Lietuvos statistika
Statistics Lithuania

SUMMER SCHOOL ON SURVEY STATISTICS 2021

BNU Network on Survey Statistics

Virtual Sessions in English: Friday 3, 10, 17 and 24 September 2021
Virtual Sessions in Russian: Saturday 4, 11, 18 and 25 September 2021

Statistics
Lithuania

ISBN 978-9955-797-34-0

SUMMER SCHOOL ON SURVEY STATISTICS 2021

BNU Network on Survey Statistics

Virtual Sessions in English: Friday 3, 10, 17 and 24 September 2021
Virtual Sessions in Russian: Saturday 4, 11, 18 and 25 September 2021

Summer School on Survey Statistics 2021. Virtual Sessions in English: Friday 3, 10, 17 and 24 September 2021. Virtual Sessions in Russian: Saturday 4, 11, 18 and 25 September 2021. Statistics Lithuania, Vilnius, 2021, 111 p.

Compiler: Dalius Pumputis.

The publication is designed to provide a methodical tool for the summer school participants

ISBN 978-9955-797-34-0

Preface

Dear Participants of the virtual Summer School on Survey Statistics!

The Summer School on Survey Statistics 2021 is the 25st event organised by the Baltic–Nordic–Ukrainian (BNU) Network on Survey Statistics in a sequence of annual workshops, conferences and summer schools. Originally, the network's steering committee planned the summer school to be an on-site event in 2020 in Minsk, Belarus. However, the ongoing global pandemic made us move to a virtual event while still maintaining the bilingual feature of the original summer school. Four Fridays in September were devoted to educational and scientific sessions in English and four Saturdays to sessions in Russian. The virtual sessions were organized through Zoom hosted by the University of Helsinki.

The summer school aims to promote scientific and educational cooperation in survey and official statistics between statisticians interested in new trends in the area. The Programme Committee selected Data integration, Machine Learning and Small area estimation as the main topics for the event. Three widely recognized keynote speakers were invited: Shu Yang of North Carolina State University, USA, Piet Daas of Eindhoven University of Technology & Statistics Netherlands and Marcin Szymkowiak of Poznan University of Economics and Business & Statistical Office in Poznan, Poland. Four experts from Ukraine were invited to give lectures on survey statistics in Russian: Tetiana Ianevych and Iryna Rozora, both of Taras Shevchenko National University of Kyiv, Tetiana Manzhos of Kyiv National Economic University and Olga Vasylyk of National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute". The program also includes seven invited speakers from the partner organizations of the network and a selection of contributed papers on topics in modern survey statistics. Abstracts of the 35 presentations are collected in this Proceedings.

We are thankful to the International Association of Survey Statisticians (IASS) for sponsoring this event. Support of the University of Helsinki and the other partner organizations is appreciated. Thanks are due to Dalius Pumputis for collecting and compiling the materials in the Proceedings, and Mārtiņš Liberts who took care of the Slack operations. Special thanks are due to Anastasiia Volkova and Amanda Häkkinen who contributed to the organization of the event. Members of the Steering Committee of the BNU network deserve credit for their hard work.

We dedicate the summer school to the memory of Professor Seppo Laaksonen who passed away in December 2020. His experience and expertise in survey methodology was enjoyed at numerous educational and scientific events of the network. Seppo's warm personality and colorful lectures are unforgettable.

This is the final edition of the proceedings publication of the summer school. In the first edition one month ago, we wished fruitful sessions, new knowledge and useful contacts for everyone. We hope that the event successfully met the expectations of more than 200 people from several countries who attended the summer school.

On behalf of the organizers,

Risto Lehtonen, University of Helsinki

Maria Valaste, University of Helsinki

Contents

Contributed papers

Summer School on Survey Statistics 2021

Virtual sessions in English

Scientific Programme

Friday 3 September	15:00-16:00	**Session 1**	
	15.00-15.15	Opening: Risto Lehtonen and Maria Valaste	
	15.15-16.15	**Keynote Lecture**	
		Shu Yang: Data integration: a new paradigm for survey statistics	
		Chair: Risto Lehtonen	
	16:15-16:45	**Session 2**	
		Invited lecture	
	16.15-16.45	Kaja Sõstra: SAE Methods for Developing the Digital Economy and Society Index (DESI) at local level	
		Chair: Imbi Traat	
		*** BREAK***	
	17:20-18:00	**Session 3**	
		Contributed papers (no parallel sessions)	
	17.20-17.35	Vilma Nekrašaitė-Liegė: A Comparison of URL finders for online-based enterprise characteristics	
	17.40-17.55	Baiba Zukula: CAWImobile for household surveys	
		Chair: Olga Vasylyk	
Friday 10 September	15:00-16:00	**Session 4**	
		Keynote Lecture	
		Piet Daas: Identifying different types of companies via their website text	
		Chair: Maciej Beręsewicz	
	16:00-17:00	**Session 5**	
		Invited lectures	
	16.00-16.30	Signe Bāliņa: What is behind statistics?	
	16.30-17.00	Mykola Sydorov: UniDOS online with LimeSurvey	
		Chair: Mārtiņš Liberts	
		*** 5 min BREAK***	
	17:05-18:05	**Session 6**	
		Contributed papers, parallel session 1	**Contributed papers, parallel session 2**
	17.05-17.20	Milda Šličkutė-Šeštokienė: Register based census in Lithuania	Anastasiia Volkova: On the importance of conceptualization and operationalization in survey design: lessons from the Morally Debatable Behaviors scale
	17.25-17.40	Ance Cerina and Zane Matveja: Statistical Disclosure Control for Census 2021	Natalia Bokun: Enterprises Survey on Personnel Demand

	17.45-18.05	Ruāna Pavasare: Statistical Editing and Imputation of Missing values for the Population Census 2021 in Latvia	Chair: Vilma Nekrašaitė-Liegė
		Chair: Andrius Čiginas	
Friday 17 September	15:00-16:00	**Session 7**	
		Keynote Lecture	
		Marcin Szymkowiak: Small area estimation in official statistics - past, present and future directions of applications	
		Chair: Danutė Krapavickaitė	
	16:00-17:00	**Session 8**	
		Invited lectures	
	16.00-16.30	Blaise Ngendangenzwa and Joel Tolsheden: Machine Learning and Automatic Editing	
	16.30-17.00	Krista Lagus: Open-ended questions in surveys: Exploring the possibilities of NLP and data science	
		Chair: Thomas Laitila	
		*** 5 min BREAK***	
	17:05-18:05	**Session 9**	
		Contributed papers, parallel session 1	**Contributed papers, parallel session 2**
	17.05-17.20	Jelena Voronova: Observing nonresponse bias optimising data collection strategy for adaptive sample survey design	Danutė Krapavickaitė: Highlights of the WSC 2021 in Survey Statistics
	17.25-17.40	Liliāna Roze: Optimal identification of auxiliary variables in sample surveys to reduce nonresponse bias	Ulrich Rendtel, Andreas Neudecker and Lukas Fuchs: The display of Corona incidences in space and time
	17.45-18.05	Darja Goreva and Viktors Veretjanovs: Analysis of EU-SILC data depending on modes of data collection in Latvia	Maria Valaste and Hanna Wass: Data Collection Mode and Nonresponse: Practical Experiences
		Chair: Tetiana Ianevych	Chair: Milda Šličkutė-Šeštokienė
Friday 24 September	15:00-16.00	**Session 10**	
		Invited lectures	
		Sylwia Filas-Przybył and Tomasz Klimanek: Income stratification of the urban population in Poland	
		Tomas Rudys: The use of alternative data sources at Statistics Lithuania	
		Chair: Jānis Lapiņš	
	16.00-16.10	*** Group photo ***	
	16:10-17:30	**Session 11**	
		Contributed papers (no parallel sessions)	
	16.10-16.25	Mārtiņš Liberts: Unequal Probability Sampling for the European Interview Health Survey in Latvia	
	16.30-16.45	Ieva Burakauskaitė and Vilma Nekrašaitė-Liegė: Selective Editing Using Contamination Model	

	16.50-17.05	Yana Bondarenko: A Sequential Probability Ratio Test for Online Experiments	
	17.10-17.25	Andrius Čiginas: On design mean square error estimation for model-based small area estimators	
		Chair: Maria Valaste	
	17:30-18:00	**Session 12**	
		Closing of the Summer School and Farewell Party	
		Chair: Kaur Lumiste	

Summer School on Survey Statistics 2021

Virtual sessions in Russian

Scientific Programme

Суббота, 04.09.	11:00-12:30	**Занятие 1**
Saturday 4 September		**Session 1**
Chair: Natalia Bokun		Открытие конференции
		Opening
		Лекция «Выборочные обследования: основные методы оценивания»
		(Татьяна Яневич)
		Invited lecture “Sample Surveys: Main Estimation Methods” (Tetiana Ianevych)
	13:00-14:30	**Занятие 2**
		Session 2
		Татьяна Яневич продолжает
		Invited lecturer continues (Tetiana Ianevych)
Суббота, 11.09.	11:00-12:30	**Занятие 3**
Saturday 11 September		**Session 3**
Chair: Natalia Bokun		Лекция «Оценивание доменов и малые области оценивания» (Ольга Василик)
		Invited lecture “Estimation for Domains and Small Areas” (Olga Vasylyk)
	13:00-14:30	**Занятие 4**
		Session 4
		Ольга Василик продолжает
		Invited lecturer continues (Olga Vasylyk)
Суббота, 18.09.	11:00-12:30	**Занятие 5**
Saturday 18 September		**Session 5**
Chair: Natalia Bondarenko		Лекция
		«Использование калибрации с целью сокращения смещения в результате неответов» (Ирина Розора)
		Invited lecture “Calibration Estimation for Nonresponse Bias Reduction” (Irina Rozora)
	13:00-14:30	**Занятие 6**
		Session 6
		Выступления
		Contributed papers
		Natalia Bokun: Labor Market Surveys in Belarus (Выборочные обследования на рынке труда в Беларуси)
		Sharilova Eugenia: Sample surveys in assessing the main determinants of fertility decline in the Republic of Belarus / Выборочные обследования в оценке основных детерминантов снижения рождаемости в Республике Беларусь
		Nataliya Pekarskaya: Household survey to measure the agricultural activity

Суббота, 25.09.	11:00-12:30	**Занятие 7**
Saturday 24 September		**Session 7**
Chair: Natalia Bondarenko		«Big data и выборочные обследования: проблемы использования» (Татьяна Манжос)
		Invited lecture “Big Data and Sample Surveys: Problems of Use” (Tetiana Manzhos)
	13:00-14:30	**Занятие 8**
		Session 8
		Выступления
		Contributed papers
		Sakovich N.: Consumer Prices Sample Surveys in Belarus (Выборочные обследования потребительских цен в Беларуси)
		Korolenok A.: Problems of survey of unpaid activities
		Liudmila Soshnikava: Using logistic regression to analyze the results of statistical observations

IDENTIFYING DIFFERENT TYPES OF COMPANIES VIA THEIR WEBSITE TEXT

Piet Daas

Eindhoven University of Technology, Statistics Netherlands, Netherlands
e-mail: pjh.daas@cbs.nl

Abstract

The internet and especially web pages are a very interesting source of data. It has very interesting potential applications such as providing novel insights on the activities of companies, to inform policy makers and also for official statistics, especially when performed at large scale. However, extracting relevant and reliable information from big data sources in a reproducible way is not an easy task. In this presentation results of Machine Learning based classifications of web sites texts are discussed in relation to the identification of innovative, platform economy and AI companies.

Keywords: Big Data, website text, machine learning, classification, bias correction.

References

Daas, P.J.H., van der Doef, S. (2020) Detecting Innovative Companies via their Website. Statistical Journal of IAOS 36(4), pp. 1239-1251, doi/10.3233/SJI-200627.

Puts, M.J.H., Daas, P.J.H.. (2021) Unbiased Estimations Based on Binary Classifiers: A Maximum Likelihood Approach. Abstract for the 2021 Symposium on Data Science and Statistics, Machine Learning session. Archive link: https://arxiv.org/abs/2102.08659

Puts, M., Daas, P. (2021) Machine Learning from the Perspective of Official Statistics. The Survey Statistician 84. pp. 12-17.

SMALL AREA ESTIMATION IN OFFICIAL STATISTICS – PAST, PRESENT AND FUTURE DIRECTIONS OF APPLICATIONS

Marcin Szymkowiak

Poznan University of Economics and Business, Statistical Office in Poznan, Poland
e-mail: marcin.szymkowiak@ue.poznan.pl

Abstract

Small area estimation methodology (SAE) has been developed to produce reliable estimates of different characteristics of interest, such as means, counts, quantiles or ratios for domains for which only small samples are available. From that point of view SAE has become a topic of great importance due to the growing demand for reliable small area statistics. SAE methodology is used by different national statistical institutes in different areas, in particular to estimate quantities that are related to the labour market, agriculture or business statistics. It is also useful for mapping poverty. For instance, the World Bank has used SAE methodology to prepare poverty maps for tens countries all over the world. The main purpose of this presentation is to provide a review of the main applications in SAE methods, mainly in official statistics. Presentation will be based on earlier and present applications which serve as a necessary background for the new directions of the SAE development, including using big data sources as an example.

Keywords: small area estimation, calibration, poverty mapping, disability, big data.

References

Marchetti, S., Beręsewicz, M., Salvati, N., Szymkowiak, M., & Wawrowski, Ł. (2018), The use of a three-level M-quantile model to map poverty at local administrative unit 1 in Poland, Journal of the Royal Statistical Society: Series A, (Statistics in Society), Vol. 181, No. 4, 1-28.

Szymkowiak, M., Młodak, A., & Wawrowski, Ł. (2017), Mapping Poverty at the Level of Subregions in Poland Using Indirect Estimation, Statistics in Transition – New Series, 18 (4), 609-635.

Rao, J. N., & Molina, I. (2015). Small area estimation. John Wiley & Sons.

DATA INTEGRATION: A NEW PARADIGM FOR SURVEY STATISTICS

Shu Yang

North Carolina State University, USA
e-mail: syang24@ncsu.edu

Abstract

Finite population inference is a central goal in survey sampling. Probability sampling is the gold standard statistical approach to finite population inference. Challenges arise due to high costs and increasing non-response rates. Data integration provides a timely solution by leveraging multiple data sources to provide more robust and efficient inference than using any single data source alone. The technique for data integration varies depending on the types of samples and available information to be combined. This talk provides a systematic review of data integration techniques for combining probability and non-probability samples and for combining probability and big data samples. A wide range of integration methods will be covered such as calibration weighting, inverse probability weighting, mass imputation, and doubly robust methods. Finally, I will highlight important questions for future research.

Keywords: Calibration weighting; data integration and fusion; double robustness; mass imputation; variable selection.

References

S. Chen, S. Yang, and J.K. Kim (2020). Nonparametric mass imputation for data integration. *Journal of Survey Statistics and Methodology*, doi.org/10.1093/jssam/smaa036.

S. Yang, J. K. Kim, and R. Song (2020). Doubly robust inference when combining probability and non-probability samples with high-dimensional data, *Journal of the Royal Statistical Society*: Series B, 82, 445–465.

S. Yang, J. K. Kim, and Youngdeok Hwang (2021). Integration of data from probability surveys and big found data for finite population inference using mass imputation. *Survey Methodology*, 47, 29–58.

SAMPLE SURVEYS: MAIN ESTIMATION METHODS

Tetiana Ianevych

Taras Shevchenko National University of Kyiv, Ukraine
e-mail: tetianayanevych@knu.ua

Abstract

The lecture will be devoted to some basic estimators widely applied in Survey Sampling. I start from Horvitz-Thompson estimator and its forms under different sampling designs. Then I explain which parameters are nonlinear and how they can be estimated using as an example *ratio* parameter. Some regression estimators will be also discussed, in particular, those based on models of common ratio and simple liner regression.

I shall illustrate how and when we can use this estimators for analyzing the results from sample survey using StatVillage. This is a hypothetical village in Canada. It is free online tool developed by Schwarz (1997) and based on real data taken from the 1991 census of Canada.

The data from Statvillage can be downloaded as a txt-file and we will try to process the data using free statistical software `R`. It is desirable to install `R` in advance. This can be done from https://www.r-project.org/.

Since the timing for the lecture is limited I shall have possibility to say only few words on different estimation techniques. So, I would recommend some classical books on Survey Sampling for deeper learning for those who become interesting: Cochran (1977) (translated into Russian), Lohr (1999), Säarndal et al. (1992), Vasylyk and Yakovenko (2010) (in Ukrainian).

Keywords: Horvitz-Thompson estimator, estimation of ratio, regression estimators.

References

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

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

Säarndal, C.E., Swensson, B., Wretman, J. (1992) *Model Assisted Survey Sampling*.SpringerVerlag, New York.

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

Василик О. І., Яковенко Т.О. (2010) *Лекції з теорії і методів вибіркових обстежень*. Видавництво Київський університет, Київ. (In Ukrainian).
https://probability.knu.ua/userfiles/yakovenko/Vasylyk_Yakovenko_book_rev.pdf

ВЫБОРОЧНЫЕ ОБСЛЕДОВАНИЯ: ОСНОВНЫЕ МЕТОДЫ ОЦЕНИВАНИЯ

Татьяна Яневич

Киевский национальный университет имени Тараса Шевченко, Украина
e-mail: tetianayanevych@knu.ua

Аннотация

Лекция будет посвящена некоторым базовым оценочным методам, которые широко используются в теории выборочных обследований. Начну я из рассмотрения оценки Горвица-Томпсона и ее форм при разных методах отбора. Потом я объясню какие параметры не есть линейными и как их можно оценивать на примере параметра *отношение*. Мы также рассмотрим регрессионные оценки, в частности, те, которые строятся на модели общего отношения и простой линейной регрессии.

Я проиллюстрирую как и когда можно применять эти оценки для анализа результатов выборочных обследований используя Статвиладж. Это искусственное поселение в Канаде. Этот онлайн-инструмент специально был создан для учебных целей Шварцом (1997) и для него были использованы реальные данные из переписи населения Канады 1991 года.

Данные полученные в Статвилидже можно загрузить в виде txt-файла. Я попробую показать, как их можно обработать используя бесплатное статистическое программное обеспечение `R`. Желательно установить `R` загодя. Это можно сделать следуя по ссылке https://www.r-project.org/.

Поскольку время лекции ограничено, я не буду иметь возможности детально рассказать обо всех методах оценивания. По этому я порекомендую некоторые классические книги по теории выборочных обследований для углубленного изучения для тех, кого это заинтересовало: Кокран (1977) (переведена на русский язык), Лор (1999), Сарндал и др. (1992), Василик и Яковенко (2010) (на украинском языке).

Ключевые слова: Оценка Горвица-Томпсона, оценивание отношения, регрессионные оценки.

Литература

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

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

Säarndal, C.E., Swensson, B., Wretman, J. (1992) *Model Assisted Survey Sampling*.SpringerVerlag, New York.

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

Василик О. І., Яковенко Т.О. (2010) *Лекції з теорії і методів вибіркових обстежень*. Видавництво Київський університет, Київ. (In Ukrainian).
https://probability.knu.ua/userfiles/yakovenko/Vasylyk_Yakovenko_book_rev.pdf

BIG DATA AND SAMPLE SURVEYS: PROBLEMS OF USE

Tetiana Manzhos

Vadym Hetman Kyiv National Economic University, Ukraine
e-mail: tmanzhos@gmail.com

Abstract

Machine learning (ML) methods are gaining popularity across various scientific, technical and business spheres, and survey research is no exception. Different types of ML models (such as LASSO, SVM, CART, Random Forest etc.) have been used for responsive/adaptive designs, data processing, nonresponse adjustment and weighting. More detailed review of various ML methods and their uses for survey management has been discussed by Buskirk et al. (2018) and Kern et al. (2019).

Advantages of ML approaches are ability to automatically investigate the data, to find dependency between target variable and predictors without prior knowledge about functional form, to detect nonlinearities and to identify interactions automatically, to adapt to new data. The main disadvantage is that such models are not always interpretable. Due to this fact, for tasks which require result descriptions in terms of user understanding and making decisions based on them, it is needed to use interpretable models (i.e. Decision Tree or LASSO) or apply special methods for explanation the outcomes. To dig deeper into the topic, see Molnar (2019).

There are two types of ML techniques: supervised and unsupervised learning. The difference between them is that in supervised learning it is needed to provide the correct results in terms of labeled data. In supervised learning, model needs to find the function to map the input variables with the output variable (labels). Supervised learning can be used for two types of problems: classification (binary/discrete label) and regression (continuous label). In unsupervised machine learning, the data are not labeled. The goal of unsupervised learning is to find the structure from the input data (i.e. clustering task). In a survey research, the task of predicting unit nonresponse is an example of supervised classification task. Predictors for the model can be both respondent-related and interview-related variables, target variable is a binary label (indicator of nonresponse). Tree-based classification model in this case can be used for constructing nonresponse weights (for detailed overview, see Phipps & Toth, 2012).

Keywords: predictive models, machine learning, unit nonresponse, tree-based models

References

Buskirk, T. D., Kircher, A., Eck, A., Signorino, C.S. (2018) An introduction to machine learning methods for survey researchers. *Survey Practice*, **11**(1), 1-10.

Kern, C., Klausch, T., Kreuter, F. (2019) Tree-based machine learning methods for survey research. *Survey Research Methods*, **13**(1), 73-93.

Molnar, C. (2019) *Interpretable machine learning. A Guide for Making Black Box Models Explainable.* https://christophm.github.io/interpretable-ml-book/

Phipps, P., Toth, D. (2012) Analyzing establishment nonresponse using interpretable regression tree model with linked administrative data. *The Annals of Applied Statistics*, **6**(2), 772-794.

BIG DATA И ВЫБОРОЧНЫЕ ОБСЛЕДОВАНИЯ: ПРОБЛЕМЫ ИСПОЛЬЗОВАНИЯ

Тетяна Манжос

Київський національний економічний університет імені Вадима Гетьмана, Украина
e-mail: tmanzhos@gmail.com

Аннотация

Методы машинного обучения становятся все более популярными в научной, технической и бизнес сферах и выборочные исследования здесь не исключение. Разные модели машинного обучения, такие как LASSO, SVM, CART, Random Forest и др., используются для адаптивного дизайна, предсказания неответов, построения весов. Более детально рассматривается вопрос использования машинного обучения в сфере выборочных исследований в статьях Бускирк и др. (2018) и Керн и др. (2019).

Преимуществами подходов машинного обучения являются построение зависимости между целевой (независимой) переменной и предикторами без предположения о виде функции, возможность исследования структуры данных, обнаружение нелинейностей, адаптация к новым данным. Одним из недостатков является в большинстве случаев невозможность интерпретировать результаты. В случае, когда интерпретация важна в смысле понимания исследователем результатов моделирования и принятия решений на их основании, используются интерпретируемые модели (деревья решений или LASSO) или специальные методы для интерпретации. Детальное описание таких методов рассмотрено в книге Молнара (2019).

Существует два способа машинного обучения: обучение с учителем и без учителя. Обучение с учителем требует размеченных данных, т.е. в этом случае имея зависимую переменную и набор независимых переменных строится функция-соответствие. В свою очередь, существует два типа задач обучения с учителем: задачи классификации (бинарная или дискретная целевая переменная) и регрессии (непрерывная целевая переменная). Обучение без учителя не использует метки, основной целью здесь является нахождение структуры и паттернов в данных (например, в задачах кластеризации). В случае выборочных исследований, примером задачи машинного обучения с учителем (а именно, задачи классификации) является предсказание неответов. Здесь целевой переменной выступает индикатор неответа (есть неответ или нет), множество предикторов может включать как общие сведения о респондентах, так и связанные с интервьюированием. Модели, построенные на деревьях решений в данном случае, могут использоваться в дальнейшем при построении весов (см. Фиппс и др. (2012)).

Ключевые слова: предиктивные модели, машинное обучение, неответы, деревья решений

References

Buskirk, T. D., Kircher, A., Eck, A., Signorino, C.S. (2018) An introduction to machine learning methods for survey researchers. *Survey Practice*, **11**(1), 1-10.

Kern, C., Klausch, T., Kreuter, F. (2019) Tree-based machine learning methods for survey research. *Survey Research Methods*, **13**(1), 73-93.

Molnar, C. (2019) *Interpretable machine learning. A Guide for Making Black Box Models Explainable.* https://christophm.github.io/interpretable-ml-book/

Phipps, P., Toth, D. (2012) Analyzing establishment nonresponse using interpretable regression tree model with linked administrative data. *The Annals of Applied Statistics*, **6**(2), 772-794.

CALIBRATION ESTIMATION FOR NONRESPONSE BIAS REDUCTION

Iryna Rozora

Taras Shevchenko National University of Kyiv, Ukraine
e-mail: rozora.iryna@gmail.com

Abstract

Calibration is a hot topic in many recent articles on estimation in survey sampling. Calibration provides a systematic way to incorporate auxiliary information to estimate finite population parameters.

During my lecture, I will give the basic concepts of calibration approach considering a point estimator of total and mean based on calibration and a corresponding variance estimator. They are general in regard to both the sampling design and the form of the auxiliary information.

We also should pay attention to different calibrated robust estimators of such population parameter as mean with comparing them on the example.

Both the users and scientists of statistics know that nonresponse can greatly reduce the quality of the estimates. Weighting is widely applied in surveys to adjust for nonresponse and correct other nonsampling errors. We will discuss calibration estimation in the presence of nonresponse with a focus on the linear calibration estimator.

Keywords: Calibration, weighting, nonresponse, mean.

References

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

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

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

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

Särndal, C.-E. and Lundström, S. (2005). *Estimation in Surveys with Nonresponse.* Wiley, New York.

Bookwriter A. B., Author, A. (1999) Another important paper. In: *Proceedings of the 4th Conference in Statistical Sciences*, (ed. E. Theeditor), Publisher, City, 159–167.

ИСПОЛЬЗОВАНИЕ КАЛИБРАЦИИ С ЦЕЛЬЮ СОКРАЩЕНИЯ СМЕЩЕНИЯ В РЕЗУЛЬТАТЕ НЕОТВЕТОВ

Ірина Розора

Киевский национальный университет имени Тараса Шевченко, Украина
e-mail: rozora.iryna@gmail.com

Аннотация

Калибровка в последние года является очень востребованной темой в научно-исследовательских статьях, посвященных выбоочным наблюдениям. При калибровки исспользуется систематический подход для включения вспомогательной информации для построения оценки параметров совокупности.

Во время лекции я дам основные понятия о методе калибровки, рассматривая точечные оценки суммы и среднего и соответствующие оценки дисперсий.

Мы также обратим наше внимание на откалибованные разные робастные оценки такого показателя как среднее и сравним их на примере.

И статистики, и ученные знают, что отсутствие ответа в выборочных наблюдениях может значительно снизить качество оценок. Взвешивание широко применяется для корректировки отсутствия ответов и для исправления других ошибок. Мы рассмотрим оценку при отсутствии ответов (nonresponse), используя линейную связь в калибровке.

Ключевые слова: Калибровка, взвешивание, отсутствие ответа, среднее.

Литература

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

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

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

Harms, T., and Duchesne, P. (2006). On calibration estimation for quantiles. *Survey Methodology,* **32**, 37-52.
Särndal, C.-E. and Lundström, S. (2005). *Estimation in Surveys with Nonresponse.* Wiley, New York.

Bookwriter A. B., Author, A. (1999) Another important paper. In: *Proceedings of the 4th Conference in Statistical Sciences*, (ed. E. Theeditor), Publisher, City, 159–167.

ESTIMATION FOR DOMAINS AND SMALL AREAS

Olga Vasylyk

National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Ukraine
e-mail: vasylyk@matan.kpi.ua

Abstract

Estimation for domains is concerned with the development of statistical procedures for producing efficient estimates for population parameters, such as totals, means, proportions, medians, quantiles, percentiles for population subgroups called domains. Domains are defined by the cross-classification of geographical districts by social, economic, demographic characteristics. Small area estimation (SAE) means estimation for domains whose sample size is small or very small (even zero).

There are several classifications of SAE methods, in particular, direct and indirect methods, design-based methods and model-based methods, area-level models and unit-level models. Direct methods use only domain-specific data. Indirect methods borrow information from all the data. Design-based methods often use a model for the construction of the estimators (model assisted), but the bias, variance and other properties of the estimators are evaluated under the randomization (design-based) distribution. Model-based methods usually condition on the selected sample, and the inference is with respect to the underlying model. Area-level models relate small area direct estimators to area specific covariates. Such models are necessary if unit (or element) level data are not available. Unit-level models relate the unit values of a study variable to unit-specific covariates.

In the lecture, we shall consider classical and new approaches to estimation for domains and small areas. Also a brief overview of SAE software will be presented and some examples of computing the estimates will be given.

Keywords: design-based methods, domains, model-based methods, small area estimation.

References

Ghosh, M., Rao, J.N.K. (1994). Small area estimation: an appraisal. *Statistical Science*, **9**, 55-93.

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

Lehtonen, R., Veijanen, A. (2009). *Design-based methods of estimation for domains and small areas.* In Sample Surveys: Inference and Analysis (D. Pfeffermann and C. R. Rao, eds.). Handbook of Statistics, **29B**, 219-249. North-Holland, Amsterdam.

Lehtonen, R., Veijanen, A. (2020). Hybrid calibration methods for small domain estimation. *Statistics and Applications.* **XVII**. 201-235.

Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science*, **28**, 40-68.

Pratesi, M., Giusti, C. (2015) Small area estimation I. *1st EMOS Spring School*, Trier, Pisa, Manchester, Luxembourg, 23-27 March 2015.

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

ОЦЕНИВАНИЕ ДОМЕНОВ И МАЛЫЕ ОБЛАСТИ ОЦЕНИВАНИЯ

Ольга Василик

Национальный технический университет Украины "Киевский политехнический институт имени Игоря Сикорского", Украина

e-mail: vasylyk@matan.kpi.ua

Аннотация

Оценивание для доменов связано с разработкой статистических процедур для получения эффективных оценок параметров, таких как суммарные значения, средние значения, пропорции, медианы, квантили, процентили для подгрупп популяции (генеральной совокупности), называемых доменами. Домены определяются перекрестной классификацией географических районов по социальным, экономическим и демографическим характеристикам. Оценивание для малых областей (SAE) означает оценивание для доменов, размер выборки из которых мал или очень мал (даже может быть равен нулю).

Существует несколько классификаций методов оценивания для малых областей, в частности, прямые и косвенные методы оценивания; методы, основанные на дизайне выборки и методы, основанные на моделях; модели уровня области и модели уровня единицы. Прямые методы используют только данные, относящиеся к домену. Косвенные методы заимствуют информацию из всех данных. Методы, основанные на дизайне, часто используют модель для построения оценок, но смещение, дисперсия и другие свойства оценок оцениваются в рамках рандомизационного (основанного на дизайне) распределения. Методы, основанные на моделях, обычно зависят от полученной выборки, а выводы делаются относительно базовой модели. Модели уровня области связывают прямые оценки для малых областей с ковариатами, характерными для этих областей. Такие модели необходимы, если данные на уровне единицы (элемента) недоступны. Модели уровня единицы связывают значения исследуемой переменной для элемента с ковариатами, специфичными для элемента.

В лекции мы рассмотрим классические и новые подходы к оцениванию для доменов и малых областей. Также будет представлен краткий обзор программного обеспечения для SAE и приведены некоторые примеры расчета оценок.

Ключевые слова: домены; методы, основанные на дизайне выборки; методы, основанные на моделях; оценивание для малых областей.

Литература

Ghosh, M., Rao, J.N.K. (1994). Small area estimation: an appraisal. *Statistical Science*, **9**, 55-93.

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

Lehtonen, R., Veijanen, A. (2009). *Design-based methods of estimation for domains and small areas.* In Sample Surveys: Inference and Analysis (D. Pfeffermann and C. R. Rao, eds.). Handbook of Statistics, **29B**, 219-249. North-Holland, Amsterdam.

Lehtonen, R., Veijanen, A. (2020). Hybrid calibration methods for small domain estimation. *Statistics and Applications.* **XVII**. 201-235.

Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science*, **28**, 40-68.

Pratesi, M., Giusti, C. (2015) Small area estimation I. *1st EMOS Spring School*, Trier, Pisa, Manchester, Luxembourg, 23-27 March 2015.

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

INCOME STRATIFICATION OF THE URBAN POPULATION IN POLAND

Sylwia Filas-Przyby[1,2] **and Tomasz Klimanek**[1,3]

[1] Statistical Office in Pozna, Poland
e-mails: s.filas@stat.gov.pl,
t.klimanek@stat.gov.pl

[2] Adam Mickiewicz University in Pozna, Poland
e-mail: sylfil2@ext.amu.edu.pl

[3] Pozna University of Economics and Business, Poland
e-mail: tomasz.klimanek@ue.poznan.pl

Abstract

Income stratification of urban population in Poland is a proposal of a new methodological approach in Statistics Poland to the study of personal incomes at the local level. It was designed to be in line with the modern paradigm of statistical data collection, which stipulates that instead of burdening respondents with the obligation of completing multiple questionnaires, national statistical institutions should make the widest possible use of information contained in administrative registers. One of the basic variables describing the populations standard of living is income. Personal income earned by individuals enables them to meet their various needs. Because Statistics Poland processes and publishes income data (especially for households) from sample surveys, the resulting statistics are usually available only at the level of province or even higher levels of spatial aggregation, which are of little use to researchers interested in conducting more detailed socio-economic analyses. The lack of data for lower levels of spatial aggregation is particularly aggravating precisely because the variation in personal incomes becomes evident mainly at these lower levels. So far, official statistics, especially concerning cities and inner-city areas, have not included information about the characteristics describing the level of and variation in personal incomes. The presentation covers a general description of the methodology applied to obtain income stratification of urban population in Poland, including the review of the literature on measures of income, methods of classification and spatial analysis and finally selected statistics about the level and variation in the distribution of incomes earned by inhabitants of Polish towns and cities. Selected results of the study are shown in the form of choropleth maps and tables.

Keywords: income stratification, personal income tax register, cities and inner-cities areas.

References

Atkinson, A. B., Bourguignon, F. (Eds.). (2014) *Handbook of income distribution*, **vol. 2**, Elsevier.

Canberra Group, (2011) *Handbook on Household Income Statistics*, United Nations Economic Commission for Europe, Geneva.

Van Kerm, P. (2007). Extreme incomes and the estimation of poverty and inequality indicators from EU-SILC.

Mastronardi, L., Cavallo, A. (2020) The spatial dimension of income inequality: An analysis at municipal level. *Sustainability*, **12(4), 1622**.

Moser, M., Schnetzer, M. (2014). The Geography of Average Income and Inequality: Spatial Evidence from Austria, *Department of Economics Working Paper Series*, **191**. WU Vienna University of Economics and Business, Vienna.

Organisation for Economic Co-operation and Development, (2013). *OECD framework for statistics on the distribution of household income, consumption and wealth.* OECD Publishing.

Organisation for Economic Co-operation and Development, (2013). *OECD guidelines for micro statistics on household wealth.* OECD Publishing.

MACHINE LEARNING AND AUTOMATIC EDITING

Blaise Ngendangenzwa and Joel Tolsheden

Abstract

This presentation concerns application of Machine Learning (ML) algorithms for imputation in the data editing process. Imputations are made automatically without manual intervention in the process. Imputations are made for missing observations and observations flagged as potentially erroneous by the selective editing program SELEKT. Three different ML algorithms are tested and evaluated on data from the short-term employment survey. The algorithms are trained on "cold" data sets and the selected variants of the algorithms are used for imputation in a "hot" data set. Results imply ML algorithms are useful in the editing process of replacing missing or potentially erroneous observations. Final estimates are close to those obtained under traditional manual imputation. Results also show the importance of selecting the specific algorithm.

THE USE OF ALTERNATIVE DATA SOURCES AT STATISTICS LITHUANIA

Tomas Rudys

Statistics Lithuania, Lithuania
e-mail: tomas.rudys@stat.gov.lt

Abstract

We are giving a short overview of the projects related to the use of alternative data sources at Statistics Lithuania. In this digitalization age huge amount of data is available from different data sources including privately held data. In order to use these kind of data and make inference its important to understand the data generation mechanizms which are usualy not very clear. In this talk first atemps to use alternative data sources and chalanges at statistics Lithuania will be presented. Particularly the talk will focus on legislative, technical aspects of alternative data sources, importance of the need for mew methods and techniques for data integration. Curently projects releated the use of web scraped data for online vacancies and enterprise characteristics is under develompment. Aditionaly engitiation with private trade companies is giving first results to obtain scaner data fro price statististics. Furthemore analysis of posibilities to use satelite imagery data for agriculture statistics is under way. Despite methodological difficulties the posibilities to obtain data from private sector remains. These isues will be shortly discused during the presentation.

Keywords: official statistics, alternative data sources, data integration.

UniDOS ONLINE WITH LIMESURVEY

Mykola Sydorov[1] and Oleksiy Sereda[2]

[1] Taras Shevchenko National University of Kyiv, Ukraine
e-mail: myksyd@knu.ua

[2] Taras Shevchenko National University of Kyiv, Ukraine
e-mail: as_sereda@knu.ua

Abstract

In this paper we showed that it is easy to conduct anonymous surveys in web with LimeSurvey. The UniDOS survey 2020 was made with direst email sending of invitations and remainders. There was formulated 2 most problems: low email rate and low response rate and described how to reduce theirs.

Keywords: web survey, LimeSurvey.

1 Introduction

From 2009 at Taras Shevchenko National University of Kyiv we conduct monitoring survey about different sides of students' life. The last couple of year in before pandemic period we performed solid surveys for 1st year students and multilevel sampled surveys for 2+ year student (bachelors 2-4 years and masters 1-2 years of education) (Sydorov and Sereda 2018). That surveys were conducted with self-filling paper based questionnaires in classrooms during lectures or other lessons. The sampling error did not exceed 4.2% with a confidence level of 0.95.
In 2020 there was impossible to conducts UniDOS (UniDOS, 2021) in same way because of pandemic restrictions: direct contacts between interviewers and respondents was prohibited, all lessons in classrooms were transferred to online mode.
Due to the fact that we have been conducting online surveys since 2006 and switched to LimeSurvey (LimeSurvey, 2021) use (Sydorov, 2009) in 2008, UniDOS 2020 was decided to hold online in LimeSurvey.

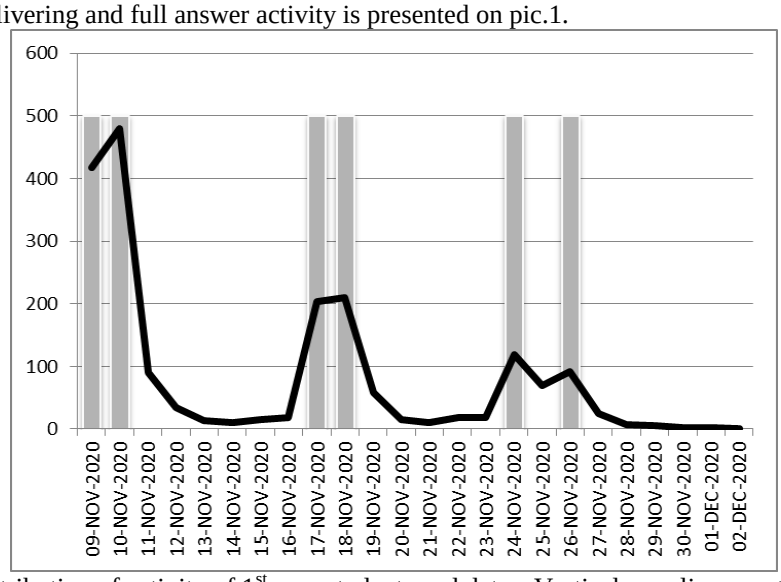
2 Survey methodology and respondent anonymity

In our former paper-based solid surveys of 1st year students in 2013-2019 the response rate was about 50-60%. This was due to the fact that some students during the survey were on internship or assistant training practice, expressed a reluctance to participate in the study or simply absent in classrooms. Since response rate in web surveys is lower than in paper-based, in 2020 we conducted solid web surveys as for 1st year students so as for 2+.
LimeSurvey gives a possibility to create respondents data base and use it for direct mailing of invitations and remainders. This approach is very important for some reasons: we could control does respondent have finished survey, estimate response rate and errors, avoid third-party respondents.
Before survey we have got email addresses of all students (except some students due to local problems in the faculties) that had to take part in the monitoring research 2020. Then we had created email data base and automatically generated unique assess keys (tokens) for each respondent. Every of these tokens could be used just once in survey. This this does not allow third-party respondents to take part in the survey and registered respondents to take the survey twice.
The first question students asked us during survey was about respondents' anonymity. In Lime Survey anonymizer of responses is present. It gives even not confidentiality but real anonymity and existence

of full response control in one time. In the opposite of not anonymous surveys in our case creates 2 data tables. In one there are exists just answers of the respondents and random seeds, in other – mail data base, tools for sending invitations and reminders and indicators of questionnaire completion. It is impossible to determine which respondent indicate which answer but possible to know has fixed respondent completed questionnaire or not. By the way such tool we use for online secret ballot at the Academic Council etc.

3 Survey organization and data collection

In first step letters with invitation was sent to first part of appropriate students (different questionnaire for 1st and 2+ years of study and there are limitation of number of letters to send in one day) with unique token for each person. From our previous researches we decided that the best time for this is from Monday evening to Tuesday evening (Wednesday morning). For 1st year students' emails and remainders delivering and full answer activity is presented on pic.1.



Pic. 1. Distribution of activity of 1st year students and dates. Vertical gray lines are the dates of sending invitations (first pair of gray lines) and remainders (other pairs). Curve – number of complete answers per day.

The amount of 1st year unique students' mails was 4754. We used external mailing service SMTP from ukr.net mail portal and there is a restriction of not more than about 3000 letters per day for sending from one address. That is why we used 2 days for inviting and reminding – 1st and 2nd weeks after inviting.

The same situation was with 2+ surveys, but the former paper based questionnaire for 2+ was quite big as for online (about 276 mono questions). For solving this problem we divided questionnaire for 3 parts: 1 core and 2 crosses. Each of new questionnaires was consist of core and one of cross parts.

Inviting and reminding was organized in same way: Monday evening to Wednesday – inviting, after a week – we have sent reminders. The respondents had the opportunity to refuse of participation in the survey. In this case, they were not reminded again.

After collecting of survey data in LimeSurvey it is easy to export whole data to SPSS, R, SAS, xlsx etc. formats. In exported dataframe all data are anonymous.

4 Results

After collecting the data we received complete answers of 1st year students with response rates mean 0.51 that is very good for web survey (see Table 1). There were some faculties with low email rates – proportion between given emails and number of students.

Table 1. Survey paradata of 1st year of study students.

Faculty	Number of students	Emails	Answers	RRate	Wages	Weighted number
faculty 1	207	204	149	73,04%	0,6	89
faculty 2	633	639	264	41,31%	1,03	273
faculty 3	231	213	121	56,81%	0,83	100
faculty 4	173	172	80	46,51%	0,94	75
faculty 5	372	316	163	51,58%	0,98	160
faculty 6	334	294	125	42,52%	1,15	144
faculty 7	233	181	83	45,86%	1,2	100
faculty 8	255	27	21	77,78%	5,24	110
faculty 9	176	100	57	57,00%	1,33	76
faculty 10	95	91	68	74,73%	0,6	41
faculty 11	126	126	81	64,29%	0,65	53
faculty 12	105	65	42	64,62%	1,07	45
faculty 13	516	496	197	39,72%	1,13	222
faculty 14	218	80	63	78,75%	1,49	94
faculty 15	347	297	128	43,10%	1,17	150
faculty 16	530	518	255	49,23%	0,89	228
faculty 17	761	631	352	55,78%	0,93	328
faculty 18	74	74	18	24,32%	1,72	31
faculty 19	287	230	175	76,09%	0,71	124
Total	5673	4754	2442			2443

As we can see the high wages (more than 1.4) are presented mostly for the faculties with low email rates.

For 2+ students we received the same situation with a bit less response rates (mean 0.33) by faculties (see table 2) but it is usual situation that 2+ students' response rate is lower than of 1st.

Table 2. Survey paradata of 2+ year of study students.

Faculty	Number of students	Emails	Answers	RRate	Wages	Weighted number
faculty 1	845	735	390	53,06%	0,50	195
faculty 2	653	329	152	46,20%	0,99	151
faculty 3	337	320	89	27,81%	0,88	78
faculty 4	1706	1661	437	26,31%	0,90	394
faculty 5	278	200	81	40,50%	0,79	64
faculty 6	971	760	257	33,82%	0,87	224
faculty 7	1982	1337	400	29,92%	1,14	457
faculty 8	2328	1168	457	39,13%	1,18	537
faculty 9	666	552	177	32,07%	0,87	154
faculty 10	608	353	164	46,46%	0,85	140
faculty 11	588	231	64	27,71%	2,13	136
faculty 12	1071	1047	265	25,31%	0,93	247
faculty 13	998	786	183	23,28%	1,26	230
faculty 14	606	399	126	31,58%	1,11	140
faculty 15	207	207	124	59,90%	0,39	48
faculty 16	468	430	136	31,63%	0,79	108
faculty 17	711	66	19	28,79%	8,63	164
faculty 18	399	394	186	47,21%	0,49	92
faculty 19	1872	1262	281	22,27%	1,54	432
	17294	12237	3988			3991

As we can denote – there are 2 most problem for data quality of our survey: low email rate and low response rate. But these problems could be solved or reduced.

5 Conclusion and discussion

We could conclude that such approach of online survey gives appropriate results of data collection procedure. The analysis of biases show that difference between 2019 paper-based and 2020 web-based surveys data collection mostly becomes because of problems with email rates and response rates.
As we noted, there were 2 most problems for data collection quality in our survey: low email rates by faculties (and years) and low response rate by faculties (and years).
First one was caused because of insufficient level of completeness of e-mail addresses lists provided by faculties. This problem could be solved when we will send requests to faculties about collecting students' email addresses earlier and note, that we need emails of all students of fulltime education.
The second problem could be reduced when before survey runs advertising campaign of this survey, present results of previous waves and declare how results of previous waves influence to students' life.

References

LimeSurvey (2021) Limesurvey GmbH. / LimeSurvey: An Open Source survey tool /LimeSurvey GmbH, Hamburg, Germany. URL http://www.limesurvey.org

Sydorov, M. (2009) Using the LimeSurvey shell to conduct online surveys. (in Ukrainian) *I Congress of the Sociological Association of Ukraine "Sociology in a situation of social uncertainty" October 15-16, 2009,* (Kharkiv, Ukraine)

Sydorov, M. and Sereda, O. (2018) Sample models in monitoring survey UniDOS. *Workshop of Baltic-Nordic-Ukrainian Network on Survey Statistics 2018* (Latvia, Jelgava)

UniDOS (2021) *University Social Research*. . URL http://unidos.univ.kiev.ua/?q=uk/zvity_pro_doslidzhennya

SAE METHODS FOR DEVELOPING THE DIGITAL ECONOMY AND SOCIETY INDEX (DESI) AT LOCAL LEVEL

Kaja Sõstra

Statistics Estonia, Estonia
e-mail: kaja.sostra@stat.ee

Abstract

Digital Economy and Society Index (DESI) is a composite index summarising progress on connectivity, digital skills, use of internet by citizens, integration of digital technology by businesses and digital public services on national level. The aim of the current research was to test small area estimation (SAE) methods for estimating some DESI components on local government level. The following four human capital indicators were selected for analyses and testing: share of frequent internet users, communication skills above basic, share of persons who used online banking and share of persons who ordered goods or services online.

Data sources for the local DESI were sample survey of the Information technology of households, statistical population register and employment register. Administrative registers provided auxiliary information for small area estimation models. Four estimators were tested in the present research: Direct, GREG, Synthetic and EBLUP (EURAREA, 2004). The performance of estimators was tested using Monte-Carlo simulation. Artificial population was used for the simulation study. Samples with size 4000 persons were selected from the population by systematic sampling. The sampling design was similar to the real survey design. All possible 246 samples were selected with starting point from the 1st person to the 246th person. Four indicators were estimated from every sample for local governments. The simulation study showed that the Synthetic and EBLUP estimators are reliable estimation methods for local DESI components for small and medium municipalities. The choice between GREG and EBLUP for large municipalities depends on if one prefers unbiased estimator where it is reliable or to use the same method for all areas for better comparability.

The research report is published in Sõstra (2021). The data analysis and simulation study demonstrated that DESI human capital components could be well explained by demographic and socio-economic variables available for all population. Therefore small area estimation methods give reliable results for local DESI if administrative register data or/and some alternative data sources are available for statistical purposes.

Keywords: DESI index, small area estimation (SAE).

References

EURAREA (2004) Enhancing Small Area Estimation Techniques to meet European Needs.

Sõstra, K (2021) Developing the Digital Economy and Society Index (DESI) at local level - "DESI local". *https://futurium.ec.europa.eu/en/urban-agenda/digital-transition/library/desi-local-developing-digital-economy-and-society-index-desi-local-level*

ENTERPRISES SURVEY ON PERSONNEL DEMAND

Natallia Bokun

Belarusian State University (BSEU), Belarus
e-mail: nataliabokun@rambler.ru

Abstract

The In recent years, the growing number of small enterprises, problems of labor force and labor market, have motivated the development of specialized methodology and software for enterprises sample surveys, whose purpose is to study the demand in personnel by occupation and profession.

Since 2018 and until 2020 sample surveys of enterprises employees structure spent without special software. Survey objects were organizations by regions and kinds of activity.

Nowadays, the Ministry of Labor and Social Affairs of the Republic of Belarus together with Research Institute of Labor develop sample survey algorithms and make the preparatory work on implementation of the enterprises sample surveys. In October-November 2021 a test sample survey is being planned. The first results of adaptation of Sample Survey methodology indicated the appearance of significant organizational and methodological problems: non-responses, the need for localization of the sample, using a combination of selection methods, samples in small domains.

This paper has the next parts:

1) sampling frames that incorporate files of enterprises by regions and kinds of activity;

2) questionnaire: number of employees, total and by occupation, profession;

3) sample design: territorial stratified univariate samples are used;

4) statistical weighting that includes traditional Horvitz-Thomson estimator, annual sample updating.

The use of combination of different univariate sample methods (proportional, optimal allocation and others), weighting methods will provide very reliable information over total indicators of employees number. However, standard errors, calculated by separate indicators by occupation in the context of different kinds of activity at regional level, are rather high. To improve the representativeness by region weighting procedure can be complicated by usage of auxiliary calibration estimators.

Keywords: employees, personnel, enterprises, sample survey, sampling frame, weighting, estimator.

References

Bokun, N. (2014) Micro-entities sample survey: problems of design, formation and usage. In: Workshop of Baltic – Nordic – Ukrainian network on Survey Statistics, Tallinn, Estonia, August 25-28, 2014. – P. 25-31.

Bokun, N. (2016) Micro-entities and small enerprizes surveys in Belarus. In: Proc. of the XI Intern. Conf. “Compute Data Analysis and Modeling”, Minsk, Sept. 6-10, 2016. – P. 240-245.

Särndal, C.-E., Swensson, B., Wretman, J. (2003). Model assisted survey sampling. Springer Verlag.
Bookwriter, A. B. (1980) *Title of the Book*. Publisher, City.

LABOR MARKET SURVEYS IN BELARUS

Natallia Bokun

BSEU, Belarus
e-mail: nataliabokun@rambler.ru

Abstract

The experience of conducting special surveys on the labor market in Belarus is considered: Labor Force Survey, Survey of employers about demand for personnel, Survey of wages of employees by category. The design and statistical weighting of surveys are analyzed. It is proposed to use a combination of univariate and multivariate selection methods.

Keywords: labor market, sample, statistical weighting, sampling error, sample size.

1 Introduction

The labor market is a complex system of relationships: employee - state - employer. The development of market relations, on the one hand, contributes to the development of human potential, on the other, to the growth of labor productivity. The importance of labor resources, labor force, labor market is increasing in the context of globalization and pandemic, which is due to a number of factors: aging of the population, the formation of special market segments (caring for the elderly, sick, information technology), integration processes. The tools for assessing the real situation on the labor market are not perfect: special subsystems of labor market indicators that comprehensively reflect its state (demand, supply of labor, market infrastructure, efficiency) are not distinguished, the possibilities of existing information support are not fully used. So, there are no data on latent, cyclical, structural unemployment, retrospective time series of actual unemployment indicators. The directions for the development of labor market statistics are associated with improving the methodology of household surveys on employment issues, surveys of wages of employees by profession and occupation, conducting surveys of employers on personnel structure and demand.

2 Households survey to study the problems of employment

In Belarus, the main sources of information on the labor market up to 2012: current reporting of organizations, administrative data and the population census, - made it possible to measure in detail the labor force indicators, but did not provide annual and monthly estimates of actual unemployment, which, according to the 2009 census, was 6-7 times exceeded its officially registered level, there was no distribution of employed and unemployed by age, profession, employment status, part-time employment was not determined. The above factors led to the need for a special labor force survey, which has been carried out by the National Statistical Committee of the Republic of Belarus on a regular basis since 2012. The main objectives of the survey: to study the state and dynamics of demand - supply of labor, the formation of official statistical information on the number of employed, unemployed, causes and duration of unemployment. Survey objects are private households, residents aged 15-74.
The survey is carried out quarterly, in each region and separately in Minsk. Taking into account possible non-responses, the selection share is 0.9%, or 37.2 thousand households. The territorial probabilistic three-stage sampling is used. At the first stage, the selection units are cities, urban-type settlements, village councils, at the second - the enumeration areas formed during the last census, and rural settlements, at the third - households. Rotation of 25% of the sampled households is carried out annually. The selection procedure for administrative-territorial units is carried out once every 4 years.

The methodology for statistical weighting and dissemination of data to the general population is based on assigning an appropriate weight to each holding (Bi):

$$B_i = \frac{1}{p_1 \cdot p_2 \cdot p_3}, \quad (1)$$

where p_1 - probability of the selection of each city (village council);

p_2 - probability of selection of each polling district in cities, village council;

p_3 - probability of selecting a household.

Individual weights of respondents are calculated based on the results of iterative weighing: weights are calculated separately by gender, urban and rural areas; adjustments are made to the initial coefficients, first in the context of urban and rural areas, then - for five-year age groups.
Final individual weight for the respondent in each 5-year group:

$$K_i = B_в \cdot k_1, \quad (2)$$

where $B_в = \frac{S_j}{s_j}$; $k_1 = \frac{S_t}{S_\varepsilon}$;

S_j, s_j – population size in the j-th age and sex group based on the results of Census and survey;

S_t – population size in the t-th group by urban (rural), sex (on the Census data);

S_ε – extrapolated population size in the t-th group (by $B_в$).

To increase the representativeness of the data (by region, urban and rural areas, age and gender groups), it is possible to increase the number of iterations, use alternative weighing schemes.

3 Survey of wages of employees by profession and occupation

The survey is carried out twice every 5 years. Enterprises are surveyed by type of activity. A two-stage sampling is used: at the first stage, using a combination of one-dimensional and multidimensional methods, organizations are selected, at the second - at each selected enterprise, in turn, workers are mechanically selected. As a sampling frame, an array of organizations in the form 12-t "Labor report" is used, as well as the payroll number of employees for October. The approximate sampling fraction of organizations is 35-40%, the relative sample error in the country as a whole is up to 2%, by type of activity - up to 6-7%.
Data extrapolation is performed using aggregated and final weights:

$$k_{ai} = k_n \cdot k_{ui};\ k_n = \frac{N_m}{n_m};\ k_{ui} = \frac{T_i}{t_i}, \quad (3)$$

where k_n – weight of the organization;

k_{ui} – individual weight of an employee of the *i*-th category;

N_m, n_m – number of organizations of the *m*-th group in general and sample populations accordingly

T_i –total number of employees of the *i*-th category in organization;

t_i – number of employees of the *i*-th category who were included in sample population of this organization.

4 Survey of employers on the personnel demand

The survey is carried out annually, starting from 2018-2019, jointly by the Ministry of Labor and Social Affairs and the Research Institute of Labor. The goal is to study the composition of the workforce at the level of organizations by main types of activity in the context of categories and occupations. As separate sections of the questionnaire, there are: payroll number of employees in the context of categories and occupations, skills and abilities of employees. Questions are asked not only about the available number of employees, but also about the needs in the coming years. During 2020-2021 the development of algorithms and specialized software is carried out. They are:
- formation of sample populations of organizations by region and type of activity, using a combination of random selection without stratification and simple, proportional and optimal stratification;
- statistical weighting using the Horwitz-Thompson estimator;
- updating sample population for non-responses by imputing data (replacing non-responses, duplicating organizations, etc.).
The sampling frame is an array of organizations reporting in the form 4-fund; excluded sections O, T, I, housing, construction, garage cooperatives, parties, confessions.
The main problems of sampling are associated with a high number of non-responses (up to 50-60%), the presence of atypical units, and fragmentation of the population as a result of regional and sectoral distribution.

Conclusion

The experience of conducting employment surveys, surveys of wages, surveys of employers in Belarus has shown:
- survey problems are mainly associated with the presence of non-responses, the need of localization of sampling, regional subsamples construction; the need to use different weighing and extrapolation schemes;
- the most optimal model for selecting households is a three-stage stratified sampling; for organizations is a combination of univariate and multivariate samples;
- the recommended sampling fraction of organizations is 20-35%; households is 0.4-0.9%.

References

Инструкция по организации и проведению выборочного обследования домашних хозяйств в целях изучения проблем занятости // Постановление Белстата №10 от15.03.2013 с дополнениями от 17.04.2020, 18.12.2020. – Минск: Белстат, 2020. – 18 с.

Bokun, N. Labour Force Survey in Belarus: determination of sample size, sample design, statistical weighting / N.Bokun / Workshop of Baltic-Nordic-Ukrainian Network on Survey Statistics, August 24-28, 2012, Valmiera, Latvia. – p. 37-45.

Инструкция по организации и проведению выборочного государственного статистического наблюдения о заработной плате работников по профессиям и должностям // Постановление Белстата №123 от 28.07.2014 с дополнениями от 13.06.2016, 14.06.2019. – Минск: Белстат, 2021. – 11 с.

Разработка прогноза потребности экономики в кадрах на пятилетний период по профессионально-квалификационным группам: Отчет о НИР (заключительный). – Минск: НИИ труда, 2019. – 190 с.

ВЫБОРОЧНЫЕ ОБСЛЕДОВАНИЯ НА РЫНКЕ ТРУДА В БЕЛАРУСИ

Наталия Бокун

БГЭУ, Беларусь
e-mail: nataliabokun@rambler.ru

Аннотация

Рассмотрен опыт проведения в Беларуси специальных обследований на рынке труда: обследование рабочей силы, опрос нанимателей о потребности в кадрах, обследование условий заработной платы работников по категориям. Проанализированы дизайн и статистическое взвешивание обследований. Предложено использовать сочетание одномерных и многомерных методов отбора.

Ключевые слова: рынок труда, выборка, статистическое взвешивание, ошибка выборки, объем выборки.

1 Введение

Рынок труда представляет собой сложную систему взаимоотношений: работник – государство – наниматель. Развитие рыночных отношений, с одной стороны, содействует развитию человеческого потенциала, с другой – росту производительности труда. Значимость трудовых ресурсов, рабочей силы, рынка труда усиливается в условиях глобализации и пандемии, что обусловлено рядом факторов: старение населения, формирование особых сегментов рынка (уход за престарелыми, больными, информационные технологии), интеграционные процессы. Инструменты оценки реальной ситуации на рынке труда не совершенны: не выделяются специальные подсистемы индикаторов рынка труда, комплексно отражающих его состояние (спрос, предложение рабочей силы, инфраструктура рынка, эффективность), не полностью используются возможности существующего информационного обеспечения. Так, отсутствуют данные о скрытой, циклической, структурной безработице, ретроспективные динамические ряды показателей фактической безработицы. Направления развития статистики рынка труда связаны с совершенствованием методологии проводимых обследований домашних хозяйств по вопросам занятости, обследований заработной платы работников по профессиям и должностям, проведением опросов нанимателей о составе и потребности в кадрах.

2 Выборочное обследование домашних хозяйств в целях изучения проблем занятости населения

В Беларуси основные источники информации о рынке труда до 2012 года: текущая отчетность организаций, административные данные и перепись населения, – позволяли детально измерить показатели рабочей силы, но не давали годовых и ежемесячных оценок фактической безработицы, которая по данным переписи 2009 года в 6-7 раз превышала ее официально зарегистрированный уровень, отсутствовало распределение занятых и безработных по возрасту, профессиям, не определялся статус занятости, частичная занятость. Перечисленные факторы обусловили необходимость специального обследования рабочей силы, которое с 2012 года проводится Национальным статистическим комитетом Республики Беларусь на регулярной основе. Основные цели обследования: изучение состояния и динамики спроса – предложения рабочей силы, формирование официальной статистической информации о

численности занятых, безработных, причинах и продолжительности безработицы. Опрашиваются частные домашние хозяйства, резиденты в возрасте 15-74 лет.
Обследование проводится ежеквартально, по каждой области и отдельно в г. Минске. С учетом возможных неответов доля отбора составляет 0,9 %, или 37,2 тысячи домашних хозяйств. Используется территориальная вероятностная трехступенчатая выборка. На первой ступени единицы отбора – города, поселки городского типа, сельсоветы, на второй – счетные участки, сформированные при проведении последней переписи, и сельские населенные пункты, на третьей – домашние хозяйства. Ежегодно осуществляется ротация 25 % домашних хозяйств, попавших в выборку. Процедура отбора административно-территориальных единиц осуществляется 1 раз в 4 года.
Методология статистического взвешивания и распространения данных на генеральную совокупность основана на присвоении каждому хозяйству соответствующего веса (B_i):

$$B_i = \frac{1}{p_1 \cdot p_2 \cdot p_3}, \quad (1)$$

где p_1 – вероятность отбора каждого города (сельсовета);

p_2 – вероятность отбора счетного участка или сельского населенного пункта;

p_3 – вероятность отбора домашнего хозяйства.

Индивидуальные веса респондентов рассчитываются по результатам итеративного взвешивания: веса рассчитываются отдельно по полу, городской и сельской местности; осуществляются корректировки начальных коэффициентов сначала в разрезе городской и сельской местности, затем – по пятилетним возрастным группам.
Конечный индивидуальный вес для респондента каждой пятилетней группы:

$$K_i = B_в \cdot k_1, \quad (2)$$

где $B_в = \frac{S_j}{s_j}$; $k_1 = \frac{S_t}{S_\varepsilon}$;

S_j, s_j – численность населения в j-й половозрастной группе по результатам последней переписи и выборки;

S_t – численность населения в t-й группе городской или сельской местности по полу;

S_ε – экстраполированная численность населения в t-й группе (по $B_в$).

Для повышения репрезентативности данных (в разрезе областей, городской и сельской местности, половозрастных групп) возможно увеличение числа итераций, использование альтернативных схем взвешивания.

3 Выборочное обследование заработной платы работников по профессиям и должностям

Обследование проводится два раза в 5 лет. Обследуются предприятия по видам деятельности. Используется двухступенчатая выборка: на первой ступени с использованием комбинации одномерного и многомерного методов отбираются организации, на второй – на каждом отобранном предприятии, в свою очередь, механическим способом отбираются работники. В

качестве основы выборки применяется массив организаций по форме 12-т «Отчет по труду», а также списочная численность работников за октябрь. Примерная доля отбора организаций – 35-40 %, относительная стандартная ошибка выборки по республике в целом – до 2 %, по видам деятельности – до 6-7 %.
Экстраполяция данных осуществляется с использованием агрегированного и конечного весов:

$$k_{ai} = k_n \cdot k_{ui};\ k_n = \frac{N_m}{n_m};\ k_{ui} = \frac{T_i}{t_i}, \quad (3)$$

где k_n – вес организации;

k_{ui} – индивидуальный вес работника *i*-й категории;

N_m, n_m – число организаций *m*-й группы соответственно в генеральной и выборочной совокупностях;

T_i – общее число работников *i*-й категории в организации;

t_i – число работников *i*-й категории, попавших в обследование по данной организации.

Полученная выборочная совокупность выступает основой для формирования статистической отчетности по форме 6-т (профессии). Исходная информация формируется на региональном уровне, обобщается и экстраполируется – на республиканском.

4 Опрос нанимателей о потребности в кадрах

Обследование проводится ежегодно, начиная с 2018-2019 гг., совместно Министерством труда и социальной защиты и НИИ труда. Цель – изучение состава рабочей силы на уровне организаций по основным видам деятельности в разрезе категорий и занятий. В качестве отдельных разделов опросника выступают: списочная численность работников в разрезе категорий и занятий, умения и навыки работников. Задаются вопросы не только о наличной численности работников, но и потребности в последующие годы. В течение 2020-2021 гг. осуществляется разработка алгоритмов и специализированного программного обеспечения, которое предполагает:
- формирование выборочных совокупностей организаций в пределах регионов и видов деятельности с использованием комбинации собственно-случайного отбора, простого, пропорционального и оптимального расслоения;
- экстраполяцию выборочных показателей по методу Горвица-Томпсона;
- корректировку состава выборочной совокупности на неответы путем импутации данных (замена неответов, дублирование организаций и т.д.).
Основа выборки – массив организаций, отчитывающихся по форме 4-фонд; исключены секции О, Т, И, жилищные, строительные, гаражные кооперативы, партии, конфессии.
Основные проблемы проведения выборки связаны с высоким количеством неответов (до 50-60 %), наличием нетипичных единиц, дроблением совокупности в результате регионального и отраслевого распределения.

Заключение

Опыт проведения в Беларуси обследований по вопросам занятости, обследований заработной платы, опросов нанимателей показал:
- проблемы обследований в основном связаны с наличием неответов, необходимостью локализации малых выборок, построением региональных подвыборок; необходимостью использования различных схем взвешивания и экстраполяции;
- наиболее оптимальная модель отбора домашних хозяйств – трехступенчатая расслоенная выборка; для организаций – комбинация одномерных и многомерной выборок;

- рекомендуемая доля отбора организаций – 20-35 %; домашних хозяйств – 0,4-0,9 %.

Список использованных источников

Инструкция по организации и проведению выборочного обследования домашних хозяйств в целях изучения проблем занятости // Постановление Белстата №10 от15.03.2013 с дополнениями от 17.04.2020, 18.12.2020. – Минск: Белстат, 2020. – 18 с.

Bokun, N. Labour Force Survey in Belarus: determination of sample size, sample design, statistical weighting / N.Bokun / Workshop of Baltic-Nordic-Ukrainian Network on Survey Statistics, August 24-28, 2012, Valmiera, Latvia. – p. 37-45.

Инструкция по организации и проведению выборочного государственного статистического наблюдения о заработной плате работников по профессиям и должностям // Постановление Белстата №123 от 28.07.2014 с дополнениями от 13.06.2016, 14.06.2019. – Минск: Белстат, 2021. – 11 с.

Разработка прогноза потребности экономики в кадрах на пятилетний период по профессионально-квалификационным группам: Отчет о НИР (заключительный). – Минск: НИИ труда, 2019. – 190 с.

A SEQUENTIAL PROBABILITY RATIO TEST FOR ONLINE EXPERIMENTS

Yana Bondarenko

Oles Honchar Dnipro National University, Ukraine
e-mail: yana.bondarenko@pm.me

Abstract

We propose an extension of the sequential probability ratio test (SPRT) for online randomized experiments. The hallmark of SPRT is the formulation of the composite alternative hypothesis on the conversion rate difference and Bayesian inference for estimation of unknown parameters using pre-experiment data.

Keywords: Conversion Rate Optimization, Sequential Analysis, Bayesian Inference

1 Introduction

According to the optimization glossary, conversion rate optimization is the process of increasing the percentage of conversion from website or mobile app. Conversion rate optimization involves generating ideas for elements on the site or app that can be improved and then validating those hypotheses through A/B testing and multivariate testing (Kohavi, Henne, et.al., 2007; Kohavi, Longbotham, et.al., 2009; Kohavi, Deng, et.al., 2013; Kohavi, Deng, et.al., 2014; Stucchio, 2015; Pekelis, Walsh, et.al., 2015; Johari, Koomen, et. al., 2017; Abhishek & Mannor, 2017; Johari, Pekelis, et. al., 2019; Bondarenko & Kravchenko, 2019).

A simple online randomized experiment uses a statistical method of comparing two versions of a webpage or app against each other to determine which one performs better. Versions A and B are exposed for the baseline and experimental group of visitors, respectively. We need to identify visitors during testing for clear experiment and suggest them the same version that they viewed earlier in case of repeated visits. Each visitor can belong to the baseline or experimental group with probability $1/2$. Visitor behavior is determined with two outcomes: success – conversion action is done, failure – conversion action isn't done. If visitor belongs to the baseline group, success will happen with probability p_1, if visitor belongs to the experimental group, success will happen with probability p_2. Conversion rate difference $p_1 - p_2$ fluctuates with unknown mean and unknown variance (Fig. 1).

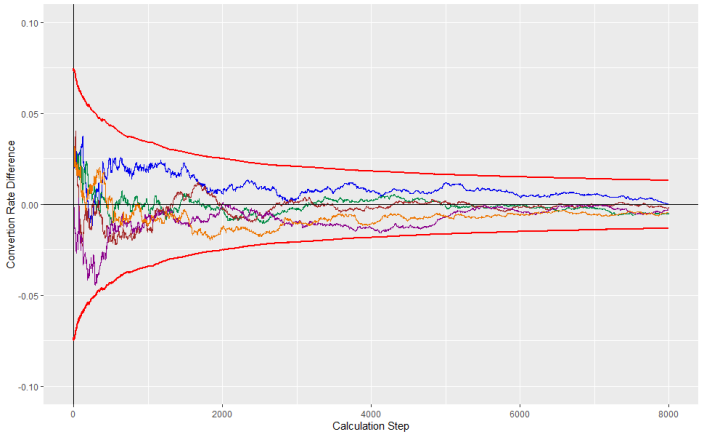


Fig. 1. Fluctuation of conversion rate difference

Denote the conversion rate difference by θ. Let θ be a random variable whose distribution is normal distribution with unknown mean μ and variance σ^2. In this paper, Bayesian estimation of parameters μ and σ^2 is presented and mixture SPRT as a technique of decision-making by successively gathering and processing data is proposed.

2 A Sequential Probability Ratio Test

2.1 Bayesian estimation of mean μ and variance σ^2

Consider a pre-experiment data of conversion rate differences. Let $\theta_1, \theta_2, ..., \theta_n$ be independent and identically distributed normal random variables with unknown mean μ and unknown variance σ^2. The likelihood function of the parameters (μ, σ^2) is

$$L(\theta_1, ..., \theta_n; \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(\theta_i - \mu)^2}{2\sigma^2}\right\}$$

$$\propto \tau^{n/2} \exp\left\{-\frac{\tau}{2}\left(ns^2 + n(\bar{\theta} - \mu)^2\right)\right\}, \quad (1)$$

where $\bar{\theta}$ is sample mean and s^2 is sample variance, $\tau = \sigma^{-2}$.
Let prior distribution on parameters (μ, τ) be normal-gamma distribution with known parameters $(\mu_0, \lambda_0, \alpha_0, \beta_0)$:

$$\pi(\mu, \tau) = \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \tau^{\alpha_0 - 1} \exp\{-\beta_0 \tau\} \frac{\sqrt{\lambda_0 \tau}}{\sqrt{2\pi}} \exp\left\{-\frac{\lambda_0 \tau (\mu - \mu_0)^2}{2}\right\}$$

$$\propto \tau^{\alpha_0 - 1/2} \exp\{-\beta_0 \tau\} \exp\left\{-\frac{\lambda_0 \tau (\mu - \mu_0)^2}{2}\right\}. \quad (2)$$

Posterior distribution of the parameters (μ, τ) can be determined by Bayes' theorem:

$$\pi(\mu, \tau \mid \theta_1, ..., \theta_n) \propto L(\theta_1, ..., \theta_n; \mu, \tau)\pi(\mu, \tau)$$

$$\propto \tau^{n/2} \exp\left\{-\frac{\tau}{2}\left(ns^2 + n(\bar{\theta} - \mu)^2\right)\right\} \tau^{\alpha_0 - 1/2} \exp\{-\beta_0 \tau\} \exp\left\{-\frac{\lambda_0 \tau (\mu - \mu_0)^2}{2}\right\}$$

$$\propto \tau^{n/2 + \alpha_0 - 1/2} \exp\left\{-\tau\left(\frac{ns^2}{2} + \beta_0\right)\right\} \exp\left\{-\frac{\tau}{2}\left(\lambda_0(\mu - \mu_0)^2 + n(\bar{\theta} - \mu)^2\right)\right\}$$

$$\propto \tau^{n/2 + \alpha_0 - 1/2} \exp\left\{-\tau\left(\beta_0 + \frac{ns^2}{2} + \frac{\lambda_0 n(\bar{\theta} - \mu_0)^2}{2(\lambda_0 + n)}\right)\right\} \exp\left\{-\frac{\tau}{2}(\lambda_0 + n)\left(\mu - \frac{\lambda_0 \mu_0 + n\bar{\theta}}{\lambda_0 + n}\right)^2\right\}. \quad (3)$$

We move away from normal-gamma probability density function with parameters $(\mu_0, \lambda_0, \alpha_0, \beta_0)$ to normal-gamma probability density function with parameters:

$$\left(\frac{\lambda_0 \mu_0 + n\bar{\theta}}{\lambda_0 + n}, \lambda_0 + n, \alpha_0 + \frac{n}{2}, \beta_0 + \frac{ns^2}{2} + \frac{\lambda_0 n(\bar{\theta} - \mu_0)^2}{2(\lambda_0 + n)}\right). \quad (4)$$

Bayesian estimators

$$\hat{\mu} = \mu_0,\ \hat{\tau} = \alpha_0 \beta_0^{-1} \tag{5}$$

can be computed after n observations:

$$\hat{\mu} = \frac{\lambda_0 \mu_0 + n\bar{\theta}}{\lambda_0 + n}, \tag{6}$$

$$\hat{\tau} = \left(\alpha_0 + \frac{n}{2}\right)\left(\beta_0 + \frac{ns^2}{2} + \frac{\lambda_0 n(\bar{\theta} - \mu_0)^2}{2(\lambda_0 + n)}\right)^{-1}. \tag{7}$$

Bayesian estimator $\hat{\mu}$ can be explained as a weighted average of prior mean μ_0 and sample mean $\bar{\theta}$ of conversion rate difference, where parameters λ_0 and n can be considered as sample sizes for calculation of prior mean and sample mean, respectively. Bayesian estimator $\hat{\tau}$ can be interpreted as ratio of shape to rate of posterior gamma distribution.

2.2 A sequential test of a simple hypothesis against a set of infinitely many alternatives

Consider a sequential probability ratio test with strength (α, β) for testing $H_0 : p_1 - p_2 = 0$ against $H_\theta : p_1 - p_2 = \theta$. Let p_{0n} be joint distribution of $(X_1, Y_1),...,(X_n, Y_n)$ that null hypothesis H_0 is true after n observations have been made:

$$p_{0n} = L(0; x_1,..., x_n, y_1,..., y_n), \tag{8}$$

and let p_{1n} be weighted average of joint distributions of $(X_1, Y_1),...,(X_n, Y_n)$ that correspond a set of paramater points θ in rejection region ω_0 of null hypothesis H_0:

$$p_{1n} = \int_{\omega_0} L(\theta; x_1,..., x_n, y_1,..., y_n)\,\omega(\theta)\,\mathrm{d}\theta, \tag{9}$$

where $\omega(\theta)$ is non-negative weight function satisfying

$$\int_{\omega_0} \omega(\theta)\,\mathrm{d}\theta = 1. \tag{10}$$

Ratio of distributions has the following form:

$$\frac{p_{1n}}{p_{0n}} = \frac{\int_{\omega_0} L(\theta; x_1,..., x_n, y_1,..., y_n)\,\omega(\theta)\,\mathrm{d}\theta}{L(0; x_1,..., x_n, y_1,..., y_n)}. \tag{11}$$

Let weight function $\omega(\theta)$ be normal probability density function with Bayesian estimators $(\hat{\mu}, \hat{\sigma}^2)$, $\hat{\sigma}^2 = \hat{\tau}^{-1}$. Then statistic for mixture Sequential Probability Ratio Test

$$\frac{p_{1n}}{p_{0n}} = \frac{\int_{-\infty}^{+\infty} \exp\left\{-\frac{(\hat{p}_1 - \hat{p}_2 - \theta)^2}{2\hat{p}_0(1-\hat{p}_0)/n}\right\} \cdot \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \exp\left\{-\frac{(\theta - \hat{\mu})^2}{2\hat{\sigma}^2}\right\} \mathrm{d}\theta}{\exp\left\{-\frac{(\hat{p}_1 - \hat{p}_2 - \hat{\mu})^2}{2\hat{p}_0(1-\hat{p}_0)/n}\right\}} \tag{12}$$

takes the closed form solution:

$$\frac{p_{1n}}{p_{0n}} = \sqrt{\frac{V}{V+\hat{\sigma}^2}} \exp\left\{\frac{1}{2}\frac{\left(\hat{p}_1 - \hat{p}_2 - \hat{\mu}\right)^2 \hat{\sigma}^2}{\left(V+\hat{\sigma}^2\right)V}\right\}, \tag{13}$$

$$V = \frac{\hat{p}_0\left(1-\hat{p}_0\right)}{n}, \tag{14}$$

$$\hat{p}_0 = \frac{\hat{p}_1 + \hat{p}_2}{2}. \tag{15}$$

The sequence of statistics p_{1n}/p_{0n} forms a martingale under the null hypothesis H_0. According to Doob's martingale inequality, type I error is controlled at any time during sequential testing:

$$P\left\{\max_{0\le k\le n}\frac{p_{1k}}{p_{0k}} \ge \frac{1}{\alpha}\right\} \le \alpha, \quad \alpha > 0. \tag{16}$$

A stopping rule for mixture Sequential Probability Ratio Test is

$$\inf\left\{n > 0 : \frac{p_{1n}}{p_{0n}} < \frac{1}{\alpha}\right\}. \tag{17}$$

3 Conclusions

We combine a prior information (a prior distribution for parameters μ and σ^2) with information about successive observations of visitor actions during online pre-experiment, which gradually begins to prevail in a posterior distribution for parameters μ and σ^2. We propose the extension of the sequential probability ratio test for online A/B testing.

An immediate consequence of our result is well-known mixture Sequential Probability Ratio Test (Pekelis, Walsh, et.al., 2015) under the null hypothesis $H_0 : p_1 - p_2 = 0$ and a prior $N(0,\sigma^2)$, where σ^2 is determined from the outcomes of extensive analysis of historical online experiments run on Optimizely's platform.

References

Abhishek, V., Mannor, S. (2017) A nonparametric sequential test for online randomized experiments, arXiv: 1610.02490v4.

Bernardo, J.M., Smith, A.F.M. (1994) Bayesian Theory, Wiley.

Bondarenko, Ya.S., Kravchenko, S.V. (2017) On the frequentist approach to multivariate landing page testing. *Visnyk DNU. Series: Modelling*, **9**, 142–151.

Bondarenko, Ya.S., Kravchenko, S.V. (2019) Bayesian approach to landing page testing. *Problems of Applied Mathematics and Mathematical Modelling*, **20**, 3-16.

Bondarenko, Ya. (2019) Sequential A/B Testing. In: *Proceedings of the 2019 IEEE International Conference on Advanced Trends in Information Theory*, Kyiv, Ukraine.

Goodson, M. (2014) Most winning a/b test results are illusory. Whitepaper, Qubit.

Johari, R., Pekelis, L., and Walsh, D. (2016) Always valid inference: Bringing sequential analysis to A/B testing, arXiv:1512.04922v36.

Johari, R., Koomen, P., Pekelis, L., and Walsh, D. (2017) Peeking at a/b tests: Why it matters, and what to do about it. In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1517–1525.

Kohavi, R., Henne, R.M., Sommerfield, D. (2007) Practical guide to controlled experiments on the web: listen to your customers not to the hippo. In: *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, 959–967.

Kohavi, R., Longbotham, R., Sommerfield, D., Henne, R.M. (2009) Controlled experiments on the web: survey and practical guide. *Data mining and knowledge discovery*, vol. 18, no. 1, 140–181.

Kohavi, R., Deng, A., Frasca, B., Walker, T., Xu, Y. and Pohlmann, N. (2013) Online controlled experiments at large scale. In: *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1168–1176.

Kohavi, R., Deng, A., Longbotham, R., and Xu, Y. (2014) Seven rules of thumb for web site experimenters. In: *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*.

Pekelis, L., Walsh, D., Johari, R. (2015) The new stats engine. Whitepaper, Optimizely.

Robbins, H. (1970) Statistical methods related to the law of the iterated logarithm. *The Annals of Mathematical Statistics*, 41(5): 1397-1409.

Stucchio, C. (2015) Bayesian A/B Testing at VWO. Whitepaper, Visual Website Optimizer.

Wald, A. (1945) Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2): 117-186.

SELECTIVE EDITING USING CONTAMINATION MODEL

Ieva Burakauskaitė[1] and Vilma Nekrašaitė-Liegė[2]

[1] Statistics Lithuania, Lithuania
e-mail: ieva.burakauskaite@stat.gov.lt

[2] Vilnius Gediminas Technical University, Statistics Lithuania, Lithuania
e-mail: vilma.nekrasaite-liege@vilniustech.lt

Abstract

Selective editing was applied to the data editing process of the quarterly statistical survey on service enterprises (turnover indicator) of Statistics Lithuania. Predictions of the target variable were obtained using the contamination model. An impact of a potential error on a sample estimate was evaluated using a score function with a standard structure âĂŞ a difference between the observed value of the target variable and its prediction multiplied by a sample weight and a suspicion component. A discrete and a continuous suspicion components were used and an impact of the suspicion component on the effectiveness of selective editing was investigated.

Keywords: contamination model; selective editing; data validation; statistical survey; official statistics.

Introduction

An appropriate accuracy of sample estimates is one of the most important results to be achieved using sampling methods in official statistics. Accuracy of sample estimates depends not only on sampling strategy (a sampling plan and an estimator) but on the quality of statistical data as well. Commonly, an unknown part of statistical data contains errors. According to various studies, in order to achieve a desired accuracy of a sample estimate, it is unnecessary to edit all of the detected errors. The main idea of selective editing is to identify and sort errors according to the influence they have on the sample estimate (Lawrence and McDavitt 1994; Lawrence and McKenzie 2000). It is also worth noting that error detection is usually carried out before the calculation of sample estimates. Therefore, it is important to identify only the part of erroneous data that must be edited. Selective editing remains an important, uncommon topic for research in Lithuania.

The first part of the paper introduces the contamination model and the selective editing method that form the base for the practical study of the outlier detection. The second part of the paper shortly presents a study that was carried out using statistical data. During the study some randomly selected values of statistical data were replaced with errors. The detection of randomly introduced errors were then carried out using a few versions of selective editing. The comparison of results as well as its summary are presented in the Conclusions. Calculations were carried out with the statistical programming language R and its package `SeleMix` that has been designed to execute the selective editing method (RDocumentation 2020).

1 Methodology on Selective Editing

1.1 Contamination Model

Suppose that true (unobserved) data are independent realizations of p-variate random vectors $\mathbf{Y}_i^* = (\mathbf{Y}_{i1}^*, \ldots, \mathbf{Y}_{ip}^*)'$, $i = 1, \ldots, n$, with a Gaussian distribution with mean vectors $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{ip})'$ and common covariance matrix $\boldsymbol{\Sigma}$. Also, a set of q covariates $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iq})'$ exists for every sampled unit i and $\boldsymbol{\mu}_i = \mathbf{B}'\boldsymbol{x}_i$ where $\mathbf{B}$ is a $q \times p$ matrix of unknown coefficients (Di Zio and Guarnera 2013). The corresponding true data model can be expressed as

$$\mathbf{Y}^* = \mathbf{XB} + \mathbf{U} \tag{1}$$

where $\mathbf{Y}^*$ is the $n \times p$ true data matrix, $\mathbf{X}$ – $n \times q$ covariate matrix and $\mathbf{U}$ – $n \times p$ matrix of normal residuals. Rows of the $\mathbf{U}$ matrix are independent realizations of Gaussian random vectors with mean equal to $\mathbf{0}$ and a covariance matrix $\boldsymbol{\Sigma}$.
The generic marginal probability distributions of the ith sampled unit of matrices $\mathbf{Y}^*$ (true data) and $\mathbf{U}$ (residuals) are denoted as

$$f(\boldsymbol{y}_i^*) = N(\boldsymbol{y}_i^*; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}), \quad f(\boldsymbol{u}_i) = N(\boldsymbol{u}_i; \mathbf{0}, \boldsymbol{\Sigma}), \quad i = 1, \ldots, n. \tag{2}$$

In general form $N(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a marginal probability distribution of the p-variate random vector $\mathbf{Y}$ with mean equal to $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$.
It is assumed that the presence of errors in data is described by independent Bernoulli random variables. Therefore the observed (erroneous) data can be expressed as

$$\mathbf{Y} = \mathbf{Y}^* + \mathbf{I}\boldsymbol{\epsilon} \tag{3}$$

where $\mathbf{I}$ is a diagonal $n \times n$ matrix with its diagonal elements equal to Bernoullian variables $I_1, \ldots, I_n$ ($I_i = 1$ if the corresponding sampled unit is erroneous and $I_i = 0$ otherwise, $i = 1, \ldots, n$). A marginal probability distribution of the p-variate random vector $\boldsymbol{\epsilon}_i$ (random noise) can be expressed as

$$f(\boldsymbol{\epsilon}_i) = N(\boldsymbol{\epsilon}_i; \mathbf{0}, \boldsymbol{\Sigma}_\epsilon), \quad \boldsymbol{\Sigma}_\epsilon = (\alpha - 1)\boldsymbol{\Sigma}, \tag{4}$$

with a numeric constant $\alpha > 1$.
$f(\boldsymbol{y}|\boldsymbol{y}^*)$ denotes a conditional marginal probability distribution of random variables $\mathbf{Y}$ and $\mathbf{Y}^*$. Therefore, model (3) can be expressed equivalently:

$$f(\boldsymbol{y}|\boldsymbol{y}^*) = (1 - \pi)\delta(\boldsymbol{y} - \boldsymbol{y}^*) + \pi N(\boldsymbol{y}; \boldsymbol{y}^*, \boldsymbol{\Sigma}_\epsilon) \tag{5}$$

where π is âĂIJa prioriâĂİ probability of contamination and $\delta(\boldsymbol{y} - \boldsymbol{y}^*)$ is the delta function with mass at $\boldsymbol{y}^*$.
Furthermore, a marginal probability distribution of the observed data can be expressed as

$$f(\boldsymbol{y}_i) = (1 - \pi)N(\boldsymbol{y}_i; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}) + \pi N(\boldsymbol{y}_i; \boldsymbol{\mu}_i, \alpha\boldsymbol{\Sigma}). \tag{6}$$

Coefficients of the later observed data model can be obtained by the maximum likelihood estimation.

1.2 Selective Editing

Selective editing is based on the comparison between the observed data and predictions of the true (unobserved) data. The later can be obtained from a conditional marginal probability distribution $f(\boldsymbol{y}_i^*|\boldsymbol{y}_i)$ (Di Zio and Guarnera 2013). An application of the Bayes formula provides:

$$f(\boldsymbol{y}_i^*|\boldsymbol{y}_i) = \tau_1(\boldsymbol{y}_i)\delta(\boldsymbol{y}_i^* - \boldsymbol{y}_i) + \tau_2(\boldsymbol{y}_i)N(\boldsymbol{y}_i^*; \tilde{\boldsymbol{\mu}}_i, \tilde{\boldsymbol{\Sigma}}) \tag{7}$$

where $\tilde{\boldsymbol{\mu}}_i = \frac{\boldsymbol{y}_i+(\alpha-1)\boldsymbol{\mu}_i}{\alpha}$ and $\tilde{\boldsymbol{\Sigma}} = \left(1-\frac{1}{\alpha}\right)\boldsymbol{\Sigma}$, $\delta(\boldsymbol{y}_i^* - \boldsymbol{y}_i)$ is the delta function with mass at $\boldsymbol{y}_i$, $\tau_1(\boldsymbol{y}_i)$ and $\tau_2(\boldsymbol{y}_i)$ are posterior probabilities that the ith sampled unit with observed values $\boldsymbol{y}_i$, $i = 1, \ldots, n$, is not erroneous and that it is contaminated respectively:

$$\begin{aligned} \tau_1(\boldsymbol{y}_i) &= P(\boldsymbol{y}_i = \boldsymbol{y}_i^* | \boldsymbol{y}_i) = \frac{(1-\pi)N(\boldsymbol{y}_i; \boldsymbol{\mu}_i, \boldsymbol{\Sigma})}{(1-\pi)N(\boldsymbol{y}_i; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}) + \pi N(\boldsymbol{y}_i; \boldsymbol{\mu}_i, \alpha\boldsymbol{\Sigma})}, \\ \tau_2(\boldsymbol{y}_i) &= P(\boldsymbol{y}_i \neq \boldsymbol{y}_i^* | \boldsymbol{y}_i) = 1 - \tau_1(\boldsymbol{y}_i). \end{aligned} \tag{8}$$

Posterior probabilities (8) are defined in terms of the conditional expected value $\tilde{\boldsymbol{y}}_i = E(\boldsymbol{y}_i^* | \boldsymbol{y}_i)$, $i = 1, \ldots, n$. Therefore, the expected error can be defined as

$$\boldsymbol{y}_i - \tilde{\boldsymbol{y}}_i = \tau_2(\boldsymbol{y}_i)(\boldsymbol{y}_i - \tilde{\boldsymbol{\mu}}_i). \tag{9}$$

In practice, formula (9) is usually applied by using maximum likelihood estimates instead of the corresponding true data values.

1.2.1 Definition of Score Function

Hereinafter $\hat{p}$ denotes a maximum likelihood estimate of some parameter p.
Suppose one seeks to estimate a sum of the variable Y_j^*, $j = 1, \ldots, p$, with a sampling weight w_i of the ith sampled unit, – $T_j^* = \sum_{i=1}^n w_i y_{ij}^*$. A ratio between the expected error (9) with a sampling weight w_i multiplied by a suspicion component s_{ij} (probability that the ith sampled unit is erroneous) and target parameter estimate $\hat{T}_j = \sum_{i=1}^n w_i \hat{y}_{ij}$ denotes the conditional error of the ith sampled unit:

$$r_{ij} = \frac{s_{ij} w_i (y_{ij} - \hat{y}_{ij})}{\hat{T}_j}. \tag{10}$$

The local score function for the variable Y_j is denoted as $S_{ij} = |r_{ij}|$. Separate local scores can be combined into one global score GS_i in a few different ways: $GS_i = \max_j S_{ij}$ or $GS_i = \sum_j S_{ij}$. In order to identify an optimal number of observations to be edited, the corresponding sampled units are sorted descendingly according to the GS_i. First $\tilde{k}$ observations are then chosen for the editing procedure:

$$\tilde{k} = \min\{k^* \in 1, \ldots, n \mid \max_j R_{kj} < \eta, \ \forall\ k > k^*\} \tag{11}$$

where $R_{ij} = |\sum_{k \geq i}^n r_{kj}|$ with an accuracy level η.
The suspicion component s_{ij} can take on a discrete form ($s_{ij} \in \{0, 1\}$) and a continuous form ($s_{ij} \in [0, 1]$). In the paper the later continuous suspicion component is defined according to Norberg et al. (2010). An additional test variable should be defined prior to defining the suspicion component:

Definition 1 (Test variable) *Test variable* can be a combination of variables from a statistical survey and (or) additional information. Statistical errors can then be identified by checking whether a value of the test variable $\mathbf{t}_{j'}$, $j' = 1, \ldots, p'$, for the ith sampled unit falls into some chosen acceptance region $\left(\hat{t}_{ij'}^{(L)}, \hat{t}_{ij'}^{(U)}\right)$.

Definition 2 (Discrete suspicion component) *Discrete suspicion component* equals to 1 when a value of the jth survey variable of the ith sampled unit y_{ij} is a non-statistical error or a value of the jth test variable of the ith sampled unit $t_{ij'}$ is a statistical error ($t_{ij'} \notin \left(\hat{t}_{ij'}^{(L)}, \hat{t}_{ij'}^{(U)}\right)$). The later case gives $s_{ij} = 1$ for every survey variable y_{ij} that is a part of the combination $t_{ij'}$. Otherwise $s_{ij} = 0$.

Nonetheless, it is important to take into consideration different distances between observations that do not fall into the chosen acceptance region $\left(\hat{t}_{ij'}^{(L)}, \hat{t}_{ij'}^{(U)}\right)$ and the corresponding bound of the region. A continuous suspicion component should convey the information on the later distance more effectively.

Definition 3 (Continuous suspicion component) Hereinafter $\hat{t}_{ij'}$ denotes a prediction of the test variable $t_{ij'}$.

1) $s_{ij} = 1$ if a value of the jth survey variable of the ith sampled unit y_{ij} is a non-statistical error;

2) $\tilde{s}_{ij'} = \frac{\hat{t}_{ij'} - \kappa \cdot \left(\hat{t}_{ij'} - \hat{t}_{ij'}^{(L)}\right) - t_{ij'}}{\max\left\{\left(\hat{t}_{ij'}^{(U)} - \hat{t}_{ij'}^{(L)}\right),\ \alpha \cdot \hat{t}_{ij'}\right\}}$ if $t_{ij'} < \hat{t}_{ij'} - \kappa \cdot \left(\hat{t}_{ij'} - \hat{t}_{ij'}^{(L)}\right)$;

3) $\tilde{s}_{ij'} = \frac{t_{ij'} - \hat{t}_{ij'} - \kappa \cdot \left(\hat{t}_{ij'}^{(U)} - \hat{t}_{ij'}\right)}{\max\left\{\left(\hat{t}_{ij'}^{(U)} - \hat{t}_{ij'}^{(L)}\right),\ \alpha \cdot \hat{t}_{ij'}\right\}}$ if $t_{ij'} > \hat{t}_{ij'} + \kappa \cdot \left(\hat{t}_{ij'}^{(U)} - \hat{t}_{ij'}\right)$;

4) $\tilde{s}_{ij'} = 0$ if $\hat{t}_{ij'} - \kappa \cdot \left(\hat{t}_{ij'} - \hat{t}_{ij'}^{(L)}\right) < t_{ij'} < \hat{t}_{ij'} + \kappa \cdot \left(\hat{t}_{ij'}^{(U)} - \hat{t}_{ij'}\right)$.

Continuous suspicion component then equals to $s_{ij'} = \frac{\tilde{s}_{ij'}}{\tau + \tilde{s}_{ij'}}$ with parameters $\kappa \geq 0$, $\alpha > 0$ and $\tau > 0$. $s_{ij} = \max_{j'} s_{ij'}$ for every survey variable y_{ij} that is a part of the combination $t_{ij'}$.

2 Selective Editing Application on Statistical Survey Data

The outlier detection study was carried out using statistical data from the quarterly statistical survey on service enterprises of Statistics Lithuania. Enterprise turnover[1] of the accounting period was the target variable of the study. Predictor variables and the corresponding number of observations in data sets are given in Table 1.

Table 1: Number of Observations in Statistical Data Sets

Predictor variable	Number of observations (n)
Turnover from VAT declarations	4085
Turnover from the quarterly F-01 questionnaire	574
Average number of employees	4867
Total hours worked	4931

Before applying selective editing on statistical data it was important to ensure that all items for the target and predictor variables are not missing and greater than 0. Therefore, a number of observations in data sets (primary populations) varies according to the chosen predictor variable. In order to control the data contamination process, detected outliers in primary populations were replaced with contamination model predictions. The following procedure was then applied to every primary population:

1. Data were contaminated in 3 different ways:
   (a) 1,5 percent of observations were multiplied by 100,
   (b) 2 percent of observations were trimmed leaving only the first and the last digits,

[1] *Enterprise turnover* – enterprise income gained during the accounting period for sold goods and granted services. It does not include value-added tax (hereinafter referred to as VAT), income for long-term material assets, income for financial and investment activities, dividends, etc. (Official Statistics Portal, 2015).

(c) 20000000 was added to $1,5$ percent of observations;

2. Estimation of model coefficients and outlier (potential error) detection were carried out using the statistical programming language R and its package `SeleMix` (function `ml.est`);

3. Values of the target variable were sorted descendingly according to estimates of the global score function. An estimate of the global score function is close to 0 when a value of the target variable is not identified as an outlier and therefore has no major impact on the accuracy of the sample estimate, and greater than 0 when a value of the target variable is identified as an outlier;

4. The part of outliers that have a major impact on the accuracy of the sample estimate (influential errors) were chosen for the editing procedure.

The later influential error detection procedure was repeated in two different ways – by calculating estimates of the score function (1) with a discrete suspicion component that is the same among all observations ($s_i = 1$), and (2) with a continuous suspicion component. The later suspicion component was designed using an acceptance region between the first and the third quartiles $\left(\hat{t}^{(L)}, \hat{t}^{(U)}\right)$ where $\hat{t}_i = \hat{y}_i$ $(i = 1, \ldots, n)$, parameters κ and τ varies, $\alpha = 0,05$.
Selective editing with different accuracy levels gives a different number of influential errors. If all of the detected influential errors were introduced by the data contamination procedure, the corresponding accuracy level was chosen for the following study (see Table 2).

Table 2: Levels of Accuracy (Threshold Values) for Statistical Data Sets

Predictor variable	Level of accuracy
Turnover from VAT declarations	0.011
Turnover from the quarterly F-01 questionnaire	0.004
Average number of employees	0.027
Total hours worked	0.026

The results of selective editing were then compared by estimating the relative absolute bias after every edit of an influential error. This way a number of influential errors to be edited in order to achieve the desired accuracy of sample estimates was determined (see Table 3).

Table 3: Number of Influential Errors in Statistical Data Sets

Predictor variable	Total number of influential errors	Number of influential errors to be edited
(1) Selective editing with a discrete suspicion component		
Turnover from VAT declarations	134	92
Turnover from the quarterly F-01 questionnaire	23	14
Average number of employees	90	> 90
Total hours worked	111	> 111
(2) Selective editing with a continuous suspicion component		
Turnover from VAT declarations	93	92
Turnover from the quarterly F-01 questionnaire	15	14
Average number of employees	136	121
Total hours worked	124	123

It is important to note that selective editing with predictor variables such as average number

of employees and total hours worked gives a lower number of influential errors with a dicrete suspicion component compared to the case when a continuous suspicion component is used. Nonetheless, the chosen accuracy level is not achieved even after editing all of the identified influential errors. The main reason is a weak dependency between the target variable of the study and the corresponding predictor variables (correlation coefficient estimates are lower than 0.6). The later aspect causes greater differences between true values of the target variable and its contamination model predictions. Applications of selective editing with different predictor variables have shown an effectiveness of a continuous suspicion component on the outlier detection procedure as this approach to selective editing lets to identify a lower number and more important influential errors.

Conclusions

After calculations of the relative absolute bias dependency on the number of edited influential errors, selective editing with a continuous suspicion component was determined to be an optimal method of the outlier detection procedure. The later version of selective editing prevents from the unnecessary statistical data editing.
Turnover from VAT declarations and turnover from the quarterly F-01 questionnaire were identified as the most suitable predictor variables for the outlier detection procedure. The main property of a suitable predictor variable turned out to be a high correlation between the later predictor variable and the target variable of the study.

References

Di Zio, M., Guarnera, U. (2013) A Contamination Model for Selective Editing. *Journal of Official Statistics*, **29**(4), 539-555.

Lawrence, D., McDavitt, C. (1994) Significance Editing in the Australian Survey of Average Weekly Earnings. *Journal of Official Statistics*, **10**, 437-447.

Lawrence, D., McKenzie, R. (2000) The General Application of Significance Editing. *Journal of Official Statistics*, **16**, 243-253.

Norberg, A., Adolfsson, C., Arvidson, G., Gidlund, P., Nordberg, L. (2010) *A General Methodology for Selective Data Editing*. Statistics Sweden, Stockholm.

Official Statistics Portal (2015). PaslaugÅş ÄŕmoniÅş veiklos statistinio tyrimo metodika. Retrieved from `https://osp.stat.gov.lt/documents/10180/687662/Metodika_2012DI121.pdf`.

RDocumentation (2020). Functions in SeleMix (1.0.2). Retrieved from `https://rdocumentation.org/packages/SeleMix/versions/1.0.2`.

STATISTICAL DISCLOSURE CONTROL FOR CENSUS 2021

Ance Ceriņa[1] and Zane Matveja[2]

[1] Central Statistical Bureau of Latvia, Latvia
e-mail: ance.cerina@csp.gov.lv

[2] Central Statistical Bureau of Latvia, Latvia
e-mail: zane.matveja@csp.gov.lv

Abstract

To provide a harmonised Census data protection within the European Union (EU), Eurostat`s working group has developed common suggestion rules for the EU countries to implement. Data confidentiality has been named as one of the priorities of the working group. Census tables and hypercubes contain many cells with small values, which can disclose sensitive data, therefore, should be controlled. The working group has suggested two Statistical Disclosure Control (SDC) methods: Target record Swapping (TRS) for microdata and Cell Key Method (CKM) for tabular data. While CSB Latvia plans to use both methods for Census tables, so far only CKM has been tested and implemented. TRS will be tested and implemented for more detailed tables, which are not published yet.

As a part of the transition to the annual Censuses (from 2024), CSB Latvia has implemented Census 2021 tables and SDC methods into official Population statistics. This approach is helping to prevent confusion for data users by not publishing two different Population statistics tables with a similar content. By having the same approach for all Population statistics data tables, greater data harmonisation, consistency, and lower disclosure risk can be achieved.

SDC parameters should be determined to yield minimal loss of information while avoiding disclosure risks to a level that is considered acceptable. If possible, National Statistics Institutes should publish *unchanged* data to provide more precise data for data users. Since both SDC methods suggested by Eurostat`s working group involve changes in data, other SDC methods can be used, for instance, data categorisation in larger category groups. As is shown in the table below, data on the country of birth can be published with no categories, with some categorisation or with big categorisation groups. In addition, indicators can be published on a different geographical unit level – national, regional, urbanisation level, county, parish, or grid level. Depending on the indicator category level and geographical unit level, the table yields a different disclosure risk.

	Category level 1 - no categorisation	Category level 2 - some categorisation	Category level 3 - big categorisation groups
Country of birth	Latvia	Latvia	Latvia
	Estonia	Estonia	European Union country
	Lithuania	Lithuania	other countries
	Germany	Germany	
	United States of America	Spain	
	China	CIS countries	
	etc.	other countries	

CSB Latvia is implementing CKM only in tables, which contain detailed information on sensitive indicators. Furthermore, tables with sensitive indicators such as nationality, ethnicity, country of birth

are perturbated only in tables when they are published in small geographical units and contain little or no categorisation.
CMK has been tested on the table that contains these indicators such as age, sex, ethnicity, county, marital status and is published on a county level. CSB Latvia is using R software for CMK implementation.

Keywords: Social Disclosure Control, Census.

References

Antal, L., Enderle, T., Giessing, S. (2017) *Harmonised protection of census data in the ESS*, Deliverable D3.1 of Work Package 3 'Development and testing of recommendations; identification of best practices' within the Specific Grant Agreement 'Harmonised protection of census data in the ESS', URL: https://ec.europa.eu/eurostat/cros/system/files/methods_for_protecting_census_data.pdf

Giessing, S., Schulte Nordholt, E. (2017) *Harmonised protection of census data in the ESS*, Deliverable D3.3 of Work Package 3 'Development and testing of recommendations; identification of best practices' within the Specific Grant Agreement 'Harmonised protection of census data in the ESS', URL: https://ec.europa.eu/eurostat/cros/system/files/recommendations_for_the_protection_of_hypercubes.pdf

Giessing, S., Tent, R. (2019) *Concepts for generalising tools implementing the cell key method to the case of continuous variables*, Joint UNECE/ Eurostat Work Session on Statistical Data Confidentiality (The Hague, 29-31 October 2019), URL: https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2019/mtg1/SDC2019_S2_Germany_Giessing_Tent_AD.pdf

Meindl, B. (2020) *Introduction to the cellKey-Package*, URL: https://sdctools.github.io/cellKey/articles/introduction.html

ON DESIGN MEAN SQUARE ERROR ESTIMATION FOR MODEL-BASED SMALL AREA ESTIMATORS

Andrius Čiginas[1,2]

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

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

Abstract

Estimating the means or totals in domains of a survey population with small sample sizes, indirect model-based estimators are often more efficient than direct ones. In practice, it is important to have mean square error (MSE) estimators for the former estimation derived under a design-based approach, which is typical for direct estimation applied to domains with large samples. We consider the design MSE estimation for empirical best linear unbiased predictors based on the Fay–Herriot model. In this case, unbiased MSE estimators are known as unstable in the literature. We combine them with some biased but less variable estimators of the design MSEs and show the gain in the simulation study.

Keywords: conditional mean square error, area-level model, empirical best linear unbiased predictor, composite estimation.

1 Introduction and some results

We estimate the means of a survey variable in M sampled domains (areas) of a finite population. Let $\hat{\theta}_i^{\mathrm{d}}$ be a design-unbiased estimator of the mean θ_i in the ith area with $\mathrm{E}(\hat{\theta}_i^{\mathrm{d}} \,|\, \theta_i) = \theta_i$ and the sampling variance $\mathrm{var}(\hat{\theta}_i^{\mathrm{d}} \,|\, \theta_i) = {}_{\psi i}$ is assumed to be known. This variance can be large if the domain sample size is small.

Suppose that, for each area, the auxiliary information is available as a vector $\mathbf{z}_i$ of known characteristics, which are linearly associated with unknown parameter θ_i. Then, to improve the direct estimation, famous area-level Fay–Herriot model can be used to build the best linear unbiased predictors (Rao and Molina, 2015, Section 6.1.1)

$$\tilde{\theta}_i^{\mathrm{H}} = \gamma_i \hat{\theta}_i^{\mathrm{d}} + (1-\gamma_i)\mathbf{z}_i'\tilde{\boldsymbol{\beta}} \quad \text{with} \quad \gamma_i = \sigma_v^2/({}_{\psi i} + \sigma_v^2), \qquad i = 1, \dots, M, \tag{1}$$

and

$$\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}({}_{\psi i}, \sigma_v^2) = \left[\sum_{i=1}^{M} \mathbf{z}_i\mathbf{z}_i'/({}_{\psi i} + \sigma_v^2)\right]^{-1} \left[\sum_{i=1}^{M} \mathbf{z}_i\hat{\theta}_i^{\mathrm{d}}/({}_{\psi i} + \sigma_v^2)\right],$$

where σ_v^2 is the variance of random area effects, which is assumed to be known. Predictors (1) are the linear combinations of the direct estimators $\hat{\theta}_i^{\mathrm{d}}$ and the regression-synthetic estimators $\tilde{\theta}_i^{\mathrm{S}} := \mathbf{z}_i'\tilde{\boldsymbol{\beta}}$ with the weights γ_i.

Replacing σ_v^2 by an estimator $\hat{\sigma}_v^2$ in (1), we obtain empirical best linear unbiased predictors (EBLUPs) $\hat{\theta}_i^{\mathrm{H}}$ of the means θ_i, $i = 1, \dots, M$. In practice, the design variances ${}_{\psi i}$ are also unknown and therefore they are evaluated from external sources or by smoothing their direct estimates. Let us denote the evaluated variances by $\hat{{}_{\psi i}}^{\mathrm{s}}$.

Model MSE of EBLUPs $\hat{\theta}_i^{\mathrm{H}}$ is often used to measure the variability of the predictors. On the other hand, if the accuracy of the direct estimators is evaluated using the design MSE in domains with sufficiently large sample sizes, then it makes sense to use the same measure also for EBLUPs applied in the survey (Rao et al., 2018). However, estimation of the design (conditional) MSE

$$\mathrm{MSE}(\hat{\theta}_i^{\mathrm{H}}) = \mathrm{E}[(\hat{\theta}_i^{\mathrm{H}} - \theta_i)^2 \,|\, \theta_i] \tag{2}$$

is also a small area estimation problem because (approximately) design-unbiased estimators of (2) can be very unstable and take negative values for small sample sizes. It happens for the estimators of (2) proposed in Rivest and Belmonte (2000), Datta et al. (2011), and for elementary estimators considered in Pfeffermann and Ben-Hur (2019).

As an alternative to the unbiased estimators, one can use, according to Pfeffermann and Ben-Hur (2019), the naïve estimators

$$\mathrm{mse}_{\mathrm{n}}(\hat{\theta}_i^{\mathrm{H}}) = \hat{\gamma}_i^2 \hat{\psi}_i^{\mathrm{s}} + (1-\hat{\gamma}_i)^2 (\hat{\theta}_i^{\mathrm{H}} - \mathbf{z}_i' \tilde{\boldsymbol{\beta}}(\hat{\psi}_i^{\mathrm{s}}, \hat{\sigma}_v^2))^2, \qquad i = 1, \ldots, M, \tag{3}$$

of (2), where $\hat{\gamma}_i = \hat{\sigma}_v^2/(\hat{\psi}_i^{\mathrm{s}} + \hat{\sigma}_v^2)$. These estimators are biased but more stable and positive.

We propose another estimation of (2). We apply the results of Čiginas (2021) to design-based compositions (1) and then replace the unknown parameters by their empirical versions. First, we derive the estimator $\hat{\gamma}_i(1-\hat{\gamma}_i)\hat{\psi}_i^{\mathrm{s}}$ of the part of the squared bias of (2). Second, in line with the assumptions used in Čiginas (2021), we approximate $\mathrm{var}(\tilde{\theta}_i^{\mathrm{H}} \,|\, \theta_i) \approx \gamma_i^2 \psi_i + (1-\gamma_i)^2 \,\mathrm{var}(\tilde{\theta}_i^{\mathrm{S}} \,|\, \theta_i)$. Estimating this approximation and adding it to the estimated bias part, we arrive to

$$\mathrm{mse}_{\mathrm{b}}(\hat{\theta}_i^{\mathrm{H}}) = \hat{\gamma}_i \hat{\psi}_i^{\mathrm{s}} + (1-\hat{\gamma}_i)^2 \hat{\sigma}^2(\hat{\theta}_i^{\mathrm{S}}), \qquad i = 1, \ldots, M, \tag{4}$$

where $\hat{\sigma}^2(\hat{\theta}_i^{\mathrm{S}})$ denotes an estimator of the design variance $\mathrm{var}(\tilde{\theta}_i^{\mathrm{S}} \,|\, \theta_i)$.

The reference Rao et al. (2018) suggests linearly combine unbiased MSE estimators with biased ones like (3) using the estimated $\hat{\gamma}_i$ in the weighting. We compare some of that combinations numerically and present these results at the conference.

References

Čiginas, A. (2021) Design-based composite estimation rediscovered. `arXiv:2108.05052 [stat.ME]`.

Datta, G. S., Kubokawa, T., Molina, I., Rao, J. N. K. (2011) Estimation of mean squared error of model-based small area estimators. *Test*, **20**, 367–388.

Pfeffermann, D., Ben-Hur, D. (2019) Estimation of randomisation mean square error in small area estimation. *International Statistical Review*, **87**, 31–49.

Rao, J. N. K., Molina, I. (2015) *Small Area Estimation.* John Wiley, New Jersey.

Rao, J. N. K., Rubin-Bleuer, S., Estevao, V. M. (2018) Measuring uncertainty associated with model-based small area estimators. *Survey Methodology*, **44**, 151–166.

Rivest, L.- P., Belmonte, E. (2000) A conditional mean squared error of small area estimators. *Survey Methodology*, **26**, 67–78.

ANALYSIS OF EU-SILC DATA DEPENDING ON MODES OF DATA COLLECTION IN LATVIA

Darja Goreva[1] and Viktors Veretjanovs[2]

[1] Central Statistical Bureau of Latvia
e-mail: Darja.Goreva@csp.gov.lv

[2] Central Statistical Bureau of Latvia
e-mail: Viktors.Veretjanovs@csp.gov.lv

Abstract

Objective of the study is analysing the main indicators of survey "European Statistics on Income and Living Conditions (EU-SILC)" depending on data collection mode, particularly the impact of CAWI (i.e. analysis of non-response). The analysis is based on data of Latvian EU-SILC 2017, 2018, 2019 and 2020 surveys.

As in recent times demand for use of the latest data collection modes (CAWI) is increasing, there is a need for data quality analysis depending on collection mode. This issue became even more important during the Covid-19 crisis.

The analysis is prepared by experts from Central Statistical Bureau of Latvia.

Keywords: modes of data collection, EU-SILC survey.

References

Computer-assisted web interviewing.

Computer-assisted personal interviewing.

Computer-assisted telephone interviewing.

EU statistics on income and living conditions (EU-SILC) methodology.

Regulation (EU) 2019/1700 of the European Parliament and of the Council of 10 October 2019 establishing a common framework for European statistics relating to persons and households, based on data at individual level collected from samples, amending Regulations (EC) No 808/2004, (EC) No 452/2008 and (EC) No 1338/2008 of the European Parliament and of the Council, and repealing Regulation (EC) No 1177/2003 of the European Parliament and of the Council and Council Regulation (EC) No 577/98.

Commission Implementing Regulation (EU) 2019/2242 of 16 December 2019 specifying the technical items of data sets, establishing the technical formats and specifying the detailed arrangements and content of the quality reports on the organisation of a sample survey in the income and living conditions domain pursuant to Regulation (EU) 2019/1700 of the European Parliament and of the Council.

PROBLEMS OF SURVEY OF UNPAID ACTIVITIES

Alesia Korolenok

BSEU, Belarus
e-mail: Alesia_tar@mail.ru

Abstract

In the world statistical practice in recent years, much attention has been paid to the development of concepts, methodological standards for obtaining indicators of labor statistics in the context of forms of labor activity. In October 2013, the 19th ICLS (Resolution I) new standards were adopted for labor statistics, which are in line with the general production frontier formulated in the 2008 System of National Accounts (hereinafter - 2008 SNA). This creates the prerequisites for identifying and statistically assessing the volume of labor activity in the national production accounts, including existing satellite accounts, and measuring the contribution of all forms of labor activity to economic development, household income and the welfare of individuals and society as a whole.

The new conceptual framework for work statistics is fully aligned with the general production boundary of the 2008 SNA. Employment is a part of labor activity within the boundaries of the sphere of production. Also it includes the production of goods for own use; unpaid work of trainees and persons undergoing vocational technical training; labor activity of volunteers, as well as in households producing goods and other forms of labor activity. All these forms of labor activity form the basis for the preparation of national production accounts within the production boundaries according to the 2008 SNA. Providing services for own use complements the national production accountsю It is outside the boundaries of the sphere of production, but within the general boundary of the sphere of production.

Unpaid household activities are an important aspect of economic activity and an indispensable contributor to the well-being of individuals, their families and communities. At the same time, this activity is not taken into account in economic monitoring systems. However, disregard for unpaid work in cooking, caring for and teaching children, cleaning their own homes, etc. leads to incorrect conclusions in different areas of socio-economic analysis. Assessment of unpaid activities and their inclusion in national accounts is an urgent direction in the development of a macroeconomic accounting system.

In the Republic of Belarus, the main source of information for calculating these indicators is a sample survey of households to study the use of the daily time fund of population, which is periodically conducted by organizations of state statistics.

Keywords: system of national accounts, forms of employment, sample survey of households.

References

Руководство по стоимостной оценке неоплачиваемой трудовой деятельности по оказанию домашних услуг (2017) https://www.unece.org/fileadmin/DAM/stats/publications/2018/ECECESSTAT20173_ru.pdf

Resolution concerning statistics of work, employment and labour. 19th International Conference of Labour Statisticians (2013) http:// www.ilo.org/global/statistics-and-databases/meetings-and-events/international-conference-of-labour-statisticians/19/lang--en/index.htm

Руководство по подготовке статистических данных об использовании времени для оценки оплачиваемого и неоплачиваемого труда (2007). https://unstats.un.org/unsd/publication/SeriesF/SeriesF_93R.pdf

ПОДХОДЫ К ИЗУЧЕНИЮ НЕОПЛАЧИВАЕМОЙ ДЕЯТЕЛЬНОСТИ ДОМАШНИХ ХОЗЯЙСТВ

Алеся Королёнок

Белорусский государственный экономический университет, Республика Беларусь
e-mail: Alesia_tar@mail.ru

Аннотация

В мировой статистической практике в последние годы уделяется большое внимание разработке понятий, методологических стандартов для получения показателей статистики труда в разрезе форм трудовой деятельности. В октябре 2013 г. 19-я МКСТ (Резолюция I) приняла новые стандарты, касающиеся статистики трудовой деятельности, которые соответствует общей границе производственной деятельности, сформулированной в Системе национальных счетов 2008 г. (далее – СНС-2008). Это создает предпосылки для идентификации и статистической оценки объема трудовой деятельности в национальных счетах производства, включая существующие «сателлитные» счета, и измерения вклада всех форм трудовой деятельности в экономическое развитие, доходы домашних хозяйств и благосостояние отдельных лиц и всего обществ.

Новые концептуальные рамки статистики трудовой деятельности полностью совпадают с общей границей сферы производства СНС-2008. Занятость является частью трудовой деятельности в границах сферы производства, куда также входит производство товаров для собственного использования; неоплачиваемый труд стажеров и лиц, проходящих профессиональную техническую подготовку; трудовая деятельность волонтеров, а также в домашних хозяйствах, производящих товары и другие формы трудовой деятельности. Все эти формы трудовой деятельности формируют основу для подготовки национальных счетов производства в границах производственной деятельности согласно СНС-2008. Оказание услуг для собственного использования дополняет национальные счета производства, т.е. находится за пределами границы сферы производства, однако в пределах общей границы сферы производства.

Неоплачиваемая деятельность домашних хозяйств является важным аспектом экономической деятельности и незаменимым фактором, способствующим благополучию людей, их семей и общества. В тоже время эта деятельность во многом остается неучтенной в системах экономического мониторинга и, не смотря на обновленные международные стандарты в области статистики труда. Тем не менее, пренебрежение к неоплачиваемому труду по приготовлению пищи, уходу за детьми и их обучению, уборке собственного жилья и т.п. приводит к неверным выводам в различных областях социально-экономического анализа. Оценка неоплачиваемой деятельности, находящейся за пределами границы сферы производства в денежном выражении и включение её в национальные счета является актуальным направлением развития системы макроэкономического учета.

В Республике Беларусь основным катализатором измерения неоплачиваемой деятельности домашних хозяйств по оказанию услуг является разработка обследований использования времени, которые содержат данные, необходимые для определения масштабов и характера работы, выполняемой дома. Основным источником информации для расчета данных показателей является выборочное обследование домашних хозяйств по изучению использования суточного фонда времени, которое периодически проводится органами государственной статистики.

Ключевые слова: система национальных счетов, формы трудовой деятельности, выборочное обследование домашних хозяйств

Список используемых источников

Руководство по стоимостной оценке неоплачиваемой трудовой деятельности по оказанию домашних услуг (2017) https://www.unece.org/fileadmin/DAM/stats/publications/2018/ECECESSTAT20173_ru.pdf

Resolution concerning statistics of work, employment and labour. 19th International Conference of Labour Statisticians (2013) http:// www.ilo.org/global/statistics-and-databases/meetings-and-events/international-conference-of-labour-statisticians/19/lang--en/index.htm

Руководство по подготовке статистических данных об использовании времени для оценки оплачиваемого и неоплачиваемого труда (2007). https://unstats.un.org/unsd/publication/SeriesF/SeriesF_93R.pdf

HIGHLIGHTS OF THE WSC 2021 IN SURVEY STATISTICS

Danutė Krapavickaitė

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

Abstract

The 63rd World Statistics Congress was held on July 11-16, 2021, virtually. There were about 1600 participants from 104 countries, 770 authors submitted their abstracts/papers/posters. The number of presentations was higher than the number of papers.

Scientific program included 18 IASS supported invited sessions, among them two special invited sessions. There were more presentations devoted to the survey statistics. The most popular survey statistics topics discussed in the WSC:

- population census;
- imputation of missing data (Lee and Kim 2020);
- machine learning (classification example of the random forest usage and section on nonprobability sampling included in Valliant et al. 2018);
- population size estimation.

The main motif of the WSC presentations in survey statistics – data integration:

- macro-integration: time series of temporary employment - combining quarterly sample survey (LFS) and monthly employment register (ER) data;
- nonprobability samples; survey using Facebook sample data (Kreuter et al. 2020, Bradley et al. 2021);
- data integration by combining probability sample and nonprobability sample data, probability sample survey data and big data for finite population inference (Beaumont 2020, Beaumont and Rao 2021, Hill et al. 2020, Meng 2018, Tam and Holmberg 2020, Tam 2015, Rao 2020, Yang and Kim 2020);
- small area estimation in the case of probability sample and non-probability data set (Beaumont and Rao 2021, Rao 2020).

The Facebook data based sample survey carried out by a big team of statisticians from the Carnegie Mellon University (CMU), University of Maryland (UMD) and Facebook, US, will be discussed in more detail.

References

Beaumont, J.-F. (2020). Are probability surveys bound to disappear for the production of official statistics? *Survey Methodology*, **46**, 1-28.

Beaumont, J.-F., J. N. K. Rao. (2021) Pitfalls of making inferences from nonprobability samples: Can data integration through probability samples provide remedies? *The Survey Statistician*, **83**, 11-22.

Bradley, V. C., S. Kuriwaki, M. Isakov, D. Sejdinovic,X.-L. Meng, S. Flaxman. (2021). Are We There Yet? Big Data significantly overestimates COVID-19 Vaccination in the US. *MedRxiv, The preprint server for health sciences* (has not been peer-reviewed), https://doi.org/10.1101/2021.06.10.21258694; https://www.medrxiv.org/content/10.1101/2021.06.10.21258694v1.

Hill, C. A., P. Biemer, T. D. Buskirk, L. Japec, A. Kirchner, S. Kolenikov, L. E. Lyberg (editors). (2020). *Big Data meets survey science: A collection of innovative methods.* Wiley.

Kreuter, F. et al. (2020). Partnering with Facebook on a university-based rapid turn-around global survey. *Survey Research Methods.* **14**(2), 159-163. https://doi.org/10.18148/srm/2020.v14i2.7761.

Lee, D., J. K. Kim. (2020). Semiparametric imputation using conditional Gaussian mixture models under item nonresponse. *Biometrica*. https://doi.org/10.1111/biom.13410.

Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox and the 2016 US presidential election. *Annals of Applied Statistics*, **12**, 685-726. https://doi.org/10.1214/18-AOAS1161SF

Tam, S.-M., A. Holmberg. (2020). New Data Sources for Official Statistics – A Game Changer for Survey Statisticians? *The Survey Statistician,* **81**, 21-35.

Tam, S-M. (2015). A statistical framework for analyzing big data. *The Survey Statistician* **72**, 36-51.

Rao, J. N. K. (2021). On making valid inferences by integrating data from surveys and other sources, *Sankhyā B,* **83,** 242–272. https://doi.org/10.1007/s13571-020-00227-w.

Yang, S., J. K. Kim. (2020). Statistical data integration in survey sampling: a review. *Japanese Journal of Statistics and Data Science*, **3**(1), 625–650. https://doi.org/10.1007/s42081-020-00093-w.

Valliant**,** R., J. A. Dever, F. Kreuter. (2018). *Tools for designing and weighting survey samples*, Springer.

UNEQUAL PROBABILITY SAMPLING FOR THE EUROPEAN INTERVIEW HEALTH SURVEY IN LATVIA

Mārtiņš Liberts

Central Statistical Bureau of Latvia
e-mail: martins.liberts@csp.gov.lv

Abstract

A common requirement for a large scale sample survey is to deliver sufficiently precise estimates at population and domain level. Often study domains are with unequal size. Some of study domains can be much smaller then others. Those are contradicting requirements regarding the choice of an optimal sampling design to fulfil those requirements. One of the examples is the latest European Health Interview Survey which has been done in Latvia during 2019/2020. There are quality requirements at population level defined by the corresponding regulation. At the same time there are national requirements defined at domain level. An experimental sampling design with unequal sampling probabilities was proposed and implemented to fulfil those requirements.

Keywords: Unequal probability sampling, European Health Interview Survey.

1 European Health Interview Survey

The European Health Interview Survey 2019 (EHIS-2019) was organised in European Statistical System according to the Commission Regulation (EU) 2018/255 of 19 February 2018 (European Commission, Eurostat, 2018). The implementation of the survey is guided by the Methodological manual (European Union, Eurostat, 2020).

1.1 Precision Requirements of Eurostat

The Annex II of the regulation (European Commission, Eurostat, 2018) defines the precision requirements. This is a citation from the regulation:

1. Precision requirements for all data sets are expressed in standard errors and are defined as continuous functions of the actual estimates and of the size of the statistical population in a country.
2. The estimated standard error of a particular estimate $\hat{SE}(\hat{p})$ shall not be bigger than the following amount:

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{f(N)}}$$

3. The function $f(N)$ shall have the form of $f(N) = a\sqrt{N} + b$.
4. The following values for parameters N, a and b shall be used:

- $\hat{p}$: Percentage of population severely limited in usual activities because of health problems (age 15 years or over).
- N: Country population aged 15 years or over residing in private households, in million persons and rounded to 3 decimal digits.
- a: 1200
- b: 2800

1.2 National Survey and Precision Requirements

National survey and precision requirements were defined in the following form:

- Sample size: 11,000 persons.
- Two-stage sampling with geographical clustering should be used to optimise fieldwork cost where two main cost components are travelling expenses and time required for fieldwork operation.
- The main variables of interest:
  - General health self-assessment.
  - Health problems limiting activities.
  - Financial obstacles for receiving health care services.
  - Height and weight.
- The main population domains of interest:
  - Gender split by age groups (15-24, 25-34, 35-44, 45-54, 55-64, 65-74, 75+).
  - Economic activity (employed, unemployed, economically inactive).
  - Education (according ISCED: 0-1, 2, 3-4, 5-8).
  - Household net monthly equalised income quintile groups (five groups).
  - Region (NUTS-3, six regions).

2 Methodology

2.1 Sampling Frame

The sampling frame was created in three sequential steps.

The statistical register of dwellings and persons is a monthly updated statistical register maintained by the Central Statistical Bureau of Latvia. It contains information about all registered persons and inhabited dwellings in Latvia. The administrative data sources are used for updating the register. The main administrative data sources are the Population Register and the Address Register.

The general sampling frame of all registered residents living in private dwellings is a general sampling frame which can be used as an initial data source for any sampling frame for a sample survey where persons are sampled. The general frame is created with a standardized procedure. It is created whenever a sampling frame is necessary for any sample survey where persons are sampled. The data sources for creating the general sampling frame are the statistical register of dwellings and persons, the Address Register, phone lists, samples of other previous surveys, and other data sources.

The EHIS-2019 sampling frame is a frame which was used for sampling of persons for the EHIS-2019. It was created specifically for the needs of the EHIS-2019. For example,

population coverage was reduced to persons aged 15+, persons likely to be *de-facto* non-residents were excluded, extra variables for the needs of the EHIS-2019 were added. The data sources for creating the EHIS-2019 sampling frame were:

- the general sampling frame of all registered residents living in private dwellings (2019-06-27),
- micro data of the population statistics (2016 2019),
- administrative data about persons who have received state medical services in 2018,
- yearly income of persons in 2017,
- administrative data about self-employed persons in the first quarter of 2019,
- administrative data about employees and employers in May 2019,
- administrative data about registered unemployed persons in the first two quarters of 2019,
- the highest attained education level on 2019-01-01 from the Population Census data base.

2.2 Sample Size and Sampling Design

The total sample size for the EHIS-2019 was fixed to 11,000 persons. It was required to use two-stage sampling with geographical clustering of sampled persons to optimise fieldwork cost.

2.2.1 Sampling of Survey Areas

There were survey areas (*iecirki* in Latvian) available for sampling. The survey areas were created as compact geographical clusters of inhabited private dwellings with a purpose to be used for sample surveys where geographical clustering of sample units is necessary. The size of survey areas is measured by the number of inhabited private dwellings. The survey areas were created with similar size in urban and rural territories (300 for urban and 150 for rural territories). The survey areas were redesigned in 2019. The EHIS-2019 was the first survey to use the 'new' survey areas.

The survey areas were used as the first stage sample units. So, stratification for the first stage sample could be done using geographical information only. Stratification of persons (and sample areas) was done according to the declared living place of persons. There were five strata:

1. Persons with declared living place in Riga (the capital of Latvia).
2. Persons with declared living place in cities under state jurisdiction (eight cities excluding Riga).
3. Persons with declared living place in towns.
4. Persons with declared living place in parishes (rural areas).
5. Persons with cancelled or erroneous declared living place address (for those persons phone number was available to make the first contact over phone or to do data collection over phone using computer assisted telephone interview approach).

Sample allocation by strata was calculated according to the Neyman allocation (Neyman, 1934), where standard deviation was calculated according to the binary variable describing if person had received the state funded medical services in 2018. The sample size of the first four strata was rounded to the closest multiple of six (sample size of persons at the second stage for each PSU). The sample size for the 5th strata was calculated as a reminder (11,000 minus the total sample size of the strata 1–4). Sample allocation is available in Table 1 where:

- **strata**: strata identification
- **N**: frame population size
- **P**: proportion of frame persons who have received the state funded medical services in 2018
- **n_prop**: proportional sample allocation (only as a reference)
- **n_neim**: Neyman optimal sample allocation calculated using **N** and **S**
- **n_SSU**: Neyman optimal sample allocation rounded to the closest multiple of 6 (for the strata 1–4) and sample size in strata 5 is a reminder to the total sample size
- **n_PSU**: number of sampled PSUs
- **f**: sampling fraction

Table 1: Sample allocation by strata

strata	N	P	S	n_prop	n_neim	n_SSU	n_PSU	f
1	523 724	0.753	0.431	3 596	3 738	3738	623	0.007137
2	301 081	0.808	0.394	2 067	1 963	1962	327	0.006517
3	262 583	0.810	0.392	1 803	1 706	1704	284	0.006489
4	505 062	0.771	0.420	3 467	3 512	3516	586	0.006962
5	9 791	0.466	0.499	67	81	80	80	0.008171
Total	**1 602 241**			**11 000**	**11 000**	**11 000**	**1 900**	**0.006865**

Two-stage sampling was used for the first four strata. Single stage sampling was used for the 5th stratum (geographical clustering was not possible for persons with unknown declared living place).

The mentioned survey areas (clusters of persons) were used as the primary sampling units for strata 1-4. Survey areas were sampled in each stratum with systematic sampling with probabilities proportional to area size. The area size was calculated as number of persons available for sampling (there is a negative coordination with samples of other surveys with an aim to reduce the burden of respondents) associated to a respective area. Survey areas have been ordered in each stratum geographically so that contiguous areas in the survey area frame are also geographically close in space.

2.2.2 Sampling of Persons

The secondary sampling units in strata 1–4 and the primary sampling units in stratum 5 were persons. There were six persons sampled in each sampled area for strata 1–4. Sample size for the 5th stratum was 80 persons.

One of the survey requirements for national needs was to optimise the survey to produce reliable survey estimates for several population domains of interest. Those domains were defined as:

- gender and age groups (14 domains),
- economic activity status (3 domains),
- household income (5 domains),
- NUTS-3 regions (6 domains),

- highest achieved education level (4 domains).

It was possible to create those domains in the population frame according to the available external data sources (administrative data were used in most cases; exception is education level were different data sources were used including administrative, sample survey and the last census 2011 data).

Obviously the correspondence of those frame (*de jure*) domains with real (*de facto*) domains differ. For example, gender and age group domains in frame are almost 100 % equal to the real gender and age group domains. There is some level of misclassification errors for all other domains. However, it was assumed that those frame domains are good auxiliary information representing the main domains of interest. So, this information could be used to improve the precision of survey estimates in those target domains.

The population size for those domains differ quite a lot. For example the largest of those domains is persons with secondary education (corresponding to the ISCED 3 or ISCED 4). The size of this domain in the frame was 825,022 making 0.515 share of all frame persons. On the opposite side the smallest domain was unemployed persons. The size of this domain was 41,267 making only 0.026 share of all frame persons.

Assume we are using equal probability sampling. This approach would provide estimates with acceptable precision for large domains. For example, the expected sample size for persons with secondary education would be $11,000 \cdot 0.515 = 5664$, which should be enough to provide estimates with acceptable precision. However, the expected sample size for unemployed persons would be only $11,000 \cdot 0.026 = 283$. Taking non-response and over-coverage into account the expected net-sample size could be close to 155 which would not be enough to provide estimates with acceptable precision.

It would be necessary to over-sample small size domains while large size domains should be under-sampled to keep the total sample size fixed. Such approach would allow to improve the expected precision for estimates in small size domains.

Assume full response and equal variance in all domains, namely

$$S_d^2 = \frac{1}{N_d - 1} \sum_{i \in U_d} (y_i - \bar{y}_d)^2 = S, \text{ for } \forall d,$$

where:

- U_d is subset of target population belonging to domain d, $U_d \subset U$ where U is a set of units belonging to the target population and the size of U is constant.
- N_d is domain d population size,
- y_i is a value of a study variable for a population unit i,
- $\bar{y}_d = \frac{1}{N_d} \sum_{i \in U_d} y_i$.

Assume D non-overlapping domains covering U completely:

$$U_k \cap U_j = \emptyset \text{ for } \forall k, j$$

and

$$\bigcup_{d=1}^{D} U_d = U.$$

The optimal sample allocation in this case would be equal sized sample allocation by domains, namely $n_d = \frac{1}{D} n$, where n is the total sample size. Hence the optimal sampling probabilities would be

$$\pi_{i|d} = \frac{n}{DN_d},$$

where $\pi_{i|d}$ is a sampling probability for a population unit i under assumption $i \in U_d$. Those sampling probabilities provide equal sample size in all domains:

$$\sum_{i \in U_d} \pi_{i|d} = \sum_{i \in U_d} \frac{n}{DN_d} = \frac{n}{D} \text{ for } \forall d.$$

There are 32 target domains which are overlapping. So, this approach cannot be used directly. Those 32 domains can be ordered in five sets of domains where domains from one set are non-overlapping and covering U completely. Those domain sets are:

1. gender and age groups (14 domains),
2. economic activity status (3 domains),
3. household income (5 domains),
4. NUTS-3 regions (6 domains),
5. highest achieved education level (4 domains).

The exception are domains by education which do not cover frame population completely. There are persons with unknown education in a frame. Such domain exists only in a frame (because of missing information). However, such domain does not exist in a target population.

For each of those domain sets an optimal sampling probabilities were calculated as

$$\pi_{i|d_g} = \frac{n}{D_g N_{d_g}},$$

where d_g is a domain d from the domain set g, D_g is a number of domains in a domain set g, $\pi_{i|d_g}$ is a sampling probability for a person i under assumption $i \in U_{d_g}$.

Five sampling probabilities were calculated for each frame person according to each of five domain sets. Obviously those five sampling probabilities differ, so the final sampling probability for each frame person was calculated as an average of those five sampling probabilities:

$$\pi_i = \frac{1}{5} \sum_{g=1}^{5} p_{i|d_g}.$$

Those sampling probabilities π_i would be possible to use directly for a single stage sampling. In case of two stage sampling π_i cannot be used directly for sampling. But π_i were used as a "size measure" for persons to calculate second stage sampling probabilities proportional to π_i. For example:

- The lowest "size measure" was for an employed woman aged 55-64 living in Riga with the secondary education (ISCED 3–4). Those persons represent large size domains, so we want to sample those persons with low sampling probability.
- The highest "size measure" was for an unemployed woman aged 15-24 living in the Vidzeme region without primary education (ISCED 0–1). Those persons represent small size domains, so we want to sample those persons with high sampling probability.

2.2.3 Sampling Algorithm

The systematic sampling method (unequal probabilities, without replacement, fixed sample size) was used for both sampling stages. The function `UPsystematic` from the R (R Core Team, 2021) package `sampling` (Till & Matei, 2021) was used to implement the sampling method.

2.3 Weighting

The survey weights were calculated in three steps:

- Design weights.
- Non-response adjustment.
- Calibration of weights.

Design weights were calculated as inverse of the corresponding sampling probabilities.

Response probabilities were estimated using a logit model. Response probabilities were estimated only for the eligible persons (non-eligible persons were excluded from the estimation, persons with unknown eligibility status were assumed to be eligible). Model regressors were constructed using variables as:

- Type of territory (according to the declared living place): Riga (the capital city), cities under state jurisdiction (excluding Riga), towns, parishes,
- NUTS-3 region (according to the declared living place),
- Sex,
- Age group (15-24, 25-34, 35-44, 45-54, 55-64, 65-74, 75+),
- The highest achieved education level (four groups according to the ISCED 2011: 0-1, 2, 3-4, 5-8),
- Usage of the state funded medical services during the year 2018,
- Equalised yearly household income (2017, five quintile groups),
- Economic activity status (employed, unemployed, inactive).

There were 38 variables included in the response logit model. Non-response adjusted weight for each respondent was computed as design weight divided by the estimated response probability.

Only respondents were used in the weight calibration. Calibration variables were constructed using variables as:

- Type of territory (according to the declared living place): Riga (the capital city), cities under state jurisdiction (excluding Riga), towns and parishes,
- NUTS-3 region (according to the declared living place),
- Sex,
- Age group (15-24, 25-34, 35-44, 45-54, 55-64, 65-74, 75+),
- The highest achieved education level (three groups according to the ISCED 2011: 0-2, 3-4, 5-8).

There were 52 variables included in the calibration equation. Linear calibration from the R (R Core Team, 2021) package `surveyweighting` (Breidaks, 2020) was used.

Table 2: Survey outcome

No	Name	Non-weighted	Weighted
0	Sample size	11,000	1,357,679
1	Respondents	6,033	741,635
2	Non-respondents	4,636	574,336
3	Non-eligible	331	41,708
4	Over-coverage rate ([3] / [0])	0.030	0.031
5	Response rate ([1] / [1] + [2])	0.565	0.564

3 Results

3.1 Statistics

The statistics produced using the EHIS-2019 data are available at the Official Statistics Portal of Latvia (`https://stat.gov.lv/en`).

3.2 Non-sampling Errors

The total design weighted response rate was 0.564 and the total design weighted over-coverage rate was 0.031. See more details in Table 2.

3.3 Sampling Errors

The sampling error estimates for the main population parameter estimates are provided in Table 3. The description of the main population parameters:

- HS1: Proportion of persons aged 15+ in good or very good health
- HS2: Proportion of persons aged 15+ with longstanding illness or health problem
- HS3: Proportion of persons aged 15+ that were severely limited in activities people usually do because of health problems for at least past 6 months (this is the parameter used to define the Eurostat precision requirement)
- HO1: Proportion of persons aged 15+ having been hospitalized in the past 12 months
- BMI: Proportion of persons aged 18+ who are obese (BMI equal or above 30, where BMI (body mass index) is calculated as weight in kg divided by height in meters squared)

The Eurostat precision requirements defined at the regulation (European Commission, Eurostat, 2018) were verified. The precision requirement was: *the standard error estimate for the estimated proportion of severely limited in usual activities because of health problems (age 15 years or over, HS3) shall not be bigger than*

$$\sqrt{\frac{\hat{p}\,(1-\hat{p})}{f\,(N)}}$$

where:

- $f\,(N) = a\sqrt{N}+b$ is the minimum effective sample size necessary to fulfil the corresponding precision requirements.

Table 3: Estimates of the main parameters of interest with respective precision measures

variable	gender	respondents	estimate	SE	confidence interval	deff
HS1	All	5848	0.496	0.004216	0.487 — 0.504	0.418
HS1	Women	3420	0.454	0.005353	0.443 — 0.464	0.375
HS1	Men	2428	0.548	0.006679	0.535 — 0.561	0.475
HS2	All	6025	0.732	0.003984	0.725 — 0.740	0.483
HS2	Women	3495	0.775	0.004845	0.766 — 0.785	0.444
HS2	Men	2530	0.680	0.006468	0.667 — 0.692	0.515
HS3	All	6023	0.091	0.002581	0.086 — 0.096	0.501
HS3	Women	3492	0.106	0.003533	0.099 — 0.113	0.446
HS3	Men	2531	0.072	0.003516	0.065 — 0.079	0.524
HO1	All	6024	0.115	0.003046	0.109 — 0.121	0.558
HO1	Women	3495	0.123	0.004182	0.115 — 0.131	0.545
HO1	Men	2529	0.106	0.004582	0.097 — 0.115	0.616
BMI	All	5528	0.230	0.004014	0.222 — 0.238	0.523
BMI	Women	3265	0.257	0.005278	0.247 — 0.268	0.466
BMI	Men	2263	0.196	0.006105	0.184 — 0.208	0.603

- $\hat{p}$: Percentage of population severely limited in usual activities because of health problems (age 15 years or over). The estimated proportion was 0.091 for Latvia (see the line HS3 "All" in Table 3).
- N: Country population aged 15 years or over residing in private households, in million persons and rounded to 3 decimal digits. It was 1.585 million for Latvia.
- a: 1200
- b: 2800

The threshold value for the estimated standard error is equal to

$$\sqrt{\frac{\hat{p}\,(1-\hat{p})}{f\,(N)}} = \sqrt{\frac{\hat{p}\,(1-\hat{p})}{a\sqrt{N}+b}} = \sqrt{\frac{0.091\,(1-0.091)}{1200\cdot\sqrt{1.585}+2800}} = 0.004373.$$

We can see the estimated standard error for the respective population parameter estimate was 0.002581 (see the line HS3 "All" in Table 3) which is lower then the threshold value (0.004373). We can conclude that precision requirements defined by the regulation (European Commission, Eurostat, 2018) have been fulfilled for the EHIS-2019 in Latvia.

4 Conclusions

The paper presents an empirical work with an implementing of an unequal probability sampling for the European Interview Health Survey in Latvia. The aim of this approach was to over-sample target domains with small population size. It was expected to provide an optimal sampling design to fulfil national and European precision requirements.

The overall precision requirements defined by the regulation (European Commission, Eurostat, 2018) have been satisfied. The precision of the estimates of other main population parameters are good in general.

The precision of domain estimates have not been derived yet (exception is gender). Hopefully some of those results will be available for presenting at the Summer School on Survey Statistics 2021.

Acknowledgement

Author has used some of the material which he has provided to Eurostat for the quality report of the European Interview Health Survey 2019. The EHIS-2019 quality report has not been published yet.

References

Breidaks, J. (2020). surveyweighting: Survey weighting [Computer software manual]. Retrieved from `https://github.com/CSBLatvia/surveyweighting` (R package version 0.7)

European Commission, Eurostat. (2018). *Commission regulation (EU) 2018/255.* Retrieved from `http://data.europa.eu/eli/reg/2018/255/oj`

European Union, Eurostat. (2020). *European health interview survey (EHIS wave 3) methodological manual (re-edition 2020).* Retrieved from `https://ec.europa.eu/eurostat/` (DOI: 10.2785/135920)

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, *97*(4), 558–625. Retrieved from `http://www.jstor.org/stable/2342192`

R Core Team. (2021). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Till, Y., & Matei, A. (2021). sampling: Survey sampling [Computer software manual]. Retrieved from `https://cran.r-project.org/package=sampling` (R package version 2.9)

A COMPARISON OF URL FINDERS FOR ONLINE-BASED ENTERPRISE CHARACTERISTICS

Vilma Nekrašaitė-Liegė

Vilnius Gediminas technical University, Statistics Lithuania, Lithuania
e-mail: vilma.nekrasaite-liege@vilniustech.lt

Abstract

One of the fields where the integration of big data in the regular production of official statistics might be possible is Online-based Enterprise Characteristics. Currently two URL finder programs are suggested by ESSnet, thus overview and comparison of these two URL finders will be presented here.

Keywords: Online-based Enterprise Characteristics, URL finder.

1 The web as a statistics data source

Web scraping is easy, however if you want to use it as a statistics data source, you want it to be automated, methodologically sound, transparent, robust, consistent and efficient. For this reason the ESSnet Web Intelligence Network project was created and Statistics Lithuania is a member of it. The project goal is to contribute to establishing the Web Intelligence Network (WIN) across the ESS and to make use of the Web Intelligence Hub (WIH) services for the production of statistics with web data. This project started in 2021 and is going to last 4 years. Currently, there are two main fields (Online Job Advertisements (OJA) and Online-based Enterprise Characteristics (OBEC)) where the initial steps are made. This article will focus on the work done in the OBEC field.

The use of OBEC data would support the official statistics with more recent data, it would improve the Statistical Business Register (SBR) and could be used as additional information in Information and Communication Technology (ICT) surveys.

The first OBEC goal is to create a database containing URLs for each enterprise in the target population, where there can be one, many or zero URLs for a given enterprise. For some enterprises the URLs might already be available in a SBR or obtained from other sources. It can also be built up from scratch by searching for enterprises via search engines like Bing or Google. Of course, web scraping can be used as a verification tool. Thus, the search results can help to answer two main questions:

- Does an enterprise have a website?
- Which URL is most likely to belong to that enterprise?

Commonly, a web search will return several results leading to different base URLs for one enterprise. The different machine learning methods and algorithms like logistic regression or random trees can be used to identify a valid URL.

The previous projects (ESSnet Big Data I and ESSnet Big Data II) also investigated this field and two URL finder softwares were created:

- **UrlSearher**: `https://github.com/SummaIstat/UrlSearcher`

- **URLsFinder**: `https://github.com/EnterpriseCharacteristicsESSnetBigData/StarterKit/tree/master/URLsFinder`

A more detailed analysis of these programs is provided in the next section.

2 Comparison of URL finders

UrlSearher was created by Donato Summa and his team (Italian National Institute of Statistics) (Barcaroli G., Scannapieco M., Summa D. (2016)). UrlSearcher is a Java application where a strategy for solving the URL retrieval problem is adopted. It consists of 5 steps:

- *Step 1: Building the input training dataset.* Combining different sources the list of enterprises with several indicators (enterprise name, city, telephone number and ect.) is created. This step must be done outside the URL finder, because each country can use different sources and different indicators and there is no possibility to automate this process.
- *Step 2: URLs Searching.* In this step for each unit in the input training dataset the first 10 URLs were stored from the search engine, where the search was done using the enterprise name.
- *Step 3: URLs Crawling.* For each row of the seed file, if the URL is not in the list of the domains to filter out, the program tries to acquire the HTML content of the page. From each acquired HTML page the program extracts just the textual content of the HTML fields with useful information (for example, contact information) and write a line in a TSV file.
- *Step 4: URLs Scoring.* A score vector is computed and a score is assigned for each line in the TSV file. The elements/characteristics that were considered in a score vector by default (it is possible to adapt it to each country) are these:
  - Simple URL (is the URL in the form www.name.lt or not?);
  - VAT (is it present in the page or not?);
  - city (is it present in the page or not?);
  - province code (is it present in the page or not?);
  - link position (from 0 to 9);
  - telephone number (is it present in the page or not?);
  - zip code (is it present in the page or not?).

  A score is calculated as a sum of assigned points for each element/characteristic.
- *Step 5: Using a Machine Learning approach to associate URLs to enterprises.* The easiest way to assign the valid URL for each unit is to select that with the maximum score, but knowing that not all units have a URL, the more precise algorithm must be used. That is why in this step three methods (neural networks, random forest and logistic model) are used to determine if the URL with the highest score is valid or not.

The other program **URLsFinder** was created by Kostadin Georgiev (Bulgarian National Statistical Institute) and is a part of a Starter Kit package. The URLsFinder is written in Python and it contains two main modules:

- *URLsFinderWS* - defines methods for scraping information for the enterprises' URLs from the internet with the help of search engine Duck Duck Go.

- *URLsFinderMLLR* - defines methods for determining the enterprises' URLs or characteristics from the scraped information from the internet by using logistic regression machine learning.

As the UrlSearher, the URLsFinder has a similar course of action, still there are some differences, which are presented in table 1.

Table 1: A comparison of URL finders

	UrlSearher	StarterKit
Language	Java	Python
Search engine	Bing	Duck Duck Go
Characteristics included in a score vector (by default)	Simple URL VAT city province code link position telephone number zip code	Simple URL ID city address link position telephone number name equal domain
Machine Learning methods	neural network random forest logistic	logistic

More detailed comparison and adaptation to Statistics Lithuania needs will be presented during the presentation.

3 Some observations

ESSnet WIN project is still at the early stage, thus the main results will be obtain in the future, however, some observations regarding OBEC field can be already made:

- Even if we agree that the web scraping is a powerful tool to obtain the information, still at this moment it won't change the traditional survey sampling, but it can provide useful up-to-date additional information, which could be integrated in the survey sampling procedures.
- It is necessary to define country specific steps and stages for collecting the data, thus the programs must be easily updated.
- To validate that suggested URL is correct the machine learning methods are used, where there is a need to have a train set. Unfortunately not always there is a possibility to construct an appropriate train set.

References

Barcaroli G., Scannapieco M., Summa D. (2016) On the use of internet as a data sourse for official statistics: a strategy for identifying enterprises on the web. *Rivista Italiana di Economia Demografia e Statistica*, **LXX**, 25-41.

ESSnet Big Data I. WP2 led by Monica Scannapieco/ISTAT (OBEC) `https://ec.europa.eu/eurostat/cros/content/wp2-webscraping-enterprise-characteristics_en`

ESSnet Big Data II. WPC led by Galia Stateva/BNSI (OBEC) `https://ec.europa.eu/eurostat/cros/content/WPC_Enterprise_characteristics_en`

STATISTICAL EDITING AND IMPUTATION OF MISSING VALUES FOR THE POPULATION CENSUS 2021 IN LATVIA

Ruāna Pavasare

Central Statistical Bureau of Latvia, Latvia
e-mail: Ruana.Pavasare@csp.gov.lv

Abstract

Although population census 2021 will be register-based, it is not possible to obtain all required information about all residents of Latvia from administrative data. For this reason, it is necessary to apply the statistical imputation and editing methods for various census variables. The development of the census imputation and editing methodology has been underway since 2016. Imputation and editing methods are evaluated and improved every year. Two methods are currently used: *k*-nearest neighbours and the classification tree method (*rpart*).
The classification tree method is used to edit the status of economic activity because it is not possible to determine unregistered employment and unemployment from administrative data. Labour Force Survey data is used as training and benchmark data for this task. The classification tree method is used also to impute the status of employment and the location of the workplace (Latvia or abroad). The imputation of occupation, industry and education variables is carried out using the *k*-nearest neighbours method.

Keywords: census, imputation, editing.

References

Hasler, Caren, and Yves Tillé. ”Balanced *k*-nearest neighbor imputation.” arXiv preprint arXiv:1501.07622 (2015) https://arxiv.org/pdf/1501.07622.pdf

Terry M. Therneau, Elizabeth J. Atkinson. ”An introduction to Recursive Partitioning Using the RPART Routines.” https://stat.ethz.ch/R-manual/R-patched/library/rpart/doc/longintro.pdf

HOUSEHOLDS SURVEY TO MEASURE THE AGRICULTURAL ACTIVITY

Natalia Pekarskaya

BSEU, Belarus
e-mail: npekarskaya@list.ru

Abstract

Current agricultural statistics in Belarus provides full coverage of the main producers (agricultural organizations and farmer households), which account for about 78% of total agricultural production. The obtaining of statistical data on agricultural activities of personal subsidiary plots of citizens, permanently residing in rural areas, since 2011 is carried out on the basis of a sample survey.

However, it was not possible to obtain information on agricultural activities of other population from sample surveys. Also uncovered by the sample survey are households that carry out agricultural activities in urban areas, in garden and dacha associations, and vegetable gardening cooperatives.

The source of all information on agricultural production is the agricultural census. According to the recommendations of the Food and Agriculture Organization of the United Nations (FAO), agricultural censuses should be conducted once every 10 years.

During the recent ten-year period, an agricultural census has been carried out in almost all CIS countries. An event of this kind and scale was not held in the Republic of Belarus until 2019.

The organization and conduct of an agricultural census requires significant labor and financial resources, therefore, an agricultural census as part of the population census would avoid additional budget expenditures and ensure a complete coverage of households engaged in agricultural activities.

In this regard, an agricultural census was carried out as part of the 2019 population census. Thus, for the first time in Belarus, complete information on agricultural activities of the population has been obtained. In the course of this survey, information was obtained that will be used in the future to conduct various types of sample surveys and one-time accounting in agriculture; detailed data was obtained on the state of development of personal subsidiary plots of citizens both in the country as a whole, and in individual regions and territories.

Keywords: agricultural census, sample household survey, population census.

References

Agriculture: statistical collection, 2021 / National Statistical Committee of the Republic of Belarus, 2021 http://www.belstat.gov.by

FAO. 2020. FAO at 75 – Grow, nourish, sustain. Together. Rome. https://doi.org/10.4060/cb1182en

ОБСЛЕДОВАНИЕ ДОМАШНИХ ХОЗЯЙСТВ ДЛЯ ОЦЕНКИ ОБЪЕМОВ СЕЛЬСКОХОЗЯЙСТВЕННОЙ ДЕЯТЕЛЬНОСТИ

Natalia Pekarskaya

БГЭУ, Беларусь
e-mail: npekarskaya@list.ru

Аннотация

Ведение текущей статистики сельского хозяйства в Беларуси обеспечивает полный охват основных товаропроизводителей (сельскохозяйственные организации и крестьянские (фермерские) хозяйства), на долю которых приходится около 78% общего объема производства сельскохозяйственной продукции. А сбор статистических данных о сельскохозяйственной деятельности личных подсобных хозяйств граждан**,** постоянно проживающих в сельской местности, начиная с 2011 г., проводится на основе выборочного обследования.

Однако получить информацию о сельскохозяйственной деятельности иного населения из выборочных обследований не представлялось возможным. Так, не охвачены выборочным обследованием хозяйства населения, осуществляющие сельскохозяйственную деятельность в городской местности, в садовых и дачных товариществах, огороднических кооперативах.

Источником всей информации о сельскохозяйственном производстве выступает сельскохозяйственная перепись. Согласно рекомендациям Продовольственной и сельскохозяйственной организации Объединенных наций (ФАО) сельскохозяйственные переписи должны проводиться один раз в 10 лет.

За последние десять лет сельскохозяйственная перепись была проведена практически во всех странах СНГ. В Республике Беларусь мероприятие подобного рода и масштаба до 2019 г. не проводилось.

Для организации и проведения сельскохозяйственной переписи требуются значительные трудовые и финансовые ресурсы, поэтому проведение сельскохозяйственной переписи в рамках переписи населения позволило бы избежать дополнительных расходов бюджета и обеспечить сплошной охват хозяйств населения, осуществляющих сельскохозяйственную деятельность.

В связи с этим в рамках переписи населения 2019 года проведена и сельскохозяйственная перепись. Таким образом, в Беларуси впервые получена полная информация о сельскохозяйственной деятельности населения. В ходе данного обследования получена информация, которая в дальнейшем будет использована для проведения различного рода выборочных обследований и единовременных учетов в сельском хозяйстве; получены детализированные данные о состоянии развития личных подсобных хозяйств граждан как в целом по стране, так и по отдельным регионам и территориям.

Ключевые слова: сельскохозяйственная перепись, выборочное обследование домашних хозяйств, перепись населения.

Список использованных источников

Agriculture: statistical collection, 2021 / National Statistical Committee of the Republic of Belarus, 2021 http://www.belstat.gov.by

FAO. 2020. FAO at 75 – Grow, nourish, sustain. Together. Rome. https://doi.org/10.4060/cb1182en

THE DISPLAY OF CORONA INCIDENCES IN SPACE AND TIME

Ulrich Rendtel[1], Marcus Gross[2], Andrea Neugebauer[3], Lukas Fuchs[4] and Jingying Shang[5]

[1] FB Wirtschaftswissenschaft, Freie Universität Berlin, Germany
e-mail: ulrich.rendtel@fu-berlin.de

[2] INWT Statistics GmbH, Berlin, Germany
e-mail: Marcus.gross@inwt-statistics.de

[3] INWT Statistics GmbH, Berlin, Germany
e-mail: Anrdreas.neugebauer@inwt-statistics.de

[4] Joint Berlin Master Program Statistics, Berlin, Germany
e-mail: Lukas.fuchs@student.hu-berlin.de

[5] Joint Berlin Master Program Statistics, Berlin, Germany
e-mail: shanjing@hu-berlin.de

Abstract

The representation of the spacial and temporal dispersion of the Corona pandemic is a key issue of epidemiologic research but also of public media. This issue is often realized via maps which are often animated. The web-application which is presented here (https://www.inwt-statistics.com/read-blog/covid-19_heat-map_of-local_7-day_incidences_over_time.html) uses an alternative statistical concept for the display of Corona incidences. Instead of the standard assumption of a uniform distribution over the reference area we use the approach of Gross et al. (2020). The gain of this approach is the joint analysis of neighboring areas.

This general statistical approach is applied here for the estimation of local Corona incidences in Germany. The approach avoids the discontinuities at the borderlines of counties which appear in standard maps by a joint analysis of neighboring counties. The focus of the presentation is the realization of this concept by a web-application and its use. By three examples we demonstrate that during the second Corona wave there exist in Germany fixed local clusters which may broaden over time and which may also merge.

Keywords: Corona Incidence, Internet Maps, Choropleths, Kernel density estimation, Simulated EM-Algorithm.

References

Groß M, Kreutzmann A-K, Rendtel U, Schmid T, Tzavidis N (2020): Switching between different area systems via simulated geo-coordinates: a case study for student residents in Berlin. J Off Stat 36:297–314. https://doi.org/10.2478/JOS-2020-0016

Rendtel, U.; Neudecker, A.; Fuchs, L. (2021): Die Darstellung von Inzidenzgebieten
mit simulierten Geokoordinaten. (The display of incidence areas by simulated geo-coordinates. In German) AStA Wirtschafts- und Sozialstatistisches Archiv, 15, Online under https://doi.org/10.1007/s11943-021-00288-x

OPTIMAL IDENTIFICATION OF AUXILIARY VARIABLES IN SAMPLE SURVEYS TO REDUCE NONRESPONSE BIAS

Liliāna Roze

Central Statistical Bureau of Latvia, Latvia
e-mail: Liliana.Roze@csp.gov.lv

Abstract

One of the tasks in survey statistics is to reduce the bias of the sample survey estimates. Nonresponse in sample surveys can lead to biased survey estimates. One of the solutions to deal with the nonresponse bias is the usage of a calibration estimation with auxiliary information. The question that is being discussed in the paper is how to select auxiliary variables for optimal bias reduction. The main goal of the paper is to calculate and interpret nonresponse bias indicators in practical application for two sample surveys.

Keywords: Nonresponse, auxiliary variables, calibration estimator, bias indicators, optimal auxiliary variables.

1 Introduction

The need for accurate data and statistics is only growing. However, 100% data accuracy is impossible. There will be missing or useless data because of the nonresponse.

There are many methods to reduce nonresponse bias after a completed survey. This paper is mainly about auxiliary information. But of course, to identify optimal auxiliary information, the user needs to have knowledge about calibration factors and indicators, that can determine that.

2 Theory

Before approaching the theory about indicators. Worth mentioning some information on nonresponse, study and auxiliary variables, calibration estimator, and factor.

2.1 Nonresponse

One of the problems in sample surveys that make bias *(bias = real value - expected value*) bigger is nonresponse (Holton, 2014). Nonresponse occurs when the respondent is not reachable, respondents do not want to answer specific questions, or the survey is damaged somehow (Särndal et al., 1992).

So if we have probability sample *s* from a population $U = \{1,2, \dots, k, \dots, N\}$, and *r* is the respondent data set, that means that the nonresponse set is $nr = s - r$. If $nr = \emptyset$ is empty, that means that all sample units are respondents. It is a perfect situation. Sadly, in practice, it is impossible to get (Särndal et al., 2009).

The main goal is to reduce bias which is a result of nonresponse. To do that, apply a calibration estimator, which is dependent on study and auxiliary variables (Särndal et al., 2009).

2.2 Study and auxiliary variables

Often there are many variables in sample surveys. Attention is on the study and auxiliary variables. The study variable is also known as the research variable - it is a variable that we want to explore or estimate. We can explore many or just one study variable. Study variable value to unit *k* denoted as y_k (Särndal et al., 2009).

On the other hand, the auxiliary variable is a variable that is not the supreme exploring variable but can improve the study variable estimated value. Auxiliary variable value to unit k will be denoted as x_k. But the set of auxiliary variables - $\boldsymbol{x_k}$ (Särndal et al., 2009).

Both (study, auxiliary) variables if k unit belongs to sample, is denoted as $k \in s$. If $k \in r$, it means that all units are in response set (Särndal et al., 2009).

In the previous subsection, it was mentioned that study and auxiliary variables affect calibration estimator and factor.

2.3 Calibration estimator and factor

A calibration estimator is adjusted, to get an estimated value closer to the real value in the case of nonresponse. Estimator requires a calibration factor, which is dependent on design weights and auxiliary variables x_k. The design weight formula is (1). Where $\pi_k > 0$, π_k is the probability of unit k to obtain into the sample, when $k \in U$ (Särndal et al., 2009).

$$d_k = {}^{1}/_{\pi_k} \tag{1}$$

The calibration factor is in the formula (2). The symbol $'$ means transpose matrix.

$$m_k = (\sum_{k\in s} d_k \boldsymbol{x_k})'(\sum_{k\in r} d_k \boldsymbol{x}_k \boldsymbol{x'}_k)^{-1} \boldsymbol{x}_k \tag{2}$$

Also, the equation $\sum_{k\in r} d_k\, m_k \boldsymbol{x_k} = \sum_{k\in s} d_k\, \boldsymbol{x_k}$ is true. This equation was used in practice to check if the calibrations factor calculation is correct. The calibration estimator is visualized in (3). Exactly (3) formula, because it is more focused on reducing nonresponse bias than the variance (Särndal et al., 2009).

$$\tilde{Y}_{CAL} = \sum_{k\in r} d_k\, m_k y_k \tag{3}$$

To calculate the indicator, which shows us the optimal set of auxiliary variables, we will use the calibration factor m_k.

2.4 Indicators

The stepwise forward procedure is used to find an optimal set of auxiliary variables (stepwise backward procedure can be used also). Then the first set of auxiliary variables is $\boldsymbol{x_k} = x_{1k}$. The second will be $\boldsymbol{x_k} = (x_{1k},\, x_{2k})$ and the third set - $\boldsymbol{x_k} = (x_{1k},\, x_{2k}, x_{3k})$. The algorithm repeats till all possible auxiliary variables are in the set of auxiliary variables $\boldsymbol{x_k} = (x_{1k},\, x_{2k}, \dots, x_{lk})$ (Särndal et al., 2009).

Accordingly, to (Särndal et al., 2009), there are two indicators. First is H_1 indicator (4) is suitable when there is only one study variable.

$$H_1 = |R_{y;m}| \times cv_m \tag{4}$$

The second indicator is H_3, (5) is suitable when there are many variables, starting with two. Both indicators maximized value detects an optimal set of auxiliary variables.

$$H_3 = cv_m \tag{5}$$

Where cv_m is the coefficient of variation, of the calibration factor m_k.

$$cv_m = \frac{S_m}{\bar{m}_{r;d}} = \sqrt{\frac{\bar{m}_{s;d}}{\bar{m}_{r;d}} - 1} \tag{6}$$

Where S_m is the weighted standard deviation of the calibration factor m_k.

$$S_m^2 = \bar{m}_{r;d}\ (\bar{m}_{s;d} - \bar{m}_{r;d}) \quad (7)$$

Where $\bar{m}_{r;d}$ is the weighted mean value of the calibration factor m_k, $k \in r$. Besides $\bar{m}_{s;d}$ is the weighted mean of the calibration factor $m_k, k \in s$.

$$\bar{m}_{r;d} = \frac{\sum_{k\in r} d_k m_k}{\sum_{k\in r} d_k} = \frac{\sum_{k\in s} d_k}{\sum_{k\in r} d_k} \quad (8)$$

$$\bar{m}_{s;d} = \frac{\sum_{k\in s} d_k m_k}{\sum_{k\in s} d_k} \quad (9)$$

Returning to the formula (4). $R_{y;m}$ is the correlation coefficient depending on the set of study variables y_k and calibration factor m_k. $R_{y;m}$ satisfies $-1 \le R_{y;m} \le 1$ inequality.

$$R_{y;m} = Cov(y,m)/S_y S_m \quad (10)$$

Where $Cov(y,m)$ is the covariance coefficient depending on the set of study variables y_k and the calibration factor m_k.

$$Cov(y,m) = Cov(y,m)_{r;d} = \frac{1}{\sum_{k\in r} d_k} \sum_{k\in r} d_k\,(m_k - \bar{m}_{r;d})(y_k - \bar{y}_{r;d}) \quad (11)$$

In the formula (10) (Särndal et al., 2009), there is also S_y that is the standard deviation of the set of study variables y_k.

$$S_y^2 = S_{y|r;d}^2 = \frac{1}{\sum_{k\in r} d_k} \sum_{k\in r} d_k (y_k - \bar{y}_{r;d})^2 \quad (12)$$

Where $\bar{y}_{r;d} = \sum_{k\in r} d_k y_k / \sum_{k\in r} d_k$ is the weighted mean of the set of study variables, $k \in r$ (Särndal et al., 2009).

3 Indicators in practice

The programming language R was used in practice to compute calibration factors, indicators, and other estimated values.

Two different sample surveys were reviewed in practice –

- 2020 sample survey of turnover retail sales in Latvia. The period of this survey is a month (data of January). The response rate to this specific sample survey is 0.856. There is only one main study variable for this sample survey. Along with theory, an optimal set of auxiliary variables can find out with H_1 indicator formula (4) (Särndal et al., 2009).

Table 1 shows that the optimal set of auxiliary variables can be created from more than one variable. Four of them have a maximal H_1 indicator value of 0.0738. However, the choice of three sets of auxiliary variables is unnecessary complexity in this case, so an optimal set of auxiliary variables for this survey is chosen as $\boldsymbol{x_1}$ = *turnover in previous year*.

Table 1: Sample survey of turnover retail sales with one study variable (turnover of January), mean estimated values cv_m, S_m, $Cov(y,m)$, $R_{y;m}$ and indicator H_1; the results are derived from 100 iterations of a missing data simulation.

Auxiliary variable vectors	$\boldsymbol{cv_m}$	$\boldsymbol{S_m}$	$\boldsymbol{Cov(y,m)}$	$\boldsymbol{R_{y;m}}$	$\boldsymbol{H_1}$
$\boldsymbol{x_1}$ = *turnover in previous year*	0.997	1.176	134202	0.073	**0.0738**
$\boldsymbol{x_2}$ = *NACE classification group*	0.090	0.107	1754	0.010	0.0023
$\boldsymbol{x_3}$ = *strata group*	0.199	0.235	399	0.002	0.0027
$\boldsymbol{x_4}$ = *(turnover in previous year; NACE classification group)*	0.997	1.176	134202	0.073	**0.0738**
$\boldsymbol{x_5}$ = *(turnover in previous year; strata group)*	0.997	1.176	134202	0.073	**0.0738**

$\boldsymbol{x_6}$ = *(turnover in previous year; NACE classification group; strata group)*	0.997	1.176	134202	0.073	**0.0738**

- 2019 European health interview survey. The period is a year. The response rate is 0.564. The sample survey has 6 main study variables and more possible sets of auxiliary variables combinations than turnover retail sales. On this survey data computed H_3 indicator (5) because there is more than one study variable.

Table 2 visualizes a piece, how auxiliary variables were chosen (with a stepwise forward procedure) and that the optimal set of auxiliary variables, in this case, included all auxiliary variables ($\boldsymbol{x_6}$). Worth mentioning that the order of auxiliary variables in set does not change H_3 the result.

Table 2: Fragment of computed H_3 indicator results of a different set of auxiliary variables using European health interview survey data

Auxiliary variable vectors	$\boldsymbol{H_3}$
$\boldsymbol{x_1}$ = *(education group; gender)*	0.120
$\boldsymbol{x_2}$ = *(education group; gender; age group by gender)*	0.140
$\boldsymbol{x_3}$ = *(education group; gender; age group by gender; household income)*	0.145
$\boldsymbol{x_4}$ = *(education group; gender; age group by gender; household income; economical status)*	0.180
$\boldsymbol{x_5}$ = *(education group; gender; age group by gender; household income; economical status; age group)*	0.180
$\boldsymbol{x_6}$ = *(education group; gender; age group by gender; household income; economical status; age group; declared city, region of residence)*	**0.261**
$\boldsymbol{x_7}$ = *(education group; age group)*	0.114
$\boldsymbol{x_8}$ = *(education group; age group; declared city, region of residence)*	0.222

4 Conclusion

Information about nonresponse, study and auxiliary variables, calibration estimator and factor, indicator H_1 and H_3 is presented in the paper. Also, described the scenarios for nonresponse indicators usage. Two sample surveys were used in practice – 2020 sample survey on retail sales turnover in Latvia and 2019 European health interview survey.

Main conclusions –

- Computing calibration factor, the sum of design weights with auxiliary variable matrix in response set cannot create a singular matrix. If it is singular, then cannot calculate the inverse matrix.
- Indicator H_1 should be used if we have a sample survey with only one main study variable. If there are more than two study variables, H_3 indicator (5) should be used. If both indicators are used in the same survey, different results for the optimal auxiliary variable set can be observed.
- A user who calculates indicators should choose which indicator to use in different sample surveys.
- For a sample survey of retail sales turnover (estimating a monthly turnover in January), an optimal set of auxiliary variables consists of retail sales turnover in the previous year. However, H_1 indicator maximal value was achieved for 4 different sets of auxiliary variables.
- An optimal set of auxiliary variables for the European health interview survey includes all the auxiliary variables. Besides the order of auxiliary variables is not important regarding the nonresponse indicator value.

References

Holton, Glyn A. 2014. *Value at risk second edition.* https://www.value-at-risk.net/bias/.

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

Särndal C.-E., Lundström S. 2009. *Design for estimation: Identyfing auxiliary vectors to reduce nonresponse bias.* Statistics Sweden.

CONSUMER PRICES SAMPLE SURVEYS IN BELARUS

Natalia Sakovich

Belarus State Economic University, Belarus
e-mail: sakovichn11@gmail.com

Abstract

The paper considers the main questions of sample survey 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.

Keywords: consumer prices index, sample size, sample design.

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, for which business registers or telephone directories do provide frames, at least in some countries, notably in Western Europe, North America and Oceania. There is also the possibility of constructing tailor-made frames in a limited number of cities or locations, which are sampled as clusters in a first stage. For products, it may be noted that the product assortment exhibited in an outlet provides a natural sampling frame, once the outlet is sampled as a kind of cluster, as in the BLS sampling procedure presented above. So the absence of sampling frames is not a good enough excuse for not applying probability sampling.
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). Both simulated cut-off sampling of products within item groups. Dalén looked at about 100 groups of items sold in supermarkets and noted large biases for the sub-indices of many item groups, which however almost cancelled out after aggregation. De Haan, Opperdoes and Schut used scanner data and looked at three categories (coffee, babies'napkins and toilet paper) and, although the bias for any one of these was large, the mean square error (defined as the variance plus the squared bias) was often smaller than that for pps sampling. Biases were in both directions and so could be interpreted to support Dalén's findings. The large biases for item groups

could, however, still be disturbing. Both Dalén and De Haan, Opperdoes and Schut report biases for single-item groups of many index points.
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. Against this, it may happen that short-lived products have a different price movement from the price movement of long-lived ones and constitute a significant part of the market, so leaving them out will create bias.
4) A probability sample with respect to the base period is not a proper probability sample with respect to the current period. It is certainly true that 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. It is, of course, cheaper to collect prices near the homes of the price collectors, and it would be difficult and expensive to recruit and dismiss price collectors each time a new sample is drawn. This problem can be reduced by having good coverage of the country in terms of price collectors. One way to achieve this is to have a professional and geographically distributed interviewer organization within the national statistical agency, which works on many surveys at the same time. Another way of reducing the problem is to have a first-stage sample of regions or cities or locations which changes only very slowly.
6) The sample size is too small. Stratification is sometimes made so fine that there is room for only a very small sample in the final stratum. A random selection of 1–5 units may sometimes result in a final sample that is felt to be skewed or otherwise to have poor representativity properties. Unless the index for this small stratum is to be publicly presented, however, the problem is also small. The skewness of small low-level samples will even out at higher levels. The argument that sample size is too small has a greater validity when it relates to first-stage clusters (geographical areas) that apply to most subsequent sampling levels simultaneously.
7) Sampling decisions have to be taken at a low level in the organization. Unless price collectors are well versed in statistics, it may be difficult for them to perform probability sampling on site. Such sampling would be necessary if the product specification that has been provided centrally covers more than one product (price) in an outlet. Nevertheless, in the United States (U.S. BLS, 1997) field representatives do exactly this. In Sweden, where central product sampling (for daily necessities) is carried to the point of specifying well defined varieties and package sizes, no sampling in the outlets is needed. In countries where neither of these possibilities is at hand, full probability sampling for products would be more difficult.

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 survey in Belarus

The survey of consumer prices in Belarus has been conducted since 1992. National Statistical Committee, 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).

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 500 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.

To maintain the relevance of a selective set of products (services) -representatives, its gradual rotation is carried out by excluding certain goods (services) -representatives and including new ones for the following reasons:

- the product (service) is no longer representative, since its share in consumer spending of the population is gradually decreasing;
- the sale of goods (services) in the consumer market is not carried out (for example, as a result of technology changes or for other reasons);

New goods (services) are included in the consumption bundle in those cases when the share of expenses for their acquisition is at least 0,01% of the total consumer spending of the population of the republic.

At present, consumption bundle in Belarus is substantially expanded: are included insurance services, in particular, vehicles, financial and legal services, health services and others.

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. The weights are meant to reflect the relative importance of the goods and services as measured by their shares in the total consumption of households. The weight attached to each good or service determines the impact that its price change will have on the overall index. The weights should be made publicly available in the interests of transparency, and for the information of the users of the index.
A sample survey of households is formed on a territorial principle, households are selected proportionally to their number in the general population. The general population for the sampling procedure comprises the total number of households living in the Republic of Belarus (according to the most recent population census) excluding institutional households (residing in residential care facilities for the elderly, boarding schools, etc.) and students residing in student residence halls. While extrapolating the survey results on the general population, statistical weighting is carried out by means of assigning a statistical weight to every surveyed household. The statistical weight characterizes a represented number of households.
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 2021, weights in 2019. In cases of the unstable economic situation and consumer behavior atypical of the population, 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.

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 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).

Instructions for organizing and conducting selective state statistical monitoring of prices and tariffs for consumer goods and paid services provided to the population // Resolution of the National Statistical Committee of the Republic of Belarus No. 114 dated November 15, 2019 with amendments and additions dated October 16, 2020 No. 106.

ВЫБОРОЧНЫЕ ОБСЛЕДОВАНИЯ ПОТРЕБИТЕЛЬСКИХ ЦЕН В БЕЛАРУСИ

Наталья Сакович

БГЭУ, Беларусь
e-mail: sakovichn11@gmail.com

Аннотация

Рассмотрены основные этапы, особенности и проблемы проведения выборочных обследований потребительских цен в официальной статистике Беларуси. Отмечается, что в практике исчисления ИПЦ в основном применяются такие невероятностные методы как метод репрезентативных продуктов и метод отсечения.

Ключевые слова: индекс потребительских цен, выборка, репрезентативность, товар-представитель, невероятностный отбор.

1 Введение

Для составления индекса потребительских цен (ИПЦ) национальные органы статистики осуществляют сбор данных о ценах с помощью выборочного обследования. В практике многих стран эту процедуру следует рассматривать как состоящую из множества различных обследований, каждое из которых охватывает разные подсовокупности продуктов, включаемых в индекс.

Генеральная совокупность обычно рассматривается как трехмерный показатель. Во-первых, она имеет измерение в отношении продуктов; во-вторых, измерение в отношении географического местоположения и торговых точек, в-третьих, временное измерение, охватывающее все субпериоды того периода, к которому относится индекс.

При проведении обследований потребительских цен могут использоваться методы вероятностного и невероятностного отбора. Однако традиционно при составлении ИПЦ для отбора торговых точек или продуктов используются в основном методы невероятностного отбора. Для отбора продуктов особенно широко применяется метод репрезентативных продуктов. Для формирования выборки также используются методы отсечения и квотного отбора. В некоторых случаях эти два метода применяются в сочетании; например, торговые точки отбираются с использованием вероятностных методов, а продукты отбираются методом репрезентативных продуктов.

2 Методы невероятностного отбора

В международном стандарте по статистике цен приводятся основные причины использования невероятностного отбора:

1) Отсутствие основы выборки. Такая ситуация обычно наблюдается в отношении отбора продуктов, реже – в отношении отбора торговых точек (в качестве основы выборки которых обычно выступают регистры предприятий или справочники)

2) Систематическая ошибка в результате невероятностного отбора пренебрежимо мала, особенно это касается индексов высокого уровня агрегирования, что подтверждается в работах Далена (Dalén,1998b) и Де Хаана, Оппердуса и Шута (De Haan, Opperdoes and Schut, 1999).

3) Необходимость поддержания выборки в течение определенного времени, что связано с исчезновением ряда товаров и проблемой замены, которая также влечет за собой риски систематической ошибки.
4) Вероятностная выборка, составленная для базисного периода, не является надлежащей вероятностной выборкой для текущего периода.
5) Сбор данных о ценах должен производиться там, где есть регистраторы цен.
6) Объем выборки слишком мал. Стратификация иногда производится настолько детально, что из конечной страты может быть составлена лишь очень небольшая выборка, имеющая низкую репрезентативность.
В практике обследования потребительских цен используют следующие основные виды невероятностного отбора:
1) Отбор методом отсечения – когда n крупнейших единиц выборки отбираются с определенностью, оставляя нулевую возможность включения в выборку прочих единиц. В данном контексте «крупность» определяется некоторым показателем размера, который тесно коррелирует с целевой переменной. Под «отсечением» понимается пограничное значение между включаемыми и не включаемыми в выборку элементами. В выборку отбираются все крупные единицы, а средние и мелкие отбираются пропорционально значению заданного параметра (например, стоимостного объема продукции)
2) Квотный отбор – в полученной выборке единицы должны быть представлены в той же пропорции, что и в генеральной совокупности, с точки зрения ряда известных характеристик, таких как подгруппа продуктов, тип торговой точки и местоположение. Ограничением квотного отбора, как и других невероятностных методов, является невозможность определить стандартную ошибку оценки.
3) Метод репрезентативных продуктов – это традиционный для ИПЦ метод, при котором Центральное учреждение готовит перечень видов продуктов, содержащий их спецификации. Эти спецификации могут быть строгими, или свободными.
Метод строгих спецификаций может приводить к снижению репрезентативности, поскольку в индекс не войдут продукты, не отвечающие спецификации. Еще один недостаток – отсутствие некоторых товаров в торговых точках, что приводит к сокращению выборки. Основное преимущество метода его простота.
Метод свободных спецификаций дает регистраторам цен возможность корректировать выборку в соответствии с местными условиями и обычно приводит к более высокой репрезентативности выборки. Однако здесь присутствует проблема субъективного подхода при замене товаров.
Многие страны в практике обследования потребительских цен широко используют методы вероятностного отбора. Так, например, в США и Швеции применяются модификации отбора с вероятностью, пропорциональной размеру (ВПР), во Франции проводится случайный отбор в два этапа, в Люксембурге применяется стратифицированная целевая выборка, в Великобритании и Финляндии проводятся экспериментальные работы по составлению выборки.

3 Обследование потребительских цен в Беларуси

В 1992 году статистические службы Беларуси наряду с остальными странами СНГ перешли на выборочное наблюдение в области статистики цен с целью адекватного отражения уровня инфляции. Методология наблюдения за потребительскими ценами и расчета ИПЦ разработана при участии экспертов Международного валютного фонда и других международных организаций (ОЭСР, МВФ, Евростат) и в целом соответствует международным стандартам. Расчет ИПЦ базируется на двух информационных массивах: 1) данных ежемесячной регистрации цен по заранее определенному набору товаров представителей; 2) годовых данных выборочного обследования домашних хозяйств о структуре потребительских расходов за базисный год.
Выборочное наблюдение за потребительскими ценами включает в себя следующие этапы: 1) отбор населенных пунктов; 2) отбор базовых организаций; 3) отбор товаров-представителей; 4) регистрация цен (тарифов).

При *отборе населенных пунктов* учитывается их географическая представительность и насыщенность потребительского рынка товарами и услугами. В республике обследуется 31 город, где проживает свыше 50 % населения страны. Перечень городов остается неизменным в течение всего периода обследования. Данное обстоятельство способствует сопоставимости информации, однако снижает ее репрезентативность. Сельские населенные пункты не участвуют в наблюдении по причине невысокой насыщенности потребительскими товарами, а также в связи с недостатком достаточно подготовленных специалистов.
Отбор базовых торговых организаций основан на выборочном методе наблюдения. Выборочная совокупность включает около 7000 организаций. Для отбора базовых организаций используются данные статистической отчетности об объемах реализации товаров населению. Базовые организации должны быть представительными с различных точек зрения: форм торговли и форм собственности, размера и месторасположения. Обновление выборки базовых организаций производится ежегодно, при этом возможны замены базовой организации в случае ее ликвидации или прекращения работы на срок более 6 месяцев. Частая замена базовых организаций ухудшает качество выборки и снижает сопоставимость результатов наблюдения.
Отбор товаров (услуг) представителей. Потребительский набор, на основании которого рассчитывается ИПЦ, представляет собой репрезентативную выборку товаров и услуг, наиболее часто потребляемых населением, и в настоящее время включает около 500 наименований. В Беларуси в корзинку не включены товары бывшие в употреблении, покупки в кредит, услуги страхования, однако некоторые страны учитывают данные наименования. Потребительский набор формируется с использованием невероятностного отбора – метода репрезентативных продуктов с учетом свободных спецификаций. Большое значение имеют вопросы обновления выборки в связи с изменениями в структуре потребительского спроса, появлением новых модификаций товаров и принципиально новых товаров.
Регистрация цен и тарифов проводится в период с 10 по 30 число ежемесячно, при этом необходимо соблюдать установленные сроки регистрации цен (тарифов) с тем, чтобы выдерживать интервал между двумя регистрациями в один месяц.
Формирование весов для расчета ИПЦ осуществляется на основе данных выборочного обследования домашних хозяйств, а также дополнительной информации о розничном товарообороте, производстве и импорте отдельных товаров. Обновление весов рекомендуется производить не реже чем раз в пять лет. В Беларуси, как и в большинстве стран, обновление весов производится ежегодно, с 1 января, причем используется структура периода (t-2), то есть года, предшествующего предыдущему. В настоящее время для расчета индексов цен в 2021 г. используются веса 2019 г., однако на взгляд автора, в условиях нестабильной экономической ситуации следует использовать усредненные за ряд лет веса (например, за три года), что позволить сгладить резкие изменения в структуре потребительских расходов.

Заключение

С целью совершенствования выборочного наблюдения за потребительскими ценами в Беларуси следует:
- сочетать методы вероятностного и невероятностного отбора, расширяя применение вероятностных выборок;
- проводить географическую ротацию городов, участвующих в выборке, по возможности включать крупные сельские населенные пункты;
- расширить перечень товаров и услуг, включаемых в потребительских набор, в частности, товаров, продаваемых в кредит, бывших в употреблении, финансовых, банковских услуг, услуг страхования и ряда других;
- при формировании весов использовать средние веса за ряд лет с целью исключения влияния случайных факторов.

Список использованных источников

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).

Инструкция по организации и проведению выборочного государственного статистического наблюдения за ценами и тарифами на потребительские товары и платные услуги, оказываемые населению // Постановление Национального статистического комитета Республики Беларусь № 114 от 15.11.2019 с изменениями и дополнениями от 16.10.2020 № 106.

SAMPLE SURVEYS IN THE ASSESSMENT OF THE MAIN DETERMINANTS OF THE DECLINE IN THE BIRTH RATE IN THE REPUBLIC OF BELARUS

Eugenia Sharilova

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

Abstract

In recent decades, the demographic development of the Republic of Belarus is considered exclusively as a crisis process, one of the components of which is an intensive decline in the birth rate. Thus, in 1990-2019, the total fertility rate decreased by 27.8% and throughout the entire time period did not even reach the level of simple replacement of generations.

It is proposed to consider the problem of reducing the birth rate in the Republic of Belarus from the position of determining the reasons for the deviation of the actual birth rate from its specific norm. The age-related birth rates of the Hutterite sect are used as a standard of natural fertility in demographic practice . Based on the calculations, it should be concluded that the degree of use of the childbearing potential by women in Belarus in 2000 was 10.1%, and in 2019 – 11.4%.

To assess the direct factors of the identified negative trend, we use the data of sample surveys conducted in the Republic of Belarus. We will consider the reproductive attitudes of women in Belarus on the basis of a special survey of the reproductive health of the population conducted by Larchenko A.V. (603 women aged 15-49 years were examined in Minsk) (Larchenko, A.V. (2014)). It should be concluded that the modal value of the desired number of children in real conditions is 1 child, and in ideal conditions - 2 children. Most women are focused on a one-child family, moreover, every fifth of the respondents do not want to have children in the current conditions.

The implementation of reproductive attitudes is carried out through means of intra-family birth control, the most dangerous of which are artificial abortions. It should be noted that in the Republic of Belarus, in 2000-2019, the number of abortions per 1000 women aged 15-49 years decreased by 79% and amounted to 9.7% in 2019.

The use of contraceptives by women of the Republic of Belarus will be considered on the basis of the results of a Multi-indicator cluster survey to assess the situation of children and women conducted in 2012 and 2019. (MICS (2012. 2019)). More than 50% of women of reproductive age use contraceptives to regulate the number of children in the family and the time of their birth.

The results of sample surveys conducted in the Republic of Belarus serve as a justification for the low birth rate in the Republic of Belarus, reflecting the extremely low reproductive attitudes of the population in combination with the active use of contraceptives. The desire of modern women to get a high-quality education, a well-paid job, and lead an active lifestyle is in contradiction with the national demographic interests of the Republic of Belarus in the field of fertility.

Keywords: sample surveys, the determinants of the decline in the birth rate.

References

Larchenko, A.V. (2014) Special survey of the reproductive health of the population: analysis and results, *Voprosy statistiki*, **10**, 28-33.

Multiple Indicator Cluster Surveys to assess the situation children and women in the Republic of Belarus (MICS) (2012).

Multiple Indicator Cluster Surveys to assess the situation children and women in the Republic of Belarus (MICS) (2019).

ВЫБОРОЧНЫЕ ОБСЛЕДОВАНИЯ В ОЦЕНКЕ ОСНОВНЫХ ДЕТЕРМИНАНТОВ СНИЖЕНИЯ РОЖДАЕМОСТИ В РЕСПУБЛИКЕ БЕЛАРУСЬ

Евгения Шарилова

Белорусский государственный экономический университет, Республика Беларусь
e-mail: sharilovaee@mail.ru

Аннотация

В статье на основе результатов выборочных обследований населения Республики Беларусь рассматриваются прямые факторы, определяющие отклонение фактического уровня рождаемости от естественного уровня. В качестве таковых выступают внутрисемейные меры ограничения рождаемости, а именно: искусственные аборты и использование мер контрацепции.

Ключевые слова: выборочные обследования, детерминанты снижения рождаемости.

В последние десятилетия демографическое развитие Республики Беларусь рассматривается исключительно как кризисный процесс, одной из составляющих которого выступает интенсивное снижение рождаемости. Графическое подтверждение данного факта представлено на рисунке 1.

Рисунок 1 – Суммарный коэффициент рождаемости населения Республики Беларусь за 1990-2019 гг.

Примечание – Источник: собственная разработка на основе данных Белстата.

Так, за 1990-2019 гг. суммарный коэффициент рождаемости снизился на 27,8% и на протяжении всего временного периода не достигал даже уровня простого замещения поколений.

Следует отметить, что уровень рождаемости одновременно детерминируют факторы природной, демографической и социальной среды, действующие одновременно, с разной интенсивностью и направленностью. Достаточно сложно выделить наиболее значимые из них.

Предлагается рассмотреть проблему снижения рождаемости в Республике Беларусь с позиции определения причин отклонения фактической рождаемости от ее видовой нормы.

Здоровый человек репродуктивного возраста способен к воспроизведению потомства. Большинство статистических показателей, характеризующих интенсивность рождаемости, рассчитываются относительно женских репродуктивных контингентов (15-49 лет). Не будем отступать от этой традиции и вдаваться в полемику относительно значимости исчисления аналогичных величин для мужского населения. Тогда перефразируем первое предложение. Каждая здоровая женщина репродуктивного возраста способна к воспроизводству, то есть реализации видовой плодовитости, которая по оценкам ученых составляет порядка 10–12 живорождений за всю жизнь. Реализованная видовая плодовитость представляет собой естественную рождаемость, то есть рождаемость, не ограничиваемую противозачаточными мерами и искусственными абортами. Таким образом, гипотетически возможный уровень естественной рождаемости «преломляется» через средства внутрисемейного ограничения рождаемости, определяемые репродуктивными установками партнеров, и трансформируется в фактический уровень рождаемости (см. рисунок 2).

Рисунок 2 – Трансформация естественной рождаемости в фактическую
Примечание – Источник: собственная разработка.

Следует отметить, что измерение уровня естественной рождаемости представляет интерес с позиции его сравнительного анализа с фактическим уровнем рождаемости и определения степени использования потенциально возможного уровня видовой рождаемости, а также масштабов распространения методов намеренного внутрисемейного ограничения рождаемости. В качестве эталона естественной рождаемости в демографической практике используются возрастные коэффициенты рождаемости секты гуттеритов[1]. На основании данных о возрастной интенсивности рождаемости женщин секты гуттеритов и белорусских женщин построен рисунок 3, который наглядно отражает значимые различия в уровнях рассматриваемых показателей.

[1] Закрытая секта, в которой запрещены все методы ограничения рождаемости.

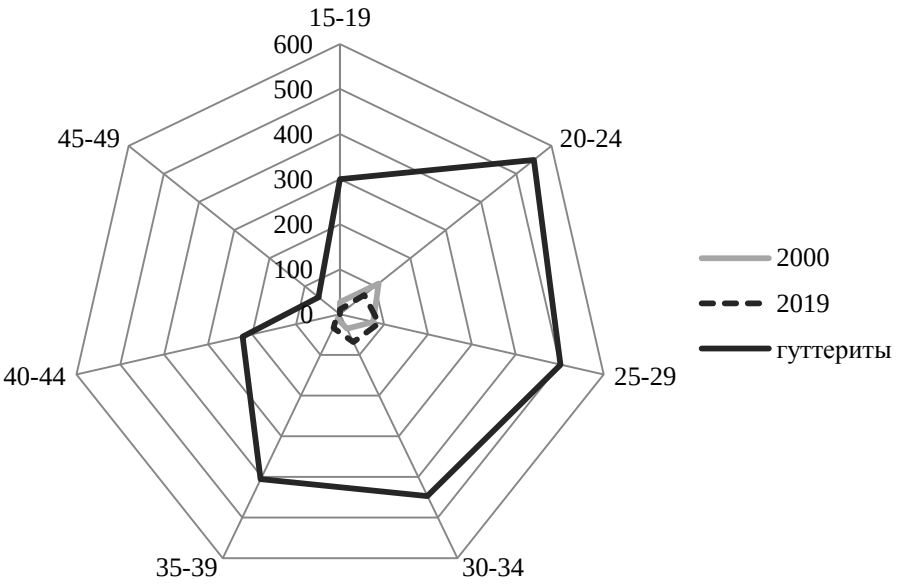


Рисунок 3 – Возрастные коэффициенты рождаемости женщин секты гуттеритов и белорусских женщин за 2000 и 2019 год, ‰

Примечание – Источник: собственная разработка на основе данных Белстата.

Кроме того, была исчислена степень использования детородного потенциала женщин Беларуси, которая в 2000 г. составляла 10,1%, а в 2019 г. – 11,4% (при условии сочетания повозрастной интенсивности рождаемости секты гуттеритов и возрастной структуры белорусских женщин 15-49 лет специальный коэффициент рождаемости в 2019 г. составил бы 351‰, при фактическом уровне 40‰). Таким образом, в Республике Беларусь наблюдается колоссальное недоиспользование потенциала естественной рождаемости.

Для оценки прямых факторов выявленной негативной тенденции используем данные выборочных обследований, проводимых в Республике Беларусь. В соответствии с рисунком 2 в качестве отправной точки следует определить оценку репродуктивных установок населения, которые выступают первопричиной использования средств внутрисемейного ограничения рождаемости. Рассмотрим репродуктивные установки женщин Беларуси, исчисленные на основе специального обследования репродуктивного здоровья населения, проведенного Ларченко А.В. (обследовано 603 женщины в возрасте 15-49 лет в г. Минске) (см. рисунок 4).

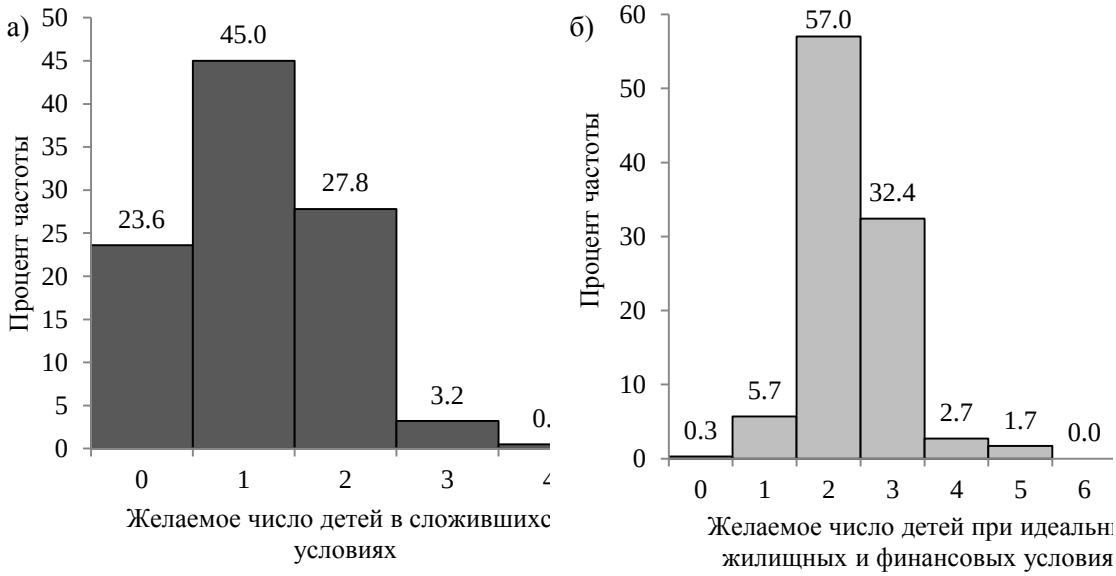


Рисунок 4 – Распределение опрошенных женщин г. Минска по желаемому числу детей: а - при сложившихся на данный момент финансовых и жилищных условиях; б - при идеальных жилищных и финансовых условиях

Примечание – Источник: [Ларченко, с. 30].

На основе данных рисунка 4 следует заключить, что модальное значение желаемого числа детей в реально сложившихся условиях составляет 1 ребенок, а в идеальных условиях - 2 ребенка. По нашему мнению, в расчет следует принимать результаты опроса в реальных условиях. Большинство женщин ориентированы на однодетную семью, более того, каждая пятая из опрошенных не желают иметь детей в сложившихся условиях. Для решения кризиса низкой рождаемости необходимо, чтобы абсолютное большинство белорусских семей было 3-4-детными. Следовательно, данные обследования показывают, что фактические репродуктивные установки населения далеки от демографических интересов страны.

Реализация репродуктивных установок осуществляется через средства внутрисемейного ограничения рождаемости, наиболее опасным из которых являются искусственные аборты. Следует отметить, что в Республике Беларусь, за 2000-2019 гг. число абортов на 1000 женщин в возрасте 15-49 лет снизилось на 79% и составило в 2019 г. 9,7‰.

Использование средств контрацепции женщинами Республики Беларусь рассмотрим на основе результатов Многоиндикаторного кластерного обследования для оценки положения детей и женщин, проводимого в 2012 г. и 2019 г. (см. рисунок 5).

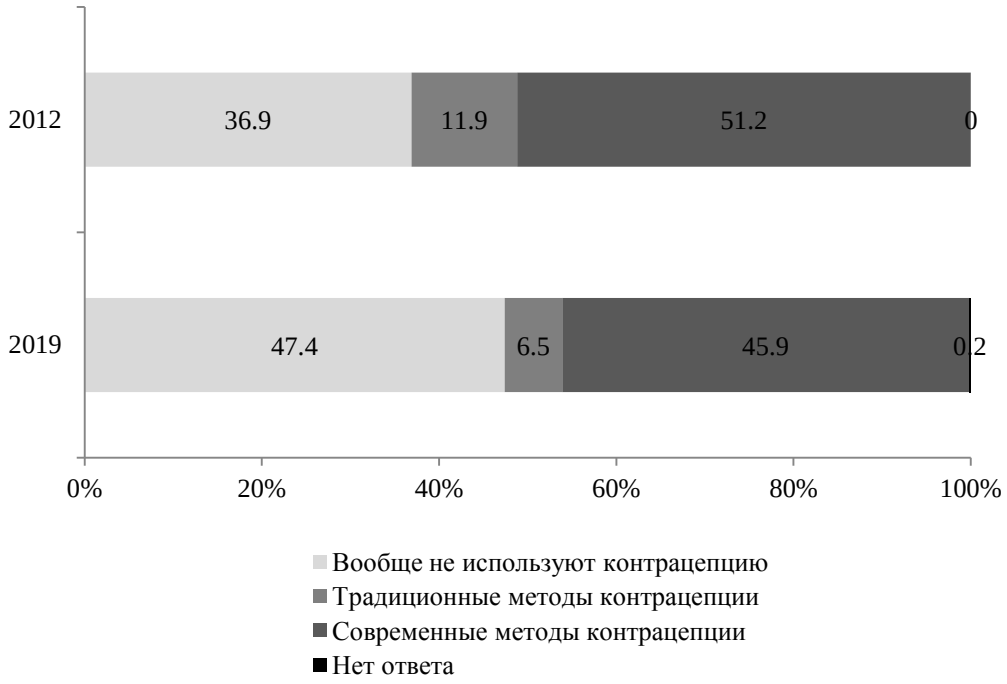


Рисунок 5 – Методы контрацепции, используемые замужними/ состоящими в незарегистрированных отношениях женщинами в возрасте 15-49 лет в Республике Беларусь в 2012 г. и 2019 г.

Примечание – Источник: собственная разработка на основе данных [Многоиндикаторное кластерное обследование 2012, с. 91; Многоиндикаторное кластерное обследование 2019, с. 72].

Данные рисунка 5 показывают, что более 50% женщин репродуктивного возраста используют средства контрацепции для регулирования числа детей в семье и времени их рождения. Достаточно неожиданным является факт повышения на 10,5 процентного пункта удельного веса женщин, не использующих данные средства.

Таким образом, результаты проводимых в Республике Беларусь выборочных обследований выступают обоснованием низкой рождаемости в Республике Беларусь, отражая крайне низкие репродуктивные установки населения в сочетании с активным использованием средств контрацепции. Желание современных женщин получить качественное образование, высокооплачиваемую работу, вести активный образ жизни входит в противоречие с национальными демографическими интересами Республики Беларусь в сфере рождаемости.

Список используемых источников

Ларченко А.В. Специальное обследование репродуктивного здоровья населения: анализ и результаты /А.В. Ларченко // *Вопросы статистики*. – 2014, №8. – С. 28-33.

Многоиндикаторное кластерное обследование 2012: Итоговый отчетстат / Нац. стат. комитет Респ. Беларусь, ООН (ЮНИСЕФ). – Минск, 2013. – 330 с.

Многоиндикаторное кластерное обследование 2019: Отчет о результатах обследования/ Нац. стат. комитет Респ. Беларусь, ООН (ЮНИСЕФ). – Минск, 2021. – 427 с.

USING LOGISTIC REGRESSION TO ANALYZE THE RESULTS OF STATISTICAL OBSERVATIONS

Ludmila Soshnikava

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

Abstract

The article discusses the issues of statistical analysis using logistic models of ordered multiple choice, which are based on the results of statistical observations, assuming the presence of a categorical dependent variable. It is advisable to use this group of models when the discrete dependent variable takes several alternative values. For example, an assessment of the level of student progress (excellent, good, satisfactory, unsatisfactory), an assessment of living conditions, an assessment of health, etc. To estimate the parameters of such models, algorithms based on elements of the theory of probability are used. The purpose of building a multiple choice model is to determine which factors and to what extent affect the probability of an event occurring, the probability of choosing one or another alternative. The article describes in sufficient detail the algorithms for calculating the logit models of binary and multiple choice, and then, using the example of a specific problem of statistical analysis of the results of self-assessment of the health of household members, the solution of such a model using the SPSS package is demonstrated.

Assessment of population health includes an objective assessment of the health status of the population according to official statistics on the prevalence of diseases among the population and an aggregate subjective assessment of the individual health status based on the results of sociological research. It is important to know to what extent the objective assessment of the health of the population and the subjective awareness of the health status of the subjects are in agreement. Since the primary files of a sample survey of households are confidential, to construct a multiple choice model, the author used conditional data, which in their characteristics are close to real values. During the sample observation, such variables as place of residence, gender, age, health assessment, sports, smoking, income were registered. For the construction, the variable health (good, fair, poor) was used as a dependent variable, gender and education were used as categorical variables; the covariates were age and income. After the model was built and its recognizing power (the correctness of predicting the dependent variable) was evaluated, the specification of this model was saved in a special file for its subsequent reconstruction.

Keywords: logit model, binary choice, multiple choice, logarithm of chance, maximum likelihood method, self-reported health.

References

Ayvazyan S.A., Mkhitaryan V.S. (1998). Applied statistics and foundations of econometrics. Textbook for universities.−M.: YUNITI,−1022 p.

Anatoliev, S. (2009). Nonparametric regression, // Quantile, , No. 7, pp. 37-52.

Byul, A. (2002). SPSS: The Art of Information Processing. Analysis of statistical data and recovery of hidden patterns. / A. Buul, P. Cefel; lane with German / ed. V. E. Momota. SPb: LLC "DiaSoftUP", - 608 p.

Voishcheva, O.S. (2006). Econometric models of qualitative variables in predictive marketing problems. / O.S. Voishcheva // Vestnik VSU, Series: Economics and Management, №2.–P. 261.

Green, William G. (2016). Econometric Analysis. Book 1 /. William Green; per. from English under scientific. ed. S.S. Sinelnikov and M.Yu. Turuntseva.—M.: Delo RANEPA,–760 p.

Eliseeva, I.I., (2007).Econometrics / I.I. Eliseeva, S.V. Kurysheva, T.V. Kosteeva et al. - Moscow: Finance and Statistics, - 576 p.

Zolt Sh. (2009). Multinomial discrete choice models. // Quantile. - No. 7–URL: http://quantile.ru/07/07-ZS.pdf

Pautova, N.I., Pautov, I.S. (2015). Gender characteristics of self-assessment of health and its perception as a sociocultural value (according to the 21st RLMS-HSE wave). / N.I. Pautova, I.S. Pautov // Woman in Russian society. No. 2.–P. 60.

Perova, MB (2016). Objective and subjective assessment of the health status of the population of Russia. - URL: http://sisupr.mrsu.ru/2016-1/PDF/Perova_2016-1.pdf - Access date: 01.12. 2019.

REGISTER-BASED CENSUS IN LITHUANIA

Milda Šličkutė-Šeštokienė

Statistics Lithuania, Lithuania
e-mail: milda.slickute@stat.gov.lt

Abstract

Statistics Lithuania, like other National Statistics Institutes, is constantly moving towards a wider usage of administrative sources. Administrative sources help to spare the costs as well as to improve the quality of the results. In Statistics Lithuania 43 percent of the published results are based on administrative sources.
Administrative sources were also widely used for Population Census 2011, but only as auxiliary information. In 2021 Population Census will be for the first time completely register-based, all the micro data will be obtained by linking number of administrative sources, no fieldwork will be carried out.

Keywords: census, register-based, administrative sources.

References

Wallgren A. and Wallgren B. (2014): Register-Based Statistics: Administrative Data for Statistical Purposes. John Wiley & Sons, Ltd

DATA COLLECTION MODE AND NONRESPONSE: PRACTICAL EXPERIENCES

Maria Valaste[1] **and Hanna Wass**[2]

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

[2] University of Helsinki, Finland
e-mail: hanna.wass@helsinki.fi

Abstract

Missing data appears in almost all survey research. There are two types of nonresponse in surveys: unit nonresponse and item nonresponse. Unit nonresponse is the failure to obtain any information from a sample unit. Item nonresponse refers to the failure to obtain information for one or more questions in a survey, given that the other questions are completed. (de Leeuw et al. 2008; Laaksonen 2018.)

In surveys data collection can be carried out using several methods. When the data collection is implemented using more than one mode, then it is a multi-mode or mixed-mode survey. A good mixed-mode strategy could lead to higher response rates and lower nonresponse bias (Laaksonen and Heiskanen 2014).

Our survey includes three different modes of data collection: face-to-face ($n = 995$), web survey ($n = 2400$), and online survey panel ($n = 681$). The survey is part of the Tackling the Biases and Bubbles in Participation (BIBU, `https://bibu.fi/en/`, project number 312710) project funded by the Academy of Finland's Strategic Research Council. The survey data is supplemented with register data taken from the administrative registers of Statistics Finland for respondent that gave their permission to combine survey and register data. We will treat permissions as response indicators. This provides us an opportunity to assess the effect of data collection mode on nonresponse and also other characteristics of the respondents. For the future studies, it may be beneficial to know in advance, who are willing to give permission to combine different data sources.

Keywords: Survey mode, mixed-mode, nonresponse, register data.

References

de Leeuw, E.D., Hox, J.J. & Dillman, D.A. (2008) The cornerstones of survey research. In de Leeuw E.D., H.J.J. & D.D.A., eds., *The international handbook of survey methodology.* Erlbaum/Taylor & Francis, New York/London.

Laaksonen, S., Heiskanen M. (2014) Comparison of three modes for a crime victimization survey. *Journal of Survey Statistics and Methodology*, **2**, 459–483.

Laaksonen, S. (2018). *Survey Methodology and Missing Data: Tools and Techniques for Practitioners.* Cham, Switzerland: Springer.

ON THE IMPORTANCE OF CONCEPTUALIZATION AND OPERATIONALIZATION IN SURVEY DESIGN: LESSONS FROM THE MORALLY DEBATABLE BEHAVIORS SCALE

Anastasiia Volkova

University of Helsinki, Finland
e-mail: anastasiia.volkova@helsinki.fi

Abstract

It is believed that nowadays research papers in social sciences are not paying enough attention to the critical issues of questionnaire design. But even using highly reputed cross-national longitudinal surveys does not protect against construct and validity problems. The present paper demonstrates the importance of conceptualization and operationalization by looking at the implementation of the Morally Debatable Behaviors Scale (MDBS) in the European Values Study (EVS). It considers the history of the survey and scale, the current approaches to the MDBS usage, and the following problems with interpretation of the results. The paper concludes with some suggestions for survey methodologists.

Keywords: Conceptualization, Operationalization, Morally Debatable Behaviors Scale, European Values Study, Reliability and Validity.

1 Introduction

Researchers strive to reach the maximum reliability and validity for their object of study. Without proper conceptualization of a phenomenon, attempts at operationalizing, testing, and predicting it can be meaningless (Babbie 2020). While new measurement tools are constantly being developed in the social and psychological sciences, sometimes the statistical analysis of the indicators and the obtained results seems to be discussed in more detail rather than the embedded in the scale meaning itself. Yet occasionally even longstanding and highly reputed surveys hide problems with constructs and scales back from their early days.

The European Values Study (EVS) was launched in 1981, with successive cross-national longitudinal waves every nine years. Despite its current scientific reputation, the first wave was designed not only by academicians but also by politicians, business executives, and priests (Kropp 2017). Many items were collected from various sources, resulting in little homogeneity in measures and formulations throughout the first questionnaire (Schwarz 1997; Kropp 2017). While some questions were replaced in the following waves, some remained untouched, such as the Morally Debatable Behaviors Scale (MDBS).

This scale is a fitting example of why a solid theoretical and methodological foundation is crucial for any survey. The MDBS tries to measure moral values by asking justifications of different actions and events, from claiming social benefits to euthanasia (Harding & Phillips 1986). However, there is no statement on why these exact phenomena were selected and why they are a priori defined as moral issues. And as morality is a latent construct that cannot be easily observed nor measured, the precise definitions of the concepts' meanings under study are essential.

2 Results

2.1 The implementation of MDBS in the EVS

The EVS began to document the survey more thoroughly only from the third wave, facing increasing demands for quality control (Halman 2001). Hence, there is very little documentation about any aspects of the pre-fieldwork: no nominal definitions, literature reviews, or any other hints of the underlying theoretical framework – although it is implied, judging by the detected dimensions (Harding & Phillips 1986).

According to several researchers (e.g., Vauclair & Fischer 2011), the MDBS had been redeveloped and adjusted to fit in the EVS from the early Crissman's scale on moral judgments (1942). But even though these scales have a similar idea, Steven Harding and David Phillips (1986) do not cite any of Paul Crissman's studies. Moreover, even if they used this scale to develop the instrument for the EVS, Crissman himself tells the readers that some methodological steps are skipped: "No special justification can be given for the employment of this particular scale" (1942, p. 29).

Furthermore, while Crissman used concrete scenarios (e.g., "Taking one's own life (assuming no near relatives or dependents)"), the methodologists of the EVS formulated the questions with vague and ambiguous indicators, often using only one word (e.g., "suicide"). It can cause differences in interpretations not only among the EVS respondents but also in further research.

2.2 Current usage and interpretation of the MDBS

Since 1981, the EVS version of the MDBS has been constantly used to measure people's moral judgments about what people should do or expected to do in a specific culture. So far, its items have been interpreted as, for instance, values (Braithwaite & Scott 1991), moral values (Halpern 2001), moral beliefs (Halman 1996), moral attitudes (Vauclair & Fischer 2011), or social attitudes (Schwartz 2006). The understanding of fundamental dimensions also differs – there are theories about one, two, and three dimensions (Harding & Phillips 1986; Halman 1996).

With such a spread, one will inevitably ponder what this scale should measure, according to the original plan. Could it be a sign of a specification error? Given the lack of proper thought and documentation, it would not be surprising if the construct implied in the survey question differs from the intended construct that should be measured (de Leeuw et al., 2008). The relationship between values, attitudes, justifications and behaviors is complicated, and that is why it is important to specify at the very beginning what exactly is being studied.

Unfortunately, this methodological confusion could be one of the causes why the EVS data on the MDBS is full of abnormalities. Most of the indicator distributions are highly skewed, making the range from 1 to 10 almost dichotomous. Also, respondents refused to answer many of the questions, which led to a large percentage of missing values (up to 23% in specific countries). Moreover, information on the measurement validity of this scale, especially cross-national, is incomplete, differs from paper to paper, and is not well-reported (Vauclair & Fischer 2011).

3 Conclusions

The European Values Study has similar problems with other scales, like, for example, the block of questions on religion. Some of the questions there are not formulated in the best way, but they cannot be easily rephrased or replaced for the sake of comparative strength (Kropp 2017). Unfortunately, this makes the usage of something like the Morally Debatable Behavior Scale complicated, albeit possible. That is why it is essential to build a solid methodological foundation from the starting point – discuss possible concepts and questions, test different measures and scales, and provide the interchangeability of indicators for better concept coverage.

And it is also crucial to document every step and decision so that other scientists can use the obtained data in the future or replicate a similar study. A suitable example of such practices is another

famous cross-national longitudinal survey, European Social Survey (ESS), which invited experts to write recommendations for possible themes, and their conceptualization and operationalization (Kropp 2017). Following these suggestions and proposals, the questionnaire has been constantly changed until it reached a state that satisfies the quality criteria (European Social Survey 2001).

References

Babbie, E. (2020) *The Practice of Social Research, 15th ed.* Cengage Learning, Wadsworth.

Braithwaite, V. A., and Scott, W. A. (1991) Values. In: *Measures of personality and social psychological attitudes* (eds. J. P. Robinson, P. R. Shaver, and L. S. Wrightsman), Academic Press, San Diego, CA, 661–753.

Crissman, P. (1942) Temporal change and sexual difference in moral judgments. *Journal of Social Psychology*, **16**, 29–38.

de Leeuw, E., Hox, J. J., and Dillman, D. A. (2008) The cornerstones of survey research. In: International Handbook of Survey Methodology, (ed. E. de Leeuw, J. J. Hox, and D. A. Dillman), Taylor & Francis, London, 1–17.

European Social Survey. (2001) *European Social Survey Core Questionnaire Development*. City University London, London.

Kropp, K. (2017) The cases of the European Values Study and the European Social Survey — European constellations of social science knowledge production. *Serendipities. Journal for the Sociology and History of the Social Sciences*, **2(1)**, 50–68.

Halman, L. (1996) Individualism in individualized society? Results from the European values surveys. *International Journal of Comparative Sociology*, **37**, 195–214.

Halman, L. (2001) *The European Values Study: A Third Wave: Source Book of the 1999/2000 European Values Study Surveys*. WORC of Tilburg University, Tilburg.

Halpern, D. (2001) Moral Values, Social Trust and Inequality. *British Journal of Criminology*, **41(2)**, 236–251.

Harding, S., and Phillips, D. (1986) *Contrasting values in Western Europe. Unity, diversity and change*. Macmillan, London.

Schwarz, N. (1997) Questionnaire design: The rocky road from concepts to answers. In: *Survey measurement and process quality* (eds. L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, et al.), Wiley, New York, 29–45.

Schwartz, S. H. (2006) A theory of cultural value orientations: Explication and applications. *Comparative Sociology*, **5**, 136–182.

Vauclair, C. M., and Fischer, R. (2011) Do cultural values predict individuals' moral attitudes? A cross- cultural multilevel approach. *European Journal of Social Psychology*, **41(5)**, 645–657.

OBSERVING NONRESPONSE BIAS AND OPTIMISING DATA COLLECTION STRATEGY FOR ADAPTIVE SAMPLE SURVEY DESIGN

Jelena Voronova

Central statistical bureau of Latvia, Latvian University, Latvia
e-mail: jelena.voronova@csp.gov.lv

Abstract

Survey sampling theoretical base knowledge determines the rules for organizing survey, drawing up sample, processing data. Theory face with non-response, various restrictions and issues during methodological implementation in the real world, justifying a growing of interest in the possibility of adapting the survey design. The flexible organization of the survey during data collection, using additional information and reacting to the achievement of the stated objectives, is called adaptive design. This work looks at designing of data collection strategy for a short-term survey using R-indicators as representativeness measures. R-indicators show the degree of difference between responding and non-responding sample groups, identify bias risk. Adjustments into data collection can be incorporated based on the analyses of the R-indicators.

Keywords: Nonresponse, bias, R-indicator, adaptive design

1 Introduction

National Statistical Institute (NSI) usually tends to achieve sufficient number of respondents or predefined response level at the end of data collection as it used to be the most important indicator of data collection quality. It is expected to achieve better data quality with higher response level, despite some studies has shown [2] high level of response does not always indicate a high quality of the data collected or measuring of impact of nonresponse of a survey.
The analysis of non-response includes portrait creating of the responding and non-responding units, identifying differences and thus potential bias in the data. Respondents form groups with a low level of response rate are additionally addressed after identification, or, using response correction or calibration methods, reduce their impact in post-processing. The essence of the R-indicator is to turn a qualitative analysis into quantitative indicator. This quality measure shows the level of difference between two sets – responding and non-responding units.
A responsive or adaptive survey data collection design is intended for active survey control, which aims to improve the chances of obtaining a representative set of final responses, reducing the variation in the weight of the final survey.

2 Response propensity and R-indicator

Using the notation and definition of response propensities as set out in Schouten, Cobben and Bethlehem (2009) [9] and Shlomo, Skinner and Schouten (2012) [14], denote U the set of units in the population $U = 1,2,..,k,..N$ and s the set of units in the sample $s = 1,2,..,k,..n$. Denote a response indicator variable R_i which takes the value 1 if unit i in the population responds and the value 0 otherwise. The response propensity is defined as the conditional expectation of R_i given the vector of values x_i of the vector X of auxiliary variables:

$$\rho_x(x_i) = E(R_i = 1 \mid X = x_i) = P(R_i = 1 \mid X = x_i) \quad (1)$$

and also denote this response propensity by ρ_x .

The quantitative quality indicators (or R-Indicators) measure the degree of difference of two sets, i.e. respondents and nonrespondents. Define the R-indicator as:

$$R(\rho_x) = 1 - 2S(\rho_x) \quad (2)$$

Estimation of the response propensity is based on logistic regression model and estimator of the variance of the response propensities:

$$\hat{S}^2(\hat{\rho}_x) = \frac{1}{N-1}\sum_s d_i(\hat{\rho}_x(x_i) - \overline{\hat{\rho}_x})^2 \quad (3)$$

where $d_i = \pi_i^{-1}$ is the design weight or inverse inclusion probabilities and $\overline{\hat{\rho}_x} = \frac{1}{N}\sum_s d_i \hat{\rho}_x(x_i)$. Thereby, estimation of the R-indicator $\hat{R}(\hat{\rho}_x) = 1 - 2\hat{S}(\hat{\rho}_x)$.

As in variance analysis, R-indicator has the same characteristics and could be split into unconditional partial indicators, which measures the distance to representative response for single auxiliary variables and are based on the between variance given a stratification with categories of Z and conditional partial R- indicators measure the remaining variance due to variable Z within sub-groups formed by all other remaining variables as in Schouten, Shlomo and Skinner (2011) [12].

3 STS retail adaptive design

Monthly data collection is done for the retail survey, where sampling design is stratified simple random sample, strata having two parameters – NACE Rev. 2 groups and size groups by turnover. Response level for business statistics usually is higher than in social statistics and reach up to 96%. Despite high level of response in STS, phenomena of unbalanced responding units were observed. STS response propensity variates and increases within time with a maximum on the last day of data collection. As well the response propensities depend on a reporting period (month). It was decided to develop a response propensity model for each selected date from data collection period (seven time points were chosen after the end of a reference period). Reference period (month) is used as one of the dependent variables in a model.

Several variables were evaluated for explaining response propensities with logistic regression. Various approaches were used for variable selection, including correlation analysis, evaluation of the amount of available data, level of explanation of the propensity to respond.

Response propensities were estimated a generalized linear model (GLM), a generalization of the classical linear model, with the binomial family logistic-regression model (logistic link function), using categorical and continuous numerical variables with different value scales and different distributions. Box-Cox conditional transformation was performed for numerical type variables.

Selection of the final model specification evaluated by the automatic *stepAIC* procedure from the *MASS* package [15], thus iteratively review all possible models from the initially passed parameters and leave only those variables where the AIC criteria is the smallest. As a result, seven models have been developed for each of the chosen seven time points in a reporting period.

The R-indicator as response rate ideally tended to be 1, both partial R-indicators ideally aimed to be 0 for the unbiased set of respondents. The values of the unconditional partial R-indicator are bounded in [-1; 1], where values with a minus sign indicate the group is underrepresented in the response set and in population estimates, due to estimated response propensity values are lower than the population average. Similarly, values with a positive sign indicate the groups where estimated response propensity is higher than the population average and can be overestimated in population estimates.

Conditional partial R-indicator compares response propensities within groups with the average response propensity of the same strata (group).

Algorithms were developed in the free software environment R [5], which allows to observe in dynamics the risk of bias due to non-response. Estimated R-indicator can help identify groups that

potentially cause a bias due to non-response, thus reducing the representativeness of the collected data. Other population parameters allow targeted planning of data collection and implementation strategy.

The auxiliary variables X are usually used in the weight adjustment stage, but their use in adaptive data collection also adds value. Monitoring the non-response bias with the help of R-indicators provides potential improvements in the implementation of effective data collection, planning and organization, potentially reducing the bias of the obtained estimates and variance of the final weights.

References

Bethlehem J., Cobben F., Schouten B. (2011) *Handbook of nonresponse in household surveys*, New York: Wiley.

Groves M., Peytcheva E. (2008) *The impact of nonresponse rates on nonresponse bias: A meta-analysis*.

Little R.J.A. (1986) *Survey nonresponse adjustments for estimates of means*, International Statistical Review, 54, 139-157.

Pengfei L, *Box-cox transformations: An overview* pengfei li, department of statistics, https://www.ime.usp.br/abe/lista/pdfm9cJKUmFZp.pdf , University of Connecticut.

R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rosenbaum P.R., Rubin D.B. (1983) *The central role of the propensity score in observational studies for causal effects*, Biometrika, 70, 41-55.

Rubin D.B. (1976*) Inference and missing data*, Biometrika 63 (3) 581–90.

Särndal C.-E., Lundquist P. (2019) *An assessment of accuracy improvement by adaptive survey design*.

Schouten B., Cobben F., Bethlehem J. (2009) *Indicators for the representativeness of survey response*, Computer Science, Chemistry Dalton Transactions.

Schouten B., Peytchev A., Wagner J. (2018) *Adaptive survey design*, Chapman & Hall/CRC Statistics in the Social; Behavioral Sciences.

Schouten B., Shlomo N. (2015*) Selecting adaptive survey design strata with partial R-indicators*, CBS, https://www.cbs.nl/-/media/imported/documents/2015/51/2015-selecting-adaptive-survey-design-strata-with-partial-r-indicators.pdf?la=nl-nl

Schouten B., Shlomo N., Skinner C.J. (2011) *Indicators for monitoring and improving representativeness of response*, Journal of Official Statistics, Vol. 27, No. 2, 231-253.

Shlomo N., Schouten B., de Heij V. (2013) *Designing adaptive survey designs with R-indicators*, NTTS 2013, https://ec.europa.eu/eurostat/cros/system/files/NTTS2013fullPaper_63.pdf

Shlomo N., Skinner C., Schouten B. (2012) *Estimation of an indicator of the representativeness of survey response*, Volume 142, Issue 1, January 2012, 201-211.

Venables, W. N. & Ripley, B. D. (2002) *Modern Applied Statistics with S. Fourth Edition*. Springer, New York. ISBN 0-387-95457-0

CAWI-mobile FOR HOUSEHOLD SURVEYS

Baiba Zukula

Central Statistical Bureau of Latvia, Latvia
e-mail: Baiba.Zukula@csp.gov.lv

Abstract

Individual-level and household surveys conducted by National statistical offices are constantly facing new challenges – the response rate is decreasing due to the changes in social habits and behavior of the individuals. In addition, because of the global COVID-19 pandemic in 2020 and 2021, the state of emergency and obligatory social distancing was announced in Latvia and other countries. As a result, all face-to-face interviews were suspended.

Because of the reasons mentioned above, it is necessary to provide the most convenient and safest approach for the respondents to participate in the surveys, i.e., on time and in an acceptable manner. The Central Statistical Bureau (CSB) of Latvia uses the Metadata-driven Integrated Statistical Data Processing and Management system (ISDAVS-CASIS) to process and manage a household and person-level survey data. In this system, it is possible to create a data entry form for face-to-face data collection (CAPI), telephone interviews (CATI) (since 2007) and online interviews (CAWI) (since 2014).

Learning from the ongoing social changes CSB Latvia has realized that it is necessary to accommodate the 'new normal'. Undoubtedly, traditional face-to-face interviews will remain as one of the data collection methods. However, it no longer can be sustained as the main data collection method. In 2019, face-to-face interviews accounted for about 60% of the largest survey data collections in Latvia.

Already in 2007, CSB Latvia established a telephone interview center to evolve the CATI data collection method. Using multiple modes for data collection is supposed to increase response rates since different modes are preferred by specific population groups, improve sample balance, and allow to reduce costs (Stadtmüller, Beuthner & Silber 2021, p. 2). Unfortunately, there is no unified mobile telephone number register in Latvia where the mobile telephone numbers of the individuals would be stored. In 2020, additional efforts were made to obtain additional telephone numbers from administrative registers and mobile phone operators, which allowed data to be collected only through telephone interviews and online interviews.

The biggest challenge for online interviews (CAWI) is that the so-far developed data entry program is not 'small screen friendly' (for phones and tablets). The survey can be easily completed on a computer, but it could be challenging to fill out on a mobile phone.

In 2021, within the framework of the EU-SILC Eurostat grant project, CSB Latvia started the work to develop a CAWI-mobile program. It will allow the respondent to fill out surveys (also as complex as the Labor Force Survey and the EU-SILC survey) on a mobile phone. Currently, the developments are based on the structure of the EU-SILC questionnaire while bearing in mind the specificities and needs of other surveys. It is planned that the designed product could be used for other surveys as well (individual-level, household, and business surveys).

When developing CAWI-mobile several limitations and challenges must be considered:

• Developments for surveys that are carried out to provide official statistics - questions and answers are strictly regulated by the legislation of the European Union, therefore the modification or shortening of the questions is rather limited. Often the number and length of possible answers are long while one screen view is not.

• In several surveys, a single questionnaire for each household member should be completed, and opportunities should be provided to move from one questionnaire to another in a way that is understandable and convenient for the respondent.

• To ensure data quality, data entry programs use classifications that make it easier for the respondent to choose the appropriate value. However, many of them are very long and includes detailed breakdowns (e.g. National Classification of Occupations, Address List).

• Many validations are already built into the data entry program to improve the quality of the data. However, often they are very complex, difficult to summarize (that is necessary in the case of small screen) and manage.

Considering the challenges identified, CSB of Latvia is planning to test CAWI-mobile for EU-SILC in 2022 and start to use it in surveys in 2023.

Keywords: household surveys, CAWI-mobile, EU-SILC.

References

Stadtmüller S., Beuthner C. & Silber H. (2021) *Mixed-Mode Surveys.* Mannheim, GESIS – Leibniz Institute for the Social Sciences (GESIS – Survey Guidelines).

Speakers

Name	Surname	Country	Institution	e-mail
Natallia	Bokun	Belarus	Belarusian State University (BSEU)	nataliabokun@rambler.ru
Yana	Bondarenko	Ukraine	Oles Honchar Dnipro National University	yana.bondarenko@pm.me
Ieva	Burakauskaitė	Lithuania	Statistics Lithuania	ieva.burakauskaite@stat.gov.lt
Ance	Cerina	Latvia	Central Statistical Bureau of Latvia	ance.cerina@csp.gov.lv
Andrius	Čiginas	Lithuania	Vilnius University, Statistics Lithuania	andrius.ciginas@mif.vu.lt, andrius.ciginas@stat.gov.lt
Piet	Daas	Netherlands	Eindhoven University of Technology, Statistics Netherlands	pjh.daas@cbs.nl
Sylwia	Filas-Przybył	Poland	Statistical Office in Poznan, Adam Mickiewicz University in Poznan	s.filas@stat.gov.pl, sylfil2@ext.amu.edu.pl
Lukas	Fuchs	Germany	Joint Berlin Master Program Statistics, Berlin	Lukas.fuchs@student.hu-berlin.de
Darja	Goreva	Latvia	Central Statistical Bureau of Latvia	Darja.Goreva@csp.gov.lv
Marcus	Gross	Germany	INWT Statistics GmbH, Berlin	Marcus.gross@inwt-statistics.de
Tetiana	Ianevych	Ukraine	Taras Shevchenko National University of Kyiv	tetianayanevych@knu.ua
Tomasz	Klimanek	Poland	Statistical Office in Poznan, Poznan University of Economics and Business	t.klimanek@stat.gov.pl, tomasz.klimanek@ue.poznan.pl
Alesia	Korolenok	Belarus	Belarusian State University (BSEU)	Alesia_tar@mail.ru
Danutė	Krapavickaitė	Lithuania	Vilnius Gediminas Technical University	danute.krapavickaite@vilniustech.lt
Mārtiņš	Liberts	Latvia	Central Statistical Bureau of Latvia	martins.liberts@csp.gov.lv
Tetiana	Manzhos	Ukraine	Kyiv National Economic University	tmanzhos@gmail.com
Zane	Matveja	Latvia	Central Statistical Bureau of Latvia	zane.matveja@csp.gov.lv
Vilma	Nekrašaitė-Liegė	Lithuania	Vilnius Gediminas Technical University, Statistics Lithuania	vilma.nekrasaite-liege@vilniustech.lt
Andrea	Neugebauer	Germany	INWT Statistics GmbH, Berlin	Anrdreas.neugebauer@inwt-statistics.de
Blaise	Ngendangenzwa	Sweden	Statistics Sweden	Blaise.Ngendangenzwa@scb.se
Ruāna	Pavasare	Latvia	Central Statistical Bureau	Ruana.Pavasare@csp.gov.lv
Natalia	Pekarskaya	Belarus	Belarusian State University (BSEU)	npekarskaya@list.ru
Ulrich	Rendtel	Germany	FB Wirtschaftswissenschaft, Freie Universität Berlin	ulrich.rendtel@fu-berlin.de
Liliāna	Roze	Latvia	Central Statistical Bureau of Latvia	Liliana.Roze@csp.gov.lv
Iryna	Rozora	Ukraine	Taras Shevchenko National University of Kyiv	rozora.iryna@gmail.com
Tomas	Rudys	Lithuania	Statistics Lithuania	tomas.rudys@stat.gov.lt
Natallia	Sakovich	Belarus	Belarusian State University (BSEU)	sakovichn11@gmail.com
Oleksiy	Sereda	Ukraine	Taras Shevchenko National University of Kyiv	as_sereda@knu.ua
Jingying	Shang	Germany	Joint Berlin Master Program Statistics, Berlin	shanjing@hu-berlin.de
Eugenia	Sharilova	Belarus	Belarus State Economic University	sharilovaee@mail.ru
Liudmila	Soshnikava	Belarus	Belarus State Economic University	ludmila_sosh@mail.ru
Kaja	Sõstra	Estonia	Statistics Estonia	kaja.sostra@stat.ee
Mykola	Sydorov	Ukraine	Taras Shevchenko National University of Kyiv	myksyd@knu.ua
Marcin	Szymkowiak	Poland	Poznan University of Economics and Business, Statistical Office in Poznan	marcin.szymkowiak@ue.poznan.pl
Milda	Šličkutė-Šeštokienė	Lithuania	Statistics Lithuania	milda.slickute@stat.gov.lt

Joel	Tolsheden	Sweden	Statistics Sweden	Blaise.Ngendangenzwa@scb.se
Maria	Valaste	Finland	University of Helsinki	maria.valaste@helsinki.fi
Olga	Vasylyk	Ukraine	National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"	vasylyk@matan.kpi.ua
Viktors	Veretjanovs	Latvia	Central Statistical Bureau of Latvia	Viktors.Veretjanovs@csp.gov.lv
Anastasiia	Volkova	Finland	University of Helsinki	anastasiia.volkova@helsinki.fi
Jelena	Voronova	Latvia	Central statistical bureau of Latvia, Latvian University	jelena.voronova@csp.gov.lv
Hanna	Wass	Finland	University of Helsinki	hanna.wass@helsinki.fi
Shu	Yang	USA	North Carolina State University	syang24@ncsu.edu
Baiba	Zukula	Latvia	Central Statistical Bureau of Latvia	baiba.zukula@csp.gov.lv

Tyrimų statistikos vasaros mokykla 2021.
Virtualios sesijos anglų kalba: 2021 m. rugsėjo 3, 10, 17 ir 24 d.
Virtualios sesijos rusų kalba: 2021 m. rugsėjo 4, 11, 18 ir 25 d.

Summer school on survey statistics 2021.
Virtual Sessions in English: Friday 3, 10, 17 and 24 September 2021
Virtual Sessions in Russian: Saturday 4, 11, 18 and 25 September 2021

ISBN 978-9955-797-34-0

Už išleidimą atsakingas Andrius Čiginas
Responsible for the publication
Tel. (+370 5) 2364949

Maketavo Dalius Pumputis
Responsible for the layout

Išleido Lietuvos statistikos departamentas
Gedimino pr. 29, LT-01500 Vilnius
https://www.stat.gov.lt/
El. p. statistika@stat.gov.lt