TARAS SHEVCHENKO NATIONAL UNIVERSITY OF KYIV

PROCEEDINGS OF BALTIC-NORDIC-UKRAINIAN SUMMER SCHOOL ON SURVEY STATISTICS

AUGUST 22 - 26, 2016

KYIV, UKRAINE

Proceedings of the Baltic-Nordic-Ukrainian Summer School on Survey Statistics. – Kyiv, 2016. – 80 p.

Organizing institutions

University of Tartu, Estonia
University of Helsinki, Finland
University of Latvia
Vilnius University, Lithuania
Vilnius Gediminas Technical University, Lithuania
Taras Shevchenko National University of Kyiv, Ukraine
Institute for Demography and Social Research, Ukraine
Statistics Estonia
Central Statistical Bureau of Latvia
Statistics Lithuania
State Statistics Service of Ukraine

Organizing committee

Mykhaylo Gorodnii (Taras Shevchenko National University of Kyiv, Dean of the Faculty of Mechanics and Mathematics, Chair)
Yuliya Mishura (Taras Shevchenko National University of Kyiv, Co-chair)
Tetiana Ianevych (Taras Shevchenko National University of Kyiv, Vice-chair)
Vitalii Golomozyi (Taras Shevchenko National University of Kyiv)
Danute Krapavickaite (Vilnius Gediminas Technical University)
Risto Lehtonen (University of Helsinki)
Iryna Rozora (Taras Shevchenko National University of Kyiv)
Volodymyr Sarioglo (Institute for Demography and Social Research)
Olga Vasylyk (Taras Shevchenko National University of Kyiv, Secretariat)

Programme committee

Risto Lehtonen (University of Helsinki, Chair)
Natallia Bokun (Belarus State Economic University)
Tetiana Ianevych (Taras Shevchenko National University of Kyiv)
Danute Krapavickaite (Vilnius Gediminas Technical University)
Thomas Laitila (Örebro University)
Janis Lapins (Bank of Latvia)
Natalja Lepik (University of Tartu)
Martins Liberts (Statistics Latvia)
Imbi Traat (University of Tartu)
Olga Vasylyk (Taras Shevchenko National University of Kyiv)

Sponsors

International Association of Survey Statisticians (IASS)
The organizing institutions

July 2016

PREFACE

Dear participants!

We congratulate you on the 20th anniversary event of the Baltic-Nordic-Ukrainian Network on Survey Statistics!

This Summer School is the second event within the Network hosted by Ukraine. Co-operation between Baltic and Nordic countries in the field of survey statistics began in 1992 — between universities as well as statistical agencies. A Baltic-Nordic network for co-operation on education and research in survey statistics has grown continuously since 1996 to include six partner universities in Estonia, Finland, Latvia, Lithuania and Sweden: the Universities of Helsinki, Latvia, Stockholm, Tartu, Umeå and Vilnius. The National Taras Shevchenko University of Kyiv, as the first representative from Ukraine, joined the Network in 2007 and then, the first Summer School in Ukraine was arranged in 2009.

The Network activity was initiated, expended and guided during long time by Professor Gunnar Kulldorff from the University of Umeå (Sweden). We regret that he passed away last year and miss the presence and energy of this outstanding person.

The main objectives of the School are to provide an opportunity for university teachers, research students and survey practitioners from different countries to discuss their problems and to learn from each other.

The School starts with an opening session by welcoming speeches by Risto Lehtonen (University of Helsinki), Mykhaylo Gorodnii, Dean of the Faculty of Mechanics and Mathematics (Taras Shevchenko National University of Kyiv), Yuliya Mishura, Head of the Department of Probability Theory, Statistics and Actuarial Mathematics (Taras Shevchenko National University of Kyiv) and Vadym Pishcheiko, Advisor to Head of the State Statistics Service of Ukraine.

The Programme Committee invited four keynote speakers: Jelke Bethlehem (University of Leiden), Vassili Levenko (Statistics Estonia), Kaija Ruotsalainen (Statistics Finland) and Imbi Traat (University of Tartu). There are also nine invited speakers who will deliver special lectures covering different topics of the theory and application of survey statistics and provide some practical knowledge of working with R-packages. There are 40 registered participants at the BNU Summer School. Some of them will present contributed papers included into this book. All presentations will be followed by discussions.

And last, but not the least, the financial support kindly given by the International Association of Survey Statisticians (IASS) is very much appreciated.

We wish everybody the inspiring Summer School and enjoyable stay in Kyiv!

On behalf of the Organizing Committee,
Yuliya Mishura
Tetiana Ianevych
Olga Vasylyk

CONTENT

Abstracts of lectures

Keynote speakers

Invited speakers

Contributed papers

ESTIMATION PROBLEMS IN WEB SURVEYS

Jelke Bethlehem

Leiden University, The Netherlands
e-mail: jelkeb@xs4all.nl

Survey data collection has undergone radical changes over the last decades. First, there was traditional data collection that came in three modes: face-to-face surveys, telephone surveys, and mail surveys. In the 1980s, there was the advent of computer-assisted interviewing. There were also three modes: CAPI (for face-to-face surveys), CATI (for telephone surveys) and CASI (the electronic form of mail surveys). And then, in the 1990s , the internet emerged, and it became possible to carry out web surveys.

Web surveys became rapidly very popular. This popularity is not surprising as a web survey is a simple means of getting access to a large group of potential respondents. Questionnaires can be distributed at very low costs. No interviewers are needed, and there are no mailing and printing costs. Surveys can be launched very quickly. Little time is lost between the moment the questionnaire is ready and the start of the fieldwork. Web surveys also offer new, attractive possibilities, such as the use of multimedia (sound, pictures, animation and movies).

Online surveys not only have advantages. There are also a number of methodological problems that may have a negative effect on the quality of the outcomes of web surveys. Because of these problems, the researcher runs a serious risk that the estimates of population characteristics are seriously biased.

This presentation describes a number of these methodological problems: under-coverage, sample selection, nonresponse, and measurement errors. Some practical examples show how important it is to be able to distinguish good and bad web surveys.

NONRESPONSE PROBLEMS IN SURVEYS: DETECTION AND CORRECTION

Jelke Bethlehem

Leiden University, The Netherlands
e-mail: jelkeb@xs4all.nl

Nonresponse occurs in a survey when people selected in the sample do not provide the requested information, or that the provided information is unusable. Nonresponse can have a serious impact on the outcomes of a surveys. Estimates of population characteristics may be seriously biased. Unfortunately, response rates decrease in many countries, so that the nonresponse problems increase.

This presentations gives more insight in the nonresponse problem. It shows in what way it can have an impact on estimates. Because of the serious consequences, it is important to realise already in the design stage of the survey that there will be nonresponse. Consequently, it is important to have auxiliary variables. These are variables that are measured in the survey, and for which the population distribution is available. The auxiliary variables can be used to analyse the nonresponse, to estimate response probabilities, and to correct for the negative effects of nonresponse.

The response rate is an important indicator of the survey response. Such an indicator is, however, not sufficient. An additional indicator is discussed. This is the R-indicator. It is an indicator of the representativity of the survey response.

Weighting adjustment is a frequently used technique to correct for a possible nonresponse bias. Auxiliary variables are an important ingredient of weighting adjustment. Unfortunately, not every auxiliary variable is effective. So weighting is no guarantee for success.

USE AND ABUSE OF GRAPHS

Jelke Bethlehem

Leiden University, The Netherlands
e-mail: jelkeb@xs4all.nl

A graph is an effective instrument to show a statistical message that is hidden in the data. Particularly for the general public, graphs work better than plain text, or data in tabular form. Therefore it is not surprising that many publications contain graphs. They work especially well when showing patterns in large amounts of data

On the one hand, a graph is a powerful way to convey the message in the data. On the other hand, there are also caveats. Errors in the design of graphs (either knowingly or unknowingly) may cause wrong conclusion to be drawn.

This presentation provides a set of guidelines for good graphs. It starts with a historic overview. Some iconic examples are given of graphs that 'tell a story'.

The presentation continues by describing the various ingredients of a graph, such as the data and the metadata. The metadata include titles, axes, tick marks, scales, colours, legends, etc.

Some guidelines are discussed in more detail, such as (1) not messing with the axes, (2) presenting data in the proper context, (3) careful use of color, (4) avoiding three-dimensional graphs, and (5) avoiding chartjunk. Many examples illustrate the proper use of guidelines.

PREPARING TO CONDUCT REGISTER-BASED POPULATION AND HOUSING CENSUS IN ESTONIA IN 2020

Vassili Levenko

Statistics Estonia, Estonia
e-mail: Vassili.Levenko@stat.ee

The last Population and Housing Census (PHC) in Estonia in 2011 was traditional sensus, incl e-census. The next PHC at the end of 2020 is intended to be register-based.

The main idea of register-based PHC is to find answers to formulated by Eurostat questions in existing register not disturbing Estonian residents.

In that connection many questions arise:

- Which registers to use;
- Do we have enough information in our registers;
- What is the quality of the information in registers;
- Is legislation base complete for the census purpose;
- Do we have communication channels to get the registers;
- Do we have software to process the information and get the result etc.

To answer these and other questions Statistics Estonia started preparations to register-based census in 2010 with the Methodology Project. The project was finished in 2013 and we understood that register-based PHC in Estonia is possible. Our starting point compared to other countries in similar situation is not worse, it may be even better.

Plans of preparations include three pilot PHC in 2014, 2016 and 2018. There are two main methodology tasks within the current pilot PHC: how to define the total population and the full set of dwellings using existing registers.

The lecture will be focused on these and other related issues discussing, e.g. the results of current pilot PHC.

Register-based population census methodology in Finland

Kaija Ruotsalainen

Statistics Finland
e-mail: kaija.ruotsalainen@stat.fi

Abstract

In Finland the use of administrative data and registers already has a history of over 40 years. The decisive step towards a register-based population system was taken at the end of the 1960s when the Central Population Register was established. By means of this system an identifying personal code was issued to each resident of Finland. The same personal identity code was taken into use in other administrative registers, such as in taxation and in the employment pension insurance system. The register data was used for the first time for the 1970 Population Census: personal data were obtained from the Central Population Register and income data from the taxation register.

The use of registers increased in the 1970s and 1980s so that the register-based data on all census related statistics were first produced for the year 1987; since then the statistics have been compiled on an annual basis. Each year, Statistics Finland produces demographic and employment statistics, building, dwelling, household and family statistics and statistics on housing conditions. Finland was the second country in the world to start using register data for population census purposes in 1990. The first was Denmark in 1981.

The source materials used for the census are mainly administrative registers and other register-based data materials. Direct data collection is made only for determining establishment data for those working for multi-establishment enterprises. Also, the data on occupation is collected from those enterprises whose employees' occupations are not available in any register. In all, data from about 30 different registers or data materials are usually used for completing the statistical data for the census.

The most important of the exploited registers are the Population Information system and the Business Register. Additional registers used include registers of employment pensions, taxation, unemployed, pensioners, and students.

As already earlier mentioned, the annual statistics on census related statistics are produced annually since 1987. They are available besides by national level but also by small areas like provinces and municipalities and by coordinate based areas. Also, it is possible to combine this annual data with the earlier census data from the years 1970-1985. This gives huge possibilities for research. With this longitudinal data it is possible to monitor e.g. the changes of the employment status of the population, the influence on the occupation for the causes of death or how students find a job after their graduation.

SAMPLING AND ESTIMATION IN SURVEYS

Imbi Traat

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

Summary

The presentation gives a brief overview of the basic ingredients of sample surveys − sampling and estimation. The development of the field until 1990 is well covered by the landmark papers collected into one volume Nathan et al. (2000). An influential book by Särndal et al. (1992) has been a basic source for many survey statisticians for the last 20 years. It covers in a unified manner the design-based theory, and brings auxiliary information and model building under this framework. The auxiliary information has become an important keyword for new developments. Under growing nonresponse rates it was effectively used to reduce the accompanying bias in the estimates. An even newer question has arisen − can data collection be monitored and directed to give a balanced response set, the balance measured with auxiliary information. The answer on the benefits from the balancing activity to estimation phase is still open (Särndal, et al. 2016).

The presentation gives reflections to the topic through my own research and supervision. It is concerned more with the sampling theory than with sampling applications. The classical approach is reminded along with jumps to other viewpoints realized in my and my co-authors' works (e.g. Traat et al. 2004).

References

Nathan, G. et al. editors. (2000) Landmark papers in Survey Statistics. *IASS Jubilee Commemorative Volume*. International Assotiation of Survey Statisticians.

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

Särndal, C.-E., Lumiste, K., Traat, I. (2016) Reducing the response imbalance: Is the accuracy of the survey estimates improved?*Survey Methodology*, to appear.

Traat, I., Bondesson, L., Meister, K. (2004) Sampling design and sample selection through distribution theory. *Journal of Statistical Planning and Inference*.

SMALL BUSINESS SURVEYS IN BELARUS

Natallia Bokun

Belarusian State Economic University, Belarus
e-mail: nataliabokun@rambler.ru

Abstract

The paper briefly describes the sampling methodology of micro-entities and small enterprises, problems of introduction of the micro-entities sample survey in practice of Belarusian official statistics. The sampling frame, sampling design and precision estimation are considered.

This paper on small business sampling has the next parts:

1) history of development of branch sample surveys;
2) small enterprises sample survey;
3) micro-entities sampling frames that incorporate two files of economic units: micro-entities and private farms;
4) micro-entities sample design; territorial stratified univariate and multivariate (multidimensional) samples are used. The algorithm to receive optimal sample size for *i*-th kind of activity and *j*-th region is presented;
5) statistical weighting that includes three methods: traditional Horvitz-Thomson estimator and calibration (GREG- and SYN-estimators).

Keywords: micro-entities, sample survey, sampling frame, weighting, small enterprises.

SAMPLE SELECTION IN R

Danutė Krapavckaitė

Vilnius Gediminas Technical University, Lithuania
e-mail: danute.krapavickaite@vgtu.lt

Abstract

The paper reviews the fields of official statistics implemented into the environment for statistical computing and graphics **R**. Packages with possibilities to draw a probability sample are discussed. The contents of the Summer School lecture are described.

1 Introduction

Computer software **R** is a language and an environment for statistical computing and graphics. It is free software, containing many packages for various statistical techniques, which can be highly extended.
Statistical packages are classified into groups according to the field of application (*task views*); one of them is called *Official statistics*. It includes:

- editing and visual inspection of micro-data,
- imputation of missing data,
- statistical matching and record linkage,
- indices and indicators and visualisation of indicators,
- seasonal adjustment,
- complex survey design: sample selection and estimation,
- micro-simulation.

The possibilities of using **R** in official statistics are reviewed in Templ and Todorov, 2016. Quite a number of packages are devoted to survey statistics. There are connections between some of the packages. The package `survey` seems to be central and has connections with many other packages.

2 R packages for sample selection

In this section, we review the packages enabling sample selection.

The package `base` is supplied with the **R** distribution. It contains a function `sample` for the selection of a sample with or without replacement, with equal or unequal selection probabilities. The sample selected is presented by a vector of length equal to the sample size containing indices of the selected elements.

Various algorithms for drawing an *unequal probability* sample are implemented in the package *sampling*: Brewer, Midzuno, with probabilities proportional to size, systematic, Sampford, balanced (cluster or stratified) sampling via the cube method, etc. Second-order inclusion probabilities are computed and presented for unequal probability sampling designs in order to be used for the estimation of variance for estimators of the population parameters. The sample selected is presented by a vector of length equal to the population size and sampled elements denoted by the components of the vector equal to 1; otherwise, the components of the vector are equal to 0. The algorithms for unequal probability sampling are described in Brewer &Hanif, 1983, and Tillé, 2006.

The `pps` package contains functions to select *samples with probabilities proportional to size* (pps). Stratified simple random sampling and computation of joint inclusion probabilities for Sampford's algorithm of pps sampling are also possible. The sample selected is presented by a vector of length equal to the sample size containing indices of selected elements.

The package `stratification` may be used for the *construction of a stratified sampling design*. It contains functions for univariate stratification of survey populations. A generalisation of the Lavallee-Hidiroglou method of stratum construction is used. The generalised method takes into account a discrepancy between the stratification variable and the survey variable. The determination of the optimal boundaries incorporates, if desired, an anticipated non-response, a take-all stratum for large units, a take-none stratum for small units, and a certainty stratum to ensure that some specific units are in the sample. The well-known cumulative root frequency rule of Dalenius and Hodges and the geometric rule of Gunning and Horgan are also implemented.

The package `SamplingStrata` offers an approach for choosing the best stratification of a sampling frame in a multivariate and multi-domain setting, where the sample sizes in each stratum are determined in order to satisfy the *minimum sample cost under the accuracy constraints*. This approach is based on the use of the genetic algorithm: each solution (i.e. a particular partition in the strata of the sampling frame) is considered as an individual in a population; the fitness of all individuals is evaluated applying the Bethel-Chromy algorithm to calculate the sample size satisfying the precision constraints on the target estimates. The functions of the package allow (a) analysing the results of the optimisation step; (b) assigning the new strata labels to the sampling frame; (c) *selecting a sample* from the new frame according to the best allocation. The functions for the execution of the genetic algorithm are a modified version of the functions in the `genalg` package. The sample selected is a data frame consisting of the sampled elements.

For estimation purposes and to work with the survey samples already drawn, the package `survey` can be used once the given survey design has been specified. The sample selected is presented as a vector of length equal to the sample size with indices giving the sampled elements. The function `svydesign` creates a `survey.design` object. It combines a data frame and all the survey design information needed to analyse it. This object is used by the survey modelling and summary functions. The book of the package author Lumley, 2010, presents methods of data analysis from complex surveys and provides documentation and explanation of the functions of the *survey* package

for the **R** statistical environment. The book supplements the reference manual of the *survey* package. It is designed for the readers who have some experience with applied statistics, especially in social and health sciences.

The `survey` package has a function to select a *stratified sample*. The conspicuous feature of the statistical data analysis in this package, beside the usual estimation methods for means, totals and ratios, is linear regression and generalised linear regression modelling, calibration of design weights, adjustment for non-response, including multiple imputation, and other statistical topics.

3 Contents of the lecture

Sample selection using the function `sample` and `survey` package possibilities will be shown, but attention will be mainly focussed on sample selection using the `sampling` package. The main features of the unequal probability sampling algorithm will be presented, the syntax of the corresponding functions will be given, and the examples of their application will be shown. The listeners will have a possibility to run these examples themselves. Exercises for self-study and practice will be given. The audience will be equipped with the materials of the lecture in electronic form.

References

Brewer, K. R.; Hanif, M. (1983) *Sampling with Unequal Probabilities*. Series: *Lecture Notes in Statistics*, **15**. Springer, New York.

Lumley, T. (2010) *Complex Surveys*. John Wiley & Sons, Hoboken.

Templ, M.; Todorov, V. (2016) The Software Environment R for Official Statistics and Survey Methodology. *A Austrian Journal of Statistics*, **45**, 97-124. *http://www.ajs.ot.ut/* `doi:10.17713/aja.v4511.100`

Tillé, Y. (2006) *Sampling Algorithms*. Springer, New York.

The Comprehensive R Archive Network. *URL http://CRAN.R-project.org*

RADIATION RISK ESTIMATION UNDER MEASUREMENT ERRORS IN EXPOSURE DOSES

Alexander Kukush[1]

[1] Taras Shevchenko National University of Kyiv, Ukraine
e-mail: alexander.kukush@gmail.com

We study the effect of measurement errors in exposure doses in a regression model with binary response. Recently it has been recognized that uncertainty in exposure dose is characterized by measurement errors of two types: classical additive errors, and Berkson multiplicative errors. In a simulation study based on data from radio-epidemiological research of thyroid cancer in Ukraine caused by Chornobyl accident, it is shown that ignoring measurement errors in doses leads to overestimation of background prevalence and underestimation of excess relative risk.

We propose several methods to reduce bias: (a) new Regression Calibration,
(b) SIMEX (simulation-extrapolation) that takes into account errors of both types, and
(c) novel Corrected Score method.
The SIMEX method is the most flexible and plausible one.

The results are joint with Prof. I. A. Likhtarev, Dr. S. V. Masiuk, Dr. L. N. Kovgan (Radiation Protection Institute of the Academy of Technological Sciences of Ukraine), and Dr. S. V. Shklyar (Taras Shevchenko National University of Kyiv).

ANTICIPATED AND REALIZED DESIGN EFFECTS OF THE EUROPEAN SOCIAL SURVEY

Seppo Laaksonen

University of Helsinki, Finland
e-mail: Seppo.Laaksonen@Helsinki.Fi

The European Social Survey (ESS) has aimed to control the sample designs used by specifying sampling guidelines that should have been followed in each participating country. The main requirements are the use of probability sampling and the achievement of a minimum effective sample size that is determined by ineligibility rate, nonresponse rate, inclusion probabilities and clustering effects. Several sampling strategies have been used over the first seven rounds.

This paper does not cover everything in sampling designing but it is focused on one demanding question of it. How to anticipate or predict the key components of the design effects at the sampling designing stage so that the final estimates are as accurate as expected? This cannot be made ever perfectly but as well as it is possible so that the comparisons between countries and other domains are reliable reasonably. The sampling expert team is responsible for this task (I am one member of it). It has prepared the guidelines for country experts that should be followed before the approval of the sampling design. The guidelines give the realistic requirements for the following factors that all should be anticipated: (i) ineligibility, (ii) unit nonresponse, (iii) the variability of the inclusion probabilities and consequently of the sampling weights that leads to the design effects (DEFF) due to varying weights, (iv) intra-class correlation and net cluster size that are the key components of the DEFF due to clustering. When these two components are multiplied, the final design effect are obtained. Currently, this DEFF has not been tried to anticipate and apply by strata that could lead to a slightly lower DEFF. This could be however useful since the realized estimates should be calculated including the stratification variable. Unfortunately, it is not now possible since the variables stratum and primary sampling unit (PSU) are not in the public website files. The main reason is that some countries do not allow to publish them for confidentiality reasons.

The DEFFs are naturally dependent also on the missingness since they the net sample sizes is required to anticipate as well. Currently, ineligibility rates are more difficult to know since the migration between countries is hard to predict. However, the foreign based people who are residents of the country belong to the target population and are more and more interest for researchers.

The paper includes many details about these problems, and most recent results as well. The ESS data of the first seven rounds are now available, and the eight round is at the designing phase, including both sampling designing and fieldwork designing.

SELECTIVE EDITING

Thomas Laitila

Örebro university, Sweden, and Statistics Sweden
e-mail: thomas.laitila@oru.se

Abstract

A major activity in a survey is the editing process. Developments of new theories and methods useful for reducing the resources spent on editing are of interest as it may provide with substantial cost savings and improved timeliness. One alley for reducing editing is development of more efficient tools for identification of erroneous observations. Another is to reduce the number of observations edited. Indeed the traditional approach to edit all observations is not generally necessary for appropriate statistical inference.

One approach towards a more efficient editing process is selective editing, defined as editing methods where only a subset of the response set is selected for editing (Granquist and Kovar, 1997). In the selective editing literature the leading idea is to spend resources only on those observations which will have potential effects on the estimates. For this selective editing is based on the calculations of "global scores" expressing a combined measure of importance in estimation and suspicion of measurement error. These global scores can then be used for ranking of observations in the response set and those observations with the largest scores are selected for editing. Here either a predetermined number of observations are edited or all observations with a score larger than a threshold are edited.

Suggested selective editing procedures are largely based on ad hoc methods developed from pure common sense reasoning, and there is yet no accepted theory developed (de Waal, 2014). In particular it does not rest on randomization theory and it is not possible to use traditional statistical methods for generalizing the results from the edited set to the set of non-edited observations and a corresponding part of the population. Ilves and Laitila (2009) and Laitila and Ilves (2012) therefore suggest selective editing based on random selection of units for editing and derives, with respect to measurement errors, unbiased estimators of population parameters.

This talk presents the general idea of selective editing and its implementation using scores. The inferential limitations with this metthod are discuseed and the method of random selection is introduced as a solution. An alternative method based on modelling and predition of measurement errors is also presented, including a discussion of models of random measurement errors.

References

De Waal, T. (2014). Selective Editing: A Quest for Efficiency and Data Quality, *Journal of Official Statistics,* 29:4, 473-488.

Granquist, L. and J.G. Kovar, (1997). Editing of Survey Data: How Much is Enough?, In: *Survey Measurement and Process Quality*, L.E. Lyberg, P. Biemer, M. Collins, E.D. de Leeuw, C. Dippo, N. Schwarz and D. Trevin (eds.), Wiley, Hoboken, NJ, pp. 416-435.

Ilves, M. and T. Laitila (2009. Probability-Sampling Approach to Editing, Austrian Journal of Statistics, 38:3, 171-182.

Laitila, T. and M. Ilves (2012). Probability Editing, Paper prepared for the UNECE Work Session on Statistical Data Editing, Oslo, Norway, 24-26 September, 2012.

CALIBRATION METHODS FOR DOMAIN AND SMALL AREA ESTIMATION

Risto Lehtonen[1] and Ari Veijanen[2]

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

[2] Statistics Finland, Finland
e-mail: ari.veijanen@stat.fi

ABSTRACT

Calibration refers to a family of methods commonly used in official statistics to incorporate auxiliary information on the population in the estimation of sample-based statistics for the population or sub-populations. Calibration methods discussed in the paper include the traditional model-free calibration (Deville and Särndal 1992), several variants of model calibration (Wu and Sitter 2001, Lehtonen and Veijanen 2016a,b) and a more recent method we call hybrid calibration (Lehtonen and Veijanen 2015). Hybrid calibration represents an attempt to combine some of the favorable properties of model-free calibration and model calibration into a single method that uses modeling and auxiliary data at the unit level and at an aggregate level. All these methods are design based and share the standard properties of design-based estimators such as approximate design unbiasedness and improved efficiency over the Horvitz-Thompson estimator if certain favorable conditions are met. Model calibration and hybrid calibration constitute model-assisted methods because an explicit assisting model is postulated, such as a member of the family of generalized linear mixed models, in contrary to model-free calibration that only involves an implicit linear relationship between the study variable and the auxiliary variables. In estimation for domains, calibration methods involve the construction of calibration weights that reproduce the domain totals of the calibration variables, when the weights are applied to the sample. In model-free calibration, the auxiliary variables constitute the vector of calibration variables, whereas predictions from the assisting model serve in this role in model calibration. In hybrid calibration, a set of auxiliary variables and predictions from a model fitted with a set of auxiliary variables are inserted in the calibration vector, constituting the model-free calibration part and the model calibration part of the calibration vector. The two sets of auxiliary variables can be separate or they can overlap. In the methods referred above, calibration is defined at the domain level. We extend hybrid calibration to cases where the model calibration part is defined at the domain level (e.g. NUTS4) and the model-free calibration part is defined at a higher hierarchical level (e.g. NUTS3). We call the new method as two-level hybrid calibration. We discuss the relative merits of each calibration method. Statistical properties of the methods are examined by simulation experiments using artificially generated data and real data obtained from Statistics Finland.

References

Deville J.-C. and Särndal C.-E. (1992). Calibration estimators in survey sampling. *JASA* 87, 376–382.

Lehtonen R. and Veijanen A. (2015). Small area estimation by calibration methods. World Statistics Conference, July 2015, Rio de Janeiro.

Lehtonen R. and Veijanen A. (2016a). Model-assisted methods for small area estimation of poverty indicators. In: Pratesi M. (ed.) *Analysis of Poverty Data by Small Area Estimation*. Chichester: Wiley, 109–127.

Lehtonen R. and Veijanen A. (2016b). Estimation of poverty rate and quintile share ratio for domains and small areas In: Alleva G. and Giommi A. (eds.) *Topics in Theoretical and Applied Statistics*, New York: Springer, 153–165.

Wu C. and Sitter R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *JASA* 96, 185–193.

R TOOLS IN SURVEY ESTIMATION (R-PACKAGE: SAMPLING)

Natalja Lepik

University of Tartu, Estonia
e-mail: natalja.lepik@ut.ee

Summary

The computer lab will concentrate on the possibilities of R-package 'sampling' (Tillé and Matei, 2015) for the estimation of the population total and mean in survey sampling. The design-based approach by Särndal et al. (1992) will be used. The lab will start with the Horvitz-Thompson and the Hájek estimators of the population total. The stratified design will be considered as a special case. Nowadays different data sources are available and the auxiliary information can be implemented to the estimation process. Well-known calibration approach will be also practised during the lab. In addition, exercises on the variance estimation will be looked through.

The important topic in today's sampling surveys is growing nonresponse rate. This leads to the biased estimates. In our lab some nonresponse adjustment methods will be covered.

Overview of the relevant R-functions of package 'sampling' will be given with some practical examples. In our lab we will use the RStudio (Open Source Edition) as a user-friendly R-software. The special format R-markdown will be introduced to create dynamic documents from R.

References

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

Tillé, Y., Matei, A. (2015) Package 'sampling'.
https://cran.r-project.org/web/packages/sampling/sampling.pdf

R TUTORIAL

Mārtiņš Liberts[1] and Juris Breidaks[2]

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

[2] Central Statistical Bureau of Latvia
e-mail: juris.breidaks@csb.gov.lv

Abstract

Four sessions "Use of R in survey statistics" will be organised during the Baltic – Nordic – Ukrainian Summer School on Survey Statistics 2016. A short practical R training will be done during the first session – "R tutorial". The aim of the session is to provide practical knowledge necessary to work with R packages *surveyplanning*, *survey*, and *sampling* which will be used during the next three R sessions.

1 Introduction

R (R Core Team 2016) is becoming more and more popular tool for statisticians (Muenchen 2016). The similar trend is observable also in the field of survey statistics. Most of the young statisticians have gained the first experience with R form the studies in university so it is easier for them to start using R also in their research or work environment.

2 The content of the session

The overall aim of the four sessions "Use of R in survey statistics" organised during the Baltic – Nordic – Ukrainian Summer School on Survey Statistics 2016 is to show how R can be used for three important stages of survey sampling process, namely:

- Sample designing;
- Sampling;
- Estimation.

Three R packages will be used during the sessions:

- *surveyplanning* (Breidaks, Liberts, Jukams 2016);
- *survey* (Lumley 2014);
- *sampling* (Tillé, Matei 2015).

The aim of the session "R tutorial" is to provide practical knowledge necessary to work with the mentioned R packages. The content of the session will be:

- What is R and how does it look like?

- Where to get R?
- Data objects in R;
- Data processing in R;
- Data import and export in R.

2.1 What is R and how does it look like?

R is a free software environment for statistical computing and graphics (R Core Team 2016). R is available for a wide variety of UNIX platforms, Windows and MacOS. You can use R for free – there is no charge. R is an open source project – the source code is publicly available. You can get it, explore it, test it and modify for your own needs if necessary.
There are many ways how you can interact and work with R. The main component of R is the R engine – the core executable who does the calculations. You can run the R engine from the console or terminal, but this is not the usual way how you work with R.
You can write R scripts. An R script is a plain text file containing the commands which will be executed by the R engine. You can copy commands from one script to another. You can reuse previously prepared scripts. The usual work of statistician is to write and maintain R scripts. R script should contain all the data processing steps done by a statistician.
Usually we work with a graphical user interface (GUI) or an integrated development environment (IDE). GUI or IDE is a software which makes work with R scripts easier and more productive. The main features of GUI or IDE is:

- Simple access to the R engine. With a key combination or a button you can transmit commands from a script to the R engine and see the result directly in an output screen;
- Advanced text editor (with syntax highlighting, command completion, code diagnostics and many other features);
- Simple access to the R help system;
- Integration with other useful tools (for example, version control tools like *git* or *SVN*).

The most common GUI for R is the RGui which is provided together with the R installation for Windows. Very popular IDE for R is RStudio (RStudio Team 2015).

2.2 Where to get R?

The usual installation of R consists of the R engine, core packages and additional packages. All of the files necessary for R installation are distributed through the Comprehensive R Archive Network (CRAN). This is a network of servers all around the world which are providing identical copies of R source code and installation files. The access to the CRAN is available through https://cran.r-project.org/.

RStudio is a separate project from the R project. The information and installation files are available at https://www.rstudio.com/.

2.3 Data objects in R

R programming language is an object–oriented programming language. Information in R environment is organised as individual objects. The user can create new objects, modify existing objects and save the objects in R data file. There are several types of objects. The following objects will be considered during the session:

- `vector` – the basic R object;
- `matrix` – a vector with a dimension attribute;
- `data.frame` – a list of vectors, `data.frame` is the most similar data structure to the SPSS, SAS, or Stata data file;
- `data.table` – an extension of `data.frame` (Dowle, Srinivasan, et al. 2015);
- `function` – with function you can create or modify objects in R.

2.4 Data processing in R

There are many ways how a user can interact with R objects. The following processing commands will be considered during the session:

- creating of a new object;
- printing – the most common procedure to observe the content of an R object;
- subset or extract – you can extract elements from a vector or a matrix, rows or columns from any two dimensional data object (`matrix`, `data.frame`, or `data.table`);
- arithmetic operations.

2.5 Data import and export in R

It is possible to exchange information between R and many other data formats and information systems. Some data exchange examples:

- The most common and robust data exchange format is a text file formatted as comma–separated value (CSV) file. There are several function in R to read data from CSV files and to save data as CSV files.
- MS Excel data files are not recommended as data exchange format but still is very popular. There are many R packages allowing to exchange data between R and Excel. We will look at the *openxlsx* package (Walker 2015).
- It is possible to work with SPSS, SAS and Stata data files with R (R Core Team 2015; Wickham, Miller 2015).
- Using *RODBC* package it is possible to read and write data to any ODBC capable data system, for example, MS SQL Server (Ripley, Lapsley 2016).

3 Conclusions

R is the new language of statisticians including survey statisticians. The R world is very flexible. The most recent statistical procedures are available in R. If there is something missing, you can become an author and write your own R functions to implement the missing statistical procedure. If you think it could be useful also for others, create an R package and make it available. R project is a community project – anybody is welcome to contribute.

References

Juris Breidaks, Martins Liberts, and Janis Jukams (2016) surveyplanning: Survey Planning Tools. R package version 1.6. https://CRAN.R-project.org/package=surveyplanning

Matt Dowle, Arun Srinivasan, et al. (2015) data.table: Extension of Data.frame. R package version 1.9.6. https://CRAN.R-project.org/package=data.table

Thomas Lumley (2014) survey: analysis of complex survey samples. R package version 3.30

Thomas Lumley (2004) Analysis of complex survey samples. Journal of Statistical Software 9(1): 1-19

Bob Muenchen (2016, June 8) R Passes SAS in Scholarly Use (finally) [Blog post]. Retrieved from http://r4stats.com/2016/06/08/r-passes-sas-in-scholarly-use-finally/

R Core Team (2015) foreign: Read Data Stored by Minitab, S, SAS, SPSS, Stata, Systat, Weka, dBase, ... R package version 0.8-66. https://CRAN.R-project.org/package=foreign

R Core Team (2016) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Brian Ripley and Michael Lapsley (2016) RODBC: ODBC Database Access. R package version 1.3-13. https://CRAN.R-project.org/package=RODBC

RStudio Team (2015) RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/

Yves Tillé and Alina Matei (2015) sampling: Survey Sampling. R package version 2.7. https://CRAN.R-project.org/package=sampling

Alexander Walker (2015) openxlsx: Read, Write and Edit XLSX Files. R package version 3.0.0. https://CRAN.R-project.org/package=openxlsx

Hadley Wickham and Evan Miller (2015) haven: Import SPSS, Stata and SAS Files. R package version 0.2.0. https://CRAN.R-project.org/package=haven

R TOOLS IN SURVEY DESIGN

Mārtiņš Liberts[1] and Juris Breidaks[2]

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

[2] Central Statistical Bureau of Latvia
e-mail: juris.breidaks@csb.gov.lv

Abstract

Four sessions "Use of R in survey statistics" will be organised during the Baltic – Nordic – Ukrainian Summer School on Survey Statistics 2016. The aim of this session is to present R procedures useful for sample survey planning stage. The R packages *surveyplanning* will be used as an example.

1 Introduction

The planning stage of sample surveys is probably the most critical stage from the whole process. The quality of the survey results depends very much from the decisions made at the planning stage. If you make a mistake at the planning stage, it will be very hard – even impossible to make any error correction in latter stages.
The usual questions we are dealing with at the planning stage is:

- The sample size for survey;
- Sample allocation by strata;
- Expected precision for the estimates of population parameters we will achieve.

The R (R Core Team, 2016) package *surveyplanning* (Breidaks, Liberts, Jukams 2016) can help to answer these questions. The aim of the session is to provide practical knowledge necessary to work with the R package *surveyplanning*. The following R functions will be considered during the session:

- *s2* for population variance estimation;
- *expsize* for sample size calculation;
- *optsize* and *dom_optimal_allocation* for optimal sample size allocation;
- *expvar* for expected precision for the estimates of totals.

2 The functions

The R package *surveyplanning* contains several functions useful at the sample planning stage.

2.1 Function *s2*

The function *s2* can be used to compute or estimate the population variance. This is a universal function which can be applied to the population frame, sample file or the file of respondents. The population variance is estimated or computed using the formula

$$\sigma^2 = \frac{N-1}{N}\frac{n}{n-1}\frac{1}{N-1}\left(\sum_{i=1}^{n} y_i^2 w_i - \frac{1}{N}\left(\sum_{i=1}^{n} y_i w_i\right)^2\right), \tag{1}$$

$$N = \sum_{i=1}^{n} w_i, \tag{2}$$

where

- n is the number of records in data file (it can be size of frame, sample size or number of respondents);
- y_i is the value of the target variable for the record i;
- w_i is the weight for the record i.

In case of population frame $n = N$ and $w_i = 1$ and the formula is equivalent to the usual formula for the population variance

$$\sigma^2 = \frac{N-1}{N}\frac{n}{n-1}\frac{1}{N-1}\left(\sum_{i=1}^{n} y_i^2 w_i - \frac{1}{N}\left(\sum_{i=1}^{n} y_i w_i\right)^2\right), \tag{3}$$

$$= \frac{N-1}{N}\frac{N}{N-1}\frac{1}{N-1}\left(\sum_{i=1}^{N} y_i^2 - \frac{1}{N}\left(\sum_{i=1}^{N} y_i\right)^2\right), \tag{4}$$

$$= \frac{1}{N-1}\left(\sum_{i=1}^{N} y_i^2 - \frac{1}{N}\left(\sum_{i=1}^{N} y_i\right)^2\right) = S^2. \tag{5}$$

In case of simple random sample $w_i = \frac{N}{n}$ and the formula is equivalent to the usual formula for the sample variance

$$\sigma^2 = \frac{N-1}{N}\frac{n}{n-1}\frac{1}{N-1}\left(\sum_{i=1}^{n} y_i^2 w_i - \frac{1}{N}\left(\sum_{i=1}^{n} y_i w_i\right)^2\right), \tag{6}$$

$$= \frac{N-1}{N}\frac{n}{n-1}\frac{1}{N-1}\left(\frac{N}{n}\sum_{i=1}^{n} y_i^2 - \frac{1}{N}\left(\frac{N}{n}\sum_{i=1}^{n} y_i\right)^2\right), \tag{7}$$

$$= \frac{N-1}{N}\frac{n}{n-1}\frac{1}{N-1}\frac{N}{n}\left(\sum_{i=1}^{n} y_i^2 - \frac{1}{N}\frac{N}{n}\left(\sum_{i=1}^{n} y_i\right)^2\right), \tag{8}$$

$$= \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2\right) = s^2. \tag{9}$$

2.2 Function *expsize*

The aim of the function *expsize* is to calculate minimal sample size for each stratum to achieve defined precision for the estimates of totals in each stratum. The minimal sample size is calculated as

$$n_h = \frac{N_h^2 S_h^2 \,\mathrm{deff}_h}{R_h((Y_h \frac{\mathrm{CV}_h}{100})^2 + N_h S_h^2 \,\mathrm{deff}_h)}, \tag{10}$$

where

- N_h is the population size in stratum h;
- Y_h is the population total in stratum h;
- S_h^2 is the population variance in stratum h;
- R_h is the response rate in stratum h;
- deff_h is the design effect in stratum h;
- CV_h is the necessary precision to be achieved in stratum h defined as coefficient of variation (in percentage).

The population size N_h usually can be computed from the population frame. The population total Y_h and the population variance S_h^2 can be computed from the population frame using some good auxiliary variable (if available) or estimated from the previous or similar other sample survey. The response rate R_h and the design effect deff_h has to be guessed or estimated from the similar survey. The necessary precision is set by a user.

2.3 Function *optvar*

The aim of the function is to calculate the optimal sample allocation to estimate the population total $Y = \sum_U y_i$ with minimum variance. The following assumptions need to be taken into account:

- sample size is provided as n;
- stratified simple random sampling will be used;
- population is broken down in H strata;
- Horvitz–Thompson estimator with non-response correction will be used in each stratum (assumption on constant response probability for all elements in stratum).

Optimal sample allocation, which ensures minimal variance for $\hat{Y}$, can be calculated as

$$n_h = n \frac{N_h S_h \sqrt{\frac{\mathrm{deff}_h}{R_h}}}{\sum_{i=1}^{H} N_i S_i \sqrt{\frac{\mathrm{deff}_i}{R_i}}}, \tag{11}$$

where

- N_h is the population size in stratum h;
- S_h is the population standard deviation in stratum h;
- R_h is the response rate in stratum h;
- deff_h is the design effect in stratum h.

Theoretical proof that the given arrangement is optimal – ensures minimal dispersion for estimation Y - is not available at the moment. If the response rate and design effect in all strata is the same ($R_h = R$ and $\text{deff}_h = \text{deff}$), the optimal sample allocation is equal to Neyman's allocation (Neyman, 1934)

$$n_h = n\frac{N_h S_h}{\sum_{i=1}^{H} N_i S_i}. \tag{12}$$

If the response rate, design effect, and population standard deviation in all strata is the same ($R_h = R$, $\text{deff}_h = \text{deff}$, and $S_h = S$), the optimal sample allocation is equal to proportional allocation

$$n_h = n\frac{N_h}{\sum_{i=1}^{H} N_i} = n\frac{N_h}{N}. \tag{13}$$

2.4 Function *dom_optimal_allocation*

The function *dom_optimal_allocation* is function for computing the sample size and the optimal sample allocation. This function is useful when precision requirements are set for the estimates at population domains not strata. Note that any stratum should be a subset of a domain.
The sample size in each stratum is computed by Neyman allocation (Neyman, 1934) for each domain separately, taking into account the number of statistical units per stratum, and variable variance of the respective stratum. Computation of sample size and allocation is done using iterative process:

1. Initial sample size is set for a stratum h in domain d:

$$n_{h,d}^{min} = min(N_{h,d}; n^{min}), \tag{14}$$

   where $N_{h,d}$ is the population size for a stratum h in domain d and n^{min} is minimum sample size for each stratum;

2. The sample size for domain d is computed as:

$$n_d = \sum_{h=1}^{H_d} n_{h,d}^{min}, \tag{15}$$

   where H_d is the number of strata in domain d;

3. Sample size for a stratum h in domain d is computed as Neyman optimal allocation in each domain independently:

$$n_{h,d}^{opt} = n_d \frac{N_{h,d} S_{h,d}}{\sum_{i=1}^{H_d} N_{i,d} S_{i,d}}, \tag{16}$$

where $N_{h,d}$ is population size for stratum h in domain d, $S_{h,d}$ is population standard deviation for stratum h in domain d.

4. The strata sample size is corrected if it is larger than population size:

$$n'_{h,d} = min(N_{h,d}; n_{h,d}^{opt}) \tag{17}$$

5. The expected precision CV_d is calculated for each domain d using the sample allocation computed in the previous step.

6. For domains where $\mathrm{CV}_d > \mathrm{CV}_d^{max}$ the domain sample size is increased $n_d := n_d + 1$ and steps 3 to 6 are repeated until $\mathrm{CV}_d \leq \mathrm{CV}_d^{max}$ for all domains d. CV_d^{max} is the required precision set by user.

2.5 Function *expvar*

The function *expvar* can be used to compute the expected precision for the estimates of totals. Estimates for population, domains and strata are considered. Note that any stratum should be a subset of a domain in case for domain estimates. Variance for stratum, domain and population total are computed as (Särndal, 1992)

$$\mathrm{var}\left(\hat{Y}_{h,d}\right) = \sum_{U_{h,d}} N_{h,d}^2 \frac{\left(1 - \frac{n_{h,d}\tau_{h,d}}{N_{h,d}}\right)}{n_{h,d}\tau_{h,d}} S_{h,d}^2 \,\mathrm{deff}\left(\hat{Y}_{h,d}\right), \tag{18}$$

$$\mathrm{var}\left(\hat{Y}_d\right) = \sum_{h=1}^{H_d} \mathrm{var}\left(\hat{Y}_{h,d}\right), \tag{19}$$

$$\mathrm{var}\left(\hat{Y}\right) = \sum_{d=1}^{D} \mathrm{var}\left(\hat{Y}_d\right), \tag{20}$$

where

- $U_{h,d}$ is the set of population elements of stratum h of domain d;
- $\hat{Y}_{h,d}$ is the estimate of total in stratum h of domain d;
- $\hat{Y}_d$ is the estimate of total in domain d;
- $\hat{Y}$ is the estimate of total in population;
- $N_{h,d}$ is the population size of stratum h of domain d;
- $n_{h,d}$ is the sample size of stratum h of domain d;

- $\tau_{h,d}$ is the expected response rate in stratum h of domain d;
- $S^2_{h,d}$ is population variance in stratum h of domain d;
- deff $\left(\hat{Y}_{h,d}\right)$ is the design effect for the estimate of total in stratum h of domain d;
- H_d is the number of strata in domain d;
- D is the number of domains.

If simple random sample will be used in each stratum, design effect deff $\left(\hat{Y}_{h,d}\right)$ is equal to 1. If another sample design will be used, the expected design effect has to be estimated. If similar sample design was used in a previous survey, the design effect can be estimated from the data of the previous survey. Estimation of design effect from above can be used. Setting higher value for the estimate of design effect will result with more conservative estimate of variance.
The expected coefficient of variation is calculated as

$$\mathrm{CV}(\hat{Y}_{h,d}) = 100\frac{\sqrt{\mathrm{var}(\hat{Y}_{h,d})}}{\hat{Y}_{h,d}}, \tag{21}$$

$$\mathrm{CV}(\hat{Y}_{d}) = 100\frac{\sqrt{\mathrm{var}(\hat{Y}_{d})}}{\hat{Y}_{d}}, \tag{22}$$

$$\mathrm{CV}(\hat{Y}) = 100\frac{\sqrt{\mathrm{var}(\hat{Y})}}{\hat{Y}}. \tag{23}$$

The function *expvar* can be used also for the ratio of two totals $R = \frac{Y}{Z}$. Taylor linearisation is applied to the statistic $\hat{R} = \frac{\hat{Y}}{\hat{Z}}$ (Särndal, 1992) and linearised variable

$$\hat{u}_i = \frac{1}{\hat{Z}}\left(y_i - \hat{R}z_i\right) \tag{24}$$

is derived. The population variance $S^2_{h,d}$ is computed or estimated using the values of $\hat{u}_i$ then.

3 Conclusions

We have seen how sample size, sample allocation and expected precision can be computed. Please note that we never now the true information about the population when planning a sample. A lot of assumption has to be made and this is the hardest task – to make a good assumptions. It requires some experience certainly.

References

Juris Breidaks, Martins Liberts, and Janis Jukams (2016) surveyplanning: Survey Planning Tools. R package version 1.6. https://CRAN.R-project.org/package=surveyplanning

Jerzy Neyman (1934) On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society*, vol. 97 (**4**), p. 558–625, available at: http://www.jstor.org/stable/2342192

R Core Team (2016) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Carl-Erik Särndal, Bengt Swensson, and Jan Wretman (1992) *Model Assisted Survey Sampling.* Springer, New-York

FACTORIAL SURVEYS IN R

Mykola Sydorov

Sociology faculty of Taras Shevchenko National University of Kyiv, Department of Methodology and Methods of Sociological Research. Ukraine
e-mail: ms123@ukr.net

Abstract

Factorial Design is a quantitative method comes from Social Psychology and enables disentangling of causal factors which are confounded in reality and enables evaluations of rare situations. The procedure of crating vignette set and data analysis are presented in the paper.

Idea of Factorial Design

Factorial Design is method of collecting data and data analysis very similar to Conjoint Analysis but suitable for Quantitative research. The idea is the same as: presents to respondent a number of experimental situations and ask about his (her) opinion. Then usually builds the regression equation with dependent "answer for experiment situation" and independent "characteristics of object".

The example of question is:

Read the description of a person below. Could you imagine this person as your friend? Please mark your answer on a scale from (-3) "Very unlikely" to (+3) "Very likely".

A woman who is the same age as you, speaking mostly Ukrainian. Her hobby is literature and music. She is from Donetsk region. She has the same attitude to studies at University as you. This person helped you when you needed help. She helped you She was indifferent to Euromaidan.

-3 «Very unlikely» ... +3 «Very likely»)

Each question is an experimental situation in which some levels of some characteristics of object we interested are presented. Each characteristic we name "dimension". Our goal is to try all different combination of levels and dimensions to fix the respondent opinion for each experimental situation (vignette) (Auspurg and Hinz (2015)).

Using of Factorial Design

This method is very popular in marketing research but is not so suitable in sociological surveys because of a big number of different experimental situations (total set of possible vignettes names "vignette universe"). For example in our survey "Role of Ideological Issues in Friendship" the volume of vignette universe was 2*3*3*3*3*4*4*3=7776 - we used 8 dimensions with different number of levels:

1) 2 levels of «Gender» (male, female);

2) 3 levels of «Age» (younger than you, your age, elder than you);

3) 4 levels of «Region» (Central Ukraine, Donbas and Crimea, South-Eastern region except Donbas and Crimea, Western Ukraine).

4) 3 levels of "Language of communication" (mostly Ukrainian, mostly Russian, equally Ukrainian and Russian);

5) 3 levels of "Attitude to Euromaidan" (supported, was against, was indifferent);

6) 3 levels of "Attitude to studies at the University" (the same as you, opposite to you, undefined);

7) 4 levels of «Hobby» (reading and music, biking and tourism, dancing and clubbing, TV and computer games);

8) 3 levels of "Experience of help" (did not help you when you needed help; helped you when you needed help; you don't have the experience if this person could help in difficult situation)

Sure this number of questions is impossible to ask each respondent and for Factorial Surveys use the specific algorithms for quality reducing number of different vignettes with minimal lost of information. One of this methods names D-efficient method of creating orthogonal matrix of combination. This means that we have the sample of vignettes, not only respondents.

In R (R Core Team (2015)) there is an AlgDesign library with possibility of creating vignette universe and performing the D-efficient sampling of vignettes.

In result each respondent receives a number (we set 5 vignettes) of different vignettes and it means that the questionnaire for each respondent is different (Appendix 1).

We had a respondent list (there were all students from Sociology Faculty with a list of e-mails) and by LimeSurvey generated online questionnaire different for each respondent and depend on e-mail of respondent.

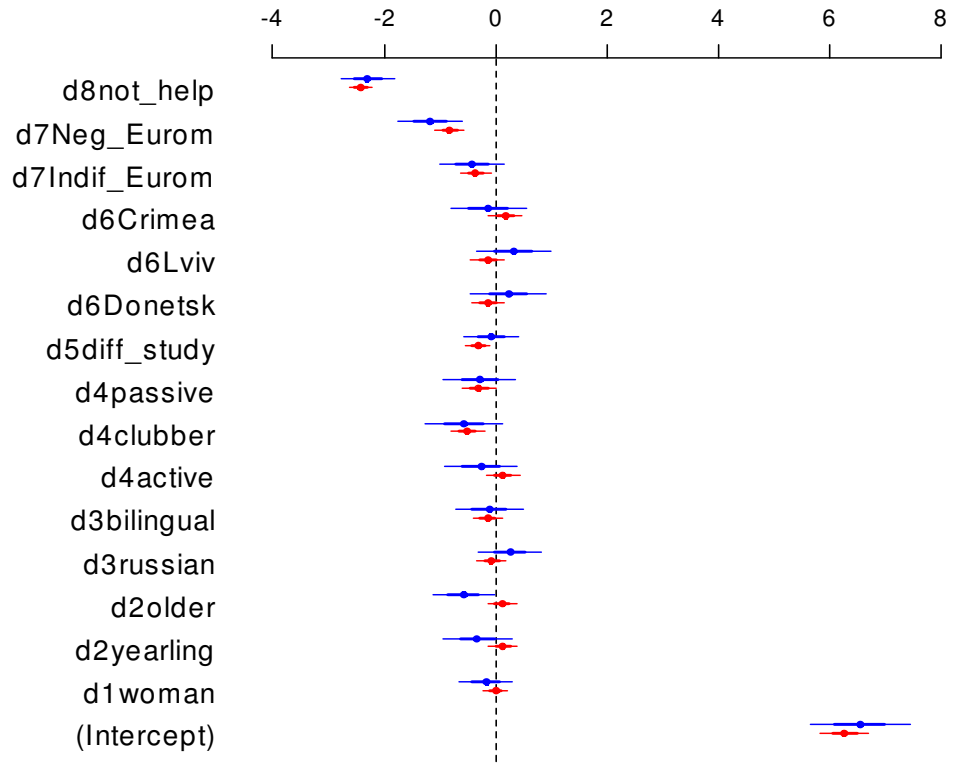
After collecting the data is easy to build the regression equation. This equation is not Best Fit equation but the goal is to find how dimensions are significant for respondent's decision. By using of "arm" library is possible to build regression model result by sex (Appendix 2)

Appendix 1.

```
install.packages("AlgDesign")

respnum<-351 #number of respondents
vgnset<-5 #number of vignettes in set (for each respondent
gets a different set)
vgnall<-respnum*vgnset #volume of vignette universe
dimnum<-8 #number of dimensions in vignette

library(AlgDesign)

#building of vignette universe
data<-gen.factorial(c(2,3,3,4,2,4,3,2), factors="all",
varNames=c("d1", "d2", "d3", "d4", "d5", "d6", "d7", "d8"))
vnum<-c(1:nrow(data)) # just number of vignettes
#data frame with vignette universe
VgnSpace<-data.frame(vnum,data)

seed<-6553555
set.seed(seed)
#Building optimal sample of vignettes
set.seed(seed)
Dsample<-optFederov(frml=~.^2,data=VgnSpace[,2:(2+dimnum-1)],
nTrials=vgnall, criterion="D")
```

Appendix 2.

```
library(arm)
model0f<-
lm(v~d1+d2+d3+d4+d5+d6+d7+d8,data=SurveyShort[SurveyShort$q1==
"Famale",])
model0m<-
lm(v~d1+d2+d3+d4+d5+d6+d7+d8,data=SurveyShort[SurveyShort$q1==
"Male",])
coefplot(model0f,xlim=c(-4, 8), col.pts="red",
intercept=TRUE,main="red - жінка")
coefplot(model0m, add=TRUE, col.pts="blue",  intercept=TRUE,
offset=0.2)
```



Pic.1. Linear model by sex

References

Auspurg Katrin, Hinz Thomas (2015). Factorial Survey Experiment. –Sage, Series: Quantitative Applications in the Social Survey .- vol.175, p.143

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

THE TIME BUDGET SURVEY IN BELARUS: METHODOLOGY AND RESULTS

Natallia Bandarenka

Belarusian State University, Belarus
e-mail: bondnata@mail.ru

Abstract

The paper considers the problem of the time budget sample surveys conducted in the Republic of Belarus. The author describes design and features of the sample survey of time budget in Belarus.

The paper has the next parts:

1) history of development the time budget sample surveys in Belarus;
2) the time budget sample surveys design;
3) the main tools using for the time budget sample surveys;
4) a number of problems which are common to surveys;
5) the main results of the time budget survey in 2015.

Keywords: sample survey, sampling design, the questionnaires, social state policy.

EDGEWORTH APPROXIMATIONS TO DISTRIBUTION OF MEDIAN IN STRATIFIED SAMPLES

Andrius Čiginas

Statistics Lithuania, Lithuania
e-mail: andrius.ciginas@stat.gov.lt

Abstract

We consider an Edgeworth type approximation to the distribution function of sample median in the case of stratified samples drawn without replacement. We give an explicit expression of this approximation, and also its empirical version based on bootstrap. We compare their accuracy with that of the normal approximation and the bootstrap approximation in a simulation study.

Formulation of the problem

Consider a population $\mathcal{X} = \{x_1, \dots, x_N\}$ of size N. Let $\mathcal{X}$ be divided into $h \geq 1$ nonoverlapping strata $\mathcal{X} = \mathcal{X}_1 \cup \dots \cup \mathcal{X}_h$, where $\mathcal{X}_k = \{x_{k,1}, \dots, x_{k,N_k}\}$, $1 \leq k \leq h$. Evidently, $N = N_1 + \dots + N_h$. Let $\mathbb{X}_k = \{X_{k,1}, \dots, X_{k,n_k}\}$ be a simple random sample of size $n_k \leq N_k$ drawn without replacement from the stratum $\mathcal{X}_k$. We assume that the samples $\mathbb{X}_1, \dots, \mathbb{X}_h$ are independent. Write $\mathbb{X} = \mathbb{X}_1 \cup \dots \cup \mathbb{X}_h$ and denote $n = n_1 + \dots + n_h$. Denote the distribution function of the stratum k and its empirical analogue as follows:

$$F_{N,k}(x) = \frac{1}{N_k} \sum_{i=1}^{N_k} \mathbb{I}\{x_{k,i} \leq x\} \quad \text{and} \quad F_{n,k}(x) = \frac{1}{n_k} \sum_{i=1}^{n_k} \mathbb{I}\{X_{k,i} \leq x\},$$

respectively. Here $\mathbb{I}\{\cdot\}$ is the indicator function. Then the distribution function of the population $\mathcal{X}$ and its estimator are

$$F_N(x) = \sum_{k=1}^{h} \frac{N_k}{N} F_{N,k}(x) \quad \text{and} \quad F_n(x) = \sum_{k=1}^{h} \frac{N_k}{N} F_{n,k}(x),$$

respectively. Consider the population median defined as follows: $F_N^{-1}(0.5) = \inf\{x : F_N(x) \geq 0.5\}$. Define its estimator

$$X_{\text{med}} = F_n^{-1}(0.5) = \inf\{x : F_n(x) \geq 0.5\}.$$

Denote $\sigma^2 = \mathbf{Var}\, X_{\text{med}}$. We are interested in approximations to the distribution function

$$F_{\text{med}}(x) = \mathbf{P}\{X_{\text{med}} - \mathbf{E}\, X_{\text{med}} \leq x\sigma\}. \tag{1}$$

The asymptotic normality of the median X_{med}, under a stratified simple random sampling (STSRS) without replacement, was considered by Shao (1994), see also Gross

(1980). Here we present an Edgeworth type approximation to $F_{\mathrm{med}}(\cdot)$ and its empirical version. Our approach is based on Hoeffding's (orthogonal) decomposition

$$X_{\mathrm{med}} = \mathbf{E}\, X_{\mathrm{med}} + L + Q + R,$$

constructed by Bloznelis (2003) for general symmetric statistics based on STSRS samples drawn without replacement. Here L and Q are called linear and quadratic parts of the decomposition, and R is a remainder term. In the case of U-statistics of degree 2, where $R \equiv 0$, one-term (short) Edgeworth expansions were constructed and their second-order correctness was shown in Bloznelis (2007). Thus, we expect that, if R is negligible, those Edgeworth expansions will also approximate (1) well. In particular, we suggest to apply

$$G(x) = \Phi(x) - \frac{\lambda}{6}\Phi'(x)(x^2 - 1), \tag{2}$$

obtained in Bloznelis (2007). Here $\Phi'(x)$ denotes the derivative of the standard normal distribution function $\Phi(x)$, and $\lambda = \lambda(\mathcal{X})$ is the population characteristic, which consists of certain moments of L and Q. The Edgeworth correction term, added to the normal approximation in (2), reflects the skewness of the distribution of the sample median.

In order to apply (2) in practice, the parameter λ must be estimated from the sample or other data. In Gross (1980), for the estimation of the variance $\sigma^2 = \sigma^2(\mathcal{X})$, a convenient plug-in rule was proposed, where strata distribution functions were replaced by their corresponding empirical counterparts. However, it appears impossible to employ an analogous method for the estimation of λ. Another way is to replace λ by its jackknife estimator, see Bloznelis (2007). But it is well known that jackknife estimators often fail in the case of parameters of sample median (or other empirical quantiles). Therefore, we construct the estimator $\hat{\lambda} = \hat{\lambda}(\mathbb{X})$ based on the finite population bootstrap of Booth et al. (1994). Then the empirical Edgeworth expansion is

$$\widehat{G}(x) = \Phi(x) - \frac{\hat{\lambda}}{6}\Phi'(x)(x^2 - 1). \tag{3}$$

Our study is based on analytical calculations of the orthogonal decomposition in Čiginas (2012). Unfortunately, we are not able to evaluate theoretically the accuracies of the constructed approximations but, at the conference, we will present a numerical comparison of 'true' distribution (1) with Edgeworth expansions (2) and (3), and also with the normal and bootstrap approximations. We stress that the proposed formal approximations may be very efficient in real surveys, where we need to measure the accuracy of the sample median in small domains of a population (for some collections of strata) and where populations are highly skewed.

References

Bloznelis, M. (2003) A note on the bias and consistency of the jackknife variance estimator in stratified samples. *Statistics*, **37**, 489–504.

Bloznelis, M. (2007) Second-order and resampling approximation of finite population U-statistics based on stratified samples. *Statistics*, **41**, 321–332.

Booth, J. G., Butler, R. W., Hall, P. (1994) Bootstrap methods for finite populations. *Journal of the American Statistical Association*, **89**, 1282–1289.

Čiginas, A. (2012) *Approximations to Distributions of Linear Combinations of Order Statistics in Finite Populations*. PhD thesis, Vilnius University, Vilnius.

Gross, S. (1980) Median estimation in sample surveys. In: *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 181–184.

Shao, J. (1994) L-statistics in complex survey problems. *Annals of Statistics*, **22**, 946–967.

COMBINING INFORMATION FROM TWO SURVEYS

Miika Honkala

Statistics Finland
e-mail: miika.honkala@stat.fi

Abstract

This paper is based on my master's thesis I wrote at Statistics Finland in 2015. I introduce methods to combine information from two surveys. I consider constructing weights for a combined dataset using response propensity model. I also introduce some methods to combine estimates calculated from two datasets separately, and compare the methods. In addition, I consider a LGREG estimator and using it with different datasets. The results of my thesis are that the response propensity weighting procedure using the combined dataset is almost similar to a normal case when there is one survey and its dataset. When using bigger, combined dataset, we will get more explanatory variables to the response propensity model. The weights are thus more accurate. The best method for combining estimates is weighted mean which takes the sizes of the datasets into account. LGREG estimates calculated from bigger and smaller dataset are close to each other when the model of the LGREG estimator is same in both datasets.

1 Introduction

Nowadays, there is a lot of surveys which have same study variables in their datasets. Researchers want to improve the estimates of these common variables by using information from different datasets. In the literature, there is two approaches to combine information: combining samples and combining estimates. When combining samples, the idea is to construct new weights for the combined dataset and calculate estimates using these weights. When combining estimates, the estimates are calculated from each survey separately. The combined estimate is usually some linear combination of these estimates, for example mean. In my thesis, I constructed weights for two separate datasets and their combined dataset using response propensity model. I also examined methods to combine estimates and compared the methods. In addition, I used LGREG estimator with different datasets. I used same assisting model in each dataset and examined if LGREG estimates calculated from different datasets were close to each other.

I used two datasets of Statistics Finland in my thesis. The first of them was the dataset of Survey on work and well-being among people of foreign origin. Its shorthand is UTH in Finland. The other was the dataset of Labour market situation of migrants and their immediate descendants. This survey was ad-hoc module of the European Union labour force survey, so I call it AHM. Both surveys were carried out in 2014. The target population of UTH was people who lived Finland and had foreign origin: their both parents were born abroad. The target population of AHM was defined in almost same way. I used only people of foreign origin (defined in exactly same way as people of dataset UTH) and removed other members from the dataset AHM. Thus, I

had two datasets relating to people of foreign origin. Sample size was 4 977 in UTH and 1 472 in AHM. These datasets had common target population. Its size was 240 801.

Research questions were as follows. When combining datasets and constructing weights using response propensity model, the question was what are the differences compared to a normal case when a dataset is based on one survey? How much does we get advantage for the model if we use bigger, combined dataset? When combining estimates, the goal was to find the best method to combine estimates. What things are important when choosing the combining method? Are the sizes of the datasets relevant? When calculating LGREG estimates, the important question was if we can produce almost same LGREG estimates for the common variables of the datasets using bigger and smaller dataset.

2 Methods and main results

2.1 Notation

Consider a finite population U which has N elements. Let s be a sample of size n drawn from U. Inclusion probability of unit k is $\pi_k = P(k \in s)$. The design weight of unit k is $d_k = 1/\pi_k$. For short, $\sum_A$ means $\sum_{k \in A}$. Then $\sum_s y_k$ means sum $\sum_{k \in s} y_k$, for example. Let the set of respondents be $r \subset s$ and m is the number of respondents ($m \leq n$). The basic weight of respondent k is $w_{basic,k} = N/m$. It is same for all respondents.

We want to estimate $t_y = \sum_U y_k$, which is the total of the target variable y. Horvitz-Thompson estimator (briefly HT estimator) of the total t_y is

$$\hat{t}_y^{HT} = \sum\nolimits_s d_k y_k. \tag{1}$$

Let's use following notation: $\hat{t}_y^w$ is adjusted HT-estimator of the total t_y:

$$\hat{t}_y^w = \sum\nolimits_r w_k y_k. \tag{2}$$

The adjusted HT estimator is almost same as the HT estimator (1), but the weights of the adjusted HT estimator have been adjusted (and possibly calibrated) because of nonresponse. In addition, the adjusted weights are only for respondents.

Coefficient of variation of weights w_k is

$$CV(w_k) = \frac{s(w_k)}{\bar{w}_k}, \tag{3}$$

where $s(w_k)$ is standard deviation of weights w_k and $\bar{w}_k$ is mean of weights w_k.

Suppose that we have binary y variable which has classes $i = 0, 1$. We want to estimate $t_{y,i}$ which is total frequency of class i. LGREG estimator of the total $t_{y,i}$ is

$$\hat{t}_{y,i}^{LGREG} = \sum\nolimits_U \hat{\mu}_k + \sum\nolimits_s d_k(y_k - \hat{\mu}_k), \quad i = 0, 1, \tag{4}$$

where $\hat{\mu}_k$ is estimated probability that individual k belongs to class i. We get probabilities $\hat{\mu}_k$ using statistical model.

I will mark my datasets as follows. Let the dataset UTH be s_{uth} and the dataset AHM s_{ahm}. The combined dataset of these two datasets is s_{com}. These datasets have common target population U. Its size is $N = 240\ 801$. Sample sizes are $n_{uth} = 4\ 977$, $n_{ahm} = 1\ 472$ and $n_{com} = 6\ 449$. The sets of respondents are r_{uth} and r_{ahm}. The numbers of respondents are $m_{uth} = 3\ 262$, $m_{ahm} = 747$ and $m_{com} = 4\ 009$. The target population U is big compared to datasets s_{uth} and s_{ahm}. Therefore, it can be supposed that the datasets s_{uth} and s_{ahm} does not include same individuals. The datasets s_{uth} and s_{ahm} include several same study variables y and auxiliary variables x.

2.2 Response propensity weighting using combined dataset

Response propensity weighting is method which utilizes response propensity model in weighting. Response propensity model is binary model, usually logistic model. The dependent variable is binary responce indicator which tells if sample person is respondent or nonrespondent. The explanatory variables of the model are auxiliary variables, for example age group, sex, region and civil status.

Constructing weights using response propensity model, similarly as Laaksonen (2013, 128-129), includes following steps (when there is one survey and its dataset):

1) Choose starting weights w_k. They must be constructed for the respondents, so design weights d_k are not useful. Basic weights or post-stratification weights are useful, for example.

2) Create response indicator (for example 1=respondent, 0=nonrespondent).

3) Construct response propensity model using good link function (logit, probit, log-log or clog-log). The good model includes as much as possible statistically significant explanatory variables for response. When you have found "the best" model, use its response probabilities $\hat{\mu}_k$.

4) Calculate new adjusted weights as follows:

$$w_{adj,k} = \frac{w_k}{\hat{\mu}_k} q. \tag{5}$$

q is scale factor which ensures that sum of weights w_{adj} corresponds the population size N. The factor q can be calculated as

$$q = \frac{\sum_r w_k}{\sum_r (w_k/\hat{\mu}_k)}. \tag{6}$$

I constructed weights $w_{adj,k}$ for the datasets s_{uth}, s_{ahm} and their combined dataset s_{com}. Constructing weights procedure for the combined dataset was almost identical as

constructing weights for s_{uth} and s_{ahm}. The only difference was in step 1. When combining samples, the combined dataset does not include any weights w_k which satisfies $\sum_r w_k = N$. I had to calculate new starting weights w_k so that their sum corresponded the size of population. Otherwise the weighting procedures using s_{uth}, s_{ahm} or s_{com} were similar.

Response propensity model for s_{com} included 13 explanatory variables, whereas model for s_{uth} included 11 and model for s_{ahm} 8 explanatory variables. Combined dataset is always bigger compared to original datasets, so we get more explanatory variables. Then the adjusted weights $w_{adj,k}$ take missingness better into account and are thus more accurate.

The dataset s_{uth} were big (sample size 4 977) compared to s_{ahm} (sample size 1 472). The sample size of s_{com} was thus 6 449. Therefore, the number of explanatory variables in the combined dataset (13) was only little bigger than it is in s_{uth} (11). If we combine two datasets which have same size, the combined dataset is two times bigger than the original datasets. Then the response propensity model for the combined dataset includes much more explanatory variables compared to models which are constructed for the original datasets.

For estimation, I calibrated these weights $w_{adj,k}$ using three auxiliary variables: sex, age group and region. I made the calibration for these three datasets:s_{uth}, s_{ahm} and s_{com}. I made the three calibrations separately. The result was that I had calibrated weights $w_{cal,k}$ in each three datasets. The weights $w_{cal,k}$ were as close the starting weights $w_{adj,k}$ as possible and their distributions by sex, age group and region were similar to population U. Laaksonen (2013, 129) has used this method where the weights constructed with response propensity model are the starting weights of calibration. Laaksonen calls this method combination of response propensity model and calibration.

2.3 Comparing combined estimates

I calculated estimates of totals for study variables (y variables) which were same in the datasets s_{uth} and s_{ahm}. For example, I estimated the total of employed and the total of unemployed people in the target population. I used five study variables. All of them were classified and these five variables had 21 classes in total. Thus, I estimated 21 totals using three datasets: s_{uth}, s_{ahm} and s_{com}.

Total estimators I used were adjusted HT-estimators like (2). The weights of the adjusted HT estimators were $w_{cal,k}$ which I constructed as I introduced in section 2.2, first using response propensity weighting and after that calibration. Total estimators were thus

$$\hat{t}^w_{y,uth} = \sum\nolimits_{r_{uth}} w_{cal,k} y_k, \tag{7}$$

$$\hat{t}^w_{y,ahm} = \sum\nolimits_{r_{ahm}} w_{cal,k} y_k, \tag{8}$$

where $\hat{t}_{y,uth}$ means estimated total of variable y calculated from r_{uth}. $\hat{t}_{y,ahm}$ means estimated total of variable y calculated from r_{ahm}, respectively. I combined estimates calculated from datasets r_{uth} and r_{ahm} using four methods:

Method M1. Simple mean of the estimates. The combined estimate of the total t_y is

$$\hat{t}_{y,c} = \frac{\hat{t}^w_{y,uth} + \hat{t}^w_{y,ahm}}{2} = 0,5\hat{t}^w_{y,uth} + 0,5\hat{t}^w_{y,ahm}. \tag{9}$$

Method M2. Weighted mean which takes the sizes of the datasets into account. Then the combined estimate is calculated as

$$\hat{t}_{y,c} = \frac{n_{uth}\hat{t}^w_{y,uth} + n_{ahm}\hat{t}^w_{y,ahm}}{n_{uth} + n_{ahm}} \approx 0,77\hat{t}^w_{y,uth} + 0,23\hat{t}^w_{y,ahm}. \tag{10}$$

I used accurate values in calculations, but there is approximation in equation (10) so that the method M2 can easily be compared to the other methods.

Method M3. Weighted mean which takes number of respondents into account. Combined estimate is

$$\hat{t}_{y,c} = \frac{m_{uth}\hat{t}^w_{uth} + m_{ahm}\hat{t}^w_{ahm}}{m_{uth} + m_{ahm}} \approx 0,81\hat{t}^w_{uth} + 0,19\hat{t}^w_{ahm}. \tag{11}$$

Response rate in UTH was better than response rate in AHM. Therefore, when using the method M3, the factor of UTH estimate (0,81) is higher than it is in the method M2 (0,77). If the response rates in UTH and AHM had been similar, the factors in the methods M2 and M3 would have been identical. In this case, the response rates in UTH and AHM were different, so I wanted to try out the both methods M2 and M3.

Method M4. Weighted mean as follows:

$$\hat{t}_c = \lambda\hat{t}^w_{uth} + (1-\lambda)\,\hat{t}^w_{ahm}, \tag{12}$$

where

$$\lambda = \frac{m_{uth}/deff_{uth}}{m_{uth}/deff_{uth} + m_{ahm}/deff_{ahm}}. \tag{13}$$

$deff_{uth}$ and $deff_{ahm}$ are design effects related to datasets s_{uth} and s_{ahm}. This method is almost same as the method that O'Muircheartaigh and Pedlow (2002) have used. The difference is that I used the numbers of respondents in equation (13) instead of sample sizes. As O'Muircheartaigh and Pedlow say, it is inconvenient to use the design effects themselves, since they are different with each variable. They have used design effects which do not depend on variables. I used such design effects:

$$deff_{uth} = 1 + [CV(w_{cal,k})]^2 = 1 + \left(\frac{s(w_{cal,k})}{\bar{w}_{cal,k}}\right)^2, k \in r_{uth} \tag{14}$$

and

$$deff_{ahm} = 1 + [CV(w_{cal,k})]^2 = 1 + \left(\frac{s(w_{cal,k})}{\bar{w}_{cal,k}}\right)^2, k \in r_{ahm} \quad (15)$$

$CV(w_{cal,k})$ means coefficient of variation of weights $w_{cal,k}$ in r_{uth} and r_{ahm}. When we set (14) and (15) to equation (13), we get $\lambda \approx 0,84$. Thus, the combined estimate (12) takes a form

$$\hat{t}_{y,c} \approx 0,84\hat{t}^w_{y,uth} + 0,16\hat{t}^w_{y,ahm}. \quad (16)$$

The factor of $\hat{t}_{uth}$ is the highest in the method M4 (0,84) and the lowest in the method M1 (0,5).

I calculated estimates of 21 class frequencies of five target variables using datasets s_{uth} and s_{ahm}. I combined these 21 estimates using the methods M1, M2, M3 and M4. In addition, I calculated same estimates using combined dataset s_{com}. We can suppose that the estimates calculated from s_{com} are the most accurate because s_{com} is the biggest dataset. In addition, its weights are the most accurate for the same reason. Therefore, I compared the combined estimates to estimates calculated from s_{com}. I compared how much the combined estimates differed from estimates which were calculated from the combined dataset. I tried to find out which method of the methods M1-M4 produce estimates which are the most similar compared to estimates calculated from the combined dataset.

Table 1. Comparison of the methods M1-M4 - results.

Method	s_{uth}	s_{ahm}	M1	M2	M3	M4
Deviations in total	22 142	67 490	25 140	7 503	9 162	10 598
Mean deviation	1 054	3 214	1 197	357	436	505
Closest to the estimate of s_{com}	4	0	2	13	1	1

The results of the methods M1-M4 are in table 1. Deviations in total tell how much 21 estimates differ in total from estimates calculated from the combined dataset. The best method were M2 (weighted mean which takes sample sizes into account). This method produced estimates which were most similar compared to estimates calculated from the combined dataset. Using M2, deviations in total were 7 503. The second best method was M3 (weighted mean which takes number of respondents into account). Its deviations in total were 9 162.

The worst of the methods M1-M4 was M1 (simple mean). Its deviations in total were 25 140. For a comparison, we see in table 1 that using only dataset s_{uth}, deviations in total were 22 142. Using only dataset s_{uth}, the estimates were closer to estimates calculated from the combined sample, than estimates using the method M1. If we have big and small datasets, simple mean can be poor alternative. It is then better to use other methods.

When using the method M4, the factor of $\hat{t}_{y,uth}$ 0,84 were the highest, but this method was not until third best. The methods M2 and M3 were better although their factors of $\hat{t}_{y,uth}$ were lower (0,77 and 0,81). When we have big and small datasets, the factor relating the estimate of the big dataset has to be suitable. It must not be too high or too low. We must also take the estimates of the small dataset into account and give them appropriate factor. When we have big and small dataset, the methods M2 and M3 are good.

These results are useful when we have big and small datasets. If we had two datasets with same sizes, the results could be different. For example, simple mean (M1) could then be better than it was in this case.

2.4 LGREG estimator using the dataset UTH and the combined dataset

Logistic generalized regression estimator (briefly LGREG estimator) is estimator which can be used with classified variables when estimating class frequencies. Lehtonen and Veijanen (1998) has considered LGREG estimator very extensively. The LGREG estimator is design-based model-assisted estimator. It utilizes additional information (auxiliary variables) with a statistical model. If study variable y is binary, the LGREG estimator can be calculated using equation (4).

The model for LGREG is similar to the responsive propensity model, which has introduced in section 2.2. In the LGREG model, the dependent variable is study variable y we are interested in. The explanatory variables are auxiliary variables. When using the LGREG estimator, we need information of the respondents in sample s. In addition, we need a dataset which covers the whole population U. This dataset must contain all auxiliary variables we use in the LGREG model. Using the LGREG model, we can construct estimated probabilities $\hat{\mu}_k = P\left(y_k = i\right)$ for all individuals in population U. The LGREG estimator utilizes y-values from respondents in sample s and estimated probabilities $\hat{\mu}_k$ which are for whole population U.

I calculated LGREG estimates using datasets s_{uth} and combined dataset s_{com}. I used binary dependent variables. First I used employment variable (1=employed, 0=not employed). I constructed logistic regression model in datasets s_{uth} and s_{com}. The explanatory variables were same in both datasets. They were sex, age group, country where the parents of the respondents were born and employment status according to job seeker register. The aim was to study how similar the LGREG estimates calculated using s_{uth} and s_{com} were, when assisting model was same. The numbers of respondents were: $m_{uth} = 3\ 262$ and $m_{com} = 4\ 009$. The set of respondents r_{com} included all 3 262 respondents in r_{uth} because $r_{uth} \subset r_{com}$.

Table 2. LGREG estimates calculated from datasets s_{uth} and s_{com}.

Dataset	Number of employed persons	Standard deviation	95 % CI
s_{uth}	140 805	1 690	137 492 - 144 118
s_{com}	140 595	1 558	137 542 - 143 648

The results are in table 2. The estimates are almost same: the difference is only 210 persons. The difference is about 0,15 % of the sizes of the estimates, so the difference is very small.

I calculated also LGREG estimates of the number of unemployed persons, using datasets s_{uth} and s_{com}. I used same four explanatory variables in the logistic regression model as I used with employment variable. LGREG estimates of the number of unemployed persons were 31 033 when using s_{uth} and 30 892 when using s_{com}. The difference of the estimates was only 141.

If we use a good estimator like LGREG estimator, we can get very similar results using datasets which have different sizes. Although dataset s_{uth} had 747 respondents less than s_{com}, LGREG estimates were almost same. Therefore, we can in a sense compensate missingness by using a good estimator. Using the LGREG estimator is however a little inconvenient because each study variable needs usually own model with certain explanatory variables. For example, if we have 100 study variables, we have a lot of work with models. It is however useful to calculate LGREG estimates at least for the most important study variables.

References

Laaksonen, S. (2013). *Surveymetodiikka: Aineiston kokoamisesta puhdistamisen kautta analyysiin.* 2. edition. Online book.

Lehtonen, R. and Veijanen, A. (1998). Logistic Generalized Regression Estimators. *Survey Methodology*, **24**, 51-55.

O'Muircheartaigh, C. and Pedlow, S. (2002). Combining samples vs. cumulating cases: a comparison of two weighting strategies in NLSY97. *ASA Proceedings of the Joint Statistical Meetings*, 2557-2562.

ANALYZING DIFFERENT ALLOCATIONS IN STRATIFIED SAMPLING

Nestor Hrabets[1] and Tetiana Ianevych[2]

[1] Taras Shevchenko National University of Kyiv, Ukraine
e-mail: nestorhrabets@ukr.net

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

Abstract

In this paper we examine different allocation for stratified simple random sampling taking into account the problem of obtaining equally precise estimates within each stratum. Some practical results have been obtained for the sample survey of capital expenditure of Ukrainian enterprises.

1 Introduction

In many surveys the objective is not only to obtaine the estimates for parameters of interest with good precision but also to have sufficiently good estimates for different domains. For example, if we want to investigate the total capital expenditure of enterprises in Ukraine we need to construct the sample in such a way in order to get good precision on the country level. But it is also very important to have the possibility to get good-quality estimetes for every region of the Ukraine or for different types of economical activities of enterprises.

Usually the stratified sampling design is used in such surveys. For more detales you may see Lehtonen and Pahkinen (2004). In stratified sampling it is important to choose sample size and allocate the sample for strata in a way to obtain accurate estimates of parameters. There are many different allocations of the stratified sample: proportional allocation, Neyman allocation etc. The Neyman allocation is optimal if the costs of the surveying of the units are equal. But usage the Neyman allocation usually leads us to domain estimates with quite different precision. For some domains (strata) the precision can be very good but for some it can be very bad. So, we cannot use such estimates for comparison.

Here the very important problem for practitioners appears – how to allocate the sample on strata in such a way that precision of estimates in every stratum be almost identical? Under "precision" in this work we understand coefficient of variation (CV) in strata. Wesolowski and Wieczorkowski (2015) had proposed Equal-Precision Allocation (EPA) allowing obtaining equally accurate estimates within domains. We will explore how works EPA in a particular case when it is needed to explore the capital expenditure in Ukraine. As the strata we consider the territorial units of Ukraine. We will analyze pros and cons of EPA and compare it with proportional and Neyman allocations.

2 Theoretical aspects

Consider a population U partitioned into strata (subpopulations, domains) $U_1,\ldots,U_H$ such that $U=\bigcup_{h=1}^{H}U_h$, $U_i\cap U_j=\varnothing$. Assume that we are interested in estimation of means of a

variable *y* in all strata. In each U_h a sample of n_h elements is chosen according to simple random sampling without replacement. Assume additionally that the total sample size $n = \sum_{h=1}^{H} n_h$ is fixed.

2.1 Neyman allocation

Let's suppose that the costs to survey one element in a sample $s = \bigcup_{h=1}^{H} s_h$, $s_h \subset U_h$ are identical in all strata U_h. Within stratified sampling, the optimal Neyman allocation minimizes overall variance in case when the total costs are fixed and costs in every strata are identical. It takes the form

$$n_h = n \frac{N_h S_h}{\sum_{k=1}^{H} N_k S_k}, \quad h = 1,..., H$$

,

To get optimal results we should know mean-squared deviation in strata (S_h). It is impossible in practice. But for the repeated surveys we can use information received from the previous surveys for obtaining approximate standard deviation values. Then we can obtain allocation close to optimal.

2.2 Proportional allocation

Proportional allocation is given by the relation:

$$n_h = n \frac{N_h}{N}, \quad h = 1,..., H \ ,$$

Here, it is also assumed that the sample size is fixed. That's why it always can be calculated. If standard deviations in all stratums are identical (S_1=…S_H), then the proportional allocation will be optimal Neyman allocation. In other cases the proportional allocation gives worse precision than Neyman allocation, especially when the values of S_h are very different.

2.3 Equal-Precision allocation

The objective of the Equal-precision allocation is to allocate the sample among subpopulations in such a way that the precision of the estimators in each of the subpopulations be the same. By equal-precision we mean equal CVs in subpopulations. That is we want to have

$$CV_h = \left(\frac{1}{n_h} - \frac{1}{N_h} \right) \gamma_h^2 = const = T, \quad h = 1,..., H$$

where γ_h^2 are the coefficient of variation for variable y in U_h. Expressing in terms of coefficient of variations CV_h the constraint on the size of the total sample $n = \sum_{h=1}^{H} n_h$ gives the equation

$$n = \sum_{h=1}^{H} \frac{N_h \gamma_h^2}{\gamma_h^2 + TN_h}$$

with unknown precision T. The equation above has a unique solution, which can be easily computed numerically (however no analytical explicit formula is available). Obviously, such a solution, T^* gives the desired allocation

$$n_h = \frac{N_h \gamma_h^2}{\gamma_h^2 + T^* N_h}$$

On the other hand if one imposes requirements on CV's of estimators in subpopulations, that is, when CV_h are given (not necessarily identical) there is no freedom in the sense that they determine uniquely the total sample size. If instead one assumes only the restriction that CV's of domain mean estimators are bounded from above by (possibly) different constraints, the minimization of the total sample size is a valid question. It has been solved recently (with additional constraint on the CVs of the estimator of the population mean) for stratified SRSWOR by Choudhry, Rao and Hidiroglou (2012) through a nonlinear programming Newton-Raphson procedure.

3 Practical results

We consider as a population the enterprises participated in the first quarter survey of capital expenditure in 2010 in Ukraine. So we have N=19087 enterprises. Total sample consists of n=5000 enterprises. The strata are the territorial units of Ukraine (H=27). We estimate mean value of the variable y – capital expenditures. In each stratum U_h a sample of n_h elements should be chosen according to simple random sampling without replacement (SRSWOR).

CVs in each stratum is given by the formula

$$CV_h = \frac{\sqrt{Var_h\left(\hat{\bar{y}}_{h\pi}\right)}}{\bar{y}}$$

where $Var_h\left(\hat{\bar{y}}_{h\pi}\right)$ is a variance of Horwitz-Thompson estimator of y in the stratum h.

The values of CV's for Neyman, Proportional and Equal-Precision allocations are placed into the Table 1 below

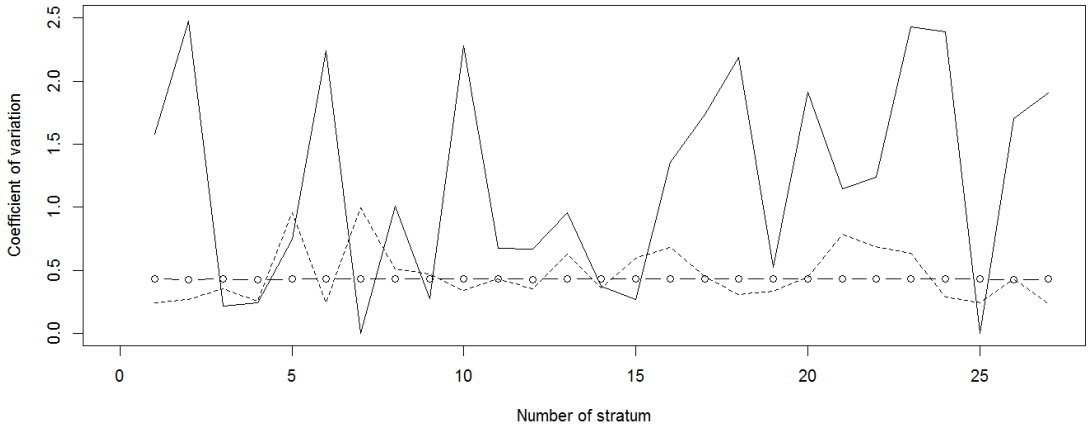
Table 1 *Coefficients of Variation for regional domains*

CV(%)	Neyman allocation	Proportional allocation	Equal-precision allocation
Autonomous Republic of Crimea	157.70	24.41	42.83
Vinnytsya Region	190.65	23.19	42.86
Volyn Region	247.63	26.84	42.78
Dnipropetrovsk Region	21.64	35.06	42.80
Donetsk Region	24.32	25.59	42.73
Zhytomyr Region	74.46	95.62	42.81
Zakarpattya Region	223.97	24.44	42.91
Zaporizhzhya Region	0.00	99.69	42.82
Ivano-Frankivsk Region	101.02	51.18	42.80
Kyiv Region	27.09	47.08	42.83
Kirovograd Region	227.79	34.27	42.83
Luhansk Region	67.07	43.05	42.90
Lviv Region	66.57	35.39	42.79
Mykolayiv Region	95.30	63.10	42.90
Odesa Region	37.31	35.94	42.83
Poltava Region	26.99	59.78	42.90
Rivne Region	135.29	67.86	42.90
Sumy Region	173.17	44.84	42.90
Ternopil Region	218.39	30.89	42.89
Kharkiv Region	52.43	33.14	42.86
Kherson Region	191.38	44.68	42.81
Khmelnytskyy Region	114.55	78.38	42.90
Cherkassy Region	123.94	68.46	42.87
Chernivtsi Region	242.60	63.53	42.88
Chernihiv Region	238.97	28.50	42.83
Kyiv City	0.00	24.25	42.88
Sevastopol City	169.91	43.02	42.20

This data is visualized on the graphic below. The solid line stands for Neyman allocation, the dashed line – for proportional allocation and the dotted line – for equal-precision allocation.

As we see on the graphic, CV's corresponding to equal-precision allocation are about the same and mostly much smaller than CV's received by Neyman allocation and proportional allocation.

Picture 1



As for CV's for total sample included to the Table 2 below, we can see that EPA gives the larger total CV comparing to the other allocations.

Table 2 *Total Coefficients of Variation*

	Neyman allocation	Proportional allocation	Equal-precision allocation
CV (%)	9.87	11.35	13.69

Conclusions

If we need to obtain equal-precision in every strata than EPA produces the needed sample sizes and the accuracy in every strata is equally good. But if we need to get the highest precision for total sample it is better to use Neyman allocation.

References

Choudhry, G.H., Rao, J.N.K and Hidiroglou, M.A. (2012) On sample allocation for efficient domain estimation. *Survey Methodology*, *Vol. 38, No. 1, pp. 23-29.*

Lehtonen, R and Pahkinen, E. (2004). *Practical Methods for Design and Analysis of Complex Surveys*, 2 edition/John Wiley & Sons Ltd.

Wesolowski, J and Wieczorkowski, R. (2015) An eigenproblem approach to optimal equal-precision sample allocation in subpopulations. *arXiv:1503.08686v1*.

DETECTION OF STRUCTURAL CHANGE IN LINEAR REGRESSION MODEL BY USING AN R-PACKAGE "STRUCCHANGE"

Sofia Lishnianska

Taras Shevchenko National University of Kiev, Ukraine
e-mail: lishnianska.sofia@gmail.com

Abstract

Testing and analyzing structural change in econometric models is a very active research area. For the last decades there were presented several theoretical results like generalized fluctuation test framework (Kuan and Hornik 1995) on the one hand and tests based on F statistics (Hansen 1992; Andrews 1993; Andrews and Ploberger 1994) on the other. This contributed paper contains a short overview of implementing two main classes of tests on structural changes by an R-package "strucchange" and its applying to macroeconomics data series of Ukraine.

1 Introduction

Structural change is a very important interest in many fields of research and data analysis: to learn if, when and how the structure of the data generating mechanism underlying a set of observations changes. Usually, it is known with respect to which quantity the structural change might occur, i.e. overtime or with the increase of a certain risk factor. But to assess whether there is evidence for such a structural change or not, we need a statistical test: given a model and it's tested whether the data support the hypothesis that there is a stable structure against the alternative that it changes over time.

2 The model

Consider the standard linear regression model

$$y_i = x_i^T \beta_i + u_i \tag{1}$$

where at time i, y_i is the observation of the dependent variable, $x_i = (1, x_{i2}, \dots, y_i)$ is a k×1 vector of observations of the independent variables, with the first component equal to unity, u_i are iid $(0, \sigma^2)$, and β_i is the k×1 vector of regression coefficients. Tests on structural change are concerned with testing the null hypothesis of "no structural change"

$$H_0 : \beta_i = \beta_0, (i=1, \dots, n) \tag{2}$$

against the alternative that the coefficient vector varies over time, with certain tests being more or less suitable (i.e., having good or poor power) for certain patterns of deviation from the null hypothesis. It is assumed that the regressors are nonstochastic with $\|x_i\| = O(1)$ and that

$$\frac{1}{n}\sum_{i=1}^{n} x_i x_i^T \to Q \tag{3}$$

for some finite regular matrix Q. These are strict regularity conditions excluding trends in the data which are assumed for simplicity. In what follows $\widehat{\beta^{(i,j)}}$is the ordinary least squares (OLS) estimate of the regression coefficients based on the observations i + 1,...,i + j, and $\widehat{\beta^{(i)}}$ =$\widehat{\beta^{(0,i)}}$ is the OLS estimate based on all observations up to i. Hence $\widehat{\beta^{(n)}}$is the common OLS estimate in the linear regression model.

3 The data

The sample of data that used for examples in this paper are macroeconomic time series from Ukraine. This data set contains the aggregated monthly personal income and personal consumption expenditures (in millions UAH) between January 2006 and December 2015. It was originally taken from the site of State Statistics Service of Ukraine: http://www.ukrstat.gov.ua

For the consumption function it could be used a simple error correction model (ECM) (Hansen 1992b):

$$\Delta c_t = \beta_1 + \beta_2 e_{t-1} + \beta_3 \Delta i_t + u_t \tag{4}$$

$$e_t = c_t - \alpha_1 - \alpha_2 i_t \tag{5}$$

where c_t is the consumption expenditure and i_t the income. In this case we estimate the cointegration equation (5) by OLS and use the residuals $\widehat{e_t}$ as regressors in equation (4), in which we will test for structural change. Thus, the dependent variable is the increase in expenditure and the regressors are the cointegration residuals and the increments of income (and a constant).

4 Generalized fluctuation tests

The class of generalized fluctuation tests (in particular the CUSUM and MOSUM tests) fit a model to the given data and derive an empirical process, that captures the fluctuation either in residuals or in estimates. The limiting process for these empirical processes are known and so that boundaries can be computed, whose crossing probability under the null hypothesis is α. If path of the empirical process crosses its boundaries. The null hypothesis should be rejected at significance level α.

In an R-package “strucchange” exists function efp, that creates an object of class "efp" which contains a fitted empirical fluctuation process of a specified type. The process itself is of class "ts" (the basic time series class in R), which either preserves the time properties of the dependent variable if this is a time series, or which is standardized to the interval [0,1]. In this section, there will be a short overview of these types, in particular CUSUM and MOSUM.

4.1. CUSUM processes:

The first type of processes that can be computed are CUSUM processes, which contain cumulative sums of standardized residuals. According to Zeileis, Leisch, Hornik, and Kleiber (2002), let's consider cumulative sums of recursive residuals:

$$W_n(t) = \frac{1}{\tilde{\sigma}\sqrt{\eta}}\sum_{i=k+1}^{k+\lfloor t_\eta \rfloor} \tilde{u}_i \quad (0 \leq t \leq 1) \tag{6}$$

Where and $\eta = n-k$ is the number of recursive residuals and $\lfloor t\eta \rfloor$ is the integer part of $t\eta$. Under the null hypothesis the limiting process for the empirical fluctuation process $W_n(t)$is the Wiener Process W(t). More precisely the following functional central limit theorem (FCLT) holds:

$$W_n(t) \Longrightarrow \mathrm{W} \tag{7}$$

as $n \to \infty$.

4.2 MOSUM processes:

Another possibility to detect a structural change is to analyze moving sums of residuals (instead of using cumulative sums of the same residuals). The resulting empirical fluctuation process does then not contain the sum of all residuals up to a certain time t but the sum of a fixed number of residuals in a data window whose size is determined by the bandwidth parameter $h \in (0,1)$ and which is moved over the whole sample period. Hence the Recursive MOSUM process is defined by

$$M_n(t|h) = \frac{1}{\tilde{\sigma}\sqrt{\eta}}\sum_{i=k+1+\lfloor N_\eta t \rfloor}^{k+\lfloor N_\eta t \rfloor+\lfloor \eta h \rfloor} \tilde{u}_i \qquad (0 \leq t \leq 1\text{-h}) \tag{8}$$

$$W_n\left(\frac{\lfloor N_\eta t \rfloor+\lfloor \eta h \rfloor}{\eta}\right) - W_n^0\left(\frac{\lfloor N_\eta t \rfloor}{\eta}\right) \tag{9}$$

where $N_\eta = (\eta - \lfloor \eta h \rfloor)/(1-h)$.

Similarly the OLS-based MOSUM process is defined by

$$M_n^0(t|h) = \frac{1}{\tilde{\sigma}\sqrt{\eta}}\left(\sum_{i=1+\lfloor N_\eta t \rfloor}^{\lfloor N_\eta t \rfloor+\lfloor \eta h \rfloor} \tilde{u}_i\right) \quad (0 \leq t \leq 1-\mathrm{h}) \tag{10}$$

$$= W_n^0\left(\frac{\lfloor N_n t \rfloor+\lfloor nh \rfloor}{n}\right) - W_n^0\left(\frac{\lfloor N_n t \rfloor}{n}\right) \tag{11}$$

where $N_n = (n - \lfloor nh \rfloor)/(1-h)$. As the representations (9) and (12) suggest, the limiting process for the empirical MOSUM processes are the increments of a Brownian motion or a Brownian bridge respectively. This is shown in detail in Chu, Hornik, and Kuan (1995). If a single structural shift is assumed at t_0, then both MOSUM paths should also have a strong shift around t_0.

4.3 Estimates-based processes

Fluctuation processes also can be based on estimates of the unknown regression coefficients. Generally, the technique is quite similar to CUSUM and MOSUM-type processes: the vector β (k×1) is estimated recursively with a growing number of observations or with a moving data window of constant bandwidth h and then compared to the estimates based on the whole sample. According to Ploberger, Kramer, and Kontrus (1989), the fluctuation process is defined by:

$$Y_n(t) = \frac{\sqrt{i}}{\hat{\sigma}\sqrt{n}}\left(X^{(i)T}X^{(i)}\right)^{1/2}\left(\widehat{\beta^{(i)}} - \widehat{\beta^{(n)}}\right) \tag{12}$$

Where $i = \lfloor k + t(n-k) \rfloor$ and $t \in [0,1]$. The second approach gives the moving estimates (ME) process which was described by Chu, Hornik, and Kuan (1995):

$$Z_n(t|h) = \frac{\sqrt{\lfloor nh \rfloor}}{\hat{\sigma}\sqrt{n}} \left(X^{(\lfloor nh \rfloor, \lfloor nh \rfloor)T} X^{(\lfloor nh \rfloor, \lfloor nh \rfloor)} \right)^{1/2} (\beta^{(\widehat{\lfloor nh \rfloor, \lfloor nh \rfloor})} - \beta^{\widehat{(n)}}) \tag{13}$$

where 0 ≤ t ≤ 1−h. Both are k-dimensional empirical processes. Under a single shift alternative the recursive estimates processes should have a peak and the moving estimates process should again have a shift close to the shift point t_0.

4.4 Boundaries

The common property to all generalized fluctuation tests is that when the fluctuation of the empirical process efp(t) gets improbably large compared to the fluctuation of the limiting process, then the null hypothesis of "no structural change" should be rejected. In case of the one-dimensional residual-based processes this comparison is performed by some appropriate boundary b(t), that the limiting process just crosses with a given probability α. Thus, if efp(t) crosses either b(t) or −b(t) for any t then it has to be concluded that the fluctuation is improbably large and the null hypothesis can be rejected at confidence level α. According to Chu, Hornik, Kuan (1995), both limiting CUSUM processes, the Brownian motion and the Brownian bridge, are not stationary. It would seem natural to use boundaries that are proportional to the standard deviation function of the corresponding theoretic process, i.e.:

$$b(t) = \lambda\sqrt{t} \tag{14}$$

$$b(t) = \lambda\sqrt{t(1-t)} \tag{15}$$

for the Recursive CUSUM and the OLS-based CUSUM path respectively, where λ determines the confidence level. But the boundaries that are commonly used are linear, because a closed form solution for the crossing probability is known. So the standard boundaries for the two processes are

$$b(t) = \lambda\,(1+2t) \tag{16}$$

$$b(t) = \lambda \tag{17}$$

They were chosen because they are tangential to the boundaries (16) and (17) respectively in t = 0.5.

The situation for the MOSUM processes is different though. For example, the boundaries for the MOSUM processes are constants, i.e., of form b(t) = λ, which seems natural as the limiting processes are stationary. Given a fitted empirical fluctuation process the boundaries can be computed very easily using the function boundary, which returns a time series object with the same time properties as the given fluctuation process.

5 F-tests

In this section there will be a short overview of different approach to determination of structural changes by using F-tests. According to Zeileis (2000), the important di□erence is that the alternative is specified: whereas the generalized fluctuation tests are suitable for various patterns of structural changes, the F tests are designed to test against a single shift alternative. Thus, the alternative can be formulated on the basis of the model (1):

$$\beta_i = \begin{cases} \beta_A \ (1 \le i \le i_0) \\ \beta_B \ (i_0 \le i \le n) \end{cases} \tag{18}$$

where i_0 is some change point in the interval (k,n−k). Chow (1960) was the first to suggest such a test on structural change for the case where the (potential) change point i_0 is known. He proposed to fit two separate regressions for the two subsamples defined by i_0 and to reject whenever

$$F_{i_0} = \frac{\hat{u}^T\hat{u} - \hat{e}^T\hat{e}}{\hat{e}^T\hat{e}/(n-2k)} \tag{19}$$

is too large, where $\hat{e}=(\hat{u}_A, \hat{u}_B)^T$ are the residuals from the full model, where the coe□cients in the subsamples are estimated separately, and $\hat{u}$ are the residuals from the restricted model, where the parameters are just fitted once for all observations. The test statistic F_{i_0} has an asymptotic $\chi 2$ distribution with k degrees of freedom and (under the assumption of normality). The major drawback of this "Chow test" is that the change point has to be known in advance, but there are tests based upon F statistics (Chow statistics), that do not require a specification of a particular change point.

5.1 F-statistics: function fstats

A natural idea to extend the ideas from the Chow test is to calculate the F statistics for all potential change points or for all potential change points in an interval $[\underline{i}, \bar{i}]$ and to reject if any of those statistics get too large. Therefore the first step is to compute the F statistics F_i for k< $\underline{i} \leq i \leq \bar{i} < n-k$, which can be easily done using the function Fstats. Again the model to be tested is specified by a formula interface and the parameters i and ı are respresented by from and to, respectively.

6 Practical results

After an analyzing the data set and plots, I made following conclusions:

1) According to both of these plots, we can see a significant shift between Q3 2013 and Q4 2015, that has a peak around Q1-Q2 2014. Nevertheless, OLS-based CUSUM and MOSUM paths don't exceed the boundaries (see Figure 1). This fact gives us a reason to move to the next test – moving estimates test, which is more precise.

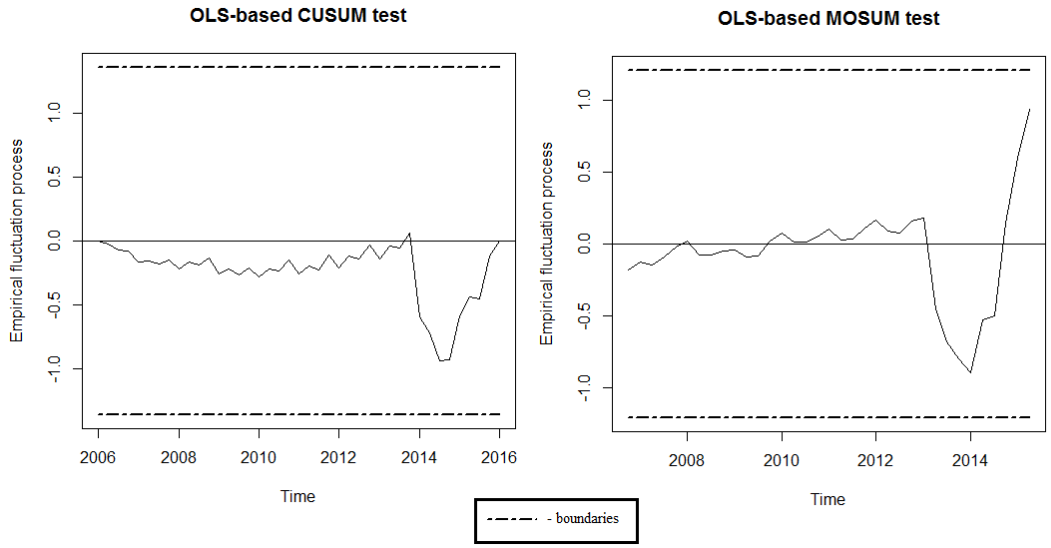


Figure 1: OLS-based CUSUM and MOSUM tests

2) According to plot, It can be seen that the moving estimates process exceed the boundary in Q1 2014. (See Figure 2); hence there is evidence for a structural change. To provide a little more information about the nature of the structural change, let's apply a 3-dimensional ME test. The output from the Figure 3, is next: we can see three parts of the plot show the processes that correspond to the estimate of the regression coe□cients of the intercept, the cointegration residuals and the increments of income, respectively. All three paths show one shift, that starts in about Q3 2013 and ends in the very end of a sample period. The shift that causes the significance seems to be the strong first shift in the process for the intercept, because the path cross its boundaries. In this case, the ME test leads to different results as the OLS-based CUSUM test.

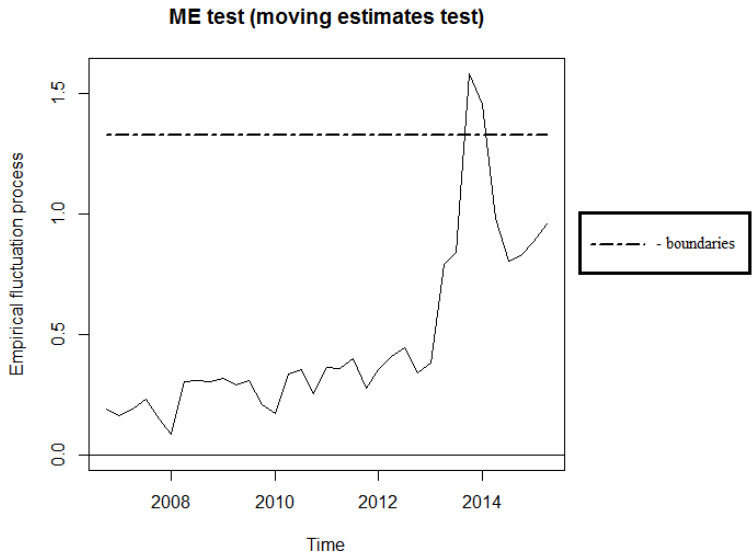


Figure 2: One-dimensional moving estimates test

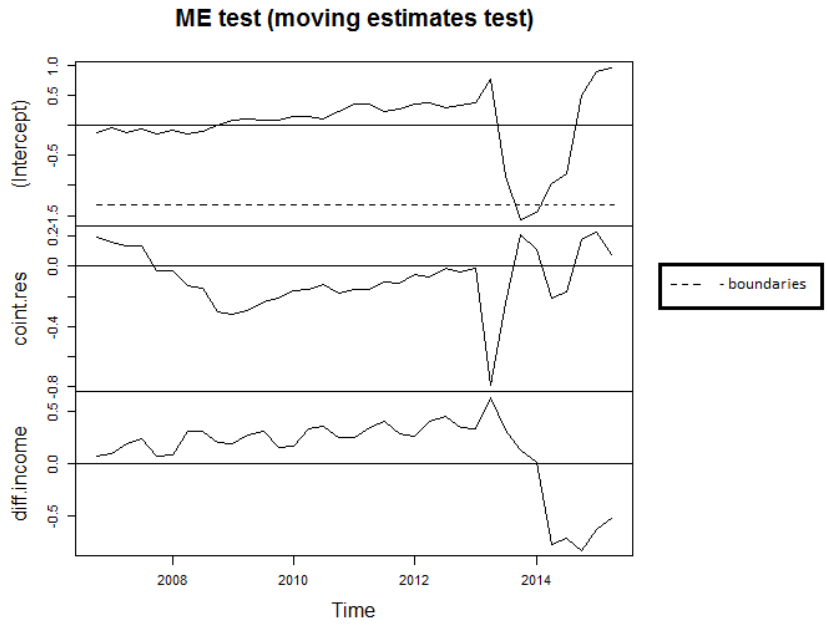


Figure 3: 3-dimensional moving estimates test

3) As the F statistics cross their boundary, there is evidence for a structural change (at the level $\alpha = 0.05$). The process has a clear peak in 2014, which mirrors the results from the analysis by empirical fluctuation processes and tests, respectively, that also indicated a break between Q4 2013 and Q3 2014.

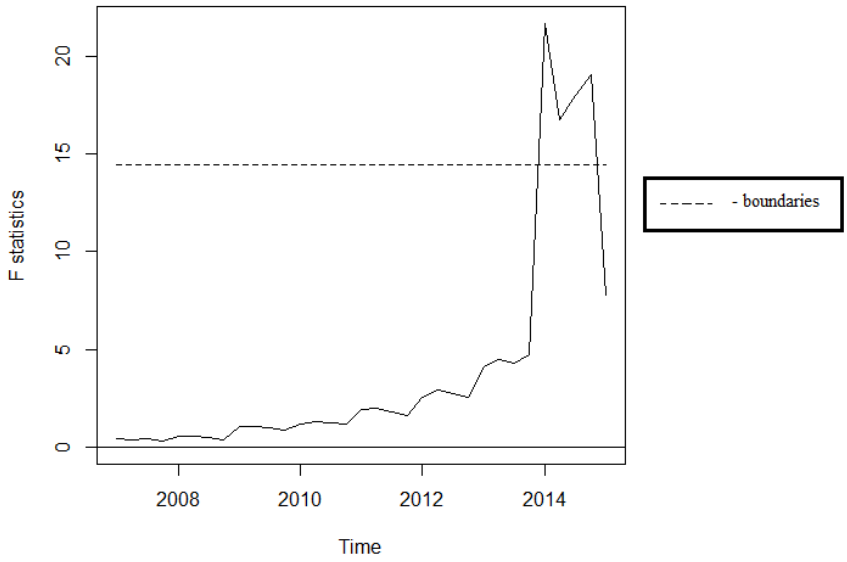


Figure 4: Using of function fstats

Conclusion

The result of analysis shows that between Q4 2013 and Q4 2015 there were structural changes in our linear regression model. These facts could be interpreted as shifts in "personal expenditure-income" model. Of course this situation was caused by political, economic and monetary policy crisis. Following fluctuations of the exchange rate of UAH and series of wrong decisions of Central Bank caused a very rapid increase of inflation rate and as a reason another wave of crisis appeared. Nevertheless, during Q3-Q4 2015 and Q1 2016 there was a trend of a stabilization of macroeconomic situation in Ukraine.

References

D. W. K. Andrews. (1993) Tests for parameter instability and structural change with unknown change point. Econometrica, **61**,821–856.

G. C. Chow. (1960) Tests of equality between sets of coe□cients in two linear regressions. Econometrica, **28**,591–605.

C.-S. J. Chu, K. Hornik, and C.-M. Kuan (1995) MOSUM tests for parameter constancy. Biometrika, **82**, 603–617.

C.-S. J. Chu, K. Hornik, and C.-M. Kuan. (1995) The moving-estimates test for parameter stability. Econometric Theory, **11**, 669–720.

B. E. Hansen. (1992) Testing for parameter instability in linear models. Journal of Policy Modeling, **14**, 517–533.

C.-M. Kuan and K. Hornik. (1995) The generalized fluctuation test: A unifying view. Econometric Reviews, **14**, 135–161.

F. Leisch, K. Hornik, and C.-M. Kuan. (2000) Monitoring structural changes with the generalized fluctuation test. Econometric Theory, **16**, 835–854.

A. Zeileis, F. Leisch, K. Hornik, and C. Kleiber (2002) Strucchange: An R package for testing for structural change in linear regression models. Journal of Statistical Software, **7(2)**, 1–38.

SURVEY OF THE INTERNALLY DISPLACED PERSON'S CONDITIONS

Tetiana Lukovych[1]

[1] Taras Shevchenko National University of Kyiv, Ukraine
e-mail: tan_luk@ukr.net

Abstract

In the light of the situation in Ukraine, the new kind of migration people appeared. This group is classified as Internally Displaced Person (IDPs). The main aim of the Survey is to get complex information about IDPs's condition. This paper investigates the estimates of sampling errors.
Keywords: Household survey, estimates of sampling errors.

1 Introduction

The IDPs Survey was organized by International Organization for Migration and was implemented by Ukrainian Center for Social Reforms. By reason of special situation in Eastern Ukraine and Crimea the appreciable changes of population distribution had been occurred. According to the information produced by the Ministry of Social Policy the total number of IDPs in Ukraine is over 1.7 million. Now housing, funding, employment, access to medical care, social services are not all problems facing the IDPs, so investigations IDPs's conditions is very important.

2 IDPs Survey

2.1 Sampling design and design weight

The sample for this Survey was designed to ensure reliable estimates of main survey indicators for study domains: Ukraine as whole, urban and rural areas at the national level, and the following regions formed on the based of the distance from the conflict zone:

- Region 1 includes Donetsk Region, Luhansk Region, Dnipropetrovsk Region, Kharkiv Region, Zaporizhzhya Region;
- Region 2 includes Cherkassy Region, Kherson Region, Kirovograd Region, Mykolayiv Region, Poltava Region, Sumy Region;
- Region 3 includes Chernihiv Region, Kyiv City, Kyiv Region, Odesa Region, Vinnytsya Region, Zhytomyr Region;
- Region 4 includes Chernivtsi Region, Ivano-Frankivsk Region, Khmelnytskyy Region, Lviv Region, Rivne Region, Ternopil Region, Volyn Region, Zakarpattya Region.

The Survey objects are households with IDPs in 24 administrative regions of Ukraine and Kyiv city, where they are currently living.

Survey subjects are the average household size; the percent of household with children; average income per household member; percent of employed IDPs; the distribution of IDPs by type of accommodation they have: rented accommodation, hostel/collective accommodation center, relatives/host family. Also the problems concerning IDPs's relocation, future resettlement or returning home were investigated.

The Survey was carried out in four rounds, one per month. The main Survey aim is to get complex information about IDPs's condition.

The main information and datasets sources are:

- administrative data, in particular, from Ministry of Social Policy of Ukraine;
- data from the sample households interview survey;
- data from key informant interviews;
- data from focus group in which participants are key informants and IDPs;
- available relevant information from other sources, for instance information about IDPs's current place of residence.

The target population for the survey was defined as all registered IDPs until December 2015. The target population was stratified by regions and within regions by types of settlement places: large cities, towns and rural areas. Within each stratum, the sample was selected in two stages. On the first stage, territorial units (TU) were chosen. The selection of the TU was done with probability proportional to number of registered IDPs in each region. In the second stage, a fixed number of two household was selected in each TU. On the second stage random sampling was applied. In total, 300 TU were selected, therefore the overall sample size equals 600 households.

For estimating characteristics, familiar formulas were used, for example to estimate sample proportion the following formula was used:

$$\hat{P} = \frac{\sum_{i=1}^{n} w_i \; q_i}{\sum_{i=1}^{n} w_i}$$

In which w_i is statistical weight for i-th person; q_i is value of binary variable, which takes value of 1 if a person is employed and 0 if not; n is the Survey sample size.

Calculation of the statistical weight w_i for i-th household and each person in it, was based on probability of primary TU selection (P_{1i}) and probability of household selection (P_{2i}). This calculation includes computation of the household basic weights as inverse values of the general probabilities of household selection and adjustment of the basic weights in order to take into account the actual household and individuals' response rate. Therefore, the basic weight of the i-th selected household defined as:

$$w_{Bi} = \frac{1}{P_{1i} \cdot P_{2i}}$$

To be able to obtain all needed estimates, the cumulate array was created by combining micro data in micro level and the statistical weight was calculated in the way described above. After combining the arrays, the statistical weight was corrected.

2.2 Estimates of sampling errors

One of the key characteristics of the indicator's quality is sampling error. Sampling errors measure the variability between the estimates from all possible samples. To measure the quality of estimates the following values were used:

1. standard error

$$SE = \sqrt{\frac{\sigma^2}{n}}$$

For the samples that have complex design, variance is obtained by inputting the design-effect component ($deff$)

$$\sigma^2 = deff \cdot \frac{\sum_{i-1}^{n} (y_i - \bar{y})^2}{n}$$

where y_i is value of characteristic in a focus for i-th household; $deff$ is the ratio of the actual variance of an indicator, under the sampling method used in the survey, to the variance calculated under the assumption of simple random sampling. In this survey it was taken from external sourses.

2. value of limit sample error (ME) and confidence limits

$$ME = t \cdot SE$$

$$\bar{y}_L = \bar{y} - ME$$

$$\bar{y}_R = \bar{y} + ME$$

t is quantile of standard normal distribution, which equals 1.96 for confident level 95%.

3. coefficient of variation (CV)

$$CV = \frac{SE}{\bar{y}} \cdot 100\%$$

CV is widely used for analyzing suitability of the data sets. If $CV \le 5\%$ then the estimation is reliable, if $5\% \le CV \le 10\%$ then the estimation is suitable for quantity analyzing but its reliability is not good enough, if $10\% \le CV \le 25\%$ then the estimation is suitable for quality analyzing but it needs careful use.

The next indicators were estimated and sampling errors were calculated for them:

Table 1: Sampling errors: National level

Indicator	Value	Standard error	Confidence limits		CV
			Lower limit	Upper limit	
the average household size	2.65	0.0371	2.58	2.72	1.40%
the percent of household with children (%)	49.56	1.3026	47.00	52.11	2.63%
average income per household member	1420.37	21.2961	1378.63	1462.11	1.50%
percent of employed IDPs	35.12	1.2351	32.70	37.54	3.52%
rented accommodation (%)	65.32	1.2185	62.93	67.71	1.87%

hostel/collective accommodation center (%)	11.54	0.7386	10.09	12.99	6.40%
relatives/host family (%)	20.31	1.0378	18.27	22.34	5.11%

Table 2: Sampling errors: Region level

Indicator		Value	Standard error	Confidence limits		CV
				Lower limit	Upper limit	
the average household size	Region 1	2.65	0.0431	2.56	2.73	1.63%
	Region 2	2.98	0.1208	2.74	3.21	4.06%
	Region 3	2.54	0.0922	2.35	2.71	3.64%
	Region 4	3.05	0.1294	2.79	3.30	4.24%
the percent of household with children (%)	Region 1	51.10	1.5065	48.14	54.05	2.95%
	Region 2	49.23	4.0419	41.31	57.15	8.21%
	Region 3	38.22	3.3300	31.69	44.74	8.71%
	Region 4	65.11	4.9048	55.49	74.72	7.53%
average income per household member	Region 1	1345.52	23.1353	1300.17	1390.86	1.72%
	Region 2	1719.63	70.4662	1581.51	1857.74	4.10%
	Region 3	1798.04	76.2310	1648.63	1947.45	4.24%
	Region 4	1490.96	91.2616	1312.09	1669.84	6.12%
percent of employed IDPs	Region 1	32.84	1.4162	30.06	35.62	4.31%
	Region 2	44.44	3.9103	36.78	52.10	8.80%
	Region 3	45.31	3.3398	38.77	51.86	7.37%
	Region 4	36.47	4.7450	27.17	45.77	13.01%
rented accommodation (%)	Region 1	65.04	1.4188	62.26	67.82	2.18%
	Region 2	58.69	4.1211	50.61	66.77	7.02%
	Region 3	68.37	2.9358	62.61	74.12	4.29%
	Region 4	72.87	2.8723	67.24	78.50	3.94%
hostel/collective accommodation center (%)	Region 1	11.14	0.8333	9.51	12.78	7.47%
	Region 2	12.74	2.,9290	7.00	18.48	22.98%
	Region 3	14.04	2.1273	9.87	18.21	15.14%
	Region 4	7.58	1.6850	4.28	10.88	22.20%
relatives/host family (%)	Region 1	22.10	1.2312	19.69	24.52	5.56%
	Region 2	20.09	3.4048	13.42	26.76	16.94%
	Region 3	10.35	2.0850	6.26	14.44	20.13%
	Region 4	14.38	2.1552	10.16	18.61	14.98%

3 Conclusions

Coefficient of variation is less than 6.4% in national level and its maximum value for each region variates in 7.47% - 22.98%. Therefore, estimations of indicators are reliable enough in national level, so they can be used in the analysis.

The results of analysis shows that the average income per household member in Ukrainian (4Q 2015) is 1925 UAH. While the same income of IDPs variates from 1378.63 UAH to 1462.11 UAH (95% confidence). Taking into account the average household size the IDPs household income is less than the average household income of Ukraine and such a difference is 1338 UAH.

References

Kish L. (1995) *Survey sampling,* Wiley, New York, 643.

Sarndal Carl-Erik, Swensson Bengt, Wretman Jan (1992) *Model Assisted Survey Sampling*.

MEAN ESTIMATION WITH ROBUST CALIBRATED ESTIMATORS

Iryna Rozora[1] and Olga Lukovych[2]

[1] Taras Shevchenko National University of Kyiv, Ukraine
e-mail: irozora@bigmir.net

[2] Taras Shevchenko National University of Kyiv, Ukraine
e-mail: lukolga@ukr.net

Abstract

Calibration and robustness is hot topic in many recent articles on estimation in survey sampling. The article deals with estimation of population average by applying different calibration approaches. Two new calibrated estimators of mean are introduced.

1 Introduction

The problem of outliers is an important one in all branches of statistics. In sampling theory, the background is different from that of parametric statistics since the objective is often to estimate the mean of a variable of interest *y*. An outlier may have its full weight within the population mean. Therefore, the presence of such outliers in the sample may introduce bias and increase the variance of estimator of the selected model parameters. Outliers could also be the consequence of a highly skewed distribution. The presence of outliers in the sample could also be the result of measurement errors. However, it is assumed in the rest of this paper that the data have been verified and corrected, if necessary, and that there is no measurement error left in the data. Lee (1995) has provided an overview of robustness developments within sampling theory. Many of the first robust alternatives to the mean were based on M estimators and GM-estimators. Nevertheless, much interest has been shown recently for estimators that also provide good overall robustness, as measured by the breakdown point of an estimator. The breakdown point measures the percentage of outliers within the sample that the estimator can tolerate while providing nonetheless a good estimate of a given characteristic of the population.
Section 2 is devoted to such property of estimators as asymptotical normality. The definitions of asymptotic vatiance and relative asymptotic efficiency are given. We consider some estimators of population mean such as sample mean as well as median. A special attention should be paid to trimmed mean for which there exists the breakdown point that measures the percentage of outliers. And on the other hand this estimator provides a good efficiency. We also consider the median of Walsh averages (Walsh median) as estimator of the mean. It's shown that all these estimators are asymptotically normal. For trimmed mean and Walsh median the theorems are written that give the lowest level of relative asymptotic efficiency compared with sample mean.
The third section deals with different calibration approaches to estimate population mean. The basic design estimator of population mean that is calibrated is the Horwitz-Thompson estimator. We also study the model based calibrated estimators as calibrated median. Calibrated trimmed mean and calibrated median of Walsh averages are introduced and investigated. They are new estimators of population mean.
In last section the example with application of all calibration technique is considered.

2 Asymptotic normality of the estimators

Consider a sample $(X_1, X_2, ..., X_n)$, where random variables X_k, $k = \overline{1,n}$, are independent identically distributed.

Definition. T is called a statistics (estimator) if it is an arbitrary borelean function of the sample $(X_1, X_2, ..., X_n)$.

Definition. The statistics $T = T_n$ is called *asymptotically normal* if there exist such numeric sequences a_n and b_n that the value $\frac{T_n - a_n}{b_n}$ tends to standard normal random variable by distribution as $n \to \infty$:

$$\frac{T_n - a_n}{b_n} \xrightarrow{d} Z \sim N(0,1).$$

In the case of asymptotically normal estimator $\hat{\theta}_n$ of the parameter θ in regular statistical models the typical order of the smallness for the coefficient b_n equals $1/\sqrt{n}$. Then the condition of asymptotic normality can be rewritten as

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \xi \sim N(0, \sigma^2(\theta)) \text{ as } n \to \infty.$$

Evidently that from asymptotic normality follows asymptotic unbiaseness and consistency of the estimator.

Definition. The quantity $\sigma^2(\theta)$ is called *asymptotic variance* of asymptotically normal estimator $\hat{\theta}_n$.

Example. Assume that $0 < VarX_1 < \infty$. According to Central Limit Theorem (CLT) the following relationship holds true for sample mean $\bar{X}$ as an estimator of mathematical expectation EX_1

$$\sqrt{n}(\bar{X} - EX_1) \xrightarrow{d} \xi \sim N(0, \mathrm{Var}\, X_1) \text{ as } n \to \infty.$$

Hence, the value $VarX_1$ is asymptotic variance of sample mean $\bar{X}$ which is asymptotically normal estimator of theoretical mean EX_1.

Consider other asymptotic normal estimator of the mean.

2.1 Sample median

Hereinafter we will consider only symmetric distributions. Let's give a formal definition.

Definition. Cumulative distribution function (cdf) F belongs to the class of *symmetric continuously differentiable distributions* (Ω_s) if there exists such constant c: $0<c\leq \infty$ that $F(-c) = 0$, $F(c) = 1$ and on the domain (-c;c) the function F has even continuous and positive density function p(x).

Definition. A *relative asymptotic efficiency* of asymptotically normal estimator $\hat{\theta}_1$ to asymptotically normal estimator $\hat{\theta}_2$ is called the value

$$e_{\hat{\theta}_1,\hat{\theta}_2} = \frac{\sigma_2^2(\theta)}{\sigma_1^2(\theta)}.$$

Consider the elements of sample in increasing order: $X_{(1)} \le X_{(2)} \le ... \le X_{(n)}$. The estimator

$$MED = \begin{cases} X_{(k+1)}, & \text{if } n = 2k+1, \\ (X_{(k)} + X_{(k+1)})/2, & \text{if } n = 2k \end{cases}.$$

is called sample median. In the case of symmetric distribution MED can be applied as an estimator of mean. The median is not so sensitive to outliers than sample mean.

Theorem. Suppose that the elements of sample X_i have cdf $F(x-\theta)$ with density function $p(x)$, where $F \in \Omega_s$. Let $p(\theta) > 0$ then

$$\sqrt{n}(MED-\theta) \xrightarrow{d} \xi \sim N\left(0, \frac{1}{4p^2(\theta)}\right).$$

Remark. If the sample with normal distribution $N(\theta,1)$ is considered then it's easy to show that the relative asymptotic efficiency between sample mean and sample median is equal to $e_{MED,\bar{X}} = \frac{2}{\pi} \approx 0.64$. This means that sample mean is on 36% more effective than sample median.

2.2 Trimmed mean

As a compromise between robustness of median and effectiveness of sample mean can be considered a trimmed mean. Let's give a definition.

Definition. Let $\alpha \in (0,1/2)$, $k = [\alpha N]$, where [ ] is an integer of the number, n is sample size. A *trimmed mean* is called an estimator

$$\bar{X}_\alpha = \frac{1}{N-2k}\left(X_{(k+1)} + ... + X_{(N-k)}\right),$$

where $X_{(1)} \le X_{(2)} \le ... \le X_{(n)}$ are order statistics.

The bounded values $\alpha = 0$ and $\alpha = \frac{1}{2}$ corresponds to $\bar{X}$ and MED respectively (see fig.1).

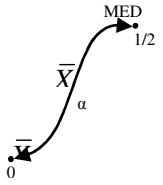


Fig.1 The trimmed mean

Theorem. Suppose that the elements of sample X_i have cdf $F(x-\theta)$ with density function $p(x)$, where $F \in \Omega_s$. Then

$$\sqrt{n}(\bar{X}_\alpha - \theta) \to \xi \sim N\left(0, \sigma_\alpha^2\right), \quad n \to \infty,$$

where $\sigma_\alpha^2 = \frac{2}{(1-2\alpha)^2}\left[\int_0^{x_{1-\alpha}} t^2 p(\mathrm{t})\,\mathrm{dt} + \alpha x_{1-\alpha}^2\right]$, $x_{1-\alpha}$ is a (1-α) -quantile of cdf F.

Example.
In table below shows how decreases the asymptotic relative efficiency with increasing of α in the case of normal sample (We suppose that F is cdf of N(0,1)).

α	0	1/20	1/8	1/4	3/8	1/2
$e_{\bar{X}_\alpha,\bar{X}}$	1,00	0,99	0,94	0,84	0,74	0,64

In particular case for α=1/8 (12,5% data protection from outliers) the loss of efficiency consists of 6%.

The following theorem gives the lowest bound for relative efficiency.

Theorem. For any distribution $F \in \Omega_s$ the next inequality holds:

$$(1-2\alpha)^2 \le e_{\bar{X}_\alpha,\bar{X}}(\mathrm{F}) \le \infty.$$

Some values of $(1\text{-}2\alpha)^2$ are presented in the next table.

α	0	1/20	1/8	1/4	3/8	1/2
$(1\text{-}2\alpha)^2$	1,00	0,81	0,56	0,25	0,06	0,00

α=1/8 provides the loss of efficiency on the level 44% that is too much compared with normal distribution.

2.3 Median of Walsh averages

By sample items $X_1, X_2, ..., X_n$ let's construct $M = \frac{n(\mathrm{n}-1)}{2}$ new random variables $Z_k = \frac{1}{2}\left(X_i + X_j\right)$, $i \le j$. These random variables are called Walsh averages. The quantity $W = MED\{Z_1, ..., Z_M\}$ is a median of Walsh averages.

Theorem. Suppose that the elements of sample $\mathrm{X_i}$ have cdf $F(\mathrm{x}-\theta)$ with density function $p(\mathrm{x})$, where $F \in \Omega_s$. Then

$$\sqrt{n}(\mathrm{W}-\theta) \xrightarrow{d} \xi \sim \mathrm{N}\left(0, \sigma_F^2\right),$$

where asymptotic variance in equal to $\sigma_F^2 = \frac{1}{E(\mathrm{F})}$, $E(\mathrm{F}) = 12\left(\int_R p^2(\mathrm{t})\,\mathrm{d}t\right)^2$.

In the case of normal sample N(0,1) relative asymptotic efficiency between median of Walsh averages and sample mean is $e_{\mathrm{W},\bar{X}} \approx 0{,}955$. Hence, Walsh median provides only 4.5% the loss of efficiency compared with sample mean.
The following theorem gives the lowest bound for relative efficiency.

Theorem. For any distribution $F \in \Omega_s$ the next inequality holds:

$$e_{\mathrm{W},\bar{X}}(\mathrm{F}) \geq 108/125 \approx 0{,}864.$$

This means that for any distribution the loss of efficiency for Walsh median could not be greater than 14%.

3 Calibration approach to estimation

A probability sample s is drawn from the finite population $U = \{1,2,...,k,...,N\}$. Denote by π_k the inclusion probability of unit and by the design weight of k. Let y be the variable of interest. The value y_k of the study variable y is recorded for all $k \in s$ (complete response). The objective is to estimate the unknown mean $\mu_y = \frac{\sum_U y_k}{N}$. The basic design estimator of μ_y is $\hat{\mu}_y^{HT} = \frac{\sum_{k \in s} d_k y_k}{N}$, the Horwitz-Thompson estimator.

It is unbiased for μ_y and it is the basis for construction of many estimators. It is not always a very good estimator. HT estimator may be seriously deficient. One of the question is to improve an accuracy of the estimate $\hat{\mu}_y$.

A more efficient weighting as compared with HT estimator is usually achieved by using the available auxiliary information. Denote by x_k an auxiliary variable, associated with the k-th unit. It can be a vector. Under the basic conditions we need to distinguish two different cases relative to x_k

(i) x_k is a known value for every $k \in U$. (complete auxiliary information)

(ii) $\sum_U x_k$ is known (imported) total, and x_k is known (observed) for every $k \in s$.

Case (i) gives some freedom in structuring the auxiliary vector x_k. For example, if x_k is a continuous variable value specified for every $k \in U$, then we can consider and other functions of x_k, such as x_k^2 or $\log x_k$ and so on. The sum of these values is computed without any problem. Case (ii) prevails in surveys where (i) is not met, but where $\sum_U x_k$ is imported from an outside source, and the individual value x_k is available (observed in data collection) for every $k \in s$.

3.1 The minimum distance method to find calibrated mean

Definition. (Deville and Särndal (1992)) Estimator

$$\hat{\mu}_{CAL,y} = \frac{\sum_{k \in s} w_k y_k}{N}$$

is called calibrated if

- it estimates the known mean μ_x without error: $E\hat{\mu}_{CAL,x} = \mu_x$ (or the same $\sum_{k \in s} w_k x_k = \sum_{k \in U} x_k$) and
- the distance between the weights d_k and weights w_k is minimal according to the loss function $L(w,d) = L(w_k, d_k, k \in s)$

To calculate calibrated weights we solve a minimization problem

$L(w,d) = L(w_k, d_k, k \in s) \rightarrow \min$

Note that for some loss functions the explicit solution exists, for others the iterative procedure can be used. Deville and Särndal (1992) show that a variety of distance functions satisfying

mild conditions will generate asymptotically equivalent calibration estimators. The most commonly used distance measure is the chi-squared distance

$$L(w,d) = \sum_s (w_k - d_k)^2 / (q_k d_k),$$

where the q_k's are known positive constants uncorrelated with the d_k's.

Calibrated weights can be less than 1 or negative (for some loss functions). (It may cause a real problem when, for example, estimating totals for domains.)

Questions about the existence of a solution to the calibration equation are discussed in Théberge (2000).

Considering the weights in such form $w_k = d_k(1 + q_k x_k' \lambda)$ and solving calibration equation with respect to λ, we obtain that calibrated mean estimator is equal to

$$\hat{\mu}_{CAL,y} = \frac{\sum_{k\in s} w_k y_k}{N} = \hat{\mu}_y^{HT} + \hat{\beta}(\mu_x - \hat{\mu}_x^{HT}), \qquad (1)$$

where $\hat{\beta} = (\tilde{x}\tilde{x}')^{-1}\tilde{x}'\tilde{y}$, $\tilde{x} = (\tilde{x}_k)_{k\in s}$ is a matrix with corrected vectors $\tilde{x}_k = \sqrt{q_k d_k}\, x_k$, $\tilde{y}_k = \sqrt{q_k d_k}\, y_k$.

3.2 Calibrated median as estimator of population mean

The median of the finite population is important descriptive value and the estimator of the mean, especially in economic surveys. Compared with HT estimator it's biased but asymptotically unbiased. The median is robust estimator with respect to outliers. That's why some times it' more desirable to use it.

To find median, the finite population distribution function must first be estimated. More recent articles have turned to the calibration approach for the same purpose, including Kovaĉević (1997), Wu and Sitter (2001), Ren (2002) and Rueda et al. (2007).

Let $I(t)$ be an indicator function of the event $t \geq 0$ (the Heaviside function), defined for all real t so that

$$I(t) = \begin{cases} 1, & t \geq 0, \\ 0, & t < 0. \end{cases}$$

The unknown cumulative distribution function of y is

$$F_y(t) = \sum_{k\in U} \frac{I(t - y_k)}{N}$$

Definition. The α-quantile of the finite population is defined as

$$Q_{y\alpha} = \inf\{t \in R \mid F_y(t) \geq \alpha\}$$

The median of y is the quantile with degree ½, that is $med_y = Q_{y/2}$.

The auxiliary variable x_j, that takes values x_{jk}, has the distribution function $F_{x_j}(t) = \sum_{k\in U} \frac{I(t - x_{jk})}{N}$ with median denoted med_{x_j} j = 1, 2, ..., J . An estimator of $F_y(t)$ based on the design weights $d_k = \frac{1}{\pi_k}$ is

$$\hat{F}_y(t) = \frac{\sum_{k\in s} d_k I(t - y_k)}{\sum_{k\in s} d_k}$$

A calibration estimator of $F_y(t)$ can be considered as

$$\hat{F}_{yCAL}(t) = \frac{\sum_{k\in s} w_k I(t - y_k)}{\sum_{k\in s} w_k}$$

where the weights w_k are suitably calibrated to a specified auxiliary information; then from $\hat{F}_{yCAL}(t)$ we obtain the median estimator as

$$\hat{med}_{CAL,y} = \inf\{t \in R \mid \hat{F}_{yCAL}(t) \geq 1/2\}. \qquad (2)$$

The same formulae holds for $\hat{F}_{x_jCAL}(t)$ and $\hat{med}_{x_j}$ .Without explicit reference to any model, Harms and Duchesne (2006) specify the information available for calibration as a known population size, N, and known population medians med_{x_j} for j = 1, 2, ..., J . The complete auxiliary information, with values $x_k = (x_{k1}, \ldots x_{kJ})$ known for $k \in U$, is not required. They determine the w_k to minimize the chi-square distance $L(w,d) = \sum_s (w_k - d_k)^2 / (q_k d_k)$, for specified q_k , subject to the calibration equations

$$\sum_{k \in s} w_k = N, \quad \hat{med}_{x_j} = med_{x_j}, \quad j = 1, \ldots, J .$$

The computationally simpler method of Rueda et al. (2007) is an application of model calibration, in that they calibrate with respect to a population total of predicted y values. Complete auxiliary information is required. Using the known x_k, compute first the linear predictions $\hat{y}_k = \hat{\beta}' x_k$ for $k \in U$, with $\hat{\beta} = (\tilde{x}\tilde{x}')^{-1}\tilde{x}'\tilde{y}$, where $\tilde{x} = (\tilde{x}_k)_{k \in s}$ is a matrix and values $\tilde{x}_k, \tilde{y}_k$ are corrected by multiplied scalar coefficient $\sqrt{q_k d_k}$: $\tilde{x}_k = \sqrt{q_k d_k}\, x_k$ $\tilde{y}_k = \sqrt{q_k d_k}\, y_k$ and the q_k are specified scale factors. The weights w_k are obtained by minimizing the chi-square distance subject to calibration equations stated in terms of the predictions, so as to have consistency at J arbitrarily chosen points t_j , j = 1, ..., J :

$$\frac{\sum_{k \in s} w_k I(t_j - \hat{y}_k)}{N} = F_{\hat{y}}(t_j), \quad j = 1, \ldots, J$$

where $F_{\hat{y}}(t_j)$ is the finite population distribution function of the predictions $\hat{y}_k$ evaluated at t_j. Once the w_k are determined, the median estimate is obtained from

$$\hat{F}_{yCAL} = (1/N) \sum_{k \in s} w_k I(t - y_k).$$

3.3 Calibrated trimmed mean

Calibrated trimmed mean can be constructed in the same manner as calibrated median.

A calibration estimator of cumulative distribution function $F_y(t)$ can be considered as

$$\hat{F}_{yCAL}(t) = \frac{\sum_{k \in s} w_k I(t - y_k)}{\sum_{k \in s} w_k}$$

where the weights w_k are suitably calibrated to a specified auxiliary information. Then from $\hat{F}_{yCAL}(t)$ we obtain the sequence of order calibrated statistics as

$$\hat{y}_{(k)}^{Cal} = \inf\{t \in R \mid \hat{F}_{yCAL}(t) \geq 1/k\}, \quad k = \overline{1, n}. \qquad (3)$$

Then for $\alpha \in (0, 1/2), \quad k = [\alpha n]$ the calibrated trimmed mean is defined as

$$\hat{Y}_{\alpha}^{Cal} = \frac{1}{n-2k}\left(\hat{y}_{(k+1)}^{Cal} + \ldots + \hat{y}_{(n-k)}^{Cal}\right). \quad (4)$$

3.4 Calibrated median of Walsh averages

Use the sequence (4) to define calibrated median of Walsh averages. Consider $M = \frac{n(n-1)}{2}$ new variables $Z_k^{Cal} = \frac{1}{2}\left(\hat{y}_{(i)}^{Cal} + \hat{y}_{(j)}^{Cal}\right)$, $i \le j$. Then the estimator

$$W^{Cal} = MED\left\{Z_1^{Cal}, \ldots, Z_M^{Cal}\right\} \quad (5)$$

is called calibrated median of Walsh averages.

4 Example

This example illustrates how these estimates work in practice. Using samples were loaded from the European Social Survey (ESS) website http://www.europeansocialsurvey.org. ESS is a biennial multi-country survey covering over 30 nations. The estimated variable is called "Most people can be trusted or you can't be too careful". Auxiliary information variable is "Tv watching, total time on average weekday". The researched variable is changing from 0 to 10 points and has several additional values as "refusal", "don't know" and "no answer". For all calculation a programming language and software environment for statistical computing and graphics R is used. For researched variable we calculate calibrated HT mean, calibrated median and calibrated median of Walsh averages. The results are presented in Figure 2.

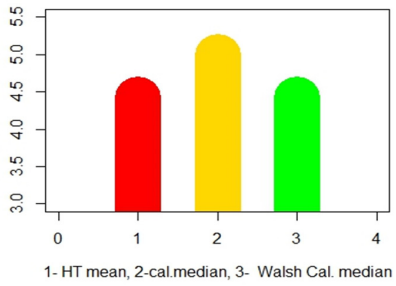


Figure 2. The results of calculation

References

Alexander, C.H. (1987). A class of methods for using person controls in household weighting. *Survey Methodology*, 13, 183-198.

Bankier, M.D., Rathwell, S. and Majkowski, M. (1992). Two step generalized least squares estimation in the 1991 Canadian Census. Working paper, *Census Operations Section, Social Surveys Methods Division,* Statistics Canada.

Bethlehem, J.G., and Keller, W.J. (1987). Linear weighting of sample survey data. *Journal of Official Statistics*, 3, 141-153.

Chambers, R.L. (1996). Robust case weighting for multipurpose establishment surveys. *Journal of Official Statistics*, 12, 332.

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

Fuller, W.A., Loughin, M.M. and Baker, H.D. (1994). Regression weighting in the presence of nonresponse with application to the 1987-1988 nationwide Food Consumption Survey. *Survey Methodology*, 20, 75-85.

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

Harms, T., and Duchesne, P. (2006). On calibration estimation for quantiles. *Survey Methodology*, 32, 37-52.

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

Kalton, G., and Flores-Cervantes, I. (1998). *Weighting methods*. In New Methods for Survey Research (Eds. A.Westlake, J. Martin, M. Rigg and C. Skinner),. Berkeley, U.K.: Association for Survey Computing.

Kovaĉević, M.S. (1997). Calibration estimation of cumulative distribution and quantile functions from survey data. *Proceedings of the Survey Methods Section*, Statistical Society of Canada, 139-144.

Lagutin M.B.(2009) *Transparent mathematical statistics.* Moscow, BINOM. (in Russian)

Lee, H. (1995). Outliers in business surveys. In *Business Survey Methods*, (Eds. B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge and P.S. Kott). New York: John Wiley & Sons, Inc.

Lehtonen, R., Särndal, C.E. and Veijanen, A. (2003). The effect of model choice in estimation for domains, including small domains. *Survey Methodology*, 29, 3344.

Ren, R. (2002). Estimation de la fonction de répartition et des fractiles d'une population finie. *Actes des journées de méthodologie statistique*, INSEE Méthodes, tome 1, 100, 263-289.

Rueda, M., Martínez, S., Martínez, H. and Arcos, A. (2007). Estimation of the distribution function with calibration methods. *Journal of Statistical Planning and Inference*, 137, 435-448.

C.-E. Särndal (2007). The calibration approach in survey theory and practice. *Survey methodology*, 33(2), 99-119.

Théberge, A. (2000). Calibration and restricted weights. *Survey Methodology*, 26, 99-107.

Särndal, C., Swensson, B. & Wretman, J. (2003). *Model assisted survey sampling*. Springer Verlag.

Wu, C., and Sitter, R.R. (2001). A modelcalibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96, 185-193.

MATCHING OF STATE SAMPLE HOUSEHOLD LIVING CONDITIONS SURVEY DATA

Svitlana Synogub

Ptoukha Institute for Demography and Social Studies of the National Academy of Sciences of Ukraine, Ukraine
e-mail: sveta.llana@gmail.com

Abstract

The state Household Living Conditions Survey (HLCS) provided by the State Statistics Service of Ukraine on quarterly basis is the main source of information for assessing the targeting performance of social programs.

Number of the observed households is clearly insufficient. The problem of the assessment reliability and forming the targeting of information source arises.

In this paper the relevant information base is formed by specially designed procedures using cumulative microdata of the state sample household living conditions surveys for 2011-2014 years and administrative dataset of GO of Ministry of Social Policy.

1. Introduction

We can see in Table 1 that the number of the observed households, which are the recipients of social assistance in annual HLCS, is clearly insufficient for the following more deep research. Annually 147 households on average are observed.

Table 1.

The number of recipients of social assistance for low-income families in HLCS data for 2011, 2012, 2013 and 2014 years

	The number of households
Data matching	589
HLCS 2011	146
HLCS 2012	116
HLCS 2013	152
HLCS 2014	175

The relevant source of information is built by specially designed procedures that will be seen further.

2. HLCS data matching

Statistical data matching enables to perform state sample survey of 589 households instead of 147.

In order to prepare targeting information source the state sample household living conditions survey microdata is formed by matching of HLCS data by households which were surveyed in 2011-2014 years. Forming of data matching for four years is performed by using the method of merging data (Tereshchenko 2010). This method is used when it is necessary to cumulate data by the same features, that is when the additional sample units are needed to be added to data set (Fig. 1). As a result, here is matched data in which all households from four

data sets are represented. It should be remarked that the matched data doesn`t contain information about households in the Crimea and Sevastopol City during these years.

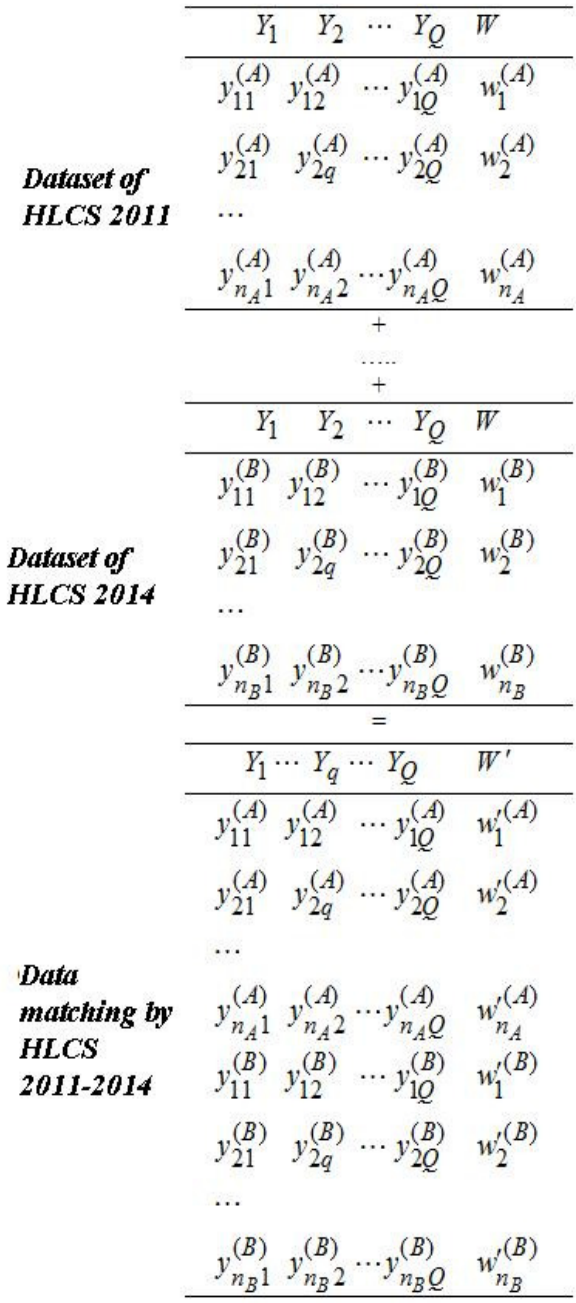


Figure 1. Scheme of matching of annual HLCS data in one data set by the method of merging data

Statistical weights adjusting of households was made during matching of data. Adjusted statistical weights of matched data are calculated as:

$$w_i' = \frac{w_i^{(r)}}{4} \tag{1}$$

w_i - statistical weight of households; $i = 1, 2, ..., n$; n - size of selective totality in r year;

r = 2011, 2012, 2013, 2014.

The statistical weights adjusting of households was used for making all estimates of the number of households as the ones that are in 2014, particularly for taking account of changes in the distribution of households due to their displacement from the area of anti-terrorist operations to other regions of Ukraine. The calibration procedure (Sarioglo 2005) of statistical weights of households was implemented using external information about the distribution of households by regions in 2014.

It should be said that in the matched data incomes indicators of households were adjusted for each year in such way that the matched data of households reflected the incomes of households for 2014 year (see. Fig. 2).

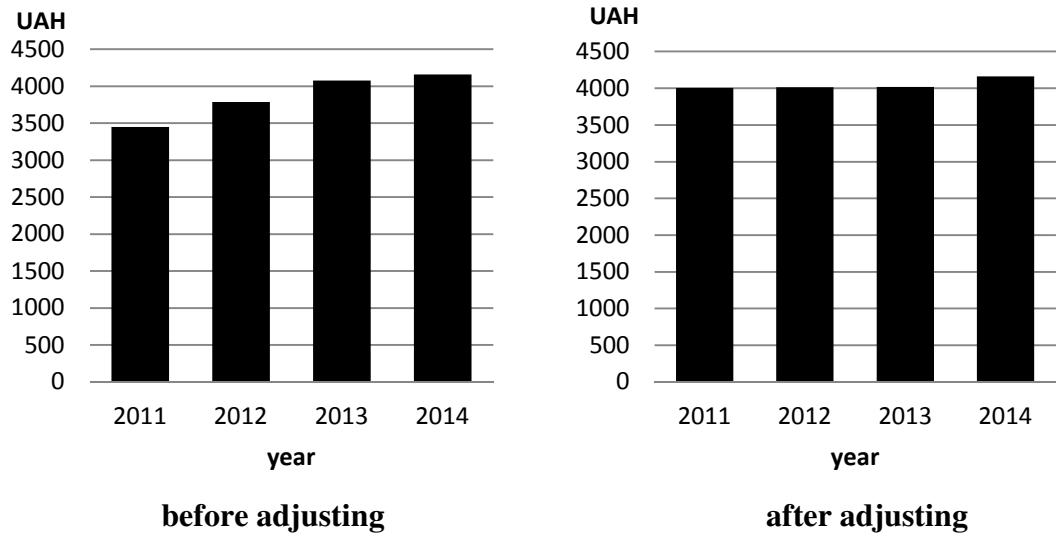


Figure 2. Average monthly cash incomes of households in 2011-2014 before and after adjusting

3. Administrative data

Also data set of GO of the Ministry of Social Policy of Ukraine is built with the registers of families that participate in Program by regional Departments of Labor and Social Protection.

The input data set for each region contains information about family (recorded in one row):

• code region (first column) – took the one value for each file;

• number of family within the region (from 1 to 1236);

• date of the last applying;

• family size (number of members) (including only 14 members);

• total family income for 6 months (UAH);

• the average family income (UAH);

• date of last repeated applying;

• date of initial appointment of social assistance;

• size of social assistance (UAH);

• percentage of social assistance size (%);

• size of social assistance payment (UAH);

• reason for refusal (text variable);

• code number of the applicant, as a member of the family ("0");

• code of the other social payments (43 variants of codes);

• name of other social payments;

• size of the other social payments (UAH).

The input data set in section where information about family members was (each family member was in the new row) included such information:

• Ordinal number of the family member (from 0 (the code of the applicant) to 13);

• Birthday of the family member;

• Sex of the family member (binary variable that takes the value 1 or 2);

• Code of family ties of the family member (in regard to the applicant);

• The name of the family ties of the family member (in regard to the applicant) (text variable);

• Code of the category of the family member (socioeconomic status);

• The name of the category of family member (socioeconomic status) (text variable).

In a result of data processing of the output files it were generated two files in «* .sav» (statistical package SPSS): data set with information about families and data set with information about family members.

Record for the each family and its member was assigned a unique number. Each number corresponds to one record (one row in data). So information about other social payments (code, name, size) in data set of families and information about incomes (code, name, size) in data set of family members placed in one row. This led to the fact that 200 variables were built in data set of families and 197 variables were built in data set of family members.

In such way, according to the applied procedures information base was formed, which in general is relevant.

References

Терещенко Г. І. Сучасні методологічні підходи до статистичногоо б'єднання даних / Г. І. Терещенко // Статистика України. — 2010. — № 3. — С. 23-29.

Саріогло В.Г. Проблеми статистичного зважування вибіркових даних. – К.: ІВЦ Держкомстату України, 2005.

PARTICIPANTS

Surname	Name	Organization & country	e-mail
Bandarenka	Natallia	Belarusian State University, Belarus	bondnata@mail.ru
Bethlehem	Jelke	Leiden University, Netherlands	jelkeb@xs4all.nl
Bokun	Natalia	Belarusian State Economic University, Belarus	nataliabokun@rambler.ru
Bondarenko	Iana	Oles Honchar Dnipropetrovsk National University, Ukraine	iana.s.bondarenko@gmail.com
Breidaks	Juris	Central Statistical Bureau of Latvia, Latvia	juris.breidaks@csb.gov.lv
Chebanova	Mariia	Taras Shevchenko National University of Kyiv, Ukraine	M_Chebanova@ukr.net
Čiginas	Andrius	Statistics Lithuania, Lithuania	andrius.ciginas@stat.gov.lt
Donchenko	Volodymyr	Taras Shevchenko National University of Kyiv, Ukraine	voldon@bigmir.net
Honkala	Miika	Statistics Finland, Finland	miika.honkala@gmail.com
Hrabets	Nestor	Taras Shevchenko National University of Kyiv, Ukraine	maxor2000@ukr.net
Ianevych	Tetiana	Taras Shevchenko National University of Kyiv, Ukraine	yakovenkot@gmail.com
Krapavickaite	Danute	Vilnius Gediminas Technical University, Lithuania	danute.krapavickaite@vgtu.lt
Kukush	Alexander	Taras Shevchenko National University of Kyiv, Ukraine	alexander.kukush@gmail.com
Laaksonen	Seppo	University of Helsinki, Finland	seppo.laaksonen@helsinki.fi
Laitila	Thomas	Örebro University, Sweden	thomas.laitila@oru.se
Lapins	Janis	Bank of Latvia, Latvia	janis.lapins@bank.lv
Lehtonen	Risto	University of Helsinki, Finland	risto.lehtonen@helsinki.fi
Lepik	Natalja	University of Tartu, Estonia	natalja.lepik@ut.ee
Levenko	Vassili	Statistics Estonia, Estonia	vassili.levenko@stat.ee
Liberts	Mārtiņš	Central Statistical Bureau of Latvia, Latvia	martins.liberts@csb.gov.lv

Surname	Name	Organization & country	e-mail
Lishnianska	Sofia	Taras Shevchenko National University of Kyiv, Ukraine	lishnianska.sofia@gmail.com
Lukovych	Tetiana	Taras Shevchenko National University of Kyiv, Ukraine	tan_luk@ukr.net
Lukovych	Olha	Taras Shevchenko National University of Kyiv, Ukraine	lukolga@ukr.net
Lysa	Olha	Kyiv International Institute for Sociology, Ukraine	olysa@ukr.net
Mishura	Yuliya	Taras Shevchenko National University of Kyiv, Ukraine	yumishura@gmail.com
Nemyrovska	Maryna	Taras Shevchenko National University of Kyiv, Ukraine	marina.n777@gmail.com
Rozora	Iryna	Taras Shevchenko National University of Kyiv, Ukraine	irozora@bigmir.net
Rozora	Natalia	Nielsen Co, Russia	rozora@ukr.net
Rudys	Tomas	Statistics Lithuania, Lithuania	tomas.rudys@gmail.com
Ruotsalainen	Kaija	Statistics Finland, Finland	kaija.ruotsalainen@stat.fi
Sarioglo	Volodymyr	Ptoukha Institute for Demography and Social Studies, Ukraine	sarioglo@idss.org.ua
Skendere	Inga	Rīga Stradiņš University, Latvia	inga.skendere@rsu.lv
Solovei	Anna	Taras Shevchenko National University of Kyiv, Ukraine	abrikoska@inbox.ru
Sydorov	Mykola	Taras Shevchenko National University of Kyiv, Ukraine	ms123@ukr.net
Sydorova	Daryna	Taras Shevchenko National University of Kyiv, Ukraine	dashasyd@i.ua
Sydorova	Kateryna	Taras Shevchenko National University of Kyiv, Ukraine	kateryna.sydorova@i.ua
Synogub	Svitlana	Ptoukha Institute for Demography and Social Studies, Ukraine	sveta.llana@gmail.com
Traat	Imbi	University of Tartu, Estonia	imbi.traat@ut.ee
Trotsko	Maria	Taras Shevchenko National University of Kyiv, Ukraine	mary.trotsko1@gmail.com
Vasylyk	Olga	Taras Shevchenko National University of Kyiv, Ukraine	olva75@gmail.com

НАУКОВЕ ВИДАННЯ

МАТЕРІАЛИ БАЛТІЙСЬКО-СКАНДИНАВСЬКО-УКРАЇНСЬКОЇ ЛІТНЬОЇ ШКОЛИ ЗІ СТАТИСТИКИ ОБСТЕЖЕНЬ

СЕРПЕНЬ 22 – 26, 2016

КИЇВ, УКРАЇНА