Integrating a non-probability sample and its complementary probability sample

Andrius Čiginas[1]

(joint work with Jae-Kwang Kim[2] & Ieva Burakauskaitė[1])

[1]Vilnius University, Lithuania
[2]Iowa State University, U.S.A.

2024 Joint Statistical Meetings

Motivation

- The Lithuanian Population and Housing Census 2021 was based mainly on administrative data.
- Sampling was used to collect the rest of the necessary data on ethnicity, native language and religion.
- The sample was obtained in two steps:
  1. about 2% of the population provided their data voluntarily;
  2. a probability sample was drawn from non-respondents and covered about 6% of the census population.
- The complete data on the same variables from previous censuses helped to apply data integration approaches from Chen, Li, and Wu (2020).
- Now, we assume a *non-ignorable* selection mechanism for the non-probability sample.

Data framework

Let y be the study variable with the fixed values $y_1, \ldots, y_N$ in a survey population $U = \{1, \ldots, N\}$. We aim to estimate the population mean

$$\mu = \frac{1}{N} \sum_{i \in U} y_i.$$

The sample S measuring the variable y is collected in two steps.

1. At first, a non-probability sample A of size n_A is obtained from U.
2. Then, a sample B of size n_B is drawn from the set $U \backslash A$ by the probability sampling design $p(\cdot)$ with known inclusion into the sample probabilities $\pi_i^{(B)} = \mathrm{P}_p(i \in B) > 0$, $i \in U \backslash A$.

This way, the sample $S = A \cup B$ of size $n = n_A + n_B$ is obtained.

We assume that the vector values $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{im})'$ of the auxiliary variables $\mathbf{x}$ are known for $i \in S$.

Model for propensity scores

We model the unknown selection mechanism of the non-probability sample A. Let

$$\delta_i = \begin{cases} 1 & \text{if } i \in A, \\ 0 & \text{if } i \in U \backslash A, \end{cases} \qquad i \in U,$$

be the inclusion indicators, which are assumed to be independent random variables. The propensity scores are the probabilities

$$\pi_i^{(A)} = \mathrm{P}_q(\delta_i = 1 \,|\, \mathbf{x}_i, y_i), \quad i \in U,$$

where the subscript q refers to a model. The model that we postulate to estimate $\pi_i^{(A)}$, $i \in A$, is the parametric logistic regression

$$\pi_i^{(A)} = \pi^{(A)}(\mathbf{x}_i, y_i; \boldsymbol{\phi}) = \frac{\exp(\boldsymbol{\phi}_1' \mathbf{x}_i + \phi_2 y_i)}{1 + \exp(\boldsymbol{\phi}_1' \mathbf{x}_i + \phi_2 y_i)}, \quad i \in U,$$

with the parameter $\boldsymbol{\phi} = (\boldsymbol{\phi}_1', \phi_2)'$ and data $(\mathbf{x}_i', y_i)'$ known for $i \in S$. This model implies that the selection mechanism is *non-ignorable*.

Inverse probability weighted estimation

After we have the estimates $\hat{\pi}_i^{(A)}$ of the propensity scores, we can construct the inverse probability weighted (IPW) estimator

$$\hat{\mu}_{IPW} = \frac{1}{\widehat{N}_A} \sum_{i \in A} \frac{y_i}{\hat{\pi}_i^{(A)}}, \quad \text{where} \quad \widehat{N}_A = \sum_{i \in A} \frac{1}{\hat{\pi}_i^{(A)}},$$

to estimate the population mean μ.

A simple alternative to the IPW estimator of μ is the design-based post-stratified estimator (Kim & Tam, 2021)

$$\hat{\mu}_{dir} = \frac{1}{N} \left[\sum_{i \in A} y_i + \frac{N_B}{\widehat{N}_B} \sum_{i \in B} \frac{y_i}{\pi_i^{(B)}} \right], \quad \text{where} \quad \widehat{N}_B = \sum_{i \in B} \frac{1}{\pi_i^{(B)}}$$

and N_B is the size of $U \backslash A$. It is a benchmark for comparisons.

Estimation of $\pi_i^{(A)}$ following Chen et al. (2020)

Treating the sample A as fixed, one can say that $S = A \cup B$ was drawn by a probability sampling design with the inclusion probabilities

$$\pi_i = \begin{cases} 1 & \text{if } i \in A, \\ \pi_i^{(B)} & \text{if } i \in U \backslash A. \end{cases}$$

Then the Chen et al. (2020) maximum likelihood (ML) estimation approach, which assumes the ignorable selection mechanism, can be extended to the non-ignorable selection situation. That is, the ML estimator of $\pi_i^{(A)}$ is $\hat{\pi}_i^{(A)} = \pi^{(A)}(\mathbf{x}_i, y_i; \hat{\boldsymbol{\phi}})$, where $\hat{\boldsymbol{\phi}}$ maximizes the estimated (pseudo) log-likelihood function

$$l(\boldsymbol{\phi}) = \sum_{i \in S} \frac{1}{\pi_i} \left(\delta_i (\boldsymbol{\phi}_1' \mathbf{x}_i + \phi_2 y_i) - \log\{1 + \exp(\boldsymbol{\phi}_1' \mathbf{x}_i + \phi_2 y_i)\} \right).$$

The ML estimator $\hat{\boldsymbol{\phi}}$ is obtained by applying a standard iterative procedure to solve the score equations $\partial l(\boldsymbol{\phi}) / \partial \boldsymbol{\phi} = \mathbf{0}$.

Alternative estimation of $\pi_i^{(A)}$

The inclusion into the sample $S = A \cup B$ probabilities are decomposed as follows:

$$\mathrm{P}(i \in S) = \mathrm{P}(i \in A) + \mathrm{P}(i \in B \,|\, i \in U \backslash A)\, \mathrm{P}(i \in U \backslash A), \quad i \in U.$$

Then, conditional on the observed pooled sample S, the inclusion into the sample A probabilities are expressed as

$$p_i(\boldsymbol{\phi}) := \mathrm{P}(i \in A \,|\, i \in S) = \frac{\pi_i^{(A)}}{\pi_i^{(A)} + \tilde{\pi}_i^{(B)}(1 - \pi_i^{(A)})}, \quad i \in S.$$

Here the conditional probabilities

$$\tilde{\pi}_i^{(B)} = \mathrm{P}(i \in B \,|\, i \in U \backslash A)$$

have a simple expression only if $p(\cdot)$ is the simple random sampling. In such a case, $\tilde{\pi}_i^{(B)} = n_B / N_B$ for all $i \in S$.

Alternative estimation of $\pi_i^{(A)}$ (cont.)

For complex sampling design $p(\cdot)$, we first fit a parametric weight model

$$\tilde{w}_i^{(B)} = \tilde{w}^{(B)}(\mathbf{x}_i, y_i; \boldsymbol{\theta})$$

with the parameter $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \theta_2)'$ using the data $w_i^{(B)} = 1/\pi_i^{(B)}$ and $(\mathbf{x}_i, y_i)$ available in the sample B, and obtain the smoothed weights $\hat{w}_i^{(B)} = \tilde{w}^{(B)}(\mathbf{x}_i, y_i; \hat{\boldsymbol{\theta}})$ for all $i \in S$.

Second, due to the relation between conditional inclusion probabilities and sample weights (Pfeffermann & Sverchkov, 1999), we can compute the estimates $\hat{\pi}_i^{(B)} = 1/\hat{w}_i^{(B)}$ of the quantities $\tilde{\pi}_i^{(B)}$, $i \in S$.

The alternative ML estimator $\hat{\pi}_i^{(A)} = \pi^{(A)}(\mathbf{x}_i, y_i; \hat{\boldsymbol{\phi}})$ of $\pi_i^{(A)}$ is obtained by finding the value $\hat{\boldsymbol{\phi}}$ maximizing the log-likelihood function

$$l(\boldsymbol{\phi}) = \sum_{i \in S} \left(\delta_i \log p_i(\boldsymbol{\phi}) + (1 - \delta_i) \log\{1 - p_i(\boldsymbol{\phi})\} \right),$$

that is, numerically solving the score equations $\partial l(\boldsymbol{\phi})/\partial \boldsymbol{\phi} = \mathbf{0}$.

Remark

The alternative estimation method combines a two-phase sampling framework with conditional likelihood.

- Two-phase sampling framework: we can view the data integration problem (combining A and B) as a two-phase sampling problem, where $S = A \cup B$ is the first-phase sample and A is the second-phase sample.
- Conditional likelihood approach: we use the conditional selection probability $p_i(\boldsymbol{\phi}) = \mathrm{P}(i \in A \,|\, i \in S)$ and estimate $\boldsymbol{\phi}$ by maximizing the conditional log-likelihood $l(\boldsymbol{\phi})$ for $\boldsymbol{\phi}$.

Simulations: estimators, population and sampling

Estimators to compare:

- $\hat{\mu}_{CLW}$ – the IPW estimator of the mean μ based on the estimators $\hat{\pi}_i^{(A)}$ constructed as in Chen et al. (2020);
- $\hat{\mu}_{new}$ – the IPW estimator based on the alternative $\hat{\pi}_i^{(A)}$;
- $\hat{\mu}_{dir}$ – the post-stratified estimator.

The real Lithuanian Census 2021 sample data and the ML method extending the approach of Chen et al. (2020) were used to find (estimate) the 'true' propensity score model.

Generating the sample $S = A \cup B$ from the Census 2011 population:

1. the non-probability sample A is selected by the Poisson sampling with inclusion probabilities from the 'true' model;
2. two scenarios are considered for the probability sample B:
   - the simple random sampling design;
   - a complex stratified cluster sampling close to the real design of the Census 2021.

Simulations: sample sizes and model covariates

Generating independently $R = 1000$ sample pairs $\{A, B\}$:

1. two scenarios are considered for the sample A:
   - it covers 2% of the population U;
   - it takes up 20% of U;
2. the probability sample B covers about 6% of U.

Considering the binary study variable *is-the-person-Roman-Catholic*, the covariates in the propensity score model also include the following binary auxiliary sociodemographic variables:

- gender (male);
- residency status (living in Vilnius);
- identification with the Polish national minority;
- level of education (higher education);
- employment status (employed).

These covariates are the same as in the 'true' model.

Simulations: weight smoothing models

For complex sampling design $p(\cdot)$, the conditional probabilities $\tilde{\pi}_i^{(B)}$ are needed to be estimated.

The two candidate models could be:

- the smoothing model (Kim & Skinner, 2013)

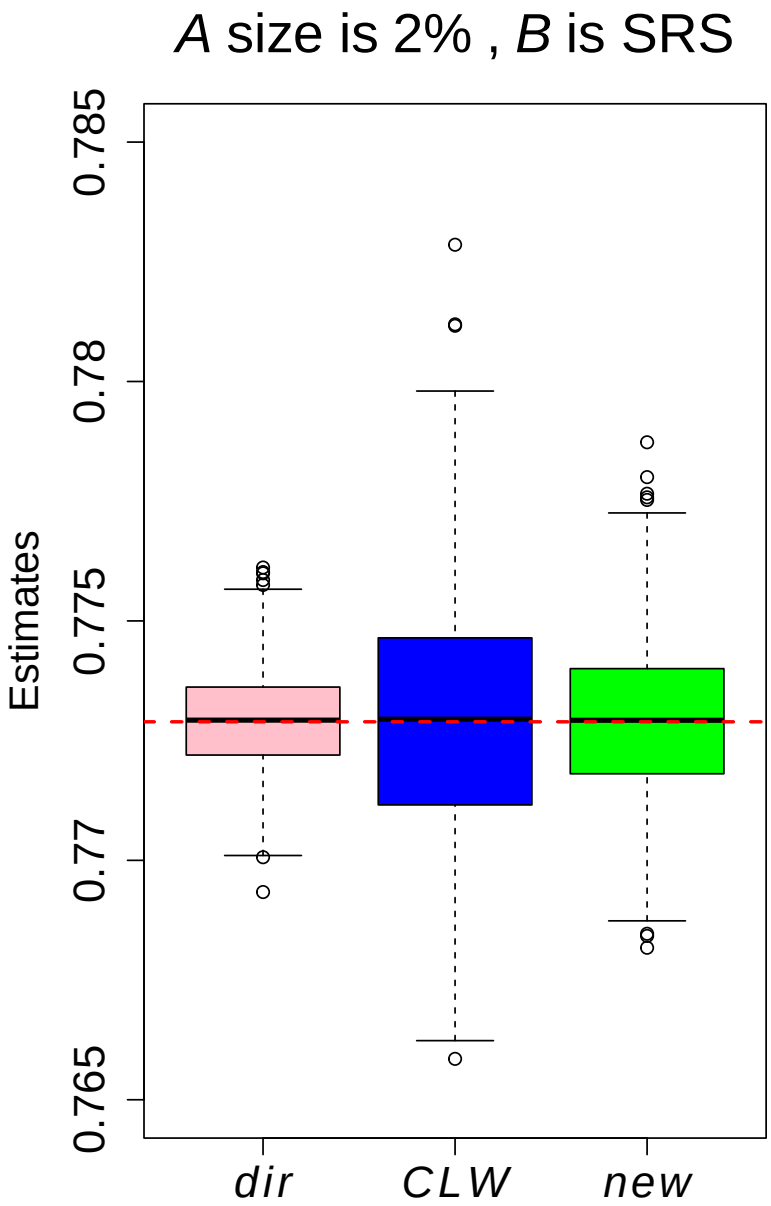
$$\tilde{w}^{(B)}(\mathbf{x}_i, y_i; \boldsymbol{\theta}) = 1 + \exp(-\boldsymbol{\theta}_1' \mathbf{x}_i - \theta_2 y_i) \qquad \text{(KS)}$$

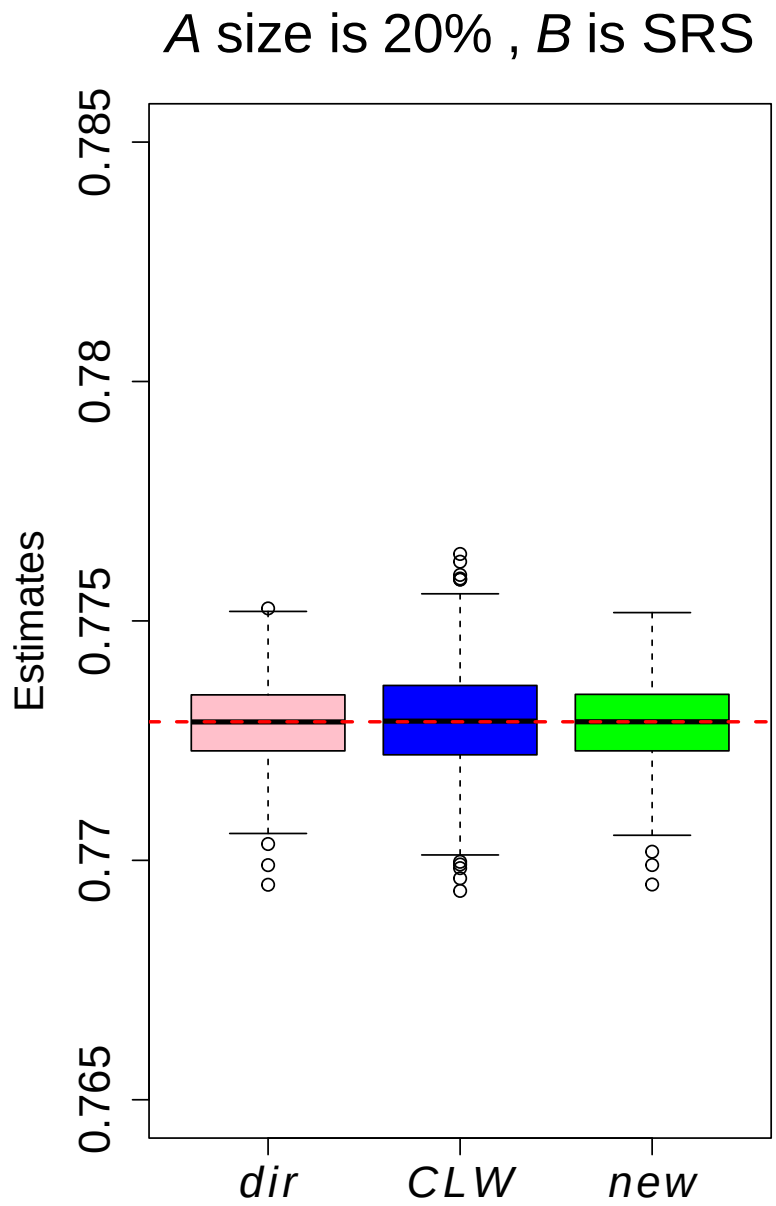
  with the same covariates that were used in the propensity score model and 14 additional dummy covariates describing the household size;

- the negative binomial model, which uses the same covariates as model (KS).

Both models work similarly, so we apply the smoothing model (KS).

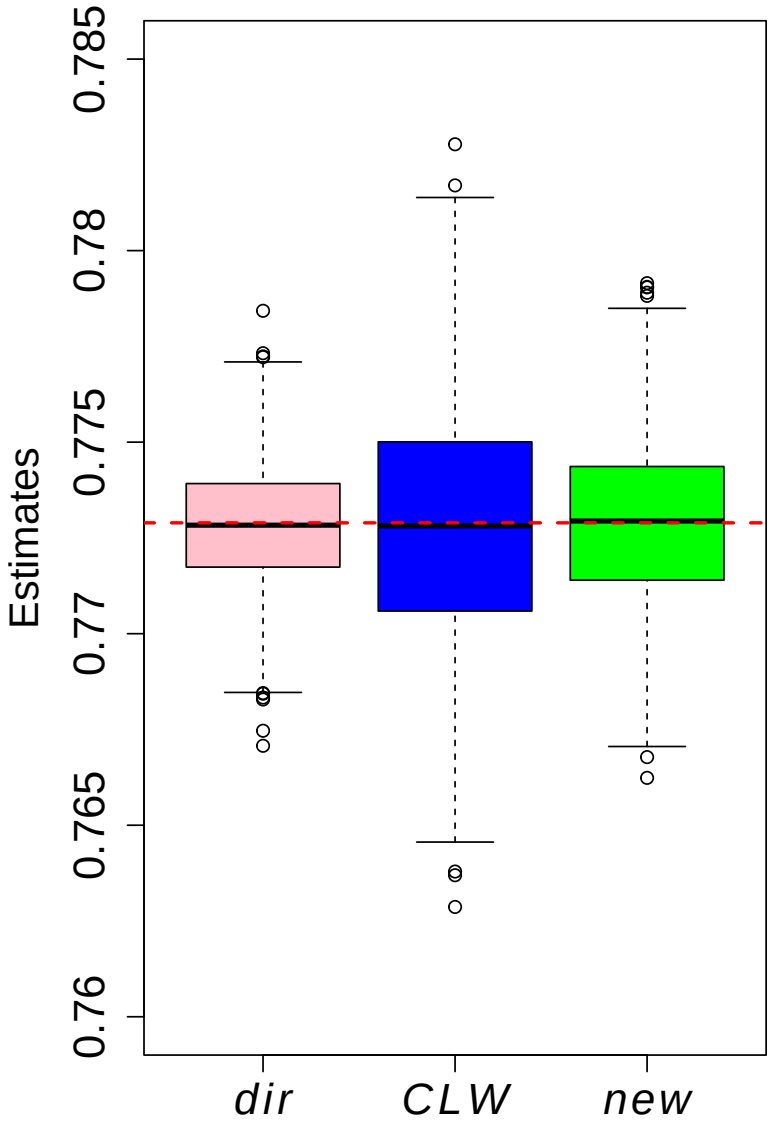
Simulations: results for simple random sampling case



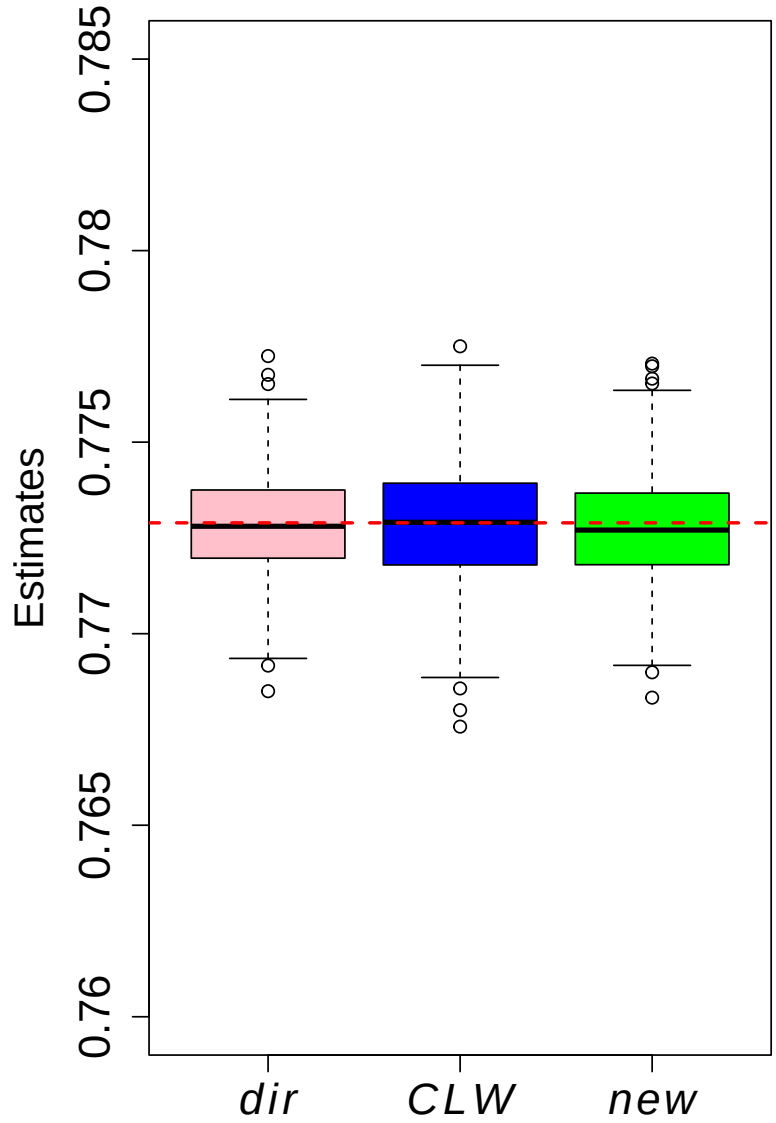


Simulations: results for complex sampling case

A size is 2% , *B* is complex



A size is 20% , *B* is complex



Variance estimation

A pseudo population bootstrap procedure for the estimator $\hat{\mu}_{new}$ is similar to that suggested in Liu et al. (2023).

Algorithm (Pseudo population bootstrap)

1. Estimate the probabilities $\pi_i^{(S)} = \mathrm{P}(i \in S)$, $i \in S$, using

$$\hat{\pi}_i^{(S)} = \hat{\pi}_i^{(A)} + \hat{\pi}_i^{(B)}(1 - \hat{\pi}_i^{(A)}),$$

   and define the weights $w_i = 1/\hat{\pi}_i^{(S)}$, $i \in S$.
2. Multiply each w_i by $N/\sum_{i \in S} w_i$ to obtain $\sum_{i \in S} w_i = N$.
3. Randomly round w_i to its ceiling with probability $w_i - \lfloor w_i \rfloor$ and to its floor otherwise to obtain $[w_i]$, and ensure that $\sum_{i \in S}[w_i] = N$.
4. Create a pseudo population by copying unit i $[w_i]$ times.

Variance estimation (cont.)

5. Draw a bootstrap non-probability sample A from the pseudo population by Poisson sampling design with probabilities $\hat{\pi}_i^{(A)}$.
6. Draw a bootstrap probability sample B of size n_B from the rest of the pseudo population according to the sampling design $p(\cdot)$ with new inclusion into the sample probabilities.
7. Get the bootstrap estimate of the target parameter.
8. Repeat steps 5–7 for T times to obtain T bootstrap estimates.

Using the bootstrap estimates $\hat{\mu}_{new}^{(t)}$, $t = 1, \ldots, T$, of the mean μ,

$$\widehat{V}(\hat{\mu}_{new}) = \frac{1}{T-1} \sum_{t=1}^{T} (\hat{\mu}_{new}^{(t)} - \bar{\mu}_{new})^2 \quad \text{with} \quad \bar{\mu}_{new} = \frac{1}{T} \sum_{t=1}^{T} \hat{\mu}_{new}^{(t)}$$

is the bootstrap estimate of the variance $\mathrm{Var}(\hat{\mu}_{new})$.

Simulations: validity of bootstrap procedure

A twenty times smaller population is used. We evaluate:

- relative biases (RB) of the variance estimator $\widehat{V}(\hat{\mu}_{new})$;
- coverage rates (CR) of the 95% confidence intervals.

Table: Bootstrap characteristics for simple random sampling case.

size of A	RB (%)	normal CR (%)	empirical CR (%)
2%	6.68	99.8	99.7
20%	1.23	94.9	93.9

Table: Bootstrap characteristics for complex sampling case.

size of A	RB (%)	normal CR (%)	empirical CR (%)
2%	0.56	97.6	97.4
20%	8.06	95.6	94.8

Conclusions and further work

- Assuming the *non-ignorable* selection mechanism, two IPW estimators of the population mean are constructed:
  - (i) a modification of Chen et al. (2020) approach;
  - (ii) an alternative approach based on the conditional ML for propensity scores.
- Simulating a census survey, alternative estimator (ii) is more efficient than estimator (i), especially when a non-probability sample is smaller.
- A pseudo population bootstrap is an option to estimate the variance of estimator (ii). An alternative will be to apply the Taylor expansion to obtain a variance estimator.
- One can try to combine the IPW estimator (ii) and a simple post-stratified estimator of Kim & Tam (2021) in some way to benefit from both of them.

References

Chen, Y., Li, P., Wu, C. (2020). Doubly robust inference with non-probability survey samples. *J. Am. Stat. Assoc.* 115:2011–2021.

Kim, J.-K., Skinner, C.J. (2013). Weighting in survey analysis under informative sampling. *Biometrika* 100:385–398.

Kim, J.-K., Tam, S.-M. (2021). Data integration by combining big data and survey sample data for finite population inference. *Int. Stat. Rev.* 89:382–401.

Liu, A.-C., Scholtus, S., de Waal, T. (2023). Correcting selection bias in big data by pseudo-weighting. *J. Surv. Stat. Methodol.* 11:1181–1203.

Pfeffermann, D., Sverchkov, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhya B* 61:166–186.