for the share of the unemployed estimated by results of LFS and HLFS, and also values of average total per capita expenditures, estimated on the HLFS data for the level of Ukraine are resulted. As it is clear from the shown data on cells there are essential distinctions in values of the unemployed shares on LFS and HLFS. Also distinctions in values of cumulative expenditures are observed. The procedure of the data matching, which is developed by results of researches, allows reflecting full enough the pointed distinctions on cells at the data reproduction.

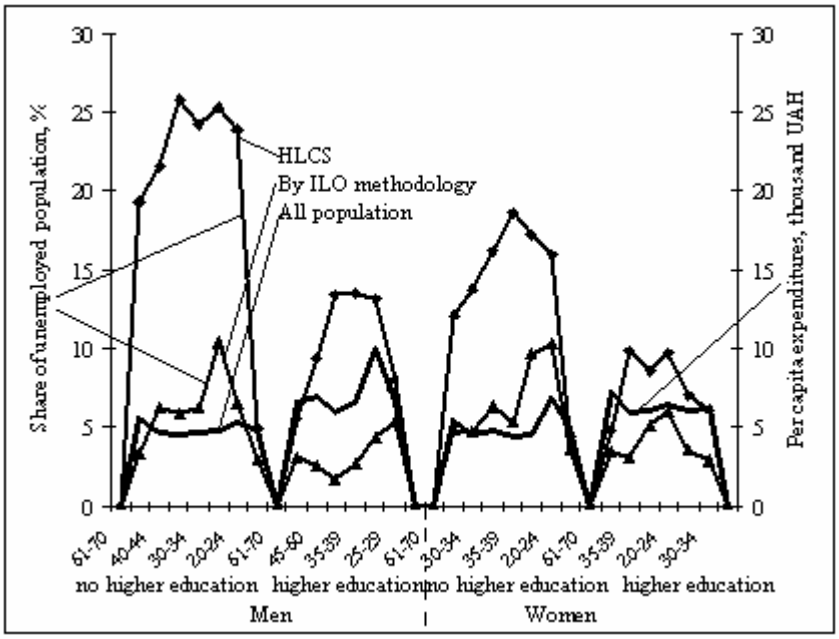


Fig. 3. Share of unemployed population on groups, 2005

By results of the data matching of HLCS and LFS surveys .distribution of population including the unemployed by the ILO methodology, on the annual total per capita expenditures (see fig. 4) is constructed. As it is clear from the shown data, this distribution for the unemployed essentially differs from distribution for all population at able-bodied age. The value of average total expenditures for the unemployed is equal 4,7 thousand UAH per capita in one year, for all population – 5,8 thousand UAH per capita in one year.

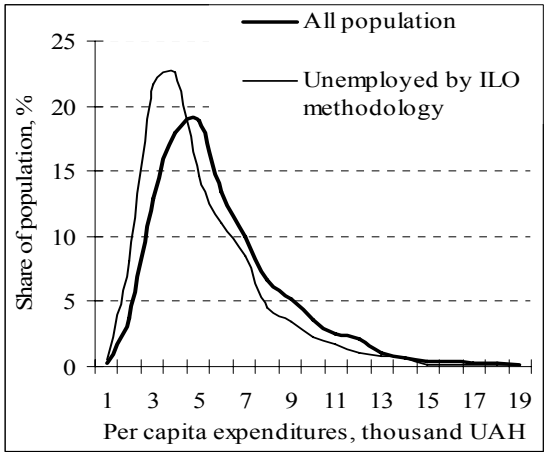


Fig. 4. Distribution of population including the unemployed by the ILO methodology, on the annual total per capita expenditures, 2005

Still bigger distinctions are observed in distributions of expenditures of the unemployed and employed population. It is obvious, that reproduction of the status of economic activity on ILO methodology in HLCS file opens ample opportunities for the analysis of various social and economic processes. However consideration of these questions is beyond this work.

Merging of the labour force survey data, obtained on samples with different design has allowed improving the reliability level of employment and unemployment indicators estimation in rural area of Ukraine. The methodological problem of adequate determination

of indicators estimates bias value is revealed; it is connected with differences of populations and with features of the separate surveys organization.

The methodological approach to matching on microlevel the separate attributes of labour force and household living conditions sample surveys is offered with use of harmonized indicators system. Procedure is worked at the data matching concerning status of household members' employment. It has allowed establishing and analyzing distribution of unemployed, defined by ILO methodology, on the level of average cumulative expenditures per capita.

References

Rässler S., Fleischer K. (1998) Aspects concerning data fusion techniques. ZUMA Nachrichten Spezial, № 4, 317—333.

Rässler S. (2003) How accurate can data fusion be? Bulletin of the International Statistician Institute 54-th Session, Aug. 13—20, 2003. Berlin, Vol. LX, B. 1, 372—375.

D'Orazio M., Di Zio M., Scanu M. (2006) Statistical Matching: Theory and Practice. Chichester: John Wiley&Sons, 256 p.

D'Orazio M., Di Zio M., Scanu M. (2002) Statistical Matching and Official Statistics. Rivista di Statistica Ufficiale, 1, 5—24.

Sarioglo V.G. (2006) Problems of the sample data statistical weighting. Kyiv: State Statistics Committee of Ukraine, 264 p.

MEASUREMENT ERRORS IN SURVEY DATA

Maria Valaste[1]

[1] Maria Valaste, University of Helsinki, Finland
e-mail: maria.valaste@helsinki.fi

Abstract

In sample surveys, the uncertainty of parameter estimates comes from two main sources: sampling and measuring the study units. When estimating the parameters from survey data, it is important to have control over the sources of uncertainty in the estimation procedure. Often the data is collected by a complex sampling design involving stratification, clustering and unequal inclusion probabilities. The first source of error then comes from the implementation of the complex sampling design and generalizing the results to the population. Another source of error is present when measuring the study units. When assessing the quality of the collected and measured data set, we end up with the following questions: Are we measuring the right thing? How accurate our measurements are? The former question leads us to the concept of validity which is the most important property of the quality of measurement. The latter question is related to the concept of reliability.

This paper focuses on measurement errors in sample surveys. A lot of the literature on overall measuring originates in classical test theory from psychology. Despite of the existence of the measurement errors in surveys, little attention is paid to their effect in survey design and analysis. A literature review of the central concepts and of development in measurement studies is provided. Also a plan for further research including simulation studies is introduced and discussed.

Participants

Name		Organization	e-mail
Babrova	Nastassia	Economic Institute of National Academy of science, Belarus	nastassiabobrova@mail.ru
Bartkus	Ignas	Vilnius Pedagogical University, Lithuania	ignas.bartkus@gmail.com
Bokun	Natalia	Belarussian State Economis University	nataliabokun@rambler.ru
Bondarenko	Yana	Dnipropetrovsk National University, Ukraine	yanabondarenko@ua.fm
Budkina	Natalja	University of Latvia	budkinanat@gmail.com
Bülow	Erik	University of Gothenburg, Sweden	bulow@student.chalmers.se
Chadyšas	Viktoras	Vilnius Gediminas Technical University, Lithuania	Viktoras.chadysas@fm.vgtu.lt
Chernyak	Oleksandr	Taras Shevchenko National University of Kyiv, Ukraine	chernyak@univ.kiev.ua
Chystsenka	Katsiaryna	National Statistics Committee of the Republic of Belarus	chistenko@gmail.com
Čiginas	Andrius	Vilnius University, Lithuania	andrius.ciginas@mif.vu.lt
Dariychuk	Illya	Taras Shevchenko National University of Kyiv, Ukraine	elijadar@rambler.ru
Fedorianych	Tetiana	Uzhhorod National University, Ukraine	fedoryanicht@gmail.com
Fisenko	Andris	Central Statistic Bureau of Latvia	andris.fisenko@csb.gov.lv
Hindrikson	Merike	University of Tartu, Estonia	mercs@ut.ee
Jukams	Janis	Central Statistical Bureau of Latvia	janis.jukams@csb.gov.lv
Kemzūraitė	Edita	Vilnius Gediminas Technical University (VGTU), Lithuania	edita.kemzuraite@gmail.com
Kolosov	Alexander	Donetsk National University, Ukraine	kolosov@dongu.donetsk.ua
Krapavickaitė	Danutė	Institute of Mathematics and Informatics, Gediminas Technical University, Statistics Lithuania	krapav@ktl.mii.lt
Kulldorff	Gunnar	University of Umeå, Sweden	gunnar@matstat.umu.se
Lapiņš	Jānis	Bank of Latvia	Janis.Lapins@bank.lv, lapinsj@inbox.lv
Lehtonen	Risto	University of Helsinki, Finland	risto.lehtonen@helsinki.fi
Lepik	Natalja	University of Tartu, Estonia	natalja.lepik@ut.ee
Liberts	Mārtiņš	Central Statistical Bureau of Latvia	Martins.Liberts@csb.gov.lv, Martins.Liberts@gmail.com
Lumiste	Kaur	Tartu University, Estonia	Kaur.Lumiste@gmail.com
Lysa	Olha	Institute for Demography and Social Research of the National Academy of Sciences of Ukraine	olysa@ukr.net
Magnusson	Måns	Swedish National Council for Crime Prevention, Sweden	mons.magnusson@gmail.com

Name		Organization	e-mail
Manzhos	Tetyana	Vadim Getman Kyiv National Economic University, Ukraine	tmanzhos@gmail.com
Masiulaitytė	Inga	Statistics Lithuania, Vilnius University	inga.masiulaityte@stat.gov.lt
Mishura	Yulia	National Taras Shevchenko University of Kyiv, Ukraine	myus@univ.kiev.ua
Orlova	Julia	Belarussian State Economic University	55xx@mail.ru
Osypchuk	Mykhaylo	Precarpatian university, Ukraine	myosyp@ukr.net
Pettersson	Nicklas	Stockholm University, Sweden	Nicklas.pettersson@stat.su.se
Plikusas	Aleksandras	Institute of Mathematics and Informatics, Vilnius University, Lithuania	Plikusas@ktl.mii.lt
Pumputis	Dalius	Vilnius Pedagogical University, Lithuania	dpumputis@vpu.lt, dpumputis@yahoo.co.uk
Rozora	Iryna	Taras Shevchenko National University of Kyiv, Ukraine	irozora@bigmir.net, rozorchik@ukr.net
Rozora	Natalia	Nielsen Ukraine	Natalia.Rozora@nielsen.com
Rychka	Svitlana	Bohdan Khmelnitsky University of Cherkasy, Ukraine	Rychka-svetlana@ukr.net
Rässler	Susanne	University of Bamberg, Germany	susanne.raessler@uni-bamberg.de
Sarioglo	Volodymyr	Institute for Demography and Social Research of the National Academy of Sciences of Ukraine	sarioglo@idss.org.ua
Semenovs'ka	Nataliya	Taras Shevchenko National University of Kyiv, Ukraine.	semenovsky@hotmail.ru
Shafie	Termeh	Stockholm University, Sweden	termeh.shafie@stat.su.se
Shcherbina	Artem	Taras Shevchenko National University of Kyiv, Ukraine	artshcherbina@gmail.com
Šličkutė-Šeštokienė	Milda	Statistics Lithuania	milda.slickute-sestokiene@stat.gov.lt
Sugakova	Olena	Taras Shevchenko National University of Kyiv, Ukraine	sugak@univ.kiev.ua
Tereshchenko	Ganna	Institute for Demography and Social Research of the National Academy of Sciences of Ukraine	a_tereschenko@ukr.net
Thorburn	Daniel	Stockholm University, Sweden	Daniel.Thorburn@stat.su.se
Traat	Imbi	University of Tartu, Estonia	Imbi.traat@ut.ee
Valaste	Maria	University of Helsinki, Finland	maria.valaste@helsinki.fi
Vasylyk	Olga	Taras Shevchenko National University of Kyiv, Ukraine	ovasylyk@univ.kiev.ua
Vlasenko	Nataliya	State Statistics Committee of Ukraine	vlasenko@ukrstat.gov.ua
Yakovenko	Tetyana	Taras Shevchenko National University of Kyiv, Ukraine	yata452@univ.kiev.ua