Summer School of Baltic-Nordic-Ukrainian Network on Survey Statistics 2013

PROCEEDINGS

Summer School was held on 14-18 of June, 2013,
near Minsk Sea in the studying center “Forest”.

Preface

Dear Participant,

We congratulate you for coming to the 17th event of the Baltic-Nordic-Ukrainian Network on Survey Statistics. The Summer School in Minsk 2013 is the first event within Network that takes place in Belarus. The main objectives of the School are to provide an opportunity for university teachers, research students and survey statisticians to discuss their problems and to learn from the experiences of the other countries.

The School starts with the day in Russian for Belarussian statisticians, students, investigators. The next five days will be devoted to different problems of theory and application of survey statistics. The Programme Committee invited four main speakers: Aleksandras Plikusas (Lithuania), Imbi Traat (Estonia), Risto Lehtonen (Finland), Seppo Laaksonen (Finland) to give series of lectures of teaching nature. There are five more invited speakers: Gunnar Kulldorff (Sweden), Danute Krapavickaite (Lithuania), Oksana Honchar (Ukraine), Martins Liberts (Latvia), Natalia Bokun (Belarus). They will deliver special lectures covering a lot of sample surveys' problems.

There are 38 participants at the School. Most of them will represent contributed papers included in this book. All presentations will be followed by discussions.

Finally we would like to thank International Association of Survey Statisticians, Institute of Economics of National Academy of Sciences of Belarus and Belarus State Economic University for a given support.

We wish you successful work, to get a lot of new ideas for your research. We wish you to have a pleasant time in Belarus and hope that the participation at the Summer School will be fruitful and enjoyable for everyone.

On behalf of the Organizing Committee,

Natalia Bokun

Anna Larchenko

Anastacia Bobrova

Katsiaryna Chystsenka

Contents

Invited papers

Analysis of complex survey data

Risto Lehtonen

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

Stratification, clustering and unequal probability sampling are typical properties of complex survey data sets used in scientific research and official statistics production. Good examples in social statistics are the EU's Statistics on Income and Living Conditions (SILC) survey and the European Social Survey (ESS). In many countries, these surveys are based on stratified multi-stage cluster sampling designs. In addition to the production of descriptive statistics, complex survey data are increasingly used for analytical purposes. For proper statistical inference, the various complexities need to be properly accounted for in the analysis phase. The lectures aim at giving a brief overview of the main approaches (design-based and model-based methods) for accounting for the complex survey design in statistical analysis. Topics that are treated in more detail include design-based analysis for two-way tables and logistic and linear regression and analysis of covariance. Real-world examples are given, as well as brief software overview and suggestions for further reading.

References

Lehtonen, R. and Pahkinen, E. (2004). Practical Methods for Design and Analysis of Complex Surveys. Second Edition. Chichester: John Wiley & Sons Ltd. (With Web extension)

Lohr S.L. (2009). Sampling: Design and Analysis. 2nd Edition. Boston, MA:Brooks/Cole.

Lumley, T. (2010). Complex Surveys: A Guide to Analysis Using R. Hoboken, NJ: John Wiley & Sons Inc.

Web references

European Social Survey (ESS), see at:http://www.europeansocialsurvey.org/

Statistics on Income and Living Conditions (SILC), see at:

http://epp.eurostat.ec.europa.eu/portal/page/portal/income_social_inclusion_living_conditions/introduction

Virtual Laboratory in Survey Sampling VLISS, see at: http://vliss.helsinki.fi/

Software for survey sampling and analysis

Danutė Krapavickaitė

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

Abstract

An overview of the computer software for probabilistic sample selection, imputation of missing values, estimation of finite population parameters and other kinds of statistical data analysis will be made.
Keywords: sample selection, complex survey data, multiple imputation, calibration, estimation.

1 Contents

The following computer programs for survey data analysis will be overwied.

Sample selection and general estimation problems

1. SAS has procedures for sample selection (procedure *surveyselect*), estimation and data analysis when complex sample design is taken into account: procedures *surveymeans, surveyfreq, surveyreg, surveylogistic, surveyphreg*. SAS macro program CLAN (Statistics Sweden) is written for estimation needs of statistical office, including calibration of design weights.

2. SPSS Complex Samples Selection module may be used to select a sample under a complex sampling design, other modules for estimation and survey data analysis are included into the Complex Samples group. g-Calib Release 2.0 software for calibration of design weights by Camille Vanderhoeft, Statistics Belgium, is used under SPSS.

3. R packages. Package sampling includes many different algorithms for drawing survey samples and calibrating the design weights. Package survey can also handle moderate data sets and is the standard package for dealing with already drawn survey samples in R, point and variance estimates can be computed. Package simFrame is designed for performing simulation studies in official statistics. It provides a framework for comparing different point and variance estimators under different survey designs as well as different conditions regarding missing values, representative and non-representative outliers.

Special topics

4. Imputation of missing values. Analysis of missing values and single imputation methods are implemented in SPSS Missing Values modules, R package VIM. Multiple imputation – imputation of random values with subsequent estimation of increase in variance due to imputation – has achieved the highest attention in statistical software, and is realised in SAS procedures MI and MIANALYSE, SPSS Missing Values modules, R packages mi, mice, Amelia, and SAS macro (and self-dependent software) IVEware from Michigan University.

5. Small area estimation. The SAS macro program EBLUPGREG by Statistics Finland is designed to estimate mean and total of quantitative variable in the population domains and small areas. The super-population linear regression model is supposed. Generalized regression, synthetic and composit estimators are used.

Overviews of the software for survey statistics are presented in *The Survey Statistician*. All software mentioned here except of main SAS and SPSS products are free.

References

Summary of Survey Analysis Software. http://www.hcp.med.harvard.edu/statistics/survey-soft/

SPSS Software. http://www.spss.com/

Statistics Belgium. g-calib.software.informer.com/

The Comprehensive R Archive Network. http://cran.r-project.org/

IVEware: Imputation and Variance Estimation Software. http://www.isr.umich.edu/src/smp/ive/

Trivellore E.Raghunathan, James M. Lepkowski, John Van Hoewyk, Peter Solenberger. A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. *Survey Methodology*, June 2001. Vol. 27, No. 1, pp. 85-95. Statistics Canada, Catalogue No. 12-001.

Yang Yuan. Multiple Imputation Using SAS Software. *Journal of Statistical Software*. December 2011, Volume 45, Issue 6. http://www.jstatsoft.org/

EURAREA http://www.ons.gov.uk/ons/guide-method/method-quality/general-methodology/spatial-analysis-and-modelling/eurarea/index.html

The Survey Statistician. http://isi.cbs.nl/iass/allUK.htm

Households' Sample Survey: rotational arrangements

Natallia Bokun

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

Abstract

In this paper the problems of rotational arrangements for Households' Sample Surveys are considered. Different rotation mechanisms, used in EU countries are analysed and systematized. Several variants of households sample frame are proposed to use in Belarus.
Keywords: sampling, repeated sample, rotation design, change over time, overlap

1 Introduction

In a number of sample surveys conducted by government statistical agencies, sample units are observed more than once in a specified rotation design. For example, in a quarterly survey, households might be interviewed several times over a number of quarters and then retired from the sample, while being replaced by another set of units. A rotation design is considered as a compromise between a complete sample overlap (fixed sample), where the units remain in sample indefinitely or for an extended time period, and independent sample, where respondents are contacted only once. Each from these two extreme approaches has advantages and disadvantages.

With a complete sample overlap the variance of an estimator of change may be reduced if there is a strong correlation between estimates at consecutive time points. Cost of data collection might be only raised when contacting a household for the first time. But repeated interviews are connected with a potentially heavier respondent load and possible nonresponses.

Rotation design allows to realize some of the variance reduction of fixed sample and to reduce excess load. Besides repeated measurements within a set rotation design allows to develop more efficient estimators by applying composite estimation.

In Belarus until recent times, fixed and independent (yearly, for enterprises) samples have been used. Widening of sampling practice, stage-by-stage implementation of sampling in practical statistics, constant renovation of the load has caused the need for a specialized rotation design. Rotation schemes for households' surveys are presented.

2 Rotation schemes in EU countries

European countries use different rotation designs for households' surveys, particularly for Labour Force Survey (LFS) (Table 1).

Most of countries apply the rotation scheme 2-(2)-2:

1. Respondents are surveyed four times – firstly in the initial quarter, then in following quarter, a third time a year apart from the initial quarter and the final (the fourth) time – in a quarter thereafter. Households for the sample for two consecutive quarters then are temporary removed for the next two quarters and included again for the following two quarters, thereafter being definitely removed from the survey.

2. Each quarter ¼ of primary sample units is replaced from households sample for a new one.

3. Accordingly with a rotation plan every quarter 50% of the households were also interviewed in the previous quarter and 50% were interviewed in the same quarter of previous year (including 25% of the households, observed in previous quarter and also in the same quarter of previous year). Thus an overlap between consecutive quarters is 50%, between consecutive years is also 50% ($\frac{4-2}{2} = 2/4 = 0.5$).

Rotation schemes used in Czech Republic, Ireland, Netherlands, Austria, United Kingdom, Slovakia (5-), Greece, Spain, France, Hungary (6-) are interesting. These rotation systems are composed of five or six waves (quarters). Each quarter sample contains of five or six sub-samples. For example, 1/5 is surveyed for the second time, 1/5 – for the third, 1/5 – for the fourth, 1/5 – the fifth (and the last) time. Analogous procedure is done for 6 rotation groups. Each household remains in the sample for five (six) quarters.

Rotation plan, used in Germany (4-annual), consists of four waves (rotation quarters). Each sampling district remains in the sample for four years and 25% of the sample is replaced every year. The degree of overlapping between two consecutive yearly samples is 75%. No one is investigated more than once a year.

The choice of optimal rotation system for carrying out of particular households' survey depends on a survey objective, character of the auxiliary information: degrees of noninformation, the sample size, load, availability of the correlated auxiliary information.

3 Rotation schemes in Belarus

Rotation schemes for two households sample surveys, conducted by statistics in Belarus are considered: 1) survey of incomes, expenses and consumption of households (Households Sample Survey); 2) Labour Force Survey.

Households' Sample Survey. Until 2012 fixed sample was used. In 2012 rotation scheme (2-

annual) was adopted. According to this rotation pattern each year ¾ the observed HHs are replaced by new units, ¼ of the selected HHs remains in the survey. Annual (within year) overlap is 25%, quarter overlap (within year) is 100%.

Table 1. Stratification and rotation scheme in the EU – LFS by countries

Country*	Stratification	Frequency of the results	Rotation scheme
Be	Region	Quarterly and annually	1-
BG	Administrative districts: urban/rural	Quarterly	2-(2)-2
CZ	Region	Quarterly	5-
DK	Registered unemployment	Quarterly	2-(2)-2
DE	Region, size of building	Quarterly	4-(annual)
EE	Group of regions by population size	Quarterly	2-(2)-2
IE	Region, urbanization	Quarterly	5-
EL	Region	Quarterly	6-
ES	Population size of municipality, socio-economic characteristics of the population	Quarterly	6-
FR	Region, type of urban unit	Quarterly	6-
IT	Region, size categories of municipalities	Quarterly	2-(2)-2
CY	Districts, urban/rural	Quarterly	6-
LV	Region, urban/rural	Quarterly and annually	2-(2)-2
LT	-	Quarterly	2-(2)-2
LU	Cantons, households' size classes	Quarterly and annually	2-
HU	Administrative units	Monthly and quarterly	6-
MT	-	Quarterly	2-(2)-2
NL	Region	Quarterly	5-
AT	Bundesland	Quarterly	5-
PL	Region, urban/rural	Quarterly	2-(2)-2
PT	Region	Quarterly	6-
RO	Region, urban/rural	Quarterly	2-(2)-2
SL	Region, size and type of settlements (urban/rural)	Quarterly and yearly	3-(1)-2
SK	Region	Quarterly	5-
FI	Region	Monthly, quarterly, yearly	3-(1)-2
SE	Sex, age group, region	Monthly	8-
UK	By frame	Quarterly	5-
HR	Counties, Zagreb	Quarterly	2-(2)-2
MK	Region, urban/rural, size of enumeration district	Quarterly and yearly	2-(2)-2
TR	Region, urban/rural, density	Monthly	2-(2)-2
IS	-	Quarterly	3-(2)-2
NO	County	Quarterly	8-
CH	Region / size of region	Annually	5-(annual)

* Be – Belgium, BG – Bulgaria, CZ – Czech Republic, DK – Denmark, DE – Germany, EE – Estonia, IE– Ireland, EL – Greece, ES – Spain, FR – France, IT – Italy, CY – Cyprus, LV – Latvia, LT – Lithuania, LU – Luxembourg, HU – Hungary, MT – Malta, HL – Netherlands, AT – Austria, PL – Poland, PT – Portugal, RO – Romania, SL – Slovenia, SK – Slovakia, FI – Finland, SE – Sweden, UK – United Kingdom, HR – Croatia, MK – Macedonia, TR – Turkey, IS –Iceland, NO – Norway, CH – Switzerland.

Labour Force Survey. In 2011-2012 the National Statistical Committee of the Republic of Belarus together with some foreign and national experts made a preparatory work on the implementation of the Labour Force Survey. One of the most important topics was rotation design.

Several types of the rotational systems were executed: a) 2-(2)-2; b) 5-; c) 4-(annual). The first variant (Figure 1) is not enough acceptable under a given annual sample size (28000 HHs), a given quarter sample size (7000 HHs) and a limited number of interviewers (200).

Figure 1. Rotation scheme 2-(2)-2

4 rotation waves, overlaps between years is 50%

Year, quarter	Rotation groups							
	A	B	C	D	E	F	G	H
2012 I	1			4	3			2
II	2	1			4	3		
III		2	1			4	3	
IV			2	1			4	3
2013 I	3			2	1			4
II	4	3			2	1		
III		4	3			2	1	
IV			4	3			2	1
2014 I	1			4	3			2
II	2	1			4	3		
III		2	1			4	3	
IV			2	1			4	3

If the quarter sample size (7000 HHs) and the load of interviewer are not changed, year sample size will reduce under the quarter replacement, equal ¼ of the sample, we'll got only 12250 HHs ($7000+3\cdot0.25\cdot7000 = 12250$). As a result, it is not possible to receive representative sample parameters by regions, gender-age groups. If the quarter sample size will increase to 28000 HHs, it results to excess of interviewer's load (more than 2 times exceed normal level).

Figure 2 illustrates a rotation design (5-) in which HHs remain in sample for five consecutive quarters. According to the rotation scheme year sample size (under quarter size – 7000 HHs) will reduce from 28000 to 11200 ($7000+3\cdot0.2\cdot7000 = 12200$). It is less than in rotation scheme (2-(2)-2) and cannot be used for LFS.

The third variant of rotation scheme is more acceptable (Figure 3). Proposed rotation system consists of four year waves. Each surveyed household remains in the sample for four consecutive years. Household is interviewed four times once a year. This rotation design has different samples for each quarter of a year and each sample is generally interviewed occasionally one, two and three years apart. Overlap between quarters is absent; between years it is ¾ or 75%.

Figure 2. Rotation scheme (5-)

5 rotation waves, quarter overlaps is 80%, year overlap – 40%

Year, quarter	Rotation groups							
	A	B	C	D	E	F	G	H
2012 I	1				5	4	3	2
II	2	1				5	4	3
III	3	2	1				5	4
IV	4	3	2	1				5
2013 I	5	4	3	2	1			
II		5	4	3	2	1		
III			5	4	3	2	1	
IV				5	4	3	2	1
2014 I	1				5	4	3	2
II	2	1				5	4	3
III	3	2	1				5	4
IV	4	3	2	1				5

The sample population is formed for four waves (2012-2015). It consists of 28 rotation groups: in each quarter four rotation groups are observed. Every group includes 1750 HHs (7000:4=1750). In 2012 16 groups are surveyed, and then in each following year ¼ of the sample is replaced. The first sample cycle will be finished in 2015.

The sample rotation scheme adopted for LFS balances two concerns: 1) estimation of year-on-year changes and 2) increasing the sample size for regional estimation.

4 Concluding remarks

The use of multistage territorial sampling and rotation design allows receiving representative sample population, to estimate quarter-on quarter and year-on-year changes. After testing several rotation variants we have got that acceptable rotation designs for Households Surveys in Belarus are 2-(annual) and 4-(annual). However, standard errors calculated for the level of unemployment, unemployed in the context of gender-age groups at regional level are rather high (10-15%). To improve the representativeness by regions quarterly indicators of the survey can be formed on the basis of the three samples, the average for three consecutive quarters. It is possible to use alternative rotation schemes, for example, quarter rotation.

Figure 3. Rotation design for LFS. January 2012– December 2019 (4-(annual))

Year, quarter	Rotation groups																											
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB
2012 I	1																4				3				2			
II		1																4				3				2		
III			1																4				3				2	
IV				1																4				3				2
2013 I	2				1																4				3			
II		2				1																4				3		
III			2				1																4				3	
IV				2				1																4				3
2014 I	3				2				1																4			
II		3				2				1																4		
III			3				2				1																4	
IV				3				2				1																4
2015 I	4				3				2				1															
II		4				3				2				1														
III			4				3				2				1													
IV				4				3				2				1												
2016 I					4				3				2				1											
II						4				3				2				1										
III							4				3				2				1									
IV								4				3				2				1								
2017 I									4				3				2				1							
II										4				3				2				1						
III											4				3				2				1					
IV												4				3				2				1				
2018 I													4				3				2				1			
II														4				3				2				1		
III															4				3				2				1	
IV																4				3				2				1
2019 I	1																4				3				2			
II		1																4				3				2		
III			1																4				3				2	
IV				1																4				3				2

References

Bokun, N., Cherhysheva, T. (1997). *Metody vyborochnyh obsledovanij*. Minsk.

Hussmans, R., Mehran, F., Vermo, V. (1994). *Surveys of economically active population, employment, unemployment and underemployment*. Geneva.

Labour Force Survey in the EU, candidate and EFTA countries. Main characteristics of the national surveys, 2008 (2010). Luxemburg: Publications Office of the European Union.

The Choice of a Sampling Design Regarding Survey Cost Efficiency

Mārtiņš Liberts

Central Statistical Bureau of Latvia, e-mail: martins.liberts@gmail.com

Abstract

The aim of sample surveys is to obtain sufficiently precise estimates of population parameters with low cost. The expected precision of estimates and the expected data collection cost are usually unknown making the choice of sampling design a complicated task. Analytical methods can not be used often because of the complexity of the sampling design or data collection process. The aim of this research is to develop a mathematical framework to compare sampling designs of interest with respect to their expected precision of estimates and data collection cost. As a result a framework is developed, which employs artificial population data generation, survey sampling techniques, survey cost modelling, Monte Carlo simulation experiments and other techniques. The framework is applied to analyse the cost efficiency of the Labour Force Survey.
Keywords: cost efficiency; simulation study; survey cost estimation; survey methodology; variance estimation

1 Introduction

The inspiration for this research work comes from pure practical necessity. National Statistical Institutes (NSIs) are the main providers of official statistics in most countries. A large proportion of official statistics produced by NSIs are done using data collected via sample surveys, with the main customer of official statistics being the general public (or tax payers, in other words). These days, cost efficiency is an essential consideration in all government spending; the question is, are NSI sample surveys cost efficient?

There is not a simple answer to the question posed. A sample survey can possess one of many different sampling designs. The simplest sampling designs do not necessarily provide the lowest data collection cost. More complex sampling designs are considered in theory and applied in practice to obtain statistical information with an acceptable precision at a lower cost.

In designing a sample survey, the following considerations should be decided upon: *What is the expected precision of the estimates of population parameters? What is the expected data collection cost? Which sampling design should be chosen in order to minimise sampling errors*

under a fixed data collection cost? These are commonly asked questions during the planning stage of a sample survey. In most cases, the answers to the questions posed cannot be gained through analytical means and NSIs are usually reliant on expert's judgement to some extent.

The relation between the precision of estimates and survey cost has been discussed in literature for at least 70 years, though the topic has not been comprehensively addressed. Different aspects of the relationship have been analysed and different goals of analysis have been set by authors but it is possible to observe the lack of common foundations for the topic. One of the first papers devoted to the topic are by Mahalanobis (1940) and Jessen (1942). The topic is extensively discussed by Hansen (1953) and Kish (1965). Significant book regarding the topic is by Groves (1989). The author advocates simulation studies to be the best-suited for a sample design analysis because of usual complexity of cost and precision functions.

The research of survey field operations is a brand new topic in the scope of statistical research. Several research activities have been devoted to the topic only recently (Chen 2008, Cox 2012). Several events have been organised recently, in the United States of America, devoted to the topics of survey cost estimation and simulation models for survey fieldwork operations, for example “Survey Cost Workshop” (2006, Washington, D.C.) and “Workshop on Microsimulation Models for Surveys” (2011, Washington, D.C.).

In general, the total price for a survey, where data are collected directly from respondents, is increasing. There are several reasons for the increase of the price, but one significant reason is the decreasing of response level. In today's world, either much more effort is needed to increase the cost efficiency of surveys, or a higher price must be paid, in order to produce the same quality of statistics as in times when non-response was not such a big problem. However, given the current economic climate, in most cases it is simply not possible to spend more since most government budgets for surveys are reducing, or at best being kept the same as the previous year's. It is clear, therefore, that increased cost efficiency is crucial to maintain the production of high quality statistics under a decreasing or fixed budget. Since survey sampling emerged as a methodology, problem with non-response and budget restrains has not been met so often. This is one of the main reasons why survey cost efficiency has not been a very important research topic until recently.

Another conclusion is that, simulation experiments are getting more and more attention as a tool used in the designing of production systems for official statistics. The expansion of the method is possible because of a cheap computer power available currently. Even a desktop or a laptop computer nowadays can be set up to solve large scale simulation experiments.

2 Outline of the Lecture

The aim of this lecture is to present a framework which can be used to compare arbitrary sampling designs by their cost efficiency. The framework can be used to analyse selected sampling designs and determine the sampling design that leads to the highest overall precision of estimates under a fixed survey budget. The main points to be discussed during the lecture are:

- The introduction to survey cost efficiency.
- The presentation of the framework for survey cost efficiency analyses using simulation techniques.
- Some results from the study where the framework has been applied in the case of Latvian Labour Force Survey.

References

Chen, B.-C. (2008). *Stochastic simulation of field operations in surveys* (Research Rep.) Washington: U. S. Census Bureau. Retrieved from https://www.census.gov/srd/www/byyear.html

Cox, L. (2012). The case for simulation models of federal surveys. In *Research conference papers of federal committee on statistical methodology research conference 2012*. Washington. Retrieved from http://www.fcsm.gov/events/papers2012.html

Groves, R. M. (1989). *Survey errors and survey costs*. New Jersey: Wiley.

Hansen, M. H., Hurwitz, W. N., & Madow, W. G. (1953). *Sample survey methods and theory* (Vol. I). New-York: Wiley.

Jessen, R. J. (1942). *Statistical investigation of a sample survey for obtaining farm facts* (Research Bulletin No. 304). Iowa State College of Agriculture and Mechanic Arts.

Kish, L. (1965). *Survey sampling.* New-York: John Wiley & Sons.

Mahalanobis, P. C. (1940). A sample survey of the acreage under jute in Bengal. *Sankhyā: The Indian Journal of Statistics*, **4**(4), 511–530.

Contributed papers

Combining sample of three household surveys in Estonia

Julia Aru

University of Tartu and Statistics Estonia, e-mail: julia.aru@stat.ee

Abstract

Statistics Estonia conducts several social surveys with very similar or identical target populations. These surveys focus on different topics, but still contain a block of common questions. Thanks to that, it is possible to combine samples of different surveys to produce more detailed output on these common questions and increase precision. We discuss combining of the samples of three surveys: Survey on Income and Living Conditions, Labour Force Survey and Household Budget Survey. First two are longitudinal surveys with rotational design, and the third one is a purely cross-sectional survey, so the main challenge is computing weights for the combined sample. We describe previous experiment with two surveys, discuss calculation of weights by cumulating probabilities of the three surveys used in analysis and non-response correction models. Empirical comparison of estimates will be presented during the workshop.

Keywords: combining samples, cumulating probabilities

1 Introduction

Statistics Estonia, like any other national statistical office, conducts a lot of social surveys. These surveys focus on various topics and may differ in terms of target population and design, but there is always some overlap in questionnaires, e.g. education, socio-economic status, living conditions etc. In this situation, estimates for common questions can be derived from several surveys. By combining samples of several surveys and estimating common questions from the bigger combined sample, NSI can avoid publishing different estimates for the same phenomenon, which is confusing for users, as well as increase precision of output. This approach is already used for years in several European countries, like Netherlands and UK, while this paper will describe the results of the second exercise of this kind in Statistics Estonia.

In the first experiment (reported last year during the BNU Valmiera workshop) we combined samples of two surveys (Survey on Income and Living Conditions 2010, EU-SILC, and Household Budget Survey 2010, HBS) and tested two approaches to weighting: adjusting existing weights and cumulating probabilities (or calculating weights over from the begging). The surveys had identical target populations and were both yearly surveys, i.e. every household was interviewed just once during the yearly survey circle. The comparison of estimates using different weighting methods showed that method of cumulating probabilities was yielding just marginally lower variance estimates, while being much more difficult to calculate. The final recommendation was to adjust existing weights.

In the present experiment we use data of 2011 and introduce the third survey – Labour Force Survey (LFS). In Estonia LFS has rotational design 2-(2)-2: selected household stays in the survey for two consecutive quarters, then skips two quarters and finally is again in the sample for two quarters. As a result, a large part of households is included twice in the yearly survey sample. To combine LFS with EU-SILC and HBS we first need to form a yearly LFS sample with each household included only once. For the present experiment, a first (earlier) quarter was selected for the households that were included twice. It was done to ensure better comparison with EU-SILC having fieldwork in the first two quarters of the year (HBS sample is divided between all 12 months). The alternative would be to select a quarter randomly.

Since we had to form a yearly sample of LFS, the existing survey weights were not applicable and so it was not possible to use the method of adjusting existing weights. Another obstacle to using a simpler method of adjusting existing weights is the difference in target population. LFS covers only population aged 15 to 74 years old, while EU-SILC and HBS cover all persons in private households.

Other features of three surveys used in this analysis are summarized in Table 1.

Table 1: HBS, EU-SILC and LFS

Feature	HBS	EU-SILC	LFS
Target population	All persons in private households		Persons aged 15-74 in private households
Longitudinal component	No longitudinal component, purely cross-sectional	4-year rotation: every year ca ¼ of the sample is dropped and the new sub-sample is introduced	Rotation pattern 2–(2)–2: two quarters in survey, then skips two quarters and then again two quarters in survey.
Yearly sample	Households not repeated in one year		Some households are included twice
Sampling design	Stratified systematic sampling of persons from the Population Register, with whole household included along with selected person, which results in PPS sampling for households		
Age restriction for sampled persons	15+	14+	15-74
Sample size in 2011 (yearly sample, households)	3600	5000	4700
Common questions in 2011 questionnaires	Living conditions: tenure status, size, type and quality of dwelling, presence of water, electricity etc. Material deprivation: presence of mobile phone, PC, TV, washing machine etc Household members: mother tongue, health condition		-
	Size and type of household, place of residence Welfare: amount of money in disposal of the household in one month Household members: ethnicity, education, country of birth, citizenship, marital status, socio-economic status, employment status, occupation, field of economic activity of employer, type on contract, managerial position, number of working hours in one week, present studies		

As mentioned, it was not possible to use the simpler method of weighting, i.e. adjust existing weights with a scaling factor. So we had to use more complicated method of cumulating probabilities, which in some sense starts from the beginning and calculates the probabilities to be included into the combined sample for each household. In the following section we will describe this method in more detail.

2 Method of cumulating probabilities

With this method, for each household we calculate the probability to be included (and respond) in the combined sample. Here, for each household, we calculate the probability to be included in the combined sample as a whole, i.e. to be included in one of the samples, independently on which it was really included in. So, for example, for a household from HBS we need to calculate the probability that it would have been included in EU-SILC and the probability that it would have been included in LFS, and vice versa, taking into account survey-specific response pattern. The same method of weighting the combined sample was applied in Iachan *et al.* (2003) and O'Muircheartaigh & Pedlow (2002), but without modelling the response mechanism.

For surveys with longitudinal part, as EU-SILC and LFS, it is useful to model the response separately for those who have responded to survey request at least once and are approached repeatedly and those who have never responded before, because of different response rates and different amount of information available for non-respondents. Thus, EU-SILC and LFS are divided into the new part and the repeated part. In what follows we treat the combined sample as the concatenation of five (instead of three) surveys: EU-SILC repeated part, EU-SILC new part, LFS repeated part, LFS new part and HBS. We call those five computational surveys to distinguish from the three real surveys being combined.

For repeated parts of LFS and EU-SILC, probability to respond in 2011 is modelled as the product of probability to respond in the first year (in the year of first selection into the sample) and probability to respond in 2011 (given the household has responded in the year of selection).

We use following notation:

S – combined sample;

R – response set for the combined sample;

S_{SR}, R_{SR}, R'_{SR} – respectively the sample, first year response set and 2011 response set for EU-SILC repeated part

S_{SN}, R_{SN} – sample and 2011 response set for EU-SILC new part;

S_{LR}, R_{LR}, R'_{LR} – respectively the sample, first year response set and 2011 response set for LFS repeated part

S_{LN}, R_{LN} – sample and 2011 response set for LFS new part;

S_H, R_H – sample and 2011 response set, HBS;

As $R = R'_{SR} \cup R_{SN} \cup R'_{LR} \cup R_{LN} \cup R_H$ and surveys are negatively coordinated (households already participating in one of the surveys are dropped prior to the sample selection for the other) the probability of household *i* to be included into the combined sample and respond can be written as:

$$\begin{aligned}\Pr(i \in R) &= \Pr(i \in R'_{SR}) + \Pr(i \in R_{SN}) + \Pr(i \in R'_{LR}) + \Pr(i \in R_{LN}) + \Pr(i \in R_H) = \\ &= \Pr(i \in R'_{\mathrm{SR}} \mid i \in R_{SR}) \Pr(i \in R_{SR} \mid i \in S_{SR}) \Pr(i \in S_{SR}) + \\ &+ \Pr(i \in R_{SN} \mid i \in S_{SN}) \Pr(i \in S_{SN}) + \\ &+ \Pr(i \in R'_{\mathrm{LR}} \mid i \in R_{LR}) \Pr(i \in R_{LR} \mid i \in S_{LR}) \Pr(i \in S_{LR}) + \\ &+ \Pr(i \in R_{LN} \mid i \in S_{LN}) \Pr(i \in S_{LN}) + \\ &+ \Pr(i \in R_H \mid i \in S_H) \Pr(i \in S_H).\end{aligned} \quad (1)$$

That is, for every household in the combined sample we need to calculate the inclusion probability and the response probability to every one of five computational surveys (i.e. five inclusion probabilities and five response probabilities).

3 Response models

Inclusion probabilities can easily be calculated, because we know the sampling design of each survey and have all the required design variables (in our case county and household size). Response probabilities need to be estimated based on the response model that still needs to be specified.

Since we no longer need weights to be comparable with existing weights, we are able to test different response models. Auxiliary variables have to be selected among those available in all three surveys and for both respondents and non-respondents.

For households never responded before (HBS, new parts of EU-SILC and LFS) we can use register information available for sample persons and municipality-level averages available from 2011 Census. For interviewed households we can additionally use variables in common for all the tree surveys (see Table 1).

Separate response models had to be fitted for every one of five computational surveys, since surveys had different response rates and patterns. For repeated parts of EU-SILC and LFS we decided to test two models: short model using only register variables and area-level averages and long model using also the information available from the last interview. While it is clear that the long model would fit better potentially, we still need to explore the effect of that on the final estimates, since in practice it is much easier to use just register and area-level variables for all the computational surveys than link the datasets of repeated surveys additionally to previous years.

Thus, a probability that a household was included in the combined sample and responded was

calculated for every household in the combined sample according to equation (1). Reciprocals of these probabilities are non-response corrected weights, which were calibrated by sex, 5-years age group and county before the calculation of estimates. Finally, we got two sets of weights: with short non-response models and with long non-response models in the repeated parts of EU-SILC and LFS.

3 Comparison of weights and estimates

Empirical comparison of resulting weights and comparison of estimates they yield to Census 2011 totals will be presented during the workshop.

References

Iachan, R., Saaverda, P. & Robb, W. (2003). Two weighting schemes for combining sample in the Health Behaviors in School-age Children Survey. *2003 Joint Statistical Meetings - Section on Survey Research Methods*

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

European Social Survey mixed mode experiment in Estonia: CAWI and CAPI sequential design

Mare Ainsaar [1], Laur Lilleoja [2], Kaur Lumiste[3] and Ave Roots[4]

[1] Univeristy of Tartu, e-mail: mare.ainsaar@ut.ee
[2]Univeristy of Tartu, e-mail: laur.lilleoja@ut.ee
[3]Univeristy of Tartu, e-mail: kaur.lumiste@ut.ee
[4]Univeristy of Tartu, e-mail: ave.roots@ut.ee

Abstract

In autumn of 2012 the European Social Survey (ESS) conducted a mixed mode survey experiment under the supervision of Centre for Comparative Social Surveys at City University London, UK. Alongside UK, Sweden and Slovenia, Estonia was one of the participating countries. In this mixed mode survey, respondents were asked to answer the ESS questionnaire online and if there was no reaction to the postal invitation and the two reminders, the respondent was referred to CAPI mode. Current paper presents the first results and preliminary analysis of that experiment in Estonia.
Keywords: European Social Survey, mixed-mode survey, sequential design, CAWI.

1 Introduction

The European Social Survey (the ESS) is an academically-driven social survey designed to chart and explain the interaction between Europe's changing institutions and the attitudes, beliefs and behaviour patterns of its diverse populations. The ESS was established in 2001, and was led by its founder and coordinator Roger Jowell until his death in December 2011.

Currently in the midst of its sixth round, this biennial cross-sectional survey covers more than thirty nations and employs the most rigorous methodologies. The ESS has received funding from the European Commission's Framework programmes, from the European Science Foundation, and from national funding councils in participating countries. In January 2013 the ESS applied for selection as a European Research Infrastructure Consortium (ERIC).

The experiment in general was carried out as a part of the ESS Mixed-mode Design research programme, the participating countries were the UK, Sweden, Slovenia and Estonia, and different settings of three data collection modes were considered: CATI, CAWI and CAPI. The

aim of the experiment was to investigate the feasibility of using a mixed-mode data collection method in the ESS and which modes of data collection should be allowed.

We present an overview of the setup and preliminary results of the ESS mixed-mode experiment held in Estonia in 2012. Estonian part of the experiment involved only CAWI and CAPI modes, respondents were asked to fill in the ESS questionnaire online and if this had not been done within 5 weeks, the respondent received a visit from an interviewer to do a face-to-face interview.

Data collection was carried out simultaneously with the main ESS Round 6 survey in autumn and winter of 2012. All stages of the survey were prepared and carried out by the same fieldwork agency and all materials and questions resembled the main ESS as much as possible. The survey was carried out in Estonian and Russian languages, as usual in the Estonian ESS.

2 Setup and fieldwork of the experiment

The mixed-mode experiment was run alongside the main ESS Round 6 data collection, so that the results of the experimental method could be compared with the results of the traditional method. This gives an opportunity to detect possible mode effects and assess the mixed mode design as feasible data collection method.

At first stage the main ESS questionnaire was translated into Russian and Estonian following all ESS translation requirements. A large portion of the process had to be done again because the experiment needed a modified ESS questionnaire because of CAWI specifics:

1. Change of written orders („please write“ in web-survey, instead of „please tell me“; “Please use card 45” etc.);
2. Transformation of show cards (cards with answer options, shown to the respondents in CAPI) to CAWI mode specifics;
3. Extra texts that informed the respondent if (s)he had left a question unanswered or might have mistyped or gave an illogical answer (e.g. indicated that (s)he works more than 168 hours a week or years of completed education was greater than the persons age);
4. Block of questions, usually answered by the interviewer, needed a new version compared with CAPI.

All together about 85% of a main ESS questionnaire needed additional treatment as a result of these amendments.

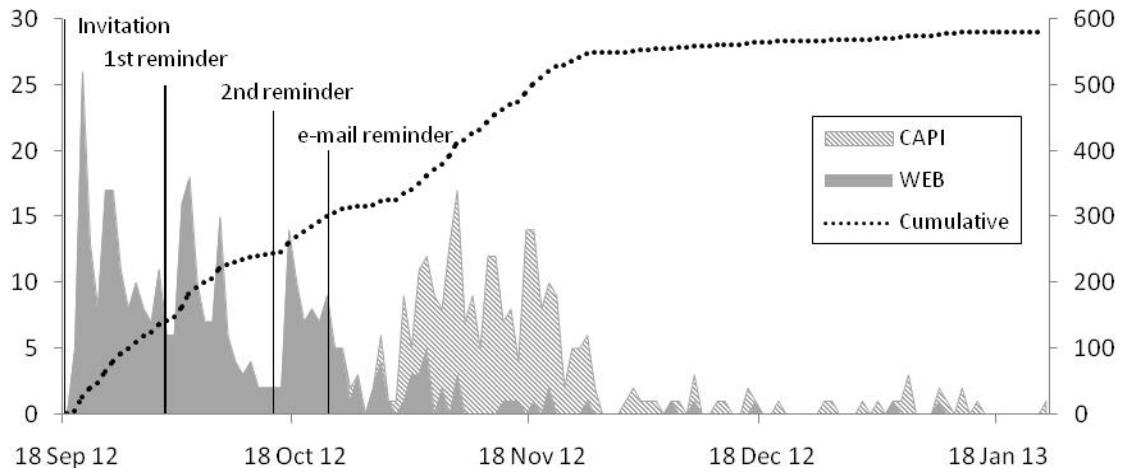
The questionnaires for both modes were programmed using Confirm it surveying platform and although the questions were similar, two separately acting questionnaires had to be programmed.

Population register based individual sample was formed and an invitation letter was sent to sampled persons’ postal address on 18 September 2012. The letter included information an individual password to logon to the online survey. The letter also included information that if

the respondent will not/cannot fill the survey on the Internet he/she will be recruited for a face-to-face interview and will be visited by an interviewer during the fieldwork period.

If sampled persons did not finish the questionnaire or did not give any information after two weeks of the invitation, they received a reminder letter on 1 October 2012. The reminder had a clear effect on response activity (Figure 1).

Figure 1: Completed questionnaires in a day by survey mode



Second reminder letter was sent out again after two weeks on 16 October 2012. By that time 241 respondents had finished their survey in internet, 30 had started, but had not finished and 27 persons had passed information about themselves to the fieldwork agency. In 22 October the third reminder was sent to e-mail addresses of 15 respondents who had started filling in the questionnaire online, gave their e-maid address, but did not manage to finish the survey.

Table 1: Response statistics of the main ESS and the mixed mode survey in Estonia

	ESS main survey	ESS mixed mode	
		Total	CAWI mode
Initial sample	3707	927	927
Number of completed interviews	2 380	587	356
Response rate	67.8%	66.0%	40.3%
Refusal rate	12 %	8.2%	
Non-contacts	20.2%	25.8%	-
Ineligibles	5.3%	4.0%	-
Survey period	6. Sept 2012 – 13. Jan 2013	19. Sept 2012 – 13. Jan 2013	19. Sept 2012 – 10. Jan 2013

On 22 October CAPI stage started – sampled persons who had not responded in CAWI mode an interviewer was sent to interview the respondent. However, the online survey environment remained open, and all together 49 respondents filled in the questionnaire online after the start

of the CAPI phase and before an interviewer made a visit (Figure 1). Final response statistics can be seen on Table 1.

3 Sampling and surveyed sample

The sampling procedure for mixed mode experiment was similar to main ESS sampling in Estonia. The frame was taken from a population register and was individual random sample with full coverage of the population 15 and older. The net sample of mixed mode was planned to have not less than500 persons. The target response rate was set at 65% and 17% of the sample was estimated to consist of ineligibles.

3.1 Sampling design

Explicit and implicit stratified systematic random sampling with proportional allocation in stratum was used. Register of persons aged 15 and above with permanent residence was divided into explicit strata by their registered address according to NUTS III areas (5 regions). Each explicit stratum was then ordered by ID numbers (that reflect gender and age). The size (number of persons aged 15+) of each explicit stratum and its proportion were found from the extracted register mentioned above. The gross sample size was multiplied by respective proportion in order to get the sample sizes for respective explicit stratum (proportional allocation). Systematic random samples were drawn from each explicit stratum resulting in implicit stratification of each stratum (age and gender). For implicit stratum, proportional allocation was approximate, respectively.

3.2 Surveyed sample

We expected that with the application of mixed survey mode there would be a rise of overall response rate:

- Younger and highly mobile respondents were expected to be more active in participating in the CAWI mode;
- An online survey might be more attractive and motivate respondents to participate – secure home environment and absence of a face-to-face contact with the interviewer might encourage middle aged reluctant respondents to participate.

The final response rate for the mixed mode survey was 66% which did not differ much from the ESS main survey response rate. The final achieved total mixed-mode sample was representative according to region of residence, age groups, education and gender in the sample. Differences between Estonian population census 2011 data and the collected sample were very small (Tables 2, 3, 4). Mixed-mode sample was also very similar to the main ESS achieved sample.

Table 2: Distribution of respondents in the mixed-mode experiment according to gender and NUTS3 Estonia region and comparison with population census 2011 and main ESS.

Region	Gender	Census	Mixed mode sample	Mixed-mode experiment respondents			ESS main respondent
				Total	Web	CAPI	
North	Male	19%	18%	15%	18%	11%	15%

	Female	23%	23%	22%	24%	19%	22%
West	Male	5%	6%	5%	5%	6%	5%
	Female	6%	6%	7%	6%	9%	7%
Central	Male	4%	5%	5%	4%	6%	5%
	Female	5%	5%	6%	6%	6%	6%
Nort-West	Male	5%	5%	6%	5%	6%	6%
	Female	7%	7%	7%	5%	11%	7%
South	Male	11%	11%	11%	12%	10%	12%
	Female	13%	13%	14%	14%	15%	16%

The main difference of the survey modes was that younger, more educated and living in North-Estonia (more urbanized) tended to prefer the CAWI mode. There are no gender differences in mode preferences.

Table 3: Distribution of respondents in the mixed-mode experiment according to gender and age group and comparison with population census 2011and main ESS.

Age group	Gender	Census	Mixed-mode sample	Mixed-mode experiment respondents			ESS main respondents
				Total	Web	CAPI	
15-24	Male	8%	7%	7%	9%	4%	7%
	Female	7%	7%	8%	10%	6%	7%
25-34	Male	8%	9%	6%	7%	4%	6%
	Female	8%	9%	8%	10%	5%	7%
35-44	Male	8%	8%	6%	8%	4%	7%
	Female	8%	8%	8%	9%	7%	8%
45-54	Male	8%	8%	8%	8%	8%	7%
	Female	8%	9%	9%	8%	9%	9%
55-64	Male	7%	7%	7%	6%	7%	7%
	Female	9%	8%	10%	10%	10%	10%
65+	Male	7%	7%	9%	6%	13%	8%
	Female	14%	14%	14%	7%	23%	17%

Table 4: Distribution of respondents in the mixed-mode experiment according to gender and education and comparison with population census 2011 and main ESS.

Education (ISCED)	Gender	Census	Mixed-mode experiment respondents			ESS main respondents
			Total	Web	CAPI	
ISCED 0	Male	0%	0%			0%
	Female	0%	0%		0%	0%
ISCED 1	Male	2%	2%	1%	3%	1%
	Female	3%	2%	1%	3%	2%
ISCED 2	Male	11%	6%	6%	7%	8%
	Female	9%	8%	6%	11%	10%
ISCED 3	Male	18%	17%	16%	17%	18%
	Female	18%	20%	17%	24%	18%
ISCED 4	Male	3%	5%	4%	6%	6%
	Female	4%	7%	5%	9%	10%
ISCED 5	Male	11%	12%	16%	6%	9%
	Female	20%	20%	25%	13%	18%
ISCED 6	Male	0%	1%	1%		0%
	Female	0%	1%	1%		0%

Conclusions

As a conclusion we can state that CAWI mode helped to capture those respondents who usually

tend to be difficult to reach, but it did not improve the final respondent structure of the mixed mode experiment compared with the main ESS respondent structure. One explanation might be that the addresses received from the register, in many cases, were not correct – people do not live in their registered addresses, especially younger persons (usually live in rented apartments, registered address is kept the same with parents or are away from home studying in another town, etc.) – and therefore invitations to participate in CAWI did not reach them, but were later tracked down in the CAPI mode. Substantive analysis is yet to be made.

The selection of causes of death, related to alcohol consumption, on the basis of current statistics and sample survey

Anastacia Bobrova

Institute of Economic of National Academy of Sciences of Belarus

Abstract

It is believed that the hazardous alcohol consumption is one of the main factors influencing the past and recent mortality in Belarus. The present paper concentrates on identifying the main alcohol causes of death. The demographic losses related to alcohol consumption represented by the number of people who had died because of hazardous drinking. The direct and indirect causes of death are taken into consideration in the estimations of the total number of deaths. The findings were compared with the results of special survey conducted in RSPC "Mental Health".

At first, using the method proposed by Nemtsov and Terehin (2007), the number of deaths from direct and indirect alcohol causes are evaluated. The results of the factor analysis of mortality and alcohol consumption in Belarus during 1970-99, conducted by Razvodovsky (2003), are taken into account.

The data on the reconstructed causes of deaths (Grigoriev P., Meslé, F., Vallin, J, 2012) for the period 1990-2010 are used, separately for men and women. Additionally, the population exposure is taken from the Human Mortality Database. For the standardization the WHO method is applied. To evaluate the alcohol contribution to premature mortality the annual total standardized death rates on the i-th cause in the j-th year are used.

The first step of research is the selection of certain causes of death that could be connected to alcohol consumption. The following causes are considered: hypertension, cerebrovascular diseases, respiratory diseases with the exception of influenza and pneumonia, peptic ulcer, pulmonary tuberculosis, ischemic heart disease, alcohol poisoning, accidents with vehicles, alcoholism, alcoholic psychosis, liver cirrhosis, other heart disease, pancreatitis and other external reasons. The correlation results reveal that only 4 of them are associated with the per capita alcohol consumption (ischemic heart disease, pulmonary tuberculosis, liver cirrhosis and pancreatitis). The regression results also show the significant association between these causes of deaths and the alcohol consumption.

According to the method proposed by Nemtsov and Terehin, the alcohol contribution to

mortality can be expressed as the ratio of the difference between mortality rate at the maximum and the minimum levels of consumption to mortality rate at the maximum level. During the 1990-2010 in Belarus, the minimum level of consumption was 5.7 liters, while the maximum – 12.5 liters. Taken this in account, the mortality rates by each cause of death at both limits of alcohol consumption are calculated and the differences between the results are evaluated. The alcohol contribution to the selected causes of death is the product of the average number of deaths from each cause to the share of alcohol consumption in these causes. The table below summarizes the results for men.

Table 1. The evaluation of the annual mortality caused by alcohol consumption for men, Belarus

The causes of death	constant term	β	Mortality when alcohol consumption is 12.5l	Mortality when alcohol consumption is 5.7l	Differe-nce	Ratio	Share of mortality due to alcohol consumption, %	Average number of deaths per year	among them due to alcohol
1.Alcoholism	-6,891156	1,429425	11,69	1,26	10,4	0,8925218	100	277	277
2.IHD	399,6713	22,05989	686,45	525,41	161,0	0,2345942	26,28442	19368	5091
3.Tuberculosis	6,192962	0,8980971	17,87	11,31	6,5	0,3669144	41,10985	644	265
4.Liver cirrhosis	-20,30238	4,876312	43,09	7,49	35,6	0,8261161	92,55977	1049	971
5.Pancreatitis	-2,557801	0,9089804	9,26	2,62	6,6	0,7166645	80,29658	259	208
							In a year	Sum 2-5	6535

Analogously, the evaluation of alcohol contribution to female mortality is done. The results are similar but in cases of ischemic heart disease and pulmonary tuberculosis no significance is found. To some extent, this is due to the lower prevalence of alcohol consumption among women and the persistent high mortality in women from cardiovascular disease. Compared to men, women are less likely to die from external causes, respiratory diseases, digestive diseases, etc. because of the different lifestyle. In addition, the life expectancy for women is sign ificantly higher than for men, which also increases the risk of dying from heart disease. In the table below only 3 indirect causes of death are presented.

Table 2. The evaluation of the annual mortality caused by alcohol consumption for women, Belarus

The causes of death	Constant term	B	Mortality when alcohol consumption is 12.5l	Mortality when alcohol consumption is 5.7l	Differe-nce	Ratio	Share of mortality due to alcohol consumption, %	Average number of deaths per year	Among them due to alcohol
1.Alcoholism	-1,386308	0,2775051	2,22	0,19	2,03	0,912	100	60	60
2.Liver cirrhosis	-12,57565	2,706823	22,61	2,85	19,76	0,874	95,8	690	661
3.Pancretitis	-0,483757	0,2259788	2,45	0,80	1,65	0,672	73,7	128	94
4.Tuberculosis	0,791019	0,1242589	2,41	1,50	0,91	0,377	41,3	85	35
							In a year	Sum	791

Thus, the total losses caused by alcohol consumption are about 225 thsd. people, including 64 thsd. from the direct causes and 161 thsd from the indirect. The alcohol mortality for women is by 8 times lower than for men.

Due to results of special one-time survey conducted in RSPC "Mental Health" in July 2011 it became possible to identify which problem was felt by alcoholics themselves. According to survey program were examined 8% of the total cumulated number of patients. Sample - repetition-free, quasi-random. Preference for self-random sample was given by the fact that this type of sample selected in strict accordance with the theory of probability and reflects the variability of trait in the general population (Bokun, Chernysheva, 1997).The general population - patients with alcoholism under treatment. The sample was single-stage, big and single-phase. The results of calculation of the amount of sampling error of 5% suggest adequate representation.

One of the question was “Which problem with health do you feel?”. The next answers were suggested:

1. Cardiovascular disease;
2. Diseases of the digestive system;
3. Headache, depression;
4. Neoplasm;
5. Other.

The respondents could choose more than one answer.

The results have confirmed the established relation between alcohol consumption and certain diseases.

So as of women 60 % of respondents had diseases of the digestive system including the liver disease. More than 40 % felt bad because of headache and depression. About 15% had some cardiovascular disease including ischemic heart disease.

The most common problem of health among men was headache and depression – more than 40 %. About 30% of respondents had diseases of the digestive system. The share of men had problem with heart and vascular system the same. Also among answers was the problem with respiratory and nervous systems (15% each).

Given the results of the calculation and the survey, it can be argued that alcohol consumption has a significant impact on public health and is one of the reasons for the high mortality rate in Belarus. The main indirect “alcoholic” problem are cardiovascular disease; diseases of the digestive system; headache, depression.

To sum up, serious actions should be taken instantly at national and local levels in order to prevent both the alcohol production and consumption. For instance, promotion of healthy life style, particularly among the younger generations, should become a priority in all governmental programs.

References

Grigoriev. P., Meslé. F., Vallin.J. (2012). Reconstruction of continuous time series of mortality by cause of death in Belarus. 1965–2010.MPIDR Working Paper WP-2012-023.

Nemtsov A., Terehin A. (2007). Razmery i diagnostichesliisostavalcogolnoismertnosti v Rossii (Dimention and diagnostic structure of the alcohol mortality in Russia) // Narcologiya. – 2007. - №12. - p.29-36.

Razvodovsky Y. (2003). Alcogol i smertnost' v Belarusi (Alcohol and mortality in Belarus). Grodno medicine university. Grodno. – 2003. – 75 p.

Survey Sampling Student Priorities in Selecting Specialty

Iana Bondarenko[1], Valery Turchyn[2], Elizabeth Burchak[3]

[1] Oles Honchar Dnipropetrovsk National University, Ukraine
e-mail: iana.s.bondarenko@gmail.com
[2] Oles Honchar Dnipropetrovsk National University, Ukraine
e-mail: vnturchyn@gmail.com
[3] Master Student at the Oles Honchar Dnipropetrovsk National University, Ukraine
e-mail: elizabethglory@mail.ru

Abstract

The aim is planning, conducting and analysis of the results of survey sampling ″Students motivation in selecting Oles Honchar Dnipropetrovsk National University for higher education and training on the chosen specialty″. Methods of study are stratified simple random sampling, parameter estimation, construction the confidence intervals for the unknown parameters, testing statistical hypothesis.
Keywords: population, sample, stratified simple random sampling, confidence interval, accuracy, reliability

1 Introduction

Applicants motivation in selecting Oles Honchar Dnipropetrovsk National University (DNU) to study the chosen specialization and higher education is a key issue that constantly examines as University leaders and lecturers to improve existing programs of disciplines, to introduce modern information technologies for learning, to achieve requirements that apply in Ukraine and in the world to classic universities, to preserve and to strengthen position of the University as a leading University central region of Ukraine. Thereby it is necessary to investigate the opinions of students about educational process in DNU for adapting Oles Honchar Dnipropetrovsk National University new social conditions.

2 Planning Survey

Questionnaire "Students motivation in selecting Oles Gonchar Dnipropetrovsk National University for higher education and training on the chosen specialty" was developed during September and October 2011 at the Department of Statistics and Probability Theory. Questionnaire consists of five blocks of questions. The first set of questions is devoted the criteria for selecting our University. The second block of questions includes criteria for selecting direction (specialty). The third set of questions concerns the educational process in the direction

(specialty). The fourth set of questions contains questions about plans after graduation. The fifth block of questions consists of questionnaire data of student. Each question provides the answers. The student can choose one variant that reflects his position and experience.

Continuous pilot survey of second-year students of all specialties of Faculty Ukrainian and Foreign Languages were conducted for the purpose of testing and improving the questionnaire. The final version of the questionnaire was prepared on the results of testing questionnaire. This version was used in conducting survey sampling in Mechanics and Mathematics Faculty and Faculty of Biology, Ecology and Medicine.

Let us consider planning a survey on Mechanics and Mathematics Faculty. The population consists of $N = 525$ students the Mechanics and Mathematics Faculty. A stratification of the population of the course: first, second, third, fourth, fifth and specialty: Mechanics (Mech), Mathematics (Math), Statistics (Stat), Heat Power Engineering (HPE) are presented in Table 1.

Table 1: Stratification of the population of students the Mechanics and Mathematics Faculty

Direction (Specialty)	Course					Total
	1	2	3	4	5	
Mechanics	20	48	28	36	37	169
Mathematics	39	75	53	50	39	256
Statistics	15	–	–	–	12	27
Heat Power Engineering	11	13	17	19	13	73
Total	85	136	98	105	101	525

One of the main issues in the planning of sampling is the issue of sample size to obtain the survey results of the required accuracy and reliability. Construct a confidence interval in which the unknown parameter p – the proportion of the population units with a certain property – is located with the required accuracy e and reliability $1-\alpha$

$$P\{\hat{p} - e \le p \le \hat{p} + e\} = 1 - \alpha \tag{1}$$

where accuracy e is calculated according to the formula

$$e = x_{1-\alpha/2}\sqrt{D\hat{p}} \tag{2}$$

where $x_{1-\alpha/2} - (1-\alpha/2)$ - quantile of normal distribution $N_{0;1}$.

The sample size n for estimation the proportion of the population units with certain property is

$$n = \frac{N \cdot x_{1-\alpha/2}^2 \cdot p \cdot (1-p)}{Ne^2 + x_{1-\alpha/2}^2 \cdot p \cdot (1-p)}. \tag{3}$$

The sample size n depends on the unknown p. Population proportion p of units with certain properties takes the values from 0,05 to 0,95 according to the results of the pilot survey. We choose the largest sample size (222) of all the calculated sample sizes for a population of 525 students the Mechanics and Mathematics Faculty.

The resulting sample size must be allocated to strata. Calculate the sample size in each stratum according to the proportional allocation. Proportional allocation is determined according to

$$n_h = \frac{nN_h}{N}, \quad h = 1,...,H, \tag{4}$$

where n_h – number of sampled units in the *h*th stratum; N_h – number of units in the *h*th stratum in the population**.** Stratification results of the sample students of Mechanics and Mathematics Faculty are presented in Table 2.

Table 2: Stratification of the sample of students the Mechanics and Mathematics Faculty

Direction (Specialty)	Course					Total
	1	2	3	4	5	
Mechanics	8	20	12	15	16	72
Mathematics	17	32	22	21	17	108
Statistics	6	–	–	–	5	11
Heat Power Engineering	5	6	7	8	6	31
Total	36	58	41	44	43	222

The sample for the survey selected from each stratum using Simple Random Sampling. We use the following scheme.
1. Systematize alphabetically the last names of students in each stratum.
2. Simulate sequence of uniformly distributed random variables in the interval (0, 1).
3. Each student is associated with uniformly distributed random variable in the interval (0, 1).
4. Organize uniformly distributed values in the interval (0, 1) in ascending order.
5. The first n students are selected from a list for the interview.

Planning survey for Faculty of Biology, Ecology and Medicine is similar.

3 Estimation proportion of students which has chosen the certain variant of answer to the question

π-estimate for proportion of units with a certain property is

$$\hat{p} = \frac{1}{N}\sum_{h=1}^{H} N_h p_h = \sum_{h=1}^{H} \frac{N_h}{N} p_h = \sum_{h=1}^{H} W_h p_h, \tag{5}$$

where $p_h = \frac{a_h}{n_h}$ is proportion of units with a certain property in the *h*th stratum.

For each specialty estimate the proportion of students which has chosen the certain variant of the answer to the question was calculated as the weighted sum estimates the proportion of students each course which has chosen the certain variant of the answer to the question

$$\hat{p} = \frac{N_1}{N}\frac{a_1}{n_1} + \frac{N_2}{N}\frac{a_2}{n_2} + \frac{N_3}{N}\frac{a_3}{n_3} + \frac{N_4}{N}\frac{a_4}{n_4} + \frac{N_5}{N}\frac{a_5}{n_5} \quad (6)$$

where a_i – number of students of the specialty in the *ith* course, which has chosen the certain variant of the answer; n_i – number of students of the specialty in the *ith* course in sample; N_i – number of students of the specialty in the *ith* course in population; N – number of students of the specialty in population.

For each course estimate the proportion of students which has chosen the certain variant of the answer to the question was calculated as the weighted sum estimates the proportion of students each specialty which has chosen the certain variant of the answer to the question

$$\hat{p}_j = \frac{N_{Mech}}{N}\frac{a_{Mech}}{n_{Mech}} + \frac{N_{Math}}{N}\frac{a_{Math}}{n_{Math}} + \frac{N_{MS}}{N}\frac{a_{MS}}{n_{MS}} + \frac{N_{HPE}}{N}\frac{a_{HPE}}{n_{HPE}}, \quad j = 1,\ldots,5 \quad (7)$$

where a_{Mech} – number of students *Mechanics* in the *jth* course, which has chosen the certain variant of the answer; n_{Mech} – number of students *Mechanics* in the *jth* course in the sample; N_{Mech} – number of students *Mechanics* in the *jth* course in the population; N – number of students in the *jth* course in population. Notations for *Math, MS, HPE* are similar.

We used the χ^2-criterion in order to compare proportions of students which has chosen the certain variant of the answer to the question. A hypothesis is $H_0 : p_1 = p_2$ (proportions of students in Mechanics & Mathematics Faculty and Faculty of Biology, Ecology and Medicine which has chosen the certain variant of the answer to the question, are equal). Namely, priorities for the selection of University for higher education and training on the chosen specialty are the same. Alternative hypothesis $H_1 : p_1 \neq p_2$ (priorities are different).

4 Analysis of results

We will consider the results of sampling survey by the specialties.

The question "Which did you use criteria when you were taking decision about joining the DNU?" was asked to students of all specialties the Mechanics and Mathematics Faculty. The

preference was given for "reputation of the University", "the quality of education and learning conditions", "the existence of specialty at the University" (61% Math, 76% Stat, 56% Mech, 58% HPE). "Advice of the parents, relatives, school teachers, University lecturers" were decisive for 26% Math, 17% Stat, 27% Mech and 33% HPE. "As luck would have it" 6% Mech and 4% Math had joined to the University. For these students were indifferently in which University they had to go for studying, just for joining. This criterion for selecting University can be considered not significant (results are within the boundaries of the accuracy 5%). Thus, the reputation of the University (important component is lecturers reputation), the quality of education and the learning conditions are working effectively for attracting the applicants.

To obtain information about the University students had used reference books about Universities and University official site on the Internet (http://www.dnu.dp.ua/) (totally so say 70% Math, 72% Stat, 78% Mech and 85% HPE). The Open Days provided the necessary information about DNU for 23% Math, 7% Stat, 17% Mech, 9% HPE. Advertising in the media is the smallest proportion in student answers (1,4% Math, 2,8% Stat, 1,4% Mech, 6,2% HPE). This source of information can be considered not significant (results are within the boundaries of the accuracy 5%).

More than half of students specialties Math, Stat, Mech did not attend Open Days in DNU (53% Math, 73% Stat, 58% Mech, 47% HPE). All still Open Days have remained useful event on which applicants can directly communicate with the lecturers and the Heads of Faculties. Thanks to this event 18% Math, 9% Stat, 13% Mech and 31% HPE had been decided to go to University. Certain proportion of students attended Open Days, but this event did not crucial in selection DNU for training.

Popular answer for the question "Whether had been influenced on your selection University attendance of DNU lecturers to your school?" was "Lecturers did not attend my school", the second most popular answer was "No, did not affect" (totally so say 83% Math, 64 % Stat, 87% Mech, 78% HPE). Perhaps it is necessary to think about the effectiveness of career guidance. It should be paid more attention not passive guidance (presentation of the DNU in schools), and active form (to teach innovative methods for solving problems on elective courses, to attract applicants to competitions, tournaments conducted by University).

Students of all specialties Faculty of Mechanics and Mathematics are expecting a classic university education and not highly specialized education, are looking for deep theoretical knowledge and practical skills in training with highly qualified lecturers. Diploma DNU is guarantee such education for 79% Math, 74% Stat, 65% Mech and 60% HPE. It should be noted that for the quarter of Mech and HPE students more important for learning is a fun and friendly student group.

"Obtaining quality education" and "perspective of the prestigious work in Ukraine" is more expectative for students among the learning outcomes (totally so say 78% Math, 82% Stat, 58% Mech, 52% HPE). "Possibility of having the prestige work abroad" and "new friends and useful familiarity" were received lower proportion of answers.

Students of different specialties have different plans after graduation. 60% of Stat students are

planning to work in the specialty. This proportion is twice as much as at the speciality of Math, Mech, HPE. It should be noted that specialty Statistic is a single specialty, in which all students have plans for the future as compared with Math (12%), Mech (21%), HPE (19%). After graduation 9% Math, 7% Mech, 12% HPE want to continue their training at the University and to attend postgraduate studies (actually it's not so little). 25% Math, 18% Mech, 14% Stat, 6% HPE are planning to work, but not by specialty. 25% Math, 25% Stat, 25% Mech and 41% HPE are going to obtain a second specialty.

Students of all specialties consider that the most perspective method of job search is find jobs via the Internet (so say more than half of Math and Stat, 38% Mech, 31% HPE). 29% Math, 29% Stat, 37% Mech and 47% HPE are expecting to get a job by the results passing practices.

Selection of the University, of course, is closely connected with the selection specialty. In this regard, the selection criteria of the specialty play great importance. Students of all specialties gave preference for "*requests for profession*" and "*prestige of specialty in society*" (so say 24% Math, 59% Stat, 36% Mech, 46% HPE). Interestingly, 15% Math, 19% Mech and 12% HPE and no statistics randomly had chosen specialty for studying. Advice of the parents, school teachers, University lecturers were crucial to 46% Math, 27% Stat, 35% Mech and 36% HPE.

The question "Would you have chosen the specialty on which you are studying, if you had to choose the profession again?" was asked to students of all specialties. A lot of students would be back to their preferred specialty. The highest percentage of "Yes" presented Statisticians (82%), the lowest percentage of "Yes" indicated Mechanics (60%). Opinions were divided among those students which would not choose specialty if you had to join again (Almost 50% of Math, Stat, Mech consider that the disciplines are not focused on the applying in future careers. Quarter of Math and HPE called unfair the assessment of attainments.)

Students of all specialties assessed teaching level as high. Thus 100% Stat find that lecturers are highly qualified in their field, for Math, Mech, HPE these proportions are 83%, 80% and 50% respectively.

Employers, business environment is rarely involved in the learning process (so say 95% Math, 79% Stat, 94% Mech, 65% HPE). Students of all specialties partially satisfied with the applying of new information technologies in education process (40% Math, 55% Stat, 57% Mech, 63% HPE), the largest proportion of dissatisfied students in Mathematics (50%).

Students of all specialties partially satisfied with availability of modern tutorials in the studied disciplines (67% Math, 27% Stat, 57% Mech, 63% HPE), the largest proportion of fully satisfied students in Statistics (46%).

Head of the Faculty of Mechanics and Mathematics presented the results of sampling survey at the meeting of Leaders of Dnipropetrovsk National University in September 2012. The results of the survey were lively discussed. Due to the sampling survey, University official site has been substantially updated, Career Days (meetings students with the employers) were carried out, activity lecturers to attract applicants to competitions conducted by University has been activated. University Leaders have expressed a desire to discuss the results of sampling survey of student opinions on actual issues annually.

References

Bethlehem, J. G. (2009) *Applied survey methods : a statistical perspective.* John Wiley & Sons, Inc., Hoboken, New Jersey.

Lohr, S.L. (2010) *Sampling: Design and Analysis*, Second Edition. Brooks/Cole, Cengage Learning.

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

Vasylyk, O., Yakovenko, T. (2010) *Lectures on Theory and Methods of Survey Sampling.* Kiev (in Ukrainian).

Survey Sampling Reputation of the University Through Students Eyes

Iana Bondarenko[1], Valery Turchyn[2], Evgeniya Hrebto[3], Inna Chernyshenko[4]

[1] Oles Honchar Dnipropetrovsk National University, Ukraine
e-mail: iana.s.bondarenko@gmail.com
[2] Oles Honchar Dnipropetrovsk National University, Ukraine
e-mail: vnturchyn@gmail.com
[3] Master Student at the Oles Honchar Dnipropetrovsk National University, Ukraine
e-mail: hrebto90@mail.ru
[4] Master Student at the Oles Honchar Dnipropetrovsk National University, Ukraine
e-mail: inna.chernyshenko@mail.ru

Abstract

The aim is planning, conducting and analyzing the results of survey sampling "Reputation of the University through students eyes" to identify factors that affect the reputation of Oles Honchar Dnipropetrovsk National University. Research methodology is a two-stage cluster sampling with probabilities proportional to size, parameter estimation, construction the confidence intervals for the unknown parameters, testing statistical hypothesis.
Keywords: population, sample, two-stage cluster sampling with probabilities proportional to size, confidence interval, accuracy, reliability

1 Introduction

Department of Statistics and Probability Theory was planned and conducted sampling survey "Students motivation in selecting Oles Honchar Dnipropetrovsk National University for higher education and training on the chosen specialty" in the 2011-2012 academic year. It has been found that the University's reputation, quality of education and learning conditions are working more effectively for attracting students to the University. Reputation represents the success of the University, helps to attract renowned scientists, lecturers, students, research partners, investments on a highly competitive market of educational services. The questions naturally arises: What factors form (determine) the reputation of the University from the viewpoint of students? What factors form (determine) the quality of education from the viewpoint of students? How these factors can be managed? New fsurvey sampling "Reputation of the University through students eyes" is devoted to research of these issues.

2 Planning Survey

Questionnaire "Reputation of the University through students eyes" was developed during September and October 2012 at the Department of Statistics and Probability Theory. Questionnaire consists of three blocks of questions. The first set of questions devoted to awareness (knowledge) of students about the features of the Classic University. The second set of questions concerns the factors that affect the level of the University's reputation and quality education. The third block of questions consists of questionnaire data of student. Each question provides the answers. The student can choose one variant that reflects his position and experience.

Continuous pilot survey of students of all specialties at the Mechanics and Mathematics Faculty were conducted for the purpose of testing and improving the questionnaire. The final version was prepared on the results of approbation questionnaire. This version was used in conducting survey sampling at the Faculty of Applied Mathematics, Faculty of Physics, Electronics and Computer Systems, Faculty of Chemistry, Physical and Technical Faculty. Also, students of social and economic, humanities specialties of University were surveyed. (They are students at the Faculty of Economics, Faculty of Law, Faculty of Social Sciences, Faculty of Psychology, Faculty of International Economics.)

One of the main issues in the planning of survey is the issue of sample size to obtain the survey results of the required accuracy and reliability. Construct a confidence interval in which the unknown parameter p – the proportion of the population units with a certain property – is located with the required accuracy e and reliability $1-\alpha$

$$P\{\hat{p}-e \le p \le \hat{p}+e\} = 1-\alpha \tag{1}$$

where accuracy e is calculated according to the formula

$$e = x_{1-\alpha/2}\sqrt{D\hat{p}} \tag{2}$$

where $x_{1-\alpha/2} - (1-\alpha/2)$ - quantile of normal distribution $N_{0;1}$.

The sample size n for estimation the proportion of the population units with certain property is

$$n = \frac{N \cdot x_{1-\alpha/2}^2 \cdot p \cdot (1-p)}{Ne^2 + x_{1-\alpha/2}^2 \cdot p \cdot (1-p)}. \tag{3}$$

The sample size n depends on the unknown p. Population proportion p of units with certain properties takes the values from 0,05 to 0,95 according to the results of the pilot survey.

Let us consider planning a survey at the Faculty of Applied Mathematics. The population consists of $N = 519$ students. The sample size must be equal to 221 to get the results with

required accuracy 5% and reliability 0,95. A stratification of the population in specialty: Applied Mathematics (AM), Informatics (IT), System Analysis (SA), Program Engineering (PI) are presented in Table 1.

Table 1: Stratification of the population of students at the Faculty of Applied Mathematics

Direction (Specialty)	Total
Applied Mathematics	170
Informatics	138
System Analysis	100
Program Engineering	111
Total	519

The resulting sample size must be allocated to strata. Calculate the sample size in each stratum according to the proportional allocation. Proportional allocation is determined according to

$$n_h = \frac{nN_h}{N}, \quad h = 1,...,H, \tag{4}$$

where n_h – number of sampled units in the hth stratum; N_h – number of units in the hth stratum in the population. Stratification results of the sample students at the Faculty of Applied Mathematics are presented in Table 2.

Table 2: Stratification of the sample of students at the Faculty of Applied Mathematics

Direction (Specialty)	Total
Applied Mathematics	72
Informatics	59
System Analysis	43
Program Engineering	47
Total	221

The sample for the survey selected from each stratum using two-stage cluster sampling with probabilities proportional to size. On the first stage academic groups of students are selected with probability proportional to the size of group. On the second stage students are selected from the groups with probability inversely proportional to the size of group. Therefore, the inclusion probability of student in the sample is the same for all students from population.

Units (academic groups) are representative with the number of students above fixed threshold of representativeness. Inclusion probability is equal 1 for such groups. Threshold value of students is determined according to

$$n_p = (N / n) \cdot m \tag{5}$$

where n – sample size of students; m – load of questioner (we assume that *m* is average number of students in academic groups certain specialty); N – population size.

Unrepresentative academic groups are selected from each stratum separately (d unrepresentative groups must be selected in stratum). Quantity of unrepresentative groups d is determined by the division total number of students in specialty on threshold representativeness

$$d = N / n_p \tag{6}$$

Inclusion probability P_{i1} of academic student groups for each specialty to the sample is determined according to

$$P_{i1} = d \cdot (a / b), \tag{7}$$

where d – number of sampled student groups; a – number of students in sampled groups; $b = \sum a$ – total number of students in specialty.

Inclusion probability P_{i2} of *i*th student from each group to the sample is determined according to

$$P_{i2} = c / a, \tag{8}$$

where c – number of students that must be sampled from this group; a – number of students in the group.

Inclusion probability P_i of *i*th student from population to the sample is determined according to

$$P_i = P_{i1} \cdot P_{i2} \tag{9}$$

The final probability must equal proportion $f = n / N$. Value $(1/f)$ shows how many students of population are presented one sampled student at the planning survey sampling.

Let us consider the planning of two-stage cluster sampling with probabilities proportional to size in specialty Informatics.

Total number of students in specialty Informatics is N = 138; sample size is n = 59; the average number of students in academic groups is m = 23. Threshold of representativeness is equal $n_p = (138/59) \cdot 23 = 54$. So, academic groups with the number of students more than 54 are included in the sample with probability equal 1. The number of unrepresentative academic

groups are $d = N / n_p = 138/54 = 3$. Sampling interval is $SI = N / d = 138/3 = 46$. Random start is $RS = RND() \cdot SI = 2$. The unrepresentative academic groups are sampled with step *SI*: *RS* = 2, *RS* + (1·*SI*) = 48, *RS* + (2·*SI*) = 94.

Verification of correctness in forming the sample consists in calculation the weight of each student $1/(P_{i1} \cdot P_{i2})$ and in summing weights of all sampled students. Obtained sum of weights must equal the size of population *N*.

Planning of two-stage cluster sampling with probabilities proportional to size for other specialty is similar.

Table 3: Results of two-stage cluster sampling with probabilities proportional to size in specialty Informatics

Group	Number of students in academic group a	Accumulated number of students	Sampled groups	Inclusion probability P_{i1}	Number of students that must be sampled from this group c	Inclusion probability P_{i2}	Weight per student $1/(P_{i1} \cdot P_{i2})$	Weight of the sampled students
IT1	33	33	2	0,72	20	0,60	2,3	46
IT2	29	62	48	0,63	20	0,68	2,3	46
IT3	25	87						
IT4	19	106	94	0,41	19	1,00	2,4	46
IT5	18	124						
IT6	14	138						
Total	b=138				n=59			N=138

3 Estimation proportion of students which has chosen the certain variant of answer to the question

For each academic group estimate the proportion of students, which has chosen the certain variant of the answer to the question, was calculated as

$$\hat{p}_i = \frac{\sum_{j \in s} w_{ij} y_j}{\sum_{j \in s} w_{ij}} \tag{10}$$

where w_{ij} – weight of *i*th student after questioning; $y_j = 1$ if student chose the certain variant of the answer to the question and $y_j = 1$ otherwise.

For each specialty (*AM, IT, SA, PI*) estimate the proportion of students, which has chosen the

certain variant of the answer to the question, was calculated as

$$\hat{p}_{AM} = \frac{1}{m_1}\sum_{i=1}^{m_1}\hat{p}_i\,,\ \hat{p}_{IT} = \frac{1}{m_2}\sum_{i=1}^{m_2}\hat{p}_i\,,\ \hat{p}_{SA} = \frac{1}{m_3}\sum_{i=1}^{m_3}\hat{p}_i\,,\ \hat{p}_{PI} = \frac{1}{m_4}\sum_{i=1}^{m_4}\hat{p}_i \qquad (11)$$

where m_1 – number of sampled groups of specialty *AM*, m_2 – number of sampled groups of specialty *IT*, m_3 – number of sampled groups of specialty *SA*, m_4 – number of sampled groups of specialty *PI*.

For Faculty of Applied Mathematics estimate the proportion of students, which has chosen the certain variant of the answer to the question, was calculated as the weighted sum estimates the proportion of students each specialty, which has chosen the certain variant of the answer to the question

$$\hat{p} = \frac{N_{AM}}{N}\hat{p}_{AM} + \frac{N_{IT}}{N}\hat{p}_{IT} + \frac{N_{SA}}{N}\hat{p}_{SA} + \frac{N_{PI}}{N}\hat{p}_{PI} \qquad (12)$$

where N_{AM} – number of students *AM* in the population; N_{IT} – number of students *IT* in the population; N_{SA} – number of students *SA* in the population; N_{PI} – number of students *PI* in the population; N – number of students in population.

4 Analysis of results

Let us consider the results of sampling survey at the Faculty of Applied Mathematics (FAM).

1. There were only eight Universities among higher education institutions in Ukraine to proclamation of independence. The number of Universities exceeded the one and a half hundreds in independent Ukraine. Classic Universities can be distinguish among the Universities. What distinguishes a Classic University from University, Academy, Institute in your opinion? (you can choose one of 1), 2), 3), 4)).
1) there are no differences (all graduates get a diploma about higher education)
2) differences in reputation
3) differences in rating
4) differences in education (a Classic University gives "University" education and "University" diploma)

It should be noted that a Classic University is clearly distinguished by the set of Faculties: medical, mathematical, physical, biological, chemical, humanitarian. 61% AM, 42% IT, 52% SA, 51% PI and in total 48% students understand by the Classic University something else as reputation, ranking, diploma about higher education are not features of Classic University. Only 39% AM, 58% IT, 48% SA, 49% PI and in total 52% students distinguished Classic university from other Universities.

2. Is there any difference between a Classic University and University in your opinion? (you can

choose one of 1), 2)).

1) Yes 2) No

72% AM, 83% IT, 68% SA, 75% PI and in total 75% students consider that a differences exist between a Classic University and University. 25% students do not see any difference.

3. Is there any difference between higher education in Classic University and higher education in University in your opinion? (you can choose one of 1), 2)).

1) Yes 2) No

71% AM, 77% IT, 71% SA, 44% PI and in total 67% students find that Classic University and University provide higher education of different level.

4. Rename of higher educational establishment, for example, from Institute to University, changes (you can choose one of 1), 2), 3)).

1) weight of diploma 2) training of specialists 3) does not change anything 51% AM, 58% IT, 36% SA, 21% PI and in total 44% students believe that rename of Institute in University influences the weight of diploma and training of specialists that does not answer reality in actual fact.

5. Which university would you have preferred? (you can choose one of 1), 2)).

1) "old" (University was founded a long time ago)

2) new modern University

Most of students would prefer long-established University, in total 80% of students have chosen "old" university with traditions and in total 20% of students would like to learn in the present-day modern university.

6. Which University would you give a preference: Kiev University or Dnipropetrovsk University under identical conditions? (you can choose one of 1), 2)).

1) Dnipropetrovsk 2) Kiev

64% AM, 32% IT, 45% SA, 43% PI and in total 47%students preferred the university located in Dnipropetrovsk.

7. Fundamental education (education oriented to the study of fundamental disciplines) increases the reputation of University (↑), lowers the reputation (↓), does not change the reputation (→) (you can choose one of ↑, ↓, →).

Fundamental education increases the reputation of University (so consider 65% AM, 20% IT, 65% SA, 71% PI and in total 62% students). 3% SA, 8% PI and in total 2% have chosen the

variant of answer "lowers". Perhaps the choice was random or students are not orientated on the study of fundamental disciplines, they are preferred narrow specialization. This result can be considered not significant (results are within the boundaries of the accuracy 5%).

8. Quality education increases the reputation of University (↑), lowers the reputation (↓), does not change the reputation (→) (you can choose one of ↑, ↓, →).

Quality education increases the reputation of University (so consider 94% AM, 76% IT, 94% SA, 88% PI and in total 92% students). 3% SA, 8% PI and in total 2% students have chosen the variant of answer "lowers". Possibly the choice was random or students are indifferent to the high level of education (just interested in joining to higher education institution). This result can be considered not significant (results are within the boundaries of the accuracy 5%).

9. Competitiveness of graduates at the market of labor increases the reputation of University (↑), lowers the reputation (↓), does not change the reputation (→) (you can choose one of ↑, ↓, →).

The variant of answer "increases" was chosen by 96% AM, 76% IT, 93% SA, 76% PI and in total 90% students. 5% SA, 11% PI and in total 3% students have chosen the variant of answer "lowers". Probably the choice was random or specialty is uncompetitive at the market of labor by the experience of graduates of these specialties or own experience. This result can be considered not significant (results are within the boundaries of the accuracy 5%).

10. Lecturers of University (known scientists, specialists) increase the reputation of University (↑), lower the reputation (↓), do not change the reputation (→) (you can choose one of ↑, ↓, →).

The variant of answer "increases" was chosen by 76% AM, 69% IT, 82% SA, 61% PI and in total 78% students.

11. Lecturers of University (the authors of textbooks, tutorials by which you and students of other Universities learn in Ukraine) increase the reputation of University (↑), lower the reputation (↓), do not change the reputation (→) (you can choose one of ↑, ↓, →).

The variant of answer "increases" was chosen by 50% AM, 31% IT, 41% SA, 39% PI and in total 49% students.

Listed characteristics may influence the reputation of the University: teaching clearly and accessibly, kindness and tact in relation of the students, patience, exactingness, objectivity in the estimation of student knowledge. We investigated the effect of the lecturer' characteristics on the reputation of University. Analysis of these and other results will be presented in detail at the presentation.

References

Bethlehem, J. G. (2009) *Applied survey methods : a statistical perspective.* John Wiley & Sons, Inc., Hoboken, New Jersey.

Lohr, S.L. (2010) *Sampling: Design and Analysis*, Second Edition. Brooks/Cole, Cengage Learning.

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

Vasylyk, O., Yakovenko, T. (2010) *Lectures on Theory and Methods of Survey Sampling*. Kiev (in Ukrainian).

Sampling as method of study of employment in the Republic of Belarus

Natalia Bandarenka

Belarus State University, Department of finance management, e-mail: bondnata@mail.ru

Abstract

The paper considers the problem of developing fundamentally new methodological approaches to the study of employment in the Republic of Belarus. Existing system in Belarus current account and assessment of the state of the labor market can rather adequately measure or estimate the degree of labor force participation, employment, wages, as a whole and by region, by type of activity, by forms of ownership. However, qualitative and quantitative evaluation of employees in small and micro entities of Belarus is virtually nonexistent.
Keywords: labour force survey, small entities, micro entities, individual entrepreneurs.

1 Introduction

For development of effective politics of employment and also for making regular comparisons between countries is needed reliable enough information about a labour market. Main data source about an economically active population, employment and unemployment are :

1) census of population; 2) inspections (survey) of organizations (enterprises);

3) registration records of administrative organs (tax, insurance, services of employment); 4) sampling of household surveys, or inspections of labour force (Labour Force Survey - LFS). Among them priority is from the point of view of meaningfulness and maintenances the survey of labour force, that comes true on recommendation of Advice of European Union from 1991 and serves for the receipt of comparable in an international scale data about the structure of employment and unemployment at the market.

2 Current account of employment in Belarus

Until 2012, in the Republic of Belarus data of the labor resources, the economically active population and employment in the economy were formed once a year in the calculation of the balance of labor resources as an average annual indicators. In the balance of labor resources is taken into account on a monthly basis the number of employees at large and medium-sized

enterprises (Ta), as well as the number of individual entrepreneurs. According to the current reporting of the organizations gets number of employees with labor contracts and civil-legal acts. The number of individual entrepreneurs is estimated according to the Ministry of Taxes and Levies (MTL) as the number of newly registered and liquidated entrepreneurs.

The annual rate of employment (T) is determined by:

$$Ts = Ta + Tb + Tc + Td + Ti + Tf \quad (1)$$

where Ta- the number of individual entrepreneurs and those working for them for hire (provided by MTL);

Tb - number of employees in organizations who report one time per year (small entities with average number of employees to 100 in a calendar year);

Tc - number of employees in organizations that do not represent statistical reporting (an example, representative offices of foreign organizations); information on them comes one time a year by administrative department (an example, Ministry of Foreign Affairs), which provides data to the National Statistical Committee;

Td - number of employees in confessions; information on them comes one time a year by authorized persons for Religious Affairs under the Council of Ministers of the Republic of Belarus;

Ti - number of employees in the private plots, for which this work is basic; recorded at the census (possible adjustment according households recording);

Tf - indirect estimate (additional calculations) of the number of employees in micro entities with average number of employees 15 or less in a calendar year.

As shows practical statistics, intra-variability of employment indicators is mainly determined by the dynamics of the number of individual entrepreneurs and employees of large entities and medium-sized business entities. The number of people working in small entities and micro entities is considered as a constant value (the report is available 1 time per year).

At the same time, the study of employment in small enterprises and micro-enterprises is essential. The market situation is developing in such a way that the young professionals to easier to get a job at the company, with minor amounts of production or services than in the large industrial enterprise. However, qualitative and quantitative evaluation of employees in small and micro entities of Belarus is virtually nonexistent.

Research of employment in small business are of significant interest to the public, scientific and academic community, as well as to the authorities, in one way or another in contact with the issues of economic regulation. Small business can have a significant impact on the economic and social processes taking place in the economy of the country, on the reproductive process as a whole, that is, small businesses in the near future should become a powerful tool in creating a new effective model for the Belarusian economy. For this reason, it is important study not only

the main indicators of financial and economic activity of small enterprises, but also the quality of employment in small enterprises, as well as their structure by gender, age, professional qualification structure, etc.

One of the main information sources about the activities of small businesses is official statistical data recorded by the National Statistical Committee of the Republic of Belarus. Since 2009 small businesses present to state department of statistics statistical reports on two standardized forms:

1) 1-МП "Report on the financial and economic activity of small business", this report is provided by business entities with the average number of employees during the year:

- In industry and transport - up to 100, inclusive;

- In agriculture and science and technology - up to 60, inclusive;

- In the construction and wholesale trade - up to 50, inclusive;

- In the retail trade and consumer services - up to 30 people, inclusive;

- Other branches - to 25 inclusive.

2) 1-МП (micro) "Report on the financial and economic activity of micro entities is and private (peasant) farms"; this report is provided by the micro entities with the average number of employees up to 15 persons inclusive and private (peasant) farms with the average number of employees up to 100 people inclusive.

However, official information from these sources is fixed only some aggregate indicators of financial and economic activities of small businesses and do not reflect the peculiarities of small businesses at different stages of the life cycle.

In addition, one of the problems of employment is study of the various characteristics of the employed population by sex, age, educational level, etc. Such information can be obtained according to the census, which is conducted one time in 10 years. Census of Population of the Republic of Belarus is the main source of the information resources on the size and structure of the population combination with social and economic characteristics, national and linguistic composition of the population, its level of education (Table 1).

Table 1: Employed population of Republic of Belarus by census 2009 (thousands people)

	Republic of Belarus			Urban			Rural		
	total	of which		total	of which		total	of which	
		men	women		men	women		men	women
Total employed	4613,3	2322,5	2290,8	3630,7	1791,8	1838,9	982,6	530,7	451,9
of which with education level:									
- higher	1154,2	509,7	644,5	1023,0	450,9	572,1	131,2	58,7	72,5
- secondary specialized	1561,5	698,8	862,7	1271,2	574,8	696,4	290,3	124,0	166,3

- vocational technical	646,7	393,4	253,4	463,0	275,6	187,4	183,7	117,7	65,9
- general secondary	989,2	564,5	424,6	686,4	384,4	301,9	302,8	180,1	122,7
- general basic	133,0	88,8	44,2	64,9	43,0	21,9	68,1	45,8	23,3
- general primary	10,1	6,6	3,5	5,4	3,4	2,1	4,7	3,2	1,5

Necessary to conduct a special surveys for obtain similar statistics data in the period between censuses.

3 Sample surveys of employment

2012 quarterly sample survey of households conducted in the Republic of Belarus in order to study the problems of the employment. Data from this survey provide a solid base of information and analysis that meets the requirements of international standards, it is necessary to make informed management decisions to improve state policy in the sphere of employment. According to the results of the survey obtained data sets on the size and structure of the employed, unemployed, the unemployed by a combination of age and gender and professional qualifications as a sign of the whole country and by region, as well as separately for urban and rural areas (Table 2).

Table 2: The results of the test sample survey of employment in the Mogilev region. (2011)

Indicators	feature value		factual error	
	extrapolated, $Э_x$	By census, x	in absolute terms, $\Delta a = \lvert x - Э_x \rvert$	в % $\Delta_{отн} = \frac{\lvert x - Э_x \rvert}{x}$
Total employed, people	506231,11	515876	9644,89	1,87
Urban	402333,2	412962	10628,8	2,57
- men	194657,81	205508	10850,2	5,28
- women	207675,39	207454	221,39	0,11
Rural	103897,91	102914	983,91	0,96
- men	55227,66	55228	0,34	0,0006
- women	48670,25	47686	984,25	2,06
Total employed, people				
- men	249885,05	260736	10851	4,16
- women	256345,64	255140	1205,64	0,47

The preliminary results of an iterative weighting of employment test sample survey of the Mogilev region (Table 2) indicate the representativeness of the constructed sample. The actual relative error does not exceed 8.7%, including the number of employees - 1.8%.

However, the purpose of the labor force survey is the formation of complete reliable information on the situation on the labor market in the whole of the Republic of Belarus, not separately the small business sector.

At the same time, the statistical practice shows that in most countries there are special sample

survey of small businesses.

In Belarus, the sample survey of small businesses conducted on a quarterly basis during the years 2004-2008. Using a combination of univariate and multivariate selection (optimum stratification, proportional stratification, multi-dimensional sampling based on the cluster selection) by regions and sectors of the economy. The volume of the sample massif in the Republic of Belarus as a whole and for each area formed within 20-35% of the total of small businesses. By industry - different proportion of selection was formed based on the number of MP and a given estimation accuracy on the basis indicators (output): the relative standard error of the country was 2%, by area - no more than 5% in small trades - not more than 8-10%. However, with the abolition of intra-sample surveys such statements are no longer held.

Thus, given the trends in the development of market relations in Belarus to develop a fundamentally new methodological approaches to the study of employment in the country is very important. In particular, there is an objective need for sample survey of small businesses, entrepreneurs, micro entities, during which will be received not only the data on the results of the financial and economic activities, but also the structure of employment in this field.

Concluding remarks

Existing system in Belarus current account and assessment of the state of the labor market can rather adequately measure or estimate the degree of labor force participation, employment, wages, as a whole and by region, by type of activity, by forms of ownership. Conducted since 2012 labor force surveys allow to obtain data sets on the number and structure of the employed, unemployed, unoccupied at the level of the republic and by regions, which makes it possible not only to show the real picture of the level of unemployment and the labor market in Belarus, but also allow identify the structure of the unemployed actually a combination of age and gender and professional qualifications of features.

Perspective directions research in the field of employment in the Republic of Belarus are:

- Extension of the use of sample surveys, including: a survey of small businesses (in particular, micro entities) by type of activity and region, a survey of individual entrepreneurs;

- Expansion of the program to the labor force survey of self-employment, gender and age and educational structure of the self-employed.

References

Bandarenka N. The problems of labor market statistics in the Republic of Belarus // *Questions of Statistics*. - 2012. - № 6. - S. 10-13.

Bandarenka N., Bokun N. Experience in conducting the Labour Force Survey // *Questions of Statistics*. - 2012. - № 6. - S. 13-23.

The results of census of Republic of Belarus in 2009. Mode of access: http://belstat.gov.by/homep/ru/perepic/2009/itogi1.php. Accessed on 20.04.2011

An estimation method in a finite population domain where sample size is small or even zero

Andrius Čiginas[1] and Tomas Rudys[2]

[1] Vilnius University Institute of Mathematics and Informatics, e-mail: andrius.ciginas@mif.vu.lt
[2] Vilnius University Institute of Mathematics and Informatics, e-mail: tomas.rudys@mii.vu.lt

Abstract

Consider a finite population $U = \{1, \dots, N\}$ of size N. Assume that, in order to estimate a parameter (characteristic) of the population, the sample $s = \{i_1, \dots, i_n\}$ of size n is drawn from U, according to a sampling design $p(\cdot)$. Here $p(s)$ is the probability to get the particular s. Let $\pi_k = \mathbf{P}(k \in s) > 0$ be the inclusion into the sample probability for the population element k. It is usual situation when a construction of the sample design $p(\cdot)$ is closely related to an auxiliary variable, say, z with values $\{z_1, \dots, z_N\}$ known for all units of U. In particular, e.g. for a stratified simple random sampling design and for probability proportional-to-size sampling the inclusion probability π_k represents an importance of the population unit k by the relative size of z_k. Let y be the variable of interest with the values $\{y_1, \dots, y_N\}$ in the population, and, we aim to estimate the total

$$t_{y,D} = \sum_{i \in D} y_i \qquad (1)$$

where $D \subseteq U$ is any non-empty set. If the particular estimation domain D is known before the sample selection, then, as it is common in practice, the sampling design $p(\cdot)$ is constructed in order to get a sufficient (for quality requirements of estimates) sample size in that domain. But if we are interested in D after sample selection and collection of the data $\{y_{i_1}, \dots, y_{i_n}\}$, the sample size in D can be too small or even equal zero and this leads to get bad estimates in the sense of their quality, if the estimators of (1) are usual Horvitz-Thompson (H-T), generalized regression estimators, etc.

Assume that at the estimation stage, we have an auxiliary variable x and all its values $\{x_1, \dots, x_N\}$ are known. In practice, the meaning of variable x can be, for instance:

x is simply a better predictor of y than z (in cases where the sample survey has more variables of interest and z was important for all of them at the sample planning moment);

x is observed in the time period between the sample selection and estimation, e.g. we get it from administrative data sources.

Then it is also meaningful to incorporate a relation between y and x into the estimation process. But in the presentation we will discuss only a linear dependence between x and z where, for instance:

x can have the same definition as z but it is observed at the different moment of time than z, and thus their values are different;

definitions of z and x slightly differ;

x is an alternative measure of size of the population units which was not used at the sample planning stage;

z can be measured roughly (quickly by a different way) and so we get x.

Thus, an origin of x can be various, but we keep in mind that both z and x more or less represent a size of y. Our idea is to describe how the sizes of the population units „like" to change or, in other words, how the study variable y „likes" to change. We formalize the changes by numbers $\theta_{i;D}, i \in s$ which are probabilities of the sample s units to „belong" to the domain D. Then we introduce the estimator of (1) which is

$$\hat{t}_{y;D} = \sum_{i \in s} \theta_{i;D} \frac{y_i}{\pi_i} \qquad (2)$$

This estimator is constructed so that it is usual H-T estimator if D coincides with U. In comparison to the known but naive special case of (2), where $\theta_{i;D} \equiv \frac{N_D}{N}$ and N_D is the size of D, estimator (2) is robust in the sense that it is not sensitive to an arbitrarily chosen domain D (homogeneous or not, with respect to the population).

We note that our estimation aproach is model-assisted similarly as, e.g. it is in popular cases of regression type estimators. In the presentation, we will explain the construction and estimation methods of the numbers $\theta_{i;D}, i \in s$. We will give the mean square error of the estimator (2) and its estimator, and demonstrate an efficiency of (2) by a simulation study.

Keywords: small area estimation, auxiliary information, linear regression, robustness.

References

Longford, N. T. (2005). Missing Data and Small-Area Estimation. Springer Verlag.

Rao, J. N. K. (2003). Small Area Estimation. Wiley-Intercience.

Särndal, C., Swensson, B. & Wretman, J. (2003). Model Assisted Survey Sampling. Springer Verlag.

The selection of respondents for the monitoring of enterprises in the National Bank of the Republic of Belarus

Katsiaryna Chystsenka

National bank of the Republic of Belarus, e-mail: katsiaryna.chystsenka@gmail.com

Abstract

In the paper short characteristic of monitoring of the enterprises in the National bank of the Republic of Belarus is considered. Sample frame, calculation of sample fraction and methodology of sample are presented.
Keywords: Monitoring of the enterprises, enterprises sample, fixed sample.

1 Introduction

Information provision is an important element for an effective monetary policy. It helps realistically and effectively estimates the right course of chosen policy. Central banks must constantly monitor the economic development, to predict the possible options for its development, and to develop corrective tools to ensure monetary and financial stability, to support balanced economic growth. An important component of monitoring is the survey of enterprises. Therefore, to solve the above problems in contemporary practice, along with the official statistics and other external information they actively use the monitoring data of enterprises.

In the Republic of Belarus, the National Bank, other government and commercial organizations use the monitoring of enterprises. However, the National Bank carries out the monitoring of the enterprises on a wider range of respondents and indicators.

The main purpose of sampling is the selection of enterprises of different sizes, with different financial situation to identify the most significant relationship between production, employment, prices, financial situation and market conditions at micro level, at macro level - the trends of economic processes for analysis and forecasting in interaction with the tools of monetary and credit policy and other elements of economic policy.

The object of the survey is enterprises of all forms of ownership (including their division) representing the state statistical report on the financial results (population).

The National Bank carried out sample survey of enterprises for monitoring of their economic and financial situation for the first time in 2001.

2 Sample size

A sample size was calculated in the National Bank. It was more than 30% of enterprises representing the state statistical report on the financial results for all selected economic branches. Calculation of the sample size consisted of four stages:

At the first stage the number of industrial enterprises was determined. In each industrial branch three sub-sectors (initial subsectors) that had the fraction in the industrial revenue from sales of goods, products, works and services in the region of 3% or less, were defined.

The number of industrial enterprises (N_{ind}) for the survey was calculated as:

$$N_{ind} = \sum\left(\frac{3 \cdot w_{ij}}{\Sigma w_{ijk}}\right), \tag{1}$$

where *3* is fixed number of enterprises in industrial initial sub-sectors,

w_{ij} is the fraction of revenue from sales of goods, products, works and services of the *i*-th industrial branch in revenue from sales of goods, products, works and services of all industrial branches in the *j*-th region as a whole,

w_{ijk} is the fraction of revenue from sales of goods, products, works and services in the *k*-th initial subsector of the *i*-th industrial branch in revenue from sales of goods, products, works and services of all industrial branches in the *j*-th region as a whole.

At the second stage the sample size for the other selected branches (*m*) in region was calculated. It depended on the sample size industrial enterprises selected for the survey in the first phase. Such a method of calculation was chosen to keep the ratio of fractions of economic branches.

The number of enterprises of other branches (N_a) was calculated as:

$$N_a = \sum\left(\frac{N_{indj} \cdot w_{mj}}{w_{indj}}\right), \tag{2}$$

where w_{mj} is the fraction of revenue from sales of goods, products, works and services of the *m*-th branch in revenue from sales of goods, products, works and services of all branches in the *j*-th region,

w_{indj} is the fraction of revenue from sales of goods, products, works and services of industrial branches in revenue from sales of goods, products, works and services of all branches in the *j*-th region.

At the third stage the donor pool has been created to replace all possible non-responses.

Also, the following enterprises were included in the sample frame (enterprises are not included in the sampling in the previous *phases*), *the fourth stage*:

- enterprises that are actively involved in the formation of the authorized capital of credit institutions, in particular, the founders of the local commercial banks;

- large borrowers;

- major exporting enterprises.

3 Sample design

The main characteristics of the sample design for the monitoring of the enterprises are presented in Table 1.

Table 1: The main characteristics of the sample design

Sample model:	stratified
	probability
	systematic
Unit survey:	enterprise
Sample variables for population stratification:	region
	kind of activity
	economic indicator "Revenue from sales of products, goods, works and services"

Sample frame is a list of enterprises stratified into groups by:

regions of the Belarus and Minsk City (7 units),

selected economic branches:

- industry;
- transport;
- construction;
- trade.

Sampling was carried out by economic branches which set the trends of the region's economy and the fraction of revenue from sales of goods, products, works and services in the total revenue for the region's economy in these branches were significant. In addition the sum of fractions of economic branches was not less than 75% of the total revenue for the region's economy.

The sampling frame changed due to the transition of the Republic of Belarus from the use of Classification of economic branches to NACE. Since 2011, monitoring of the enterprises is carried out by economic activity: mining industry; manufacturing industry, including the most

significant economic activity subsection (2 digit for NACE) for the region; production and distribution of electricity, gas and water; transport; construction; trade.

Then stratified enterprises were ranked in ascending index selection "Revenue from sales of goods, products, works and services".

A sample survey was carried out by systematic method in each region and in selected economic branches with the calculated interval (S_{ij}):

$$S_{ij} = \frac{N_{ij}}{n_{ij}}, \tag{3}$$

where N_{ij} is number of enterprises in the population in the *i*-th branch of the *j*-th region,

n_{ij} is number of enterprises in the sample population in the *i*-th branch of the *j*-th region.

Number of the initial enterprise (IN_{ij}) to begin selection is calculated as:

$$IN_{ij} = \frac{S_{ij}}{2}. \tag{4}$$

The sample size for the monitoring of enterprises was about 2,000 units (Table 2).

Table 2: Distribution of the surveyed enterprises by economic activity in 2011*

	Number of enterprises:		Sample fraction, % *f*
	in the population, *N*	in the sample population, *n*	
Mining industry	30	17	56,7
Manufacturing industry	1806	650	36,0
Production and distribution of electricity, gas and water	189	7	3,7
Construction	1433	474	33,1
Trade	1134	470	41,4
Transport	593	247	41,7
Total	*5185*	*1994*	*38,5*

*Number of enterprises representing the questionnaires of the conjunctural situation.

The simple direct estimators based on the Horvitz-Thompson estimator were used.

Four types of questionnaires for instrumental monitoring of enterprises are used by the National Bank of the Republic of Belarus:

- questionnaire of the conjunctural situation. These data are needed to assess changes in economic conditions. Questionnaire is filled in on a monthly basis.

- the investment questionnaire is necessary to assess the motives and forms of investment activities of enterprises. Questionnaire is filled in on a quarterly basis.

- financial questionnaire provides information on the sources of the formation of self-financing enterprises and determination the need for extra resources. Questionnaire is filled in on a quarterly basis.

- questionnaire of the demand for banking services is needed to study the enterprise demand for banking services and the degree of satisfaction. Questionnaire is filled in 2 times a year.

The number of enterprises filling in the questionnaires varies depending on its type. The reason for this is the reluctance of some enterprises to show certain information regardless of the status of confidentiality. The largest number of enterprises fills in the questionnaire about the conjunctural situation.

The considered sample population is fixed. The main set of the enterprises in the sample population is kept during the whole period of monitoring to detect the dynamics of the financial situation of each enterprise.

4 Sample problems

Some problems exist during the enterprises' survey.

1. *Non-responses*. The participation of enterprises in the survey is voluntary and there is the probability of failure of some selected enterprises to participate in the monitoring, so the donor pool was created (10 − 15%) for replacement of such enterprises in the sample population. This donor pool was included in the sample population.

2. *Sample updating*. The liquidation of enterprises sample, changes in their economic activity, their union, etc. create the problems of using a sample population for a long time. In the National Bank updating of the sample population is carried out annually to reduce non-responses. To do this, the regional statistics departments annually send a new stratified by area, economic activity and ranked by revenue from sales of goods, products, works and services list of enterprises representing the state statistical report on the financial results (population) to the National Bank. Updating consists of replacement of the enterprise to another enterprise with a similar amount of revenue replaced enterprise.

5 Results

The results of the monitoring of enterprises allow the identification of trends in the Belarusian economy, quickly get information about the economic situation and its possible changes, and to quickly analyze the financial condition of enterprises and the most important factors in investment activity, as well as to receive information on the development of demand for banking services and the degree of satisfaction.

The specialists prepare consolidated analytical and statistical materials on the basis of the results of the survey of enterprises, publish them on the official website and sent to interested users.

References:

The rules of enterprise monitoring by the National Bank of the Republic of Belarus: Resolution of the Board of Directors of the National Bank of the Republic of Belarus (2003), № 376. Consultant: Belarus. Tech 3000 [electronic resource] / National Center of Legal Information of the Republic of Belarus.

Milevsky, P., Zubovich, A. (2012). Monitoring of non-financial enterprises as a function of the National Bank. *Bank Bulletin* **25**, 60 - 63.

Bokun, N., Cherhysheva, T. (1997). *Metody vyborochnyh obsledovanij*. Minsk.

Monitoring of the real sector of the economy of the Republic of Belarus (2011). Minsk: *Analytical Review of the National Bank*, January-December 2011.

Administrative data in survey sample

Andris Fisenko

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

Abstract

The goal of this paper is to show practical view usage of administrative data sources in different steps of survey samples in Central Statistical Bureau of Latvia (CSB).
Keywords: Survey sampling, administrative data, auxiliary information.

1 Introduction

A database is a compilation of information on characteristics and events that is stored in an organized manner. An administrative database in CSB must ensure the integrity of the data to comply with legal and administrative requirements for supporting statistical and historical information. In CSB administrative databases are used for different stage in surveys. This paper describes some stages of using an administrative database. These stages are: Sampling, weighting and imputation.

2 Description

2.1 Sampling

All sample surveys beginning with sample. What we need to get a sample? Most likely at beginning we have some demands. Suppose that we have received information what kind of survey it is, but are it enough? No. To see full picture we need to know lot more information. Key issues are – costs, what indicators we should survey, how big sample we need to choose, what about data accuracy we want to achieve? What is expected response rate? The question is more than enough, but all this questions should be answered to build sample.

Let's assume, what we have necessary information about survey sample. Before we start with practical examples of administrative data usage let's see short introduction in CSB dwelling surveys.

Probability-sampling principles are used in the surveys of households and persons. Mainly the samples of surveys of households and persons are obtained using stratified two-stage probability sampling. The sample allocation between strata is made proportional to the population sizes of strata in all those surveys. Households or persons are stratified by the degree of urbanisation and in some survey also by the geographical location. Counting areas of population census 2000 are

used as the primary sampling units (PSU). At the second stage in survey sample the simple random sampling is used. Such kind of survey sample design is a more cost-effective sampling scheme than simple random sampling or stratified sampling.

2.1.2 Primary sampling Unit

Primary sampling units is unique geographical breakdown made by CSB. Latvian territory is covered by ~4000 counting areas of population census 2000. Each counting area contains ~200 dwellings. There are several administrative data source is used to create this counting areas. Putting together Population register and Building register CSB created Address register which is used as a basis for future activities including survey sampling.

PSU are grouped in 4 strata by the degree of urbanisation (Riga, the capital city; 8 other largest cities; other towns; rural areas). Each group of PSU is numerated. The primary sampling units are listed by geographical region, and within a geographical region in a serpentine order that places units containing similar types of people together.

Image 1 show how is realised serpentine order.

Image 1: Serpentine over

Latvia

The basic information that contains the PSU frame from register includes:

- Starta
- Counting area number
- Number of dwellings
- Administrative territory code and name
- Old district code and name
- Region code and name.

PSU's were selected by systematic sampling with inclusion probabilities proportional to population size (number of households) of PSU's. PSU sample is created for period 2007-2014. There is special rotation schema for each PSU. The annual sample is evenly distributed over time (same numbers of dwellings participate in the survey within each of 52 weeks of the year). The developed sampling procedures guarantee that within each quarter sample of PSUs is evenly distributed over space, too.

For continuity of PSU sample addition information is include:

- Year, quarter, week.
- Survey wave (time), Labour force survey each PSU include 4 time, before leave sample.
- Number of persons
- Indicators who show which household survey belong. (3 surveys are connected).

Each sampled PSU are used for 3 surveys in this connected survey situation. There are: Labour force survey, Household Budget survey and Survey of Resident Travellers. This decision reduces the travel costs of interviewer and total costs for all three surveys. More about cost function for survey can be read in Mārtiņš Liberts Doctoral Dissertation[1].

2.1.3 Secondary sampling unit

At the second stage within each sampled PSU dwellings are selected by a simple random sampling.

The information that contains the SSU frame from register includes:

- ID number of address
- Geographical coordinate
- Number of persons living in address
- Contact person information (ID number, name, surname, DoB)
- Dwelling status according Census 2011
- Dwelling type according to Buildings and Structures register.

After finale sampling in sample of dwelling are added additional information:

- Full address with postal code and phone number
- Breakdown survey mode (CAPI or CATI)

Using selected sample unit ID it's possible to link information about each declared person in dwelling. For this purpose from register additional information is include:

- Person ID, name, surname
- Sex
- Citizenship,
- Nationality,

[1] MārtiņšLiberts "The optimisation of sampling design" University of Latvia, 2013

- Marital status,
- Country of birth.

This example shows how useful are information is for preparation of survey sample. Also additional information from different administrative source can help reduce the burden of respondents.

2.2 Weighting

From survey questioner we get lot of information. Also about selected dwellings. Register is not ideal there are still has an errors and after survey is finished we have a little better information. The nonresponse cases are, dwellings not exist, dwelling is burn down, address is incorrect, interviewer can't find the address, and address is public institution. Taking all this information in to account it is necessary to reweighting design weights and keeps in mind this information for next waves of LFS.

The method of Calibration in sample survey is used. Calibration is a procedure than can be used to incorporate auxiliary data. In Labor force survey (LFS) there is different auxiliary information. One is number of person living is private households and second information from State Employment Agency (SEA). From SEA we get following information:

- Person ID
- Activity (Unemployment) status
- Exact day of activity in last quarter

There are 3 possibilities for unemployment status:

- Get
- Lost
- Has not changed since previous quarter

2.3. Imputations

If weighting dealing with unit nonresponse, then imputation dealing with item nonresponse. But it's tricky situation in survey. Why ask question if we have auxiliary information about it? One reason can be to check how good quality is this auxiliary information. Second reason can be that auxiliary information is only additional information and not always match survey methodology, but still can be used as in imputation. Good example is income. From State Revenue Service we get person income form salary, bet from survey we can get additionally illegal income. There are lots of different situations what information we take into account or what not for calculation total income of person or household.

3. Conclusions

Administrative data can be used as information to preparing survey sample as auxiliary information at the end of survey. Nowadays the administrative data is used for population estimates and target is that next Census 2021 will be based only an administrative data.

References

Lapiņš,J., Vaskis E., Priede Z. &Bāliņa S. (2002). *Houshold surveys in Latvia.*Central Statistical Bureau of Latvia, Riga.

Using robust regression for capital expenditure estimation

Tetiana Ianevych[1]and Olga Vasylyk[2]

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

Abstract

In this work we tried to incorporate the robust regression inference into estimation of capital expenditure in Ukraine. The short overview for the types of robust regression is provided and some simulation results are presented.
Keywords: capital expenditure, robust regression, small area estimation

1 Introduction

The annual and quarterly surveys of capital expenditure collect important information for the Ukrainian National Accounts. This information is used by Ukrainian government and agencies, trade associations, universities and international organizations for policy development and as a measure of regional activity.

The aim of annual surveys is the estimation of the structure whereas quarterly surveys have the estimation of changes as an objective. For the annual survey all enterprises should be observed. In the quarterly survey only enterprises that have significant size of capital expenditure are usually surveyed. Before, only the enterprises which take part in other quarter surveys (mostly in structural business survey) could be observed quarterly. So, the quarterly capital expenditure was not adjusted for enterprises that were not observed. In order to avoid underestimation of the capital expenditure value in the quarterly surveys it was decided to implement the probabilistic sampling into their investigation. Since the large and medium-sized enterprises are more valuable they are observed with probability one. Thereby we focus on the investigation of small enterprises and perceive them as population.

In the paper by Honchar & Ianevych (2012) there was made an attempt to incorporate into the estimation process the information from the previous surveys by the means of regression estimator and obtain more accurate estimates for domain. But this attempt was not very successful and we did not gain much efficiency in exploitation regression estimator comparing to Horwitz-Tompson estimator. So, we continue to search and decided to use the robust regression approach.

During our investigation we have analyzed the data from the annual 2009 and 2010 surveys and have come to the conclusion that it is not possible to build any reliable models for the small area

estimation directly. We took notice that only 7% of the small enterprises had capital expenditures in 2009 as well as in 2010. That is why, all the positive values are regarded as outliers within the robust regression inference. But these values are of major interest in the survey. So, before the construction the model we need to distinguish them and only after that to analyze. Let us also remark that investigation the annual survey data gives us the possibility to understand the nature of the data and helps in designing sample for quarterly surveys.

In section 2 we present some basic notions and ideas of robust regression and cite its main types. The section 3 is devoted to analysis of efficiency of robust regression for the total and for the domain estimation of capital expenditure of the agricultural enterprises.

2 Robust regression

2.1 What is robust regression?

Robust regression is a form of regression analysis designed to circumvent some limitations of traditional parametric and non-parametric methods. Regression analysis seeks to find the relationship between one or more independent variables and a dependent variable. Certain widely used methods of regression, such as ordinary least squares, have favourable properties if their underlying assumptions are true, but can give misleading results if those assumptions are not true; thus ordinary least squares is said to be not robust to violations of its assumptions. Robust regression methods are designed to be not overly affected by violations of assumptions by the underlying data-generating process. In particular, least squares estimates for regression models are highly non-robust to outliers. While there is no precise definition of an outlier, outliers are observations which do not follow the pattern of the other observations. In the presence of outliers that do not come from the same data-generating process as the rest of the data, least squares estimation is inefficient and can be biased. Because the least squares predictions are dragged towards the outliers, and because the variance of the estimates is artificially inflated, the result is that outliers can be masked. In many situations, including business survey, it is precisely the outliers that are of interest.

2.2 Methods for robust regression

Despite their superior performance over least squares estimation in many situations, robust methods for regression are still not widely used. Several reasons may help explain their unpopularity. One possible reason is that there are several competing methods and the field got off to many false starts. Also, computation of robust estimates is much more computationally intensive than least squares estimation; in recent years however, this objection has become less relevant as computing power has increased greatly. Another reason may be that some popular statistical software packages failed to implement the methods Although uptake of robust methods has been slow, modern mainstream statistics text books often include discussion of these methods. Also, modern statistical software packages such as R and S-PLUS include considerable functionality for robust estimation (see, for example, the book by Fox (2002)).

The simplest methods of estimating parameters in a regression model that are less sensitive to

outliers than the least squares estimates, is to use least absolute deviations. Even then, gross outliers can still have a considerable impact on the model, motivating research into even more robust approaches.

Here are some approaches which lead to the robust regression models.

M-estimation. In 1973, Huber introduced *M-estimation* for regression. The M in M-estimation stands for "maximum likelihood type". The method is robust to outliers in the response variable, but turned out not to be resistant to outliers in the explanatory variables (leverage points). In fact, when there are outliers in the explanatory variables, the method has no advantage over least squares. The idea of M-estimation is in following.

Consider the linear model for the *i*-th of *n* observations

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$$

The fitted model is

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_k x_{ik} + e_i = \mathbf{x}_i'\mathbf{b} + e_i$$

The general M-estimator minimizes the objective function

$$\sum_{i=1}^{n} \rho(e_i) = \sum_{i=1}^{n} \rho(y_i - \mathbf{x}_i'\mathbf{b})$$

where the function ρ gives the contribution of each residual to the objective function. For example, for least-squares estimation, $\rho(e_i) = e_i^2$. Another objective functions were proposed by Huber and Tukey and they produced the corresponding estimators called the *Huber estimator* and *bisquare* (or *biweight) estimator*.

Least trimmed squares. In the 1980s, several alternatives to M-estimation were proposed as attempts to overcome the lack of resistance. See the book by Rousseeuw& Leroy (2003) for a very practical review. For example, the *least trimmed squares* (LTS) is a viable alternative to M-estimation.

The residuals from the fitted regression model are

$$e_i = y_i - (b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_k x_{ik}) = y_i - \mathbf{x}_i'\mathbf{b}$$

Let us order the squared residuals from smallest to largest:

$$(e^2)_{(1)}, (e^2)_{(2)}, \ldots, (e^2)_{(n)}$$

The LTS estimator chooses the regression coefficients to minimize the sum of the smallest *m* of the squared residuals,

$$\mathbf{b}) = \sum_{i=1}^{m} (e^2)_{(i)}$$

While the LTS criterion is easily described, the mechanics of fitting the LTS estimator are complicated. Moreover, bounded-influence estimators can produce unreasonable results in certain circumstances, and there is no simple formula for coefficient standard errors.

Quantile regression. *Quantile regression* is another type of regression analysis used in statistics. Whereas the method of least squares results in estimates that approximate the conditional mean of the response variable given certain values of the predictor variables, quantile regression aims at estimating either the conditional median or other quantiles of the response variable. Advantage of quantile regression, relative to the ordinary least squares regression, is that the quantile regression estimates are more robust against outliers.

The mathematical forms arising from quantile regression are distinct from those arising in the method of least squares and M-estimation. The method of least squares leads to a consideration of problems in an inner product space, involving projection onto subspaces, and thus the problem of minimizing the squared errors can be reduced to a problem in numerical linear algebra. Quantile regression does not have this structure, and instead leads to problems in linear programming that can be solved by the simplex method. The fact that the algorithms of linear programming appear more esoteric to some users may explain partially why quantile regression has not been as widely used as the method of least squares.

The idea of estimating a median regression slope, a major theorem about minimizing sum of the absolute deviances and a geometrical algorithm for constructing median regression was proposed in 1760 by Ruđer Josip Bošković, a Jesuit Catholic priest from Dubrovnik. Median regression computations for larger data sets are quite tedious compared to the least squares method, which historically generated a lack of popularity among statisticians, until the widespread use of computers in the latter part of the 20th century.

Let Y be a real valued random variable with cumulative distribution function $F_Y(y) = P(Y \leq y)$. The τ-th quantile of Y s given by $Q_Y(\tau) = F_Y^{(-1)}(\tau) = inf\{y: F_Y(y) \geq \tau\}$.

Suppose the τ-th conditional quantile function is $Q_{(Y|X)(\tau)} = \mathbf{X}\boldsymbol{\beta}_\tau$. Given the distribution function of Y, β_τ can be obtained by solving

$$\beta_\tau = \underset{\beta \in R^k}{\mathbf{argmin}} E(\rho_\tau(Y - \mathbf{X}\boldsymbol{\beta}_\tau))$$

where a loss function $\rho_\tau(y) = y(\tau - I\{y < 0\})$. Solving the sample analog gives the estimator of β_τ.

$$\hat{\beta}_\tau = \underset{\beta \in R^k}{\mathbf{argmin}} \sum_{i=1}^{n} (\rho_\tau(Y_i - \mathbf{X}\boldsymbol{\beta}_\tau)).$$

This minimization problem can be formulated and solved as a linear programming problem.

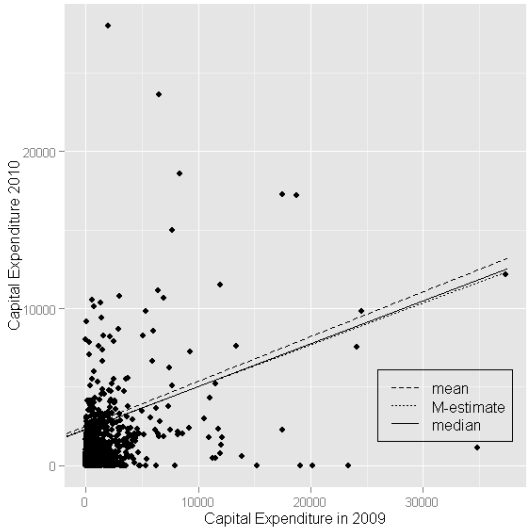
3 Estimation using quantile regression

In this section we apply the robust regression for estimation the total capital expenditure for the agricultural enterprises and corresponding estimates for regional domains. So, we consider as a population U the small agricultural enterprises, $N = 37094$. The variable that we are interested in is capital expenditure y_i of the enterprise i. At our disposal is the information on capital expenditure from annual 2009 and 2010 surveys and revenue in 2009. We have chosen as auxiliary variables x_{i1} the capital expenditure in 2009 and revenue as x_{i2}.

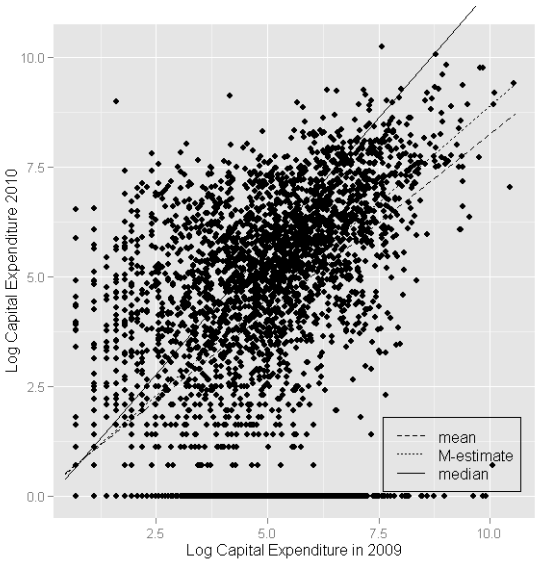
First of all we have extracted the items that have a good chance to have positive capital expenditure in 2010 using our auxiliary variables. We obtained two subpopulations: U_p is the set of enterprises that had positive capital expenditure and revenue in 2009 and U_c is its complement. The number of elements in U_p we denote as N_p=3920 and the percentage of elements that had positive capital expenditure in 2010 among them is equal to 66%. Meantime the subpopulation U_c has $N_c = 33174$ elements and only 8% of them had capital expenditure in 2010. Design effect of this stratification is equal to 86.9% and we gained in efficiency.

Let us analyse the data from subpopulation U_p. The *Picture 1* shows the dependence between variables y and x_1 with lines that correspond to ordinary linear regression (mean), robust regression based on M-estimation (M-estimate) and quantile regression (median). The *Picture 2* shows the dependence between logarithms of variables y and x_1 decreased previously by 1 with corresponding regression lines. This pictures were made in R with help of *ggplot2* package (see Wickham (2009)) using lm() for ordinary linear regression line, rlm() for robust regression line based on M-estimation and rq() for quantile regression line.

Picture 1.



Picture 2.



As we see there is no big difference between ordinary regression and M-regression. So, we

decided to compare only estimators based on ordinary and quantile regression. We have made 10000 Monte Carlo simulation of the regression estimator based on ordinary and quantile regression for the unchanged variables and their log-transformation using one auxiliary variable from two and both. Results of this simulation study for the total capital expenditure estimates in U are presented in the *Table 1*.

Table 1.

	Ordinary regression	Quantile regression	Log-transformation and ordinary regression	Log-transformation and quantile regression
x_1				
ARB, %	0.08	0.10	0.13	0.03
RRMSE, %	15.74	15.69	15.92	15.83
x_2				
ARB, %	0.12	0.21	0.13	0.11
RRMSE, %	15.86	15.87	15.98	15.93
$x_1 + x_2$				
ARB, %	0.18	0.13	0,13	0.05
RRMSE, %	15.71	15.68	15.90	15.83

For comparison, the coefficient of variation for the total capital expenditure in U in the case of SRS is equal to 18.44%, and in the case of stratification into U_p and U_c equals to 15.99 %. So, we haven't gained a big efficiency but this regression models can be used for domain estimation.

If we divide our population U into 27 regional domains and use simple stratified estimator (see Särndal *et al.* (2003)) the coefficient of variation for the population total will be equal to 18.38%. But if we use the population model based on quantile regression for domain estimation in U_p then our simulation results give us 16.47% of RRMSE. The estimates for domains will not gain much efficiency, but design of the survey can be simplified a little and the population estimates will be better.

References

Fox, J. (2002). *An R and S-plus companion to applied.*SAGE Publications.

Honchar, O. & Ianevych, T. (2012). A question of use of the previous information for sampling surveys. *Applied Statistics: Issues of Theory and Practice* **11**, 105 – 113. (In Ukrainian)

Koenker, R. (2005) *Quantile Regression*, Cambridge University Press.

Rousseeuw, P. J.& Leroy, A. M. (2003). *Robust Regression and Outlier Detection.*Wiley.

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

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis.*Springer New York.

Reproductive Health Survey: determination of sample size and design

Anna Larchenko

Belarus State Economic University, e-mail: annalarchenko@gmail.com

Abstract

In the paper the process of calculation sample size for Reproductive Health Survey is described for the samples of different kinds. Sample design is described and several variants of sampling frame are proposed to use.
Keywords: reproductive health, sample size, sample frame, sample design

1 Introduction

Under the prevailing socio-demographic conditions existing in Belarus (such as aging of population, low birth rates and relatively low life expectancy, high mortality rate, etc.) studying of reproductive health state is very important.

Official statistics of Belarus calculates and estimates only some indicators of reproductive health (for example, demographic characteristics, the incidence of STIs and HIV infection, etc.). At the same time social indicators reflecting health system efficiency are not calculated, medical literacy of the population is not explored, general characteristic of reproductive health is not provided.

Thus, the need of conducting Reproductive Health Survey comes.

2 Determination of sample size

Sample size affects, first of all, on the representativeness of the survey, on its cost as well as on its duration. For determining sample size we need to take into account: population size, pre-set value of permissible error, probability of response.

Required sample size for Reproductive Health Survey should be representative for the macro and regional level, and in the context of demographic and socio-economic groups (by gender, age, education, etc.). As the foreign practice shows, in the process of sampling multistage territorial sampling, primary and secondary sample units, clusters of households are often used. Considering this when calculating sample size we have to take into account additional factors:

- The design effect (*deff*) which reflects the impact of stages of the selection, stratification, weighting;

- *A key indicator* which is used for calculation of sample size.

In survey practice suggested value of *deff* is 1.5 (Multiple Indicator Cluster Survey Manual, 2005), which appears the maximum possible value of this quantity. Variability of most indicators in samples does not exceed the specified value.

As *a key indicator* it is recommended to take one of the most important for the survey. On its basis maximum possible sample size is calculated, which makes it possible to obtain a representative estimate for a small or minimal layer, covering about 2.5% of the population. As a possible indicator it is appropriate to take one of the indicators which are typically low.

To calculate sample size several variants are possible:

1. The sample size calculation in terms of simple random sampling and design effect:

$$n = \frac{4r \cdot (1-r) \cdot 1.5 \cdot 1.1}{(0.12r)^2 \cdot p \cdot n_h}, \quad (1)$$

where n is a required sample size; 4 is the coefficient, providing 95 percent confidence level; r – predicted or expected prevalence (coverage rate) of the key indicator; 1.1 – the coefficient that is required to increase the sample size by 10% for non-response compensation; f – deff; $0.12r$ – the margin of error acceptable at the 95-percent confidence level, defined as 12% of r (a relative sampling error for the r); p – proportion in the total population, which is based on the parameter r; n_h is the average household size.

2. The sample size calculation in terms of stratified simple random sampling of households by regions:

$$n = \frac{1.5 \cdot 1.1 \cdot 4 \cdot \sum N_i^2 \cdot w_i(1-w_i)}{N^2 (0.12w)^2 n_h^{'}}, \quad (2)$$

where n is a required sample size; 4 is the coefficient, providing 95 percent confidence level; $0.12w$ – the margin of error acceptable at the 95-percent confidence level, defined as 12% of w (a relative sampling error for the w); N_i – number of sample units the i-th group of the total population; N – total population size; w_i – share in the i-th sample group; $n_h^{'}$ – number of women aged 15-49 years per one household.

At the first step we need to find a proper key indicator. After some calculations the best variant is the share of first births up to 30 years in the number of live births. According to this indicator different possible sample sizes have been calculated (with or without *deff*, for various sample errors – Tab. 1-2).

Table 1: The calculation of the annual sample size for option 1 (key indicator – the share of first births up to 30 years in the number of live births).

The Republic of Belarus

Region	Key indicator, r	Design effect, f *(deff)*	Share of women 15-29 in women 15-49, P	Average household size, n_h	Non-response rate, %	Sample size, $n = \frac{4r \cdot (1-r) \cdot 1.5 \cdot 1.1}{(0.12r)^2 \cdot p \cdot n_h}$ with error Δ_h	
						Δ_h = 0.12	Δ_h = 0.08
Belarus, including:	0.464	1.5	0.421	2.43	10	3687	8296
Brest region	0.421	1.5	0.402	2.49	10	629	1415
Gomel region	0.468	1.5	0.410	2.32	10	532	1197
Grodno region	0.454	1.5	0.403	2.39	10	573	1290
Minsk	0.460	1.5	0.484	2.39	10	518	1010
Minsk region	0.469	1.5	0.390	2.48	10	544	1223
Mogilev region	0.490	1.5	0.401	2.44	10	483	1087
Vitebsk region	0.502	1.5	0.409	2.46	10	479	1078

Table 2: The calculation of the annual sample size for option 2 (key indicator – the share of first births up to 30 years in the number of live births).

The Republic of Belarus

Region	The share of first births up to 30 years in the number of live births	Number of women 15-29 years	Sample size, *n*, number of households		
			μ_{rel}=0.06, Δ_{rel}=0.12, (without *deff*)	μ_{rel} =0.06, Δ_{rel} =0.12, (with *deff*)	μ_{rel} =0.04, Δ_{rel} =0.08, (with *deff*)
Brest region	0.421	137868	24	38	87
Gomel region	0.468	149187	29	46	104
Grodno region	0.454	106204	15	23	52
Minsk	0.460	260977	88	140	315
Minsk region	0.469	131881	23	36	81
Mogilev region	0.490	108082	15	24	55
Vitebsk region	0.502	123561	20	32	72
Belarus	**0.464**	**1017758**	**214**	**340**	**765**

If we will calculate sample size without territorial stratification (taking into account that key indicator is 0.464 and average household's size is 2.43) we have: 518-1166 households. The result of calculation (Tab. 1) shows that for Belarus the sample size of 3700-8300 households is acceptable if we make stratification; for regions – 480-1400 households (it depends on limit error). When using stratified random sampling by regions households' sample size is significantly reduced (Tab 2) – 214-765 (it depends on limit error and design-effect). For the survey territorial stratification is very important. Considering this the most acceptable sample size is 6000-8000 households.

3 Sample design

For science-based sample design it is necessary to form sampling frame properly. Sampling frame is a list of all the units of the target population. For all mentioned units their coordinates have to be known: name and address, telephone number or email address. This list of units with their coordinates may be submitted in hard copy or electronic form. In its absence geographic maps may be used.

In order to form a sample that is representative in terms of area coverage and the socio-demographic composition of the population at the national level, the subject of survey is all the areas of the Republic of Belarus and Minsk. The sampling frame is Census data of 2009 and current account data:

1) the set of urban settlements and rural councils within each area;

2) the set of census enumeration areas in cities and towns, the sets of settlements according to the census (village account) in rural areas;

3) the set of flats and houses (according to the census), which are used for selection of final units – households.

In constructing the sampling frame territorial stratification is made: by urban and rural areas. In turn, in each stratum may allocate sub-stratus of households by type of dwellings, composition, size, age of household's members. An additional stratum consists of the persons who moved into new dwellings during the year.

Sampling methods. For Reproductive Health Survey as the statistical practice of developed countries shows us (USA, Australia, the Netherlands, etc.), and the taking into account Belarussian experience (Households' Sample Survey of living standards) the most suitable model is a multi-stage random territorial sample.

Option I. Given the number of similarities between the Labour Force Survey and Reproductive Health Survey (sampling method, sample size, etc.), it is possible to use a sampling frame for the Labor Force Survey. However, in this case a necessary condition is the exclusion of women over 50 and men aged 60 years or more.

Option II. Forming a new sampling frame. For this case the most appropriate way is to use the same three-stage sample survey which are already used in national practice: Households' Sample Survey of living standards, Labour Force Survey.

4 Concluding remarks

Using several sample variants we've got that acceptable sample size for Reproductive Health Survey in Belarus is 6000-8000 households. Several variants of sampling frame: sampling frame for the Labor Force Survey or new sampling frame can be used. On the author's opinion the most acceptable variant is forming a new sampling frame. The most suitable model of Reproductive Health Survey is a multi-stage random territorial sample.

References

Demographic Yearbook. Minsk: National Statistical Committee of Republic of Belarus, 2012.

Multiple Indicator Cluster Survey Manual, 2005

Statistical Yearbook. Minsk: National Statistical Committee of Republic of Belarus, 2012.

Women and men of the Republic of Belarus: statistical book. Minsk: National Statistical Committee of Republic of Belarus, 2010.

World Health Organization [electronic resource]. *Reproductive Health*. WHO, 2011. Mode of access: http://www.who.int/topics/reproductive_health/ru/index.html.

Experience of teaching survey sampling with the Moodle environment

Natalja Lepik

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

Abstract

Aspects of e-learning are observed. The Moodle environment as the possibility for distance or partially distance learning is described and will be demonstrated during the presentation. The experience of teaching course on survey sampling (as a partially supported by Moodle) is represented.
Keywords: Moodle, distance learning, teaching survey sampling

1 Introduction

Nowadays, the distance or partially distance learning becomes more popular and natural way for the teaching process. Some years ago, we used Excel tables for marking student's results, or some web environments for uploading teaching materials (pdf-s, slides or even video lectures). Today it is possible to organize the whole teaching process through the web browser. It means communication between students and teacher, tests (with automatical control and feedback for students), e-books and of course the evaluating system – all this without any need for programming skills.

This year, the partial support of e-learning of the basic course on survey sampling was implemented for the first time. One of the goals was to find out if the e-learning gives non-worse (or even better!) results compared to the traditional in-class-method. More precisely, it concerns the following futures:

- Does the e-learning give more benefits for students? For example, is it important for them to receive a feedback of the homework more quickly? Is it really important for students to have a possibility of choosing the time and place for studying?
- Will the knowledge level fall down?
- Is it possible to organize some group-work (say project) through the internet?
- How to compose the knowledge control tests? How to be sure that student is doing it alone and without any material?
- From the teacher's point of view – does the creating of such course and managing the teaching process bring any time benefits? Or oppositely, it takes much more time compared to the traditional system.

We decided that creating such web-supplement for the basic survey sampling course will be carried out during the teaching of it. So, it will be at the beginning of so called "partial support" with the majority of lectures in the classroom. Together with students we will try to understand the needs of the web-supplement.

2 The structure of the course

2.1 Traditional approach

The traditional basic course on survey sampling lasts through all the spring semester and gives 6 ECP (*European Credit Points*) for students with 1 ECP = 26 hours of work. It has 64 hours of class work (lectures and practical exercises in the computer class) and 92 hours for self-studying, during of which students have a lot of theoretical and practical tasks to solve.

As a group-work students have project-tasks with real data to investigate and find out the best solution for it. Then, they exchange solved projects between the groups and every group write a review on the received project. At least, all groups present and defend their projects in a class.

The course ends with the theoretical exam, which is carried out in the classroom. The final course result consists of the exam score and defended project work. Besides, all tasks given in lectures should be passed.

2.2 Aims of the partial web-support

From this approach we want that the teaching process will be partially organized through the internet. Some lectures can be given as a video lecture or e-books. But these methods do not guarantee that students have really read or seen those lectures. The better variant could be some kind of interactive lesson, where students read the material and answer some control-questions that allow them to go further with the lesson or send them back to the previous topic.

The web-support should include some evaluation system that controls itself if student has passed certain criterions or not. Students should see their own current results from this system as well.

During our lectures we are giving very many small tasks to our students. All of them should be examined, results marked to the evaluating table, and controlled answers should be given back to the students. Sometimes, students need to correct their answers and submit them again. This process produces the enormous paperwork. So, the web support should simplify this routine.

So far, we have had a course list (with course participants' e-mails in it) for messages and questions. Unfortunately, it wasn't very popular among the students. For sending messages of teaching stuff it worked well, but it was impossible to use e-mail system for typing formulas. So, web-forums with formula support are needed.

It would be nice to hold all features that are related to the course in one place, where students and teaching stuff could easily access it.

3 The Moodle environment

According to the official web-site [1], the "**Moodle** (abbreviation for Modular Object-Oriented Dynamic Learning Environment) is a free source e-learning software platform, also known as a Learning Management System, or Virtual Learning Environment".

Moodle can be developed and adopted to the needs of concrete organization. For the University of Tartu it is important that Moodle is related with our web Study Information System (SIS) that contains all information about all courses in the University, also information about students and their results. So, when students are registering themselves to the course through the SIS (it is required by Study Regulations), we can easily add them to the correspondent e-course. Students can access the e-course with the same user name and password that they have in the SIS. If somebody decides to cancel the course (it is allowed at the beginning of the course), the Moodle synchronizes the list of students with the SIS and corrects the list of students in the e-course.

Our Moodle is translated into Estonian, but can be also accessed in English, Russian and some other languages. The English version can be found in [2].

3.1 Features of Moodle

The main features of Tartu University's Moodle (like of the standard Moodle as well) are

- **Exercise submission** - answers can be easily downloaded or written directly in the Moodle by students, where teaching stuff can easily access them and check, also comment if needed; results are sent directly to the evaluating table;

- **Different forums** – discussions can be organized at so called "open level", which means that all students can see and participate in this forum; as well as "private forums" for workgroups, where only the group members can access discussions;

- **File downloads** – both students and teachers can upload files with lecture materials or tasks answers;

- **Tests** – can be quite easily created like Hot Potato tests with automatic evaluation and feedback for students;

- **Online calendar** – standard feature nowadays, but very useful within the e-course; all important dates are recorded there and the system informs course participants about upcoming events;

- **Wiki** – can be used for developing particular topic together with all students; everyone can complete or correct the text on some topic; as a result, the summarising table on survey sampling methods can be formed (description, pluses and minuses of the methods);

- **Vocabulary** – can be used for definitions used in the course;

- **Evaluating system** – special forms or even matrices can be used for obtaining a final result of some project, exam, test and so on. For example, the credits for the project-work can be defined as a sum of its components (e.g. correct form, content, defending and so on). All these components and

feedback of these parts are also seen by the students.

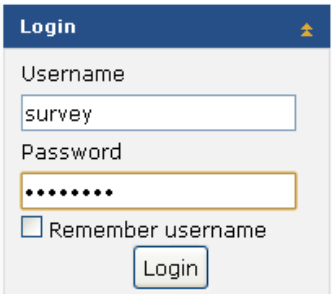
- **E-books** – quite easily can be created within the Moodle (can be compared with writing in Word), but still it is the set of web-pages that are related with each other having it own table of contents.

- **Interactive lesson** – exciting future for the self-studying. Students read one part of the lesson, and then they need to answer some questions. If all answers are correct, then the next part of the lesson is given for them. Otherwise they can be guided back to the same topic or to the page with more precise explanation of the topic. Creating such lesson needs to be deeply thought through, usually it take time, but technically it is easy to implement in the Moodle.

3.2 Course example

The course can be accessed as a student to see and try the features described above. First, go to the main web-page: http://moodle.ut.ee , where you should choose the English version (or Russian as well). Then login with the username **survey** and password **sampling** (see the Figure 1).

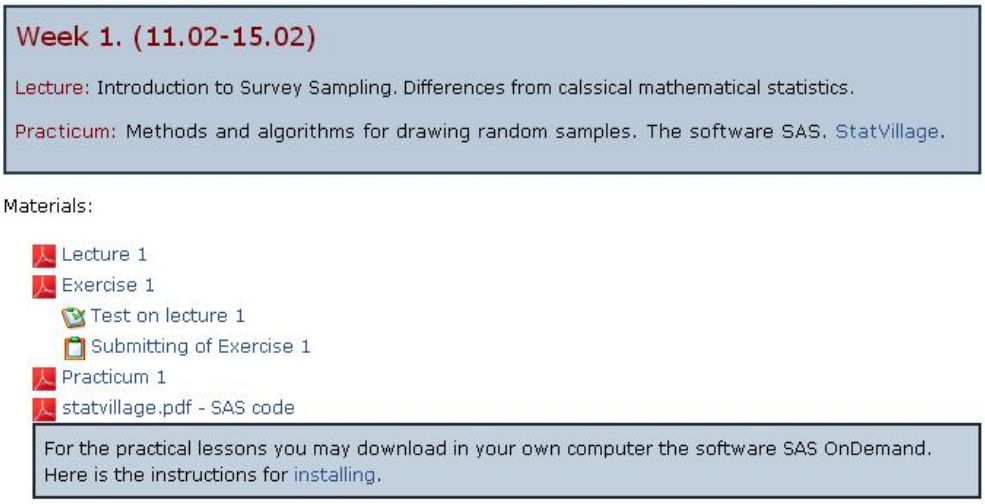
Figure 1: Login to the Moodle course



You should see then the Estonian name of the course *Valikuuringute teooria I*, click it and then you can use all materials and features that you see there. The interface of the Moodle should be in English, but other options depends on the language that disainer of the course has used. For this course they are in Estonian.

The course is devided into 16 blocks which corresponds to the 16 weeks of the course period. Each week has its own topic (in blue block, see Figure 2).

Figure 2: Typical week-block of the e-course



All weeks are listed below each other but can be easily accessed from the block of Section Links with the current week in bold (Figure 3).

Figure 3: The block for switching between weeks

On the right side of the course page is the forum block, the blocks of upcoming events and messages are also added. Due to the enormous number of different options that are used in the course, it may be unreasonable to access the material by weeks. For this purpose the block of activities on the left side is summarized by their type (see Figure 4).

Figure 4: The block of activities

The Moodle has very many possibilities for designing a course, the amount of features is huge and time-consuming at the beginning. Fortunately, we have very strong supporting team, who solves all our problems.

Some of the activities will be presented during the conference.

4 Conclusions

1. The course on survey sampling ends on the May 30th, then the feedback from the students will be received. So far, the conclusions from the lecturer's point of view can be described.

2. The creating of the e-course is very time-consuming, especially for the first time. This concrete course is completed thanks to our support team.

3. Next year with repeated course, the benefit of using Moodle's e-course is certain.

4. Managing of the teaching process is easier with the e-course than with the traditional course. A lot of paper work disappears. Very many options are automated (test results, grades).

5. This year only the partial support of distance learning is implemented. If the feedback from students will be positive, then the part of the in-class-meetings can be reduced, but still it is needed.

References

[1] The official web-page of Moodle (27.04.2013), https://moodle.org/

[2] The version of Moodle in University of Tartu (27.04.2013), https://moodle.ut.ee/?lang=en

Indirect Estimation of Monthly Unemployment Indicators for Regional Level in Ukraine

Olha Lysa

Ptukha Institute for Demography and Social Studies, National Academy of Science of Ukraine, e-mail: Olysa@ukr.net

Abstract

The paper presents two-stage approach to unemployment rate estimation for the regional level based on the SAE methods. Proposed two models for estimation by monthly results of LFS: design-based estimates – based on the realized rotation scheme and used the information from previous period of observation (1); model-based estimates – EBLUP area-specific model (2).Result of unemployment rates estimation are illustrated for all regions of Ukraine.
Keywords: indirect estimation, unemployment, contributed paper

1 Introduction

Employment and unemployment indicators are the basis for development of the well-grounded social and economic policy and estimation of its efficiency. The information needs about the condition and tendencies on the labour market constantly grow and first of all about labour force characteristics. Thus information necessity exists both on international, and at the state, regional and local levels. According to modern international standards of statistical information quality (Helsinki, 2002), one should be characterized by maximal completeness and timeliness, should corresponds to users' needs, should be reliable, accessible and clear, comparable in time and in the space, coordinated with the available comparable data from other sources. Also the important aspect is expediency, optimality of expenses financial and manpower resources on data obtaining.

The most widespread and recognized in the world the way for reception of the information concerning employment and unemployment is the sample labour force survey (LFS). Unconditional advantage of this method is integrated approach and completeness of the data, flexibility and ample opportunities for the analysis.

Getting of effective estimates of employment and unemployment indicators by LFS results for areas or groups of population with the insufficient sample size is a important question for Ukrainian official statistic. Especially sharply this problem appears for employment and unemployment indicators: (1) monthly estimates of unemployed for regional level; (2) annual estimates of employed and unemployed for local or municipal level.

2 Direct Estimates Reliability

Ukrainian Labour Force Survey is fulfilled by State Statistics Service of Ukraine (SSSU) on the constant basis from 1995. From 2004 the LFS is conducted on the monthly basis according the account ILO and EC recommendations, survey data quality requirements and the fullest satisfaction of users' needs. The sample design was constructed with accounting a possibility to get reliable estimates of main indicators on the national level but for the regional and local levels the question is still opened.

Direct estimates of indicators which calculated directly by the survey results take into account statistical weights of respondents. The estimator for calculation of employment and unemployment indicators estimates at presence of additional information is defined as (Ghosh M., Rao J.N.K., 1994):

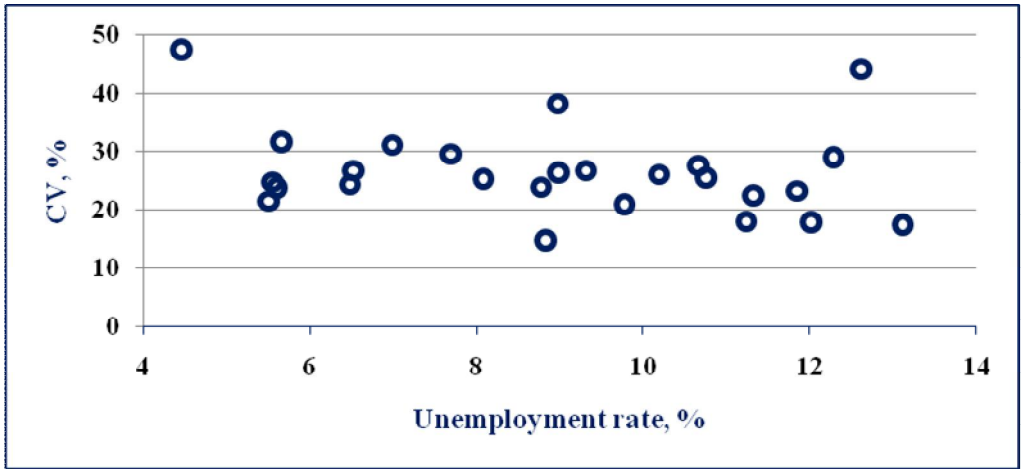
$$\hat{Y}^{(D)} = \hat{Y}^{(HT)} + \sum_{d=1}^{D} \hat{\beta}_d \left(X_d - \hat{X}_d^{(HT)} \right), \qquad (1)$$

where $\hat{Y}^{(D)}$ – direct estimator; $\hat{Y}^{(HT)}$ – Horvitz- Thompson estimator; $\hat{\beta}_d$ – estimates of regression coefficients (d =1,2,…,D – number of auxiliary data); X_d – values of indicators estimates, known from external sources; $\hat{X}_d^{(HT)}$ – Horvitz- Thompson estimates for auxiliary data.

Data of last population census, current data of demographic statistics, and current data of social statistics as for number and placement of institutional population are used as external information. One is used by the calibration of statistical weights system for what special procedures are developed (Sarioglo, 2005).

At the estimation of LFS indicators reliability method of balanced repeated replications is used. Special research of estimates quality for employment and unemployment indicators has shown that monthly estimates of unemployment can be used for quantitative analysis on nation-wide level (CV < 5%). At the regional level indicators estimates aren't suitable, in most cases CV ≥ 20% (see, for example, fig. 1).

Figure 1: Accuracy of unemployment rate monthly direct estimates (LFS, May 2010)



This figure presents variation coefficient of monthly estimates of e unemployment rates for regional levels for May 2010. From the presented data it is evident that variation of unemployment rate estimates is too high

and can't be used for analysis.

The accuracy of main indicators which are measured in the LFS is satisfied for national level. The estimates received for lower levels in many cases are insufficiently reliable and demand application of special approaches for more precise definition. It is typical also for indicators estimation of separate social and economic groups of the population. Therefore in the state statistics of Ukraine more and more attention is given to the problem of calculation of reliable estimates of these indicators for the regional level (regions, districts, separate cities).

2 Indirect Estimation

Certain measures on optimization of sample design with the purpose of receipt of more reliable estimations at regional level are constantly conducted, but these measures are unable to work out all problems. Always there is a requirement in the information for the lower and lower aggregating levels or certain small groups of population. Thus the application of the special statistical and mathematical models taking into account present external information is the most effective method receipting of reliable information for small areas.

In this paper two stages procedure for estimation the unemployment indicators on the regional level which measured on the monthly base is proposed: (1) Design-based estimator – monthly composite estimator (Design and Methodology, 2000); (2) Model-based estimator – EBLUP area-specific model (Rao, 2003)

2.1. Monthly composite estimator

Monthly composite estimator is considered with using the information from previous periods of observation for the same small area. Estimator for unemployment level (number of unemployed) in view of monthly rotation for month t is (Local Area Unemployment Statistics, 2001; Design and Methodology, 2000):

$$Y_t' = (1-K)\hat{Y}_t + K(Y_{t-1}' + \Delta_t), \qquad (2)$$

where: Y_t' – indicator composite estimate for the current month; $\hat{Y}_t$ – direct estimate for the current month; Y_{t-1}' – composite estimate of the previous month; Δ_t – estimate of changing concerning the previous month that received on the basis of 4 rotation groups data, which are the common for months t and $t-1$; K – weight coefficient.

The value of coefficient *K* is defined from the condition of variance minimization of indicator composite estimate:

$$V(Y_t') = (1-K)^2 \cdot V(\hat{Y}_t) + K^2 \cdot V(Y_{t-1}' + \Delta_t) \to \min . \qquad (3)$$

The results of the special researches showed expedience of the use of constant weighing coefficients during a year which calculated based on the data from previous year. Thus, for every region it follows to use the separate weighing coefficient for more adequate consideration of every region specification.

2.2. EBLUP area-specific model

The second model which proposed for estimation of monthly unemployment indicators on the regional level is a variety of composite estimator which uses the external data for small area – Empirical Best Linear Unbiased Predictor (EBLUP) (Rao, 2003):

$$\tilde{\bar{Y}}_d = \gamma_d \bar{Y}'_d + (1-\gamma_d)\mathbf{x}_d^{\mathrm{T}}\tilde{\beta}\,, \tag{4}$$

where: $\tilde{\bar{Y}}_d$ – EBLUP estimator for region d; $\bar{Y}'_d$ – monthly composite estimator for region d; $\mathbf{x}_d^{\mathrm{T}}\tilde{\beta}$ – synthetic estimator for region d; K – weight coefficient is defined from the condition of variance minimization.

In this type of models the synthetic estimator defines the influence external factors on the measure of the estimated indicator. The use of this model needs the deep analysis of additional information sources and correlations between estimated indicator and potential external data.

Special research of additional information defines four main sources of information (see Table 2): (1) Labour Fours Survey; (2) Business Survey; (3) Administrative Data (Register of unemployment and Pension found)

Table 2: Auxiliary information for estimation unemployment rate on the regional level

Sources	Available data	Correlation
Labour Force Survey	Labour force activity rate	-0.172
	Employment rate	-0.738
Business Survey	Part workers in population	-0.338
	No full-day employees level	0.163
	Ratio of income and outcome workers	-0.302
	Rate of workers outcome	-0.466
Administrative data	Registry unemployment rate	0.710
	Part of insured employees in population	-0.403
Demography statistic	High education level	-0.393

The results of correlation analysis showed possibility to use in the model next data: Employment rate (-0.738); Registry unemployment rate (0.710).

Not so strong correlation with Rate of wokers outcomes (-0.466) and Part of insured employees in population (-0.403) but in some conditions these data also can be used in model. In the certain case in EBLUP model we used only Registry unemployment rate as a characteristic with strong correlation.

Why didn't use Employment rate? In selecting external data for use in the EBLUP model from the parameters with similar correlation with the dependent variable should be preferred non-sampled or administrative data. This is a significant influence of sampling error on the resulting reliability level of indicators which are estimate based on statistical-mathematical models.

3 Results

Results of application of the two-stage method for estimation unemployment indicators for regional level on the basis of the data for previous survey periods and external data for current survey testify to efficiency of use of this approach to solving the problem of reliability improvement of LFS for results (see, fig. 2).

Figure 2: Comparison of models impact on the reliability of monthly estimates of unemployment rate for regions of Ukraine (LFS, May 2010)

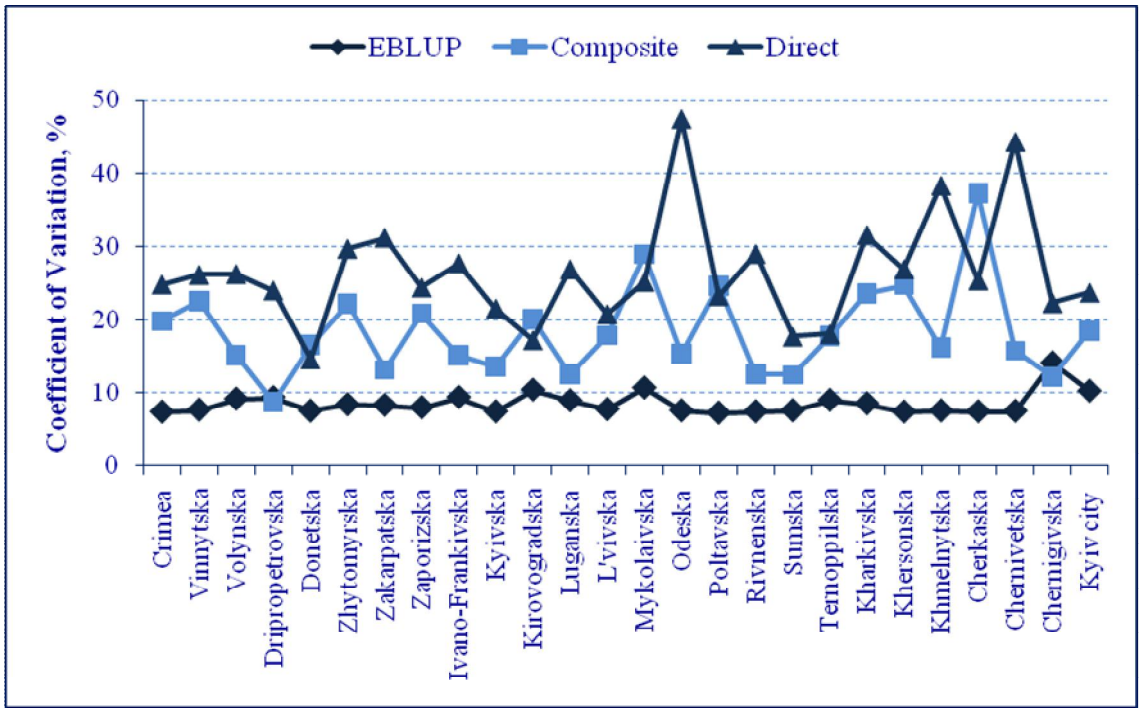


Figure 2 shows coefficient of variation (one of the main indicators of estimates reliability) of direct estimates and indirect estimates (monthly composite estimates and EBLUP) for regions of Ukraine. Comparing every stages of estimation can observe:

- for 5 regions not a significant deterioration in the estimates reliability after application of the monthly composite model, what can be explain by sharp change in parameter values from month to month. The use this model over time leads to some smoothing and increase reliability rate;
- in 2 cases the impact of monthly composite model on the reliability rate is significant and compensates the influence of the EBLUP model use.

But general improvement or estimates reliability for unemployment rate is observed for all regions of Ukraine in average in 3 times (between direct estimates and EBLUP).

The first type of models the monthly composite estimation was realized on micro-level with the procedure of reweighing which account received composite estimates of employment, unemployment and non-in-labour force levels and already used by the Labor statistics department of SSSU.

Recently, more and more research are aimed at developing models of the second type, which are used

additional information for the small area estimation of employment and unemployment indicators. Also performed in-depth analysis of potential sources of information that can be used for construction of models and analyze the quality of this information.

References

Design and Methodology. (2000) Current Population Survey: Technical Paper. Washington: U.S. Department of Labor, BLS.

Ghosh M., Rao J. N. K. (1994) Small Area Estimation. An Appraisal, Statistical Science. Vol. 9, № 1, 55–93.

Lysa O. (2009) Improvement of Reliability of LFS Indicator Monthly Estimates Proceeding of the Baltic-Nordic-Ukrainian Summer School on Survey Statistics. Kyiv, "TBiMC", p. 106 -113.

Quality Guidelines for Official Statistics. Helsinki: Statistics Finland, 2002.

Rao J.N.K. (2003) Small Area Estimation. Wiley, New York.

Sarioglo V. (2003) Methodological Approaches to Increase of Data Reliability Level of Sample Surveys of Population Economic Activity Theory of Stochastic processes. Vol. 9(25), № 3–4, p. 176–183.

Mixed mode data collection pilot survey on Consumer survey: Results on response

Saara Oinonen

Statistics Finland, e-mail: saara.oinonen@stat.fi

Abstract

Statistic Finland has a strategy that in all data collection routines concerning individuals, web response option should be available by the year 2017. Mixed mode data collection implies that there are optional means for person to respond. There has been two CATI + web mixed mode pilot projects on Finnish Consumer survey, one in 2011 and more recent in 2012. Response rates and factors affecting the response mean have been some of the objectives on these studies.
Keywords: consumer survey, data collection, mixed mode, nonresponse, response, web questionnaire

1 Introduction

Traditional data collection methods on person statistics are paper questionnaire, computer assisted personal interview (CAPI) and computer assisted telephone interview (CATI). Development of IT and networks has increased demand for web response alternatives. On mixed mode data collection method surveyed persons are offered alternative options for responding. The effect the response mean has on the results (*mode effect*) has been widely under research (see for example Kreuter, Presser and Tourangeau, 2008), but it will not be discussed here any further.

Statistics Finland have conducted several mixed mode pilot projects on person statistics and most recent is 2nd mixed mode pilot survey on Consumer survey. Project ended on May 2013 and results can be read from final project report (Heikkinen, 2013). This article will focus on response rates and factors affecting the choice of response mean on this mixed mode pilot survey.

2 Background

2.1 Consumer survey and Finnish travel joint data collection

The consumer survey studies consumers' views and intentions relating to economic matters. This monthly released statistic has very fast schedule determined by European Commission. Data collection is done on first two weeks of every month and statistics are produced within few days. At the moment data is collected by computer assisted telephone interview (CATI). Statistics Finland has a strategy that by the year 2017 all surveys done for persons, including Consumer survey, have web questionnaire as one option for responding.

The Finnish Travel Survey contains information on trips made by Finnish residents and on persons having

travelled. Data collection has been combined with Consumer survey due to similar target population and production schedules. Sampling for this joint data collection is done in every 6 months. Sample size is 2 350 persons per month and it covers population of Finland aged from 15 to 84. Sampling method is systematic random sampling and the sample is self-weighting according to regional population density. Sample is drawn from the central population register.

The response data of both surveys are expanded to the whole population with weighting coefficients, which are calculated with a calibration method and by using observations' inclusion probability. Response rate is currently around 62 % and nonresponse is gradually increasing. Since response rates are very similar for both surveys this article will focus on results of Consumer survey.

2.2 Mixed mode pilot surveys

Joint data collection of Consumer survey and Finnish travel survey has had one mixed mode pilot in March 2011. On this earlier research the prospective and design were very different than on the more recent pilot conducted in 2012. Survey sample size was then 2 200 persons per month and target population were aged 15 to 74. Pilot sample size was 4 000 and sample units were divided in groups by the mean of response and whether a phone number was found for target unit. Some groups were assigned to answer entirely on web questionnaire and one group was assigned for mixed mode response. For those who were allocated to mixed mode group the options to respond either by web questionnaire or by phone were available during the whole survey time. The main ambition was to investigate if mean of response had impact on results in these surveys (mode effect) and for that purpose it was important to gain enough results from web questionnaire. (Simpanen, 2011.)

The topic of this article is the more recent mixed mode pilot survey, which had sample size and design exactly similar to the regular monthly data collection of Consumer survey. In the beginning of 2012 sample size of the joint data collection was increased to 2 350 persons per month and population aged 75-84 were added to target population. Surveyed month was November 2012 and data collection schedule of the pilot survey was similar to the regular survey. This time the whole sample had option to answer either by web questionnaire or by phone, but the option for web questionnaire was suspended after 7 days of the beginning of data collection. This way it was not possible for respondents to refuse from phone interview by referring to the web option, which was a problem in the previous pilot survey. Now the respondents had 7 days for web questionnaire and after that remaining units were interviewed by phone as in regular survey.

3 Data collection and response

3.1 Data collection process

Data collection for pilot survey started on Thursday November 1st as the web questionnaire was opened. Sample units received information letter, including passwords for web questionnaire, on Wednesday October 31st. It was fortunate for the project that the starting day of the data collection was not on the weekend or on Monday, since mail is not delivered on weekend in Finland. Also reminder notes were sent and sample units received them on Monday November 5th. On the previous study the reminder notes were proven to be effective for gaining responses. This conclusion was confirmed on the 2nd pilot survey since over 30 % of all web responses were given a day after sample units received the reminder note. Thursday November 8th was the final day for web responses and on Friday November 9th the CATI-interviews begun. Last CATI-

interviews were done on November 21st which exceeded the regular data collection period by two days. This was due to the busy situation of CATI interviewers. Received web responses per day are shown in Table 1.

Table 1: Received web responses

Data collection day	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri[1)]
- number of web responses	15	88	45	31	32	178	94	91	8
- % of all web responses	*2,7*	*15,9*	*8,2*	*5,6*	*5,8*	*32,2*	*17,0*	*11,1*	*1,4*

[1)] Due to technical issues, web questionnaire was closed on Friday morning, so few responses were given on Friday 9th.

3.2 Response rates and nonresponse

Overall response rate for mixed mode survey was 56 % with 1 307 responses altogether. Regular Consumer survey on November 2012 had response rate 61 % which means that pilot survey had 5 percentage points more nonresponse than the regular survey. Nonresponse analysis shows that amount of persons refused to respond was 5 percentage points higher in pilot survey than in regular survey. Hence, the increased nonresponse can be covered with the increased amount of refusals. Nonresponse analysis is demonstrated on Table 2.

Table 2: Nonresponse analysis

Type of nonresponse	Refusal	Failed to reach	Other reason[1)]	All nonresponse	Response rate
Regular survey	9 %	27 %	3 %	39 %	61 %
Mixed mode	14 %	28 %	3 %	44 %	56 %

[1)] Other reasons can mean for example lack of language skills or respondents poor condition/illness.

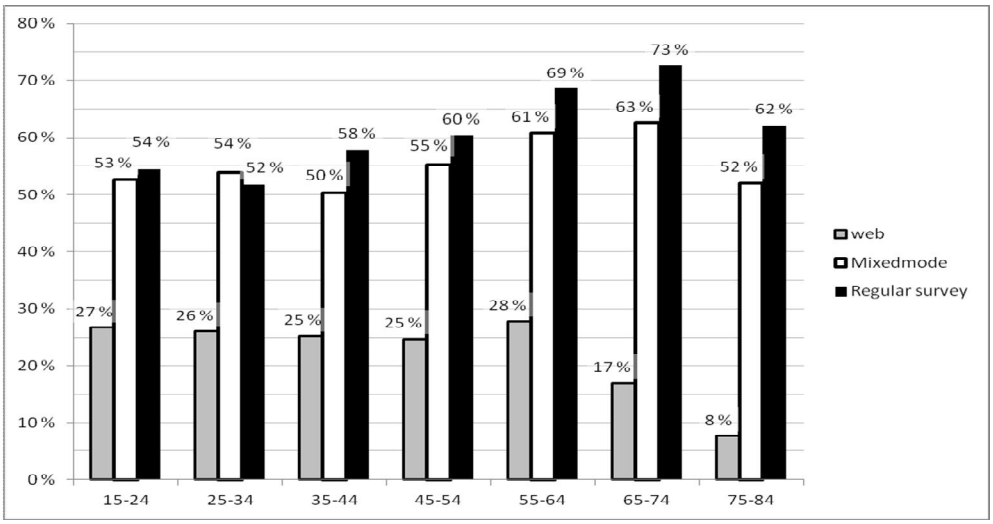
Mixed mode survey gained 552 web-responses and 755 CATI-responses. Web response rate was 24 % of the sample and 42 % of all responses. Detailed information on response rates of the surveys is shown in Table 3.

Table 3: Response rates

		Regular survey	Mixed mode survey		
			all	*CATI*	*web*
Response rate	*n*	1 435	1 307	755	552
Proprotion of sample	*%*	*61,1*	*55,6*	*32,1*	*23,5*
Proportion of responses	*%*			*57,8*	*42,2*

Response rates of these two surveys were analyzed and it became clear that two major factors resulting lower response rates for mixed mode survey were respondent's age and educational level. The effect of respondent's age was anticipated prior the survey and results show that elderly people are not willing to take part in the survey that had anything to do with computers or internet. Even when CATI interviewers explained that the subject of the survey had nothing to do with computers or internet it was very difficult to persuade elderly people to participate. Response rates by age group are described on Figure 1.

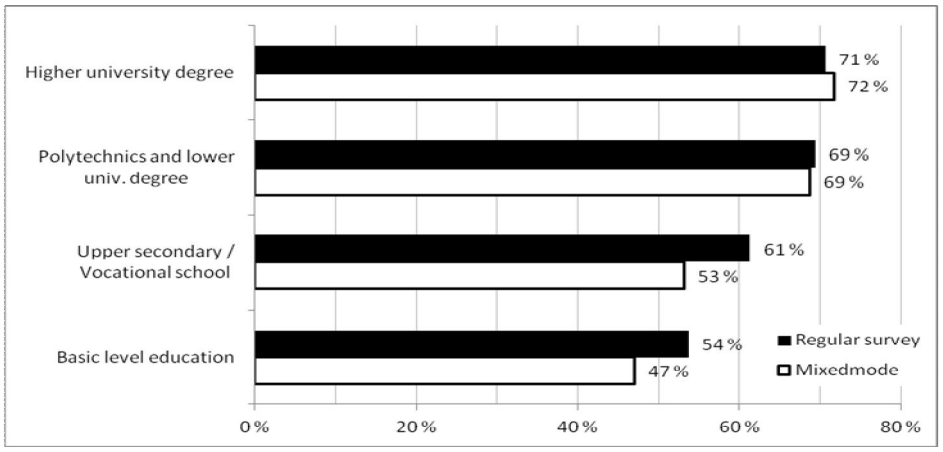
Figure 1: Response rates on mixed mode and regular survey by age groups



In addition to total response rates on mixed mode and regular surveys the web response rates on mixed mode survey are also shown (grey bar). Figure demonstrates that response rate is quite high in older age groups on regular survey, but adding the web questionnaire alternative decreases the response rate 10 percentage points. On the other hand the younger respondents did not perform any better, since it was anticipated that the younger population which normally participates surveys quite poorly would more likely to respond when alternative means are offered. Figure 1 indicates that offering web questionnaire encourages younger respondents very little. It is also notable that age group that had highest quantity of web responses (28 %) are persons aged 55–64.

It is generally acknowledged that persons with higher education level are more likely to participate in surveys than persons with less education. Yet, it was quite surprising that when alternative mean for response was offered the difference between education levels was even bigger. The response rates of mixed mode and regular surveys by education levels are illustrated on Figure 2. Mixed mode response rate drops significantly together with educational level.

Figure 2: Response rates by respondents educational levels



On basic education level, which in Finland refers to 9 years compulsory school, difference in response rates

of the two surveys is 7 percentage points. Upper secondary school and vocational school contains 3–4 years additional education. This comprises a major class of the target population, since 40 % of the sample units belong to this group. For this reason it is especially alarming that response rate on of mixed mode survey is only 53 % on this class. Together two lowest education levels cover 73 % of target population. Hence, it is very important to obtain good response rates from these groups. Third class includes graduates from polytechnics and bachelor degree graduates from universities. This covers 19 % of target population. Fourth class includes university graduates that have master degree or any higher degree, and it covers 9 % of target population. Since persons on these two upper classes tend to response well in all surveys additional web questionnaire option did not affect their response rate significantly.

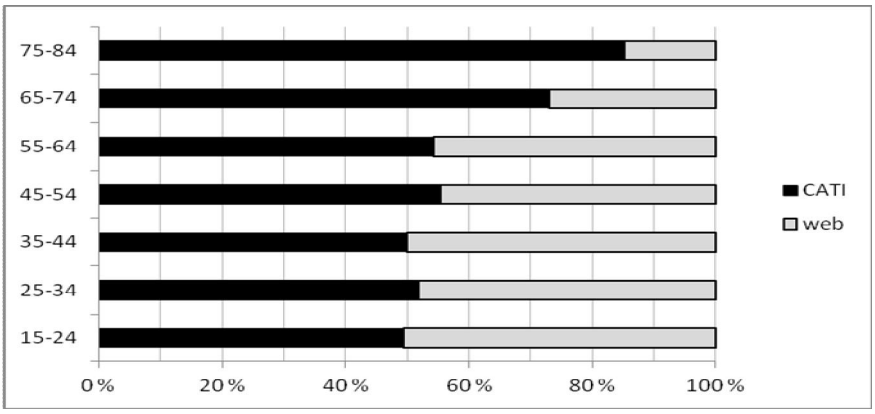
Other factors decreasing response rate of mixed mode survey were respondent's marital state and native language. However, these factors had less impact than education or age. The two survey data sets were joined and the impact of survey type on response was investigated with logistic regression analysis. Dependent factor was response state (0 = person did not respond, 1 = person did respond) and type of survey was one of the explaining variables (0 = person was on mixed mode survey sample, 1 = person was on regular survey sample). Results show that after education, marital status and age the type of survey was fourth most influential factor and that the impact of mixed mode type to the response state was negative. Results of logistic regression analysis can be found on Annex 4 of *Final report of Consumer survey and Finnish travel survey 2nd mixed mode pilot project* (in Finnish) (Heikkinen 2013).

3.3 Factors affecting the choice of mean of response

In mixed mode pilot survey, sample units had, in a way, choice of mean of response. During the first week of data collection the option was to answer on web questionnaire and on the second week it was possible to do CATI interview. Target units received two letters, one prior the data collection and one reminder of the web questionnaire alternative. Few persons considered these multiple contacts irritating but for majority it was fine. Feedback from sample units indicate that especially younger people did not even read their letters before the web questionnaire was closed. On the other hand, elderly people quite often misunderstood the letter and thought the survey was about IT-skills and internet habits. This denotes the need for more effective and subtle means for contacting the sample units. Often the person who ended up having CATI interview intended to respond via web questionnaire but web response time had run out.

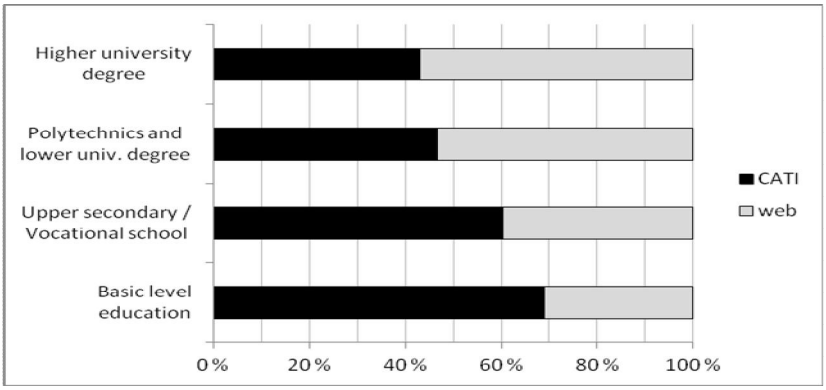
For those who did respond the factors affecting the choice of response mean were studied. Although this choice was not always made consciously, since the informing letters sometimes remained unread or misunderstood. Factors affecting the choice of response mean were much similar to the factors that affected on overall willingness to response. Again, it was anticipated that elderly persons would prefer the phone interview and age was the most influential factor. Other factors were person's educational level and sex. Some impact was discovered on the number of children under 7 years old in the household but the effect was not distinct. Response method grouped by respondent's age is described on Figure 3. It is noticeable that response methods are quite equally distributed until the age of 65, from which on the telephone interview is much more preferred option.

Figure 3: Mean of response by age groups on mixed mode survey



Differences on respondent's educational level are illustrated on Figure 4. It is apparent that preferred response mean changes after upper secondary or vocational school. Difference of chosen response mean between the lowest and the highest educational levels is almost 30 percentage points. It is not clear why person with lower education level avoids the web questionnaire. Educational level has some dependency on person's age and majority of elderly Finnish population has low educational level. On the other hand younger people under the age of 25 are almost all located on bottom two education groups and that should equalise the differences on response mean. Respondent's sex did not have great impact on the choice of response method but it was still evident that women are more likely to use web questionnaire than men.

Figure 4: Mean of response by educational level on mixed mode survey



4 Conclusions

Results indicate that offering web questionnaire as a response method will not improve response rates but it can even increase nonresponse. It the data collection time has been longer and response alternatives were more efficiently targeted to certain subgroups of population, response rates could have been much better. Also better means for contacting the sample units are required for effective mixed mode data collection. Elderly people could be excluded from mixed mode data collection entirely. For people with lower education level more effective persuasion is needed. However, mixed mode data collection method is not optimal for fast schedule surveys as hours for data editing and calculations of results increases. It is anticipated that, as time passes, web response options will be more desirable and official statistics should be able to meet that demand.

References

Heikkinen, T. & WEB04 project group (2013). *Projektin Kuluttajabarometrin ja Suomalaisten matkailu -tutkimuksen 2. mixed mode -tiedonkeruun pilotti loppuraportti,* Final report of Consumer survey and Finnish travel survey 2nd mixed mode pilot project (in Finnish). Statistics Finland.

Kreuter, F., Presser, S., Tourangeau, R. (2008) *Social Desirability Bias in CATI, IVR, and Web Surveys: The Effects of Mode and Question Sensitivity*. Public Opin Q.

Simpanen, M. & HO40W project group (2011). *Projektin EU-kuluttajabarometrin ja Suomalaisten matkailu -tutkimuksen web- ja mixed mode -tiedonkeruun pilotoinnin loppuraportti*, Final report of Consumer survey and Finnish travel survey web and mixed mode pilot project (in Finnish). Statistics Finland.

Analysis of commercial banks

Julia Orlova

Belarus State Economic University, e-mail: orlova-julia-gen@mail.ru

Abstract

The article is devoted to creating an optimal model of the banking system of the Republic of Belarus on the basis of the balance of six leading banks, whose assets constitute 80% of the assets of the system.
Keywords: Banking system, optimization model, non-probability sampling

1 Introduction

The aim is to create an optimal model of the banking system of the Republic of Belarus on the basis of the balance of six leading banks, whose assets constitute 80% of the assets of the system. Thus, the applied-probability sampling.

Each of the six leading banks to optimize their balance using nonlinear programming methods with regard to restrictions on liquidity ratios and the balance sheet. Then the distribution of the results to the general population of the 32 banks in the Republic of Belarus.

2 Optimization model

Optimization model of bank balance is a model for determining the amount of such active and passive balance sheet items, which would provide the maximum banking interest margin. The bank must ensure compliance with the National Bank sufficient capital, liquidity, the maximum allowable amount of credit.

We introduce the following notation: xi through, i = 1, m denotes ¬ denote the amount of asset i-th species in monetary units, m - number of active balance sheet of the bank; through x_{m+j}, j = 1,n denote the amount of liability j-th species in monetary units, n - number of passive balance sheet of the bank, through the d_i, i = 1,m denote the return of an asset i-th species, and by p_j, j = 1, n - acquisition costs of liability m = j-th species. Then the objective function, which expresses the interest margin, which should maximize will have the form

$$f(x) = \sum_{i=1}^{m} d_i \cdot x_i - \sum_{j=1}^{n} p_j \cdot x_{m+j} (\max) \quad (1)$$

We write restrictions.

The first limitation - the balance. The sum of active articles balance is the sum of passive articles plus own

capital

$$\sum_{i=1}^{m} x_i = \sum_{j=1}^{n} x_{m+j} + C \tag{2}$$

where C - equity capital of the bank.

The second limitation follows from the standard National Bank's capital adequacy. To write this limitation, we introduce additional restrictions through x_i, i= 1,m denote the degree of risk of the asset i-th species, through z_f, f = 1, f - off-balance sheet commitments amount Bank f-th species in the monetary units through k_f, - coefficient of equivalent credit risk outbalances obligations through r_j^b - the degree of risk as a function on the counterparty, R - the amount of reserves on assets under adherence to credit risk. The restriction on the accuracy of self- capital will take the form

$$\frac{C}{\sum_{i=1}^{m} r_i \cdot x_i + \sum_{f=1}^{F} z_f \cdot k_f \cdot r_f^b - R} \geq \begin{cases} 0{,}14(I) \\ 0{,}1(II) \end{cases} \tag{3}$$

I - bank runs less than two years;

II - Bank employs more than two years.

A third limitation follows from the norm of instant liquidity, which represents the ratio of assets to liabilities on demand and overdue.

We denote the set of assets I_{ADV} demand and past due U_{PDV} through the set of liabilities on demand, then the limit of instant liquidity will be of the form

$$\frac{\sum_{i \in I_{ADV}} x_i}{\sum_{j \in Y_{PDV}} x_{m+j}} \geq 0{,}2 \tag{4}$$

A fourth limitation of the standard should be short ¬ term liquidity (NCL) - the ratio of actual liquidity to the desired.

Denoted by l_i, i = 1, m degree of liquidity of the asset i-th species, and through r_{js} - removing the risk of liability (m + j)-th species, then the restriction on short-term liquidity will be of the form

$$\frac{\sum_{i=1}^{m} l_i \cdot x_i}{\sum_{j=1}^{n} r_j^s \cdot x_{m+j}} \geq 1 \tag{5}$$

The fifth limitation follows from the minimum value ratio of liquid assets to total assets of the bank.

We denote I_l the set of all liquid assets, then the constraint will be in the form

$$\frac{\sum\limits_{i \in I_{ADV}} x_i}{\sum\limits_{i=1}^{m} x_i - R_1} \geq 0{,}2 \tag{6}$$

where R_1 - funds set aside in the National Bank.

The sixth limitation follows from the norm as ¬ gauge risk on amounts in foreign countries that are not members of the OECD.

We denote the set of assets I_{mbk}, then corresponds to the restriction will be in the form

$$\sum_{i \in I_{mbk}} x_i \leq C \tag{7}$$

The seventh limitation write on the basis of the limited aggregate amount of all large exposures.

We denote I_{cl}. many bank customers, and through y_i, the total amount of claims to the i-th customer, while

$$\sum_{\substack{i \in I_{cl.} \\ y_i \geq 0{,}1CK}} y_i \leq 6 \cdot C \tag{8}$$

Eighth limitation follows from the standard maximum risk per customer and can be written as

$$y_i \leq \begin{cases} 0{,}2 \cdot C(I) \\ 0{,}25 \cdot C(II) \end{cases} \tag{9}$$

I - bank runs less than two years;

II - Bank employs more than two years.

Finally, we calculate the non-negativity conditions, economic meaning of which is that the amount of active and passive articles and underbalance items can not be negative:

$$\begin{aligned} x_i &\geq 0, i = \overline{1, m+1}; \\ z_f &\geq 0, f = \overline{1, F}. \end{aligned} \tag{10}$$

In the model may also include other restrictions resulting from the less important regulations or guidance offered by the bank.

References

Banking Code of the Republic of Belarus (2011). Amalfeya Minsk.

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

Small Area Estimation in Household Budget Survey 2006

Pauliina Peltonen

University of Helsinki, e-mail: pauliina.peltonen@helsinki.fi

Abstract

There are two ways to produce regional statistics in Finland: by using registers or by a sample survey. Most of the regional statistics are produced by using registers. The main point of this paper is to compare different small area estimation methods within the Household Budget Survey 2006. The three methods that are used in the comparison are Hájek estimator, GREG estimator and EBLUP estimator.
Keywords: Small area estimation, survey based estimation, Hájek estimator, GREG estimator, EBLUP estimator, standard error, analytical, bootstrap

1 Introduction

This contributed paper is based on my Master's Thesis, which I wrote in Statistics Finland during 2012-2013.

2 Small Area Estimation

In general, the term small area estimation means estimating different statistical parameters for small domains. A domain can be considered as 'small' if the size of the domain is approximately one percent of the size of the population. In this case, the sample size is not adequate to produce reliable direct estimates. Estimators can also be divided to design- or model-based estimators. Design-based estimators are based on the sample data only, as where the model-based estimators use some auxiliary information to improve the quality of the estimates. Design-based model-assisted estimators are some kind of 'intermediate form' between these two, as they can be both direct or indirect and use auxiliary information.

2.1 Direct and indirect estimators

Direct estimators are based only on the domain-specific sample data. They may also use known auxiliary information, such as the total of an auxiliary variable x, which is related to the variable of interest y. Direct estimators can be either design- or model-based.

Indirect estimators 'borrow strength' from other domains. Indirect estimators can also be either design- or model-based. If they are model-based, the statistical model is fitted to the whole sample and the estimate is then calculated by using only the values from the domain of interest.

2.2 Estimators and their equations

The three estimators, which were used in the comparison, are Hájek estimator, GREG estimator and EBLUP estimator.

Hájek estimator of domain total is the most simple estimator. It is based only on the observed values from the domain of interest. If the domain is very small, then the standard errors of the estimates can be very large. As a design-based estimator Hájek is design unbiased and the precision improves as the domain size grow.

$$\hat{t}_{Hájek} = N_d \frac{\sum_{k \in s_d} a_k y_k}{\sum_{k \in s_d} a_k}, \tag{1}$$

where N_d is the population size of domain *d*, y_k is the observed value of unit *k*, a_k is the weight of unit *k* and s_d is the domain *d* in the sample *s*.

GREG (generalized regression) estimator is design-based model assisted estimator. It uses both the observed and the fitted values, but unlike EBLUP estimator, GREG uses the residuals of the observed values to adjust for the possible bias of the synthetic part. It is nearly design unbiased (it's bias ratio (bias divided by the standard error) will reduce as the sample size grows).

$$\hat{t}_{dGREG} = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} a_k (y_k - \hat{y}_k), \tag{2}$$

where $\hat{y}_k$ is the fitted value of unit *k*, y_k is the observed value of unit *k*, a_k is the weight of unit *k*, U_d is the domain *d* in the population *U* and s_d is the domain *d* in the sample *s*. The fitted values are calculated from the statistical model, for example

$$y_k = \boldsymbol{x}_k^{'}\boldsymbol{\beta} + \varepsilon_k,$$

where $\boldsymbol{\beta}$ is the vector of fixed effects, $\boldsymbol{x}$ is the vector of auxiliary variables and ε_k is the vector of error terms. Now the fitted value of *y* is $\hat{y}_k = \boldsymbol{x}_k^{'}\boldsymbol{\beta}$.

EBLUP estimator is a fully model-based estimator. Like GREG estimator, EBLUP uses both the observed and fitted values as they are. It is design biased and the bias does not necessarily reduce as the domain size grows.

$$\hat{t}_{dEBLUP} = \sum_{k \in s_d} y_k + \sum_{k \in U_d - s_d} \hat{y}_k \quad , \tag{3}$$

where $\hat{y}_k$ is the fitted value of unit *k* and y_k is the observed value of unit *k*, U_d is the domain *d* in the population *U* and s_d is the domain *d* in the sample *s*. EBLUP estimator uses the mixed model as an assisting statistical model, where as GREG estimator usually uses fixed-effects model.

2.3 Statistical models

In model-based estimation, statistical models play a huge role. The models define the way that the related auxiliary information is incorporated in the estimation process (Rao, 2003).

The model is chosen based on the nature of the variable of interest. If the variable of interest is for example:

- Continuous, then the statistical model should be linear

- Binary, then the statistical model should be logistic

- Quantity, then the model should be logarithmic.

Models can be divided into two groups: fixed-effects models and mixed models. The difference between the two is the use of the area-specific random term.

In my thesis, all the variables of interest are continuous, so I used only linear model.

2.4 Bootstrap method

The main idea in bootstrapping is to construct resamples (bootstrap samples) from the original data. The bootstrap samples are taken from the original sample by using sampling with replacement. This process is repeated a large number of times (like 100, 500 or 1000), and the mean (or total) is calculated for each of these bootstrap samples (each of these means or totals are called bootstrap estimates). The variation of the mean (or total) can be calculated from the bootstrap estimates

$$\hat{s}_B = \left(\sum_{b=1}^{B} \frac{\left(\hat{y}_b^* - \hat{\bar{y}}^* \right)^2}{B-1} \right)^{\frac{1}{2}},$$

where $\hat{\bar{y}}^*$ is the mean of the bootstrap estimates.

2.5 Program RConsumer

RConsumer is a new program for the estimation of totals and means for population subgroups or domains and small areas. It is developed by Dr. Ari Veijanen from Statistics Finland. The program covers selected methods (Hájek for means, Horvitz-Thompson for totals, GREG and EBLUP for both) described in Lehtonen and Veijanen (2009).

RConsumer is executed with R, but it is used from SAS. Due to the features of R the program is quite slow if the sample sizes are big. One of the positive features of RConsumer is the possibility of bootstrapping.

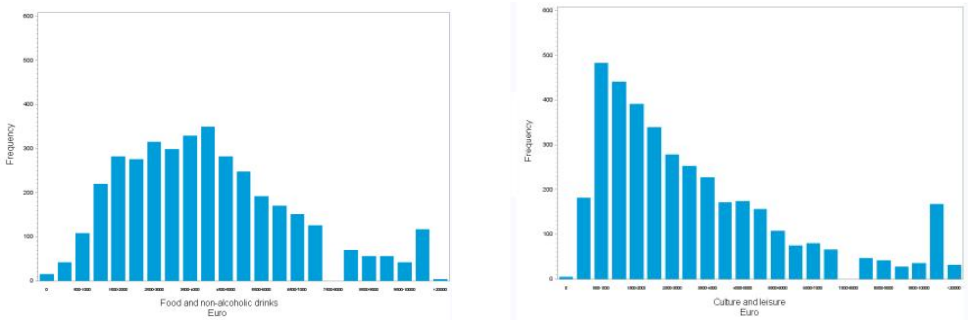
3 Data and results

3.1 Data and the auxiliary information

The data, which were used in this study, is Statistics Finland's Household Budget Survey 2006. It is a survey

about the consumption of Finnish households. The indicators which were selected for estimation process are two of the 'main classes' in the consumption survey: expenditure on food and non-alcoholic drinks and expenditure on culture and leisure. The frequencies of the indicators are shown in figure 1.

Figure 1: Frequencies of Food and non-alcoholic drinks and Culture and leisure.



The auxiliary information was taken from Statistics Finland's Total statistics on income distribution. Some very strong explanatory variables were found, such as the debt and the disposable money income of the household. Some variables, which represent the state of the household, were also selected as explanatory variables.

Multiple correlation coefficient (R^2) varied between 30 and 40 percent, which is quite high considering that the data is not longitudinal data.

3.1 Results

The comparison was carried out with different settings and area levels:

- Domains: 19 regions, 74 sub-regions and 415 municipalities

- Estimators: Hájek, GREG and EBLUP estimators

- Analytical vs. bootstrap standard errors

- Weighted vs. unweighted estimates and standard errors

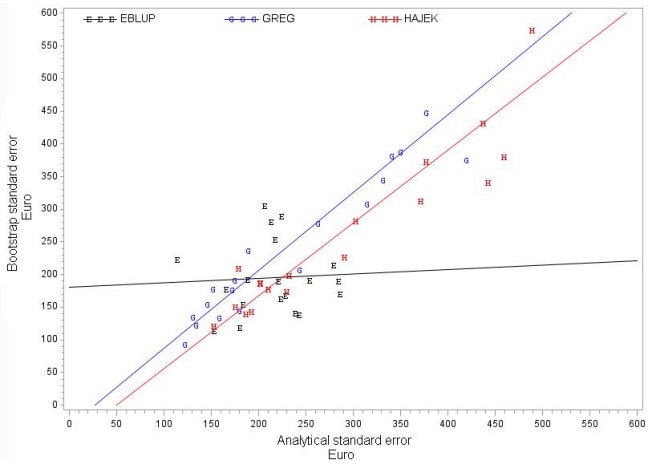
The comparison between the methods in different area levels (regions, sub-regions and municipalities) corresponds with the theory: if the sample sizes are large, as they are in regions, design based estimators work quite well. There were 415 municipalities in Finland in 2006, so most of them are quite small and the sample sizes are very low (even zero with almost 30 municipalities), so the design-based estimators do not produce good estimates. In fact, they are unusable in almost every municipality except the largest ones. Model-based estimators produce accurate estimates even if the sample size is low.

The comparison between weighted and unweighted standard error corresponds also with the theory: the weights have the bigger influence the smaller is the sample size. If the sample size is small, the weighted standard errors are usually bigger that the unweighted ones.

The most interesting result of my thesis came out with the comparison between analytical and bootstrap

standard errors. Analytical standard errors are based on Särndal, Swensson and Wretman (1992) and Lehtonen and Veijanen (2009). In figure 2 analytical standard errors are compared to bootstrap standard error at the regional level.

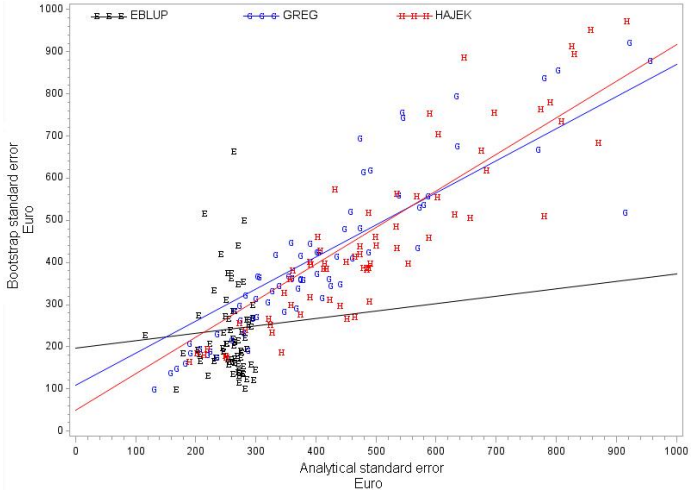
Figure 2: Analytical and bootstrap standard errors (regions).



For Hájek and GREG, analytical and bootstrap standard errors show reasonably good agreement. This is not the case for EBLUP whose analytical standard errors show too less variation.

In figure 3 analytical and bootstrap standard errors are compared at the sub-regional level.

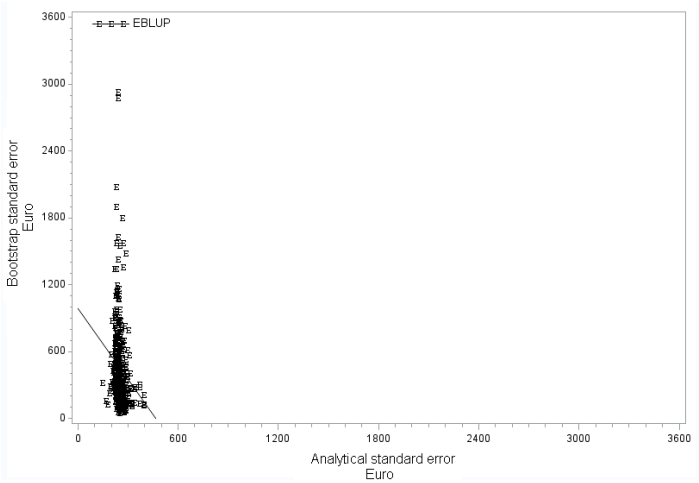
Figure 3: Analytical and bootstrap standard errors (sub-regions).



As can be seen from figure 3 Hájek and GREG standard errors are still quite equivalent, even if at the sub-regional level the sample sizes are much smaller than at the regional level. EBLUP seems to accumulate even more than before. It looks like the analytical EBLUP standard errors form some kind of ellipse shape.

Finally, in figure 4 there are analytical and bootstrap standard errors at the municipality level only for EBLUP estimator. The design-based methods Hájek and GREG are not accurate and precise enough at this area level.

Figure 4: Analytical and bootstrap standard errors (municipalities).



As can be seen from figure 4 the analytical standard errors pile up as the bootstrap standard errors vary quite well. This strongly indicates that the analytical standard errors are too optimistic.

References

Lehtonen R., Veijanen A. (2009). Design-based Methods of Estimation for Domains and Small Areas. Chapter 31 in Rao C. R., Pfeffermann D. Handbook of Statistics, vol 29B. Sample surveys: Inference and Analysis. New York: Elsevier.

Rao, J. N. K. (2003). Small Area Estimation. John Wiley & Sons.

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

Teaching Survey Sampling at Cybernetics specialities

Iryna Rozora

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

Abstract

The paper deals with teaching the theory and methodology of survey sampling at the faculty of Cybernetics of Taras Shevchenko National University of Kyiv. The short programme of the course on survey sampling is given. The problems that arose during teaching the course and the perspectives of development are mentioned.
Keywords: teaching of Survey Sampling

1 General Information about the Faculty of Cybernetics

Taras Shevchenko National University of Kyiv is a research university with a classic tradition and is the leading establishment of higher education in Ukraine. It was opened on July 15, 1834 as the Imperial University of Saint Volodymyr. Since its foundation in a structure of Philosophical Faculty there was a Department of Physics and Mathematics. Later, in 1849, on the basis of this division the faculty of the same name was organized. In 1940 on the basis of the Faculty of Physics and Mathematics of Kyiv University there appeared two new ones - the Faculty of Mechanics and Mathematics and the Faculty of Physics. An important event in the life of Kyiv University happened in 1969, when owning to the initiative V.M.Glushkov, I.J.Lyashko and I.T.Shvets' the Faculty of Cybernetics had been set up. The faculty was the first one with a computing profile, which merged the disciplines of mechanics and mathematics, economics and linguistics. The faculty is a place where a lot of remarkable scientists and fruitful researchers have worked and are working now at different aspect of mathematical and computer sciences.

At this moment the Faculty of Cybernetics consists of 9 departments:

- Department of Applied Statistics,
- Department of Systems Analysis and Decision Theory,
- Department of Computational Mathematics,
- Department of Modelling of Complex Systems,
- Department of Operations Research,
- Department of Theoretical Cybernetics,
- Department of Theory and Technology of Programming,
- Department of Mathematical Informatics,
- Department of Information Systems,

where 102 lecturers work now. There are 4 research laboratories at the faculty: Computational Methods in the Mechanics of Continuous Media, Modelling and Optimization, High Performance Data Processing

Systems, and Probability and Statistical Methods.

The students of the faculty of Cybernetics are major in such specialities:

- Systems Analysis,
- Applied Mathematics,
- Informatics,
- Social Informatics,
- Software Engineering.

Nowadays the faculty has about 900 students and 80 PhD students. The preparing of the specialists is based on the fundamental training in mathematics and computer science as well as in modern technical base.

2 The teaching of Survey Sampling

2.1 The Courses in Survey Sampling

The courses in Survey Sampling are teaching at the faculty of Cybernetics for the students of specialities "System Analysis" and "Social Informatics". The Department of Applied Statistics and Department of Systems Analysis and Decision Theory are qualified these specialities.

During different periods a lot of good mathematicians and statisticians have worked at the Department of Applied Statistics. Some of them are V. Donchenko and O. Chernyak.

In 1994 Vladimir Donchenko has specialized at survey statistics during 6 weeks at the University of Umeå (Sweden). At the present time V. Donchenko has a special course "The methods and models of data processing for social information" which is related to Survey Sampling. The course is intended for the students of the Department of Systems Analysis and Decision Theory of the fifth year of education.

Oleksander Chernyak is now a chef of Department of Economic Cybernetics at the Faculty of Economics at Kyiv University. In 1999 he visited the University of Umeå (Sweden) and specialized at survey sampling during 1 week. Two textbooks on survey sampling have been published in the Ukrainian language:

O. Chernyak (2001): Survey Sampling Technique, 248 pages;

O. Chernyak, A. Stavytskyy, H. Chornous (2006): The systems of data processing of economic information, 447 pages.

A special course "Survey Sampling Technique" at the Department of Economic Cybernetics is represented by O. Chernyak.

A course "The principles of Survey Sampling" was introduced at the Department of Applied Statistics in 2010 by I. Rozora. It was intended for the students of "System Analysis" and "Social Information" specialities in the fourth year of studies. This course consists of 36 hours of lectures.

The lectures on Survey Sampling are based mostly on the books:

- O.I. Vasylyk, T.O. Yakovenko (2010) Lectures on the theory and methods of Survey Sampling, Kyiv. (in Ukrainian);
- Lohr, S.L. (1999). Sampling: Design and Analysis. Duxbury Press, Pacific Grove;
- Särndal, C., Swensson, B. & Wretman, J. (2003). *Model assisted survey sampling.* Springer Verlag.

The experience of teaching this course at the faculty of Mechanics and Mathematics of Kyiv University was taking into account.

2.2 The programme of the course "The principles of Survey Sampling"

A short programme of the course plan is as follows:

- Introduction to sampling theory and methodology (goals and methods of survey, the main concepts and definitions of Survey Sampling theory)
- Simple random sampling with and without replacement (sample scheme, definitions, estimators of total and mean, proportion in the population and in domain, estimators of variance)
- Sampling with unequal probabilities(description of the techniques, estimators)
- Systematic Sampling
- Stratified sampling
- Single-stage and multistage cluster sampling
- Linear regression models
- Errors in Surveys, their sources and the methods of reduction

The main objective of this course is to acquaint students with the main concepts of Survey Sample Theory and the basic types of probability sampling.

The working language of the course is English. Short numeric problems that follows the lecture topic are solved without using a computer. They corroborate the theoretical results, comparing the accuracy the estimators. The sources of such problems are the textbook Vasylyk and Yakovenko (2010), Ardilly and Tillé, Y. (2006). Control works consist of short theoretical questions, definitions, proofs of the properties and the solutions of the numerical problems.

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

There were some students at the faculty of Cybernetics that wrote bachelor theses on survey sampling. The real data of statistics of Ukraine were used. This year two bachelor's theses are preparing.

3 Conclusion

The course “The principles of Survey Sampling” is very important and interesting for the students of the specialities “System Analyses” and “Social Information” at the faculty of Cybernetics of Kyiv University. It could be improved by increasing the number of hours for practical work. It could be also developed by creating advanced course and introducing new methods and modern techniques of Survey Sampling.

References

O.I. Vasylyk, T.O. Yakovenko (2010) *Lectures on the theory and methods of Survey Sampling.* Kyiv. (in Ukrainian).

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

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

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

V. Parkhomenko (2001): *Survey Sampling Methods.* Kyiv, 148 p. (in Ukrainian).

Ardilly, P. & Tillé, Y. (2006). *Sampling methods: exercises and solutions.* Springer Verlag.

The problems of consumer prices sampling in Belarus

Natallia Sakovich

Belarus State Economic University, e-mail: Sakovich-n@rambler.ru

Abstract

The article describes the main stages, characteristics and problems of sample surveys of consumer prices in the official statistics of Belarus. It is noted that in practice the calculation of the CPI are mainly non-probability methods such as the representative item method and cut-off sampling.

1 Introduction

For a consumer price index (CPI) national statistical agencies collect data on prices through a sample survey. In fact, in many countries, it might be better viewed as composed of many different surveys, each covering different subsets of the products covered by the index.

The general population usually has three dimensions: 1) product dimension, 2) geographical and outlet dimension, 3) time dimension.

In surveys of consumer prices can be used probability and non-probability sampling methods. Traditionally, however, non-probability sampling methods have mainly been used in the compilation of a CPI for choosing outlets or products. The representative item method is particularly popular for selecting items. Other methods used are cut-off sampling and quota sampling. In some cases, these two methods are used in combination, for example, outlets are selected using probability sampling techniques, whilst products are selected using the representative item method.

2 Non-probability sampling techniques

In the international standard on price statistics (Consumer price index manual: Theory and practice, 2007, p. 99) are the main reasons for using non-probability sampling:

1) *No sampling frame is available*. This is often true for the product dimension but less frequently so for the outlet dimension (as a sampling of which are usually the business registers or directories);

2) *Bias resulting from non-probability sampling is negligible,* especially for highly aggregated indexes, as evidenced in the works of Dalen (Dalén, 1998) and De Haan, Opperdusa and Jester (De Haan, Opperdoes and Schut, 1999);

3) *We need to ensure that samples can be monitored for some time.* If we are unlucky with our probability sample, we may end up with a product that disappears immediately after its inclusion in the sample. We are then faced with a replacement problem, with its own bias risks;

4) *A probability sample with respect to the base period is not a proper probability sample with respect to the current period.* The bias protection offered by probability sampling is to a large extent destroyed by the need for non-probabilistic replacements later on;

5) *Price collection must take place where there are price collectors.* This argument applies to geographical sampling only;

6) *The sample size is too small.* Stratification is sometimes done in as much detail as from the final strata can be made up only a very small sample having a low representativeness.

In practice, a survey of consumer prices using the following types non-probability techniques:

1) *Cut-off sampling* refers to the practice of choosing the n largest sampling units with certainty and giving the rest a zero chance of inclusion. In this context, the term "largeness" relates to some measure of size that is highly correlated with the target variable. The word "cut-off" refers to the borderline value between the included and the excluded units. The sample selected by all of the major units, and medium and small are selected in proportion to the value of a given parameter (eg, the value of production);

2) *The quota sampling* – in the resulting sample units should be presented in the same proportion as in the general population, in terms of number of known characteristics, such as a subset of products, type of outlet, and location. A limitation of quota sampling, as in other non-probability sampling, is that the standard error of the estimate cannot be determined;

3) *The representative item method* – it's the traditional CPI method. The central office draws up a list of product types, with product type specifications. These specifications may be tight, in that they narrowly prescribe for the price collectors what products they are permitted to select, or they may be loose, giving the price collector freedom to choose locally popular varieties.

The method with tight specifications may lead to less representative because the index will not include products that do not meet specifications. Another disadvantage with the method is that it may lead to more missing products in the outlets and thus reduce the effective sample. Its main advantage is simplicity.

The method with loose specifications gives price collectors the chance to adjust the sample to local conditions and will normally lead to greater representativity of the sample as a whole. However, here there is the problem of subjectivity in the replacement of the goods.

Many countries in the practice of the consumer price surveys are widely used methods of probability sampling. For example, in the United States and Sweden are used to modify *probability proportional to size (pps) sampling*. In France conducted *a two-stage random sample*, first of urban areas and then of a particular item (variety) in an outlet. The Luxembourg CPI can be described as *a stratified purposive sample*. In the United Kingdom and Finland are carried out experimental work on the preparation of the sample.

3 Consumer price surveys in Belarus

In 1992, National Statistical Committee of the Republic of Belarus, along with the rest of the CIS countries, have switched to a sample survey in the field of price statistics in order to adequately reflect the level of inflation. The methodology for monitoring consumer prices and CPI developed with the participation of experts from the International Monetary Fund and other international organizations (OECD, IMF, Eurostat) and broadly in line with international standards. The calculation of the CPI is based on two arrays of information: 1) the monthly data recording prices on a predetermined set of representative goods, and 2) an annual sample survey of households on the structure of consumption expenditure for the reference year.

Sample survey of consumer prices comprising the following steps: 1) selection of settlements, 2) the selection of trade organizations (or outlets), 3) the selection of representative goods (services), 4) the registration of prices (tariffs).

Table 1: The characteristics of a sample survey of consumer prices in Belarus

The characteristics	Number of survey units
1. Settlements	31
2. Trade organizations (outlets)	7000
3. Goods (services) representatives	450
4. Price quotations	50000

In *the selection of settlements* recorded their geographical representation and saturation of the consumer market with goods and services. The country surveyed 31 cities, where more than 50% of the population. The list of cities remained unchanged throughout the period of the survey. This fact contributes to the comparability of information, but reduces its representativeness. Rural communities are not involved in the observation due to the low supply of consumer goods, as well as by the lack of sufficiently trained (price collectors).

The selection of trade organizations based on the sampling method of observation. The sample population includes about 7000 organizations. For the selection of the basic statistical data used by organizations reporting on the supply of goods to the population. Basic organizations must be representative from different points of view: the forms of trade and forms of ownership, size, and location. Updating the sample of basic organizations are produced annually, and the possible replacement of the base organization in the event of liquidation or cessation of work for more than 6 months. Frequent replacement of basic organizations degrades the quality of the sample and reduces the comparability of the results of observation.

The selection of goods (services) representatives. The consumption bundle for calculating the CPI, is a representative sample of goods and services most frequently used by the public, and now includes about 450 names. In Belarus, are not included in the bundle of goods were in use, buying on credit, insurance services, but some countries allow for data items. Consumption bundle is generated using non-probability sampling – the representative item method with loose specifications. Of great importance are questions of renovation sampling due to changes in the structure of consumer demand, the emergence of new variants of goods and innovative products.

Registration prices and tariffs will be held from 10 to 30 the number of each month, and the need to adhere to deadlines registration prices (tariffs) in order to withstand the interval between two logs in one month.

Weighting for the CPI is based on a sample survey of households, as well as additional information about the retail trade, production and import of certain goods. Update the weights is recommended at least once every five years. In Belarus, as in most countries, updating the weights are produced annually, from January 1, and used the structure of period (t-2), that is the year preceding the previous year. Now for the price indices in 2013, weights in 2011, to the unstable economic situation and consumer behavior atypical of the population. In such cases, the author's opinion, should be used for a number of years, the average weight gain (eg, three years), which enable smooth out sudden changes in the structure of consumer spending.

4 Concluding remarks

In order to improve sample survey of consumer prices in Belarus should:

- Combine probability and non-probability sampling methods, expanding the use of probability sampling;

- To carry out geographical rotation of cities participating in the sample, if possible, include a large rural settlements;

- To expand the list of goods and services included in the consumer set, in particular, the products sold on credit, second-hand, financial, banking, insurance services and others;

- To use the average weights to eliminate the influence of random factors.

References

Consumer price index manual: Theory and practice (2007). Washington: International Monetary Fund.

Dalén, (1998). *Studies on the Comparability of Consumer Price Indices*, in International Statistical Review, Vol. 66, No. 1, pp. 83–113.

De Haan, E. Opperdoes, & C. Schut. (1997). *Item Sampling in the Consumer Price Index: A Case Study using Scanner Data*, Research Report (Voorburg: Statistics Netherlands).

List of participants

Name	Surname	Country	Organization	e-mail
Agnes	Andics	Hungary	Hungarian Central Statistical Office	agnes.andics@ksh.hu
Julia	Aru	Estonia	University of Tartu, Statistics Estonia	julia.aru@stat.ee
Anastacia	Bobrova	Belarus	The Institute of Economics of NAS of Belarus	nastassiabobrova@mail.ru
Natallia	Bondarenko	Belarus	Belarus State University	bondnata@mail.ru
Iana	Bondarenko	Ukraine	Dnepropetrovsk National University	iana.s.bondarenko@gmail.com
Natallia	Bokun	Belarus	Belarus State Economic University	nataliabokun@rambler.ru
Natalja	Budkina	Latvia	Riga Technical University	natalja.budkina@rtu.lv
Andrius	Ciginas	Lithuania	Vilnius University, Institute of Mathematics and Informatics	andrius.ciginas@mif.vu.lt
Katsiaryna	Chystsenka	Belarus	National Bank of the Republic of Belarus	katsiaryna.chystsenka@gmail.com
Andris	Fisenko	Latvia	Central Statistical Bureau of Latvia	andris.fisenko@csb.gov.lv
Aleksandra	Galahina	Latvia	SIA TNS Latvia	aleksandra.galahina@tns.lv
Oksana	Honchar	Ukraine	National Academy of Statistics, Accounting and Audit	ohonchar@list.ru
Tetiana	Hudyvok	Ukraine	Uzhgorod National University	fedoryanicht@gmail.com
Tetiana	Ianevich	Ukraine	Taras Shevchenko National University of Kyiv	yakovenkot@gmail.com
Danute	Krapavickaite	Lithuania	Vilnius Gediminas Technical University	danute.krapaviciate@vgtu.lt
Gunnar	Kulldorff	Sweden	University of Umea	gunnar@matstat.umu.se
Seppo	Laaksonen	Finland	University of Helsinki	Seppo.Laaksonen@Helsinki.Fi
Janis	Lapins	Latvia	Bank of Latvia	Janis.Lapins@bank.lv
Anna	Larchenko	Belarus	Belarus State Economic University	annalarchenko@gmail.com
Risto	Lehtonen	Finland	University of Helsinki	risto.lehtonen@helsinki.fi
Natalia	Lepik	Estonia	University of Tartu	natalja.lepik@ut.ee
Martins	Liberts	Latvia	Central Statistical Bureau of Latvia	martins.liberts@csb.gov.lv
Kaur	Lumiste	Estonia	University of Tartu	Kaur.lumiste@ut.ee
Olha	Lysa	Ukraine	Ptukha Institute for Demography and Social Studies	OLysa@ukr.net
Saara	Oinonen	Finland	Statistics Finland	saara.oinonen@stat.fi
Julia	Orlova	Belarus	Delta Bank	orlova-julia-gen@mail.ru
Pauliina	Peltonen	Finland	Statistics Finland / University of Helsinki	pauliina.peltonen@helsinki.fi

Name	Surname	Country	Organization	e-mail
Aleksandras	Plikusas	Lithuania	Vilnius University	aleksandras.plikusas@mii.vu.lt
Iryna	Rozora	Ukraine	Taras Shevchenko National University of Kyiv	irozora@bigmir.net
Tomas	Rudys	Lithuania	Vilnius University, Institute of mathematics and informatics	tomas.rudys@mii.vu.lt
Natallia	Sakovich	Belarus	Belarus State Economic University	sakovich-n@rambler.ru
Gennady	Shurko	Ukraine	Donetsk National University	shurko.g.k@gmail.com
Svitlana	Slobodian	Ukraine	Vasyl Stefanyk Precarpathian National University	slobodian_s@ukr.net
Imbi	Traat	Estonia	University of Tartu	imbi.traat@ut.ee