Modelling of Survey Data

Olga Vasylyk

Taras Shevchenko National University of Kyiv, Ukraine

Outline

Introduction

1. Design-based approach in survey sampling: main features
2. Model-based approach in survey sampling: main features
3. Some examples of population models
4. Design-based model assisted approach in survey sampling
5. Model-based and model-assisted estimation for domains and small areas
6. Model-based and model-assisted methods in dealing with nonresponses
7. Weighting/calibration
8. Modeling of complex survey data
9. Models for survey sampling with sensitive characteristics

Conclusions

Main references

Acknowlegements

Introduction

The main objective is to learn how models are used in sample surveys and to prepare lecture notes/ textbook for Master students specialized in Statistics at Taras Shevchenko National University of Kyiv (KNU).

This work is not finished yet. There was given one “pilot” course on “Models in Sample Surveys” for Master students specialized in Statistics, which included the following topics:

1. Design-based approach in survey sampling: main features
2. Model-based approach in survey sampling: main features
3. Some examples of population models
4. Design-based model assisted approach in survey sampling
5. Model-based and model-assisted estimation for domains and small areas

The presentation will show how the lecture notes are supposed to be organized.

Survey data may be viewed as the outcome of two random processes: The process generating the values in the finite population, often referred to as the ‘superpopulation model’, and the process selecting the sample data from the finite population values, known as the ‘sample selection mechanism’.

Brewer (1963) proposed the model-based approach in the context of the ratio model $y_i = \beta x_i + \varepsilon_i$, $i = 1,\ldots,N$.

Royall (1970) and his collaborators made a systematic study of this approach.

Valliant, Dorfmann and Royall (2000) give a comprehensive account of this theory.

1. Design-based approach in survey sampling: main features (*Little* 2004)

For a population U with N units, let $\mathbf{Y} = (y_1,\ldots,y_N)$,

where y_i is the set of survey variables for unit i,

and let $\mathbf{I} = (I_1,\ldots,I_N)$ denote a set of inclusion indicator variables,

where $I_i = 1$ if unit i is included in the sample and

$I_i = 0$ if it is not included.

Design-based inference is based on the distribution of **I**, with the survey variables **Y** treated as **fixed** quantities. For inference about a finite population quantity $Q = Q(\mathbf{Y})$ the following steps are involved:

1. Choosing an estimator $q = q(\mathbf{Y}_{\text{inc}},\mathbf{I})$, a function of the observed part $\mathbf{Y}_{\text{inc}}$ of $\mathbf{Y}$, that is unbiased or approximately unbiased for Q with respect to the distribution of $\mathbf{I}$. Here $\mathbf{q}$ is a random variable as a function of $\mathbf{I}$, and $\mathbf{Y}_{\text{inc}}$ are fixed quantities.
2. Choosing a variance estimator $v = v(\mathbf{Y}_{\text{inc}},\mathbf{I})$, that is unbiased or approximately unbiased for the variance of q with respect to the distribution of $\mathbf{I}$.

Inferences are then generally based on normal large-sample approximations.

2. Model-based approach in survey sampling: main features (*Little* 2004)

Model-based approach to survey sampling inference requires a **model** for the survey variables **Y**, which are now treated as **random**. The model is then used to predict the nonsampled values of the population, and hence finite population quantities *Q*.

There are two major variants: **superpopulation** modeling and **Bayesian** modeling.

Superpopulation modeling

Analytic inference from survey data relates to the superpopulation model, but when the sample selection probabilities are correlated with the values of the model response variables even after conditioning on auxiliary variables, the sampling mechanism becomes informative and the selection effects need to be accounted for in the inference process.

In **superpopulation** modeling the N population values of $\mathbf{Y}$ are assumed to be a random sample from a "superpopulation" and are assigned a probability distribution $p(\mathbf{Y}|\boldsymbol{\theta})$ indexed by fixed parameters $\boldsymbol{\theta}$.

Inferences are based on the joint distribution of $\mathbf{Y}$ and $\mathbf{I}$.

Bayesian modeling

Bayesian modeling requires specification of a prior distribution $p(\mathbf{Y})$ for the population values. Inferences for finite population quantities $Q(\mathbf{Y})$ are based on the posterior predictive distribution $p(\mathbf{Y}_{\text{exc}}|\mathbf{Y}_{\text{inc}})$ of the nonsampled values $\mathbf{Y}_{\text{exc}}$, given the sampled values $\mathbf{Y}_{\text{inc}}$.

The specification of $p(\mathbf{Y}|\boldsymbol{\theta})$ in Bayesian formulation is the same as in parametric superpopulation modeling, and in large samples the likelihood based on this distribution dominates the contribution from the prior distribution for $\boldsymbol{\theta}$. As a result, **large-sample inference** from the superpopulation modeling and Bayesian approaches **are often similar** (*Little* 2004).

Non-informative Sampling

Bayesian model formulations **do not involve** the distribution for **I**, basing inferences only on the distribution of **Y**. This is justified when the sampling mechanism is **“non-informative”.**

Sampling mechanism is said to be **non-informative** for a variable Y if the distribution of the sampled values of Y and the distribution of the non-sampled values of this variable are the same (*Chambers* 2003). Or, in other words, the distribution of I given Y does not depend on the values of Y (*Little* 2004).

This is the case with **probability sampling** – a method that uses a randomization device to decide which units on the frame are in the sample.

3. Some examples of population models

Main ideas and definitions used in the examples:

- Parameter to be estimated is a population total

$$t_y = \sum_U y_i = \sum_s y_i + \sum_r y_i = t_{sy} + t_{ry}$$

- An estimator of the population total is denoted by $\widehat{t_y}$

- Model-based properties of the estimator are defined by the distribution of the sample error

$$\hat{t}_y - t_y$$

under the assumed population model.

- Prediction bias of the estimator is the mean of this distribution

$$E_\xi(\hat{t}_y - t_y)$$

- Prediction variance

$$\mathrm{Var}_{\xi}(\,\widehat{t_y} - t_y)$$

- Prediction mean square error

$$E_{\xi}(\,\widehat{t_y} - t_y)^2 = \mathrm{Var}_{\xi}\big(\,\widehat{t_y} - t_y\big) + (E_{\xi}\big(\,\widehat{t_y} - t_y\big))^2$$

Homogeneous population model

Homogeneous population model – the basic “building block” for more complex models that can be used to represent real world variability (*Chambers* 2003):

$$E_{\xi}(y_i) = \mu,$$

$$Var_{\xi}(y_i) = \sigma^2,$$

$$Cov_{\xi}(y_i, y_j) = \rho\sigma^2$$

$$\hat{t}_{Hy} = N\bar{y}_s.$$

$$\hat{V}_{\xi}(\hat{t}_{Hy}) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right)\frac{1}{n-1}\sum_s (y_i - \bar{y}_s)^2 .$$

Stratified homogeneous population model (S)

$$E_{\xi}(y_i) = \mu h, i = \overline{1, N}, h = \overline{1, H}$$

$$Var_{\xi}(y_i) = \sigma_h^2,$$

$$Cov_{\xi}(y_i, y_j) = 0 \quad \forall i \neq j.$$

$$\hat{t}_{Sy} = \sum_h N_h \bar{y}_{sh} = \sum_h \hat{t}_{Hhy} .$$

$$\hat{V}_{\xi}(\hat{t}_{Sy} - t_y) = \sum_h \hat{V}_{\xi}(\hat{t}_{Hhy} - t_{hy}) = \sum_h (N_h^2 / n_h)(1 - n_h / N_h) s_h^2 .$$

Simple ratio population model (R)

$$E_\xi(y_i|x_i) = \beta x_i$$

$$Var_\xi(y_i|x_i) = \sigma^2 x_i$$

$$Cov_\xi(y_i y_j|x_i x_j) = 0 \quad \forall i \neq j$$

$$\hat{t}_{Ry} = t_{sy} + b_R \sum\nolimits_r x_i = \frac{\sum_s y_i}{\sum_s x_i} \sum\nolimits_U x_i = \frac{\bar{y}_s}{\bar{x}_s} t_x.$$

$$b_R = \frac{\sum_s y_i}{\sum_s x_i}.$$

Simple linear population model (L)

$$E_\xi(y_i|x_i) = \alpha + \beta x_i$$

$$Var_\xi(y_i|x_i) = \sigma^2$$

$$Cov_\xi\left(y_i y_j \middle| x_i x_j\right) = 0 \quad \forall i \neq j$$

$$\hat{t}_{Ly} = \sum\nolimits_s y_i + \sum\nolimits_r (a_L + b_L x_i) = \sum\nolimits_U (a_L + b_L x_i) = N\left[\bar{y}_s + b_L(\bar{x} - \bar{x}_s)\right]$$

where

$$a_L = \bar{y}_s - b_L \bar{x}_s$$

$$b_L = \frac{\sum_s (y_i - \bar{y}_s)(x_i - \bar{x}_s)}{\sum_s (x_i - \bar{x}_s)^2}.$$

Clustered population model (C)

Clustered population model (C): $E_\xi(y_{ig}) = \mu$

$$\mathrm{Var}_\xi(y_{ig}) = \sigma^2$$

$$\mathrm{Cov}_\xi(y_{ig}, y_{if}) = \begin{cases} \rho\sigma^2, & \text{if} \quad g = f \\ 0, & \text{otherwise} \end{cases}$$

$$n = \sum_s m_g, \qquad N = \sum_U M_g$$

$$\hat{t}_{Cy} = \sum_s m_g \bar{y}_{sg} + \sum_s (M_g - m_g)\left[(1-\alpha_g)\hat{\mu} + \alpha_g \bar{y}_{sg}\right] + \hat{\mu}\left(N - \sum_s M_g\right).$$

Here α_g is a weight reflecting knowledge about $\bar{y}_{rg}$ given the average value $\bar{y}_{sg}$

$$E_\xi(\bar{y}_{rg} \mid \bar{y}_{sg}) = (1-\alpha_g)\mu + \alpha_g \bar{y}_{sg}$$

$$\hat{\mu} = \frac{\sum_s m_g (1-\rho+\rho m_g)^{-1} \bar{y}_{sg}}{\sum_s m_g (1-\rho+\rho m_g)^{-1}} = \sum_s \theta_g \bar{y}_{sg}\,.$$

Option 1: Assume $\rho = 0 \Rightarrow \theta_g = n^{-1} m_g \Rightarrow \hat{\mu} = \sum_s m_g \bar{y}_{sg} / \sum_s m_g = \bar{y}_s$.

Option 2: Assume $\rho = 1 \Rightarrow \theta_g = q^{-1} \Rightarrow \hat{\mu} = q^{-1} \sum_s \bar{y}_{sg} = \bar{\bar{y}}_s$.

Option 3: Estimate ρ directly by fitting a 2-level model to sample data.

General linear model

$$Y = X\beta + \varepsilon$$

$$E_\xi(\varepsilon) = 0 \qquad V = V(X) = \begin{bmatrix} V_{ss} & V_{sr} \\ V_{rs} & V_{rr} \end{bmatrix}$$

$$\mathrm{Var}_\xi(\varepsilon) = \sigma^2 V$$

$$\hat{t}_{opt,y} = \mathbf{1}'_n \mathbf{Y}_s + \mathbf{1}'_{N-n}\left[\mathbf{X}_r\hat{\beta}_{opt} + \mathbf{V}_{rs}\mathbf{V}_{ss}^{-1}(\mathbf{Y}_s - \mathbf{X}_s\hat{\beta}_{opt})\right]$$

$$\hat{\beta}_{opt} = \left(\mathbf{X}'_s\mathbf{V}_{ss}^{-1}\mathbf{X}_s\right)^{-1}\mathbf{X}'_s\mathbf{V}_{ss}^{-1}\mathbf{Y}_s$$

4.Design-based model assisted approach in survey sampling (Särndal et al., 1992)

Design-based model-assisted approach attempts to combine the desirable features of design-based and model-based methods.

The main source:

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

In particular, in this section generalized regression (GREG) estimator will be introduced.

Generalized regression estimator is a model assisted estimator designed to improve the accuracy of the estimates by means of auxiliary information. GREG estimator guarantees the coherence between sampling estimates and known totals of the auxiliary variables, as well.

We shall see GREG estimator in the Section 5 as an estimator used for small area estimation.

5. Model-based and model-assisted estimation for domains and small areas

Lehtonen, R. (2006) *The Role of Models in Model-Assisted and Model-Dependent Estimation for Domains and Small Areas.* In: Proceedings of the Workshop on Survey Sampling Theory and Methodology (Ventspils, Latvia):

Methods available for the estimation of totals for domains and small areas include model-assisted design-based estimators, referring to the family of generalized regression (GREG) estimators (Särndal, Swensson and Wretman 1992, Estevao and Särndal 1999, 2004), and model-dependent techniques, such as the EBLUP estimator (Empirical Best Linear Unbiased Predictor) and synthetic estimators (Ghosh 2001, Rao 2003). Properties of these estimator types are discussed for example in Lehtonen and Veijanen (1998, 1999) and Lehtonen, Veijanen and Särndal (2003, 2005).

D. Pfeffermann. *Small Area Estimation: Basic Concepts, Models and Ongoing Research.* The Survey Statistician, No.62, July 2010:

SAE methods can be divided broadly into design-based and model-based methods. The latter methods use either the frequentist approach or a fully Bayesian methodology, and in some cases combine them, which is known in the SAE literature as “empirical Bayes”. Design-based methods often use a model for the construction of the estimates (model-assisted approach), but the bias, variance and other properties of the estimators are evaluated under randomization distribution over all possible samples. Model-based methods generally condition on the selected sample, and the inference is with respect to the underlying model. Design-based and model-based SAE use auxiliary information collected in the survey and in other large surveys or administrative records. This is crucial because with small sample sizes even the most sophisticated model can be of little help if it does not involve a set of covariates that provide good predictions of the small area quantities of interest.

5.1.Model-assisted design-based estimators

Let Y define the characteristic of interest and denote by y_{ij} the outcome value for unit j belonging to area i , $i = 1,\ldots, M$; $j = 1\ldots N_i$, where N_i is the area size. Let $s = s_1 \cup \ldots \cup s_m$ denote the sample, where s_i of size n_i is the sample observed for area i. Suppose that it is required to estimate the true area mean $\bar{Y}_i = \sum_{j=1}^{N_i} y_{ij} / N_i$. If no auxiliary information is available, the *direct* design unbiased estimator and its variance over the randomization distribution for given sample size n_i are given by,

$$\hat{\bar{Y}}_i = \sum_{j=1}^{n_i} y_{ij} / n_i \;\; ; \;\; Var_D[\hat{\bar{Y}}_i \mid n_i] = (S_i^2 / n_i)[1-(n_i / N_i)] = S_i^{*2}, \qquad (1)$$

where $S_i^2 = \sum_{j=1}^{N_i} (y_{ij} - \bar{Y}_i)^2 / (N_i - 1)$. Clearly, for small n_i the variance will be large, unless the variability of the *y-values* is sufficiently small. Suppose, however, that values x_{ij} of *p* concomitant variables $\mathrm{x}_1,\ldots,\mathrm{x}_\mathrm{p}$ are measured for each of the sampled units and that the area means $\bar{X}_i = \sum_{j=1}^{N_i} \mathrm{x}_{ij} / N_i$ are likewise known.

5.1.1. Regression estimator

Assuming $\mathrm{x}_{1ij} = 1$ for all (i, j), a more efficient design-based estimator in this case is the regression estimator,

$$\hat{\bar{Y}}_i^{\mathrm{Re}\,g} = \bar{X}_i'\hat{\beta}_i \quad ; \quad Var(\hat{\bar{Y}}_i^{REG} | n_i) \cong S_i^{*2}(1-\rho_i^2)\,, \qquad (2)$$

where $\hat{\beta}_i = [\sum_{j=1}^{n_i} \mathrm{x}_{ij}\mathrm{x}_{ij}']^{-1}\sum_{j=1}^{n_i} \mathrm{x}_{ij}\mathrm{y}_{ij}$ is the ordinary least square estimator and ρ_i is the multiple correlation coefficient between Y and $\mathrm{x}_1, \ldots, \mathrm{x}_\mathrm{p}$ in area i. The variance approximation in (2) assumes large n_i. Thus, by use of the concomitant variables, the variance is reduced by the factor $(1-\rho_i^2)$, illustrating the importance of using auxiliary information with good prediction power (large R^2) for SAE.

5.1.2. Synthetic regression estimator

Although the regression estimator (2) usually has smaller variance than the simple sample mean, its variance may still be large for small sample size, unless the multiple correlation is very large. However, if the regression relationships between y and x are 'similar' across the areas, a more stable estimator is the *synthetic regression* estimator,

$$\hat{\bar{Y}}_i^{Syn} = \sum_{j=1}^{N_i} \hat{y}_{ij} / N_i = \bar{X}_i'\hat{B}, \qquad (3)$$

where $\hat{y}_{ij} = \mathrm{x}_{ij}'\hat{B}$ and $\hat{B}$ may be computed as $\hat{B} = [\sum_{i,j\in s} \mathrm{x}_{ij}\mathrm{x}_{ij}']^{-1} \sum_{i,j\in s} \mathrm{x}_{ij} y_{ij}$. The prominent advantage of synthetic estimation is the substantial reduction in variance, because the estimator $\hat{B}$ uses all the sample data, but it can lead to severe biases if the regression relationships actually differ between the areas.

5.1.3. GREG estimator. Composite estimator.

In order to correct for the possible large bias of the synthetic estimator, an approximately design-unbiased estimator in common use is the GREG estimator,

$$\hat{\bar{Y}}_i^{Greg} = \sum_{j=1}^{N_i} \hat{y}_{ij} / N_i + \sum_{k \in s_i} (y_{ik} - \hat{y}_{ik}) / n_i \,. \tag{4}$$

However, this estimator may again be unstable in small samples since the second expression in the right hand side of (4) is an area sample mean. Thus, the choice between the synthetic and GREG estimators is a trade off between bias and variance. A compromise is achieved by using a composite estimator of the form,

$$\hat{\bar{Y}}_i^{Com} = \alpha_i \hat{\bar{Y}}_i^{Greg} + (1-\alpha_i)\hat{\bar{Y}}_i^{Syn} ; \quad 0 \le \alpha_i \le 1 \,. \tag{5}$$

It is common to choose the coefficient α_i (different coefficients in different areas) as some function of the achieved sample size n_i, but such choices account only partly for the relative MSE of the two estimators, as reflected also by the fact that the coefficients would be the same irrespective of the target variable of interest.

5.2. Model-based estimators

Let θ_k define the parameter of interest in area k, k=1,…,M, and let y_i, $\mathbf{x}_i$ denote the data observed for sampled area i, i = 1, ..., m, where m denotes the number of areas with data on the outcome variable. When the only available information is at the area level, y_i is commonly the direct estimator of θ_i and $\mathbf{x}_i$ is a vector of area-level covariates. When unit level information is available, y_i is a vector of individual outcomes and $\mathbf{x}_i$ is the corresponding matrix of individual covariate information.

A typical small area model consists of two parts: the first part models the distribution of $(y_i| \theta_i;\psi_{(1)})$. The second part models the distribution of $(\theta_i | \mathbf{x}_i;\psi_{(2)})$, linking θ_i to the parameters in other areas (or at different times) and to the covariates. The vector parameters $\psi_{(1)}$ and $\psi_{(2)}$ are usually unknown and are estimated from all the available data $D(s) = \{y_i, \mathbf{x}_i; i=1,\ldots,m\}$.

5.2.1. Unit level random effects model

The model, employed originally by Battese *et al.* (1988) assumes,

$$y_{ij} = \mathrm{x}'_{ij}\beta + u_i + \varepsilon_{ij}\ , \qquad (6)$$

where u_i and ε_{ij} are mutually independent error terms with zero means and variances σ_u^2 and σ_ε^2 respectively. The 'random effect' u_i represents the joint effect of area characteristics not accounted for by the covariates. Under the model, the true small area means are $\bar{Y}_i = \bar{X}'_i\beta + u_i + \bar{\varepsilon}_i$, but since $\bar{\varepsilon}_i = \sum_{j=1}^{N_i}\varepsilon_{ij} / N_i \cong 0$ for large N_i, the target parameters are often defined as $\theta_i = \bar{X}'_i\beta + u_i = E(\bar{Y}_i \mid u_i)$. For known variances $(\sigma_u^2, \sigma_\varepsilon^2)$, the best linear unbiased predictor (BLUP) of θ_i is,

$$\hat{\theta}_i = \gamma_i[\bar{y}_i + (\bar{X}_i - \bar{\mathrm{x}}_i)'\hat{\beta}_{GLS}] + (1-\gamma_i)\bar{X}'_i\hat{\beta}_{GLS}\ , \qquad (7)$$

where $\hat{\beta}_{GLS}$ is the generalized least square estimator of β computed from all the observed data and $\gamma_i = \sigma_u^2 / (\sigma_u^2 + \sigma_\varepsilon^2 / n_i)$. For areas l with no sample, $\hat{\theta}_l = \bar{X}'_l\hat{\beta}_{GLS}$.

5.2.2. Area level random effects model

This model is in broad use when the concomitant covariate information is only at the area level. It was used originally by Fay and Herriot (1979) for predicting the mean per capita income in geographical areas of less than 500 inhabitants. Denote by $\tilde{\theta}_i$ the direct sample estimator of θ_i. The model assumes that,

$$\tilde{\theta}_i = \theta_i + e_i \;\; ; \;\; \theta_i = \mathrm{x}_i'\beta + u_i \,, \tag{8}$$

such that e_i represents the sampling error, assumed to have zero mean and known design variance $Var_D(e_i) = \sigma_{Di}^2$. The model integrates therefore a model dependent random effect u_i, and a sampling error e_i with the two errors being independent. The BLUP under this model is,

$$\hat{\theta}_i = \gamma_i \tilde{\theta}_i + (1-\gamma_i)\mathrm{x}_i'\hat{\beta}_{GLS} = \mathrm{x}_i'\hat{\beta}_{GLS} + \gamma_i(\tilde{\theta}_i - \mathrm{x}_i'\hat{\beta}_{GLS}) \,, \tag{9}$$

which again is a composite estimator with coefficient $\gamma_i = \sigma_u^2 / (\sigma_{Di}^2 + \sigma_u^2)$. As with the unit level model, the variance σ_u^2 is usually unknown and is either assigned a prior distribution under the Bayesian approach, or replaced by a sample estimate in (9), yielding the corresponding EBLUP (or the EB) predictor.

5.2.3. Unit level random effects model for binary data

The previous two models assume continuous outcomes. Suppose now that y_{ij} is a binary variable taking the values 0 or 1. For example, $y_{ij}=1$ if individual j in area i is unemployed (or suffers from a certain disease) and $y_{ij}=0$ otherwise, such that $P_i = N_i^{-1}\sum_{j=1}^{N_i} y_{ij}$ is the true area unemployment rate (true area disease prevalence). The following model is often used for predicting the proportions P_i:

$$\begin{aligned} &y_{ij} \mid p_{ij} \overset{indep.}{\sim} Bernoulli(p_{ij}) \\ &\operatorname{logit}(p_{ij}) = \log[p_{ij}/(1-p_{ij})] = \mathrm{x}_{ij}'\beta + u_i\,;\; u_i \overset{indep.}{\sim} N\left(0,\sigma_u^2\right) \end{aligned} \quad , \qquad (10)$$

where, as in (6), x_{ij} is a vector of covariates, β is a vector of fixed regression coefficients and u_i is a random effect representing the unexplained variability of the individual probabilities.

6. Model-based and model-assisted methods in dealing with nonresponses

- Little, R. J. A., and Rubin, D. B. (1987, 2002) Statistical Analysis with Missing Data. New York: Wiley.

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

- Särndal, C.-E., Swensson, B. and Wretman, J. (1992) *Model Assisted Survey Sampling*. Springer-Verlag, New York. (Chapter 15. Nonresponse)
- Särndal, C.-E., Lundstrom, S. (2005) *Estimation in Surveys with Nonresponse*. Wiley, 212 p.

7. Weighting/calibration

The calibration approach to estimation for finite populations consists of

(a) a computation of weights that incorporate specified auxiliary information

and are restrained by calibration equation(s); (b) the use of these weights to compute linearly weighted estimates of totals and other finite population parameters: weight times variable value summed over a set of observed units; (c) an objective to obtain nearly design unbiased estimates as long as nonresponse and other nonsampling errors are absent.

- C. Sarndal. The calibration approach in survey theory and practice. Survey

Methodology, 33(2): 99–119, 2007.

- G. Davies. *Examination of approaches to calibration in survey sampling.* A thesis submitted for the degree of Doctor of Philosophy, March 2018.
- Montanari G.E. and Ranalli M.G. *Multiple and ridge model calibration for sample surveys.* Proceedings of the Workshop in Calibration and estimation in surveys, Ottawa, October 2007, Statistics Canada.

8. Modeling of complex survey data

- D. Pfeffermann (2011) *Modelling of complex syrvey data: Why model? Why is it a problem? How can we approach it?* Survey Methodology, Vol. 37, No. 2, pp. 115-136.
- Lehtonen R. and Pahkinen E. (2004). Practical Methods for Design and Analysis of Complex Surveys. Second Edition. Chichester: John Wiley & Sons, Ltd.

Many approaches have been proposed in the literature for estimating population models from complex survey data.

The approaches differ in the conditions underlying their use, the data required for their application, goodness of fit testing, the inference objectives that they accommodate, statistical efficiency, computational demands, and the skills required from analysts fitting the model.

Does not exists any single approach that can be considered as best in all situations.

9. Models for survey sampling with sensitive characteristics

- Jun-Wu Yu, Guo-Liang Tian, Man-Lai Tang. *Two new models for survey sampling with sensitive characteristic: design and analysis*. Metrika (2008) 67:251–263

Sensitive topics or highly personal questions are often being asked in medical, psychological and sociological surveys. This paper proposes two new models (namely, the triangular and crosswise models) for survey sampling with the sensitive characteristics.

- Liu, Y; Tian, G. *A variant of the parallel model for sample surveys with sensitive characteristics*. Computational Statistics & Data Analysis (2013), v. 67, p. 115-135. DOI: 10.1016/j.csda.2013.05.0

A new non-randomized response (NRR) model (called a variant of the parallel model) is proposed. The survey design and corresponding statistical inferences including likelihood-based methods, Bayesian methods and bootstrap methods are provided.

Conclusions

There exists vast amount of literature on the subject, therefore I still have many questions/doubts about the topics and information to be included in the lectures.
For example, what topics are more important/more useful? What applications of the theory are more interesting? How to organize some practical training corresponding to the lectures?

And the title is not final too :)

Restrictions:
Unfortunately, the KNU does not have subscription to modern journals in Statistics, therefore we mainly use open access sources or the literature that was bought within the framework of different grants.

Main references

1. Chambers, R. (2003) *An introduction to model-based survey sampling.* Seminario international de estadistica en eusradi, No.42, 90 p.
2. Lehtonen R. and Pahkinen E. (2004). *Practical Methods for Design and Analysis of Complex Surveys.* Second Edition. Chichester: John Wiley & Sons, Ltd.
3. Lehtonen, R. (2006) *The Role of Models in Model-Assisted and Model-Dependent Estimation for Domains and Small Areas.* In: Proceedings of the Workshop on Survey Sampling Theory and Methodology (Ventspils, Latvia) pp. 35-44.
4. Lehtonen, R. (2009) *Estimation for domains and small areas with design-based and model-based methods.* Lectures at the BNU Summer School on Survey Statistics (Kyiv, Ukraine).
5. Little, R. J. A., and Rubin, D. B. (1987) *Statistical Analysis with Missing Data*. New York: Wiley.
6. Little, R. J. A. (2004) *To model or not to model? Competing Mdes of Inference for Finite Population Sampling.* The Journal of the Ameriacan Statistical Association, Vol.99, No. 499. pp.546-556.
7. Lohr, S. (1999) *Sampling: Design and Analysis.* Duxbury Press, Pacific Grove.
8. Montanari G.E. and Ranalli M.G. *Multiple and ridge model calibration for sample surveys.* Proceedings of the Workshop in Calibration and estimation in surveys, Ottawa, October 2007, Statistics Canada.

Main references (cont.)

9. D. Pfeffermann (2010) *Small Area Estimation: Basic Concepts, Models and Ongoing Research.* The Survey Statistician, No.62, pp. 26-32.
10. D. Pfeffermann (2011) *Modelling of complex syrvey data: Why model? Why is it a problem? How can we approach it?* Survey Methodology, Vol. 37, No. 2, pp. 115-136.
11. Rao, J.N.K. (2005) *Interplay between sample survey theory and practice:an appraisal*. Syrvey Methodology, Vol. 31, No. 2, pp. 117-138.
12. Särndal, C.-E., Swensson, B., Wretman, J. (1992) *Model Assisted Survey Sampling*. Springer-Verlag, New York.
13. Särndal, C.-E., Lundstrom, S. (2005) *Estimation in Surveys with Nonresponse*. Wiley, 212 p.
14. Särndal, C.-E. (2010) Models in Survey Sampling. In: *Official Statistics, Methodology and Applications in Honour of Daniel Thorburn*, pp.15–27
15. Valliant, R., Dorfman, A.H. and Royall R.M. (2000). *Finite Population Sampling and Inference.* John. Wiley & Sons, New York.
16. Jun-Wu Yu, Guo-Liang Tian, Man-Lai Tang. *Two new models for survey sampling with sensitive characteristic: design and analysis*. Metrika (2008) 67:251–263

Acknowledgements

I am very grateful to the network **“European Women in Mathematics”** for supporting my participation at this Workshop.

European Women in Mathematics is an international association of women working in the field of mathematics in Europe. EWM aims at:

- encouraging women to study mathematics
- supporting women in their careers
- providing a meeting place for like-minded people
- promoting scientific communication
- cooperating with groups and organizations with similar goals
- giving prominence and visibility to women mathematicians
- spreading their vision of mathematics and science

- Founded in 1986, EWM has several hundred members and coordinators in 33 European countries.
- Every other year, EWM holds a general meeting and a summer school.

http://europeanwomeninmaths.org