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
Important contributions in statistics:

Representative strategy à la Hájek


Hájek estimator of population mean under unequal probability sampling

Motivation

METRON - International Journal of Statistics
2011, vol. LXIX, n. 1, pp. 45-65
MATTI LANGE-–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 \( z_k = (z_{1k}, z_{2k}, \ldots, z_{Lk})' \) be our auxiliary data vector for unit \( k \) in population \( U = \{1, \ldots, k, \ldots, N\} \)

Define weights \( w_k \) for \( k \in U \) such that the *representativeness equations*
\[
\sum_{k \in s} w_k z_k = \sum_{k \in U} z_k
\]
are fulfilled, where \( s \) denotes a sample from \( U \)
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, \( z_k = (z_{1k}, z_{2k}, \ldots, z_{Lk})' \) denotes the auxiliary data vector for unit \( k \) in population \( U = \{1, \ldots, k, \ldots, N\} \)

For estimation design, let \( x_k = (x_{1k}, x_{2k}, \ldots, x_{Jk})' \) 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 \( \pi_k \) that satisfy the balancing equations for any sample \( s \):

\[
\sum_{k \in s} z_k / \pi_k = \sum_{k \in U} z_k
\]

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 $s$

$$\sum_{k \in s} g_k x_k / \pi_k = \sum_{k \in U} x_k$$

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**
Penalization:

**Estimation design: Calibration**
Penalization:
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)

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

$z_k = (z_{1k}, z_{2k}, z_{3k}, z_{4k})'$, $k \in U$ 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 $y_j$: $\hat{t}_{HT}(y_j) = \sum_{k \in S} a_k y_{jk}$, $j = 1, \ldots, 6$

Calibration estimators $\hat{t}_{CAL}(y_j) = \sum_{k \in S} w_k y_{jk} = \hat{t}_{HT}(y_j) + (t_z - \hat{t}_{HTz})'B_j$

where $B_j = \left(\sum_{k \in S} a_k z_k z_k'\right)^{-1} \sum_{k \in S} a_k z_k y_{jk}$

Simulation experiments

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

Strategies for the 6 target variables $y_1, y_2, \ldots, y_6$

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

Table 1: Estimators of population total: Monte Carlo MSE relative to the MSE for non-balanced sampling with HT estimator

<table>
<thead>
<tr>
<th>Target variable</th>
<th>Horvitz-Thompson</th>
<th>Calibration</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Non-balanced</td>
<td>Balanced</td>
</tr>
<tr>
<td></td>
<td>samples</td>
<td>samples</td>
</tr>
<tr>
<td>$y_1$</td>
<td>1</td>
<td>0.90</td>
</tr>
<tr>
<td>$y_2$</td>
<td>1</td>
<td>0.91</td>
</tr>
<tr>
<td>$y_3$</td>
<td>1</td>
<td>0.80</td>
</tr>
<tr>
<td>$y_4$</td>
<td>1</td>
<td>0.21</td>
</tr>
<tr>
<td>$y_5$</td>
<td>1</td>
<td>0.15</td>
</tr>
<tr>
<td>$y_6$</td>
<td>1</td>
<td>0.26</td>
</tr>
</tbody>
</table>

Extracted from Deville & Tillé (2004) p. 909 Table 1
Table 2 Correlation of auxiliary variables with target variables in the population and R square for regression model (N=280)

<table>
<thead>
<tr>
<th>Auxiliary variables</th>
<th>Target variables</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( y_1 )</td>
</tr>
<tr>
<td>( z_1 )</td>
<td>-</td>
</tr>
<tr>
<td>( z_2 )</td>
<td>-</td>
</tr>
<tr>
<td>( z_3 )</td>
<td>-</td>
</tr>
<tr>
<td>( z_4 )</td>
<td>-</td>
</tr>
<tr>
<td>( R^2 )</td>
<td>-</td>
</tr>
</tbody>
</table>

- no data

<table>
<thead>
<tr>
<th>Correlation of aux. var. z</th>
</tr>
</thead>
<tbody>
<tr>
<td>( z_1 )</td>
</tr>
<tr>
<td>( z_1 )</td>
</tr>
<tr>
<td>( z_2 )</td>
</tr>
<tr>
<td>( z_3 )</td>
</tr>
<tr>
<td>( z_4 )</td>
</tr>
</tbody>
</table>
APPLICATION OF BALANCED SAMPLING, NON-RESPONSE AND CALIBRATED ESTIMATOR

Ieva Dirdaitė¹, Danutė Krapavickaitė²

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

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

\[ y_k = x'_k \beta + z'_k u + \varepsilon_k, \quad k \in U \]

where \( \beta \) are fixed effects and \( u \) are random effects
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

<table>
<thead>
<tr>
<th>Sampling</th>
<th>Penalized balanced sampling</th>
<th>Balanced sampling</th>
<th>Simple random sampling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimation</td>
<td>HT</td>
<td>LMM</td>
<td>HT</td>
</tr>
<tr>
<td><strong>Strategy ((x_1 : x_1)) for (y_1)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear ((m_2))</td>
<td>1</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Exponential ((m_6))</td>
<td>1</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td><strong>Strategy ((x_1 : x_2)) for (y_2)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear ((m_2))</td>
<td>1</td>
<td>0.66</td>
<td>0.99</td>
</tr>
<tr>
<td>Exponential ((m_6))</td>
<td>1</td>
<td>0.84</td>
<td>1.00</td>
</tr>
</tbody>
</table>

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 domains
Estimation of domain totals \( t_d = \sum_{k \in U_d} y_k, \ d = 1, \ldots, 40 \)
with direct and indirect Hájek and Horvitz-Thompson estimators

Auxiliary data vector \( x_k = (x_{1k}, x_{2k}, x_{3k})', \ k \in U_d, \ d = 1, \ldots, 40 \)
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 Hajék estimators for domain totals

Direct expansion type estimators

HT estimators

\[ \hat{t}_{dHT} = \sum_{k\in s_d} a_k y_k, \quad d = 1,\ldots, 40 \]

Hajek estimators

\[ \hat{t}_{dHA} = N_d \times \frac{\sum_{k\in s_d} a_k y_k}{\sum_{k\in s_d} a_k}, \quad d = 1,\ldots, 40 \]

where \( a_k = 1/\pi_k \) are design weights

Direct and indirect calibration estimators

HT type calibration estimators

\[ \hat{t}_{dCAL-HT} = \sum_{k\in s_d} w_k y_k, \quad d = 1,\ldots, 40 \]

Hajek 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

\[ \sum_{k \in s_d} w_{dk} x_k = \sum_{k \in U_d} x_k, \quad d = 1, \ldots, 40 \]

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

Calibration vectors

MFC-HT: \( x_k = (1, x_{1k}, x_{2k}, x_{3k})', \quad k \in U_d, \quad d = 1, \ldots, 40 \)

MFC-HA: \( x_k = (x_{1k}, x_{2k}, x_{3k})', \quad k \in U_d, \quad d = 1, \ldots, 40 \)

NOTE: Domain estimators are of direct type
Calibration vectors for model-assisted calibration

Calibration equations for MC

\[ \sum_{k \in s_d} w_{dk} \hat{y}_k = \sum_{k \in U_d} \hat{y}_k, \quad d = 1, \ldots, 40 \]

Calibration vectors

- **MC-HT:** \[ z_k = (1, \hat{y}_k)' , \quad k \in U_d, \quad d = 1, \ldots, 40 \]
- **MC-HA:** \[ z_k = \hat{y}_k, \quad k \in U_d, \quad d = 1, \ldots, 40 \]

Assisting model

Linear mixed model with domain-specific random intercepts

\[ y_k = x_k' \beta + u_d = (\beta_0 + u_{0d}) + \beta_1 x_{1k} + \beta_2 x_{2k} + \beta_3 x_{3k} + \epsilon_k, \quad k \in U_d \]

Predictions

\[ \hat{y}_k = x_k' \hat{\beta} + \hat{u}_d \text{ with } x_k = (1, x_{1k}, x_{2k}, x_{3k})', \quad k \in U_d \]

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

NOTE: Estimators are of **indirect** type
## 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.

<table>
<thead>
<tr>
<th>Expected domain sample size</th>
<th></th>
<th></th>
<th></th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minor</td>
<td>12</td>
<td>40</td>
<td>122</td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>Major</td>
<td>122</td>
<td>122</td>
<td>122</td>
<td></td>
</tr>
</tbody>
</table>

Horvitz-Thompson

\[
\hat{t}_{dHT} = \sum_{k \in s_d} a_k y_k
\]

<table>
<thead>
<tr>
<th></th>
<th>Minor</th>
<th>Medium</th>
<th>Major</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>29.00</td>
<td>15.77</td>
<td>8.79</td>
<td>15.80</td>
</tr>
</tbody>
</table>

Hájek

\[
\hat{t}_{dHA} = N_d \times \frac{\sum_{k \in s_d} a_k y_k}{\sum_{k \in s_d} a_k}
\]

<table>
<thead>
<tr>
<th></th>
<th>Minor</th>
<th>Medium</th>
<th>Major</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4.60</td>
<td>1.85</td>
<td>0.91</td>
<td>1.96</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Expected domain sample size</th>
<th>Minor</th>
<th>Medium</th>
<th>Major</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>12</td>
<td>40</td>
<td>122</td>
<td></td>
</tr>
</tbody>
</table>

Model-free calibration MFC
Calibration vectors \( z_k = (1, x_{1k}, x_{2k}, x_{3k})' \) and \( z_k = (x_{1k}, x_{2k}, x_{3k})' \)

<table>
<thead>
<tr>
<th></th>
<th>MFC-HT</th>
<th>MFC-HA</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>MFC-HT</td>
<td>8.82</td>
<td>1.62</td>
<td>0.78</td>
<td>1.72</td>
</tr>
<tr>
<td>MFC-HA</td>
<td>6.39</td>
<td>1.89</td>
<td>0.91</td>
<td>1.98</td>
</tr>
</tbody>
</table>

Model-assisted calibration MC
Model: \( y_k = x_k'\beta + u_d + \varepsilon_k, \ k \in U_d, \ d = 1,\ldots, D \)

Model vector \( x_k = (1, x_{1k}, x_{2k}, x_{3k})' \) Calibration vectors \( z_k = (1, \hat{y}_k)' \) and \( z_k = \hat{y}_k \)

<table>
<thead>
<tr>
<th></th>
<th>MC-HT</th>
<th>MC-HA</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>MC-HT</td>
<td>4.29</td>
<td>1.58</td>
<td>0.78</td>
<td>1.67</td>
</tr>
<tr>
<td>MC-HA</td>
<td>4.53</td>
<td>1.85</td>
<td>0.91</td>
<td>1.96</td>
</tr>
</tbody>
</table>

Extracted from Lehtonen & Veijanen (2019)
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 Hajek weights: $w_{HAdk} = N_d \times \sum_{k \in S_d} \frac{w_{dk}}{w_{dk}}$
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: Hajek 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 framework is provided by small domain estimation.
References


Thank you for your attention