On testing hypotheses in case of small sample sizes

Testing equality of means: design-based and model-based approach

Tomasz Żądło

Department of Statistics, Econometrics and Mathematics

University of Economics in Katowice

Introduction

Using classic design-based approach in survey sampling it is possible to:

- estimate population parameters (e.g. population mean),
- estimate the accuracy of estimation,
- and to test population parameters (asymptotic distributions of some tests statistics are normal).

Introduction

Using small area estimation methods (including design-based approach) it is possible to:

- estimate subpopulation parameters (e.g. subpopulation mean),
- and to estimate the accuracy of estimation

even for small or zero sample sizes in subpopulations.

We propose some tests of equality of subpopulation means which can be used even in case of small sample sizes.

Introduction

Using small area estimation methods (including design-based approach) it is possible to:

- estimate subpopulation parameters (e.g. subpopulation mean),
- and to estimate the accuracy of estimation

even for small or zero sample sizes in subpopulations.

We propose some tests of equality of subpopulation means which can be used even in case of small sample sizes.

Introduction

DIRECT estimator

To estimate a subpopulation mean, y's only from the subpopulation of interest are used.

INDIRECT estimator

To estimate a subpopulation mean, y's from different subpopulations are used.

Notations

- Ω - the population of size N
- Ω_d - the dth subpopulation of size N_d, where $d = 1, ..., D$
- $P(s)$ - sampling design, π_i - first order inclusion probabilities,
- s - the sample of size n
- s_d - the sample in the dth domain of size n_d
- set of nonsampled subpopulation elements $\Omega_{rd} = \Omega_d - s_d$ of size N_{rd}
- θ_d - the dth subpopulation mean
- $\hat{\theta}_d$ - an estimator/predictor of θ_d

Notations

- Ω - the population of size N
- Ω_d - the dth subpopulation of size N_d, where $d = 1, ..., D$
- $P(s)$ - sampling design, π_i - first order inclusion probabilities,
- s - the sample of size n
- s_d - the sample in the dth domain of size n_d
- set of nonsampled subpopulation elements $\Omega_{rd} = \Omega_d - s_d$ of size N_{rd}
- θ_d - the dth subpopulation mean
- $\hat{\theta}_d$ - an estimator/predictor of θ_d

Notations

- Ω - the population of size N
- Ω_d - the dth subpopulation of size N_d, where $d = 1, ..., D$
- $P(s)$ - sampling design, π_i - first order inclusion probabilities,
- s - the sample of size n
- s_d - the sample in the dth domain of size n_d
- set of nonsampled subpopulation elements $\Omega_{rd} = \Omega_d - s_d$ of size N_{rd}
- θ_d - the dth subpopulation mean
- $\hat{\theta}_d$ - an estimator/predictor of θ_d

Notations

- Ω - the population of size N
- Ω_d - the dth subpopulation of size N_d, where $d = 1, ..., D$
- $P(s)$ - sampling design, π_i - first order inclusion probabilities,
- s - the sample of size n
- s_d - the sample in the dth domain of size n_d
- set of nonsampled subpopulation elements $\Omega_{rd} = \Omega_d - s_d$ of size N_{rd}
- θ_d - the dth subpopulation mean
- $\hat{\theta}_d$ - an estimator/predictor of θ_d

Notations

- Ω - the population of size N
- Ω_d - the dth subpopulation of size N_d, where $d = 1, ..., D$
- $P(s)$ - sampling design, π_i - first order inclusion probabilities,
- s - the sample of size n
- s_d - the sample in the dth domain of size n_d
- set of nonsampled subpopulation elements $\Omega_{rd} = \Omega_d - s_d$ of size N_{rd}
- θ_d - the dth subpopulation mean
- $\hat{\theta}_d$ - an estimator/predictor of θ_d

Notations

- Ω - the population of size N
- Ω_d - the dth subpopulation of size N_d, where $d = 1, ..., D$
- $P(s)$ - sampling design, π_i - first order inclusion probabilities,
- s - the sample of size n
- s_d - the sample in the dth domain of size n_d
- set of nonsampled subpopulation elements $\Omega_{rd} = \Omega_d - s_d$ of size N_{rd}
- θ_d - the dth subpopulation mean
- $\hat{\theta}_d$ - an estimator/predictor of θ_d

Notations

- Ω - the population of size N
- Ω_d - the dth subpopulation of size N_d, where $d = 1, ..., D$
- $P(s)$ - sampling design, π_i - first order inclusion probabilities,
- s - the sample of size n
- s_d - the sample in the dth domain of size n_d
- set of nonsampled subpopulation elements $\Omega_{rd} = \Omega_d - s_d$ of size N_{rd}
- θ_d - the dth subpopulation mean
- $\hat{\theta}_d$ - an estimator/predictor of θ_d

Notations

- Ω - the population of size N
- Ω_d - the dth subpopulation of size N_d, where $d = 1, ..., D$
- $P(s)$ - sampling design, π_i - first order inclusion probabilities,
- s - the sample of size n
- s_d - the sample in the dth domain of size n_d
- set of nonsampled subpopulation elements $\Omega_{rd} = \Omega_d - s_d$ of size N_{rd}
- θ_d - the dth subpopulation mean
- $\hat{\theta}_d$ - an estimator/predictor of θ_d

Notations

We consider two types of subpopulations:

- strata - with fixed sample sizes,
- and domains - with random sample sizes.

Small area

is a subpopulation where sample size is too small to use direct methods

Notations

We consider two types of subpopulations:

- strata - with fixed sample sizes,
- and domains - with random sample sizes.

Small area

is a subpopulation where sample size is too small to use direct methods

Notations

We consider two types of subpopulations:

- strata - with fixed sample sizes,
- and domains - with random sample sizes.

Small area

is a subpopulation where sample size is too small to use direct methods

Tests in survey sampling

For any sampling design we test: $H_0 : \theta = \theta_0$ where $\theta = \frac{1}{N} \sum\limits_{i \in \Omega} y_i$

Test statistic:

$$\frac{\hat{\theta}^{HT} - \theta_0}{\hat{D}(\hat{\theta}^{HT})}$$

where $\hat{\theta}^{HT} = \frac{1}{N} \sum\limits_{i \in s} y_i \pi_i^{-1}$ is Horvitz-Thompson unbiased estimator of θ, θ_0 is tested value of the population mean

Tests in survey sampling

To assume that the test statistic has an approximate standard normal distribution, we should verify (Särndal, Swensson and Wretman (1992) p. 56), that:

- under null hypothesis the estimator is approximately normally distributed with mean θ_0 and variance $D^2(\hat{\theta})$,
- $\hat{D}(\hat{\theta})$ is a consistent estimator of $D^2(\hat{\theta})$.

Tests in survey sampling

To assume that the test statistic has an approximate standard normal distribution, we should verify (Särndal, Swensson and Wretman (1992) p. 56), that:

- under null hypothesis the estimator is approximately normally distributed with mean θ_0 and variance $D^2(\hat{\theta})$,
- $\hat{D}(\hat{\theta})$ is a consistent estimator of $D^2(\hat{\theta})$.

Tests in survey sampling

Small area estimation perspective:

- Although Berger (1998) proved that the asymptotic distribution of the Horvitz-Thompson estimator for any sampling design is normal but as shown by Wywiał (2017) the required sample sizes may be very large especially in case of complex sampling designs without replacement (for cases considered in Wywiał (2017) they were between 360 and 1330). Such large sample sizes are unusual for small area estimation problems.
- The second condition and the assumption of design-unbiasedness of the estimator are usually not met for small area estimators.

Tests in survey sampling

Small area estimation perspective:

- Although Berger (1998) proved that the asymptotic distribution of the Horvitz-Thompson estimator for any sampling design is normal but as shown by Wywiał (2017) the required sample sizes may be very large especially in case of complex sampling designs without replacement (for cases considered in Wywiał (2017) they were between 360 and 1330). Such large sample sizes are unusual for small area estimation problems.
- The second condition and the assumption of design-unbiasedness of the estimator are usually not met for small area estimators.

Permutation tests for equality of subpopulations means

Based on the design-based approach we consider the following hypotheses:

$$H_0 : \theta_k = \theta_l$$

$$H_1 : \theta_k \neq \theta_l$$

where $\theta_k = N_k^{-1} \sum_{i \in \Omega_k} y_i$ and $\theta_l = N_l^{-1} \sum_{i \in \Omega_l} y_i$.

Using the model-based approach we consider the following hypotheses:

$$H_0 : E_\xi(\theta_k) = E_\xi(\theta_l)$$

$$H_1 : E_\xi(\theta_k) \neq E_\xi(\theta_l)$$

where $\theta_k = N_k^{-1} \sum_{i \in \Omega_k} Y_i$ and $\theta_l = N_l^{-1} \sum_{i \in \Omega_l} Y_i$.

Permutation tests for equality of subpopulations means

The following test statistic of a permutation test is used:

$$T = |\hat{\theta}_k - \hat{\theta}_l|$$

where $\hat{\theta}_k$ and $\hat{\theta}_l$ are estimators or predictors of subpopulation means.

Permutation tests for equality of subpopulations means

Let the sample data be denoted by the matrix **Z** and the vector of subpopulation labels for both samples by **u**. The procedure is as follows (Pesarin and Salmaso 2010, p. 45):

Permutation tests for equality of subpopulations means

1. Based on the original sample data we compute the value of the test statistics $T_0 = T(\mathbf{Z}, \mathbf{u})$
2. We permute elements of vector $\mathbf{u}$ and obtain vector $\mathbf{u}^*$.
3. We compute the value of the tests statistic $T^* = T(\mathbf{Z}, \mathbf{u}^*)$
4. We repeat steps 2 and 3 B-times to obtain $T^{*b} = T(\mathbf{Z}, \mathbf{u}^{*b})$, where $b = 1, 2, ..., B$
5. We estimate p-value as $B^{-1} \sum_{1 \leqslant b \leqslant B} I(T^{*b} \geqslant T_0)$,

Permutation tests for equality of subpopulations means

1. Based on the original sample data we compute the value of the test statistics $T_0 = T(\mathbf{Z}, \mathbf{u})$
2. We permute elements of vector $\mathbf{u}$ and obtain vector $\mathbf{u}^*$.
3. We compute the value of the tests statistic $T^* = T(\mathbf{Z}, \mathbf{u}^*)$
4. We repeat steps 2 and 3 B-times to obtain $T^{*b} = T(\mathbf{Z}, \mathbf{u}^{*b})$, where $b = 1, 2, ..., B$
5. We estimate p-value as $B^{-1} \sum_{1 \leqslant b \leqslant B} I(T^{*b} \geqslant T_0)$,

Permutation tests for equality of subpopulations means

1. Based on the original sample data we compute the value of the test statistics $T_0 = T(\mathbf{Z}, \mathbf{u})$
2. We permute elements of vector $\mathbf{u}$ and obtain vector $\mathbf{u}^*$.
3. We compute the value of the tests statistic $T^* = T(\mathbf{Z}, \mathbf{u}^*)$
4. We repeat steps 2 and 3 B-times to obtain $T^{*b} = T(\mathbf{Z}, \mathbf{u}^{*b})$, where $b = 1, 2, ..., B$
5. We estimate p-value as $B^{-1} \sum_{1 \leqslant b \leqslant B} I(T^{*b} \geqslant T_0)$,

Permutation tests for equality of subpopulations means

1. Based on the original sample data we compute the value of the test statistics $T_0 = T(\mathbf{Z}, \mathbf{u})$
2. We permute elements of vector $\mathbf{u}$ and obtain vector $\mathbf{u}^*$.
3. We compute the value of the tests statistic $T^* = T(\mathbf{Z}, \mathbf{u}^*)$
4. We repeat steps 2 and 3 B-times to obtain $T^{*b} = T(\mathbf{Z}, \mathbf{u}^{*b})$, where $b = 1, 2, ..., B$
5. We estimate p-value as $B^{-1} \sum_{1 \leqslant b \leqslant B} I(T^{*b} \geqslant T_0)$,

Permutation tests for equality of subpopulations means

1. Based on the original sample data we compute the value of the test statistics $T_0 = T(\mathbf{Z}, \mathbf{u})$
2. We permute elements of vector $\mathbf{u}$ and obtain vector $\mathbf{u}^*$.
3. We compute the value of the tests statistic $T^* = T(\mathbf{Z}, \mathbf{u}^*)$
4. We repeat steps 2 and 3 B-times to obtain $T^{*b} = T(\mathbf{Z}, \mathbf{u}^{*b})$, where $b = 1, 2, ..., B$
5. We estimate p-value as $B^{-1} \sum_{1 \leqslant b \leqslant B} I(T^{*b} \geqslant T_0)$,

Permutation tests for equality of subpopulations means

Any assumptions?

Exchangeability under null hypothesis - data can be permuted (exchanged) without affecting its joint distribution.

Permutation tests for equality of subpopulations means

Any assumptions?

Exchangeability under null hypothesis - data can be permuted (exchanged) without affecting its joint distribution.

Some remarks

Values of the test statistic based on direct estimators can be computed only if at least one observation from each out of two subpopulations is available. These test statistics will be used for testing the equality of means only in strata, where the sample sizes are fixed.

Some remarks

Test statistics based on indirect estimators can be computed even in case of zero sample sizes in subpopulations but then the permutation of the sample data does not change the values of these test statistics and p-values in such cases equal 1. Hence, p-values for these tests may tend to be too large what means that they may allow to claim that the alternative hypothesis is true less often than it should be claimed.

These types of tests are called conservative tests. Their type I error probabilities are less than the assumed significance level.

Proposal

Because of this reason for the tests based on indirect estimators we propose to additionally permute values of the auxiliary variable for nonsampled population elements.

Some remarks

Test statistics based on indirect estimators can be computed even in case of zero sample sizes in subpopulations but then the permutation of the sample data does not change the values of these test statistics and p-values in such cases equal 1. Hence, p-values for these tests may tend to be too large what means that they may allow to claim that the alternative hypothesis is true less often than it should be claimed.

These types of tests are called conservative tests. Their type I error probabilities are less than the assumed significance level.

Proposal

Because of this reason for the tests based on indirect estimators we propose to additionally permute values of the auxiliary variable for nonsampled population elements.

Some remarks

Test statistics based on indirect estimators can be computed even in case of zero sample sizes in subpopulations but then the permutation of the sample data does not change the values of these test statistics and p-values in such cases equal 1. Hence, p-values for these tests may tend to be too large what means that they may allow to claim that the alternative hypothesis is true less often than it should be claimed.

These types of tests are called conservative tests. Their type I error probabilities are less than the assumed significance level.

Proposal

Because of this reason for the tests based on indirect estimators we propose to additionally permute values of the auxiliary variable for nonsampled population elements.

Permutation tests for equality of subpopulations means

Estimators $\hat{\theta}_k$ and $\hat{\theta}_l$ in

$$T = |\hat{\theta}_k - \hat{\theta}_l|$$

are:

Permutation tests for equality of subpopulations means

- HT direct estimators (HT) - only for strata
- GREG direct estimators (GREG) - only for strata
- MGREG indirect estimators and permutation is conducted only for the sample data (MGREGs)
- MGREG indirect estimators and permutation is conducted for the sample data and additionally values of the auxiliary variable for **nonsampled** population elements (MGREG**p**)
- EBLUP indirect predictors and permutation is conducted only for the sample data (EBLUPs)
- EBLUP indirect predictors with permutation conducted for the sample data and additionally for values of the auxiliary variable for **nonsampled** population elements (EBLUP**p**)
- synthetic regression indirect estimator with permutation of auxiliary variable for sampled and **nonsampled** population elements (SYN-REG)

Permutation tests for equality of subpopulations means

- HT direct estimators (HT) - only for strata
- GREG direct estimators (GREG) - only for strata
- MGREG indirect estimators and permutation is conducted only for the sample data (MGREGs)
- MGREG indirect estimators and permutation is conducted for the sample data and additionally values of the auxiliary variable for **nonsampled** population elements (MGREG**p**)
- EBLUP indirect predictors and permutation is conducted only for the sample data (EBLUPs)
- EBLUP indirect predictors with permutation conducted for the sample data and additionally for values of the auxiliary variable for **nonsampled** population elements (EBLUP**p**)
- synthetic regression indirect estimator with permutation of auxiliary variable for sampled and **nonsampled** population elements (SYN-REG)

Permutation tests for equality of subpopulations means

- HT direct estimators (HT) - only for strata
- GREG direct estimators (GREG) - only for strata
- MGREG indirect estimators and permutation is conducted only for the sample data (MGREGs)
- MGREG indirect estimators and permutation is conducted for the sample data and additionally values of the auxiliary variable for **nonsampled** population elements (MGREG**p**)
- EBLUP indirect predictors and permutation is conducted only for the sample data (EBLUPs)
- EBLUP indirect predictors with permutation conducted for the sample data and additionally for values of the auxiliary variable for **nonsampled** population elements (EBLUP**p**)
- synthetic regression indirect estimator with permutation of auxiliary variable for sampled and **nonsampled** population elements (SYN-REG)

Permutation tests for equality of subpopulations means

- HT direct estimators (HT) - only for strata
- GREG direct estimators (GREG) - only for strata
- MGREG indirect estimators and permutation is conducted only for the sample data (MGREGs)
- MGREG indirect estimators and permutation is conducted for the sample data and additionally values of the auxiliary variable for **nonsampled** population elements (MGREG**p**)
- EBLUP indirect predictors and permutation is conducted only for the sample data (EBLUPs)
- EBLUP indirect predictors with permutation conducted for the sample data and additionally for values of the auxiliary variable for **nonsampled** population elements (EBLUP**p**)
- synthetic regression indirect estimator with permutation of auxiliary variable for sampled and **nonsampled** population elements (SYN-REG)

Permutation tests for equality of subpopulations means

- HT direct estimators (HT) - only for strata
- GREG direct estimators (GREG) - only for strata
- MGREG indirect estimators and permutation is conducted only for the sample data (MGREGs)
- MGREG indirect estimators and permutation is conducted for the sample data and additionally values of the auxiliary variable for **nonsampled** population elements (MGREG**p**)
- EBLUP indirect predictors and permutation is conducted only for the sample data (EBLUPs)
- EBLUP indirect predictors with permutation conducted for the sample data and additionally for values of the auxiliary variable for **nonsampled** population elements (EBLUP**p**)
- synthetic regression indirect estimator with permutation of auxiliary variable for sampled and **nonsampled** population elements (SYN-REG)

Permutation tests for equality of subpopulations means

- HT direct estimators (HT) - only for strata
- GREG direct estimators (GREG) - only for strata
- MGREG indirect estimators and permutation is conducted only for the sample data (MGREGs)
- MGREG indirect estimators and permutation is conducted for the sample data and additionally values of the auxiliary variable for **nonsampled** population elements (MGREG**p**)
- EBLUP indirect predictors and permutation is conducted only for the sample data (EBLUPs)
- EBLUP indirect predictors with permutation conducted for the sample data and additionally for values of the auxiliary variable for **nonsampled** population elements (EBLUP**p**)
- synthetic regression indirect estimator with permutation of auxiliary variable for sampled and **nonsampled** population elements (SYN-REG)

Permutation tests for equality of subpopulations means

- HT direct estimators (HT) - only for strata
- GREG direct estimators (GREG) - only for strata
- MGREG indirect estimators and permutation is conducted only for the sample data (MGREGs)
- MGREG indirect estimators and permutation is conducted for the sample data and additionally values of the auxiliary variable for **nonsampled** population elements (MGREG**p**)
- EBLUP indirect predictors and permutation is conducted only for the sample data (EBLUPs)
- EBLUP indirect predictors with permutation conducted for the sample data and additionally for values of the auxiliary variable for **nonsampled** population elements (EBLUP**p**)
- synthetic regression indirect estimator with permutation of auxiliary variable for sampled and **nonsampled** population elements (SYN-REG)

Monte Carlo simulation study

Data:

- population of municipalities (NUTS 5), N=2476, without Warsaw and Łódź
- strata – voivodships (NUTS 2), H=16
- domains – subregions (NUTS 3), D=70, without Warsaw and Łódź
- y – the number of unemployed in 2016 (in thousands)
- x – population in 2016 (in thousands)
- assumed significance level: $\alpha = 0.1$

Monte Carlo simulation study

Simulation under H_1 hypothesis - real data

Simulation under H_0 hypothesis - modified real data:

- in case of testing strata means: to values of x and y in each strata we add a constant to obtain x-mean and y-mean in the strata equal respective population means
- in case of testing domains means: to values of x and y in each domain we add a constant to obtain x-mean and y-mean in the domain equal respective population means

Monte Carlo simulation study

Simulation under H_1 hypothesis - real data

Simulation under H_0 hypothesis - modified real data:

- in case of testing strata means: to values of x and y in each strata we add a constant to obtain x-mean and y-mean in the strata equal respective population means
- in case of testing domains means: to values of x and y in each domain we add a constant to obtain x-mean and y-mean in the domain equal respective population means

Monte Carlo simulation study

Data:

- design-based simulation study: stratified simple random sampling without replacement with optimal allocation $(n = 0.1N)$
- model-based simulation study: values of the variable of interest are generated based on the following superpopulation model with values of parameters computed based on the whole population real dataset:

$$Y_{ih} = (\beta_1 + v_h)x_{ih} + \beta_0 + e_{ih}$$

where $h = 1, 2, ..., H$, $i = 1, 2, ..., N$, $v_h \sim (0, \sigma_v^2)$, $e_{ih} \sim (0, \sigma_e^2)$ and e_{ih}'s and v_h's are mutually independent

Monte Carlo simulation study

- tests for equality of strata means (**between the first one and the others**)
- sample sizes in strata from 4 to 34
- tests for equality of domains means (**between the first one and the others**) expected sample sizes in domains from 0,07 to 8,31
- number of samples drawn in the simulation study: 1000
- number of permutations: 200

Monte Carlo simulation study

- tests for equality of strata means (**between the first one and the others**)
- sample sizes in strata from 4 to 34
- tests for equality of domains means (**between the first one and the others**) expected sample sizes in domains from 0,07 to 8,31
- number of samples drawn in the simulation study: 1000
- number of permutations: 200

Monte Carlo simulation study

- tests for equality of strata means (**between the first one and the others**)
- sample sizes in strata from 4 to 34
- tests for equality of domains means (**between the first one and the others**) expected sample sizes in domains from 0,07 to 8,31
- number of samples drawn in the simulation study: 1000
- number of permutations: 200

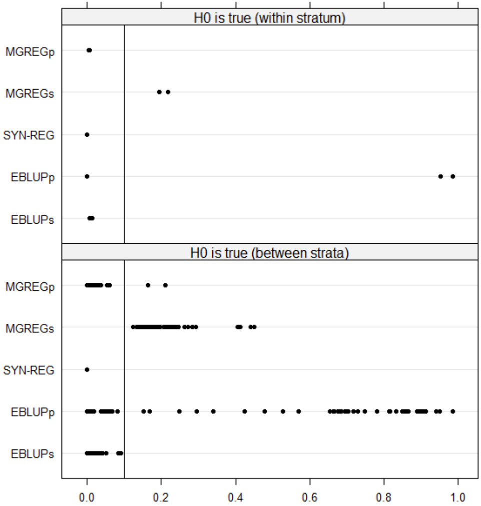
Monte Carlo simulation study

We would like to receive simulated values of type I error probabilities closer to the assumed significance level ($\alpha = 0.1$).

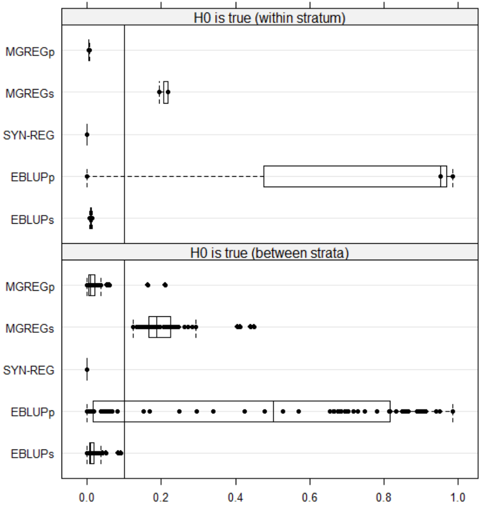
Särndal, Swensson and Wretman (1992) p. 281-284 show results of a simulation study conducted to verify the assumed nominal $1 - \alpha = 95\%$ confidence level based on confidence intervals constructed under the normality assumption and based on different estimators (not only Horvitz-Thompson estimator). They considered population of size $N = 281$ and for sample size $n = 50$ the simulation coverage rates were ca. 91%-92%.

Design-based simulation:
testing equality of domains means

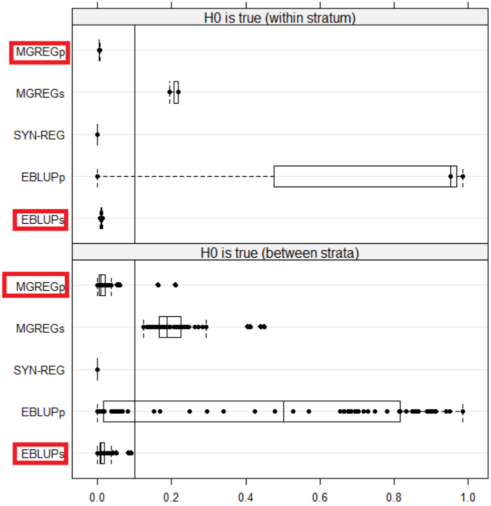
Design-based simulated values of type I error probabilities for tests of equalitys of domains means



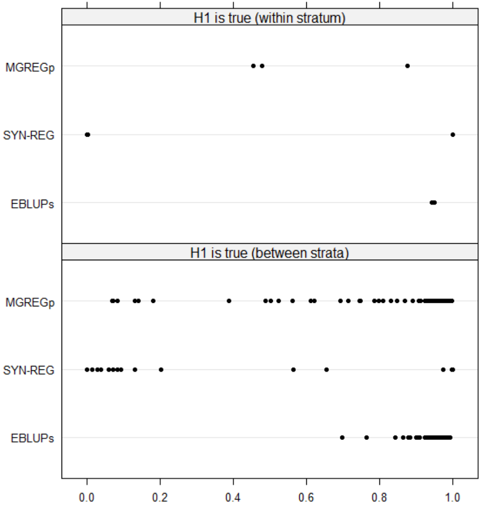
Design-based simulated values of type I error probabilities for tests of equality of domains means



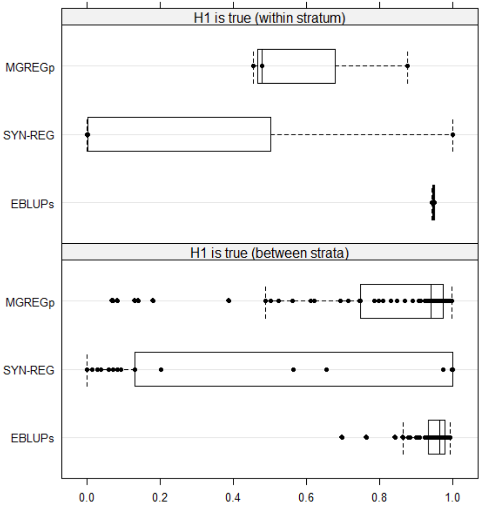
Design-based simulated values of type I error probabilities for tests of equality of domains means



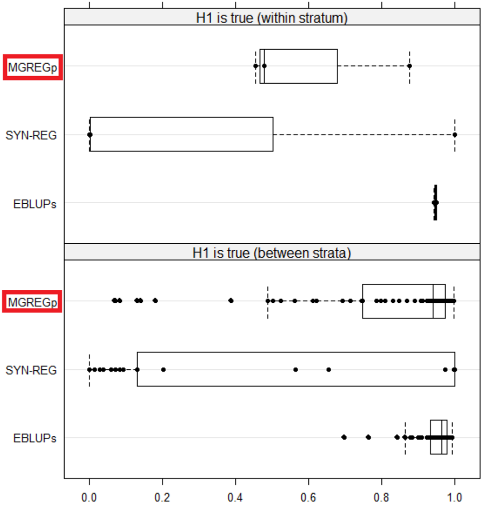
Design-based simulated values of type II error probabilities for tests of equality of domains means



Design-based simulated values of type II error probabilities for tests of equality of domains means

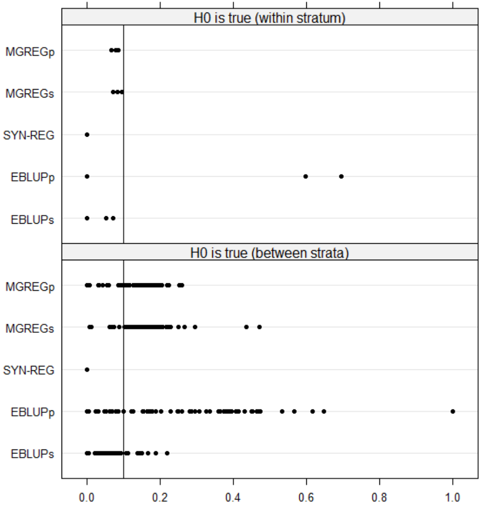


Design-based simulated values of type II error probabilities for tests of equality of domains means

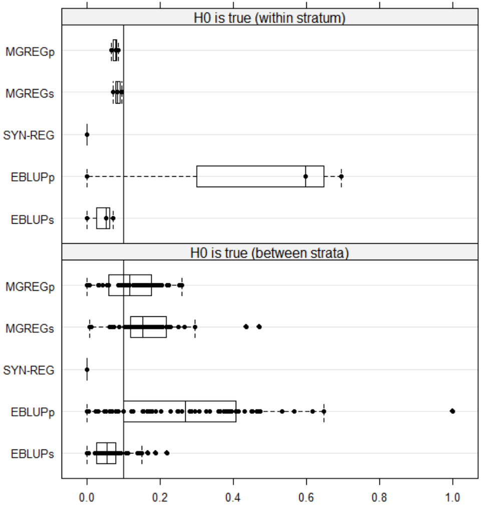


Model-based simulation: testing equality of domains means

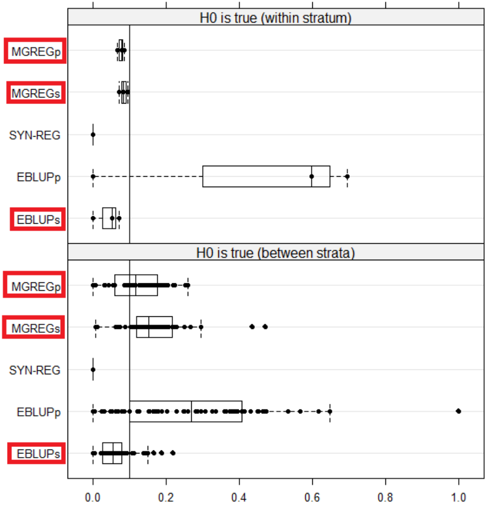
Model-based simulated values of type I error probabilities for tests of equality of domains means



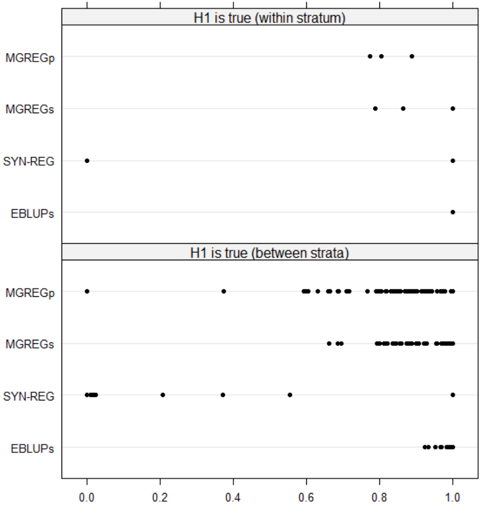
Model-based simulated values of type I error probabilities for tests of equality of domains means



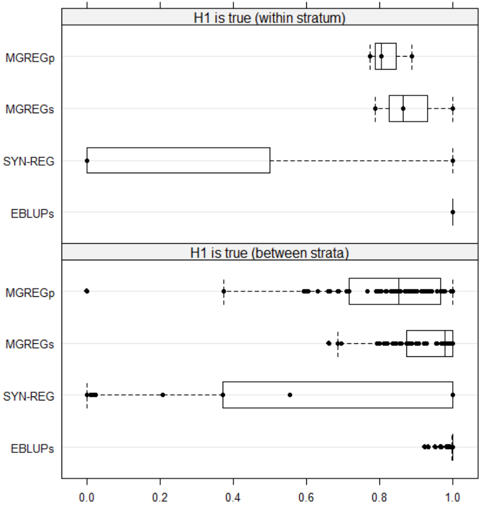
Model-based simulated values of type I error probabilities for tests of equality of domains means



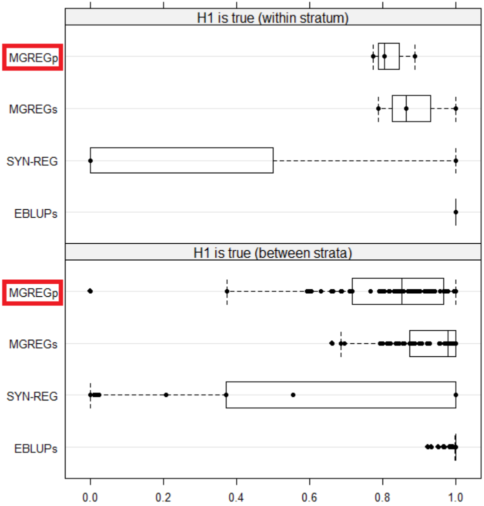
Model-based simulated values of type II error probabilities for tests of equality of domains means



Model-based simulated values of type II error probabilities for tests of equality of domains means

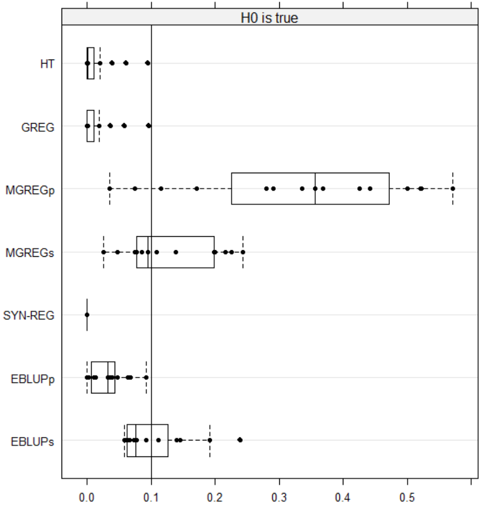


Model-based simulated values of type II error probabilities for tests of equality of domains means

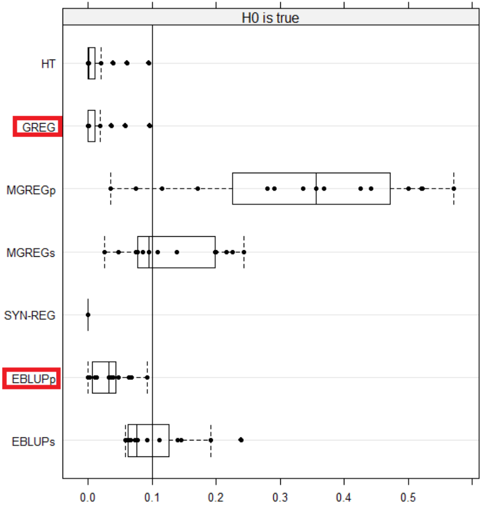


Design-based simulation: testing equality of strata means

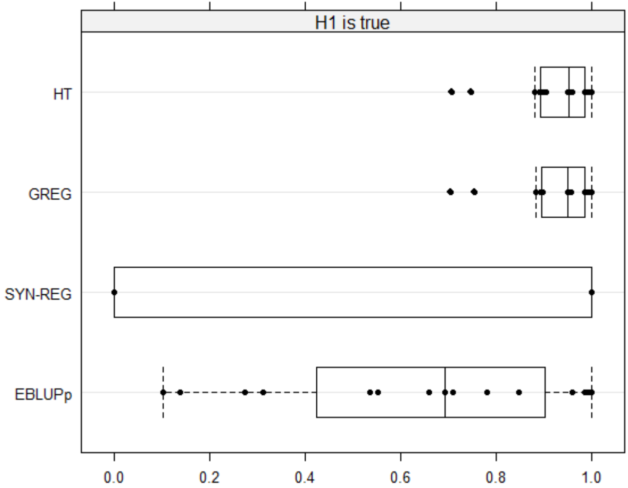
Design-based simulated values of type I error probabilities for tests of equality of strata means



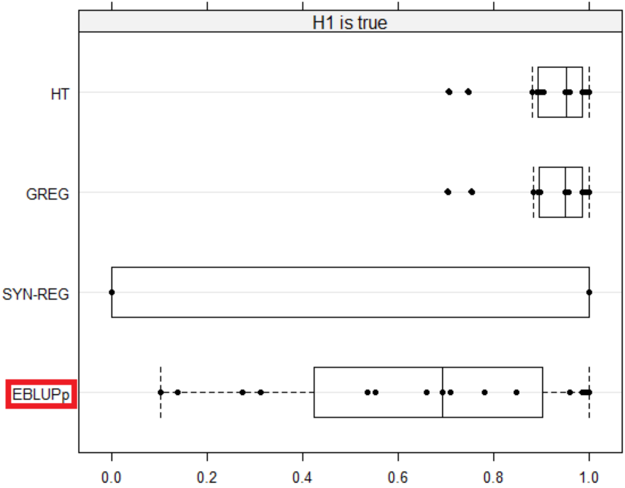
Design-based simulated values of type I error probabilities for tests of equality of strata means



Design-based simulated values of type II error probabilities for tests of equality of strata means

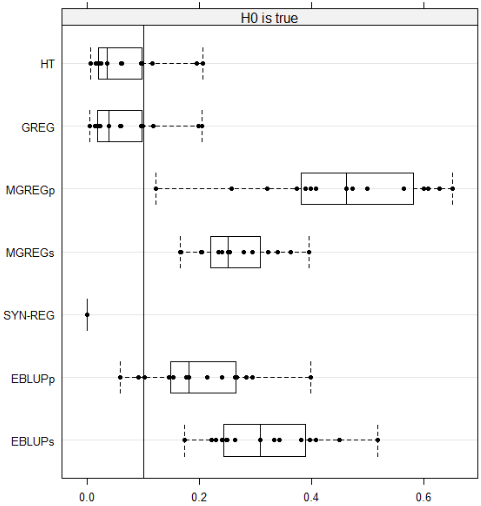


Design-based simulated values of type II error probabilities for tests of equality of strata means

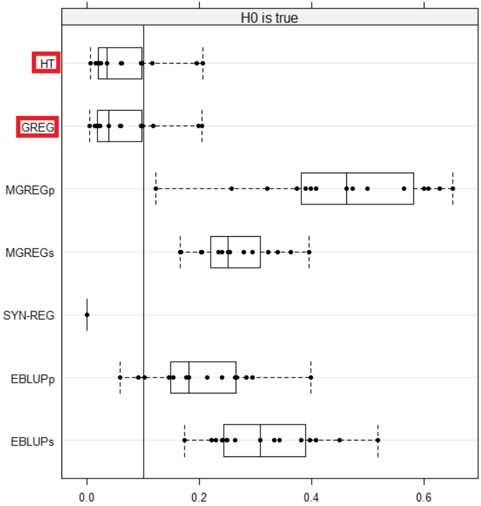


Model-based simulation: testing equality of strata means

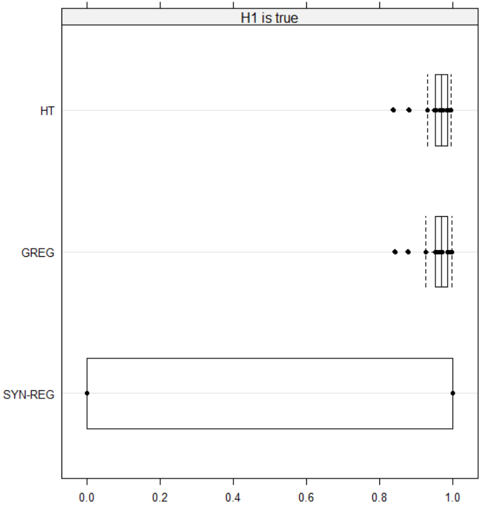
Model-based simulated values of type I error probabilities for tests of equality of strata means



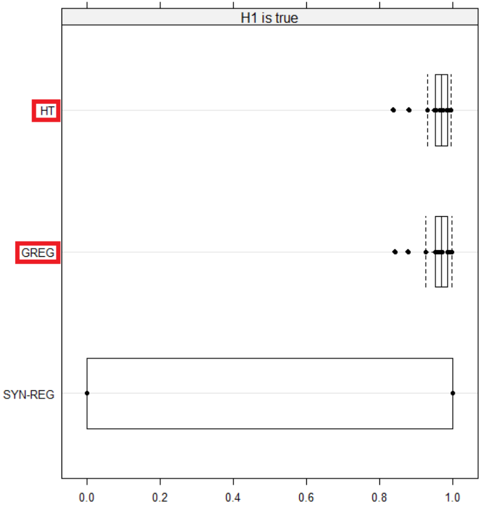
Model-based simulated values of type I error probabilities for tests of equality of strata means



Model-based simulated values of type II error probabilities for tests of equality of strata means



Model-based simulated values of type II error probabilities for tests of equality of strata means



Conclusion

- Permutation tests for equality of subpopulation means were proposed.
- The lack of exchangeability may have strong influence on tests' properties
- Simulation study: the test statistic using MGREG indirect estimators and proposed additional permutation of values of auxiliary variable for unsampled population elements can be preferred to test domains means for the consdidered data.

Literature

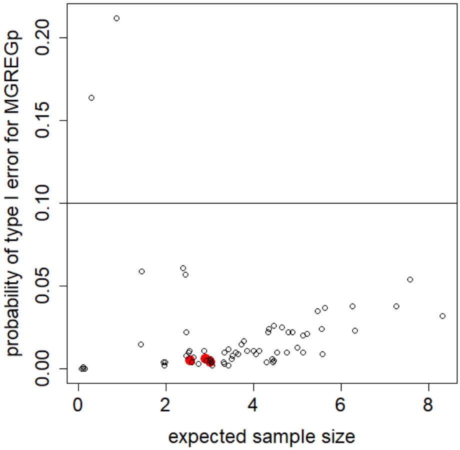
- Berger, Y (1998), Rate of convergence for asymptotic variance of the Horvitz–Thompson estimator, Journal of Statistical Planning and Inference, 74 (1), 149-168.
- Bracha, Cz. (1996), Teoretyczne podstawy metody reprezentacyjnej, PWN, Warszawa.
- Jiang, J. (1996), REML estimation: asymptotic behavior and related topics, The Annals of Statistics, 24 (1), 255-286.
- Pesarin, F, Salmaso, L. (2010), Permutation test for complex data. Theory, application and software, Wiley, Chichester.
- Rao, JNK, Molina, I (2015), Small area estimation. Second edition. Wiley, New Jersey.

Literature

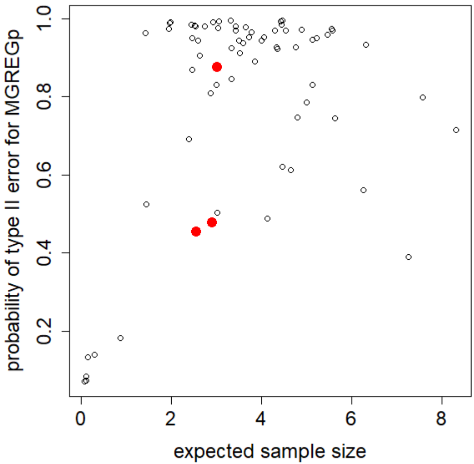
- Royall, R.M. (1976), The linear least-squares prediction approach to two-stage sampling, Journal of the American Statistical Association, 71 (335), 657-673.
- Särndal, C.E. (1981), Frameworks for inference in survey sampling with applications to small area estimation and adjustment for nonresponse, Bulletin of the International Statistical Institute, 49, 494-513.
- Särndal, CE, Swensson, B, and Wretman, J. (1992), Model assisted survey sampling, Springer, New York.
- Wywiał, J.L. (2017), On the evaluation of sample size required for a good approximation by the normal curve for some statistics, Zeszyty Naukowe Uniwersytetu Ekonomicznego w Krakowie, 5 (965), 17-29.
- Żądło, T. (2015), Statystyka małych obszarów. Podejście modelowe i mieszane, Uniwersytet Ekonomiczny w Katowicach, Katowice.

Thank you for your attention

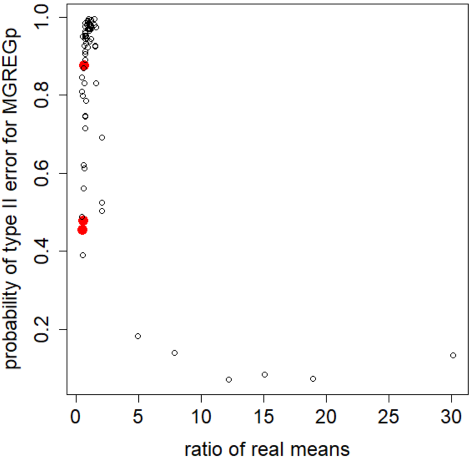
Design-based simulated values of type I error probabilities vs. expected sample sizes in domains using MGREGp



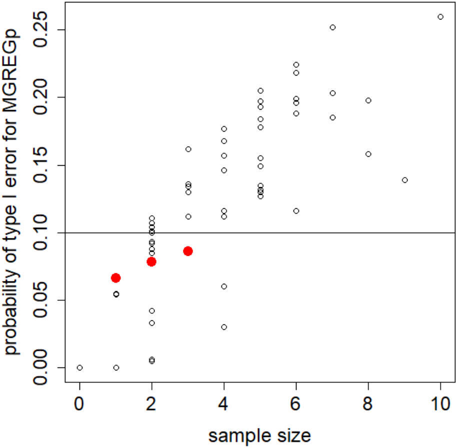
Design-based simulated values of type II error probabilities vs. expected sample sizes in domains using MGREGp



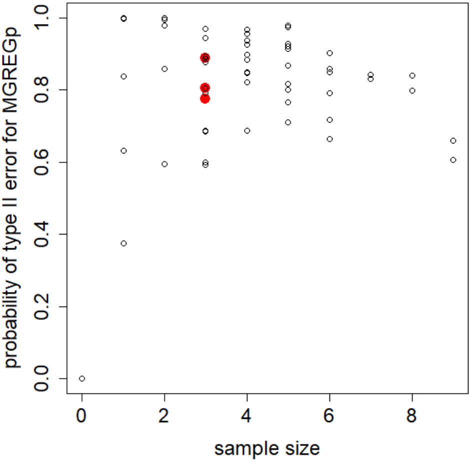
Design-based simulated values of type II error probabilities vs. ratios of real means in domains using MGREGp



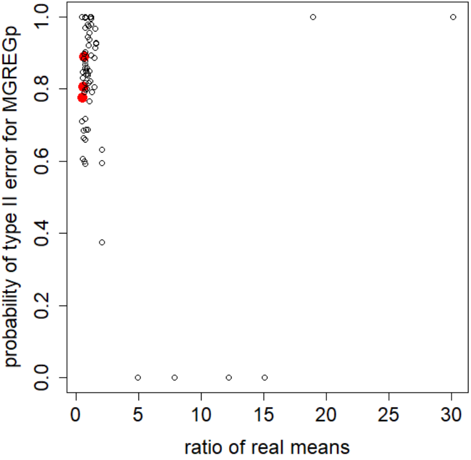
Model-based simulated values of type I error probabilities vs. sample sizes in domains using MGREGp



Model-based simulated values of type II error probabilities vs. sample sizes in domains using MGREGp



Model-based simulated values of type II error probabilities vs. ratios of real means in domains using MGREGp



Estimators of subpopulation mean used in the test statistic

The Horvitz-Thompson design-unbiased estimator:

$$\hat{\theta}_d^{HT} = N_d^{-1} \sum_{i \in s_d} \frac{y_i}{\pi_i}$$

It can be used if at least one subpopulation element is in the sample (or two elements to additionally assess the precision of estimation).

Estimators of subpopulation mean used in the test statistic

The generalized regression estimator approximately design-unbiased even if the expected domain sample size is small (Rao and Molina 2015: 18):

$$\hat{\theta}_d^{GREG} = N_d^{-1} \sum_{i \in s} w_{si} a_{id} y_i = N_d^{-1} \sum_{i \in s_d} w_{si} y_i$$

where

$$a_{id} = \begin{cases} 1 & for \quad i \in \Omega_d \\ 0 & for \quad i \notin \Omega_d \end{cases}$$

weights w_{si} are solutions of:

$$\begin{cases} \sum_{i \in s} \frac{\left(w_{si} - \pi_i^{-1}\right)^2}{\pi_i^{-1} q_i} \to \min \\ \sum_{i \in s} w_{si} \mathbf{x}_i = \sum_{i \in \Omega} \mathbf{x}_i \end{cases}$$

At least one subpopulation element in the sample in required (or two elements to additionally assess the precision of estimation).

Estimators of subpopulation mean used in the test statistic

Modified generalized regression estimator (MGREG) proposed by Särndal (1981):

$$\hat{\theta}_d^{MGREG} = N_d^{-1} \left(\sum_{i \in s_d} \pi_i^{-1} y_i + \left(\sum_{i \in \Omega_d} \mathbf{x}_i - \sum_{i \in s_d} \pi_i^{-1} \mathbf{x}_i \right)^T \hat{\mathbf{B}} \right)$$

where

$$\hat{\mathbf{B}} = \left(\sum_{i \in s} \pi_i^{-1} q_i \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} \sum_{i \in s} \pi_i^{-1} q_i \mathbf{x}_i y_i$$

It can be used to estimate the subpopulation mean even if the subpopulation sample size is zero.

Estimators of subpopulation mean used in the test statistic

Synthetic regression estimator given by (e.g. Bracha 1996, p. 260):

$$\hat{\theta}_d^{SYNT-REG} = \left(\sum_{i\in s} \pi_i^{-1}\right)^{-1} \sum_{i\in s} y_i \pi_i^{-1}$$

$$+B\left(N_d^{-1} \sum_{i\in\Omega_d} x_i - \left(\sum_{i\in s} \pi_i^{-1}\right)^{-1} \sum_{i\in s} x_i \pi_i^{-1}\right)$$

It can be used to estimate the subpopulation mean even if the subpopulation sample size is zero.

Estimators of subpopulation mean used in the test statistic

EBLUP under the following mixed model (chosen for the real data considered in the paper based on AIC)

$$Y_{id} = (\beta_1 + v_d)x_{id} + \beta_0 + e_{id}$$

where $d = 1, 2, ..., D$, $i = 1, 2, ..., N$, $v_d \sim (0, \sigma_v^2)$, $e_{id} \sim (0, \sigma_e^2)$ and e_{id}'s and v_d's are mutually independent

It can be used to estimate the subpopulation mean even if the subpopulation sample size is zero.

Estimators of subpopulation mean used in the test statistic

$$\hat{\theta}_{dh}^{BLUP} = N_d^{-1}\left(\sum_{i\in s_d} Y_i + \left[\begin{array}{cc} \sum_{i\in\Omega_{rd}} x_i & N_{rd} \end{array}\right]\hat{\beta} + \right.$$

$$\left. + \sigma_v^2 b_h^{-1} \sum_{i\in\Omega_{rd}} x_i \left(\sum_{i\in s_h} x_i y_i - \left[\begin{array}{cc} \sum\limits_{i\in s_h} x_i^2 & \sum\limits_{i\in s_h} x_i \end{array}\right]\hat{\beta}\right)\right)$$

where $\Omega_d \subset \Omega_h$, $b_h = \sigma_e^2 + \sigma_v^2 \sum\limits_{i\in s_h} x_i^2$ and $\hat{\beta}$ is the BLUE of $\left[\begin{array}{cc} \beta_1 & \beta_0 \end{array}\right]$. To obtain the EBLUP unknown σ_e^2 and σ_v^2 are replaced by their REML estimators.