HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

On balanced sampling and calibration estimation in survey sampling

Risto Lehtonen
University of Helsinki

BaNoCoSS 2019, Örebro University, 16-20 June 2019

Topics to be addressed

Motivation

Representative strategy by Hájek

Balanced sampling & calibration estimation

Hájek and HT type calibration estimators

Examples

Discussion

Jaroslav Hájek (1926-1974)

Important contributions in statistics:

Representative strategy à la Hájek

Hájek J. (1959) Optimum strategy and other problems in probability sampling, Casopis pro Pestováni Matematiky, 84, 387–423.

Hájek estimator of population mean under unequal probability sampling

Hájek J. (1971) Comment on “An essay on the logical foundations of survey sampling” by Basu, D. In Godambe V.P. and Sprott D.A. (eds.) Foundations of Statistical Inference, p. 236. Holt, Rinehart and Winston.

Motivation

METRON - International Journal of Statistics
2011, vol. LXIX, n. 1, pp. 45-65
MATTI LANGEL – YVES TILLÉ

3. REPRESENTATIVENESS

3.1. A polysemic term

The idea and concept of *representativeness* was already used in Kiaer's work (Kiaer, 1896, 1899, 1903, 1905). Because the idea of a *representative sample* is reassuring for an uninitiated audience as it provides an illusion of scientific validity, it has been an important notion in sampling ever since. However, the multiplicity of definitions to which it can be associated has been at the core of many debates and misunderstandings in the history of sampling. Thus, the term is much less used in modern survey sampling literature and in our opinion it is a term best to avoid in survey methodology.

Representative strategy

in the spirit of Jaroslav Hájek (1959, 1981)

Strategy:

a couple of *sampling design* and *estimation design*

Representative strategy:

strategy that estimates the totals of auxiliary variables exactly (without error)

Let $\mathbf{z}_k = (z_{1k}, z_{2k}, ..., z_{Lk})'$ be our auxiliary data vector for unit $k \in U$ in population $U = \{1, ..., k, ..., N\}$

Define weights w_k for $k \in U$ such that the **representativeness equations**

$$\sum_{k \in s} w_k \mathbf{z}_k = \sum_{k \in U} \mathbf{z}_k$$

are fulfilled, where s denotes a sample from U

Options

It is obvious that a representative strategy can be constructed

- under the sampling design
- under the estimation design
- under both the sampling and estimation designs

For sampling design, $\mathbf{z}_k = (z_{1k}, z_{2k}, ..., z_{Lk})'$ denotes the auxiliary data vector for unit k in population $U = \{1, ..., k, ..., N\}$

For estimation design, let $\mathbf{x}_k = (x_{1k}, x_{2k}, ..., x_{Jk})'$ be another auxiliary data vector for unit k in U

z-vectors and x-vectors may be separate or overlapping vectors

Strategy 1: Horvitz-Thompson estimation for a balanced probability sample

Representativeness through the sampling design

Auxiliary data are incorporated in the *sampling procedure*

Deville and Tillé (2004), Tillé (2011)

Sampling design : Compute inclusion probabilities π_k that satisfy the **balancing equations** for any sample *s*:

$$\sum_{k \in s} \mathbf{z}_k / \pi_k = \sum_{k \in U} \mathbf{z}_k$$

Estimation design : Horvitz-Thompson estimator

$$\hat{t}_{HT} = \sum_{k \in s} a_k y_k$$

where $a_k = 1/\pi_k$ are design weights

The sampling design is balanced on the auxiliary z-variables

Strategy 2: Calibration estimation for a (generic) probability sample

Representativeness through the estimation design

Auxiliary data are incorporated in the *estimation procedure*

Deville & Särndal (1992), Särndal (2007)

Compute adjustment factors g_k that satisfy

the **calibration equations** for the given probability sample *s*

$$\sum_{k \in s} g_k \mathbf{x}_k / \pi_k = \sum_{k \in U} \mathbf{x}_k$$

Estimation design: Model-free calibration estimator

$$\hat{t}_{CAL} = \sum_{k \in s} w_k y_k$$

where $w_k = g_k / \pi_k$ are calibration weights

The estimation design is balanced on the auxiliary x-variables

Remarks

In practical applications, the availability & share of labour between the auxiliary z-data (sampling phase) and auxiliary x-data (estimation phase) becomes an issue

Balanced sampling: z-data are needed at the sampling unit level

Calibration estimation: x-data are needed either at an aggregate level or at the unit level, depending on the calibration method

Basic developments

Sampling design: The CUBE method

Deville and Tillé (2004) Efficient balanced sampling: The cube method (Biometrika).

Penalization:

Breidt and Chauvet (2012) Penalized balanced sampling (Biometrika).

Estimation design: Calibration

Deville and Särndal (1992). Calibration estimators in survey sampling (JASA).

Penalization:

Guggemos and Tillé (2010) Penalized calibration in survey sampling: Design-based estimation assisted by mixed models (Journal of Statistical Planning and Inference).

Biometrika (2004), **91**, 4, *pp.* 893–912

Printed in Great Britain

Efficient balanced sampling: The cube method

BY JEAN-CLAUDE DEVILLE

Laboratoire de Statistique d'Enquête, CREST–ENSAI,
École Nationale de la Statistique et de l'Analyse de l'Information, rue Blaise Pascal,
Campus de Ker Lann, 35170 Bruz, France

deville@ensai.fr

AND YVES TILLÉ

Groupe de Statistique, Université de Neuchâtel, Espace de l'Europe 4, Case postale 805,
2002 Neuchâtel, Switzerland

yves.tille@unine.ch

Example 1: Deville & Tillé (2004)

$U = \{1,...,k,...,N\}$ real population (MU284), $N = 280$

$\mathbf{z}_k = (z_{1k}, z_{2k}, z_{3k}, z_{4k})'$, $k \in U$ auxiliary data vector

for both sample balancing and calibration estimation

$a_k = 1/\pi_k$ design weights

$w_k = g_k a_k$ calibration weights

HT estimators of totals of y_j: $\hat{t}_{HT}(y_j) = \sum_{k \in s} a_k y_{jk}$, $j = 1,...,6$

Calibration estimators $\hat{t}_{CAL}(y_j) = \sum_{k \in s} w_k y_{jk} = \hat{t}_{HT}(y_j) + (\mathbf{t}_z - \hat{\mathbf{t}}_{HTz})'\mathbf{B}_j$

where $\mathbf{B}_j = \left(\sum_{k \in s} a_k \mathbf{z}_k \mathbf{z}_k'\right)^{-1} \sum_{k \in s} a_k \mathbf{z}_k y_{jk}$

Simulation experiments

$K = 1000$ fixed-size samples from U, $n = 20$

...contd.

Strategies for the 6 target variables $y_1, y_2, ..., y_6$

a) Non-balanced sampling and HT estimation

b) Balanced sampling and HT

c) Non-balanced sampling and CAL estimation

d) Balanced sampling and CAL

NOTE: Actually, sampling in a) and c) is with balancing with CUBE but on a single variable (z_1)

Results on accuracy

Table1 Estimators of population total: Monte Carlo MSE relative to the MSE for non-balanced sampling with HT estimator

Target variable	Horvitz-Thompson		Calibration	
	Non-balanced samples	Balanced samples	Non-balanced samples	Balanced samples
y_1	1	0.90	0.82	0.76
y_2	1	0.91	1.02	0.87
y_3	1	0.80	0.92	0.82
y_4	1	0.21	0.11	0.11
y_5	1	0.15	0.21	0.08
y_6	1	0.26	0.15	0.14

Extracted from Deville & Tillé (2004) p. 909 Table 1

Analysis

Target variable y	Balancing & HT	Balancing & CAL
y_1	0.90	0.76
y_2	0.91	0.87
y_3	0.80	0.82
y_4	0.21	0.11
y_5	0.15	0.08
y_6	0.26	0.14

Correlation of aux. var. z				
	z_1	z_2	z_3	z_4
z_1	1.00	0.99	-	0.98
z_2	0.99	1.00	-	0.99
z_3	-	-	1.00	-
z_4	0.98	0.99		1.00

Table 2 Correlation of auxiliary variables with target variables in the population and R square for regression model (N=280)

Auxiliary variables	Target variables					
	y_1	y_2	y_3	y_4	y_5	y_6
z_1	-	0.99	0.63	0.87	0.89	-
z_2	-	0.99	0.65	0.85	0.90	-
z_3	-	-	-	-	-	-
z_4	-	0.99	0.64	0.85	0.90	-
R^2	-	0.99	0.42	0.76	0.81	-

- no data

Lithuanian Journal of Statistics *2016, vol. 55, No 1, pp. 81-90* *www.statisticsjournal.lt*
Lietuvos statistikos darbai *2016, 55 t., Nr. 1, 81-90 p.*

APPLICATION OF BALANCED SAMPLING, NON-RESPONSE AND CALIBRATED ESTIMATOR

Ieva Dirdaitė[1], Danutė Krapavickaitė[2]

[1]Pandaconnect, UAB. Address: Saulėtekio al. 15, Vilnius, 10224, Lithuania
[2]Vilnius Gediminas Technical University. Address: Saulėtekio al. 11, Vilnius, 10223, Lithuania
E-mail: [1]dirdaite.ieva@gmail.com, [2]danute.krapavickaite@vgtu.lt

COMMENT: Interesting empirical exploration on the interplay between balanced sampling and calibration estimation by simulation experiments using real survey data

Several strategies are applied by combining balanced and non-balanced sampling and Horvitz-Thompson and calibration estimators

www.statisticsjournal.lt

Remarks

The previous representative design-based strategies were *model-free* because **statistical models** did not play an explicit role

Model-assisted methods in representative design-based strategies:

- **Balanced sampling**
  Penalized balanced sampling (Breidt & Chauvet 2012)
- **Calibration estimation**
  Penalized calibration (Guggemos & Tillé 2010)
  Generalized calibration (Deville 2000)
  Model calibration (Wu & Sitter 2001)
- **Calibration in small domain estimation**
  Model-assisted calibration (Lehtonen & Veijanen 2012, 2016)
  Multiple model calibration (Montanari & Ranalli 2009)
  Two-level hybrid calibration (Lehtonen & Veijanen 2017)

Biometrika (2012), **99**, 4, *pp.* 945–958

Printed in Great Britain

doi: 10.1093/biomet/ass033
Advance Access publication 26 July 2012

Penalized balanced sampling

BY F. J. BREIDT

Department of Statistics, Colorado State University, Fort Collins, Colorado 80523-1877, U.S.A.
jbreidt@stat.colostate.edu

AND G. CHAUVET

Crest (Ensai), Campus de Ker Lann, rue Blaise Pascal, 35170 Bruz, France
chauvet@ensai.fr

Example 2: Breidt & Chauvet (2012)

Linear mixed modeling in penalized balanced sampling by relaxing some balance constraints

Analogous to the use of penalization at the estimation stage (Guggemos & Tillé 2010) for reducing some calibration constraints

Why?
Ordinary balanced samples may reduce the need for calibration weighting in the estimation phase (Deville & Tillé example)

Penalized balanced samples may reduce the need for linear mixed modeling (penalized calibration) in the estimation phase

Gain:
HT estimators for penalized balanced samples will be efficient for target variables well approximated by a linear mixed model

$$y_k = \mathbf{x}_k'\boldsymbol{\beta} + \mathbf{z}_k'\mathbf{u} + \varepsilon_k,\ k \in U$$

where $\boldsymbol{\beta}$ are fixed effects and $\mathbf{u}$ are random effects

Breidt & Chauvet contd.

Monte Carlo study including balanced sampling guided by a penalized spline expressed as a linear mixed model

Generated artificial population of $N = 1000$

Auxiliary variable $x_{1k} = (1 + z_{1k})^{-1}$, z_1 lognormal

$x_{2k} = (1 + z_{2k})^{-1}$, z_2 lognormal, independent of z_1

Target variables y_1 and y_2

Linear model $m_2 = 1 + 2(x - 0.5)$, Exponential model $m_6 = \exp(-8x)$

Sampling designs defined by x_1

Estimation designs for y_1 defined by x_1 and for y_2 by x_2

Strategy $(x_1 : x_1)$ x_1 for sampling design & estimation design

Strategy $(x_1 : x_2)$ x_1 for sampling design and x_2 for estimation design

Simulation experiments: $K = 5000$ simulated samples of size $n = 100$

Results on accuracy

Table 3 RMSE of strategies relative to the RMSE of HT estimator of total under penalized balanced sampling

Sampling	Penalized balanced sampling		Balanced sampling		Simple random sampling
Estimation	HT	LMM	HT	LMM	LMM
Strategy ($x_1 : x_1$) for y_1					
Linear (m_2)	1	1.00	1.00	1.00	1.07
Exponential (m_6)	1	1.00	1.00	0.99	1.07
Strategy ($x_1 : x_2$) for y_2					
Linear (m_2)	1	0.66	0.99	0.66	0.66
Exponential (m_6)	1	0.84	1.00	0.83	0.88

Extracted from Table 1 in Breidt & Chauvet (2010) p. 953

Example 3: Lehtonen & Veijanen (2019)

Design-based simulation experiment for finite population generated by a linear mixed model with random intercepts and slopes

Population: 1 million units and 40 unplanned domains
Estimation of domain totals $t_d = \sum_{k \in U_d} y_k, \;\; d = 1,...,40$
with direct and indirect Hájek and Horvitz-Thompson estimators

Auxiliary data vector $\mathbf{x}_k = (x_{1k}, x_{2k}, x_{3k})', \; k \in U_d, \; d = 1,...,40$
utilized in the estimation phase

Strategy: SRSWOR & model-free and model-assisted estimators

Assisting model: Linear mixed model

Monte Carlo experiments
K = 10,000 SRSWOR samples of *n* = 2000 units

HT and Hájek estimators for domain totals

Direct expansion type estimators

HT estimators $\hat{t}_{dHT} = \sum_{k \in s_d} a_k y_k, \quad d = 1,...,40$

Hájek estimators $\hat{t}_{dHA} = N_d \times \frac{\sum_{k \in s_d} a_k y_k}{\sum_{k \in s_d} a_k}, \quad d = 1,...,40$

where $a_k = 1/\pi_k$ are design weights

Direct and indirect calibration estimators

HT type calibration estimators $\hat{t}_{dCAL-HT} = \sum_{k \in s_d} w_k y_k, \quad d = 1,...,40$

Hájek type calibration estimators

$$\hat{t}_{dCAL-HA} = N_d \times \frac{\sum_{k \in s_d} w_{dk} y_k}{\sum_{k \in s_d} w_{dk}}$$

where $w_{dk} = g_{dk} a_k$ are method-specific calibration weights

Calibration vectors for model-free calibration

Calibration equations for MFC

$$\sum_{k \in s_d} w_{dk} \mathbf{x}_k = \sum_{k \in U_d} \mathbf{x}_k, \quad d = 1,...,40$$

w_{dk} calibration weight for element k in domain d

Calibration vectors

MFC-HT: $\mathbf{x}_k = (1, x_{1k}, x_{2k}, x_{3k})', \; k \in U_d, \; d = 1,...,40$

MFC-HA: $\mathbf{x}_k = (x_{1k}, x_{2k}, x_{3k})', \; k \in U_d, \; d = 1,...,40$

NOTE: Domain estimators are of **direct** type

Calibration vectors for model-assisted calibration

Calibration equations for MC

$$\sum_{k \in s_d} w_{dk} \hat{y}_k = \sum_{k \in U_d} \hat{y}_k, \quad d = 1,...,40$$

Calibration vectors

MC-HT: $\mathbf{z}_k = (1, \hat{y}_k)', \; k \in U_d, \; d = 1,...,40$

MC-HA: $\mathbf{z}_k = \hat{y}_k, \; k \in U_d, \; d = 1,...,40$

Assisting model

Linear mixed model with domain-specific random intercepts

$$y_k = \mathbf{x}_k' \boldsymbol{\beta} + u_d = (\beta_0 + u_{0d}) + \beta_1 x_{1k} + \beta_2 x_{2k} + \beta_3 x_{3k} + \varepsilon_k, \; k \in U_d$$

Predictions

$$\hat{y}_k = \mathbf{x}_k' \hat{\boldsymbol{\beta}} + \hat{u}_d \text{ with } \mathbf{x}_k = (1, x_{1k}, x_{2k}, x_{3k})', \; k \in U_d$$

$\hat{y}_k$ calculated for all $k \in U_d$

NOTE: Estimators are of **indirect** type

Accuracy of estimators

Relative root mean squared error (RRMSE)

$$\text{RRMSE}(\hat{t}_d) = \sqrt{\frac{1}{K}\sum_{i=1}^{K}(\hat{t}_d(s_i) - t_d)^2} \; / \, t_d, \quad d = 1,\ldots,D$$

where

$\hat{t}_d(s_i)$ estimate from sample s_i for domain d

t_d known parameter value in domain d

K number of simulated samples

NOTE: MFC and MC: Nearly design unbiased

Largest $ARB(\hat{t}_d) < 0.2\%$

Results on accuracy

Table 4 Median RRMSE (%) of design-based direct HT and Hájek estimators for totals for 40 domains in three domain sample size classes in a simulation experiment of 10,000 SRSWOR samples of 2000 units from a synthetic population of one million units.

	Expected domain sample size			All
	Minor 12	Medium 40	Major 122	
Horvitz-Thompson $\hat{t}_{dHT} = \sum_{k \in s_d} a_k y_k$	29.00	15.77	8.79	15.80
Hájek $\hat{t}_{dHA} = N_d \times \frac{\sum_{k \in s_d} a_k y_k}{\sum_{k \in s_d} a_k}$	4.60	1.85	0.91	1.96

Extracted from Lehtonen & Veijanen (2019)

Table 5 Median RRMSE (%) of design-based direct and indirect HT and Hájek type calibration estimators for totals for 40 domains in three domain sample size classes in a simulation experiment of 10,000 SRSWOR samples of 2000 units from a synthetic population of one million units.

	Expected domain sample size			All
	Minor 12	Medium 40	Major 122	
Model-free calibration MFC Calibration vectors $\mathbf{z}_k = (1, x_{1k}, x_{2k}, x_{3k})'$ and $\mathbf{z}_k = (x_{1k}, x_{2k}, x_{3k})'$				
MFC-HT	8.82	1.62	0.78	1.72
MFC-HA	6.39	1.89	0.91	1.98
Model-assisted calibration MC Model: $y_k = \mathbf{x}_k'\boldsymbol{\beta} + u_d + \varepsilon_k,\ k \in U_d,\ d = 1,...,D$ Model vector $\mathbf{x}_k = (1, x_{1k}, x_{2k}, x_{3k})'$ Calibration vectors $\mathbf{z}_k = (1, \hat{y}_k)'$ and $\mathbf{z}_k = \hat{y}_k$				
MC-HT	4.29	1.58	0.78	1.67
MC-HA	4.53	1.85	0.91	1.96

Extracted from Lehtonen & Veijanen (2019)

Distribution of calibrated weights

Problems of practical concern in *model-free* calibration:

- Possible large variation of weights
- Weights smaller than one, negative weights
- Positive but extremely small weights

To what extent can *model-assisted* calibration methods help?

Any differences between HT type vs. Hájek type methods?

Small simulation experiment:

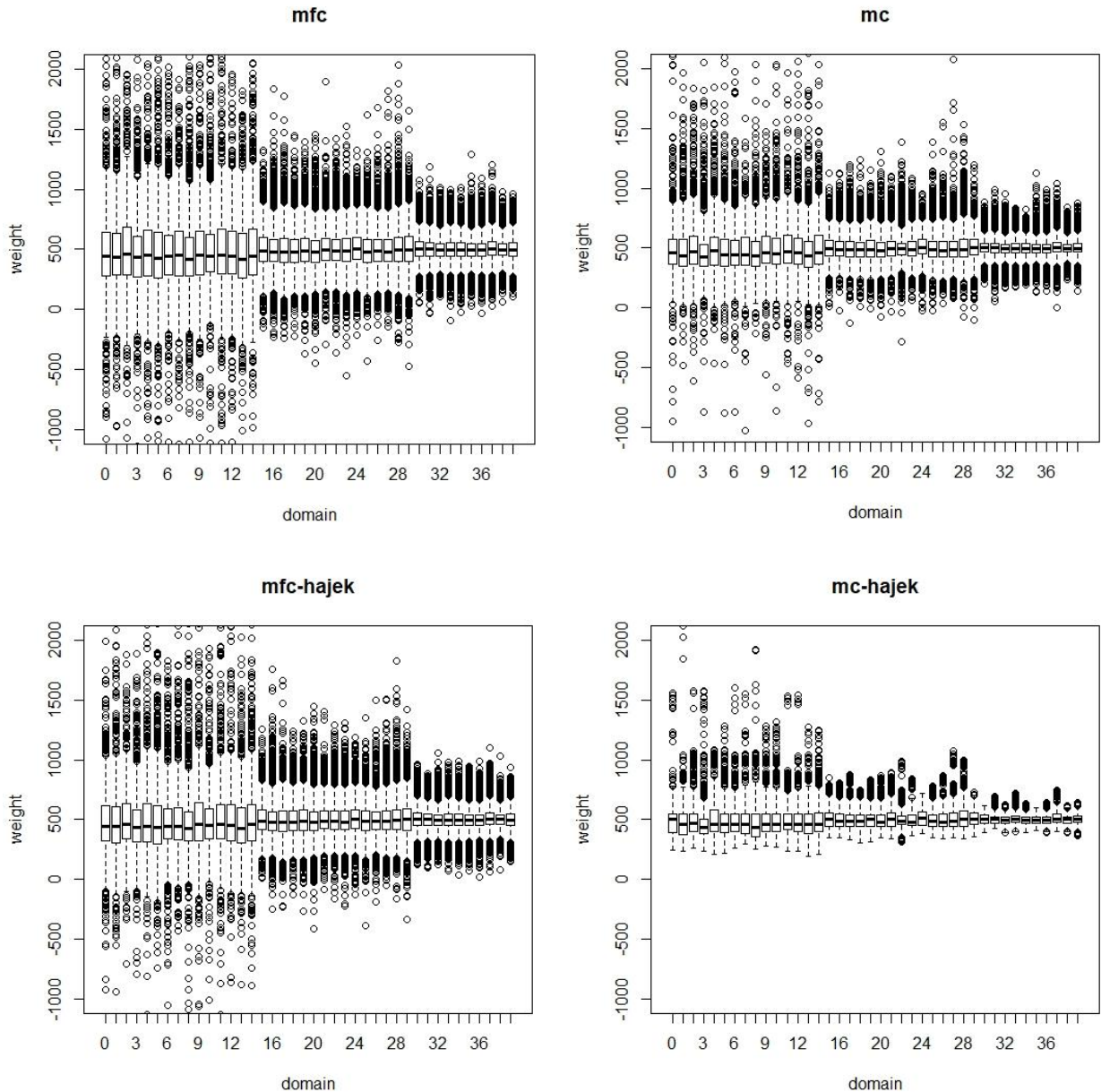
- 100 SRSWOR samples of size 2,000 elements from *U*

Results: Distribution of weights by domain size

HT weights: $w_{HTdk} = w_{dk}$

Comparable Hájek weights: $w_{HAdk} = N_d \times \frac{w_{dk}}{\sum_{k \in s_d} w_{dk}}$

Fig. 1 Distribution of weights by domain size class in simulation experiment of 100 SRSWOR samples from population *U*
Upper panel: HT type estimators, lower: Hájek type estimators



Discussion

Can strategies that combine *balanced sampling* and *calibration estimation* extend effectively the use of auxiliary data in survey strategies? What are the benefits / drawbacks?

These combined strategies may (or, may not) offer an interesting framework:

- for methodological research
- for experimentation in practical applications
- In what areas in particular?

A special interest is in strategies for sampling and estimation phases that involve approaches connected to GLMM type modelling

A challenging framework is provided by small domain estimation

References

Breidt, F.J. and Chauvet, G. (2012) Penalized balanced sampling. Biometrika, 99, 945–958.

Deville, J.-C. (2000) Generalized calibration and application to weighting for non-response. In: Bethlehem J.G. and van der Heijden, P.G.M. (eds) COMPSTAT. Physica, Heidelberg.

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

Deville, J.-C. and Tillé, Y. (2004) Efficient balanced sampling: The cube method. Biometrika, 91, 893–912.

Dirdaite, I. and Krapavickaite, D. (2916) Application of balanced sampling, non-response and calibrated estimator. Lithuanian Journal of Statistics 2016, 55, 81–90.

Guggemos, F. and Tillé, Y. (2010) Penalized calibration in survey sampling: Design-based estimation assisted by mixed models. Journal of Statistical Planning and Inference, 140, 3199–3212.

Hájek, J. (1959) Optimum strategy and other problems in probability sampling, Casopis pro Pestováni Matematiky, 84, 387–423.

Hájek, J. (1981) Sampling from a Finite Population. New York: Marcel Dekker.

Lehtonen, R. and Veijanen, A. (2012) Small area poverty estimation by model calibration. Journal of the Indian Society of Agricultural Statistics, 66, 125–133.

References

Lehtonen R. and Veijanen A. (2016) Design-based methods to small area estimation and calibration approach. In: Pratesi M. (Ed.) Analysis of Poverty Data by Small Area Estimation. Chichester: Wiley.

Lehtonen R. and Veijanen A. (2017) A two-level hybrid calibration technique for small area estimation. SAE2017 Conference, Paris, June 2017.

Lehtonen, R. and Veijanen, A. (2019) Small domain estimation with calibration methods. ITACOSM 2019 Conference, 5-7 June 2019, Florence, Italy.

Montanari G.E. and Ranalli M.G. (2009) Multiple and ridge model calibration. Proceedings of Workshop on Calibration and Estimation in Surveys 2009. Statistics Canada.

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

Tillé, Y. (2011) Ten years of balanced sampling with the cube method: An appraisal. Survey Methodology 37, 215–226.

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

Thank you for your attention