

# A general multiply robust framework for combining probability and non-probability samples in surveys

David Haziza

Department of mathematics and statistics  
University of Ottawa

Joint work with  
Sixia Chen (University of Oklahoma)

Journées de Méthodologie Statistique 2022  
Paris, France

March 29, 2022

# Combining probability and non-probability samples

- Traditionally, National Statistical Offices have collected data by means of probability sampling procedures → Design-based inference
- In recent years, there has been a shift of paradigm in NSOs that can be explained by **three main factors**:
  - (i) a dramatic decrease in response rates;
  - (ii) increasing data collection costs;
  - (iii) the availability of various types of non-probabilistic data sources that include administrative files, opt-in panels, social medias and satellite information.
- Non-probabilistic data sources provide timely data but they often fail to represent the target population of interest because of inherent selection biases.

# Combining probability and non-probability samples

- How to integrate data from non-probability samples has attracted a lot of attention in recent years; e.g., Rivers (2007), Bethlehem (2016), Elliot and Vaillant (2017), Lohr and Raghunathan (2017), Kim et al. (2019), Chen et al. (2020), Beaumont (2020) and Rao (2020).
- Estimation procedures may be classified into three broad classes :
  - (i) **Calibration weighting** of a nonprobability sample to estimated benchmarks from a probability survey;
  - (ii) **Statistical matching or mass imputation**;
  - (iii) **Propensity score weighting** of a nonprobability sample;
- Focus of this presentation: (ii) and (iii).

## Parameters of interest

- Consider a finite population  $\mathcal{P}$  of size  $N$ .
- $y$ : a survey variable
- $y_i$ :  $y$ -value attached to unit  $i$ ,  $i = 1, \dots, N$ .
- **Goal**: estimate a finite population parameter  $\theta_0$  defined as the solution of **the census estimating equation**:

$$\frac{1}{N} \sum_{i \in \mathcal{P}} U(y_i; \theta_0) = 0.$$

Parameter	$U(y_i; \theta_0)$	Explicit form of $\theta_0$
Mean	$y_i - \theta_0$	$\bar{Y} = \sum_{i \in \mathcal{P}} y_i / N$
Population $\tau_\alpha$ -th percentile	$1(y_i \leq \theta_0) - \alpha$	$\tau_\alpha = F_N^{-1}(\alpha)$

# The setup

- $S_A$  : sample, of size  $n_A$ , selected from  $\mathcal{P}$  according to a probability sampling design with first-order inclusion probabilities  $\pi_i$  (Known).
- $S_B$ : Non-probability sample, of size  $n_B$ , from  $\mathcal{P}$ .
- Typically, we would expect  $n_B > n_A$ .
- The data:

Data	$\mathbf{x} = (x_1, \dots, x_p)$	$y$	
$S_A$	✓	∅	Probability
$S_B$	✓	✓	Non-Probability

- $l_i$  : sample selection indicator such that  $l_i = 1$  if  $i \in S_A$  and  $l_i = 0$ , otherwise.
- $\delta_i$  : a participation indicator such that  $\delta_i = 1$  if  $i \in S_B$  and  $\delta_i = 0$ , otherwise.

	$y$	$x_1$	...	$x_p$	$l_i$	$1/\pi_i$
<b>1</b>	X	✓	...	✓	<b>1</b>	$1/\pi_1$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$n_A$	X	✓	...	✓	<b>1</b>	$1/\pi_{n_A}$
$n_A + 1$	X	✓	...	✓	<b>0</b>	$1/\pi_{n_A+1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$N$	X	✓	...	✓	<b>0</b>	$1/\pi_N$

Table 1: Probability sample  $S_A$  ( $n_A$ )

	$y$	$x_1$	...	$x_p$	$\delta_i$
<b>1</b>	✓	✓	...	✓	<b>1</b>
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$n_B$	✓	✓	...	✓	<b>1</b>
$n_B + 1$	X	X	...	X	<b>0</b>
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$N$	X	X	...	X	<b>0</b>

Table 2: Non-probability sample  $S_B$  ( $n_B$ )

# The setup

- $\pi_i = P(I_i = 1) = P(i \in S_A)$  is known for all  $i \in \mathcal{P}$ .
- **Unknown** probability of participation on the non-probability source:  
 $\Pr(\delta_i = 1 | \mathbf{x}_i, y_i) = \Pr(\delta_i = 1 | \mathbf{x}_i) \triangleq p(\mathbf{x}_i; \boldsymbol{\alpha}) \rightarrow$  **participation model**
- **Positivity assumption:**

$$p(\mathbf{x}_i; \boldsymbol{\alpha}) > 0 \quad \text{for all } i \in \mathcal{P}.$$

- **Outcome regression model:**

$$y_i = m(\mathbf{x}_i; \boldsymbol{\beta}) + \epsilon_i,$$

where  $\mathbb{E}(\epsilon_i | \mathbf{x}_i) = 0$  and  $\mathbb{V}(\epsilon_i | \mathbf{x}_i) = \sigma^2$ .

# The setup

- **Statistical matching (or mass imputation)**
  - Specification of an outcome regression model
  - The resulting estimator may be biased if the outcome regression model is misspecified.
- **Propensity score weighting**
  - Specification of a participation model
  - The resulting estimator may be biased if the participation model is misspecified.
- Regardless of the approach, **the validity of point estimators relies on the validity of an assumed model** → point estimators are vulnerable to model misspecification.
- **Multiply robust estimation procedures** are attractive because they provide some protection against misspecification of the model.



## Two classes of models

- Class of potential outcome regression models:

$$\mathcal{M}_1 = \left\{ m^{(j)}(\mathbf{x}; \boldsymbol{\beta}^{(j)}), j = 1, 2, \dots, J \right\}$$

$J$  models for the survey variable  $y$

- Class of potential participation models:

$$\mathcal{M}_2 = \left\{ p^{(k)}(\mathbf{x}; \boldsymbol{\alpha}^{(k)}), k = 1, 2, \dots, K \right\}$$

$K$  models for the participation probability

- The models in  $\mathcal{M}_1$  (respectively in  $\mathcal{M}_2$ ) may be based on different functionals and/or different vectors of explanatory variables.

## Estimation of the $\beta$ 's

- Estimators of  $\beta^{(j)}, j = 1, 2, \dots, J$  : obtained by solving the sample estimating equations

$$\frac{1}{N} \sum_{i \in S_B} \left\{ y_i - m^{(j)}(\mathbf{x}_i; \beta^{(j)}) \right\} \left\{ \frac{\partial m^{(j)}(\mathbf{x}_i; \beta^{(j)})}{\partial \beta^{(j)}} \right\}^\top = 0.$$

- **Special case:** The  $j$ th model is a linear regression model  $\longrightarrow$

$$\hat{\beta}^{(j)} = \left( \sum_{i \in S_B} \mathbf{x}_i \mathbf{x}_i^\top \right)^{-1} \sum_{i \in S_B} \mathbf{x}_i y_i.$$

- For each unit  $i$ , we obtain  $J$  predicted values:

$$m^{(1)}(\mathbf{x}_i; \hat{\beta}^{(1)}), m^{(2)}(\mathbf{x}_i; \hat{\beta}^{(2)}), \dots, m^{(J)}(\mathbf{x}_i; \hat{\beta}^{(J)})$$

	$y$	$x_1$	...	$x_p$	$l_i$	$1/\pi_i$
<b>1</b>	X	✓	...	✓	<b>1</b>	$1/\pi_1$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$n_A$	X	✓	...	✓	<b>1</b>	$1/\pi_{n_A}$
$n_A + 1$	X	✓	...	✓	<b>0</b>	$1/\pi_{n_A+1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$N$	X	✓	...	✓	<b>0</b>	$1/\pi_N$

Table 3: Probability sample  $S_A$  ( $n_A$ )

	$y$	$x_1$	...	$x_p$	$\delta_i$
<b>1</b>	✓	✓	...	✓	<b>1</b>
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$n_B$	✓	✓	...	✓	<b>1</b>
$n_B + 1$	X	X	...	X	<b>0</b>
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$N$	X	X	...	X	<b>0</b>

Table 4: Non-probability sample  $S_B$  ( $n_B$ )

## Estimation of the $\alpha$ 's

- If  $\mathbf{x}_i$  was available for  $i \in \mathcal{P} - S_B$ , we would estimate  $\alpha^{(k)}$ ,  $k = 1 \dots, K$ , by solving the census estimating equations

$$\frac{1}{N} \sum_{i \in \mathcal{P}} \frac{\delta_i - p_i^{(k)}}{p_i^{(k)}(1 - p_i^{(k)})} \left\{ \frac{\partial p_i^{(k)}}{\partial \alpha^{(k)}} \right\}^T = 0,$$

where  $p_i^{(k)} \equiv p^{(k)}(\mathbf{x}_i; \alpha^{(k)})$ .

- Idea in Chen et al. (2020):

$$\frac{1}{N} \sum_{i \in \mathcal{P}} \frac{\delta_i}{p_i^{(k)}(1 - p_i^{(k)})} \left\{ \frac{\partial p_i^{(k)}}{\partial \alpha^{(k)}} \right\}^T - \frac{1}{N} \sum_{i \in \mathcal{P}} \frac{1}{1 - p_i^{(k)}} \left\{ \frac{\partial p_i^{(k)}}{\partial \alpha^{(k)}} \right\}^T = 0.$$

## Estimation of the $\alpha$ 's

- The estimators of  $\alpha^{(k)}$ ,  $k = 1, 2, \dots, K$  can be obtained by solving

$$\frac{1}{N} \sum_{i \in S_B} \frac{1}{p_i^{(k)}(1 - p_i^{(k)})} \left\{ \frac{\partial p_i^{(k)}}{\partial \alpha^{(k)}} \right\}^\top - \frac{1}{N} \sum_{i \in S_A} \pi_i^{-1} \frac{1}{1 - p_i^{(k)}} \left\{ \frac{\partial p_i^{(k)}}{\partial \alpha^{(k)}} \right\}^\top = 0.$$

- For each unit  $i$ , we obtain  $K$  estimated participation probabilities:

$$p^{(1)}(\mathbf{x}_i; \hat{\alpha}^{(1)}), p^{(2)}(\mathbf{x}_i; \hat{\alpha}^{(2)}), \dots, p^{(K)}(\mathbf{x}_i; \hat{\alpha}^{(K)})$$

## Compressing the information

- For each  $i$ , define

$$\mathbf{v}_{1i} = (m^{(1)}(\mathbf{x}_i; \hat{\boldsymbol{\beta}}^{(1)}), m^{(2)}(\mathbf{x}_i; \hat{\boldsymbol{\beta}}^{(2)}), \dots, m^{(J)}(\mathbf{x}_i; \hat{\boldsymbol{\beta}}^{(J)}))^\top$$

and

$$\mathbf{v}_{2i} = (p^{(1)}(\mathbf{x}_i; \hat{\boldsymbol{\alpha}}^{(1)}), p^{(2)}(\mathbf{x}_i; \hat{\boldsymbol{\alpha}}^{(2)}), \dots, p^{(K)}(\mathbf{x}_i; \hat{\boldsymbol{\alpha}}^{(K)}))^\top.$$

- Compress the information contained in the  $J$  outcome regression models in  $\mathcal{M}_1$  by fitting a linear regression model based on the units in  $S_B$  with  $y$  as the dependent variable and  $\mathbf{v}_1$  as the vector of explanatory variables.
- The compressed score is  $\hat{m}_i = \mathbf{v}_{1i}^\top \hat{\boldsymbol{\tau}}_1$  with

$$\hat{\boldsymbol{\tau}}_1 = \left\{ \sum_{i \in S_B} \mathbf{v}_{1i} \mathbf{v}_{1i}^\top \right\}^{-1} \sum_{i \in S_B} \mathbf{v}_{1i} y_i.$$

## Compressing the information

- Compress the information contained in the  $K$  participation models in  $\mathcal{M}_2$  by fitting a linear regression model with  $\delta$  as the dependent variable and  $v_2$  as the vector of explanatory variables.
- If  $v_{2i}$  was available for all  $i \in \mathcal{P}$ , the compressed score would be  $\tilde{p}_i = v_{2i}^\top \tilde{\tau}_2$  with

$$\tilde{\tau}_2 = \left\{ \sum_{i \in \mathcal{P}} v_{2i} v_{2i}^\top \right\}^{-1} \sum_{i \in \mathcal{P}} v_{2i} \delta_i.$$

- **Solution:**

$$\hat{\tau}_2 = \left\{ \sum_{i \in \mathcal{S}_A} \pi_i^{-1} v_{2i} v_{2i}^\top \right\}^{-1} \sum_{i \in \mathcal{S}_B} v_{2i}.$$

- The compressed score is  $\hat{p}_i = v_{2i}^\top \hat{\tau}_2$ .

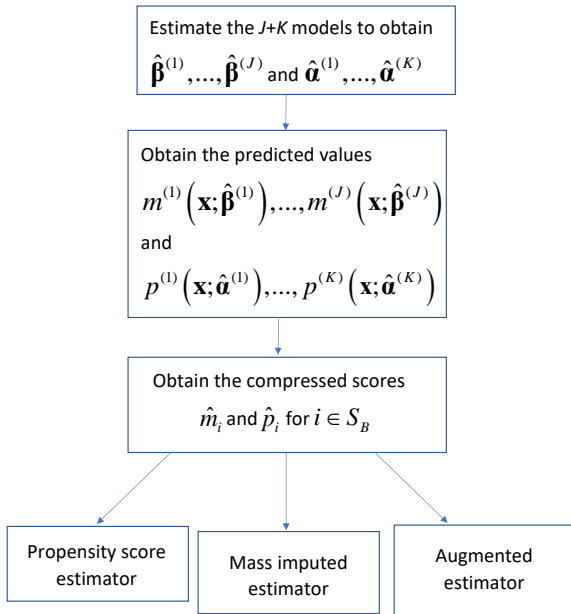


Figure 1: Steps for multiply robust estimation



## Propensity score estimation

- Inverse probability weighting estimator  $\hat{\theta}_{IPW}$ : obtained by solving the sample estimating equations:

$$\hat{U}_{IPW}(\theta) = \frac{1}{N} \sum_{i \in S_B} \frac{1}{\hat{p}_i} U(y_i; \theta) = 0.$$

- $\hat{\theta}_{IPW}$ : multiply robust in the sense that it remains consistent if one of the participation models in  $\mathcal{M}_2$  is correctly specified.
- Special case: The population mean

$$\hat{\theta}_{IPW} = \frac{\sum_{i \in S_B} \frac{y_i}{\hat{p}_i}}{\sum_{i \in S_B} \frac{1}{\hat{p}_i}}$$

## Fractionally mass imputation

- Fractionally mass imputed estimator  $\hat{\theta}_{FMI}$ :
- A consistent estimator of  $\theta_0$  is obtained by solving the following expected estimating equations:

$$\frac{1}{N} \sum_{i \in S_A} \pi_i^{-1} \mathbb{E} \{ U(y_i; \theta_0) \mid \mathbf{x}_i \} = 0.$$

- The expectation is unknown as  $f(y \mid \mathbf{x})$  is unknown.
- We want to approximate the conditional expectation by the weighted mean of the fractionally imputed estimating equations:

$$\mathbb{E} \{ U(y_i; \theta) \mid \mathbf{x}_i \} \approx \sum_{j \in S_B} w_{ij}^* U(y_i^{*(j)}; \theta),$$

where  $w_{ij}^*$  are the fractional weights such that  $\sum_{j \in S_B} w_{ij}^* = 1$  and the  $y_i^{*(j)}$ 's denote the imputed values for unit  $i$ .

	$y$	$x_1$	...	$x_p$	$l_i$	$1/\pi_i$
<b>1</b>	X	✓	...	✓	<b>1</b>	$1/\pi_1$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$n_A$	X	✓	...	✓	<b>1</b>	$1/\pi_{n_A}$
$n_A + 1$	X	✓	...	✓	<b>0</b>	$1/\pi_{n_A+1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$N$	X	✓	...	✓	<b>0</b>	$1/\pi_N$

Table 5: Probability sample  $S_A$  ( $n_A$ )

	$y$	$x_1$	...	$x_p$	$\delta_i$
<b>1</b>	✓	✓	...	✓	<b>1</b>
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$n_B$	✓	✓	...	✓	<b>1</b>
$n_B + 1$	X	X	...	X	<b>0</b>
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$N$	X	X	...	X	<b>0</b>

Table 6: Non-probability sample  $S_B$  ( $n_B$ )

## Fractionally mass imputation

- Fractionally mass imputed estimator  $\hat{\theta}_{FMI}$ :

(Step1). Obtain the weights  $\hat{w}_i$  by maximizing the empirical likelihood function

$$l = \sum_{i \in S_B} \log(w_i),$$

subject to

$$\sum_{i \in S_B} w_i = 1, \quad \sum_{i \in S_B} w_i \hat{\epsilon}_i = 0,$$

where  $\hat{\epsilon}_i = y_i - \hat{m}_i$  denotes the residual attached to  $i \in S_B$ .

(Step2). Obtain  $\hat{\theta}_{FMI}$  by solving the sample estimating equations

$$\hat{U}_{FMI}(\theta) = \frac{1}{N} \sum_{i \in S_A} \pi_i^{-1} \sum_{j \in S_B} w_{ij}^* U(y_i^{*(j)}; \theta) = 0,$$

with  $w_{ij}^* = \hat{w}_j$  denote the fractional weight and  $y_i^{*(j)} = \hat{m}_i + \hat{\epsilon}_j$ .

# Fractionally mass imputation

- $\hat{\theta}_{FMI}$ : multiply robust in the sense that it remains consistent if one of the outcome regression models in  $\mathcal{M}_1$  is correctly specified.
- **Special case:** The population mean

$$\hat{\theta}_{FMI} = \frac{\sum_{i \in S_A} \pi_i^{-1} \sum_{j \in S_B} w_{ij}^* y_i^{*(j)}}{\sum_{i \in S_A} \frac{1}{\pi_i}}$$

## Augmented estimator

- **Augmented estimator  $\hat{\theta}_{AMR}$** : can be obtained by solving the sample estimating equations

$$\begin{aligned}\hat{U}_{AMR}(\theta) &= \frac{1}{N} \sum_{i \in S_B} \frac{1}{\hat{p}_i} U(y_i; \theta) + \frac{1}{N} \sum_{i \in S_A} \pi_i^{-1} \sum_{j \in S_B} w_{ij}^* U(y_i^{*(j)}; \theta) \\ &- \frac{1}{N} \sum_{i \in S_B} \frac{1}{\hat{p}_i} \sum_{j \in S_B} w_{ij}^* U(y_i^{*(j)}; \theta) = 0.\end{aligned}$$

- $\hat{\theta}_{AMR}$ : multiply robust in the sense that it remains consistent if one of the models in either  $\mathcal{M}_1$  or  $\mathcal{M}_2$  is correctly specified.

## Simulation study

- We generated  $B = 1,000$  finite populations of size  $N = 20,000$ .
- Two auxiliary variables  $x_1$  and  $x_2$ : generated from a  $\mathcal{N}(1, 2)$
- The variable of interest  $y$  was generated according to

$$y = 0.3 + 2x_1 + 2x_2 + \epsilon,$$

where  $\epsilon \sim \mathcal{N}(0, 1)$ .

- From each finite population, a sample  $S_A$ , of size  $n_A$ , was selected using SRSWOR. We used  $n_A = 500$  and  $n_A = 1000$ .
- A non-probability sample  $S_B$  was generated using a Poisson sampling design with probability

$$p(\mathbf{x}_i; \alpha) = \frac{\exp(\alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i})}{1 + \exp(\alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i})}.$$

- The values of  $\alpha_0$ ,  $\alpha_1$ , and  $\alpha_2$  were chosen so as to lead to  $n_B$  approximately equal to 500 and 1000.

## Simulation study

- To assess the performance of the proposed methods in the presence of model misspecification, we defined **the transformed explanatory variables** as

$$z_1 = \exp(x_1/2) \text{ and } z_2 = x_2 \{1 + \exp(x_1)\}^{-1}.$$

- The **correct outcome regression and participation models** were fitted using a linear regression model and a logistic model, respectively, based on the set of explanatory variables  $\mathbf{x} = (x_1, x_2)^\top$ .
- The **incorrect outcome regression and participation models** were fitted using a linear regression model and a logistic model, respectively, based on the set of transformed explanatory variables  $\mathbf{z} = (z_1, z_2)^\top$ .
- We assume that only the variables  $\mathbf{x}$  and  $\mathbf{z}$  were available in  $S_A$ , whereas the variables  $\mathbf{x}$ ,  $\mathbf{z}$  and  $y$  were available in  $S_B$ .



# Simulation study

- **Goal:** estimate
  - The population mean of  $y$ ;
  - The population 25th percentile of  $y$ .
- We computed several estimators:
  - (1) The (unfeasible) design-weighted estimators (Benchmark) based on  $S_A$  obtained as a solution of the following estimating equations

$$\frac{1}{N} \sum_{i \in S_A} \pi_i^{-1} U(y_i; \theta) = 0.$$

- (2) The naive estimators (Naive) based on  $S_B$  obtained as a solution of the following estimating equations

$$\frac{1}{N} \sum_{i \in S_B} U(y_i; \theta) = 0.$$

## Simulation study

- (3) The parametric mass imputed estimators considered in Kim et al. (2019) using correct outcome regression model (PFMI(1000)) and the incorrect outcome regression model (PFMI(0100)).
- (4) The doubly robust estimators proposed by Chen et al. (2020): DR(1010), DR(1001), DR(0110) and DR(0101).
- (5) The MR inverse probability weighting estimator: MRIPW(0011).
- (6) The MR fractionally mass imputed estimator: MRFMI(1100).
- (7) The augmented multiply robust estimators: AMR(1110), AMR(1101), AMR(1011), AMR(0111) and AMR(1111).

## Simulation results

Parameter	Method	$(n_A, n_B = 500)$			$(n_A, n_B = 1000)$		
		RB (%)	RSE	RRMSE	RB (%)	RSE	RRMSE
	Benchmark	-0.04	1.59	1.59	-0.04	1.14	1.14
	Naive	9.37	1.56	9.50	9.14	1.16	9.21
	PFMI(1000)	0.02	1.59	1.59	-0.03	1.13	1.13
	PFMI(0100)	5.06	1.57	5.29	4.84	1.11	4.96
	DR(1010)	0.01	1.59	1.59	-0.03	1.13	1.13
	DR(1001)	0.02	1.60	1.60	-0.03	1.13	1.13
	DR(0110)	0.10	1.89	1.89	-0.04	1.34	1.34
Mean	DR(0101)	4.60	1.69	4.90	4.42	1.19	4.58
	MRIPW(0011)	0.48	1.98	2.04	0.12	1.37	1.37
	MRFMI(1100)	0.01	1.59	1.59	-0.03	1.12	1.12
	AMR(1110)	0.01	1.59	1.59	-0.03	1.13	1.13
	AMR(1101)	0.02	1.60	1.60	-0.03	1.13	1.13
	AMR(1011)	0.01	1.59	1.59	-0.03	1.13	1.13
	AMR(0111)	0.66	1.93	2.04	0.21	1.35	1.37
	AMR(1111)	0.01	1.59	1.59	-0.03	1.13	1.13

# Simulation results

Parameter	Method	$(n_A, n_B = 500)$			$(n_A, n_B = 1000)$		
		RB(%)	RSE	RRMSE	RB(%)	RSE	RRMSE
	Benchmark	0.02	2.85	2.85	-0.05	2.04	2.04
	Naive	12.54	2.89	12.87	12.15	2.1	12.33
	PFMI(1000)	1.96	2.84	3.45	1.84	2.02	2.73
	PFMI(0100)	21.98	2.48	22.12	21.74	1.81	21.82
25th percentile	MRIPW(0011)	0.44	3.69	3.72	-0.06	2.55	2.55
	MRFMI(1100)	0.03	2.52	2.52	-0.04	1.78	1.78
	AMR(1110)	0.83	3.03	3.14	0.19	2.14	2.15
	AMR(1101)	0.39	2.88	2.91	0.05	2.08	2.08
	AMR(1011)	0.78	2.99	3.09	0.19	2.14	2.15
	AMR(0111)	1.46	3.24	3.55	0.5	2.3	2.35
	AMR(1111)	0.77	2.99	3.08	0.2	2.15	2.16

## Simulation results: Bootstrap variance estimation

Table 7: Monte Carlo Percent Relative Bias (RB) of the bootstrap variance estimator, Coverage Rate (CR) %, and Average Length (AL) of confidence intervals for the proposed estimators with  $n_A = n_B = 500$ .

Parameter	Method	RB(%)	CR(%)	AL
Mean	MRIPW(0011)	6.56	95.7	0.66
	MRFMI(1100)	2.28	95.3	0.52
	AMR(1110)	2.29	95.3	0.52
	AMR(1101)	2.25	95.4	0.52
	AMR(1011)	2.30	95.3	0.52
	AMR(0111)	9.00	96.4	0.66
	AMR(1111)	2.30	95.3	0.52
25th percentile	MRIPW(0011)	9.37	95.5	0.94
	MRFMI(1100)	-0.42	94.7	0.63
	AMR(1110)	7.79	95.0	0.78
	AMR(1101)	4.41	95.3	0.74
	AMR(1011)	7.58	95.1	0.77
	AMR(0111)	9.28	96.3	0.89
	AMR(1111)	7.87	95.1	0.77

## Final remarks

- Here, we considered the case of parametric/semi-parametric models.
- Mass imputation: we can easily implement machine learning procedures. However, establishing the properties of the resulting point estimators is challenging.
- Propensity score weighting: Use of nonparametric participation models presents some challenges and requires additional research.



*La recherche consiste à établir des liens entre des choses  
qui n'ont apparemment pas de liens entre elles*

Jean-Claude Deville