On direct and indirect median estimators in survey sampling

Tomasz Stachurski

University of Economics in Katowice
Faculty of Management
Department of Statistics, Econometrics and Mathematics

10-12th of July 2018

The importance of median estimation

Median is a measure that has great importance in economic research, especially in measuring poverty:

- since 80's median is used by the European Commission (Commission of the European Communities) in comparative studies among European countries;

The importance of median estimation

Median is a measure that has great importance in economic research, especially in measuring poverty:

- since 80's median is used by the European Commission (Commission of the European Communities) in comparative studies among European countries;
- Eurostat recommends to set a poverty line as a 60% of the national median equivalised disposable income;

The importance of median estimation

Median is a measure that has great importance in economic research, especially in measuring poverty:

- since 80's median is used by the European Commission (Commission of the European Communities) in comparative studies among European countries;
- Eurostat recommends to set a poverty line as a 60% of the national median equivalised disposable income;
- Currently percentage of median equivalised disposable income in EU-SILC methodology is used to estimate the number of people being at the risk of poverty.

Main aims of the presentation

- Presenting of the proposition of a ratio synthetic estimator of a domain median.

Main aims of the presentation

- Presenting of the proposition of a ratio synthetic estimator of a domain median.
- Presenting of the proposition of the sampling design using a distance measure between values of an auxiliary variable and its median.

Main aims of the presentation

- Presenting of the proposition of a ratio synthetic estimator of a domain median.
- Presenting of the proposition of the sampling design using a distance measure between values of an auxiliary variable and its median.
- The simulation-based analysis of the properties of the considered median estimation strategies.

Definition of a median

Let $y = [y_1, y_2, \ldots, y_N]$ to be a vector of N observations of the studied variable in the population. Then, the **population median** can be defined as follows (Wywiał 2010, p. 13):

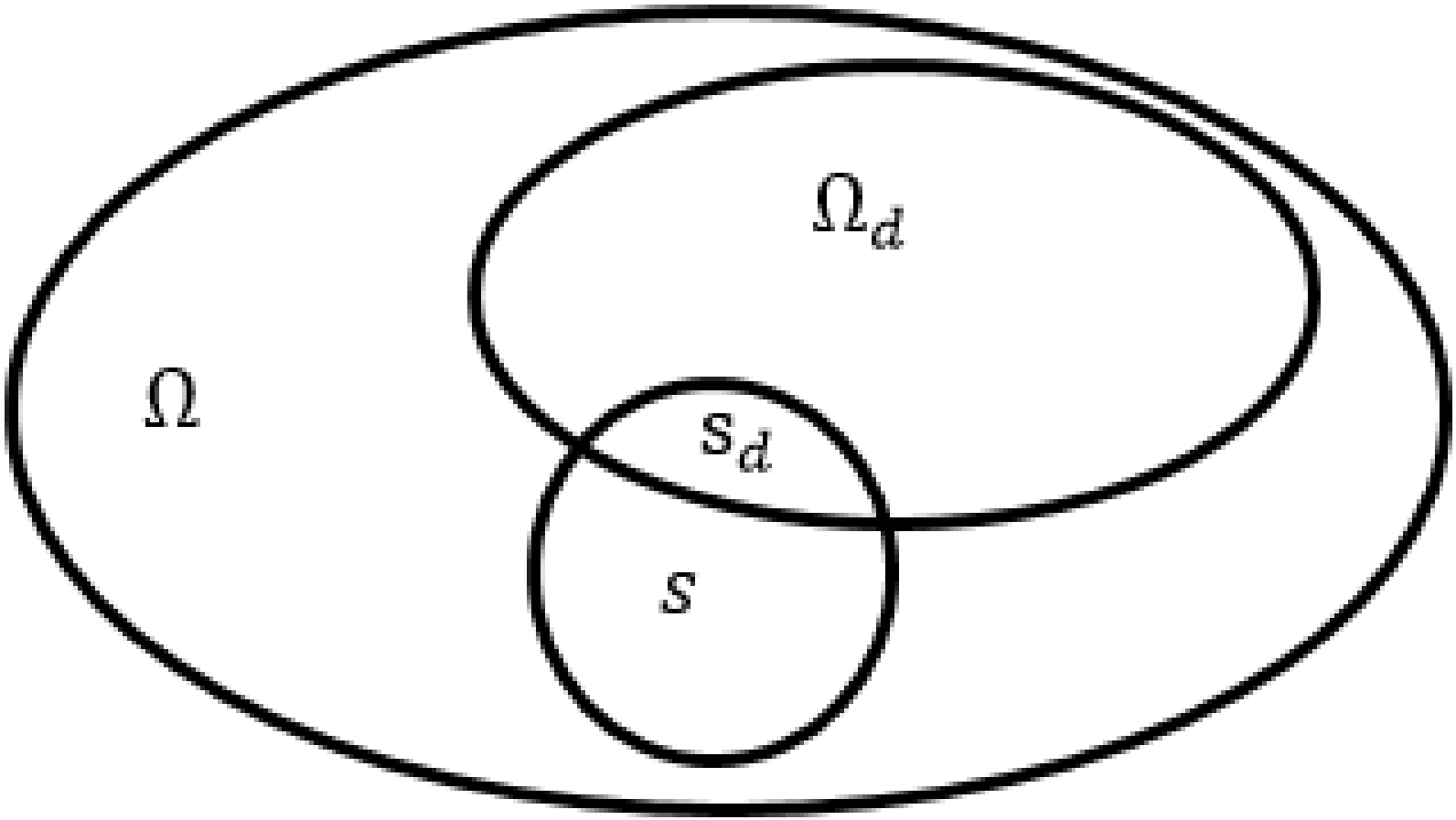
$$Me = \begin{cases} \frac{y_N + y_{N+1}}{2} & \textit{if } N \textit{ is even} \\ y_{\frac{N+1}{2}} & \textit{if } N \textit{ is odd} \end{cases}. \tag{1}$$

Median (Me) divides the population into to equal parts, so that at least half of the population elements have values smaller or equal Me, whereas at least a half have values greater or equal Me.

Basic notations

- Ω - the population of size N,
- Ω_d - the dth subpopulation (domain) of size N_d, where $d = 1, ..., D$,
- s - the sample of size n,
- s_d - the sample in the dth domain of size n_d,
- y_i - value of the study variable for the ith element of the population,
- x_i - value of the auxiliary variable for the ith element of the population.

Small area estimation



Basic definitions

- Sampling design
- First-order incusion probability
- Estimation strategy

Sampling design

Sampling design is a function:

$$p(s) : S \rightarrow [0, 1] \tag{2}$$

which assigns a probability of drawing to each sample. This function has to fulfil two conditions:

- probability of drawing each sample must be non-negative.

Sampling design

Sampling design is a function:

$$p(s) : S \rightarrow [0, 1] \tag{2}$$

which assigns a probability of drawing to each sample. This function has to fulfil two conditions:

- probability of drawing each sample must be non-negative.
- sum of probabilities for all samples s from sample space is equal to 1.

First-order inclusion probability

First-order inclusion probability is a probability that unit labeled by i belongs to the sample:

$$\pi_i = \sum_{\{s: i \in s\}} p(s) \tag{3}$$

The r-order inclusion probability is a probability that units denoted by $i_1, i_2, \ldots i_r$ belong to the sample:

$$\pi_{i_1, i_2, \ldots, i_r} = \sum_{\{s: i_1, i_2, \ldots, i_r \in s\}} p(s) \tag{4}$$

Estimation strategy

Estimation strategy for the parameter θ is a pair $(t_S, p(s))$, where t_S is an estimator of the θ and $p(s)$ is a sampling design.

Procedure of the median estimation

Let consider following procedure of the median estimation (Särndal, Swensson and Wretman (1992)):

- Let arrange sampled elements in non-decreasing order:

$$y_1 \leqslant y_2 \leqslant \ldots \leqslant y_n$$

and the corresponding inclusion probabilities π_i:

$$\pi_1, \pi_2, \ldots, \pi_n$$

.

Procedure of the median estimation

- Let calculate cumulative sums:

$$B_0 = 0$$

$$B_1 = \frac{1}{\pi_1}$$

$$B_2 = \frac{1}{\pi_1} + \frac{1}{\pi_2}$$

Procedure of the median estimation

- Let calculate cumulative sums:

$$B_0 = 0$$

$$B_1 = \frac{1}{\pi_1}$$

$$B_2 = \frac{1}{\pi_1} + \frac{1}{\pi_2}$$

- and so on; in general, for i=1,...n:

$$B_n = \sum_{i=1}^{n} \frac{1}{\pi_i} \tag{5}$$

Procedure of the median estimation

- Then, the median estimator can be obtained using (Särndal, Swensson and Wretman (1992), p. 200):

$$\hat{M} = \begin{cases} y_i & \text{if } B_{i-1} < 0,5\hat{N} < B_i \\ 0,5(y_i + y_{i+1}) & \text{if } B_i = 0,5\hat{N} \end{cases} \tag{6}$$

where $\hat{N} = \sum_{i=1}^{n} \frac{1}{\pi_i}$ is an estimator of N.

The ratio estimator of a domain median

A modification of a widely known ratio estimator of the mean or the total is **the ratio estimator for the median**, which was firstly proposed for simple random sampling by Kuk and Mak (1989) p. 262.

A simple modification of the estimator that is applicable for any sampling design is presented by Thompson and Godambe (2010) p. 86. **The ratio estimator of domain median** is given by the following formula:

$$\hat{M}^R_{\Omega_d}(y) = \frac{\hat{M}_{\Omega_d}(y)}{\hat{M}_{\Omega_d}(x)} M_{\Omega_d}(x) \tag{7}$$

where $\hat{M}^R_{\Omega_d}(y)$, $\hat{M}^R_{\Omega_d}(x)$ are obtained using (6) and $M_{\Omega_d}(x)$ is the known median of auxiliary variable X in dth domain.

The ratio estimator of a domain median

Remarks:

- If in (7) we assume that $\Omega_d = \Omega$, then we obtain the estimator of the population median.
- The ratio median estimator given by (7) is a direct estimator. It implies that it cannot be used for a domain where sample size is zero.

The synthetic ratio estimator of a domain median

We propose ratio synthetic estimator of a domain median given by formula:

$$\hat{M}_{\Omega_d}^{SR}(y) = \frac{M_{\Omega_d}(x)}{\hat{M}_{\Omega}(x)} \hat{M}_{\Omega}(y) \tag{8}$$

where $\hat{M}_{\Omega}(y)$, $\hat{M}_{\Omega}(x)$ are obtained using (6) and $M_{\Omega_d}(x)$ is the known median of auxiliary variable X in dth domain.

The synthetic ratio estimator of a domain median

Remarks

- It can be profitable to use indirect instead of direct domain estimators, because usually variances of direct estimators in domains (especially when domain sample size is small) are much more than variances of indirect estimators.
- Synthetic ratio estimator given by (8) is indirect estimator and it can be used when domain sample size is zero.

Considered estimation strategies

For each of the following estimators:

- median estimator (Särndal, Swensson and Wretman (1992), p.200) given by (6);
- ratio estimator of a domain median (Thompson and Godambe (2010) p.86) given by (7);
- proposition of the synthetic ratio domain median estimator given by (8);

there were considered three different sampling designs:

- **simple random sampling**,

Considered estimation strategies

For each of the following estimators:

- median estimator (Särndal, Swensson and Wretman (1992), p.200) given by (6);
- ratio estimator of a domain median (Thompson and Godambe (2010) p.86) given by (7);
- proposition of the synthetic ratio domain median estimator given by (8);

there were considered three different sampling designs:

- **simple random sampling**,
- sampling design **proportional to an auxiliary variable**,

Considered estimation strategies

For each of the following estimators:

- median estimator (Särndal, Swensson and Wretman (1992), p.200) given by (6);
- ratio estimator of a domain median (Thompson and Godambe (2010) p.86) given by (7);
- proposition of the synthetic ratio domain median estimator given by (8);

there were considered three different sampling designs:

- **simple random sampling**,
- sampling design **proportional to an auxiliary variable**,
- sampling design **proportional to a distance measure** between values of auxiliary variable and its median

Sampling design proportional to an auxiliary variable

The first order inclusion probabilities can be obtained using formula (Berger and Tillé (2009) p. 40):

$$\pi_i = \frac{n x_i}{\sum_{i=1}^{N} x_i} \tag{9}$$

- If, for some population elements, quantities are larger than 1, then we assign one to the value of first order inclusion probability for such elements.

Sampling design proportional to an auxiliary variable

The first order inclusion probabilities can be obtained using formula (Berger and Tillé (2009) p. 40):

$$\pi_i = \frac{n x_i}{\sum_{i=1}^{N} x_i} \tag{9}$$

- If, for some population elements, quantities are larger than 1, then we assign one to the value of first order inclusion probability for such elements.
- Then, remaining quantities are recalculated using (9).

Sampling design proportional to the distance measure

If the variable under study is highly correlated with the auxiliary variable, then it may be assumed that units which values of the auxiliary variable are close to its median, it ought to be similar for the variable of interest. That is why, we propose to use sampling proportional to a distance measure between values of auxiliary variable and its median as follows:

$$z_i = \frac{1}{|x_i - M_\Omega(x)| + c} \tag{10}$$

where c is a positive constant.

Then the first order inclusion probabilities can be obtained as follows:

$$\pi_i = \frac{n z_i}{\sum_{i=1}^{N} z_i} \tag{11}$$

Estimation the variance of the median estimator

In order to estimate the variance of the median estimator we use **the bootstrap method**.

Idea of the bootstrap method

The bootstrap method was proposed by Efron in 1979. It is a resampling method. The idea of this method is based on mutliple drawing with replacement from the original sample.

Area of applications the bootstrap method:

- estimation of confidence intervals,

Idea of the bootstrap method

The bootstrap method was proposed by Efron in 1979. It is a resampling method. The idea of this method is based on mutliple drawing with replacement from the original sample.

Area of applications the bootstrap method:

- estimation of confidence intervals,
- hypothesis testing,

Idea of the bootstrap method

The bootstrap method was proposed by Efron in 1979. It is a resampling method. The idea of this method is based on mutliple drawing with replacement from the original sample.

Area of applications the bootstrap method:

- estimation of confidence intervals,
- hypothesis testing,
- estimation the variance of estimators.

The bootstrap algorithm

Algorithm of bootstrap method consists of a few steps:

- Draw B times sample with replacement from the original sample.
- For each of B samples calculate the parameter of interest ${\theta_b}^*$ in the same way as for the original sample.
- The bootstrap estimator is the mean of estimators from B generated samples based on original one as follows:

$$\hat{\theta}^B = \frac{1}{B} \sum_{b=1}^{B} {\theta_b}^* \tag{12}$$

- The observed distribution of ${\theta_1}^*, ..., {\theta_B}^*$ can be treated as an estimation of sampling distribution of estimator and its variance can be obtained using formula (13).

The bootstrap algorithm

The bootstrap estimator of variance is given by the following formula:

$$\hat{D}^2\left(\hat{\theta}^B\right) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\theta_b{}^* - \hat{\theta}^B\right)^2 \tag{13}$$

The bootstrap algorithm for complex sampling designs

Modification of bootstrap technique that can be applied for survey sampling is presented in (Särndal, Swensson and Wretman (1992), pp.442-444).

- Construct artificial population U^*, where U^* is composed of replicates of the elements in the original sample.
- For each element in the sample we create $\frac{1}{\pi_i}$ artificial elements in U^*.
- Draw B times sample with replacement from the original sample with probabilities $p_i = \frac{{\pi_i}^*}{n}$.
- For each of B samples calculate the parameter of interest ${\theta_b}^*$ in the same way as for the original sample.
- The bootstrap estimator is the mean of estimators from B generated samples based on original one as (12) and the bootstrap variance estimator is obtained using (13).

Other bootstrap algorithms

Another bootstrap methods to estimate the variance estimator are studied by Antal and Tillé (2011, 2014) including:

- Bootstrap for unequal probability sampling without replacement using a Poisson random variable to select units from the original sample,
- Bootstrap methods using a mixture of several sampling designs.

Simulation study

In this section we presents results of the simulation study, which was conducted in R.
The aim of the study was analysis of the properties of considered median estimators under three sampling designs:

- sampling proportional to size (PPS);
- sampling proportional to a distance measure (PPD);
- simple random sampling without replacement (SWOR).

The simulation study was conducted both for real and generated data.

Simulation study

In the design-based simulation study we compute:

- the relative biases of the median estimator (in %),

Simulation study

In the design-based simulation study we compute:

- the relative biases of the median estimator (in %),
- the relative RMSE of the median estimator (in %),

Simulation study

In the design-based simulation study we compute:

- the relative biases of the median estimator (in %),
- the relative RMSE of the median estimator (in %),
- the relative biases of estimators of the design-variances of the median estimator (in %).

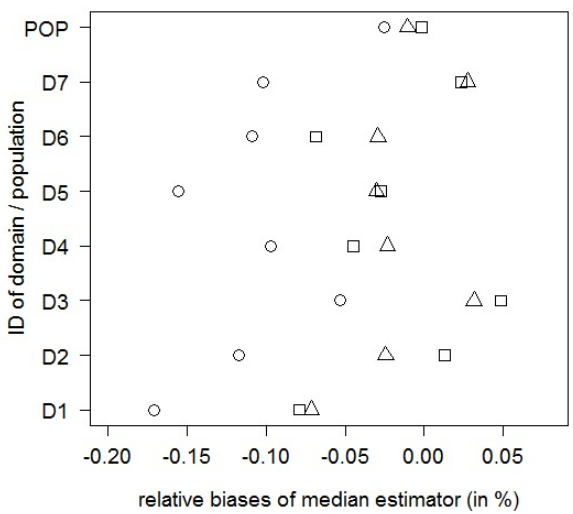
Simulation study - generated data

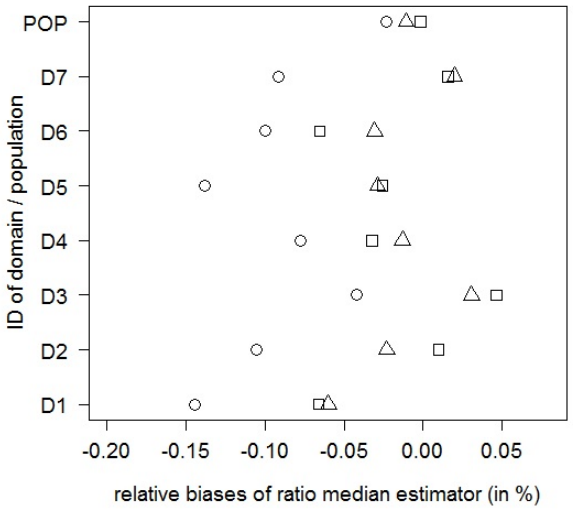
- Considered dataset consists of 281 observations and was divided into seven subpopulations.
- The sample size is equal to n = 28, which represents about 10% of the population.

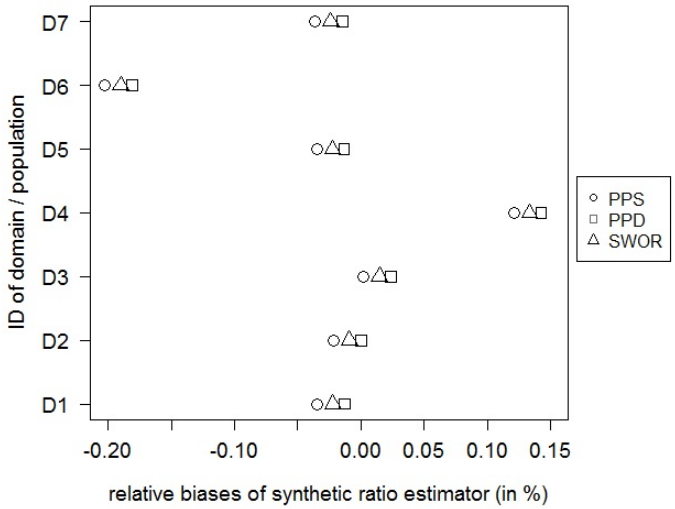
Simulation study - real data

- Considered dataset consists of 281 Swedish municipalities
- The sample size is equal to n = 28, which represents about 10% of the population.
- The population was divided into 7 regions.
- The study variable is revenue from taxes and the auxiliary variable is the number of municipal employees in 1984

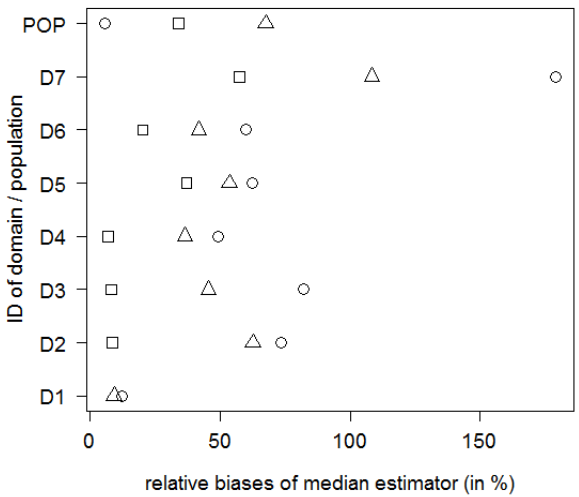
Relative biases of median estimators for the generated data

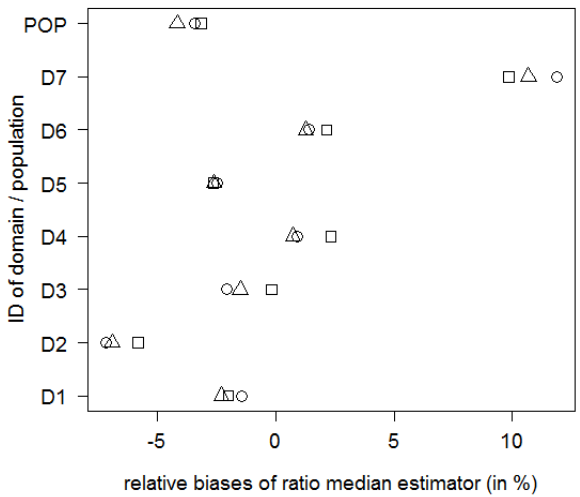


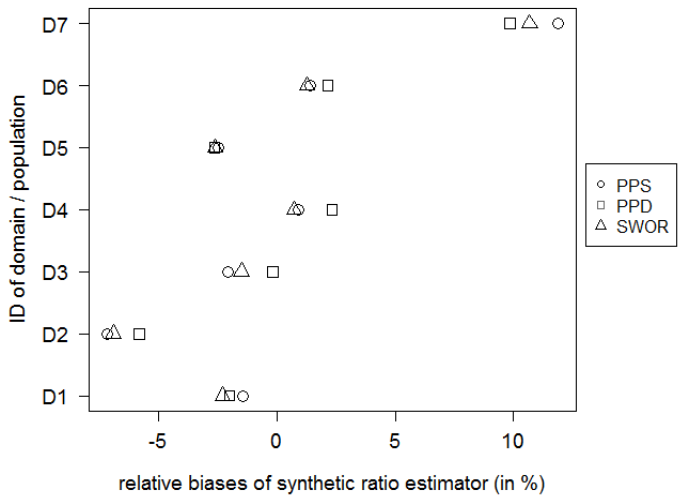




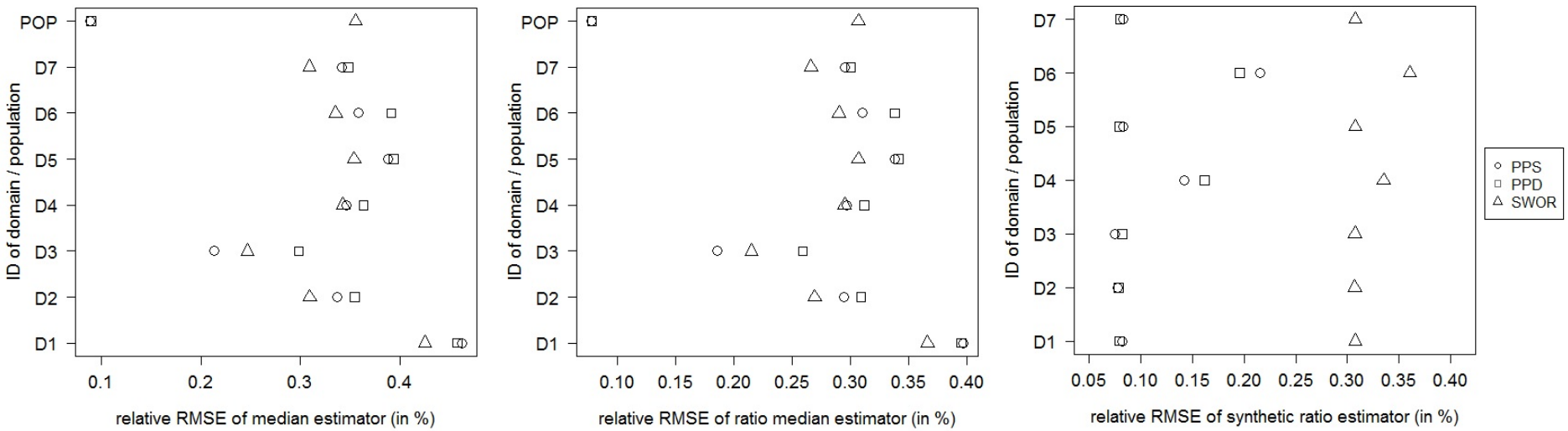
Relative biases of median estimators for the real data



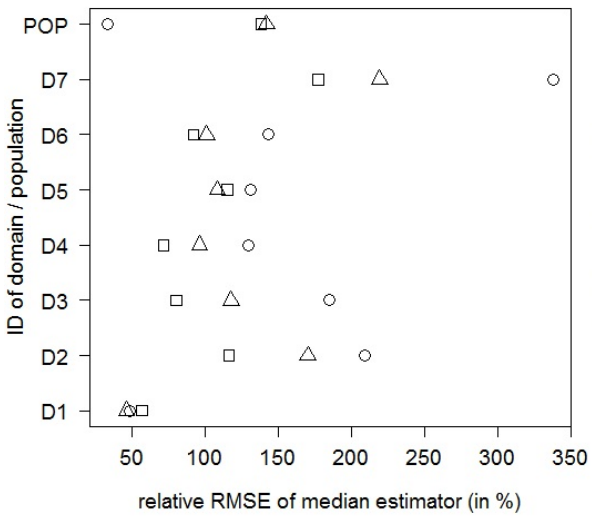


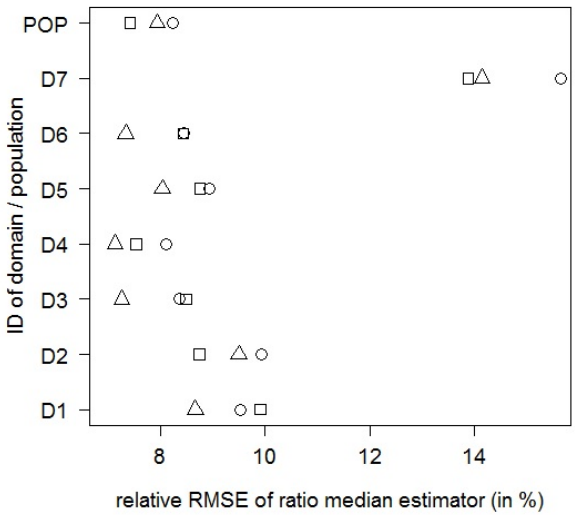


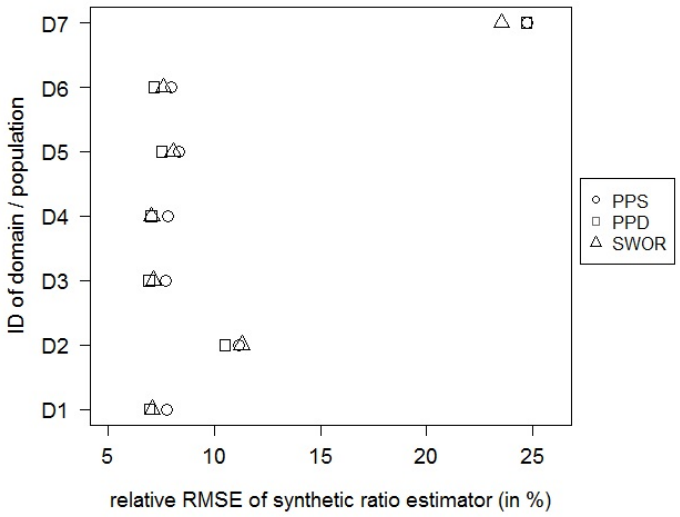
Relative RMSE of median estimators for the generated data



Relative RMSE of median estimators for the real data







Relative biases of design-variance of the ratio median estimator (bootstrap for independent variables)

ID of subpopulation/population	Relative biases
1	-37,27%
2	64.03%
3	83.77%
4	-50.09%
5	-47.99%
6	-62.48%
7	113.6%
The population	-94.45%

Relative biases of design-variance of the ratio median estimator (bootstrap for complex sampling design)

ID of subpopulation/population	Relative biases
1	-65.56%
2	222.7%
3	-40.85%
4	53.94%
5	-71.11%
6	-31.05%
7	-33.45%
The population	-83.16%

Results of another simulation studies

Antal and Tillé (2014) compared properties of some boostrap techniques in the case of median estimator.

	Relative bias	RRMSE
Bootstrap I (2011)	-6.5615%	50.6651%
Boostrap II (2014)	1.0564%	58.8889%

Conclusion

- For the asymmetric variables using ratio median estimators including the proposed synthetic ratio estimators leads to sustainable decrease of relative biases and relative root mean square error comparing to the estimator that does not use auxiliary information.
- The proposed sampling design based on a distance between values of the auxiliary variable and its median can be profitable especially in the case of the estimator which does not use auxiliary information.
- Bootstrap technique for complex sampling design in order to estimate the variance of median estimators is less biased than the bootstrap for independent variables.

References

- Antal E. and Tillé Y. (2011) A Direct Bootstrap Method for Complex Sampling, *Journal of the American Statistical Association*, 106, pp. 534-543.
- Antal E. and Tillé Y. (2014) A new resampling method for sampling designs without, *Computational Statistics* 29, issue 5, pp. 1345-1363 replacement: the doubled half bootstrap
- Berger Y. and Tillé Y. (2009) *Sampling with Unequal Probabilities*, pp. 39-54, [in:] Handbook of Statistics 29.
- Domański Cz., and Pruska K.: *Metody statystyki małych obszarów.* Wydawnictwo Uniwersytetu Łódzkiego, Łódź, 2001.
- Efron B. (1979) Bootstrap Methods: Another Look at the Jacknife, *Annals of Statistics*, nr 7, pp. 1-26.
- Kuk Y.C.A., and Mak T.K.: Median estimation in the Presence of Auxiliary Information. Jounral of the Royal Statistical Society. Series B (Methodological), Vol. 51, No. 2 (1989), 261-269.

References

- Rao J.N.K., and Molina I.: *Small Area Estimation*. John Wiley and Sons, New York, 2015.
- Särndal, C.E., Swensson, B. and Wretman J. (1992), *Model Assisted Survey Sampling*, Springer-Verlag, New York.
- Thompson M.E., and Godambe V.P.: Estimating Functions and Survey Sampling. In: *Handbook of Statistics. Sample Surveys: Inference and Analysis*, Vol. 29B (Pfeffermann D., and Rao C.R., eds.), Elsevier Sci-ence, Oxford, 2009.
- Wywiał J. (2010): *Wprowadzenie do metody reprezentacyjnej*. Wydawnictwo Uniwersytetu Ekonomicznego w Katowi-cach, Katowice, 2010.

Thank you for your attention!