Some aspects of using calibration in polish surveys

Marcin Szymkowiak

Statistical Office in Poznań
University of Economics in Poznań

Outline

1. Theoretical aspects of calibration
    - Definition of calibration
    - The problem of finding calibration weights
    - The calibration estimator for total
    - Statistical software
    - Examples
2. Calibration in NCPH 2011
    - The NCPH 2011 Methodology
    - Practical aspects of calibration in NCPH 2011
3. Assessing the feasibility of using information from administrative registers for calibration in business statistics
    - Simulation study
    - Chosen results
    - Conclusions

Theoretical background of calibration

Theoretical background of calibration

1. This technique was proposed by Devill and Särndal (1992) and is a method of searching for so called calibrated weights by minimizing distance measure between the sampling weights and the new weights, which satisfy certain calibration constraints.
2. As a consequence when the new weights are applied to the auxiliary variables in the sample, they reproduce the known population totals of the auxiliary variables exactly.
3. It is also important that the new weights should be as close as possible to sampling weights in sense of chosen distance measure (Särndal C-E., Lundström S. 2005, Särndal C-E. 2007).

Theoretical background of calibration

Theoretical background of calibration

- Let us assume that the whole population $U = \{1, 2, \dots, N\}$ consists of N elements.
- From this population we draw, according to a certain sampling scheme, a sample $s \subseteq U$, which consists of n elements.
- Let π_i denote first order inclusion probability $\pi_i = P(i \in s)$ and $d_i = 1/\pi_i$ the design weight.
- Let us assume that our main goal is estimation of the total value of the variable y:

$$Y = \sum_{i=1}^{N} y_i, \tag{1}$$

where y_i denotes the value of the variable y for i-th unit, $i = 1, \dots, N$.

Theoretical background of calibration

Theoretical background of calibration

- Let $x_1, \ldots, x_k$ denote auxiliary variables which will be used in the process of finding calibration weights and let $\mathbf{X}_j$ denote the total value for the auxiliary variable x_j, $j = 1, \ldots, k$, e.i.

$$\mathbf{X}_j = \sum_{i=1}^{N} x_{ij}, \tag{2}$$

  where x_{ij} odenotes the value of j-th auxiliary variable for the i-th unit.

- In practice it occurs that:

$$\sum_{s} d_i x_{ij} \neq \mathbf{X}_j \tag{3}$$

  so calibration is required.

Theoretical background of calibration

Theoretical background of calibration

- Let $\mathbf{w} = (w_1, \ldots, w_n)^T$ denote the vector of calibration weights.
- Our main goal is to look for new weights w_i which are as close as possible to the design weights d_i and which allow us to get known population totals from administrative registers exactly.
- The process of construction calibration weights depends on the properly chosen distance function.
- Let G denote function for which the second derivative exists and:
  - $G(\cdot) \geq 0$,
  - $G(1) = 0$,
  - $G'(1) = 0$,
  - $G''(1) = 1$.

Examples of G function

Examples of G function

$$G_1(x) = \frac{1}{2}(x-1)^2, \tag{4}$$

$$G_2(x) = \frac{(x-1)^2}{x}, \tag{5}$$

$$G_3(x) = x(\log x - 1) + 1, \tag{6}$$

$$G_4(x) = 2x - 4\sqrt{x} + 2, \tag{7}$$

$$G_5(x) = \frac{1}{2\alpha}\int_1^x \sinh\left[\alpha\left(t - \frac{1}{t}\right)\right] dt. \tag{8}$$

The choice of G function

The choice of G function

- The most common G function which can be used in the process of construction distance function is $G_1(x) = \frac{1}{2}(x-1)^2$. In this case we have:

$$D(\mathbf{w}, \mathbf{d}) = \sum_{i=1}^{n} d_i G\left(\frac{w_i}{d_i}\right) = \sum_{i=1}^{n} d_i \frac{1}{2}\left(\frac{w_i}{d_i} - 1\right)^2 = \frac{1}{2}\sum_{i=1}^{n} \frac{(w_i - d_i)^2}{d_i}. \qquad (9)$$

The problem of finding calibration weights

The problem of finding calibration weights

(C1) Find the minimum of distance function:

$$D(\mathbf{w},\mathbf{d}) = \frac{1}{2}\sum_{i=1}^{n}\frac{(w_i - d_i)^2}{d_i} \longrightarrow \min, \tag{10}$$

(C2) Calibration equations:

$$\sum_{i=1}^{n} w_i x_{ij} = \mathbf{X}_j, \quad j = 1, \ldots, k, \tag{11}$$

(C3) Calibration constraints:

$$L \leq \frac{w_i}{d_i} \leq U, \quad \text{where: } L < 1 \text{ i } U > 1, \; i = 1, \ldots, n. \tag{12}$$

The calibration estimator for total

The calibration estimator for total

The calibration estimator for total takes the form:

$$\hat{Y}_{cal} = \sum_{i=1}^{n} w_i y_i \,, \tag{13}$$

where the vector of calibration weights $\mathbf{w} = (w_1, w_2, \ldots, w_n)^T$ is obtained as the following minimization problem:

$$\mathbf{w} = \mathrm{argmin}_v D(\mathbf{v}, \mathbf{d}) \,, \tag{14}$$

$$\mathbf{X} = \tilde{\mathbf{X}}, \tag{15}$$

where

$$D(\mathbf{v}, \mathbf{d}) = \frac{1}{2} \sum_{i=1}^{n} \frac{(v_i - d_i)^2}{d_i} \,, \tag{16}$$

$$\tilde{\mathbf{X}} = \left(\sum_{i=1}^{n} w_i x_{i1}, \sum_{i=1}^{n} w_i x_{i2}, \ldots, \sum_{i=1}^{n} w_i x_{ik} \right)^T, \quad \mathbf{X} = \left(\sum_{i=1}^{N} x_{i1}, \sum_{i=1}^{N} x_{i2}, \ldots, \sum_{i=1}^{N} x_{ik} \right)^T. \tag{17}$$

Theorem

Theorem

The solution of the minimization problem is the vector of calibration weights $\mathbf{w} = (w_1, w_2, \ldots, w_n)^T$, for which

$$w_i = d_i + d_i \left(\mathbf{X} - \hat{\mathbf{X}}\right)^T \left(\sum_{i=1}^{n} d_i \underline{\mathbf{x}}_i \underline{\mathbf{x}}_i^T\right)^{-1} \underline{\mathbf{x}}_i \tag{18}$$

where

$$\hat{\mathbf{X}} = \left(\sum_{i=1}^{n} d_i x_{i1}, \sum_{i=1}^{n} d_i x_{i2}, \ldots, \sum_{i=1}^{n} d_i x_{ik}\right)^T, \tag{19}$$

$$\underline{\mathbf{x}}_i = (x_{i1}, x_{i2}, \ldots, x_{ik})^T. \tag{20}$$

Statistical software

Statistical software

- **Bascula 4.0** - the statistical tool developed in the Delphi language by Statistics Netherlands for the calculation of estimates of population totals, means and ratios.
- **Calmar/Calmar 2** - the statistical software developed by INSEE.
- **Caljack** - this is a SAS macro written and developed by Statistics Canada and is an extension of the Calmar macro.
- **CALWGT** - this is a freely distributed program for calibration written by Li-Chun Zhang in S-plus for Unix.
- **CLAN 97** - the statistical software designed to handle surveys in Statistics Sweden.
- **G-Calib 2** - the statistical software developed in the SPSS language by Statistics Belgium.
- **GES** - this is a SAS-based application with a Windows-like interface which was developed in SAS/AF by Statistics Canada.
- **R** - this is a free statistical software. The calibrate function, which can be found in the survey package, reweights the survey design weights and also adds additional information about estimated standard errors.

CALMAR

CALMAR

- Although in many statistical packages the problem of finding calibration weights was implemented using different G functions in Poland CALMAR is preferred.
- In CALMAR, which is a macro written in 4GL in SAS four distance functions were implemented: the linear method, the raking ratio metod, the logit method, the truncated linear method.
- In CALMAR 2 which is a later version of CALMAR, the distance function based on hyperbolic sinus function was also implemented.

Example 1

Example 1

- We consider an artificial population of enterprises of size $N = 1000$ from which a simple random sample of size $n = 20$ is drawn. Hence design (initial) weights are equal $N/n = 1000/20 = 50$.
- We also consider a numerical variable x_1 (for instance monthly revenue of enterprise) and one categorical variable x_2 (for instance enterprise size i.e. large - L and medium - M).
- In this example it will be only shown how to compute calibration weights. We do not take into account the variable of interest y which is not necessary to compute calibration weights and would be necessary to calculate the variance of the estimator.

Example 1 – artificial data set

Example 1 – artificial data set

Number of enterprise	Monthly revenue x_1	Enterprise size x_2	d_i
1	18	M	50
2	14	M	50
3	16	M	50
4	35	L	50
5	30	L	50
6	10	L	50
7	15	M	50
8	23	M	50
9	23	L	50
10	12	M	50
11	18	M	50
12	16	M	50
13	22	L	50
14	15	M	50
15	15	M	50
16	10	M	50
17	18	M	50
18	18	M	50
19	35	L	50
20	16	M	50

Example 1

Example 1

- The weighted sum of variable x_1 is equal to 18950.
- Number of medium and large enterprises according to this survey is equal to 700 (14 medium enterprises x 50) and 300 (6 large enterprises x 50) respectively.
- **Assumption:** The exact population total of monthly revenue is known and equals 19000 and the real number of medium and large enterprises is equal to 720 and 280 respectively.
- **Problem** We would like to change the design weights in such a way that known auxiliary totals will be reproduced. In other words, we would like to slightly modify the initial weights so that the sum of x_1 based on the new weights is equal to 19000 and weighted sum of medium and large enterprises is equal to 720 and 280 respectively.
- **Solution:** Use calibration
- The SAS code which solves the problem for creating the preliminary datasets and rucalling the macro CALMAR2 command is given on the next slide.

Example 1 – solution using CALMAR2

```
/***********Creation of input dataset with drawn units*************/
data sample;
input enterprise $ size $ revenue weight;
cards;
ent01   M            18  50
ent02   M            14  50
ent03   M            16  50
ent04   L            35  50
ent05   L            30  50
ent06   L            10  50
ent07   M            15  50
ent08   M            23  50
ent09   L            23  50
ent10   M            12  50
ent11   M            18  50
ent12   M            16  50
ent13   L            22  50
ent14   M            15  50
ent15   M            15  50
ent16   M            10  50
ent17   M            18  50
ent18   M            18  50
ent19   L            35  50
ent20   M            16  50
;
run;
```

Example 1 – solution using CALMAR2

```
/**********Creation dataset with known population totals************/
data totals;
input var $ n mar1 mar2;
cards;
size 2 280 720
revenue 0 19000 .
;
run;

/*******************Library containing CALMAR************************/
libname calm 'D:\Lamborghini\Calibration';
options mstored sasmstore=calm;

/***********************Call to CALMAR*******************************/
%CALMAR2(DATAMEN=sample, POIDS=weight, IDENT=enterprise, MARMEN=totals,
M=1,DATAPOI=wcal, POIDSFIN=cal_weights)
```

Example 1 – calibration weights

Example 1 – calibration weights

Number of enterprise	Monthly revenue x_1	Enterprise size x_2	d_i	w_i
1	18	M	50	52,275
2	14	M	50	50,5821
3	16	M	50	51,4286
4	35	L	50	50,5462
5	30	L	50	48,4301
6	10	L	50	39,9657
7	15	M	50	51,0054
8	23	M	50	54,3911
9	23	L	50	45,4675
10	12	M	50	49,7357
11	18	M	50	52,275
12	16	M	50	51,4286
13	22	L	50	45,0443
14	15	M	50	51,0054
15	15	M	50	51,0054
16	10	M	50	48,8893
17	18	M	50	52,275
18	18	M	50	52,275
19	35	L	50	50,5462
20	16	M	50	51,4286

Example 2 – register based statistics (artificial data set)

Example 2 – register based statistics (artificial data set)

No.	Enterprise size	Section	Revenue	Legal status
1	Small	Section 1	NA	A
2	Large	Section 2	Small	B
3	Large	Section 2	High	NA
4	Small	Section 2	Small	C
5	Small	Section 1	NA	C
6	Small	Section 1	High	C
7	Large	Section 1	High	C
8	Large	Section 2	Small	C
9	Large	Section 1	Small	B
10	Small	Section 1	High	B
11	Large	Section 2	Small	B
12	Small	Section 2	Small	C
13	Large	Section 1	Small	A
14	Small	Section 2	Small	NA
15	Small	Section 1	High	B
16	Large	Section 2	Small	B
17	Large	Section 1	High	C
18	Small	Section 2	High	A
19	Small	Section 1	High	NA
20	Large	Section 1	Small	B

Example 2 – register based statistics

Two-way contingency table – the problem of nonresponse

- The main goal is to create two-way contingency table which shows the structure of revenue and legal status. Because of the fact that variables revenue and legal status are affected by nonresponse final table will not be correct.
- Description of variables: **Enterprise size** (Small, Large), **Legal status** (A, B, C), **Section** (Section 1, Section 2, Section 3), **Revenue** (Small, High), NA – not available.

	Revenue		
Legal status	Small	High	Total
A	1	1	2
B	5	2	7
C	3	3	6
Total	9	6	15

- The number of enterprises in two-way contingency tables does not add up to 20.
- **Solution:** Use calibration approach to adjust numbers in particular cells.

How to find calibration weights?

How to find calibration weights?

1. Create artificial "design weights". If for any enterprise the legal status or revenue is not known than initial weight $d_i = 0$. Otherwise $d_i = 1$.
2. Choose auxiliary variables. Because for all enterprises in register information about section and enterprise size is known use theme as covariates to find calibration weights w_i. In this example three variables were taken into account: $\mathbf{x}_{i1}$, $\mathbf{x}_{i2}$, $\mathbf{x}_{i3}$.

$$\mathbf{x}_{i1} = \begin{cases} 1 & \text{if } i\text{-th enterprise is large,} \\ 0 & \text{otherwise,} \end{cases} \tag{21}$$

$$\mathbf{x}_{i2} = \begin{cases} 1 & \text{if } i\text{-th enterprise is small,} \\ 0 & \text{otherwise,} \end{cases} \tag{22}$$

$$\mathbf{x}_{i3} = \begin{cases} 1 & \text{if } i\text{-th enterprise is from section 1} \\ 0 & \text{otherwise} \end{cases} \tag{23}$$

3. Use statistical software and find calibration weights w_i.

Example 2 – register based statistics (artificial data set)

Example 2 – register based statistics (artificial data set)

No.	Enterprise size	Section	Revenue	Legal status	d_i	$\mathbf{x}_{i1}$	$\mathbf{x}_{i2}$	$\mathbf{x}_{i3}$	w_i
1	Small	Section 1	NA	A	0	0	1	1	0
2	Large	Section 2	Small	B	1	1	0	0	1,0447761
3	Large	Section 2	High	NA	0	1	0	0	0
4	Small	Section 2	Small	C	1	0	1	0	1,6069652
5	Small	Section 1	NA	C	0	0	1	1	0
6	Small	Section 1	High	C	1	0	1	1	1,7263682
7	Large	Section 1	High	C	1	1	0	1	1,1641791
8	Large	Section 2	Small	C	1	1	0	0	1,0447761
9	Large	Section 1	Small	B	1	1	0	1	1,1641791
10	Small	Section 1	High	B	1	0	1	1	1,7263682
11	Large	Section 2	Small	B	1	1	0	0	1,0447761
12	Small	Section 2	Small	C	1	0	1	0	1,6069652
13	Large	Section 1	Small	A	1	1	0	1	1,1641791
14	Small	Section 2	Small	NA	0	0	1	0	0
15	Small	Section 1	High	B	1	0	1	1	1,7263682
16	Large	Section 2	Small	B	1	1	0	0	1,0447761
17	Large	Section 1	High	C	1	1	0	1	1,1641791
18	Small	Section 2	High	A	1	0	1	0	1,6069652
19	Small	Section 1	High	NA	0	0	1	1	0
20	Large	Section 1	Small	B	1	1	0	1	1,1641791

Example 2 – register based statistics

Two-way contingency table – before calibration

	Revenue		
Legal status	Small	High	Total
A	1	1	2
B	5	2	7
C	3	3	6
Total	9	6	15

Two-way contingency table – after calibration

	Revenue		
Legal status	Small	High	Total
A	1,16	1,61	2,77
B	6,14	2,77	8,91
C	4,27	4,05	8,32
Total	11,57	8,43	20

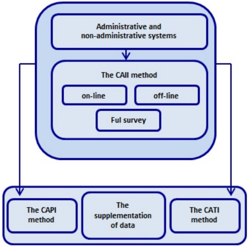
The NCPH 2011 Methodology

The NCPH 2011 Methodology

1. NCPH 2011 was carried out as a full-scale survey (administrative registers) and as a sample survey.
2. Poland used the mixed model of collecting data consisting of merging the data from administrative registers with the data obtained from direct statistical surveys.
3. Central Statistical Office in Poland decided to collect data using mixed approach because of the fact it was safer and more effective, taking into consideration the present level of development of administrative sources, their quality, and the degree of advancement of methodological work concerning the estimation and imputation of missing data in administrative sources.

The full-scale survey

- Data from administrative registers – the Master Record
- Data obtained with the CAII method
- Data supplemented with the CATI and CAPI methods

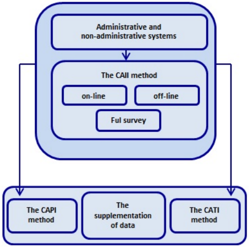


The full-scale survey

1. The full-scale survey involved population and housing, and was conducted with the use of administrative registers supplemented with a brief questionnaire to be filled in by each respondent.
2. For the first time in Poland 28 administrative sources were used in order to obtain the values of the census variables, both at the stage of creating a specification of census units (population and housing census) and for qualitative comparisons.
3. Due to a stable system of identifiers (PIN Personal Identification Number) it was possible to merge data from different registers.

The full-scale survey

- Data from administrative registers – the Master Record
- Data obtained with the CAII method
- Data supplemented with the CATI and CAPI methods

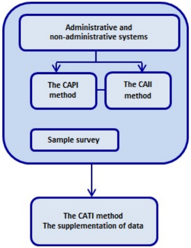

Administrative and non-administrative systems
The CAII method
on-line
off-line
Ful survey
The CAPI method
The supplementation of data
The CATI method

The full-scale survey

4. The supplementation of data was made using CATI (Computer Assisted Telephone Interview) and CAPI (Computer Assisted Personal Interviewing) methods.
5. They were used as supplementary channels, rather than the main channel for the acquisition of data. The basic method of obtaining data in the full-scale survey involved so called the „Master" record and the CAII method (Internet self-enumeration).
6. The Master record, being a set of variables derived from the registers, was the main channel supporting the collection of data, apart from Internet self-enumeration, phone interviews and direct interviews.

Sample survey

- Data obtained with CAII method
- Data obtained with the CAPI method
- Data supplemented with the CATI method (only if the survey requires a small supplementation)



Administrative and
non-administrative systems

The CAPI
method

The CAII
method

Sample survey

The CATI method
The supplementation of data

Sample survey

1. A sample survey is carried out on persons who permanently or temporarily reside in the territory of the Republic of Poland, and whose households have been sampled.
2. A sample survey was carried out using the CAII and CAPI methods. Data were supplemented with the CATI method.
3. A sample survey was carried out on a sample of 20% of dwellings and approximately 20% of population in Poland was drawn to the sample. Design weights associated with units drawn to the sample hade to be calibrated to known demographic totals from administrative registers.

Practical aspects of calibration in NCPH 2011

Practical aspects of calibration in NCPH 2011

1. Using data from many sources required on stage of generalization of results adjustment of initial weights assigned to all units drawn to a sample.
2. It was due to the fact that results from administrative registers and 20% sample should be consistent related to some basic demographic characteristic including gender, age and place of living.
3. In order to adjust design weights to reproduce known totals from administrative registers related to mentioned demographic characteristic calibration was used.
4. In the problem of finding calibration weights in NCPH 2011 G_1 function and macro CALMAR were used.

Practical aspects of calibration in NCPH 2011

Practical aspects of calibration in NCPH 2011

- In NCPH 2011 mixed approach of collecting data was used: administrative registers and survey sampling (20% of population).
- Some tables, especially related to demographic variables, were constructed using data from administrative registers (for example population in Poland in different cross-sections defined by sex, age and place of residence (urban areas, rural areas) in different territorial division from PESEL register.
- Many tables were created using data coming from the sample survey i.e. tables related to the level of education, labour market status etc.
- Design weights from the survey had to be calibrated because they did not reproduce known population totals from registers exactly.
- In NCPH 2011 design weights were calibrated in different cross-sections in different territorial division.

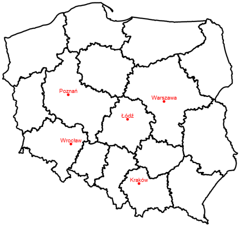
Practical aspects of calibration in NCPH 2011

- Voivodeships: sex $\times$ place of residence $\times$ individual years of age (0,1,. . . ,83,84,85+)
- Poviats: sex $\times$ place of residence $\times$ age groups (0–4,5–9,. . . ,80–84,85+)
- The biggest cities: sex $\times$ individual years of age (0,1,. . . ,83,84,85+ or 100+ for Warsaw)

Voivodeships (16) - NUTS 2

Poviats (379) - LAU 1

The biggest cities in Poland



Practical aspects of calibration in NCPH 2011

Practical aspects of calibration in NCPH 2011

- Auxiliary variables from registers taken into account in calibration process: sex, age and place of residence

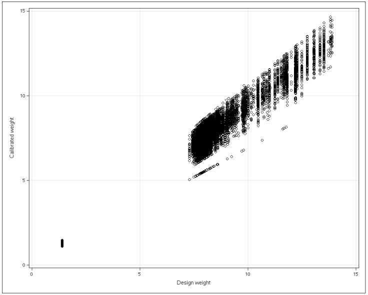
	Urban area/ Rural area	Sex	Age groups	Individual years of age	Individual years of age
	1,2	1,2	0-4, 5-9,..., 80-84, 85+	0, 1, . . . ,83, 84; 85+	0, 1, . . . ,98 99, 100+
Poland	1	1	1	1	0
Voivodeships	1	1	1	1	0
Poviats (without 5 biggest cities)	1	1	1	0	0
4 biggest cities	1	1	1	1	0
Warsaw	x	1	1	1	1
Districts of Warsaw	x	1	1	1	0
Districts of 4 biggest cities	x	1	1	1	0

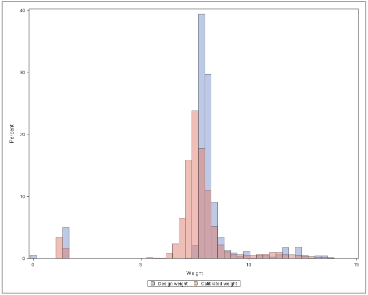
- **Legend:** 1–calibration possible, 0–calibration impossible, x–cross-section inadequate

Practical aspects of calibration in NCPH 2011

Poznanski poviat

Descriptive statistics					
Variable	Minimum	Maximum	Sum	Median	Std Dev
Design weights	1.3919308	13.8937500	350920.53	7.9896301	1.8675295
Calibrated weight	1.0884322	14.4946168	331525.00	7.5480397	1.8096110





Calibration in business statistics – assumption

Calibration in business statistics – assumption

- The simulation study investigated a few variables. The annual revenue was the response – output variable (Y).
- The list of auxiliary variables included: enterprise size (large and medium), selected PKD sections (construction, manufacturing, trade and transport) and VAT information.
- Data about the first two variables (enterprise size and PKD section) came from the DG-1 survey. The VAT variable came from the VAT register.
- To conduct the simulation study, a pseudo-population was created (further referred to as the MEETS real dataset), consisting of all enterprises included in the DG-1 survey for which information about the 3 auxiliary variables was available.
- The resulting dataset consisted of about 20,000 records containing complete information about the variables under analysis.
- Average revenue was estimated on the basis of samples of different size drawn from the MEETS real data. Simulation-based estimates were computed and evaluated at the country level, regardless of enterprise size and PKD section.

Calibration in business statistics – assumption

Calibration in business statistics – assumption

- During the simulation study, 5%, 10% and 15% samples were drawn from the MEETS real dataset, using simple random sampling without replacement.
- After obtaining a sample, information about revenue (dependent variable Y) for some enterprises was replaced with missing data.
- As a result, a given sample contained complete information about enterprise size, PKD section and VAT for each sampled unit, but incomplete data about revenue.
- 3 different approaches were used to generate missing data. In the first one missing data were generated in a random fashion (option 1). In the second (option 2) and third (option 3), missing data were attributed to enterprises with the lowest and highest revenue respectively.
- In addition, in each sample the percentages of missing data could be either 5%, 10% or 15%.
- For each sample fraction (3 options), fraction of missing data (3 options) and method of their generation (3 options) 500 iterations were performed to estimate the expected value of revenue, the expected value of the bias of the estimators and their empirical variance as well as relative estimation errors.

Calibration in business statistics – results

Calibration in business statistics – results

- The expected value of estimators of the average annual revenue for enterprises (in thousands of PLN)
- The average revenue calculated on the basis of the MEETS real data set was at the level of 45 500 (in thousand PLN).

		Horvitz-Thompson estimator			Calibration estimator		
sample size	% of missing data	Option 1	Option 2	Option 3	Option 1	Option 2	Option 3
5%	5%	46839	47388	16197	45555	44012	18718
	10%	45093	49955	11647	45411	42492	13542
	15%	45900	53392	9137	45758	40942	10684
10%	5%	46175	47290	16140	45801	44118	18264
	10%	45606	50843	11608	46079	42353	13218
	15%	45750	53303	9137	45458	40603	10502
15%	5%	45701	47862	16114	46113	44293	18078
	10%	45683	50761	11592	45802	42476	13085
	15%	45668	53254	9111	45920	40733	10404

Calibration in business statistics – results

Calibration in business statistics – results

- The expected value of the bias of estimators of the average annual revenue for enterprises (in thousands of PLN)

		Horvitz-Thompson estimator			Calibration estimator		
sample size	% of missing data	Option 1	Option 2	Option 3	Option 1	Option 2	Option 3
5%	5%	9516	8734	29353	4574	4488	26832
	10%	9222	9145	33903	4247	5225	32007
	15%	9414	10786	36413	4931	6111	34866
10%	5%	7093	6389	29410	3157	3353	27286
	10%	6442	7716	33942	3471	4302	32332
	15%	7435	8961	36413	3614	5396	35048
15%	5%	5391	5272	29436	2697	2664	27471
	10%	5860	6600	33958	2941	3592	32465
	15%	5627	8373	36439	2878	4994	35146

Calibration in business statistics – results

Calibration in business statistics – results

- The relative estimation error of estimators of the annual enterprise revenue (in percent)

		Horvitz-Thompson estimator			Calibration estimator		
sample size	% of missing data	Option 1	Option 2	Option 3	Option 1	Option 2	Option 3
5%	5%	27.77	26.16	7.26	12.33	12.36	11.84
	10%	26.99	24.48	6.88	11.81	12.85	10.36
	15%	28.15	26.67	6.00	13.49	13.72	8.53
10%	5%	19.74	18.13	5.28	8.54	8.90	7.17
	10%	17.97	19.58	4.59	9.35	9.31	5.95
	15%	20.77	18.23	4.31	9.84	9.58	5.44
15%	5%	14.85	13.46	3.98	7.30	7.09	5.27
	10%	15.93	13.81	3.52	7.94	7.12	4.38
	15%	15.69	14.62	3.37	7.87	7.63	4.07

Calibration in business statistics – conclusions

Calibration in business statistics – conclusions

- Given a certain sample size and a certain percentage of non-response, the estimators under analysis are likely to overestimate or underestimate the annual revenue of enterprises.
- Given a certain sample size and a certain non-response rate, the calibration estimator was characterized by lower bias regardless of the non-response generating scheme.
- The lowest bias for the estimators in question can be observed when non-response cases are random. Otherwise, bias generally increases, to a lesser degree, however, in the case of the calibration estimator.
- The calibration estimator is generally characterized by lower relative estimation error than the direct estimator. Relative estimation increases with the growing nonresponse rate, but to a lesser degree for the calibration estimator.
- The advantage of the calibration estimator is especially evident when the nonresponse generating mechanism is non-random. This is what often happens during surveys conducted by Central Statistical Office, where enterprises with the lowest and highest values of a given variable frequently refuse to report it.

Literature

Literature

- Särndal C-E., Lundström S. (2005), „*Estimation in Surveys with Nonresponse*", John Wiley & Sons, Ltd.
- Wallgren A., Wallgren B. (2007), „*Register-based Statistics: Administrative Data for Statistical Purposes*", Wiley.
- Deville J-C., Särndal C-E. (1992), „*Calibration Estimators in Survey Sampling*", Journal of the American Statistical Association, Vol. 87, 376–382.
- Särndal C-E. (2007), „*The Calibration Approach in Survey Theory and Practice*", Survey Methodology, Vol. 33, No. 2, 99–119.

Thank you very much for your attention!