The VIII Conference

Survey Sampling in Economic and Social Research

5-6.06.2017 Katowice

BOOK OF ABSTRACTS

Katowice 2017

THE AIM OF THE CONFERENCE

In recent years a growing need is observed for statistical data, collected quickly and at low cost. This defines the role of the survey sampling and the importance of improving survey sampling procedures from theoretical and practical perspective. The conference gives an opportunity to present latest developments in this and related fields and to exchange experience on practical applications of survey sampling.

CONFERENCE TOPICS

- Estimation of population parameters based on complex samples
- Model-based estimation
- Statistical inference using auxiliary information
- Statistical inference based on incomplete data
- Practical implementations of sampling methods
- Sample size and cost optimization in survey sampling
- Sampling in statistical quality control
- Small area estimation
- Sampling designs
- Longitudinal surveys
- Sampling in auditing

Sampling of ordered sample based on neighborhood matrix

Tomasz Bąk

A scheme of spatial sampling with a tendency to skip adjacent elements is considered, see Wywiał (1990). The idea of this sampling was based on a neighborhood matrix that contains information about neighborhoods in the population. The sampling scheme proposed by Wywiał was dedicated for unordered sample. This article presents a proposal of sampling based on neighborhood matrix for ordered sample. Two types of sampling are discussed: preferring neighboring elements and preferring non-adjacent elements. Both approaches are illustrated by an example.

Keywords: spatial sampling, neighborhood matrix, ordered sample

Modelling survey data using empirical likelihood

Yves Berger

We propose a new design-based empirical likelihood approach for regression parameters of generalised linear models. It has the advantage of taking into account of the informative effect of the sampling design. This approach can be used for point estimation, hypothesis testing and confidence intervals for the sub-vector of parameters. It can be also used for estimation of small domains means. It provides asymptotically valid inference for the finite population parameters. We will show how the approach can be extended for hierarchical models. The simulation study shows the advantages of the empirical likelihood approach over alternative approaches. We applied our approach to the Programme for International Student Assessment (PISA) survey data. We show that the proposed approach can give more accurate and different p-values than the naive restricted maximum likelihood approach, which ignores the survey weights.

Keywords: empirical likelihood, generalised linear model, sampling design

Sample size in multiple testing.

Czesław Domański, Dariusz Parys

Multiple comparisons and connected with them multiple tests involve a large number of comparisons. In this paper we proposed some sample size formulas for multiple hypotheses. These formulas will determine the sample sizes required to provide suitable power and controlling different error rates in multiple testing like familywise error rate (FWER) , false discovery rate (FDR) or positive false discovery rate (pFDR).

Keywords: Sample sizes , multiple testing , error rates.

Calibrated Edgeworth expansions of finite population L-statistics

Andrius Čiginas

We consider a short Edgeworth expansion for the distribution function of a Studentized linear combination of order statistics of a simple random sample drawn without replacement from a finite population. Using auxiliary data available for the population units, we apply a calibration method to derive new and simple empirical Edgeworth approximations. Their sensitivity to different models of the auxiliary information is analyzed in a simulation study.

Keywords: Empirical Edgeworth expansion, Auxiliary information, Sampling without replacement, Bootstrap

Quantile regression applied to small area estimation

Enrico Fabrizi, Nicola Salvati, Carlo Trivisano

Quantile and M-quantile regression methods have been applied to small area estimation in several papers (see Chambers et al., 2014). The main idea is that of using a semi-parametric regression model for quantiles, thus avoiding parametric distributional assumptions on regression residuals and random effects. Although the Bayesian literature on quantile regression is vast and rapidly growing according to a number of different lines of research, its application to small area estimation are very limited. Our contribution goes into this direction. With respect to Bayesian literature on quantile regression, we restrict our attention to methods based on the joint estimation of conditional quantiles. We avoid a parametric specification of the likelihood, while keeping the normal as a special case. Eventually, we are interested in a method that can be implemented using widely popular MCMC software such as JAGS. To the best of our knowledge there is no small area estimation strategy with these properties. We extend the quantile regression method proposed by Reich and Smith (2013) in order to apply it to small area estimation. Joint estimation of quantiles, allows for the estimation of the whole quantile function at the area level; this can provide the basis for estimating inequality measures and many other parameters. In this presentation, we will focus on the estimation of the mean of the target variable a the small area model. Posterior distribution of the mean is obtained as a function of the posterior of the quantiles. Robust alternatives, including popular L-estimators such as the tri-mean estimator proposed by Tukey (1977) are also considered. Frequentist properties of the small area predictors are explored by means of a simulation exercise.

Keywords: Bayesian inference, non-parametric statistics, robust estimation, L-estimators

Assessing accuracy of simulation-based population total estimates

Wojciech Gamrot

Sampling schemes may be characterized by sets of inclusion probabilities. These probabilities enable the assessment of estimator properties under design-based approach. However, for many schemes (such as order sampling or sequential sampling) these probabilities are not known exactly or not known at all. In such a situation it is possible to construct design-based estimates of population characteristics by using simulation-based estimates of inclusion probabilities. Hence a need arises for construction of a suitable stop rule for a simulation procedure which would provide desired accuracy of estimates. In this talk possibilities of improving over known stopping rules are explored.

Keywords: Inclusion probability, design-based approach, simulation

Adjustment of Survey Weights via Modeling

Malay Ghosh

It is well known that point estimators tend to be unstable when the survey weights are highly dispersed and exhibit a low correlation with study variables. The problem was nicely illustrated by Basu (1971) with his famous example of circus elepahants. To limit the impact of highly dispersed weights, a number of techniques has been proposed in the literature including weight trimming and weight smoothing methods. Although both types of methods are different in nature, they share the same goal, namely, modify the survey weights so that the resulting estimators have a lower mean squared error than the standard Horvitz-Thompson estimators. Mean squared error reduction is usually achieved at the expense of introducing a bias. Hence, the treatment of survey weights by either weight trimming or weight smoothing methods can be viewed as a compromise between the bias and the variance. The talk reviews several weight smoothing methods via modeling.

Keywords: survey weights, weight smoothing, mean square error

Efficiency of interval estimation in auditing – results of simulation study

Bartłomiej Janusz

In their work auditors usually test only sampled elements, not the whole populations of interest. Statistical survey sampling may be used in order to draw a sample and conclude on the whole audited area. One of sampling strategies, often used by auditors is systematic Monetary Unit Sampling scheme and confidence interval for global error amount based on Horvitz – Thompson point estimator with assumption of its asymptotic normality. The paper presents results of the simulation study on efficiency of such sampling strategy. Main evaluation criteria are actual confidence levels compared to nominal confidence levels as well as the average length of confidence intervals compared to total book amount of population. Impact of sample size and population error rate on the estimation efficiency is examined. Normality of Horvitz – Thomson point estimator is also tested.

Keywords: Auditing, Survey sampling, Interval estimation, Monetary Unit Sampling

Nonparametric bivariate control charts based on scale curve

Przemysław Jaśko, Daniel Kosiorowski

Many new phenomena appearing in so called e-economy require developing of statistical tools which are non-parametric, robust and computationally tractable at the same time. Statistical process control theory related to milestone papers of Hotelling, Hoeffding or Akaike belongs to a core of classical applied statistics and proved its usefulness in the Economics for many situations. In this paper we propose a new methodology for conducting robust multivariate economic process control, namely induced by weighted Lp depth scale curve based control charts. Our proposals directly link to an approach presented in Liu et al. (2004) but may be used in a broader class of relatively malicious data streams which may appear in modern economy. We study a performance of our proposal in case of selected models of multivariate data streams and on real example related to the Internet users behaviour monitoring. Our proposal is able to detect changes in characteristics of a process in a presence of a moderate fraction of outliers within data. Properties of the proposal are discussed using analytical and simulation arguments. We discuss a method of uncertainty evaluation of our proposal appealing to the maximal entropy bootstrap and show its implementation. We compare a performance of the proposal with selected alternatives known from the literature.

Keywords: nonparametric control charts, weighted Lp depth, scale curve

On new item count technique with continuous control variable

Barbara Kowalczyk, Wojciech Niemiro, Robert Wieczorkowski

In the talk we consider the problem of sensitive questions in surveys. Stigmatizing, illegal or socially unaccepted features usually cannot be approached directly in a questionnaire. Eliciting truthful answers in such situations is either very difficult or even impossible to achieve. Most common indirect methods of questioning about sensitive aspects are various randomized response techniques and item count techniques. The latter ones have many practical advantages and are frequently preferred by applied researchers. But so far known item count techniques possess also some disadvantage. They suffer either from the so called floor effect or from ceiling and floor effects together. We introduce a new item count technique which is not only free from both ceiling and floor effects but has also many other theoretical advantages. In the new method we propose to use control neutral variable with continuous distribution. For estimating sensitive proportion we consider both moments and maximum likelihood estimators obtained via expectation maximization algorithm. Presented theoretical developments concerning new item count technique with continupus control variable are supported by comprehensive simulation studies.

Keywords: sensitive questions, item count techniques, EM algorithm

On two sampling techniques for estimation of quantiles

Arkadiusz Kozłowski

The balanced sampling design, which is defined by the property that Horvitz-Thompson estimators of the population totals of a set of auxiliary variables equal the known totals of these variables, is a good method for decreasing the variability of estimators of totals of variables of interest. Similar idea can be utilized in the context of estimation of quantiles. If the sample is drawn in a way that simple estimators of quantiles or cumulative distribution functions of a set of auxiliary variables equal the known quantiles or cumulative distribution functions of these variables, than estimators of quantiles of variables of interest should be less variable. Some simulation experiments have been conducted that support this supposition.

Keywords: Balanced sampling, Quantile estimation, Ordered systematic sampling, Quantile groups

On the simulation study of the influence of correlation and heteroscedasticity of random effects on the MSE estimators properties in small area estimation

Małgorzata Krzciuk

We consider the problem of estimation of the mean squared error (MSE) of some domain mean predictor for special case of the general linear mixed model known as Fay-Herriot model. In the standard form of this model the independence and the homoscedasticity of random effects is assumed. The aim of the paper is to analyze the properties of eight MSE estimators for two cases of model misspecification (due to: correlation or heteroscedasticity of random effects) and different number of domains. In the simulation study we compare properties of estimators of MSE based on the jackknife, weighted jackknife (Jiang, Lahiri, Wan (2002), Chen, Lahiri (2002, 2003)) and parametric bootstrap method (Gonzalez-Manteiga et al. (2008), Butar, Lahiri (2003)). We take also into account two classic estimators of MSE – naive and estimator based on the Taylor expansion proposed by Datta and Lahiri (2000). The simulation study is prepared in R language (R development Core Team (2016)).

Keywords: estimators of MSE, jackknife, parametric bootstrap, Empirical Best Linear Unbiased Predictor

Small area estimation of household expenditures using 2-dimensional Rao-Yu model

Alina Jędrzejczak, Jan Kubacki

In the presentation, the EBLUP estimation based on bivariate Rao-Yu model, involving both- autocorrelated random effects between areas and sampling errors is considered. It was applied to the expenditures data coming from Polish HBS as well as the explanatory variables from Local Data Bank. The efficiency of this approach involves the degree of correlation between the variables utilized in the model. In the paper we discuss the advantages and limitations of bivariate models with a special attention paid to the analysis of random effects and their decomposition using both time-related and area-related random effects. The calculations were performed using the sae and sae2 packages for R-project environment. Direct estimates were performed using the WesVAR software, and the precision of the direct estimates was determined using the balanced repeated replication (BRR) method.

Keywords: small area estimation, EBLUP estimator, Rao-Yu model, multivariate analysis

Estimation under model uncertainty

Nicholas Longford

Model selection has had a virtual monopoly on dealing with model uncertainty ever since models were identified as important conduits for statistical inference. Model averaging alleviates some of its deficiencies, but does not offer a practical solution in all settings. We propose an alternative based on linear combinations of the candidate models estimators. The general proposal is elaborated for ordinary regression and is illustrated with examples. Some estimators based on invalid models contribute to efficient estimation of certain quantities.

Keywords: Basis estimator, composite estimation, model selection, ordinary regression

Estimation of the Pointwise Hölder Exponent in Spatial Phenomena Analysis

Adrianna Mastalerz-Kodzis

Development of methods which enable the description of time series (not only economic or financial ones) using stochastic processes has begun in 20th century. Among others, plane data were modelling with the use of Hölder function. At the beginning of 21st century we are able to observe dynamic development of spatial statistics and econometrics. In the article an implementation of selected methods coming from time series analysis to modelling spatial data was proposed. Therefore, the work is a generalisation of methods used for time series on multivariate case. The paper will consist of two parts. The first one will include the analysis methodology, the second one will present some pointwise Hölder exponent estimators.

Keywords: stochastic process, Hölder function, spatial modelling, spatial estimators

On influence of clustering population on accuracy of population total estimator

Janusz L. Wywiał, Grzegorz Sitek

Estimation of a population total based on cluster sample is considered. Variances of estimators depend on intra-cluster spread of the variable under study measured by homogeneity coefficient. The spread depends on a partition of the population into clusters. Accuracy of estimators are compared under several variants of clustering algorithms or sampling designs.

Keywords: cluster sampling, intra-cluster correlation coefficient

On properties of generalized regression estimators of population and subpopulations totals

Tomasz Stachurski, Tomasz Żądło

We consider the generalized regression estimator (GREG) of the population total, GREG of the subpopulation total, where auxiliary population information is used, and modified GREG of the subpopulation total. In the simulation study we analyse some properties of these estimators. We investigate the influence of using auxiliary data on accuracy of the estimators. We put particular emphasis on the dependence between auxiliary variables, as well as between auxiliary variables and the variable under study. The properties of variance estimators of GREG estimators are also taken into account.

Keywords: generalized regression estimators, correlation, simulation

Three and two-level model in poverty estimation

Łukasz Wawrowski

Poverty measures are estimated based on sample surveys, such as The European Union Statistics on Income and Living Conditions (EU-SILC) or The Household Budget Survey (HBS). Owing to small sample sizes of less than one percent of the total population, results are available at a very general level - the whole country and regions. Nevertheless, local authorities in many countries need detailed and precise information at low levels of spatial aggregation. To deliver such estimates it is necessary to deal with a large variance of direct estimates. The study aims to estimate poverty indicators for domains which the original survey was not designed at the stage of sample selection. Estimation for such unplanned domains is possible by applying small area estimation methods and using data from other sources, such as censuses and administrative registers. The presentation attempts to estimate the headcount ratio and poverty gap at the LAU 1 level in Poland. This estimation will be possible through the use of data from the EU-SILC 2011 and The 2011 Polish Census of Population and Housing. The comparison of estimates obtained with two-level and three-level nested error unit level model in EBP method will be made. Thus results for previously unpublished level of aggregation will be obtained. Estimates of these social cohesion indicators will be diagnosed in the light of differences and efficiency.

Keywords: poverty, small area estimation, unit level model, EB estimator

CONTACT INFORMATION

Bąk Tomasz	University of Economics in Katowice Department of Statistics, Econometrics and Mathematics tbak88@wp.pl
Berger Yves	University of Southampton Department of Social Statistics and Demography y.g.berger@soton.ac.uk
Domański Czesław	University of Lodz Department of Statistical Methods czedoman@uni.lodz.pl
Čiginas Andrius	Vilnius University Institute of Mathematics and Informatics Probability Theory and Statistics Department andrius.ciginas@mii.vu.lt
Fabrizi Enrico	Universita Catolica del Sacro Cuore Dipartimento di Scienze economiche e sociali Enrico.Fabrizi@unicatt.it
Gamrot Wojciech	University of Economics in Katowice Department of Statistics, Econometrics and Mathematics wojciech.gamrot@ue.katowice.pl
Ghosh Malay	University of Florida Department of Statistics ghoshm@stat.ufl.edu
Janusz Bartłomiej	Tauron PE S.A. bartlomiej.janusz@gmail.com
Jaśko Przemysław	Cracow University of Economics Computational Systems jaskop@uek.krakow.pl

Kowalczyk Barbara	Warsaw School of Economics Collegium of Economic Analysis bkowal@sgh.waw.pl
Kozłowski Arkadiusz	Uniwersytet Gdański, Wydział Zarządzania Katedra Statystyki arkadiusz.kozlowski@ug.edu.pl
Krzciuk Małgorzata	University of Economics in Katowice Statistics, Econometrics and Mathematics gosiakrzk@gmail.com
Kubacki Jan	Statistical Office in Lodz Centre for Mathematical Statistics j.kubacki@stat.gov.pl
Longford Nicholas	Imperial College Division of Medicine sntlnick@sntl.co.uk
Mastalerz-Kodzis Adrianna	University of Economics in Katowice Department of Statistics, Econometrics and Mathematics adrianna.mastalerz-kodzis@ue.katowice.pl
Parys Dariusz	University of Lodz Department of Statistical Methods parysd@wp.pl
Rudys Tomas	Statistics Lithuania Survey Sampling subdivision tomas.rudys@stat.gov.lt
Rynko Maja	Warsaw School of Economics Institute of Statistics and Demography mrynko1@sgh.waw.pl
Sitek Grzegorz	University of Economics in Katowice Department of Statistics, Econometrics and Mathematics grzegorz.sitek@ue.katowice.pl

Stachurski Tomasz
University of Economics in Katowice
Department of Statistics, Econometrics and Mathematics
tstachurski93@gmail.com

Wawrowski Łukasz
Poznan University of Economics and Business
Department of Statistics
lukasz.wawrowski@ue.poznan.pl

Wywiał Janusz
University of Economics in Katowice
Department of Statistics, Econometrics and Mathematics
janusz.wywial@ue.katowice.pl

Żądło Tomasz
University of Economics in Katowice
Department of Statistics, Econometrics and Mathematics
tomasz.zadlo@ue.katowice.pl