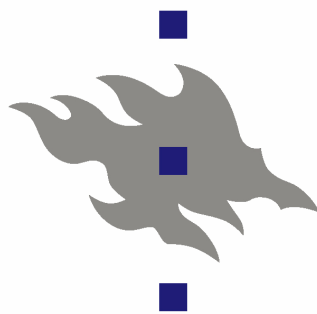


Bananas 2007  
2 - 7 June  
Kuusamo, Finland

# PROCEEDINGS



UNIVERSITY OF HELSINKI

 *Statistics Finland*



## **Second Baltic-Nordic Conference on Survey Sampling 2. – 7. June 2007, Kuusamo, Finland**

### **Scientific Committee**

Timo Alanko, Statistics Finland, Helsinki  
Signe Bāliņa, University of Latvia, Riga  
Jan Bjørnstad, Statistics Norway, Oslo  
Dan Hedlin, Statistics Sweden, Stockholm  
Annica Isaksson, Statistics Sweden, Stockholm  
Danutė Krapavickaitė, Institute of Mathematics and Informatics, Vilnius  
Gunnar Kulldorff, University of Umea  
Seppo Laaksonen, University of Helsinki  
Jānis Lapiņš, Bank of Latvia, Riga  
Risto Lehtonen (Chair), University of Helsinki  
Peter Linde, Statistics Denmark  
Aleksandras Plikusas, Institute of Mathematics and Informatics, Vilnius  
Lauri Tarkkonen, University of Helsinki  
Daniel Thorburn, University of Stockholm  
Imbi Traat, University of Tartu  
Jan Wretman, University of Stockholm

### **Organizing Committee**

Kari Djerf, Statistics Finland, Helsinki  
Terhi Hautala, University of Helsinki  
Tarja Hämäläinen, University of Helsinki  
Seppo Laaksonen, University of Helsinki  
Risto Lehtonen (Chair), University of Helsinki  
Lauri Tarkkonen, University of Helsinki  
Pasi Tuohino, University of Helsinki  
Maria Valaste (Secretary), University of Helsinki  
Kimmo Vehkalahti, University of Helsinki

### **Organizers**

Baltic-Nordic Network in Survey Sampling  
University of Helsinki, Department of Mathematics and Statistics  
Statistics Finland  
Finnish Statistical Society

### **Sponsors**

The Academy of Finland  
University of Helsinki  
Statistics Finland  
SAS Institute  
Nordic Council of Ministers  
International Association of Survey Statisticians (IASS)

## Foreword

This proceedings publication includes the abstracts of papers submitted for presentation in the Second Baltic-Nordic Conference on Survey Sampling, taking place on 2–7 June 2007 in Kuusamo, Finland. The conference belongs to a series of scientific conferences and workshops, initiated in 1997 by Professor Gunnar Kulldorff, and organized since then annually in different Baltic and Nordic countries. The main organizer has been the Baltic-Nordic Network in Survey Sampling. The network includes people from University departments, National Statistics Institutes and Statistical societies from the Nordic and Baltic countries. We now celebrate the 10<sup>th</sup> anniversary of Baltic-Nordic co-operation in research and education of survey sampling theory and methodology.

In the present conference, there are over 70 participants coming from 18 different countries. The program includes over 50 invited and contributed papers. The coverage of papers is wide: there are six sessions devoted to survey sampling and survey methodology and two sessions on business surveys. We have special thematic sessions on calibration and model-assisted methods, small area estimation, and nonresponse. Additional special sessions deal with skewed samples and longitudinal surveys, and the future of Baltic-Nordic co-operation in survey sampling. Authors of conference papers are encouraged to submit manuscripts for publication in a special issue of *Statistics in Transition Journal*; thanks are due to the Editor, Prof. Jan Kordos, for offering pages of the journal for this purpose.

The educational flavor of the conference is strong: we have a privilege to follow keynote lectures given by two prominent statisticians, Professor Harvey Goldstein of University of Bristol and Professor Carl-Erik Särndal of University of Montreal. Before the conference, we also organized a Short Course on Multilevel Modelling with Prof. Harvey Goldstein as the lecturer; there were over 60 participants in the course.

I express thanks to the members of the Scientific Committee, the Organizing Committee and the Conference Secretariat for their activity during the preparatory phases of the conference and in the event itself. Terhi Hautala edited these proceedings, Tarja Hämäläinen and Pasi Tuohino took care of administrative issues and Maria Valaste was the webmaster, all of University of Helsinki, and Hilikka Potila of Statistics Finland helped in advanced editing tasks. Several other department staff members and people of Statistics Finland assisted in practical arrangements during the conference. Thanks are due to all these people. Nordic Council of Ministers sponsored the participation of people coming from the Baltic countries. Last but not the least, this conference would not have been possible without support kindly given by The Academy of Finland, University of Helsinki and Statistics Finland.

I wish everybody an inspiring conference and enjoyable stay in Kuusamo.

Helsinki, May 2007

Risto Lehtonen

## CONTENTS

Harvey Goldstein <i>Modelling mixed response multivariate multilevel data with applications to rediction and multiple imputation</i> .....	9
Harvey Goldstein <i>An MCMC Algorithm for Estimating Multivariate Mixed Response Types at 2 Levels</i> .....	10
Carl-Erik Särndal <i>Topics in uses of auxiliary information in surveys: The role of models, Nonresponse adjustment, Estimation for (small) domains</i> .....	17
Jean-Claude Deville <i>Generalized calibration, balanced sampling and application to nonresponse</i> .....	18
Outi Ahti-Miettinen <i>Sampling Design of The Finnish Labor Cost Index</i> .....	21
Baffetta F. Fattorini L. Franceschi S. <i>A desing-based approach to K-NN technique in forest inventories</i> .....	22
Marco Ballin , Mauro Scanu , Paola Vicard <i>Efficiency of model based and model assisted estimators using probabilistic expert systems</i> .....	23
Lennart Bondesson <i>On a sampling method suitable for real time sampling</i> .....	24
Viktoras Chadyšas <i>Confidence intervals estimation for quantiles in finite population</i> .....	25
Enrico Fabrizi , Maria Rosaria Ferrante , Silvia Pacei <i>Comparing Alternative Distributional Assumptions in Mixed Models used for the Small Area Estimation of Income Parameters</i> .....	26
Lorenzo Fattorini <i>Performing Horvitz-Thompson estimation in complex sampling: a computer-intensive perspective</i> .....	27
Lorenzo Fattorini , Caterina Pisani <i>Variance estimation for measure of changes with coordinated samples</i> .....	28
Helene Feveile , Hermann Burr , Ole Olsen , Elsa Bach <i>Danish Work Environment Cohort Study 2005: Design and weighting</i> .....	29
Wojciech Gamrot <i>Estimation of Finite Population Kurtosis under Double Sampling for Nonresponse</i> 30	
Anton Grafström <i>On a generalization of poisson sampling</i> .....	31
Olga Grakoviča <i>Usage of census data as auxiliary information in survey sampling for agriculture statistics</i> .....	32
Johan Heldal <i>Ratio Estimation: When the ratio is a proportion</i> .....	33
Olga A. Vasechko , Michel Grun-Réhomme <i>Administrative and statistical registers in business statistics of Ukraine</i> .....	34
Oksana Honchar <i>Sample in Service Surveys in Ukraine: Design and Analysis</i> .....	35
Øyvind Hoveid <i>Estimation of survey weights when the frame contains information on size: Fuzzy neighbor post-stratification</i> .....	36
V C Jaunky , A J Khadaroo <i>The School-To-Work Transition for University Graduates in Mauritius: A Duration Model Approach</i> .....	38
Hans Kiesl <i>Calibrated imputation to correct for measurement error in the German Labour Force Survey</i> .....	39
Janika Konnu <i>The Effect Statistical Disclosure Control Methods Have on Data: A study on Micro Data</i> .....	41
Soile Kotala <i>Impact analysis: Grouping of Tekes-funded projects</i> .....	42
Jan Kowalski <i>Optimal linear recurrence estimators in stationary cascade rotation patterns</i> .....	43
Danutė Krapavickaitė <i>Model based estimator for a finite population total</i> .....	44
Gunnar Kulldorff <i>Ten Years of Baltic-Nordic Co-operation</i> .....	45
Seppo Laaksonen <i>Retrospective Two-Stage Cluster Sampling for Mortality</i> .....	46
Thomas Laitila <i>On-site sampling</i> .....	47
Janis Lapins , Martins Liberts <i>Estimation of monthly figures from Labour Force Survey</i> .....	48

Natalja Lepik <i>Mean square error of the general restriction estimator</i> .....	49
Daniel Thorburn and Boris Lorenc <i>Nonparametric Estimation with Double Samples</i> .....	50
Inga Masiulaitytė <i>Estimation of some inequality indexes</i> .....	51
Montanari , Giorgio E. & Ranalli , M. Giovanna <i>Calibration inspired by Semiparametric Regression as a Treatment for Nonresponse</i> .....	52
Vilma Nekrašaitė <i>Practical application of the model-based estimator for a finite population total</i> .....	54
Wojciech Niemiro and Robert Wieczorkowski <i>M-estimators and U-statistics in approximating variance of income inequality indices</i> .....	55
Kari Nissinen <i>EBLUB estimation of small area totals under linear mixed model for rotated panel data</i> .....	56
Ole Olsen , Helene Feveile , Elsa Bach <i>Changes in the psychosocial work environment in Denmark between 2000 and 2005</i> .....	57
Aleksandras Plikusas <i>Some examples of the nonlinear calibration</i> .....	58
Dalius Pumputis <i>On estimation of the variance of calibrated estimators of the population covariance</i> .....	59
Marjo Pyy-Martikainen , Leif Nordberg <i>Inverse Probability of Censoring Weighting Method in Survival Analysis Based on Survey Data</i> .....	60
Martine Quaglia , Géraldine Vivier <i>From Theory To Practice : How to Conduct Surveys Among Difficult To Reach Populations?</i> .....	61
Martine Quaglia , Cécile Lefevre , Ariane Pailhe , Anne Solaz , Ana Maria Noel, Tatiana Vichneskaïa , Bernard De Cleat <i>Interviewing both employees and employers, a mixed mode data collection for a matched survey</i> .....	62
Riku Salonen <i>Regression Composite Estimation with Application to the Finnish Labour Force Survey</i> .....	63
Nataliya Skachek <i>Sample survey of farms in Ukraine: current state and prospects</i> .....	64
Milda Slickute-Sestokiene <i>Selection of vector of auxiliary information for Generalized Regression Estimator (GREG)</i> .....	65
Kaja Sõstra <i>Restriction Estimator for Domains</i> .....	66
Silke Burestam and Daniel Thorburn <i>Correcting the regression estimator for an abundance of auxiliary variables</i> .....	67
Markus Gintas Šova , John Wood and Ian Richardson <i>Difficulties in the Estimation and Quality Assessment of Service Producer Price Indices</i> .....	68
Karolin Toompere <i>Indicator of strength of auxiliary information: a simulation study</i> .....	69
Imbi Traat <i>Estimation under restrictions</i> .....	70
Paavo Väisänen <i>Measuring item non-response of diary data in time use surveys</i> .....	71
Pieter Vlag , Koert van Bommel <i>Annual growth rates derived from short term statistics and annual structural business statistics</i> .....	72
Jacek Wesołowski <i>Linear estimation under model-design approach with small area effects</i> .....	73
Jan Wretman <i>Nonlinear Estimators of a Finite Population Total – Do They Exist?</i> .....	74
Linda Wänström <i>Sample Sizes for Two-Group Second Order Latent Curve Models</i> .....	75
Tetyana Yakovenko , Olga Vasylyk , Oksana Honchar <i>Sample Surveys in Ukraine: Education and Implementation</i> .....	76

## **KEYNOTE PAPERS**



# KEYNOTE LECTURE

## MODELLING MIXED RESPONSE MULTIVARIATE MULTILEVEL DATA WITH APPLICATIONS TO REDICTION AND MULTIPLE IMPUTATION

Harvey Goldstein <sup>1</sup>

Multilevel, hierarchically structured, data are ubiquitous in social, medical and other research, and methods for fitting models to such data are now found in standard software packages. Such models allow discrete or continuous responses, and include extensions to cross classifications, multiple membership structures and multivariate responses. The talks will focus on two extensions to current modelling procedures, with applications.

The first extension is to procedures for handling simultaneously responses that are defined at more than one level of a data hierarchy. For example, in a repeated measures ‘growth’ study we may have a series of measures on children or animals over time, that is within-individual constituting the level 1 classification, together with unchanging measures for the individual, level 2, such as their final height or their highest qualification level. Treating all these measures as responses allows the parameters of the growth ‘trajectory’ to be correlated with the individual level responses with applications, for example, to a flexible prediction system. Another application is to multiprocess models used to identify model parameters.

The second further extension, also allowing responses at several levels, is to allow both discrete and continuous responses to be modelled simultaneously, where discrete responses may be binary, ordered or unordered categorical variables. This extension is based on the idea of latent Normal variables and operates through an MCMC step which samples from such variables to replace the discrete responses. This results in a set of multivariate Normal responses which simplifies subsequent modelling steps.

An important application of these extensions is in multiple imputation. In the standard multiple imputation procedure for linear models, all the variables are treated as responses and missing responses are imputed at random from their posterior distributions. In a multilevel structure with variables at several levels it is important to allow for this structure when carrying out the imputation, thus requiring models with responses at several levels. Where there are discrete variables, current procedures typically make the often unreasonable assumption of Normality. If we use the second extension, however, this assumption can be satisfied by utilising the latent Normal variables rather than the original responses. Another important application is where we have a multivariate mixture of truly Normal responses and pseudo-continuous responses that are in fact ordered categorical variables such as rating scales.

Several examples will be discussed and further applications reviewed.

---

<sup>1</sup> University of Bristol

# APPENDIX TO KEYNOTE PRESENTATION: AN MCMC ALGORITHM FOR ESTIMATING MULTIVARIATE MIXED RESPONSE TYPES AT 2 LEVELS

Harvey Goldstein <sup>1</sup>

## The model

The model structure we consider, for a 2 level model, is as follows

$$\begin{aligned} y_{ij}^{(1)} &= X_{1ij}\beta^{(1)} + Z_{1ij}u_j^{(1)} + e_{ij}^{(1)} \\ y_j^{(2)} &= X_{2j}\beta^{(2)} + Z_{2j}u_j^{(2)} \\ e_{ij}^{(1)} &\sim MVN(0, \Omega_1), u_j = (u_j^{(1)}, u_j^{(2)})^T, u_j \sim MVN(0, \Omega_2) \end{aligned}$$

The superscripts denote the level at which a variable is measured or defined. Here  $y_{ij}^{(1)}$  is a  $p_1$  row vector containing the (latent or actual) normal responses that are defined at level 1 for level 1 unit (observation)  $i$  nested in level 2 unit  $j$ . Also,  $y_j^{(2)}$  is a  $p_2$  row vector containing the remaining responses that are defined at the higher level. We assume the same set of predictors for each response and  $X_{1ij}$  is a  $1 \times f_1$  matrix that contains the predictor variables for observation  $i$  nested in higher level unit  $j$  and  $\beta^{(1)}$  is an  $f_1 \times p_1$  matrix containing the fixed coefficients. Similarly  $Z_{1ij}$  is a  $1 \times q_1$  matrix that contains predictor variables related to  $q_1$  random effects for observation  $i$  nested in higher level unit  $j$  and  $u_j^{(1)}$  is an  $q_1 \times p_1$  matrix containing the random effects at level 2 for the level 1 responses. In the present paper we shall consider only the variance components case where  $q_1=1$ , but extensions to the general case are straightforward. Correspondingly,  $Z_{2j}$  is a  $q_2 \times p_2$  matrix for the level 2 random effects for the level 2 responses. For the level 1 residuals  $e_{ij}^{(1)}$  is a  $p_1$  row vector (calculated by subtraction).  $X_{2j}$  is a  $1 \times f_2$  vector that contains predictor variables for higher level unit  $j$  and  $\beta^{(2)}$  is an  $f_2 \times p_2$  matrix containing the fixed coefficients. The  $u_j^{(2)}$  is an  $q_2 \times p_2$  matrix of level 2 residuals (calculated by subtraction) and are correlated with the level 2 residuals for the level 1 responses. In this paper we assume  $q_2 = 1$ .

The first steps in the MCMC algorithm are concerned with how to generate the Normally distributed responses given the actual responses that may be binary, ordered, or unordered categorical. We will focus on the level 1 responses and consider each type of response in turn. The level 2 responses are generated via very similar steps.

We wish to sample Normal responses from any binary, ordered or general multicategory responses. Binary can be treated either as multicategory (*unordered*) with 2 categories or ordered with two categories. In the latter case we are effectively modelling the proportion of '1' responses and in the former the proportion of '0' responses. The latter is typically more computationally efficient.

---

<sup>1</sup> University of Bristol

## Multicategory (unordered) responses

We assume a ‘maximum indicant’ model (Aitchison and Bennett, 1970) defined as follows. Consider the multinomial vector with  $p$  categories, where the response,  $y$  is  $(0,1)$  in each category. That is, we expand the actual response for level 1 unit  $i$ , (a categorical variable with values from 1 to  $p$ ) into  $p$   $(0,1)$  variables only one of which is 1.

Thus  $y_{hi} = 1$  if response is in category  $h$  for individual  $i$ , 0 otherwise where  $h$  indexes the response. For each  $y_{hi}$  we assume an underlying latent variable  $v_{hi}$  exists and that we have the following model, where for now we omit the level 2 random effects:

$$v_{hi} = X_{1hi}\beta_{1h} + e_{hi}, \quad e_i \sim MVN(0, \Sigma)$$

$\Sigma$  is a  $p \times p$  correlation matrix,  $e_i$  mutually independent vectors (A.1)

$$X_{1hi} \text{ is } (1 \times s), \beta_{1h} \text{ is } (s \times 1), e_i \text{ is } (p \times 1), \beta_1 = \{\beta_{11}^T, \dots, \beta_{1p}^T\}^T, \text{ is } (ps \times 1)$$

For identifiability purposes we will model only the first  $p-1$  categories and assume that  $\Sigma$  is diagonal with variances equal to 1. Thus for the underlying multivariate Normal distribution the unknown parameters are the fixed coefficients with the residual covariance matrix known: this corresponds, in terms of unknown parameters, to the corresponding multinomial model where the parameters of the multinomial distribution depend only on the fixed predictor.

Let  $Y_{1hi}^*$  be the set of other responses, that is current residuals, adjusted for  $X_1$  predictors (common to all responses) and (possible) random effects at higher levels. When sampling the  $v_{hi}$  we condition on this set so that (A.1) becomes

$$v_{hi} = X_{1hi}\beta_{1h} + Y_{1hi}^*\beta_{2h} + e_{hi} \quad (\text{A.2})$$

Thus, if  $\Omega_1$  is the current residual covariance matrix for the full set of model responses, we write

$$\Omega_1 = \begin{pmatrix} \Sigma_1 & \\ \Sigma_{12} & \Sigma_2 \end{pmatrix} \quad \text{where } \Sigma_1 \text{ is the residual covariance matrix for the } Y_1^* \text{ and } \Sigma_2 = I_{p-1}. \text{ We therefore have } \beta_2 = \Sigma_{12}\Sigma_1^{-1}.$$

While the same set of model predictors  $X_1$  applies to each category, the coefficients in general are specific to each category. We therefore have

$$X_{1hi} = X_{1i}, \quad v_i = (X_{1i}^*\beta_1) + e_i, \quad v_i \text{ is } ((p-1) \times 1),$$

$$X_{1i}^* = I_{p-1} \otimes X_{1i} \text{ is } ((p-1) \times (p-1)s) \quad (\text{A.3})$$

The maximum indicant model states that we observe category  $h$  for individual  $i$  iff  $v_{hi} > v_{h^*i} \quad \forall h^* \neq h$ . Thus the category probabilities are given by

$$\pi_{hi} = pr[X_{1hi}\beta_h + e_{hi} > X_{1hi}\beta_{h^*} + e_{h^*i}] \quad \forall h^* \neq h \quad (\text{A.4})$$

If we now add level 2 random effects ( $j$  indexes level 2) (1) becomes  $v_{hij} = X_{1hij}\beta_{1i} + z_{ij}u_{hj} + e_{hij}$  where  $u_{hj}$  is  $(q \times 1)$  and we write  $u_j = \{u_{hj}\}^T$  which is a  $(q(p-1) \times 1)$  vector with  $\Omega_u = \text{cov}(u_j)$ . We also now write  $z_{ij}^* = I_{p-1} \otimes z_{ij}$  which is  $((p-1) \times q(p-1))$  and  $z_{ij}$  is  $(1 \times q)$ .

To sample the latent Normal responses  $v_{ij} = \{v_{hij}\}$  we select a sample of  $p-1$  values from  $N(X_{1i}^*\beta_1 + Y_{1i}^*\beta_2 + z_{ij}^*u_j, \Sigma_2 - \Sigma_{21}\Sigma_1^{-1}\Sigma_{12}^T)$  and accept this draw to replace the current set of  $p-1$  values if and only if the maximum of these  $p-1$  values actually occurs in the category where a response variable value of 1 is observed and if this maximum is greater than zero, or if the maximum is less than or equal to zero and a value of 1 is observed in the final category. If not, we select another sample.

## Ordered responses

Suppose we have an ordered  $p$ -category response, ordered categories numbered  $1, \dots, p$ . We consider the probit link proportional odds model

$$\gamma_h = \int_{-\infty}^{\alpha_h - (X_1\beta_1 + Y_1^*\beta_2 + ZU)} \varphi(t) dt$$

$$\gamma_h = \sum_{g=1}^h \pi_g \quad \text{categories } h = 1, \dots, p-1,$$

Where  $Y_1^*, \beta_2$  are as before, and the underlying latent Normal variable is given by

$$Y^* = e^* + (X_1\beta_1 + Y_1^*\beta_2 + ZU), \quad e^* \sim N(0, 1 - \Sigma_{21}\Sigma_1^{-1}\Sigma_{12})$$

Note that alternatively we could form  $Y^* = e^* + (X_1\beta_1 + ZU)$ ,  $e^* \sim N(0, 1)$ , but this will generally provide less efficient parameter estimates.

We assume that the intercept term is incorporated in the fixed part predictor so that  $\alpha_1 = 0$ .

We can convert this to a standard Normal model by sampling to obtain as follows (Albert and Chib, 1993). For a category 1 response we sample from the standard Normal distribution  $[-\infty, -(X_1\beta_1 + Y_1^*\beta_2 + ZU)]$

For a category  $p$  response we sample from the standard Normal distribution  $[\alpha_{p-1} - (X_1\beta_1 + Y_1^*\beta_2 + ZU), \infty]$

For every other category  $h$  we sample from the standard Normal distribution  $[\alpha_{h-1} - (X_1\beta_1 + Y_1^*\beta_2 + ZU), \alpha_h - (X_1\beta_1 + Y_1^*\beta_2 + ZU)]$

For the  $\{\alpha_h\}$ , conditional on current values of  $Y_1^*$  and other parameters we must select a new  $\alpha_h$  ( $h > 1$ ) and use MH sampling for these threshold parameters. Thus, the component of the likelihood associated with the ordered category is given by

$$P_\alpha = \prod_{i=1}^N \prod_{h=1}^p \pi_{\alpha,h}^{w_{i,h}}$$

for given  $\alpha$  where  $w_{i,h}$  is the observed (0,1) response for individual  $i$  in category  $h$ , and

$$\begin{aligned}\pi_h &= \int_{\alpha_{h-1}-(X_1\beta_1+ZU)}^{\alpha_h-(X_1\beta_1+ZU)} \varphi(t)dt, \quad 1 < h < p \quad (p \geq 3) \\ \pi_1 &= \int_{-\infty}^{-X_1\beta_1+ZU} \varphi(t)dt, \\ \pi_p &= \int_{\alpha_{p-1}-(X_1\beta_1+ZU)}^{\infty} \varphi(t)dt,\end{aligned}$$

We select a new set of values  $\alpha^*$  (one at a time) using a suitable (Normal) proposal distribution (for example derived adaptively) and set new threshold parameters  $=\alpha^*$  with probability  $\min(1, P_{\alpha^*}/P_{\alpha})$ . In addition, the order relationships among the threshold parameters must be satisfied. If the selection results in an element of  $\alpha$  that does not satisfy these relationships then that element is left at the current value. The  $\alpha$  are sampled first followed by the  $Y^*$ .

The above two steps will yield Normally distributed responses, which together with any observed Normal responses produces a multivariate Normal set. Where level 1 responses are missing we sample new responses, omitting detailed subscripts, by drawing from  $MVN(X_2^*\beta_2^* + e_1^*\beta_1^* + z_2^*u_2^*, \Sigma_2 - \Sigma_{21}\Sigma_1^{-1}\Sigma_{12})$  where  $\Sigma_2$  is the current covariance matrix of residuals for the missing responses,  $\Sigma_1$  is the covariance matrix of residuals for the observed responses and  $\Sigma_{12}$  is the matrix of covariances between the observed and missing residuals. The  $X_2^*\beta_2^*$  and  $z_2^*u_2^*$  are the fixed predictor and level 2 residual contribution for the missing responses,  $\beta_1^* = \Sigma_1^{-1}\Sigma_{12}$  and  $e_1^*$  are the level 1 residuals for the observed responses.

Following the above steps we have a new complete set of, say,  $p$  multivariate responses for each level 1 unit with the model

$$y_{hij} = X_{ij}\beta_h + z_{ij}u_{hij} + e_{hij}, \quad e_{ij} \sim MVN_p(0, \Omega_1) \quad (\text{A.5})$$

Where  $h$  indexes the response. Having sampled so that we have a set of Normal variables, we now have the following further steps. We consider first a model with only level 1 responses.

### Sampling the fixed coefficients

To sample  $\beta$  we assume a uniform prior and sample from a Multivariate Normal distribution with mean

$$[\sum_{ij} (I_{(p_1 \times p_1)} \otimes X_{ij})^T \Omega_1^{-1} (I_{(p_1 \times p_1)} \otimes X_{ij})]^{-1} \sum_{ij} (I_{(p_1 \times p_1)} \otimes X_{ij})^T \Omega_1^{-1} y_{ij}^T, \quad y_{ij} = y_{ij} - z_{ij}u_j$$

and covariance matrix  $[\sum_{ij} (I_{(p_1 \times p_1)} \otimes X_{ij})^T \Omega_1^{-1} (I_{(p_1 \times p_1)} \otimes X_{ij})]^{-1}$  where  $z_{ij}, u_j$  are defined with respect to the complete set of level 1 multivariate responses. That is

$$u_{hij} = (u_{h1j}, u_{h2j}, \dots, u_{hq_1j})^T, \quad u_j = (u_{1j}, u_{2j}, \dots, u_{p_1j})^T$$

So that  $u_j$  is  $q_1 p_1 \times 1$  with the random effects varying fastest.

### Sampling the random effects

To sample the  $u_j$  with prior  $N(0, \Omega_2)$  we note that the exponent of the likelihood for the  $j$ -th level 2 unit is  $\sum_i (y_{ij} - X_{ij}\beta - z_{ij}u_j)^T \Omega_1^{-1} (y_{ij} - X_{ij}\beta - z_{ij}u_j) + u_j^T \Omega_2^{-1} u_j$

Thus we sample  $u_j$  from the multivariate Normal distribution

$$MVN([\sum_i z_{ij}^T \Omega_1^{-1} z_{ij} + \Omega_2^{-1}]^{-1} [\sum_i z_{ij}^T \Omega_1^{-1} (y_{ij} - X_{ij}\beta)], [\sum_i z_{ij}^T \Omega_1^{-1} z_{ij} + \Omega_2^{-1}]^{-1})$$

### Sampling the level 1 (multivariate) covariance matrix

For all the categorical responses the level 1 variances are fixed to 1.0, with zero correlations among the categories of each unordered categorical variable, but non-zero correlations between these categories and other categorical and continuous variables. Thus for this set of correlations and for the unconstrained variances we use an MH sampling procedure as follows. We assume uniform priors.

Let  $\Omega_{1,lm}$  denote the  $l,m$ -th element of the covariance matrix. We update these covariance parameters using a Metropolis step and a Normal random walk proposal as follows.

At iteration  $t$  generate  $\Omega_{1,lm}^* \sim N(\Omega_{1,lm}^{(t-1)}, \sigma_{plm}^2)$  where  $\sigma_{plm}^2$  is a proposal distribution variance that has to be set for each covariance and variance. Then form a proposed new matrix  $\Omega_1^*$  by replacing the  $l,m$  th element of  $\Omega_1^{(t-1)}$  by this proposed value unless  $\Omega_1^*$  is not positive definite in which case set  $\Omega_{1,lm}^{(t)} = \Omega_{1,lm}^{(t-1)}$ . That is set  $\Omega_{1,lm}^{(t)} = \Omega_{1,lm}^*$  with probability  $\min[1, p(\Omega_1^* | e_{ij}) / p(\Omega_1^{(t-1)} | e_{ij})]$  and  $\Omega_{1,lm}^{(t)} = \Omega_{1,lm}^{(t-1)}$  otherwise. The components of the likelihood ratio are

$$p(\Omega_1^* | e_{ij}) = \prod_{ij} |\Omega_1^*|^{-1/2} \exp(-(e_{ij})^T (\Omega_1^*)^{-1} e_{ij} / 2) \text{ and}$$

$$p(\Omega_1^{(t-1)} | e_{ij}) = \prod_{ij} |\Omega_1^{(t-1)}|^{-1/2} \exp(-(e_{ij})^T (\Omega_1^{(t-1)})^{-1} e_{ij} / 2)$$

An adaptive procedure (Brown, 2004) can be used to select the proposal distribution parameters.

### Sampling the level 2 covariance matrix

We sample a new level 2 covariance matrix

$$\Omega_2^{-1} \sim \text{Wishart}(v_u, S_u)$$

$$v_u = m + v_p, \quad S_u = \left( \sum_{j=1}^m u_j u_j^T + S_p \right)^{-1}$$

Where  $m$  is the number of level 2 units,  $u_j$  is the row vector of residuals for the  $j$ -th level 2 unit and the prior,  $p(\Omega_2^{-1}) \sim \text{Wishart}(v_p, S_p)$ , where  $v_u$  is the degrees of freedom – the sum of the number of level 2 units and degrees of freedom associated with the prior. One choice is  $v_p = -3$ ,  $S_p = 0$  which is equivalent to choosing a uniform prior for  $\Omega_2$ .

The level 1 residuals are obtained by subtraction.

## Responses at both level 1 and level 2

We write the full multivariate model as follows with superscripts indicating the response level. The number of level 1 responses is  $p_1$  and there are  $p_2$  at level 2.

$$\begin{aligned} y_{hij}^{(1)} &= X_{ij}^{(1)} \beta_h^{(1)} + z_{ij}^{(1)} u_{hij}^{(1)} + e_{hij} \\ y_{hj}^{(2)} &= X_j^{(2)} \beta_h^{(2)} + z_j^{(2)} u_{hj}^{(2)} \end{aligned} \tag{A.6}$$

Note that (6) allows complex level 2 variance by specifying several random effects (Goldstein, 2003, Chapter 2), but we shall assume here that  $z_j^{(2)}$  is the constant vector =1, i.e. there are  $q_2 = p_2$  level 2 random effects for the level 2 responses. The MCMC steps are now as follows.

**Step 1:** For non-Normal level 1 responses we sample as before.

**Step 2:** For non-Normal level 2 responses we sample as for level 1 conditioning on all the remaining level 2 responses.

**Step 3:** For the level 1 covariance matrix we sample using MH as before.

**Step 4:** For the level 2 covariance matrix we sample in similar fashion to before using the full level 2 covariance matrix if all the level 2 responses are Normal. If any are categorical then, because of constraints on variances and covariances, as in sampling the level 1 covariance matrix, we need to use MH sampling element by element.

The procedure is along the same lines as for the level 1 covariance matrix but now the components of the likelihood ratio for a particular level 2 covariance matrix  $\Omega_2$  are as follows:

$$p(\Omega_2^* | u_j^{(2)}) = \prod_{ij} |\Omega_2^*|^{-1/2} \exp(-(u_j^{(2)})^T (\Omega_2^*)^{-1} u_j^{(2)} / 2)$$

$$p(\Omega_2^{(t-1)} | u_j^{(2)}) = \prod_{ij} |\Omega_2^{(t-1)}|^{-1/2} \exp(-(u_j^{(2)})^T (\Omega_2^{(t-1)})^{-1} u_j^{(2)} / 2)$$

For this step, even more than for level 1, it is important to use good starting values for the variance terms. These can be obtained, for example, from univariate 2-level variance component models.

**Step 5:** The fixed effects for the level 1 responses are estimated, as before, using the multivariate model specified by the first line of (6).

**Step 6:** The level 2 response fixed effects are estimated using the multivariate (regression) model specified by the second line of (6).

**Step 7:** The level 2 random effects for the level 2 responses are obtained by subtraction. Where level 2 responses are missing we draw a sample from  $MVN(0, \Omega_2)$ , where  $\Omega_2$  now incorporates level 2 random effects from responses at both levels. We select the random effects corresponding to these missing responses from the drawn sample.

**Step 8:** For the level 1 response level 2 random effects we sample as before, ignoring the level 2 response residuals.

Where level 2 responses are missing we sample in similar fashion to the case where level 1 responses are missing that is from  $MVN(X_2^* \beta_2^* + u_1^* \beta_1^*, \Sigma_2 - \Sigma_{21} \Sigma_1^{-1} \Sigma_{12})$  where  $\Sigma_2$  is the current covariance matrix of level 2 residuals for the missing responses,  $\Sigma_1$  is the covariance matrix of level 2 residuals for the non-missing level 2 responses and  $\Sigma_{12}$  is the matrix of covariances between the observed and missing residuals. The  $X_2^* \beta_2^*$  is the fixed predictor for the missing responses,  $\beta_1^* = \Sigma_1^{-1} \Sigma_{12}$  and  $u_1^*$  are the level 2 residuals for the observed responses.

### Imputing categories

At any cycle of the MCMC algorithm we can sample a set of category responses given the current latent responses  $Y$ . For an ordered variable we use the current parameters to sample a residual on the Normal scale and assign it to the appropriate category. For an unordered variable we sample into the category indicated by the maximum from a draw from the associated multivariate Normal distribution

# KEYNOTE LECTURE

## TOPICS IN USES OF AUXILIARY INFORMATION IN SURVEYS: THE ROLE OF MODELS, NONRESPONSE ADJUSTMENT, ESTIMATION FOR (SMALL) DOMAINS

Carl-Erik Särndal <sup>1</sup>

The lectures are set in the context of design-based estimation and inference in surveys. The use of auxiliary information is crucially important for accuracy. The lectures emphasize uses of the rich sources of auxiliary information available in Scandinavia and in the Baltic countries, through their many administrative registers. The lectures will compare (and to some extent contrast) two approaches that occupy an important place in the design-based survey sampling literature in the past twenty years: Generalized regression (GREG) estimation and calibration estimation. Design-based estimation is often qualified as model assisted; consequently, statistical models play a role in the lectures. However, model dependent estimation (non-design-based) will be referred to only incidentally. The theory will be illustrated by survey situations at Statistics Canada, Statistics Finland and Statistics Sweden.

1. The family of generalized regression (GREG) estimators. Extensions of the classical GREG estimator, to a wider context of assisting models: Mixed models, generalized linear models, logistic and others.
2. The family of calibration estimators. Extensions of the classical calibration approach. Models that may be associated with these estimators. Calibration estimators versus GREG estimators: When and how do they differ? Examples where calibration and GREG give different answers.
3. Estimation for domains and small areas. Does the choice of model matter in model assisted design-based inference for domains? In model-dependent inference?
4. Nonresponse bias. Close examination of nonresponse bias in estimators, in particular calibration estimators. How great is the bias, how damaging is it, how do we reduce it? Description of a tool (a sample-based bias indicator) for the selection of the auxiliary variables most likely to reduce the nonresponse bias.

---

<sup>1</sup> Université de Montréal

# KEYNOTE LECTURE

## GENERALIZED CALIBRATION, BALANCED SAMPLING AND APPLICATION TO NONRESPONSE

Jean-Claude Deville <sup>1</sup>

The lectures are devoted to more or less new tools in estimation theory and in sampling theory. Their application to nonresponse correction can be seen as a one of their major interest. Generalized calibration has a special interest in weighting for complete nonresponse. In particular it provides a tool for situations where nonresponse is caused by a variable of interest, for example drug consumption. Balanced sampling can be used in the case of random imputation for item nonresponse. Its effect is to reduce the extra variance due to random imputation.

1. Calibration principle and generalized calibration.
  - General calibration principle and nonlinear estimation.
  - Generalized calibration.
  - Standard metric calibration.
2. Application to the correction of total nonresponse
  - Response mechanism and response model.
  - Link with generalized calibration.
  - Bias and variance.
  - Some examples.
3. Balanced sampling
  - Balancing a sample: why ?
  - Balancing a sample: how ? The 'cube' method.
  - Approximate variance in the case of maximum entropy sampling.
4. Application to some methods of imputation for nonresponse
  - Imputation weighting like.
  - Random imputation for nonresponse.
  - The case of a (0-1) variable.
  - The case of a qualitative variable.
  - Quantitative variable: beyond the cube method.

---

<sup>1</sup> ENSAI\INSEE, France

## **INVITED AND CONTRIBUTED PAPERS**



# SAMPLING DESIGN OF THE FINNISH LABOR COST INDEX

Outi Ahti-Miettinen<sup>1</sup>

The Finnish labor cost index (FLCI) describes the development in the cost of labor for an hour worked in the private sector. Calculation of the labor cost index sets out from the forming of an index of wages and salary costs, into which an index of social costs is then added.

The FLCI system is under re-engineering that has several targets, one being to improve the estimation of the change in labor costs. This requires to pay attention to cross-sectional estimates. Our resources will allow for a sample size of around 2000. Estimates are needed both for the whole private sector and for main industries. That leads to use of stratified sampling. Because industries vary greatly from one to another, in allocating observations to strata, power allocation is been used. To get also benefits from size classification, allocation was carried in two steps: first to industries and then to size classes within industry. Cross-sectional estimates are based on standard formulae and can be estimated quite well although some informative nonresponse will appear.

The paper presents the methodology for the sampling design of FLCI. This will be based on tests against the two population level data sets, the one from 2002 and the other from 2004. These have been constructed from registers and earlier surveys, and will be rather close to real-life. An exception is that we will not have all labor cost variables in the data sets, but register based wages and salaries.

---

<sup>1</sup> Statistics Finland

# A DESIGN-BASED APPROACH TO K-NN TECHNIQUE IN FOREST INVENTORIES

Baffetta F. <sup>1</sup>, Fattorini L. <sup>1</sup>, Franceschi S. <sup>1</sup>

Several techniques for assessing natural resources employ information from satellite imagery and ground data. Between these methodologies, the non parametric k-Nearest Neighbours (k-NN) is becoming increasingly popular.

The fields in which the k-NN method found the major number of applications is related with forest attribute mapping. In this case, the data usually at disposal are the values of the spectral bands for all the study area pixels, while the interest variable values are known only for a sample of these. The k-NN estimation procedure for the interest variable at a specific pixel is made by a weighted average of the k nearest neighbours sampled pixels with respect to a distance metric adopted in the covariate space.

The effectiveness of the method is based on the fact that it makes it possible to jointly estimate the interest variable/s at any pixel as well as the total on the whole study area or in a sub-region. Notwithstanding its wide use in surveys of natural resources, the statistical properties of the k-NN estimators are not clearly delineated.

The available literature is invariably of model-based nature. Obviously the properties of the resulting predictors strictly depend on the validity of the adopted superpopulation models which are indeed frequently unrealistic. Moreover these procedures completely neglect the fact that, as usual in forest inventories, data arise from a pre-determined scheme of sampling.

In this paper, the statistical properties of the k-NN estimators are derived in a completely design-based framework, avoiding any assumption about populations. At least to our knowledge, this is the first attempt to analyse k-NN estimators from a design-based point of view. General results which hold for any sampling scheme are previously derived and subsequently applied to simple random sampling without replacement. The design-based performance of k-NN are evaluated by an extensive simulation study performed on several population whose dimension and covariate values are taken from real study cases, while the interest variables were generated from different level of relationship between covariate and interest variable.

---

<sup>1</sup> Dipartimento di metodi quantitativi, Università degli Studi di Siena

# EFFICIENCY OF MODEL BASED AND MODEL ASSISTED ESTIMATORS USING PROBABILISTIC EXPERT SYSTEMS

Marco Ballin<sup>1</sup>, Mauro Scanu<sup>1</sup>, Paola Vicard<sup>2</sup>

Two classes of estimators of a contingency table when a sample is drawn according to a stratified sampling design are introduced. The two classes are composed of model based and model assisted estimators (Särndal et al., 1992) respectively, where the model is defined in terms of probabilistic expert systems (PES, see for instance Cowell et al., 1999).

PES are widely used graphical models for the analysis in many scientific contexts, especially artificial intelligence and multivariate statistics. A graphical model describes a multivariate independence model. We show that in a model based framework the PES machinery can be easily adapted to complex sampling designs through the definition of an additional node,  $\Pi$ , representing the sample design.  $\Pi$  is a categorical variable assuming as many categories as the strata. The marginal probability attached to  $\Pi$  is defined through the sampling design as the fraction of the overall weight relative to the units of the population in each stratum. In a model assisted framework it is not necessary to include  $\Pi$  in a PES.

Both the classes include the usual Horvitz-Thompson estimator as a special case when the PES corresponds to the complete (i.e. saturated) model.

PES can be particularly useful to propagate information through a set of variables and consequently allow the definition and performance of poststratification.

Monte Carlo simulations have been carried out to evaluate and compare the two classes of estimators.

## Some references:

Cowell, R. G., Dawid, A. P., Lauritzen, S. L., and Spiegelhalter, D. J. (1999): Probabilistic Networks and Expert Systems, Heidelberg: Springer.

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

---

<sup>1</sup> ISTAT

<sup>2</sup> University Roma 3

# ON A SAMPLING METHOD SUITABLE FOR REAL TIME SAMPLING

Lennart Bondesson<sup>1</sup>

A flexible list sequential  $\pi$ ps sampling method introduced in Bondesson & Thorburn (2006) is presented. It can reproduce any given without replacement  $\pi$ ps sampling design, of fixed or random sample size. The method is a *splitting method* as defined in e.g. Tillé (2006), and uses successive linear updating of inclusion probabilities with proper restrictions. The main advantage of the method is in *real time sampling* situations, cf. Meister (2004), where it can be used as an alternative to Bernoulli and Poisson sampling and can give correlations as desired between the sampling indicators and considerably reduce the variability of the sample size. See also Grafström (2007). For fixed size  $\pi$ ps sampling, methods like Rosén's Pareto sampling, cf. e.g. Bondesson *et al.* (2006), are usually more suitable.

**Key words:** generalized Bernoulli sampling, generalized Poisson sampling, inclusion probabilities, list sequential sampling, Markov process, martingale,  $\pi$ ps sample, real time sampling, splitting method, stationary process, 3P-sampling, weak m-dependence.

*2000 Mathematics Subject Classification:* Primary 62D05; Secondary 60G10, 60G42, 60J10.

## References

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

Bondesson, L. & Thorburn, D. (2006). A list sequential sampling method suitable for real time sampling. Research Report in Mathematical Statistics No 3, Dept. of Mathematics and Mathematical Statistics, Umeå University, Umeå. .

Grafström, A. (2007). On a generalization of Poisson sampling. Research Report in Mathematical Statistics No. 2, Dept. of Mathematics and Mathematical Statistics, Umeå University, Umeå.

Meister, K. (2004). On methods for real time sampling and distributions in sampling. Ph.D thesis, Dept. of Mathematical Statistics, Umeå University, Umeå.

Tillé, Y. (2006). Sampling algorithms. Springer series in statistics. Springer science + business media, Inc., New York.

---

<sup>1</sup> Umeå University

# CONFIDENCE INTERVALS ESTIMATION FOR QUANTILES IN FINITE POPULATION

Viktoras Chadyšas<sup>1</sup>

Confidence intervals provide a way to report an estimate of a population quantile along with some information about the estimates precision. Although different settings lead to different formulas for computing confidence intervals, the basic interpretation is always the same. Understanding of the interpretation is most important. We will investigate construction of the confidence intervals for a finite population quantiles in the most basic

settings. Our main focus will be getting the correct estimates of the confidence intervals to the finite population quantiles and understanding what we can and cannot claim based on the confidence intervals. Some procedures that may be used to obtain the estimates of the confidence intervals for quantiles using a without replacement sampling design [1] in the finite population will be considered. An initial way of obtaining the confidence interval for the median when the distribution of the median estimator is asymptotically normal is described in [3]. If the population is not normally distributed, then the confidence level of such interval may be inaccurate for small sample sizes, because the distribution of the sample quantile is not well-approximated by a normal distribution. In this case we construct estimates of the confidence intervals for quantiles using re-sampling methods [2]. Thus the estimates of the confidence intervals tend to be

smaller, but they are approximately correct. Two different populations are employed in the simulation study: the first one is generated by the standard normal distribution and the second one is generated by the exponential distribution. It means that symmetric

and asymmetric population distributions are used to produce confidence intervals for the quantiles. Some simulation examples will be given during the presentation.

## References:

[1] Danutė Krapavickaitė, Aleksandras Plikusas. (2005) Imčių teorijos pagrindai. Technika, Vilnius.

[2] Sharon L. Lohr. (1999) Sampling: Design and Analysis. Pacific Grove: Duxbury Press.

[3] Carl-Erik Särndal, Bengt Swensson, Jan Wretman. (1992) Model Assisted Survey Sampling. Springer - Verlag, New York.

---

<sup>1</sup> Vilnius Gediminas Technical University

# COMPARING ALTERNATIVE DISTRIBUTIONAL ASSUMPTIONS IN MIXED MODELS USED FOR THE SMALL AREA ESTIMATION OF INCOME PARAMETERS

Enrico Fabrizi <sup>1</sup>, Maria Rosaria Ferrante <sup>2</sup>, Silvia Pacei <sup>2</sup>

Small Area Estimation is concerned with producing estimates of descriptive quantities of sub-populations whenever the portion of the sample specific to the sub-populations are so small that design-unbiased estimators are characterized by unacceptably large variances.

Linear Mixed Models provide a popular basis to produce model-based small area predictors by means of ‘borrowing strength’, i.e. by pooling together information from all areas to estimate model parameters (Ghosh and Rao, 1994). Standard applications of Linear Mixed Models to Small Area Estimation rely on the normality of both random effects and residuals, an untenable assumption when the target variable is given by household income. In these applications, the skewness (and excess of kurtosis) in the distribution of the target variable usually implies some degree of non-normality in the random effects and the residuals. One alternative is to consider a different model within the class of Generalized Linear Mixed Models (GLMM), for instance assuming that income is Gamma instead of normally distributed (Ghosh *et al.*, 1998). Another popular alternative is to apply Linear Mixed Models on the log-income. This solution requires a back-transformation to obtain predictions of descriptive quantities of the income parameters on their natural scale. Both parametric (based on log-normality assumptions) and non-parametric (smearing, RAST) can be used to solve the problem (see Chambers and Dorfman, 2003).

In this presentation we compare the design-based properties of EBLUP predictors based on ‘unit-level’ normal Linear Mixed Model, Gamma GLMM and Linear Mixed Models on log-income with different options on the back-transformation. Comparison is based on a simulation exercise that uses data from the European Community Household Panel (EUROSTAT, 2002). The considered target variable is equalised household income. The focus on design-based properties is motivated by the need to evaluate the performances of model-based predictors within the randomization framework, which represents the standard of the market in the analysis of sample surveys. Our main results is that predictors based on the normal Linear Mixed Model for the household income are the best performers. Predictors based on the Gamma GLMM are also good and compare favourably to the predictors based on Linear Mixed Models specified for the log-income.

## References:

Chambers R.L., Dorfman A.H. (2003) Transformed variables in survey sampling, *Working paper* M03/21, Southampton Statistical Sciences Research Institute.

EUROSTAT (2002) European social statistics - Income, poverty and social exclusion, 2nd report.

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

Ghosh M., Natarajan K., Stroud T.W.F, Carlin B.P (1998) Generalized linear mixed models for small-area estimation. *Journal of the American Statistical Association*, 93, 272-282.

---

1 University of Bergamo, Italy

2 University of Bologna, Italy

# **PERFORMING HORVITZ-THOMPSON ESTIMATION IN COMPLEX SAMPLING: A COMPUTER-INTENSIVE PERSPECTIVE**

Lorenzo Fattorini<sup>1</sup>

An empirical version of the Horvitz-Thompson estimator is proposed for complex sampling schemes, when the analytical determination of the inclusion probabilities is prohibitive. In accordance with Fattorini (Biometrika, 2006, 93, 269-278), the inclusion probabilities are estimated by means of independent replications of the sampling scheme. The properties of the resulting estimator are derived. An adaptive algorithm for choosing the appropriate number of replications is proposed and some applications in sequential sampling and two-phase sampling is considered. A case study for estimating the total amount of carbon in the forests of Trentino (North Italy) is discussed.

---

<sup>1</sup> Dipartimento di Metodi Quantitativi, Università di Siena, Italy

# VARIANCE ESTIMATION FOR MEASURE OF CHANGES WITH COORDINATED SAMPLES

Lorenzo Fattorini <sup>1</sup>, Caterina Pisani <sup>1</sup>

In sample surveys, there is often the need of estimating periodic changes for the total of an interest variable. If  $T_1$  and  $T_2$  denote the total at times 1 and 2, a natural estimate of the target parameter  $D = T_1 - T_2$  is given by  $\hat{D} = \hat{T}_1 - \hat{T}_2$ . In order to increase the precision of  $\hat{D}$ , positively coordinated (overlapping) samples should be selected, taking into account that populations are subject to changes in their composition and so list frames are periodically updated. This goal may be achieved using the permanent random number (PRN) techniques, which generally produce sizeable overlap between samples selected in successive temporal occasions even in presence of updated frames. The probability distribution from which random numbers are generated and the criterion adopted to select units on the basis of their associated random numbers define various sampling schemes. It is worth noting that when an auxiliary variable is available, order  $\pi$ ps sampling schemes (Rosen, *Journal of Statistical Planning and Inference*, 1997, 62, 159-191) based on permanent random numbers may be suitable. Among them, Pareto  $\pi$ ps sampling is shown to be optimal among  $\pi$ ps sampling schemes with fixed distribution shape (Rosen, 1997) and competitive for controlling sampling overlap (Ohlsson, 2000 *Proceedings of the Second International Conference on Establishment Surveys*, ASA, 255-264). Aires (*Computational Statistics*, 2004, 19, 337-345) presents Fortran procedures to compute first- and second-order inclusion probabilities when Pareto  $\pi$ ps sampling is adopted. Unfortunately, the program code proposed by Aires is no more available to be downloaded. Alternatively, PRN may be used with the Sunter sampling scheme (Sunter, *Applied Statistics*, 1977, 26, 261-268) which gives rise to first-order inclusion probabilities proportional to size for the most important part of the population and second-order inclusion probabilities which are readily computable.

However, it should be pointed out that, in order to get an unbiased variance estimator of  $\hat{D}$ , the second-order inclusion probabilities for the coordinated sampling scheme (i.e. the joint probability of two given units entering the sample at time 1 and time 2 respectively) have to be quantified. To our knowledge, this problem has been approached in the case of simple random sampling without replacement and stratified sampling (Nordberg, *Journal of Official Statistics*, 2000, 16, 363-378) but no results are available for sampling schemes based on PRN, in which the exact computation of these probabilities is numerically unmanageable. The present paper proposes a simple Monte Carlo procedure to quantify the second order inclusion probabilities of the coordinated scheme. The method is based on an algorithm by Fattorini (*Biometrika*, 2006, 93, 269-278) and works for any scheme adopted to select coordinated samples. A simulation study is performed in which coordinated samples are selected from a real population by means of the Sunter scheme.

---

<sup>1</sup> Università di Siena, Italy

# DANISH WORK ENVIRONMENT COHORT STUDY 2005: DESIGN AND WEIGHTING

Helene Feveile <sup>1</sup>, Hermann Burr <sup>1</sup>, Ole Olsen <sup>1</sup>, Elsa Bach <sup>1</sup>

The Danish Work Environment Cohort Study (DWECS) is a series of national surveys conducted in 1990, 1995, 2000 and 2005. The primary objective of the survey is surveillance, i.e. monitoring of population proportions of various occupational exposures. The secondary objective is to enable analytical, epidemiological analyses, i.e. association between occupational exposures and the subsequent development of health symptoms.

Until the 2005-wave DWECS had a split panel design combining independent, repeated cross-sectional surveys and built-in cohorts, thus enabling both the surveillance and the analytical, epidemiological aspect. In 2005, however, the sample consisted of the follow-up of the cohort (a), supplementary samples added in order to obtain a representative cross-section (b and c) as well as an additional representative, cross-sectional sample (d) and a job/trade-specific sample (e):

10,131 persons who were invited to participate in DWECS 2000 and who were still alive and living in Denmark in 2005 (they were between 23 and 74 years of age).

A simple random sample of 943 18-22-year-old persons.

A simple random sample of 236 23-59-year-old immigrants not residing in Denmark in 2000.

A simple random sample of 8,545 18-59-year-old persons.

A stratified sample of 4,183 20-59-year-old employees within 15 jobs or trades. The jobs or trades were of special interest due to specific occupational exposures or governmental plans of action - but small. In order to increase precision a larger sample of these jobs and trades were drawn.

The response rate in the follow-up part of the survey (a) was 66%, and the response rates in the subsamples, added in order to obtain a representative cross-section (b and c), were 48% and 33%, respectively. The response rate in the supplementary cross-sectional sample (d) was 61%. The overall response rate in the stratified job/trade-specific sample was 75%.

Due to the mainly descriptive objective of the study and the amount and diversity of variables, a uniform and easily comprehensible reporting of results was desirable.

The construction of weights for subsamples a – d was theoretically straightforward, since the sampling frame was a daily updated administrative national register of persons residing in Denmark. The construction of weights for subsample e was more complicated. The register holding the trade was of high quality, but the job-specific register had some problems. Thus an initial screening of respondents was necessary. The fact that employees (as opposed to all residents in Denmark) was the sampling frame for this subsample also calls for consideration. The construction of the weights will be presented.

In the presentation of population proportion estimates, subsample e is intended to increase precision for specific occupational exposures. Examples of applications of the weights will be presented.

---

<sup>1</sup> National Research Centre for the Working Environment, Copenhagen, Denmark

# **ESTIMATION OF FINITE POPULATION KURTOSIS UNDER DOUBLE SAMPLING FOR NONRESPONSE**

Wojciech Gamrot <sup>1</sup>

Estimates of the population kurtosis may be seriously affected by sample data incompleteness. The well-known procedure of double sampling for nonresponse may be utilized to compensate for this unwanted effect. In this paper an estimator of the finite population kurtosis based on a general double sampling procedure incorporating arbitrary sampling designs in both phases is proposed. Expressions for its approximate bias and variance are derived assuming a stochastic nonresponse model represented by arbitrary response distribution. An important special case is also discussed.

---

<sup>1</sup> Department of Statistics, University of Economics, Katowice, Poland

# ON A GENERALIZATION OF POISSON SAMPLING

Anton Grafström<sup>1</sup>

In real time sampling the units of a population will pass a sampler one by one in real time. Alternatively the sampler may successively visit the units of the population. Each unit passes only once and at that time it is decided whether or not it should be included in the sample. The goal is to take a sample and efficiently estimate a population parameter. Generalized Poisson sampling as defined here is a list sequential sampling method that can be used in this situation. The method is an alternative to Poisson sampling, where the units are sampled independently with given inclusion probabilities. Generalized Poisson sampling uses weights to create correlations between the inclusion indicators. In that way it is possible to reduce the variation of the sample size and to make the samples more evenly spread over the population. Simulation shows that Generalized Poisson sampling can be used to improve the efficiency in many cases.

**Key words:** list sequential sampling, Generalized Poisson sampling, Generalized Bernoulli sampling, inclusion probabilities, real time sampling, simulation, splitting method

---

<sup>1</sup> Umeå University

# USAGE OF CENSUS DATA AS AUXILIARY INFORMATION IN SURVEY SAMPLING FOR AGRICULTURE STATISTICS

Olga Grakoviča<sup>1</sup>

There will be agriculture census in 2010 in Latvia, what will give information about all farms in the country. The database of census results will give a chance to summarize statistical information in very detailed level. However it is possible to use this database not only for preparing of statistical information, but also for survey sampling organization.

The contributed paper is a practical research on census data, what is available before survey. The aim of the work is to improve the precision of sample estimates using auxiliary information. The research was based on simulations.

The first topic discussed in the paper is division of farms by size groups. There is not concrete theory about division of sampling frame units by size groups. The task of the research was to find optimal division of farms by size groups using simulation process. The test parameter is coefficient of variation (CV) of the main survey estimates.

The second topic discussed in the paper is to determine sample size and allocation by strata. It was also studied with simulation process according to the best stratification and tested by CV.

Some part of the results was used in Farm Structure Survey 2007 sampling.

---

<sup>1</sup> University of Latvia, Faculty of Physics and Mathematics

# RATIO ESTIMATION: WHEN THE RATIO IS A PROPORTION

Johan Heldal<sup>1</sup>

Ratio estimation has been a popular method to improve survey estimates of means and totals when a population total is available from a variable which is assumed to be highly correlated with the target variable. The method has a rationale from both a design and a model based perspective. An appropriate model can bring into consideration structures in the problem at hand that a design-based approach alone cannot. However, we very often see that such specific structures are disregarded in applications, even when the estimates are accompanied with model-based variances.

One such special structure occurs when some principal variable  $x$  is available for all units, for instance from a business annual accounts register, and we are interested in the total  $T_y$  for a component  $y$  of  $x$ .  $y$  has to be obtained from a survey. An appropriate model for  $y$  given  $x$  places  $z$ ,  $0 \leq z = y/x \leq 1$ , with possibly positive point probabilities at 0 and 1, i.e. a trinomial distribution on 0, (0,1) and 1 and a continuous distribution given  $z \in (0,1)$ .  $\mu = Ez$  takes the role of the ratio regression parameter.

The trinomial distribution can be estimated several ways. I will in my talk propose a model for  $z$  based on the beta distribution where  $\mu$  may be modified as a function of  $x$  and other covariates through a regression model. The results for estimating  $T_y$  using this model will be compared to that of the naïve ratio estimator.

---

<sup>1</sup> Statistics Norway,

# ADMINISTRATIVE AND STATISTICAL REGISTERS IN BUSINESS STATISTICS OF UKRAINE

Olga A. Vasechko <sup>1</sup>, Michel Grun-Réhomme <sup>2</sup>

The Ukraine is undergoing significant changes connected with fast internal development and increased participation in global markets. These changes call for improved business statistics; otherwise, public policies will bear increased risks. In particular, improvements are called for in the system of registers and sources of information used.

This paper focuses on the role of administrative registers in business surveys. We analyze the statistical burden on Ukraine enterprises with respect to the dimensions of scale and dynamics. Approaches to systems of business registers are considered

**Key words:** business statistics, operator, statistical burden, register, administrative file

---

<sup>1</sup> Scientific and Technical Complex of Statistical Research

<sup>2</sup> Université Paris 2

# SAMPLE IN SERVICE SURVEYS IN UKRAINE: DESIGN AND ANALYSIS

Oksana Honchar <sup>1</sup>

This paper presents small Ukrainian experience of sample survey methodology in service sector. Share of service sector in gross value added in Ukraine increased from 50% in 2001 to 56% in 2005. 69% of all enterprises are engaged in services. Nearly 90% from them are small. Moreover only 15% of non-financial services turnover belongs to small enterprises. Since 2001 survey in non-financial service is carried out annual and monthly.

With purpose of survey effect increasing the State Statistics Committee of Ukraine conducted pilot sample survey in September 2006. The survey units are enterprises and local units. Survey units are divided on non-active, null and active. Stratified random with Neyman allocation is applied for active enterprises.

The economic activity (59 groups) and the size of enterprise (3 groups) are used as the stratification variables. Since data are skewed extremal elements (outliers) were detected in each stratum. Extremal enterprises and small strata (which have less than ten elements) are 100% included in the sample. Systematic sampling is used within each stratum.

We analyze nonresponses and reasons of them. Weighting are used for dealing with nonresponses. We also analyze sampling errors of estimates for turnover and number of persons employed.

## References:

Cochran, William G. (1977) *Sampling Techniques*, John Wiley & Sons, New York.

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

---

<sup>1</sup> Scientific and Technical Complex of Statistical Research, Ukraine

# ESTIMATION OF SURVEY WEIGHTS WHEN THE FRAME CONTAINS INFORMATION ON SIZE: FUZZY NEIGHBOR POST-STRATIFICATION

Øyvind Hoveid <sup>1</sup>

It should be well-known that the three main methods of weight adjustment in the presence of self-selection in surveys—post-stratification, regression calibration and generalized regression calibration — are all equivalent when auxiliary variables are binary and orthogonal representing non-overlapping groups. However, when the survey frame contains variables with information on size (or order), the coding of these into binary and orthogonal ones incurs presumably both a loss of information and a certain arbitrariness with respect to the choice of coding. Since every method for weight adjustment basically is a method of weight estimation, it is desirable that this coding is automatized so that more information is retained and more of the arbitrariness of coding is considered statistical variability. Moreover, it is desirable that the estimated weights themselves are considered stochastic. Fuzzy neighbor post-stratification (FNPS) is a technique that meets these requirements. FNPS is in particular useful when the frame is rich on information and the degree of self-selection in the survey is relatively large. Actually, FNPS is developed with regard to a survey where statistical design has been informal without recorded weights at all — the Norwegian farm accounting survey.

The theoretical backdrop of FNPS is that of the super-sample and the super-design. The super-sample (a parallel to the superpopulation) is a probability distribution from which the actual sample can be assumed drawn. Information about that distribution is obtained by resampling (without replacement) from the original sample according to the statistical design. The super-design is proportionate to the ratio of the super-sample and the empirical distribution of the population. Thus, while the statistical design of the sample describes the actual draw of the sample prior to data collection, the super-design describes as-if draws at the time of data collection. The super-design can be estimated asymptotically consistent in situations where the statistician has incomplete or no control over the sampling process, and this is what fuzzy neighbor post-stratification is all about.

FNPS is based on statistical theory and utilizes as much information as possible up to a statistical criterion. The main idea is simple, and is based on relatively reliable weighted prediction models of sample variables, naturally restricted to the class of partial least squares regressions. First are nearest neighbor post-strata constructed around each single distinct sample point to comprise that part of the population space which in a certain sense has predictions closest to that of the sample point. The corresponding adjusted weights are as usual proportionate to the population count relative to the sample count within each post-stratum, and are functions of the regression weights. This simple method has the virtue that with binary and orthogonal variables, it is equivalent with the other three main methods. Nevertheless, nearest neighbor post-stratification meets two problems. First, it is affected by the uncertainty of the prediction model. Secondly, the corresponding weights are not logically consistent: If applied twice, the third set of weights are in general different from the second. The first problem is virtually eliminated when we take the expectation of nearest neighbor weights over the super-sample as adjusted weights. Logical consistency is then achieved by picking that set of weights which is identical to the expectation of its nearest neighbor weights. These are the fuzzy post stratification weights. The term "fuzzy post-strata" refers to the fact that the average stratification has the nature of a collection of fuzzy sets. Approximate fuzzy poststratification is obtained naturally by computing a series of nearest neighbor post-stratifications and forming their average. The next set of weights is computed with regard to a resample from the supersample and the average of previous weights.

---

<sup>1</sup> Norwegian Agricultural Economics Research Institute

FNPS can be compared to other methods by using the variance estimation which the resampling technique allows. In our case of application it turns out that the variances of fuzzy-weighted population means are not substantially larger than the variances of these means with standard post-stratification. On the other hand, the prediction models employed by FNPS are much more powerful taking more auxiliary variables into consideration. Hence the bias of FNPS is presumably smaller and eventual bias correction is more reliable. Only simulation experiments can decide when the computational efforts of FNPS are worth while. That depends of course on relationship between the survey self-selection mechanism and the information in the frame.

# THE SCHOOL-TO-WORK TRANSITION FOR UNIVERSITY GRADUATES IN MAURITIUS: A DURATION MODEL APPROACH

V C Jaunky<sup>1</sup>, A J Khadaroo<sup>1</sup>

This paper examines the school-to-work transition (STWT) for University of Mauritius (UoM) graduates who completed their degree during the period 1995-2000. A variety of survival frameworks is employed and the gamma frailty log-normal model is found to fit the data best. An inverted U-shaped baseline hazard is detected. A higher age at graduation and a higher father education increase the job search time for graduates. While a higher mother education and postgraduate training lead to a lower job search time. Management and engineering graduates experience a smaller job search period than science and social science graduates. In addition graduates from urban areas have a lower job search time than their rural counterparts. Male graduates and female graduates on average experience the same job search duration.

**Keywords:** Educational economics, human capital, Mauritius

---

<sup>1</sup> Department of Economics and Statistics, University of Mauritius

# CALIBRATED IMPUTATION TO CORRECT FOR MEASUREMENT ERROR IN THE GERMAN LABOUR FORCE SURVEY

Hans Kiesel<sup>1</sup>

The German LFS, which is conducted as a CAPI household survey, covering 1% of the German population, suffers from serious under-reporting of marginal employment. Compared to data from administrative files, up to 2 million marginally employed people are missing in the estimated total from the LFS. Since the LFS is mandatory in Germany, non-response rates are negligible and cannot explain these large differences. Thus we face some kind of measurement error; obviously, lots of sampled persons are "hidden" marginally employed.

There are several plausible explanations for this. First, about 20% of all interviews are conducted as proxy interviews (e.g. parents frequently answer the questions for their sons or daughters, when they are studying at university away from home). In some cases, the proxy respondents might not know about the marginal employment of the target person. Second, the interviewers are paid by numbers of interviews, not by length of interviews. Thus, some interviewers might be inclined to answer the employment questions by themselves (with "no"), if they (wrongly) consider the answer to be obvious (e.g. in case of pupils, students, pensioners or housewives). And third, the respondents might not consider their part-time work as a real kind of employment (e.g. if a schoolboy is asked, whether he is employed).

In a joint project with the German National Statistical Institute (Destatis), we are currently investigating, how information from administrative files on the number of marginally employed can be used to correct the LFS for this measurement error. Using these numbers as additional calibration totals in the weighting process (the LFS is part of the German microcensus, which is calibrated to demographic totals by means of the GREG estimator) is not recommended, since this would result in changing all weighting factors and therefore even changing the estimates of variables which are measured without error.

Beaumont (2005) proposed to use calibrated imputation in the presence of non-response. We propose to use a similar kind of calibrated imputation in the presence of measurement error, including the following steps:

Step 1: Use the LFS data to fit a binary choice model for the variable "marginally employed" and calculate a propensity score for each sampled unit. Use as many predictors as possible from the data (e.g. age, sex, region, education, household size, number of employed persons in the household, number of children in the household, dummy for proxy interview). Since there is measurement error in the data, we will get biased propensity scores.

Step 2: For the proxy interviews, re-calculate the propensity scores with the fitted model, setting "proxy = no".

Step 3: Estimate the total of the propensity scores (using the GREG weighting factors) and compare them to the data from administrative files (marginally employed by sex and age, 6 groups in total). The differences have to be imputed.

Step 4: Calibrate the propensity scores using some distance function, so that the (weighted) totals of the new propensity scores add up to the totals from administrative files. We use a distance function analogous to that used in GREG estimation, so that it is possible to use common software for this step

---

<sup>1</sup> Institute for Employment Research, Nuremberg, Germany

(we use CLAN from Statistics Sweden). Note however, that values, not weighting factors are calibrated in this step.

Step 5: Use some kind of stochastic controlled rounding procedure to transform the new propensity scores unbiasedly into a binary variable. We propose using balanced pps-sampling for this step, using the SAS macro CUBE from INSEE (Rousseau and Tardieu, 2004).

In our paper, we present our approach in more detail and show some results, e.g. how the imputation of marginal employment changes the estimated totals of employed, unemployed and people not in the labour force.

### **References:**

Beaumont, J.-F. (2005): Calibrated imputation in surveys under a quasi-model-assisted approach, *Journal of the Royal Statistical Society Series B* 67, 445–458.

Rousseau, S. and Tardieu, F. (2004): La macro SAS CUBE d'échantillonnage équilibré, *Documentation de l'utilisateur*, internal paper, INSEE.

# THE EFFECT STATISTICAL DISCLOSURE CONTROL METHODS HAVE ON DATA: A STUDY ON MICRO DATA

Janika Konnu <sup>1</sup>

The interest for Statistical Disclosure Control (SDC) methods has been increasing in statistical agencies. The demand for micro data is increasing and so is the amount of applications for micro data under contract at Statistics Finland. Data must be protected against disclosure even when they are submitted to researchers. There are many SDC methods available and they differ greatly from each other. There have been attempts to make comparison on methods but the work has proven hard. Each method has its strengths and weaknesses and how well a method perform depends on data. While studying the SDC methods, I got interested in the changes that occur in data when applying the methods. Each of the studied method is tested with different parameters to find some kind of threshold where the alteration in data is getting unacceptably largely. The first interest was the effect that the method has on categorical frequencies. Secondly we analysed fitted model on both original and protected data. The results helped us to understand what kind of effect data protecting will have on model parameters and on conclusions.

**Keywords:** Statistical Disclosure Control, Micro data, Microaggregation, PRAM

---

<sup>1</sup> Statistics Finland

# IMPACT ANALYSIS: GROUPING OF TEKES-FUNDED PROJECTS

Soile Kotala<sup>1</sup>

This paper presents an empirical study analysing the impact of publicly financed research and development projects in Finland from EPM data of Tekes (Finnish funding Agency for Technology and Innovation). Tekes is the main public funding organisation for research and development in Finland, and for instance in 2005, Tekes invested 429 million euros in research and development projects at companies, universities and research institutes. Altogether 2 134 R&D projects were provided with funding.

In 2002, Tekes assumed an Ex-Post Monitoring (EPM) system - a permanent standard of sending self-evaluation survey questionnaires ex post facto to organisations, which have implemented a Tekes-funded project three years earlier. The questionnaires contain questions, which seek to define the overall impact and effectiveness of Tekes-funded projects and Tekes's support. In the course of five years of EPM (2002 – 2006), 5564 projects have responded, which is 66 % of all projects during that time. Almost 80 % of the responds comes through EPM web survey questionnaires. The EPM data has been little exploited for statistical analysis. This paper introduces some of the earliest statistical results derived from the EPM data. The results include the grouping of Tekes-funded projects according to their overall impact and effectiveness, called the impact grouping.

The impact grouping was formed from EPM data of years 2002-2004 using cluster analysis. The cluster analysis was conducted by K-means algorithm of SPSS-program using k=5. The suitability of five groups to the data originated from the results of dendrograms. The groups were labelled as follows: 1) Influential projects, 2) Low-influential projects, 3) Autonomously implemented, financially dependent projects, 4) Tekes-steered, financially independent projects and 5) Success stories. Later on, discriminant analysis was used to assign projects from EPM data 2005-2006 into these impact groups. With classification functions, 94,5% of the new projects could be identified into their correct impact groups. The impact groups summarize 8 questions (43 sub questions) of EPM questionnaires, and serve as a feasible platform for demonstrating the effectiveness of Tekes-funded projects.

The EPM data is combinable with other project registers of Tekes, which hold information from application and project implementation phases. This information consists of e.g. the nature of the project (research projects of universities and research institutions/development projects of small and medium size companies (SMEs)/R&D projects of large companies), standard industrial classification, programme participation, finance and estimated risks of the project. The results of the study suggest e.g. that influential and autonomously implemented projects are more frequent within research projects than in company projects, whereas Tekes-steered and low-influential projects are more common within company projects, more so within R&D projects of large companies. The impact grouping, i.e. the effectiveness of the project is also connected with the standard industrial classification, for instance 'Manufacture of food products and beverages' include Success stories clearly less than would be expected if the projects had been equally distributed. Moreover, contrary to expectation, the participation in a technology programme seems to promote only the effectiveness of development projects of SMEs.

---

<sup>1</sup> SC-Research, Finland

# OPTIMAL LINEAR RECURRENCE ESTIMATORS IN STATIONARY CASCADE ROTATION PATTERNS

Jan Kowalski <sup>1</sup>

We consider a sequence of surveys involving rotation of elements in the sample. For each occasion an optimal linear unbiased estimator (BLUE) of the current population mean, based on all available previous knowledge, may be found. Our object of interest are linear recurrence relations between the BLUE estimators obtained on successive occasions. Then each estimator may be computed recursively and at a reduced cost. We aim at finding explicit formulas for the solution. This may be possible under additional assumptions, regarding for instance an exponential correlation structure. In its original version (see [1]) the problem is formulated in a fairly general manner. However, various difficulties lead to certain constraints. Firstly, we introduce a class of 'cascade' patterns, meaning that the rotation scheme should have a sufficiently regular structure. Secondly, we consider the stationary form of the recurrence. This in turn may be identified with passing to the limit with occasion number in the solution to the classical version. Our gain is the ability to apply a convenient, operator-based approach to the optimization problem. The results were obtained together with Jacek Wesolowski from the Central Statistical Office of Poland.

## **Bibliography:**

- [1] Patterson H., Sampling on successive occasions, *Journal of the Royal Statistical Society, Series B*, 12, 241-255, 1950
- [2] Rao J., Graham J., Rotation Designs for Sampling on Repeated Occasions, *Journal of the American Statistical Association*, 50, 492-509, 1964
- [3] Binder D.A., Hidioglou M.A., Sampling in Time, *Handbook of Statistics*, 6, 187-211, 1988
- [4] Kowalski J., Rotation in sampling patterns, in review for *Journal of Statistical Planning and Inference*, 2006

---

<sup>1</sup> Warsaw University of Technology, Faculty of Mathematics and Information Science

# MODEL BASED ESTIMATOR FOR A FINITE POPULATION TOTAL

Danutė Krapavickaitė<sup>1</sup>

The variable of expenditure on environmental protection is quite irregular: some of the enterprises have high expenditure and some of them have none at all. Thus, distribution of such a variable is skewed with the peak at zero and has a high population variance. For this reason the well-known design-based Horvitz-Thompson estimator of the total of such a variable in finite population has a high variance as well. We are looking for another way of estimating the total. One of the ways is to use the model-based approach of a study variable. In this case the values of variables of the elements of finite population  $U$  are assumed to be generated according to some super-population model. The non-sampled values of the study variable are predicted by this model and used for the estimation of the total. If we can find a distribution model that closely resembles the distribution of the study variable, a model based-estimator may have a smaller mean square error than a design-based one. This way of estimation is given in Valliant et. al. [4]. The tobit and Heckman models ([3]) are used in our investigation ([1], [2]) for non-negative study variable which acquires many zero values.

## References:

- [1] Krapavickaitė D. (2007). Model-based estimator of total expenditure on environmental protection, *Mathematical Modelling and Analysis*. (Sub-mitted).
- [2] Krapavickaitė D. (2007). Estimation of a finite population total for a censored regression model for a study variable. *Acta Applicandae Mathematicae* (in print). Online: <http://dx.doi.org/10.1007/s10440-007-9099-9>.
- [3] Maddala G. S. (1983). *Limited-dependent and qualitative variables in econometrics*. Cambridge University Press, Cambridge.
- [4] Valliant R., Dorfman A. H., Royal R. M. (2000). *Finite Population Sampling and Inference*, John Wiley & Sons, New York.

---

<sup>1</sup> Institute of Mathematics and Informatics; Statistics Lithuania

# TEN YEARS OF BALTIC-NORDIC CO-OPERATION

Gunnar Kulldorff<sup>1</sup>

**Survey sampling** was a non-subject in the Baltic countries before 1992. No interest. No teaching. No research. Virgin soil! What happened in 1992? What has happened since then?

After five years of awakening interest and pioneering growth – first in Estonia and then in Latvia and Lithuania – the first event of co-operation on a Baltic-Nordic scale occurred in 1997. It was a **Summer School on Survey Sampling Theory, Methodology and Practice** held in Tartu, Estonia with 49 participants. 29 came from the Baltic countries (14 from Estonia, 9 from Latvia and 6 from Lithuania) and 20 from the Nordic countries (2 from Denmark, 9 from Finland, 2 from Norway and 7 from Sweden). We had 18 invited lectures, 10 practical exercises in PC class, 16 reports and two Round Table discussions.

Then followed annual **Workshops on Survey Sampling Theory and Methodology** 1998-2001 in Jurmala, Palanga, Pärnu and Jurmala (again) with 27-30 invited participants, 8-11 invited lectures, 13-20 reports with invited discussants, and Round Table discussions. The scientific level of contributed papers grew from year to year, and in 2002 the time was ripe for a **Baltic-Nordic Conference on Survey Sampling** with open participation – also from other countries. It was held in Ammarnäs, Sweden. Among the 74 participants, 17 came from the Baltic countries and 48 from the Nordic countries. We had 21 invited lectures and 30 contributed papers.

During 2003-2006 we continued with annual **Workshops** in Palanga, Tartu, Vilnius and Ventspils with 34-42 invited participants, 9-12 invited lectures, 16-25 reports with invited discussants, and Round Table discussions.

The ten co-operative meetings 1997-2006 have been complemented by many **exchange visits** by Baltic university teachers, research students and practicing survey statisticians to some Nordic university, particularly the University of Umeå in Sweden. This university has received 97 such visits 1992-2007 with a total duration of 3038 days.

Is it possible to evaluate the effects of our co-operation? I am convinced that every participant at our meetings and every exchange visitor have learned a lot from these experiences and that every participating institution – Baltic and Nordic – has also benefited. On the Baltic side we have counted the number of **Bachelor and Master theses** on survey sampling that have been written during 1995-2006:

	Estonia	Latvia	Lithuania	Total
Bachelor theses	17	14	29	60
Master theses	6	9	25	40

---

<sup>1</sup> Department of Mathematical Statistics, University of Umeå

# RETROSPECTIVE TWO-STAGE CLUSTER SAMPLING FOR MORTALITY

Seppo Laaksonen <sup>1</sup>

Two-stage sampling has been often used in surveys for households and individuals. A standard strategy is first to stratify the frame (and target) population, then to construct regional clusters within each stratum and next to choose some of these with probability proportional to size (pps), and finally to draw sampled units randomly within each cluster. This strategy requires such statistical data that gives opportunity to correctly calculate inclusion probabilities for each stage and then for each selected sampled unit. Naturally, each selection should be really based on probability principles that are not however always guaranteed due to nonup-to-date data and due to difficulties to draw the second stage sample randomly. These problems have been met in most national surveys such as the European Social Survey (ESS) where two- or three-stage cluster sampling is usual. Moreover, most surveys have a cross-sectional nature, that is, the fieldwork is concerned just the present survey period or some recent times, alternatively. Some cross-sectional surveys are continued from the same sample basis, leading to follow-up or panel surveys. There is, in addition, an option to direct the data collection backward, that is, to use a retrospective approach. This paper illustrates such an approach using the experience of the Iraq Mortality Survey (IMS) that was conducted in summer 2006. The main study variables of the IMS are deaths due to violent or non-violent reasons. Such variables are not used in surveys of developed countries since reasonably good results are available from records of death registers or lists. Such records have not been reliable in Iraq or in other conflict countries and hence survey methodology has been attempted. The 2006 IMS was the second attempt in Iraq following a similar 'household' survey from 2004, but this latter one became well-known since October 2006, due to surprisingly high mortality estimates. Both politicians, media, citizens and scientists have participated in this discussion or debate. The paper also comments this discussion while it tries to reconstruct the estimates using initial micro data.

---

<sup>1</sup> University of Helsinki and Statistics Finland

# ON-SITE SAMPLING

Thomas Laitila<sup>1</sup>

On-site sampling means intercepting respondents at public places like shopping centers, airports, recreation sites, etc, followed up by interviews and/or address collection. Pollock, Jones and Brown (1994) present a number of on-site sampling methods for angler surveys. Here interest is in the population of fishing trips and estimation of total catch and total angler effort. Statistics Sweden also makes use of on-site sampling at e.g. airports for studies of foreign tourist visitors. Multistage sampling designs and traditional estimation procedures can be used for these types of survey problems. On-site sampling is also used for survey problems where interest is in the population of visitors. For instance, Shaw (1988) considers visitors valuation of recreation sites. Other applications are found in marketing and transportation research where interest is in populations of consumers and travelers, respectively. When interest is in the population of visitors, instead of the population of visits, standard estimation procedures are not generally valid. This talk discusses the estimation problem involved and gives a review of estimators suggested in the literature. Relations to length-biased sampling and an estimator proposed by Cox (1969) are highlighted. A proposal for a new framework for on-site sampling design and estimation of population parameters is also presented.

## References:

- Cox, D, (1969). Some Sampling Problems in Technology in *New Developments in Survey Sampling*, U. L. Johnson and H. Smith, (eds.) New York: Wiley Interscience.
- Pollock, K.H, Jones, C.M. and T.L. Brown (1994). *Angler Survey Methods and Their Applications in Fisheries Management*, American Fisheries Society Special Publication 25, American Fisheries Society, Bethesda.
- Shaw, D., (1988). On-Site Samples' Regression, Problems of Non-negative Integers, Truncation, and Endogenous Stratification. *Journal of Econometrics*, **37**, 211-223.

---

<sup>1</sup> Örebro University

# ESTIMATION OF MONTHLY FIGURES FROM LABOUR FORCE SURVEY

Janis Lapins <sup>1</sup>, Martins Liberts <sup>2</sup>

The contributed paper reflects the work and results gained during the research project “Quality Improvement in Employment Statistics”. The goal of the project was to develop some practical recommendations according to estimation of monthly figures from Labour Force Survey (LFS).

Currently LFS is organised as continuous household sample survey in Latvia. Quarterly and yearly figures of employment, unemployment and other statistics are produced. Users of statistics are interested also in monthly figures of employment and unemployment statistics. The paper presents several approaches to improve the accuracy of monthly estimates.

The legal act between Central Statistical Bureau (CSB) of Latvia and the State Employment Agency (SEA) was sign last year. The legal act allows CSB to use the register of unemployed persons maintained by SEA for statistical purposes. The data from the register of unemployed persons can be used as the source of auxiliary information to build generalised regression (GREG) estimators.

Starting from 2007 the sample size of LFS in Latvia has been increased (sampling rate is approximately 2.7% for yearly sample). Despite the increase of sample size and possibility to use auxiliary information in estimation phase the accuracy for several monthly figures cannot be obtained good enough, for example monthly employment and unemployment figures of 15-24 years old population. Several model-based estimators were considered to improve the accuracy for monthly figures.

The theoretical and practical aspects of previously mentioned challenges are discussed in the paper. It was not possible to test all theoretical solutions in practical implementation because of several practical reasons, for example it was not possible to make complete linking of survey data at the unit level with auxiliary data due to lack of appropriate key variables in the survey data. Several changes in the survey organisation of LFS have been introduced since 2007. Unique personal I.D. code is included in the LFS data set for all surveyed persons starting from January 2007. It will allow linking LFS data with SEA data at the unit level. The research will be continued as soon as data from LFS 2007 will be available.

---

<sup>1</sup> Bank of Latvia

<sup>2</sup> Central Statistical Bureau of Latvia

# MEAN SQUARE ERROR OF THE GENERAL RESTRICTION ESTIMATOR

Natalja Lepik <sup>1</sup>

Knottnerus (2003) has proposed the general restriction estimator ( $\hat{\theta}^{gr}$ ) for the parameter vector  $\theta = (\theta_1, \dots, \theta_k)'$  that satisfies certain restrictions  $R\theta = c$ , where R is a constant matrix and c is a constant vector. This estimator uses the initial vector of the unbiased estimators  $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_k)'$  and its covariance matrix.

The restriction estimator has some good properties. Besides satisfying restrictions  $R\hat{\theta}^{gr} = c$ , important property is the minimum variance of  $\hat{\theta}^{gr}$  in the class of all estimators constructed on  $\hat{\theta}$ .

Knottnerus originates in his study from the unbiased estimators  $\hat{\theta}$ . But in practice there are many of the estimators with small bias (for example generalized regression estimator that is unbiased only asymptotically). How to construct the restriction estimator in this case or how the MSE of this new estimator is related to the MSE of the biased  $\hat{\theta}$ ? What can we say about bias of the restriction estimator? Many questions and problems arise in this case. Some studies and examples on these problems will be presented.

## References:

Knottnerus, P. (2003) Sample Survey Theory: Some Pythagorean Perspectives. New York: Springer

---

<sup>1</sup> Institute of Mathematical Statistics, University of Tartu, Estonia

# NONPARAMETRIC ESTIMATION WITH DOUBLE SAMPLES

Daniel Thorburn<sup>1</sup> and Boris Lorenc<sup>2</sup>

To achieve proper inference from samples that were not selected using randomisation, a double samples setup has been proposed. It consists of complementing the nonprobability sample with a probability sample, in which only auxiliary information is measured. The setup is of use in situations where ample data are available at little cost - like in web surveys - but where the selection mechanism is unknown.

The situation of double samples has thus far been treated by modelling the process of selection into the nonprobability sample, modelling the relation between auxiliary information and the study variable or modelling both. In this paper we present a nonparametric estimator: the values of the study variable - missing by design in the probability sample - are imputed according to a distance measure on auxiliary information. Our approach is a generalisation of the nearest neighbour imputation, imputing the mean of a number of closest units with the aim of reducing the variance of the estimator. For the case of univariate auxiliary information treated in the study, we also present an expression for the variance and a variance estimator for this estimator. In a simulation study, we compared this estimator with the propensity score estimator and found the nonparametric estimator to be, under the investigated conditions, preferable because of its greater robustness to variation in the underlying conditions.

---

<sup>1</sup> Department of Statistics, Stockholm University, Sweden

<sup>2</sup> Statistics Sweden, Örebro, Sweden

# ESTIMATION OF SOME INEQUALITY INDEXES

Inga Masiulaitytė<sup>1</sup>

Suppose we have a finite population  $U$  consisting of  $N$  individuals:  $U = \{1, \dots, N\}$ . Let the probability sample  $s$  of size  $n$  is drawn from the population  $U$  according to some sampling design. Denote by  $\pi_k = P(s: k \in s)$  the inclusion probability into any of the samples. Let us investigate the variable of interest  $X$  (*income*) and  $Y$  (*expenditures*) with values in the population  $x_1, x_2, \dots, x_N$  and  $y_1, y_2, \dots, y_N$ . Let these values are arranged in the ascending order.

Let us investigate indices characterizing inequality of income and expenditures of the population: *Gini coefficient*

$$G_x = \frac{\sum_{i=1}^N (2r(i) - 1)x_i}{N \sum_{i=1}^N x_i} - 1, \quad r(i) = \sum_{l=1}^N I_{x_l \leq x_i}, \quad (1)$$

and *concentration index of expenditures*

$$C_y = \frac{\sum_{i=1}^N (2q(i) - 1)y_i}{N \sum_{i=1}^N y_i} - 1, \quad q(i) = \sum_{l=1}^N I_{y_l \leq y_i}. \quad (2)$$

Index accumulating both these measures is the index of the deviation of elasticity from unity defined by Nripesh Podder:

$$I_{\eta-1} = C_y - G_x. \quad (3)$$

The attempts to estimate this index and its accuracy are made.

## References:

Nripesh Podder. The University of New South Wales (1995) *On the relationship between the Gini coefficient and income elasticity*. Sankhya. The Indian Journal of Statistics. 1995, Volume 57, Series B, Pt. 3, pp.428–432.

---

<sup>1</sup> Faculty of Mathematics and Informatics, Vilnius University, Statistics Lithuania

# CALIBRATION INSPIRED BY SEMIPARAMETRIC REGRESSION AS A TREATMENT FOR NONRESPONSE

Montanari <sup>1</sup>, Giorgio E. & Ranalli <sup>1</sup>, M. Giovanna

Nonresponse can harm the quality of the estimates of a survey. In particular, since we have to accept that those who respond are in general different from those who do not respond, bias is introduced in the estimation of population parameters. In this paper we will not deal with imputation, but only with design weights modification to adjust for bias. Commonly, a two-phase approach is used with the response mechanism as the second phase; this is based on quasi-randomization theory where the response mechanism has corresponding response probabilities assumed to be independent of the realized sample. In practice such response probabilities have to be estimated assuming a response model. Lundström and Särndal (1999) propose a simple approach for the treatment of nonresponse based on calibration. The simplification occurs in that no explicit model has to be specified for the treatment of the nonresponse mechanism. This allows for the construction of a single set of weights for all variables of interest that are as close as possible to specified initial weights (usually the design weights), while satisfying benchmark constraints on known auxiliary information. No discrimination is made within the set of auxiliary variables available to the researcher: a single set of variables is employed at the same time for nonresponse treatment, sampling error reduction and consistency among estimates. Calibration has been shown to implicitly rely on a linear regression model between the variable of interest and the set of auxiliary variables employed (see e.g. Wu and Sitter, 2001; Montanari and Ranalli, 2005). This might be inefficient when the underlying relationship is indeed not linear. We argue that the approach in Lundström and Särndal (1999) can be usefully generalized to more complex modeling through semiparametric regression (Ruppert, Wand and Carroll, 2003) without losing in simplicity. Semiparametric regression based on penalized splines has been usefully employed for model-assisted inference in the case of complete response (Breidt, Claeskens and Opsomer, 2005). More easily than with kernel smoothing, it allows for the treatment, at the same time, of categorical and continuous auxiliary variables. The first ones can be inserted parametrically, while the second ones can be accounted for nonparametrically. This provides enhanced flexibility, especially for the implicit modeling of the nonresponse mechanism. The estimator is effectively computed by penalizing the distance measure employed in the calibration minimization (e.g. Rao and Singh, 1997).

## References:

Breidt F.J., Claeskens G. and Opsomer J.D. (2005) Model-assisted estimation for complex surveys using penalised splines, *Biometrika*, 92, 4, 831–846.

Lundström S. and Särndal C.E. (1999) Calibration as a standard method for treatment of nonresponse, *Journal of Official Statistics*, 15, 305–327.

Montanari G.E. and Ranalli M.G. (2005) Nonparametric model calibration estimation in survey sampling, *Journal of the American Statistical Association*, 100, 1429–1442.

Rao J.N.K. and Singh A.C. (1997) A ridge-shrinkage method for range-restricted weight calibration in survey sampling, in: *ASA Proceedings of the Section on Survey Research Methods*, American Statistical Association, 57–65.

---

<sup>1</sup> Università degli Studi de Perugia

Ruppert D., Wand M.P. and Carroll R. (2003) *Semiparametric Regression*, Cambridge University Press, Cambridge, New York.

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

# PRACTICAL APPLICATION OF THE MODEL-BASED ESTIMATOR FOR A FINITE POPULATION TOTAL

Vilma Nekrašaitė<sup>1</sup>

A finite population, where study variable  $y$  has zero and positive values, is simulated. The 1000 simple random samples of different sizes  $n$  have drawn. For each sample, the population total  $t_y$  was estimated in three ways:

1. design-based estimator;
2. tobit-model (Greene [1]) based estimator;
3. Heckman-model (Maddala [3]) based estimator.

Also for each sample bias and variance of design-based and tobit-model based (Valliant [4], Krapavickaitė [2]) estimators were calculated. The empirical relative mean square error and average relative mean square error of these estimators were estimated and compared. Dependency of relative mean square errors on the sample size, variance of error and percentage of zeroes in the population is demonstrated.

## References:

1. GREENE, W. H. *Econometric Analysis*. New Jersey:Prentice Hall, Upper Saddle River, 2002. 802 p.
2. KRPAVICKAITĖ, D. Estimation of a finite population total for a censored regression model for a study variable. *Acta Applicandae Mathematica*. (Delivered).
3. MADDALA, G. S. *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge University Press, Cambridge, 1983. 397 p.
4. VALLIANT, R. Nonlinear Prediction Theory and the Estimation of Proportions in a Finite Population. *Journal of the American Statistical Association*. 1985, 631-641.

---

<sup>1</sup> Vilnius Gediminas Technical University, Statistics Lithuania

# M-ESTIMATORS AND U-STATISTICS IN APPROXIMATING VARIANCE OF INCOME INEQUALITY INDICES

Wojciech Niemiro<sup>1</sup> and Robert Wieczorkowski<sup>2</sup>

There are several measures of income inequality or poverty, estimated from survey data. Let  $F(y)$  be the CDF of variable  $Y$  which describes incomes in a population and  $F(y|A)$  – the CDF of  $Y$  in some subpopulation  $A$ . Let  $\eta_q$  be the  $q$ -th quantile,  $F(\eta_q) = q$  and  $\mu = \int_0^\infty yF(dy)$ . In accordance with the EUROSTAT definitions [3] we consider the following indicators:

At-risk-of-poverty rate,  $ARPR = F(\beta\eta_q|A)$  (with  $q = 0.5$  and  $\beta = 0.6$ ).

Quintile share ratio,  $S80/S20 = \int_{\eta_q}^\infty yF(dy) / \int_0^{\eta_q} yF(dy)$  (with  $q = 0.8$ ).

The Gini index,  $GINI = \frac{1}{\mu} \int_0^\infty (2F(y) - 1) yF(dy) = \frac{1}{2\mu} \int_0^\infty \int_0^\infty |y_1 - y_2| F(dy_1) F(dy_2)$ . It has been noted [2, 1] that linearization provides simple and accurate approximations for the variance of estimators of these indices. In the classical setup of mathematical statistics, when the sample  $Y_1, \dots, Y_n$  is drawn independently from an infinite population, an estimator  $\hat{\theta}_n$  of a parameter  $\theta$  is expressed as

$$\hat{\theta}_n = \theta + \frac{1}{n} \sum_{i=1}^n L(Y_i) + o_p\left(\frac{1}{\sqrt{n}}\right).$$

We provide an alternative way of deriving linearization formulas, based on rigorous results on M-estimators and U-statistics [5, 4]. Classical results are adapted to the setup of survey statistics. We consider sampling without replacement from a finite population.

## References

- [1] Berger, Y.G. and Skinner, C.J. (2003), Variance estimation for a low-income proportion, *J. Royal Statist. Soc. ser. C* vol. 52, 457–468.
- [2] Deville, J.-C. (1999), Variance estimation for complex statistics and estimators: linearization and residual techniques, *Survey Methodology*, Vol. 25, No. 2, 193–303.
- [3] Eurostat (2004), Statistics on income, poverty & social exclusion (IPSE) and EU/SILC (Statistics on income and living conditions), Methodology of calculation of common cross-sectional EU indicators, Eurostat-Luxembourg.
- [4] Niemiro, W. (1992), Asymptotics of M-estimators defined by convex minimization, *Ann. Statist.* 20, 3, 1514–1533.
- [5] Serfling, R.J. (1980), *Approximation Theorems of Mathematical Statistics*, Wiley and Sons.

---

<sup>1</sup> Central Statistical Office of Poland and Faculty of Mathematics and Computer Science, Toruń, Poland

<sup>2</sup> Central Statistical Office, Warsaw, Poland

# **EBLUP ESTIMATION OF SMALL AREA TOTALS UNDER LINEAR MIXED MODEL FOR ROTATED PANEL DATA**

Kari Nissinen <sup>1</sup>

The paper considers EBLUP estimation of small area totals, when unit-level data from a rotated panel design is available. The data is modelled with a three-level variance component model, which is an extension of the well-known nested error regression model. The formulas of EBLUP estimator and its MSE estimator are given for the rotated panel case and the performance of the estimators is examined by a simulation study. The simulation results show that both EBLUP and its MSE estimator perform well and utilizing the rotated panel data instead of cross-sectional data highly increases the accuracy of small area estimation totals.

---

<sup>1</sup> Department of mathematics and statistics, University of Jyväskylä

Changes in the psychosocial work environment in Denmark  
**BETWEEN 2000 AND 2005**

Ole Olsen <sup>1</sup>, Helene Feveile <sup>1</sup>, Elsa Bach <sup>1</sup>

The Danish Ministry of Employment had prioritized improvements of the psychosocial work environment as one of four target areas for the period 2000-2005. The aim was that the percentage of employed persons exposed to twelve listed psychosocial risk factors should have diminished overall by 5%. It had been decided that changes should be monitored in the Danish Work Environment Cohort Study (DWECS), a study carried out every fifth year. Analysis should be straightforward, easy to comprehend and per protocol. However, several design aspects were changed between the 2000 and the 2005 data collection. The aim of our presentation is to present our analysis and the various decisions we had to make.

**Material and Methods** – In 2000 the sample was drawn from the centralised civil register to be representative for all inhabitants aged 18-69 with a permanent address in Denmark. Telephone numbers were identified and telephone interviews attempted; if these attempts were unsuccessful, face-to-face interviews at home were attempted. The total response rate was 75%. Before data collection began in 2005 it was decided primarily to use postal questionnaires instead of telephone interviews and to offer the option to reply on the internet. Because we expected differences in response patterns between telephone interviews and questionnaires, a restricted subset of the sample was randomised to telephone interview. Additional changes were imposed on the previously simple design. More details will be presented.

**Results** – The response pattern depended significantly on the data collection instrument for 8 of the 12 psychosocial risk factors. Overall the changes from 2000 to 2005 indicated an increase in the prevalence of problems; however, changes were in both directions. For some outcomes the changes were consistent across gender, age and industrial sector; for other outcomes the pattern was less obvious.

**Discussion** - Due to the political nature of the purpose, a straightforward analytic approach and presentation of results was desirable. Our chosen strategy for presentation of results will be discussed. The published report in Danish is accessible as an internet publication ([www.arbejdsmiljoforskning.dk/upload/psyk2000-2005.pdf](http://www.arbejdsmiljoforskning.dk/upload/psyk2000-2005.pdf)).

---

<sup>1</sup> National Research Centre for the Working Environment

# SOME EXAMPLES OF THE NONLINEAR CALIBRATION

Aleksandras Plikusas<sup>1</sup>

The calibrated estimators of the finite population total is widely used in a current survey practice. The definition and main properties of calibrated estimators of total is presented by Deville and Särndal [1]. We consider the estimation of more complex parameters using auxiliary variables. The main examples are the ratio of two totals and finite population covariance. The proposed estimators are constructed by some calibration equation which is nonlinear with respect to the calibrated weights we are looking for. We call this procedure the nonlinear calibration. In the case of the estimation of the ratio, the calibration equations have an explicit solution under some standard loss functions [3]. Some special case of the calibrated estimator of ratio when only one system of weights is used, is examined by Krapavickaitė and Plikusas in [2]. The second example is the finite population covariance, which is more complicated. The calibration equations may be defined in a different ways and

several systems of calibrated weights may be used. The nonlinear calibration equations has no explicit solution. The calibrated weights are expressed by iterative equations. Some calibrated estimators of the finite population covariance are studied in Plikusas and Pumputis [4]. The simulation results show that incorporation of several weighting systems can lead to a more precise estimators.

## References:

- [1] J.-C. Deville, C.-E. Särndal. Calibration Estimators in Survey Sampling, Journal of the American Statistical Association, 87, 376-382 (1992).
  
- [2] D. Krapavickaitė, A. Plikusas. Estimation of a Ratio in the Finite Population. Informatica, 2005, 16(3), p. 347-364.
  
- [3] A. Plikusas. Calibrated estimators of the ratio. Lithuanian Math. J., 41 (special issue), 457-462 (2001).
  
- [4] A.Plikusas, D. Pumputis. Calibrated estimators of the population covariance, Acta Applicandae Mathematicae (to appear, 2007).

---

<sup>1</sup> Institute of Mathematics and Informatics, Statistics Lithuania

# ON THE ESTIMATION OF THE VARIANCE OF CALIBRATED ESTIMATORS OF THE POPULATION COVARIANCE

Dalius Pumputis <sup>1</sup>

Calibrated estimators are widely used in finite population statistics to improve the quality of estimators, using auxiliary information. Such type of estimators are widely used in official statistics, especially in social surveys. The calibration technique for estimating of a finite population totals was presented by J.-C. Deville and C.-E. Särndal [1]. The estimation of more complicated parameters, using auxiliary variables, is not widely studied in the literature. One type of the calibrated estimator of the ratio of two totals was considered by D. Krapavickaitė and A. Plikusas [2], and A. Plikusas [3], [4]. Some calibrated estimators of the population covariance were introduced in A. Plikusas, D. Pumputis [5]. They are constructed using different calibration equations and different loss functions. The calibrated estimators of the population covariance are more efficient compared to the straight estimators provided the auxiliaries are well correlated with the study variables. In the case of low correlated auxiliaries, all estimators are of the similar quality.

The calibrated weights in the case of a nonlinear calibration are defined by some recurrent equations and the estimation of the variance of corresponding estimators becomes problematic. In the paper some estimators of the variance of calibrated estimators of the population covariance are considered. The estimator derived by using the approximate Taylor linearization and the Bootstrap variance estimator are examined by the simulation. The estimators considered are compared with the empirical variance.

## References:

- [1] J.-C. Deville, C.-E. Särndal, Calibration estimators in survey sampling, *Journal of the American Statistical Association*, 87, 376-382, 1992.
- [2] D. Krapavickaitė, A. Plikusas. Estimation of a Ratio in the Finite Population. *Informatika*, 16(3), 347-364, 2005.
- [3] A. Plikusas, Calibrated estimators of the ratio. *Lithuanian Math. J.*, 41 (special issue), 457-462, 2001.
- [4] A. Plikusas, Calibrated weights for the estimators of the ratio, *Lith. Math. J.*, 43, 543-547, 2003.
- [5] A. Plikusas, D. Pumputis, Calibrated estimators of the population covariance, *Acta Applicandae Mathematicae*, (to appear) 2007.
- [6] C.-E. Särndal, B. Swensson, J. Wretman, *Model Assisted Survey Sampling*, Springer-Verlag, New York, 1992.

---

<sup>1</sup> Vilnius Pedagogical University, Institute of Mathematics and Informatics, Lithuania

# INVERSE PROBABILITY OF CENSORING WEIGHTING METHOD IN SURVIVAL ANALYSIS BASED ON SURVEY DATA

Marjo Pyy-Martikainen<sup>1</sup>, Leif Nordberg<sup>2</sup>

In survival analysis based on survey data, attrition implies that the survival time of some individuals is only partially observed: it is only known that the true time exceeds that observed. These survival times are called right-censored. To simplify the analysis, it is usually assumed that the censoring process is independent. This means that the hazard of censoring does not depend on future failure time. In practice, the assumption of independent censoring may not always hold. For example, when analysing unemployment spells based on survey data, it may very well be that economically inactive persons also may not have a large interest in surveys. This latent trait, "reluctance", creates a dependency between a long unemployment duration and a high probability of attrition. Dependent censoring may cause a bias in the estimated distribution of survival times and in the estimated covariate effects. Robins (1993) introduced an inverse probability of censoring weighting (IPCW) method that adjusts for bias due to dependent censoring. Lawless (2003) considered the use of IPCW method in survival analysis based on survey data. To our knowledge, however, there are no empirical applications of the method in the survey data context. In the first phase of our study, we use Monte Carlo methods to study the performance of IPCW method in a hypothetical 2-wave panel survey. In the second phase of the study, we apply the method to the Finnish subset of ECHP data to show how the method performs in a real data set.

## References:

Lawless, J. F. (2003). Censoring and Weighting in Survival Estimation from Survey Data. SSC Annual Meeting, June 2003. Proceedings of the Survey Methods Section, 31-36.

Robins, J. M. (1993). Information Recovery and Bias Adjustment in Proportional Hazards Regression Analysis of Randomized Trials Using Surrogate Markers. Proceedings of the Biopharmaceutical Section, American Statistical Association, 24-33.

---

<sup>1</sup> Department of Economics and Statistics, Åbo Akademi University and Statistics Finland

<sup>2</sup> Department of Economics and Statistics, Åbo Akademi University

# FROM THEORY TO PRACTICE : HOW TO CONDUCT SURVEYS AMONG DIFFICULT TO REACH POPULATIONS?

Martine Quaglia <sup>1</sup>, Géraldine Vivier <sup>1</sup>

Adapting the methods developed in the US surveys among the homeless population (Burt et Cohen, 1989; Dennis et Iachan, 1993) to the French environment, M. Marpsat and J.M. Firdion (2000), answering a demand from the French National Council on Statistical Information, conducted several local surveys among users of services such as shelters, soup kitchen, day centres (1995-1998). This method, through which were met people living in shelters as well as in the streets, at friend's or relative's or in their own lodging (but using the soup kitchen), was also used by the INSEE (National Institute of Statistics) in 2001 on a national level.

These surveys were, somehow, covering a "restricted" part of the homeless population : French speaking users of services. To estimate the coverage of these studies in relation to the characteristics of the homeless population at large, two other surveys were later conducted among users of services who didn't speak French and rough sleepers met by outreach services (2002). More recently, this method was used in a survey conducted among drug users (2002, 2004). Through the necessary mediation of the services run by NGOs and institutions depending on social work and health authorities, different problems arose. Our experiences of surveys among the homeless population and drug users conducted by Ined (1995-2005) will be exposed, showing the different questions we had to cope with, and the answers provided, from sampling design to the collection of data, as well as the necessary adaptations to the field realities and to the populations we were aiming at.

**Key words:** Difficult To Reach Populations Survey Design - Indirect Sampling - Adaptation To The Field

---

<sup>1</sup> Survey Department, National Institute for Demographic Studies, France

# INTERVIEWING BOTH EMPLOYEES AND EMPLOYERS, A MIXED MODE DATA COLLECTION FOR A MATCHED SURVEY

Martine Quaglia<sup>1</sup>, Cécile Lefevre<sup>1</sup>, Ariane Pailhe<sup>1</sup>, Anne Solaz<sup>1</sup>, Ana Maria Noel<sup>1</sup>,  
Tatiana Vichneskaïa<sup>1</sup>, Bernard De Cleat<sup>1</sup>

This paper presents an original and unusual data collection, developed for a survey on the theme of conciliation between professional life and family life which was carried out in 2004-2005 by INED in partnership with INSEE (National Institute for Statistics and Economic Studies). This research was made up of two surveys: one conducted among individuals while the second addressed their employers. The survey, therefore, concerned very different statistical entities: individuals, couples, and households on one hand, public and private employer establishments on the other. The heterogeneity of the two types of respondents, the duality of the two sections, the need to obtain a high response rate in order to make the final file pairs usable led us to use several data collection methods.

## **A multi-mode collection survey: the establishment survey**

Surveying individuals is a fairly standardized procedure: CAPI was used to address around 9500 people aged between 20 and 49. When the respondent was employed, the address and size of the establishment was requested. Thus, the second part of the survey was based on establishments employing individuals who had already been interviewed, provided that they employed 20 or more employees.

A self-administered questionnaire was then sent by post to each establishment. To improve a fairly low spontaneous response rate several strategies were adopted and will be described at this session:

- A follow-up was carried out through reminder letters and telephone calls.
- Internet was added to the post as a possible mode of answer.
- Downloading the PDF questionnaire off the survey's website was possible

## **Who answers the questionnaire?**

The survey being about the day to day professional environment of the individuals (organisation of working hours, holidays, etc.) the establishment in which the respondent is working is clearly the target of the survey and has to answer the questionnaire. In fact; what seems to be evident from the point of view of the research reveals to be more complicated from the respondent point of view. Given a few examples, the question of establishment Vs institution, company or firm will also be discussed as a key question for business surveys.

**Key words:** multi-mode data collection, matched employee/employer surveys, internet surveys, business surveys

---

<sup>1</sup> INED (National Institute for Demographic Studies), France

# REGRESSION COMPOSITE ESTIMATION WITH APPLICATION TO THE FINNISH LABOUR FORCE SURVEY

Riku Salonen<sup>1</sup>

The design of the Finnish Labour Force Survey (LFS) is a complex rotating panel. Survey is repeated over time with partially overlapping samples. In the case of repeated surveys with partial overlapping, it has been seen that to utilize entire information collected in the previous waves is very advantageous (for example Singh et al., 2001). Currently, the Finnish LFS uses the generalized regression (GREG) estimator. It is based only on the current quarter's data. The LFS does not use the fact that 60 % of the LFS sample is common between consecutive quarters to improve estimates. Exploiting sample overlap over time to improve efficiency of estimates can be done via calibration by using a certain composite estimator. A method termed regression composite (RC) estimator extends the GREG-estimator in the sense that it takes advantage of the correlations over time induced by sample overlap to achieve gains in efficiency. The RC-estimator introduced by Singh, Kennedy and Wu (2001), Fuller and Rao (2001) and Gambino, Kennedy and Singh (2001). Methods have also been studied in Bocci and Beaumont (2005). The RC-estimator can be computed by adding control totals and auxiliary variables to the current GREG-estimation program (for example CLAN). Use of additional control totals based on previous quarter's estimates, and that the auxiliary variables associated with these estimated control totals.

**Key words:** Complex rotating panel; Auxiliary information; Composite estimation.

## References:

- BOCCI, C., and BEAUMONT, J.-F. (2005). A Refinement of the Regression Composite Estimator in the Labour Force Survey. Internal report, Statistics Canada.
- FULLER, W.A., and RAO, J.N.K. (2001). A Regression Composite Estimator with Application to the Canadian Labour Force Survey. *Survey Methodology*, 27, 45-51.
- GAMBINO, J., KENNEDY, B., and SINGH, M.P. (2001). Regression Composite Estimation for the Canadian Labour Force Survey: Evaluation ja Implementation. *Survey Methodology*, 27, 65-74.
- SINGH, A.C., KENNEDY, B., and WU, S. (2001). Regression Composite Estimation for the Canadian Labour Force Survey with a Rotating Panel Design. *Survey Methodology*, 27, 33-44.

---

<sup>1</sup> Statistics Finland

# **SAMPLE SURVEY OF FARMS IN UKRAINE: CURRENT STATE AND PROSPECTS**

Nataliya Skachek <sup>1</sup>

Among the agricultural enterprises in Ukraine farms make up the largest group with the share of 73.3% in total number of agricultural legal entities. But most farms are small enterprises with average agricultural lands under 100 hectares. That is why the share of farms in total agricultural lands of all producers is only about 10%. Their contribution to gross production of agriculture is also small. It was only 4.2% in 2005. Of course, it is not efficient to observe all the farms; in this case probability sampling principles provide time-cost trade-off and good results.

In Ukraine sample survey of farms has been implemented since 2002. It is conducted once in two years. At the beginning Farm Register was used as the sampling frame. Now Agricultural Register containing Farm Register as a part is used for this purpose. In each district systematic sampling is applied to the list of farms ranked by the area under crop. If the number of farms in district is less than fifty or fifty all the farms are automatically included to the sample. In districts with more than fifty farms systematic sampling is used and 20% of all farms are selected.

At the moment there are some disputes about sample design. So it is important to explore different approaches to sampling and compare their results.

## **References**

Cochran, William G. (1977) Sampling Techniques, John Wiley & Sons, New York.

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

---

<sup>1</sup> Scientific and Technical Complex of Statistical Research, Ukraine

# SELECTION OF VECTOR OF AUXILIARY INFORMATION FOR GENERALIZED REGRESSION ESTIMATOR (GREG)

Milda Slickute-Sestokiene <sup>1</sup>

Statistics Lithuania has the full range of labour statistics that meet the timeliness and demands of Eurostat and national needs. The challenge is to keep this quality and timeliness and to publish even more detailed information and at the same time spare costs.

Users need more and more statistical information and at the same time respondents want to get less and less questionnaires. That enforces Statistics Lithuania to seek for new methods for estimation of statistical information required. Administrative sources (e.g. data of Social Insurance) for statistical purposes in Lithuania become available only few years ago. As soon as they become available they started to play a significant role trying to increase the quality of the results and to diminish the burden of respondents. The more detailed breakdown is also needed.

This contribution is devoted to the usage of administrative sources at the stage of estimation. The problem of selection the vector of auxiliary information is considered. Quarterly data of earnings is used for illustration. Identification of the vector of auxiliary information causes a lot of problems when various different breakdowns are needed.

The case when different auxiliary variables are used for different study variables is also examined.

---

<sup>1</sup> Statistics Lithuania, Lithuania

# RESTRICTION ESTIMATOR FOR DOMAINS

Kaja Sõstra<sup>1</sup>

There are different small area estimation (SAE) methods developed to improve the precision of estimates. Using different methods for estimating small domains and large sub-population may cause consistency problems between the estimates. The topic of the presentation is to solve the problem of consistency implementing general restriction estimator for small domains.

Two sampling designs are briefly introduced: simple random sampling (SI) and hypergeometric (HG) sampling designs. These are common designs in official statistics. Simple random sampling and stratified simple random sampling are used for sample surveys of businesses, e.g. businesses are stratified according to number of employees and economic activity and SI design is used in every strata. Hypergeometric design describes selection mechanism in social surveys where we select individuals from population register and include all persons of their households into the sample.

It is known that the estimators of two domains are usually dependent, unless these domains are sampled independently (like strata). General forms of covariances between domain estimators and between estimators of domain and the population total are presented.

In practical situations it may occur that the same population parameter is estimated in different surveys. Often the estimates from different surveys have to obey a set of restrictions. For example the total net income of households from wage labour estimated in household budget survey has to correspond with total net wages estimated in labour force survey. Similarly different estimators from one survey have to satisfy some conditions (estimated totals of sub-populations have to sum up to the estimated population total). One solution of the described problem is general restriction (GR) estimator proposed by Knottnerus (2003).

Monte-Carlo simulations were performed to test the theoretical results. Population for simulations is based on the real LFS dataset. General restriction estimators of domains were calculated under SI- and HG-designs. Theoretical and empirical covariance matrices were calculated and compared. Simulation results are presented.

---

<sup>1</sup> Statistics Estonia

# CORRECTING THE REGRESSION ESTIMATOR FOR AN ABUNDANCE OF AUXILIARY VARIABLES

Silke Burestam <sup>1</sup> and Daniel Thorburn <sup>1</sup>

Today the amount of auxiliary information that is available for surveys has increased and also the technical possibilities to use it. Auxiliary information is not only used to decrease the random error but also when correcting for unbalanced samples due to non-response, imperfect frames and other problems. Auxiliary information is often used e.g. in post stratification or regression estimation to reduce the variance. But it is well known that this procedure may lead to the opposite, if there are too many auxiliary variables. Many simple rules have been suggested for deciding when it may be dangerous e.g. the weights should not change by a factor of more than two. These limits often require that one must subjectively select a few of the available auxiliary variables. Here we suggest a method which allows to use all auxiliary variables in regression estimators, but automatically weighs down their influence if they do not contribute much to the precision of the estimates.

Consider a finite population  $U$ , with  $N$  elements, from which a sample  $s$  with  $n$  units are taken. Every unit is characterized by an unknown study variable  $y_k$  and a column vector consisting of  $J$  known auxiliary variables  $\mathbf{x}_k = (x_{1k}, \dots, x_{Jk})'$ .

Let  $p(s)$  be the probability that the sample  $s$  is observed,  $\pi_k = \Pr(k \in s)$  and  $\pi_{kl} = \Pr(\{k, l\} \subset s)$ . The totals for the auxiliary variables are denoted by  $t$  e.g.  $t_{xi} = \sum_U \mathbf{x}_{ik}$ .

The general regression estimator (GREG) is written

$$\hat{t}_{y\pi} = \hat{t}_{y\pi} + \sum_{i=1}^J \hat{B}_i (t_{xi} - \hat{t}_{xi\pi})$$

where  $\hat{t}_{y\pi}$  is the Horwitz-Thomson estimator,  $\sum_{k \in s} y_k / \pi_k$  and where  $\hat{B}_i$  is the estimated regression coefficient.

If the regression coefficients are badly estimated the total estimator becomes bad if their impacts are large, i.e. if  $t_{xi} - \hat{t}_{xi\pi}$  is large. We suggest the following corrected regression estimator to avoid too large effects.

$$\hat{t}_{y_{areg}} = \hat{t}_{y\pi} + a \times \hat{b}(t_x - \hat{t}_{x\pi})$$

where  $a$  should be chosen so that the total variance is minimised (here given in one dimension for simplicity). Developing the estimate around the true values gives

$$\hat{t}_{y_{areg}} = \hat{t}_{y\pi} + a(\hat{b} - b)(t_x - \hat{t}_{x\pi}) + ab(t_x - \hat{t}_{x\pi}),$$

where the uncertainty in the three terms mainly depends on the uncertainty in  $\hat{t}_{y\pi}$ ,  $\hat{b}$  respektive  $\hat{t}_{x\pi}$ . We use Taylor approximation to compute its variance. It turns out that it is minimised by

$$a = \frac{b^2 \sigma_x^2}{\sigma_b^2 (t_x - \hat{t}_{x\pi})^2 + b^2 \sigma_x^2}$$

If  $b$  was known, its variance would be zero and the optimal  $a$ -value would be 1. This would give the usual regression estimator. But since  $\hat{b}$  is an estimate its variance is positive and the optimal  $a < 1$ . In order to make this expression computable we replace the parameters by their estimates which gives

$$\hat{t}_{y_{areg}} = \hat{t}_{y\pi} - \frac{1}{1 + (\hat{\sigma}_{y|x}^2 / \hat{\sigma}_{xy}^2)(t_x - \hat{t}_{x\pi})^2 / N(N-n)} \times \hat{b}(t_x - \hat{t}_{x\pi})$$

In the paper the theory is developed with many auxiliary variables and it is illustrated with simulations.

<sup>1</sup> Stockholm University

# DIFFICULTIES IN THE ESTIMATION AND QUALITY ASSESSMENT OF SERVICE PRODUCER PRICE INDICES

Markus Gintas Šova <sup>1</sup>, John Wood <sup>1</sup> and Ian Richardson <sup>1</sup>

The estimation of Service Producer Price Indices (SPPIs) faces challenges which are not encountered, or not encountered to the same extent, for Goods Producer Price Indices. It is often difficult for respondents to define individual service products for which prices can be identified and provided regularly on a constant quality basis. As a consequence, greater use is made of non-standard data collection methods, as in the use of model contract prices or unit value prices. These introduce a greater potential for bias and require careful monitoring. For some products it is difficult, if not impossible, to separate the business and retail components of turnover and prices. This complicates the calculation of separate indices for intermediate and final consumption. Some problems arise because SPPIs are still in the first stages of development. International standards for the specification and classification of service products are not well-developed. In the UK there is currently no equivalent for services of the annual ProdCom survey for goods to act as a sampling frame for price collection and as a source of timely turnover data for weights. Because of these issues, assessing the quality of SPPIs involves additional concerns: the effects of non-standard data collection methods; the suitability of indices for use in intermediate or final consumption; the reliability of sampling frames for price collection.

This paper describes how the UK Office for National Statistics is tackling the questions of measuring quality, setting quality standards and creating procedures to monitor and review the quality of SPPIs.

---

<sup>1</sup> ONS, UK

# INDICATOR OF STRENGTH OF AUXILIARY INFORMATION: A SIMULATION STUDY

Karolin Toompere<sup>1</sup>

Non-response is a very actual problem in sample surveys. It may affect the results and cause bias in the estimates. Therefore, it is very important to use an estimation technique that works well also under non-response. One possibility is to use a calibration estimator. In my study I consider the estimator:

$$\hat{Y}_w = \sum_r d_k (1 - \lambda'_r \mathbf{x}_k) y_k,$$

where  $\lambda'_r = (X - \sum_r d_k x_k)' (\sum_r d_k x_k x'_k)^{-1}$ ,  $d_k$  is a sampling weight and  $\mathbf{x}_k$  is an auxiliary vector for object  $k$ .

Effectiveness of the calibration method depends on the strength of the auxiliary information. That brings up the question which variables to use in the auxiliary vector. Särndal and Lundström (2005) introduced an indicator IND1 that helps to make this decision. In its special case

$$IND1 = \frac{\sum_r d_k (v_k - \bar{v})^2}{\sum_r d_k},$$

where  $v_k = 1 + \lambda'_r x_k$  and  $\bar{v}$  is a design weighted average of  $v_k$  over respondents.

In my simulation study I assume considerably high nonresponse (30% and 50%) with response probabilities depending on the study variable. I study the bias of the estimator  $\hat{Y}_w$ , particularly, how does adding different auxiliary variables change the bias. Paralelly, I observe the indicator IND1. Under special interest is the question, to what extent does the indicator show the strength of the auxiliary vector and how does IND1 act in different situations. It is studied whether IND1 acts differently depending on how well the auxiliary variables are correlated with the study variable, whether adding variables to the auxiliary vector always increases the non-response indicator, how does the variability of IND1 depend on the number and nature of used auxiliary variables, do the results differ for different nonresponse-rates etc.

## References:

Särndal, C.-E., Lundström, S. (2005) Estimation in surveys with nonresponse. John-Wiley and Sons.

---

<sup>1</sup> Institute of Mathematical Statistics, University of Tartu, Estonia

# ESTIMATION UNDER RESTRICTIONS

Imbi Traat<sup>1</sup>

Often, a functional relationship exists between survey parameters estimated from one or several surveys. For example domain totals have to sum up to the population total, monthly totals have to sum up to the yearly total, expenditure and savings have to be equal to the total income, parameter studied in one survey is functionally related to the parameter studied in several other surveys, etc. In spite of the known relationships between parameters, the estimators usually do not satisfy these relationships. However, the latter is often desirable. For solving this inconsistency problem Knottnerus (2003) has proposed a General Restriction (*gr*) estimator.

Let the parameter vector  $\theta = (\theta_1, \theta_2, \mathbf{K}, \theta_k)'$  satisfy restrictions  $R\theta = c$ , where  $R$  is a constant matrix and  $c$  a constant vector. Let  $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \mathbf{K}, \hat{\theta}_k)'$  be a vector of initial estimators with non-singular covariance matrix  $V$ , then the *gr*-estimator and its covariance matrix are:

$$\hat{\theta}^{gr} = \hat{\theta} + K(c - R\hat{\theta}), \quad \text{with } K = VR'(RVR')^{-1},$$

$$V(\hat{\theta}^{gr}) = (I - KR)V.$$

It is known that for a given  $\hat{\theta}$  the estimator  $\hat{\theta}^{gr}$  is optimal (in the sense of variances).

In the lecture the restriction estimator is introduced and some of its properties are displayed. Further, the relationship between restriction estimator and minimum variance greg-estimator is explained. Finally, the conditional restriction estimator is considered. The latter is meant for the case where part of the estimators under restrictions should be considered as fixed quantities. Examples are given. More specific studies on the restriction estimator are presented in the lectures of supervised students Kaja Sõstra and Natalja Lepik.

## References:

Knottnerus, P. (2003) Sample Survey Theory: Some Pythagorean Perspectives. New York: Springer.

---

<sup>1</sup> Institute of Mathematical Statistics, University of Tartu, Estonia

# MEASURING ITEM NON-RESPONSE OF DIARY DATA IN TIME USE SURVEYS

Paavo Väisänen <sup>1</sup>

National statistical offices often use non-response rates as one measure of data quality. Unit nonresponse and item non-response rates are reported for individual questions and study variables. Diary data consist of episodes where missing episodes are difficult to observe but some kind of an idea of the item non-response of diaries can be formed by observing the average numbers of episodes and the totals of time used for secondary activities. In the diaries, an episode is defined as a time slot denoted by the same code. Unobserved item non-response arises when a respondent forgets to record an activity in the diary, and this situation occurs when, for instance, a person travels home from work and stops for shopping. If the shopping stop is not recorded then the number of episodes is two episodes too low. In the Harmonised European Time Use Survey Data Base (HETUSDB), numbers of activity episodes and totals of simultaneous secondary activities are used as quality measures of diary keeping. Several reasons, such as interviewer effect, respondent's education, and motivation to keep a diary, coding, etc., influence the quality of diary data. A large number of episodes and a high total for secondary activities indicate valid diary data. The respondent's time use has an unwanted impact on this measure and, for example, persons with long working hours have usually fewer activities to report, the consequence of which is low number of episodes. The number of episodes has been analysed by using the diary data from the Finnish Time Use Survey. Comparisons can be made between the fourteen countries which will be included in the Harmonised Data Base, but comparisons with other time use surveys, in which different diaries, days, instructions for diary keeping or coding are used, are not valid. Around 20-25 activity episodes are usually regarded as a reasonable value for well filled diaries. In the HETUSDB, the number of episodes was the highest in Sweden (26 episodes) and the lowest in France (19 episodes). The average total for secondary activities was 170-190 minutes, and slight correlation (from 0.11 to 0.34) was found between these quality measures. The amount for reported secondary activities was the highest in France (351 minutes) while the lowest total for simultaneous activities was reported in Spain (82 minutes). The number of activities depended on gender, age and education level. The harmonised survey included two diaries which were rather burdensome to fill in; therefore, the numbers of episodes were lower in the second diary. The means of the episodes were 23.5 for the first diary and 22.9 for the second diary, and the totals of the respondents for secondary activities were 178 and 184 minutes, calculated with data from all the Data Base countries.

**Table** Means of episode numbers and totals for secondary activities by diary day

HETUSDB countries	Means of episode numbers			Totals for secondary activities		
	1st day	2nd day	3rd day	1st day	2nd day	3rd day
Sweden	27.0	25.7		195	187	
Norway	26.7	25.6		136	130	
Finland	26.5	25.1		194	170	
Poland	25.1	24.4		200	197	
Germany	24.7	24.2	23.0	262	250	268
UK	24.2	22.9		178	169	
Estonia	23.5	22.4		153	137	
Lithuania	22.7	21.9		119	112	
Italy	22.2	-		183	-	
Slovenia	21.4	21.1		205	202	
Spain	21.0	-		82	-	
Latvia	20.7	19.9		114	150	
Bulgaria	19.9	19.6		210	256	
France	18.8	-		351	-	

**Keywords:** Item non-response, quality measure, diary data

<sup>1</sup> Statistics Finland

# ANNUAL GROWTH RATES DERIVED FROM SHORT TERM STATISTICS AND ANNUAL STRUCTURAL BUSINESS STATISTICS

Pieter Vlag <sup>1</sup>, Koert van Bommel<sup>1</sup>

During the last five years annual growth rates derived from the short term statistics are slightly, but systematically, lower than growth rates derived from the annual structural business statistics. Further investigation revealed that this difference, observed in the Netherlands, is mainly related to a different weighting scheme. Enterprise populations, to which the survey data of both the short term statistics and the annual structural business statistics are weighted, are based on the business register of Statistics Netherlands. However, for the short term statistics the enterprise population is corrected for administrative changes in the register between two successive periods. This correction is not applied to the annual business statistics. Other differences between short term and annual statistics are related to 1) the estimation for the number of inactive enterprises, 2) outlier detection and 3) weighting of enterprises with changing activities.

A simulation study showed that annual growth rates derived from short term statistics and annual business statistics are quite similar, if the same populations and outlier procedures are used. When applying the latter, however, the month-to-month growth rates of the short term statistics are more scattered. The most likely explanation is that accidental fluctuations in the data become more visible, because the short term statistics are based on a relatively small survey. To reduce this problem, Statistics Netherlands is now conducting simulation study by combining survey data with administrative sources.

---

<sup>1</sup> Statistics Netherlands

# LINEAR ESTIMATION UNDER MODEL-DESIGN APPROACH WITH SMALL AREA EFFECTS

Jacek Wesolowski<sup>1</sup>

Assume that the population  $U = \{1, \dots, N\}$  is partitioned into  $M$  disjoint small areas  $(U_m)_{m=1, \dots, M}$ , i.e.  $U = \bigcup_{m=1}^M U_m$ . For any small area  $U_m$  a non-random vector of auxiliary variables (small area effects)  $\underline{x}_m = (x_{m,1}, \dots, x_{m,q})^T$  is given,  $m = 1, \dots, M$ .

With each element  $i \in U$  a random variable  $Y_i$  is associated. Assume that the random vector  $\underline{Y} = (Y_1, \dots, Y_N)^T$  has the following structure:

$$Y_i = \underline{x}_m^T \underline{\beta} + u_m + \varepsilon_i$$

for any  $i \in U_m$ ,  $m = 1, \dots, M$ , where  $u_1, \dots, u_M$  are iid zero mean variables with the variance  $v^2$  and  $\varepsilon_1, \dots, \varepsilon_N$  are iid zero mean variables with the variance  $\sigma^2$ . The vectors  $\underline{u} = (u_1, \dots, u_M)$  and  $\underline{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_N)$  are independent

Sampling design  $p$  is a distribution of indicators of elements being sampled, i.e.  $P(\underline{I} = \underline{\delta}) = p(\underline{\delta})$ , where  $\underline{I} = (I_1, \dots, I_N)^T$ , with  $I_i = 1$  if  $i$ th element is chosen to the sample, otherwise it is 0,  $i \in U$ , and  $\underline{\delta} = (\delta_1, \dots, \delta_N) \in \{0, 1\}^N$ . With  $p$  we associate a vector  $\underline{\pi} = (\pi_1, \dots, \pi_N)^T$  of inclusion probabilities of the first order, i.e.  $\pi_i = P(I_i = 1)$ ,  $i \in U$ , a related diagonal matrix  $\mathbf{\Pi} = \text{diag}(\underline{\pi})$  and an  $N \times N$  matrix  $\mathbf{P} = [\pi_{ij}]$  of inclusion probabilities of the second order, i.e.  $\pi_{ij} = P(I_i = 1, I_j = 1)$ ,  $i, j \in U$ .

The sampling plan is non-informative, that is  $\underline{I}$  and  $(\underline{u}, \underline{\varepsilon})$  are independent.

Additionally we introduce an  $N \times M$  matrix  $\tilde{\pi}$  whose  $m$ th column  $\underline{\pi}_m$  is a vector having  $i$ th component equal  $\pi_i$  if  $i \in U_m$  and 0 if  $i \notin U_m$ ,  $m = 1, \dots, M$ , and an  $N \times N$  matrix  $\tilde{\mathbf{P}}$  which is a block-diagonal matrix, with  $m$ th diagonal block  $\tilde{\mathbf{P}}_m$  associated to  $m$ th small area in the sense that  $\tilde{\mathbf{P}}_m = [\pi_{ij}]_{i,j \in U_m}$ , i.e. it is a matrix of second order inclusion probabilities for the restriction of  $p$  to the set  $U_m$ ,  $m = 1, \dots, M$ .

Let  $\underline{Z} = (Z_1, \dots, Z_N) = \text{diag}(\underline{I})\underline{Y}$  be the vector of observations, i.e.  $Z_i = Y_i I_i$ ,  $i \in U$ .

Note that  $\mathbb{E} \underline{Y} = \mathbf{e} \mathbf{X} \underline{\beta}$  and

$$\mathbb{E} \underline{Z} = \tilde{\pi} \mathbf{X} \underline{\beta} = \sum_{m=1}^M \underline{x}_m^T \underline{\beta} \underline{\pi}_m.$$

Moreover, on noting that  $\text{diag}(\underline{I}) \mathbf{e} \mathbf{X} \underline{\beta} = \text{diag}(\mathbf{e} \mathbf{X} \underline{\beta}) \underline{I}$ , we obtain the expression for the covariance matrix of the vector  $\underline{Z}$  of observations,  $\mathbf{K} = \text{Cov}(\underline{Z})$  as

$$\mathbf{K} = \sigma^2 \mathbf{\Pi} + \text{diag}(\mathbf{e} \mathbf{X} \underline{\beta}) (\mathbf{P} - \underline{\pi} \underline{\pi}^T) \text{diag}(\mathbf{e} \mathbf{X} \underline{\beta}) + v^2 \tilde{\mathbf{P}}.$$

The aim of the talk is to discuss linear estimation of  $\underline{x}_m^T \underline{\beta}$  and prediction of  $\theta_m = \underline{x}_m^T \underline{\beta} + u_m$  as well as of  $\bar{Y}_m = \theta_m + \bar{\varepsilon}_m$  under the strict model-design approach described above. In particular it will be shown that no best linear estimators and predictors exist in the general scheme. Rather unexpectedly this is due to the dimensionality of the auxiliary non-random small area effects ( $\underline{x}_m$ ). The optimal strategy, that is the sampling plan and estimator (or predictor) will be proposed. It will be shown that the crucial restriction which has to be imposed on sampling designs lies in fixed sample sizes. Such a restriction allows to construct BLUE and BLUPs. Also in a special case of univariate small area effects BLUE and BLUPs under a general sampling scheme can be derived. In the simplest situation of no small area effects the formulas agree with classical synthetic and composite estimators. Also empirical versions of the estimators and predictors will be discussed.

This is a report on a research which is led jointly with Wojciech Niemirow (Główny Urząd Statystyczny, Warszawa and Uniwersytet Mikołaja Kopernika, Toruń)

---

<sup>1</sup> Główny Urząd Statystyczny and Politechnika Warszawska, Warszawa, Poland

# NONLINEAR ESTIMATORS OF A FINITE POPULATION TOTAL – DO THEY EXIST?

Jan Wretman<sup>1</sup>

Estimators of a population total, suggested in sampling textbooks, are mostly *linear* estimators, as this term is usually defined. Also, most theoretical results concerning the foundations of finite population inference deal with *linear* estimators.

The question asked in the present paper is: Considering that *nonlinear* unbiased estimators of a population total are so seldom mentioned, do such estimators exist at all?

The answer to be given is that *nonlinear* unbiased estimators of a population total exist if and only if the sampling design is a nonunicluster design with all inclusion probabilities strictly positive.

---

<sup>1</sup> Department of Statistics, Stockholm University, Stockholm, Sweden

# **SAMPLE SIZES FOR TWO-GROUP SECOND ORDER LATENT CURVE MODELS**

Linda Wänström<sup>1</sup>

Sample sizes needed to detect group slope differences in different second order latent curve models (Duncan & Duncan, 1996; McArdle, 1988) are derived using Satorra and Saris' (1985) power approximation techniques as well as linear model techniques. A sample size formula that can be used by researchers prior to data collection is also presented. Sample size is found to decrease with increases in effect size, indicator reliabilities, number of indicators, and number of measurement occasions. The relative importance of these factors is also derived. An empirical example illustrates power computations using real data. Strengths and limitations of this study are presented, along with needed future research.

---

<sup>1</sup> University of Stockholm

# **SAMPLE SURVEYS IN UKRAINE: EDUCATION AND IMPLEMENTATION**

Tetyana Yakovenko <sup>1</sup>, Olga Vasylyk <sup>1</sup>, Oksana Honchar <sup>2</sup>

Survey sampling is an important branch of the statistical science. During the Soviet time very few surveys were made in Ukraine, and survey sampling was not taught in Ukrainian universities. Unfortunately, until 1996 the speciality “Statistics” in a modern sense had been absent in Ukrainian universities. At the same time this speciality played an increasingly important role in Western universities. Our situation had well-known historical reasons. Neither the non-market approach to economic and social life in the former Soviet Union, nor the command-administrative system of the government needed to train and use specialists in sample surveys. There was no interest in investigating the people’s opinions and market processes.

The current transformation of the economical system and the building of a strong independent country in Ukraine have made it necessary to train qualified specialists in statistics, in particular in survey sampling.

Two Tempus Tacis projects 1996-2000 played a very important role in the development and updating of curricula for new statistical specializations at Kyiv National University and helped to implement a new course on Survey Sampling Theory and Methodology. This course is intended for last year students specializing in Statistics in the Faculty of Mechanics and Mathematics. It consists of 36 double lectures and the same amount of practical lessons.

Certified specialists in Statistics now work as statisticians, specialists on IT-technologies, scientific workers and teachers in the fields of statistics, in different government and commercial institutions, such as the Scientific and Technical Complex of Statistical Research, the Institute for Demography and Social Research and so on.

We shall describe the current system of sample surveys in Ukraine and mention some problems that we often face. We shall also discuss the teaching programs of survey sampling in Ukrainian universities and some plans for future expansion.

---

<sup>1</sup> Department of Probability Theory and Mathematical Statistics, Kyiv National Taras Shevchenko University, Ukraine

<sup>2</sup> Scientific and Technical Complex of Statistical Research, Ukraine

# **PROGRAMME**



BaNoCoSS-2007

## Second Baltic-Nordic Conference on Survey Sampling

2-7 June 2007, Kuusamo

### Programme



#### Saturday 2 June 2007

12:00–15:00	<i>Registration</i> - Registration desk open at Hotel Holiday Club Kuusamo
16:00–19:30	<i>Registration</i>
19:30–21:00	Welcome Reception, hotel restaurant Mango

#### Sunday 3 June 2007

8:00–9:00	<i>Registration desk open</i>
<b>9:00–10:30</b>	<b>Session 1 Auditorium</b> Opening and Keynote lecture <i>Chair</i> Gunnar Kulldorff <i>Opening: Risto Lehtonen</i> (University of Helsinki) <b>Carl-Erik Särndal</b> (University of Montreal) <i>Topics in uses of auxiliary information in surveys: The role of models, Nonresponse adjustment, Estimation for (small) domains</i>
10:30–11:00	Coffee
<b>11:00–12:30</b>	<b>Session 2 Auditorium</b> Calibration and Model-assisted Techniques <i>Chair</i> Jacek Wesolowski
11:00–11:30	<b>Aleksandras Plikusas</b> (Institute of Mathematics and Informatics, Vilnius): <i>Some examples of the nonlinear calibration</i>
11:30–11:50	<b>Hans Kiesel</b> (Institute for Employment Research, Nuremberg): <i>Calibrated imputation to correct for measurement error in the German Labour Force Survey</i>
11:50–12:10	<b>Dalius Pumputis</b> (Vilnius Pedagogical University, Institute of Mathematics and Informatics): <i>On the estimation of the variance of calibrated estimators of the population covariance</i>
12:10–12:30	<b>Milda Slickute-Sestokiene</b> (Statistics Lithuania): <i>Selection of vector of auxiliary information for Generalized Regression Estimator (GREG)</i>
12:30–13:30	Lunch <i>Registration desk open</i>
13:30–	Excursion to Oulanka National Park

<b>Monday 4 June 2007</b>	
9:00–10:30	<b>Session 3 Auditorium</b> Keynote lecture <i>Chair</i> Lauri Tarkkonen <b>Harvey Goldstein</b> (University of Bristol): <i>Modelling mixed response multivariate multilevel data with applications to prediction and multiple imputation</i>
10:30–11:00	Coffee
11:00–12:30	<b>Session 4 Auditorium</b> Survey Sampling 1 <i>Chair</i> Danutė Krapavickaitė
11:00–11:30	<b>Lennart Bondesson</b> (Umeå University): <i>On a sampling method suitable for real time sampling</i>
11:30–11:50	<b>Viktoras Chadyšas</b> (Vilnius Gediminas Technical University): <i>Confidence intervals estimation for Quantiles in finite population</i>
11:50–12:10	<b>Anton Grafström</b> (Umeå University): <i>On a generalization of Poisson sampling</i>
12:10–12:30	<b>Martine Quaglia</b> and <b>Géraldine Vivier</b> (INED, National Institute for Demographic Studies, France): <i>From theory to practice: how to conduct surveys among difficult to reach populations?</i>
12:30–13:30	Lunch
13:30–14:50	<b>Session 5 Auditorium</b> Business Surveys 1 <i>Chair</i> Thomas Laitila
13:30–13:50	<b>Outi Ahti-Miettinen</b> (Statistics Finland): <i>Sampling design of the Finnish Labor Cost Index</i>
13:50–14:10	<b>Olga Grakoviča</b> (University of Latvia): <i>Usage of census data as auxiliary information in survey sampling for agriculture statistics</i>
14:10–14:30	<b>Markus Gintas Šova</b> , <b>John Wood</b> and <b>Ian Richardson</b> (Office for National Statistics, UK): <i>Difficulties in the estimation and quality assessment of service producer price indices</i>
14:30–14:50	<b>Pieter Vlag</b> and <b>Koert van Bommel</b> (Statistics Netherlands): <i>Annual growth rates derived from short term statistics and annual structural business statistics</i>
13:30–14:30	<b>Session 6 Cabinet room</b> General Methodology <i>Chair</i> Maria Valaste
13:30–13:50	<b>Helene Feveile</b> , <b>Hermann Burr</b> , <b>Ole Olsen</b> and <b>Elsa Bach</b> (National Research Centre for the Working Environment, Denmark): <i>Danish Work Environment Cohort Study 2005: Design and weighting</i>
13:50–14:10	<b>Ole Olsen</b> , <b>Helene Feveile</b> and <b>Elsa Bach</b> (National Research Centre for the Working Environment, Denmark): <i>Changes in the psychosocial work environment in Denmark between 2000 and 2005</i>
14:10–14:30	<b>Nataliya Skachek</b> (Scientific and Technical Complex of Statistical Research, Kyiv): <i>Sample survey of farms in Ukraine: Current state and prospects</i>
15:00–15:30	Coffee
15:30–17:40	<b>Session 7 Auditorium</b> Survey Sampling 2 <i>Chair</i> Johan Heldal
15:30–16:00	<b>Imbi Traat</b> (University Tartu): <i>Estimation under restrictions</i>
16:00–16:20	<b>Lorenzo Fattorini</b> and <b>Caterina Pisani</b> (Università di Siena): <i>Variance estimation for measure of changes with coordinated samples</i>
16:20–16:40	<b>Natalja Lepik</b> (University of Tartu): <i>Mean square error of the general restriction estimator</i>
16:40–17:00	<b>Jānis Lapiņš</b> (Bank of Latvia) and <b>Martins Liberts</b> (Central Statistical Bureau of Latvia): <i>Estimation of monthly figures from Labour Force Survey</i>
17:00–17:20	<b>Riku Salonen</b> (Statistics Finland): <i>Regression composite estimation with application to the Finnish Labour Force Survey</i>
17:20–17:40	<b>Tetyana Yakovenko</b> (Kyiv National Taras Shevchenko University), <b>Olga Vasylyk</b> (Kyiv National Taras Shevchenko University) and <b>Oksana Honchar</b> (Scientific and Technical Complex of Statistical Research, Ukraine): <i>Sample Surveys in Ukraine: Education and Implementation</i>
18:30–20:00	Administrative meeting of Baltic-Nordic Network in Survey Sampling

<b>Tuesday 5 June 2007</b>	
9:00–10:30	<b>Session 8 Auditorium</b> Keynote lecture <i>Chair</i> Imbi Traat <b>Carl-Erik Särndal</b> (University of Montreal) <i>Topics in uses of auxiliary information in surveys: The role of models, Nonresponse adjustment, Estimation for (small) domains (Continued)</i>
10:30–11:00	Coffee
11:00–12:30	<b>Session 9 Auditorium</b> Nonresponse <i>Chair</i> Risto Lehtonen
11:00–11:30	<b>Giorgio Montanari</b> and <b>Giovanna Ranalli</b> (Università degli Studi di Perugia): <i>Calibration inspired by semiparametric regression as a treatment for nonresponse</i>
11:30–11:50	<b>Wojciech Gamrot</b> (University of Economics, Katowice): <i>Estimation of finite population kurtosis under double sampling for nonresponse</i>
11:50–12:10	<b>Karolin Toompere</b> (University of Tartu): <i>Indicator of strength of auxiliary information: a simulation study</i>
12:10–12:30	<b>Paavo Väisänen</b> (Statistics Finland): <i>Measuring item non-response of diary data in time use surveys</i>
12:30–13:30	Lunch
13:30–15:00	<b>Session 10 Auditorium</b> Survey Sampling 3 <i>Chair</i> Jan Wretman
13:30–14:00	<b>Johan Heldal</b> (Statistics Norway): <i>Ratio Estimation: When the ratio is a proportion</i>
14:00–14:20	<b>Øyvind Hoveid</b> (Norwegian Agricultural Economics Research Institute): <i>Estimation of survey weights when the frame contains information on size: Fuzzy neighbor post-stratification</i>
14:20–14:40	<b>Janika Konnu</b> (Statistics Finland): <i>The effect statistical disclosure control methods have on data: A study on micro data</i>
14:40–15:00	<b>Jan Kowalski</b> (Warsaw University of Technology): <i>Optimal linear recurrence estimators in stationary cascade rotation patterns</i>
13:30–14:10	<b>Session 11 Cabinet room</b> Business Surveys 2 <i>Chair</i> Paavo Väisänen
13:30–13:50	<b>Olga A. Vasechko</b> (Scientific and Technical Complex of Statistical Research, Kyiv) and <b>Michel Grun-Réhomme</b> (Université Paris 2): <i>Administrative and statistical registers in business statistics of Ukraine</i>
13:50–14:10	<b>Martine Quaglia, Cécile Lefevre, Ariane Pailhe, Anne Solaz, Ana Maria Noel, Tatiana Vichneskaïa</b> and <b>Bernard De Cledat</b> (INED, National Institute for Demographic Studies, France): <i>Interviewing both employees and employers, a mixed mode data collection for a matched survey</i>
15:00–15:30	Coffee
15.30–17:40	<b>Session 12 Auditorium</b> Survey Sampling 4 <i>Chair</i> Kari Nissinen
15.30–16:00	<b>Thomas Laitila</b> (Örebro University): <i>On-site sampling</i>
16:00–16:20	<b>Lorenzo Fattorini</b> (Università di Siena): <i>Performing Horvitz-Thompson estimation in complex sampling: a computer-intensive perspective</i>
16:20–16:40	<b>Oksana Honchar</b> (Scientific and Technical Complex of Statistical Research, Ukraine): <i>Sample in service surveys in Ukraine: Design and analysis</i>
16:40–17:00	<b>Soile Kotala</b> (SC-Research, Finland): <i>Impact analysis: Grouping of Tekes-funded projects</i>
17:00–17:20	<b>Danutė Krapavickaitė</b> (Institute of Mathematics and Informatics and Statistics Lithuania): <i>Model based estimator for a finite population total</i>
17:20–17:40	<b>Vilma Nekrašaitė</b> (Vilnius Gediminas Technical University and Statistics Lithuania): <i>Practical application of the model-based estimator for a finite population total</i>
18:00–	Excursion to Ruka

<b>Wednesday 6 June 2007</b>	
9:00–10:30	<b>Session 13 Auditorium</b> Keynote lecture <i>Chair Aleksandras Plikusas</i> <b>Harvey Goldstein</b> (University of Bristol): <i>Modelling mixed response multivariate multilevel data with applications to prediction and multiple imputation</i> (Continued)
10:30–11:00	Coffee
<b>11:00–12:20</b> 11:00–11:20 11:20–11:40 11:40–12:00 12:00–12:20	<b>Session 14 Auditorium</b> Survey Sampling 5 <i>Chair Martins Liberts</i> <b>Marco Ballin</b> (ISTAT), <b>Mauro Scanu</b> (ISTAT) and <b>Paola Vicard</b> (University Roma 3): <i>Efficiency of model based and model assisted estimators using probabilistic expert systems</i> <b>Inga Masiulaitytė</b> (Faculty of Mathematics and Informatics, Vilnius University, Statistics Lithuania): <i>Estimation of some inequality indexes</i> <b>Wojciech Niemiro</b> (Central Statistical Office of Poland and Faculty of Mathematics and Computer Science, Toruń) and <b>Robert Wieczorkowski</b> (Central Statistical Office of Poland): <i>M-estimators and U-statistics in approximating variance of income inequality indices</i> <b>Marjo Pyy-Martikainen</b> (Åbo Akademi University and Statistics Finland) and <b>Leif Nordberg</b> (Åbo Akademi University): <i>Inverse probability of censoring weighting method in survival analysis based on survey data</i>
12:30–13:30	Lunch
<b>13:30–15:00</b> 13:30–14:00 14:00–14:20 14:20–14:40 14:40–15:00	<b>Session 15 Auditorium</b> Small Area Estimation <i>Chair Giovanna Ranalli</i> <b>Jacek Wesolowski</b> (Główny Urząd Statystyczny and Politechnika Warszawska) <i>Linear estimation under model-design approach with small area effects</i> <b>Enrico Fabrizi</b> (University of Bergamo), <b>Maria Rosaria Ferrante</b> (University of Bologna) and <b>Silvia Pacei</b> (University of Bologna) <i>Comparing alternative distributional assumptions in mixed models used for the small area estimation of income parameters</i> <b>Kari Nissinen</b> (University of Jyväskylä) <i>EBLUP estimation of small area totals under linear mixed model for rotated panel data</i> <b>Kaja Sõstra</b> (Statistics Estonia) <i>Restriction Estimator for Domains</i>
15:00–15:30	Coffee
<b>15:30–17:00</b> 15:30–16:00 16:00–16:30 16:30–17:00	<b>Session 16 Auditorium</b> Special Session on Correcting skewed samples and longitudinal methods <i>Organizer and Chair Daniel Thorburn</i> <b>Silke Burestam</b> and <b>Daniel Thorburn</b> (Stockholm University): <i>Correcting the regression estimator for an abundance of auxiliary variables</i> <b>Daniel Thorburn</b> (Stockholm University) and <b>Boris Lorenc</b> (Statistics Sweden): <i>Nonparametric estimation with double samples</i> <b>Linda Wänström</b> (Stockholm University): <i>Sample sizes for two-group second order latent curve models</i>
18:30–	Conference Dinner, hotel restaurant Mango

## Thursday 7 June 2007

<b>8:00–9:30</b>	<b>Session 17 Auditorium</b> Survey Sampling 6 <i>Chair</i> Pauli Ollila
8:00–8:30	<b>Seppo Laaksonen</b> (Helsinki University): Retrospective two-stage cluster sampling for mortality
8:30–8:50	<b>Federica Baffetta, Lorenzo Fattorini</b> and <b>Sara Franceschi</b> (Università degli Studi di Siena): <i>A design-based approach to k-NN technique in forest inventories</i>
8:50–9:10	<b>V. C. Jaunky</b> and <b>A. J. Khadaroo</b> (University of Mauritius): <i>The school-to-work transition for University graduates in Mauritius: A duration model approach</i>
9:10–9:30	<b>Jan Wretman</b> (Stockholm University): <i>Nonlinear Estimators of a Finite Population Total – Do They Exist?</i>
9:30–10:00	Coffee
<b>10:00–11:30</b>	<b>Session 18 Auditorium</b> Special Session on Future of Baltic-Nordic Cooperation in Survey Sampling <i>Chair</i> Seppo Laaksonen
10:00–10:30	<b>Gunnar Kulldorff</b> (Umeå University) <i>Ten years of Baltic-Nordic co-operation</i>
10:30–11:30	<i>Discussants</i> <b>Johan Heldal</b> (Statistics Norway), <b>Danute Krapavickaite</b> (Institute of Mathematics and Informatics and Statistics Lithuania), <b>Risto Lehtonen</b> (University of Helsinki), <b>Martins Liberts</b> (Statistics Latvia), <b>Daniel Thorburn</b> (Stockholm University), <b>Imbi Traat</b> (University of Tartu)
	<i>Rejoinder</i> <b>Gunnar Kulldorff</b>
	<i>Closing</i> <b>Risto Lehtonen</b>
	<i>Departure</i>
	Bus connection from hotel to airport at 12:25 (price 5 EUR)



## **PARTICIPANTS**



<b><u>Last name</u></b>	<b><u>First name</u></b>	<b><u>Organization</u></b>	<b><u>Email</u></b>
Ahti-Miettinen	Outi	Statistics Finland	outi.ahti-miettinen@tilastokeskus.fi
Aru	Julia	Statistics Estonia	julia.aru@stat.ee
Bondesson	Lennart	Umeå University, Sweden	Lennart.Bondesson@math.umu.se
Chadyšas	Viktoras	Vilnius Gediminas Technical University	viktorasch@gmail.com
Conti	Pier Luigi	University of Rome 'La Sapienza'	pierluigi.conti@uniroma1.it
Davidson	Michael	National Institute of Public Health	md@niph.dk
Fattorini	Lorenzo	Università di Siena	fattorini@unisi.it
Feveile	Helene	National Research Centre for the Working Environment	hfe@nrcwe.dk
Franceschi	Sara	University of Florence	franceschi2@unisi.it
Gamrot	Wojciech Katowice	University of Economics	gamrot@ae.katowice.pl
Garvas	Tanja	Statistical Office of the Republic of Slovenia	tanja.garvas@gov.si
Goldstein	Harvey	University of Bristol	h.goldstein@bristol.ac.uk
Grafström	Anton	Umeå University	anton.grafstrom@math.umu.se
Grakoviča	Olga	University of Latvia	olga.grakoivca@csb.gov.lv
Grun-Réhomme	Michel	University Paris 2	grun@ensae.fr
Heldal	Johan	Statistics Norway	johan.heldal@ssb.no
Honchar	Oksana	Scientific and Technical Complex of Statistical Research	ohonchar@list.ru
Hoveid	Øyvind	Norwegian Agricultural Economics Research Institute	oyvind.hoveid@nilf.no
Ilves	Maiki	Örebro University	maiki.ilves@esi.oru.se
Jaunky	Vishal	University of Mauritius	vishaljaunky@intnet.mu
Kaarna	Kai	Statistics Estonia	kai.kaarna@stat.ee
Keto	Mauno	Mikkeli University of Applied Sciences	mauno.keto@mikkeliyamk.fi
Ketoja	Elise	MTT Agrifood Research Finland	elise.ketoja@mtt.fi
Kiesl	Hans	Institute for Employment Research	hans.kiesl@iab.de
Konnu	Janika	Statistics Finland	janika.konnu@stat.fi
Kotala	Soile	SC-Research	soile.kotala@seamk.fi
Kowalski	Jan	Warsaw University of Technology	j.kowalski@mini.pw.edu.pl
Krapavickaitė	Danutė	Statistics Lithuania & Institute of Mathematics and Informatics, Lithuania	krapav@ktl.mii.lt
Kulldorff	Gunnar	University of Umeå	gunnar@matstat.umu.se
Laaksonen	Seppo	Statistics Finland and University of Helsinki	Seppo.Laaksonen@Helsinki.Fi
Laitila	Thomas	Örebro university and Statistics Sweden	thomas.laitila@esi.oru.se
Lehikoinen	Tuula	Kuntoutussäätiö	tuula.lehikoinen@kuntoutussaatio.fi
Lehtonen	Risto	University of Helsinki	risto.lehtonen@helsinki.fi
Lepik	Natalja	University of Tartu	natalja.lepik@ut.ee
Liberts	Martins	Central Statistical Bureau of Latvia	Martins.Liberts@csb.gov.lv
Lorenc	Boris	Statistics Sweden	boris.lorenc@scb.se
Masiulaityte	Inga	Vilnius University, Statistics Lithuania	inga.masiulaityte@stat.gov.lt
Masri	Azilawati	Malaysian communications & Multimedia commission	azilawati@cmc.gov.my
Nekrašaitė	Vilma	Vilnius Gediminas Technical University & Statistics Lithuania	nekrasaitė.vilma@gmail.com
Niemiro	Wojciech	Central Statistical Office of Poland and Faculty of Mathematics and Computer Science	wniemiro@gmail.com
Nikic	Boro	Statistical Office of the Republic of Slovenia	boro.nikic@gov.si
Nissinen	Kari	University of Jyväskylä	knissine@maths.jyu.fi
Olsen	Ole	National Research Centre for the Working Environment	ool@nrcwe.dk
Ollila	Pauli	Statistics Finland	Pauli.Ollila@stat.fi

Omelka	Marek	Charles University in Prague	omelka@karlin.mff.cuni.cz
Pacei	Silvia	University of Bologna	silvia.pacei@unibo.it
Pisani	Caterina	Università di Siena	pisani4@unisi.it
Pitkänen	Timo	MTT Agrifood Research Finland	timo.pitkanen@mtt.fi
Plikusas	Aleksandras	Statistics Lithuania & Institute of Mathematics and informatics	plikusas@ktl.mii.lt
Pumputis	Dalius	Vilnius Pedagogical University	dalpas@delfi.lt
Pyy-Martikainen	Marjo	Åbo Akademi / Statistics Finland	marjo.pyy-martikainen@stat.fi
Quaglia	Martine	INED	quaglia@ined.fr
Ranalli	M. Giovanna	University of Perugia	giovanna@stat.unipg.it
Randoja	Marin	University of Tartu	marin.randoja@stat.ee
Salonen	Riku	Statistics Finland	riku.salonen@stat.fi
Schjalm	Arnfinn	Statistics Norway	sch@ssb.no
Skachek	Nataliya	Scientific and Technical Complex of Statistical Research	nskachek@mail.ru
Slickute-Sestokiene	Milda	Statistics Lithuania	milda.slickute@stat.gov.lt
Šova	Markus Gintas	ONS	markus.sova@ons.gov.uk
Särndal	Carl-Erik	University of Montreal	carl.sarndal@rogers.com
Sõstra	Kaja	Statistics Estonia	kaja.sostra@stat.ee
Tamsfoss	Steinar	Synovate Norway	steinar.tamsfoss@synovate.com
Tarkkonen	Lauri	University of Helsinki	lauri.tarkkonen@gmail.com
Thorburn	Daniel	University of Stockholm	Daniel.thorburn@stat.su.se
Toompere	Karolin	University of Tartu	karolin.toompere@ut.ee
Traat	Imbi	University of Tartu	imbi.traat@ut.ee
Valaste	Maria	University of Helsinki	maria.valaste@helsinki.fi
Wesołowski	Jacek	Central Statistical Office	wesolo@mini.pw.edu.pl
Vicard	Paola	Università Roma Tre	vicard@uniroma3.it
Vivier	Géraldine	INED	vivier@ined.fr
Vlag	Pieter	Statistics Netherlands	pvag@cbs.nl
Wretman	Jan	Stockholm University	jan.wretman@stat.su.se
Väisänen	Paavo	Statistics Finland	paavo.vaisanen@stat.fi
Wänström	Linda	University of Stockholm	linda.wanstrom@stat.su.se
Yakovenko	Tetyana	Kyiv National Taras Shevchenko University	yata452@univ.kiev.ua