# Correcting for selection error

Ton de Waal

Summer school 2023

# Correcting for selection error

- Based on work by my PhD student An-Chiao (Anne) Liu (and colleague Sander Scholtus)

# Probability samples

- Probability sampling according to well-designed sampling design enables one to obtain valid estimates for population parameters of interest
- However, collection of probability samples is time-consuming and expensive

# Nonprobability samples

- Nowadays wide diversity of new data sources (e.g. big data, register data, and opt-in online) provide massive amount of information at low cost
- However, some groups of units may be overrepresented in these data sources and other groups underrepresented
- As these nonprobability samples do not come from known sampling design, it is hard to obtain unbiased estimates for population parameters of interest

# Selection error

- We are interested in estimated population mean of continuous target variable $Y$ based on nonprobability sample
- Goal is to correct for selection bias in nonprobability sample

# Approaches for correcting selection error

- There are many (classes of) approaches
  - (Re-)weighting approaches where weights are assigned to units in nonprobability sample
  - Modelling approaches where model is assumed for nonprobability data
  - Mass imputation where all non-observed units in population are imputed
- For each of these classes of approaches many different approaches have been developed

# Re-weighting approaches

- Calibration
  - Use available information on (sub)totals of background variables to correct for selectivity
- Poststratification
  - Divide observed data into (small) strata and correct for selectivity in each stratum by using stratum total
- Sample matching
  - Match each unit in nonprobability sample to nearest neighbour in probability sample and use weights associated to that nearest neighbour

# Re-weighting approaches

- Pseudo-weights
  - Construct pseudo-weights that can be used for estimation and analysis purposes
  - Often probability sample is used to “borrow” design weights from
  - There are several different variants of pseudo-weighting
  - We will look at two variants of pseudo-weighting

# How does it work?

- When we have weights, estimating population parameters is usually quite easy
- Survey weights for a probability sample with common background variables as nonprobability sample provide lot of information with respect to good weights for nonprobability sample
- We adjust survey weights from probability sample to account for differences with respect to selectivity between nonprobability sample and probability sample

# Selection error: pseudo-weighting

- Elliott & Valliant (2017) proposed weighting method for correcting selection bias
  - Appears to work well in many cases
  - Drawback: not suitable for large inclusion fractions
- We proposed variant that is suitable for larger inclusion fractions and dependency between samples
  - Suitable for (selective) administrative datasets where units can be identified

# Situation

- We have nonprobability sample NPS in which target variable $Y$ and auxiliary variables $\boldsymbol{X}$ are observed
- We also have probability sampling PS in which variables $\boldsymbol{X}$ are observed (but $Y$ is not)

# Situation

- Let $S_i^* \in \{0,1\}$ denote inclusion indicator for NPS
- Let $S_i \in \{0,1\}$ denote inclusion indicator PS

# Assumptions by Elliott & Valliant (2017)

1. For all units in population, $\Pr(i \in \mathrm{NPS})$ and $\Pr(i \in PS)$ are nonzero
2. Auxiliary variables $\boldsymbol{X}$ govern inclusion mechanism of NPS
3. Inclusion weights $d$ for inclusion in PS are available or can be computed for all units in both PS and NPS
4. Sampling fractions of PS and NPS are small so that they do not overlap
5. Inclusion in NPS and PS is independent after conditioning on $\boldsymbol{X}$

# Pseudo-weights derived by Elliott & Valliant (2017)

$$\Pr(S_i^* = 1|\boldsymbol{x}_i = \boldsymbol{x}_o) = \frac{\Pr(\boldsymbol{x}_i = \boldsymbol{x}_o|S_i^* = 1)\Pr(S_i^* = 1)}{\Pr(\boldsymbol{x}_i = \boldsymbol{x}_o)}$$

$$= \frac{\Pr(\boldsymbol{x}_i = \boldsymbol{x}_o|S_i^* = 1)\Pr(S_i^* = 1)\Pr(S_i = 1|\boldsymbol{x}_i = \boldsymbol{x}_o)}{\Pr(S_i = 1)\Pr(\boldsymbol{x}_i = \boldsymbol{x}_o|S_i = 1)}$$

$$\propto \frac{\Pr(\boldsymbol{x}_i = \boldsymbol{x}_o|S_i^* = 1)}{\Pr(\boldsymbol{x}_i = x_o|S_i = 1)}\Pr(S_i = 1|\boldsymbol{x}_i = \boldsymbol{x}_o)$$

# How does it work?

- Here differences between NPS and PS are quantified

# Pseudo-weights derived by Elliott & Valliant (2017)

- Avoiding direct estimation of $\Pr(x_i = x_o | S_i^* = 1)$ and $\Pr(x_i = x_o | S_i = 1)$ Elliott and Valliant (2017) use discriminant analysis on combination of NPS and PS
- Set $Z_i^{EV} = 1$ for units from NPS and $Z_i^{EV} = 0$ for units from PS
- Note that $Z_i^{EV} = 1$ if $S_i^* = 1$ and $Z_i^{EV} = 0$ if $S_i = 1$

# Pseudo-weights derived by Elliott & Valliant (2017)

$$\frac{\Pr\left(x_i = x_o \middle| Z_i^{EV} = 1\right)}{\Pr\left(x_i = x_o \middle| Z_i^{EV} = 0\right)}$$

$$= \frac{\Pr\left(Z_i^{EV} = 1 \middle|\, x_i = x_o\right) \Pr(x_i = x_o) / \Pr\left(Z_i^{EV} = 1\right)}{\Pr\left(Z_i^{EV} = 0 \middle|\, x_i = x_o\right) \Pr(x_i = x_o) / \Pr\left(Z_i^{EV} = 0\right)}$$

$$\propto \frac{\Pr\left(Z_i^{EV} = 1 \middle|\, x_i = x_o\right)}{\Pr\left(Z_i^{EV} = 0 \middle|\, x_i = x_o\right)}$$

# Pseudo-weights derived by Elliott & Valliant (2017)

Combining expression for

$$\Pr(S_i^* = 1 | \boldsymbol{x}_i = \boldsymbol{x}_o)$$

with expression for

$$\frac{\Pr\left(x_i = x_o \middle| Z_i^{EV} = 1\right)}{\Pr\left(x_i = x_o \middle| Z_i^{EV} = 0\right)}$$

leads to

$$\Pr(S_i^* = 1 | x_i = x_o) \propto \Pr(S_i = 1 | x_i = x_o) \frac{\Pr\left(Z_i^{EV} = 1 \middle|\, x_i = x_o\right)}{\Pr\left(Z_i^{EV} = 0 \middle|\, x_i = x_o\right)}$$

# How does it work?

- Now we are able to estimate differences between NPS and PS

# Pseudo-weights derived by Elliott & Valliant (2017)

- Estimated pseudo−weight is inverse of estimated propensity

$$w_{i,EV} = d_i \frac{\widehat{\Pr}\left(Z_i^{EV} = 0 \middle| x_i = x_o\right)}{\widehat{\Pr}\left(Z_i^{EV} = 1 \middle| x_i = x_o\right)}$$

- Estimated $w_{i,EV}$ can be plugged into, for example, Hajek estimator $\sum_{i \in NP} w_{i,EV} y_i / \sum_{i \in NP} w_{i,EV}$ for estimating population mean of target variable $y$

# Summary: pseudo-weights Elliott & Valliant (2017)

- Fit model with dependent variable $Z_i^{EV} = 1$ for units from NPS, $Z_i^{EV} = 0$ for units from PS

$$O_i = \frac{\Pr(Z_i^{EV} = 1 | \boldsymbol{x}_i = \boldsymbol{x}_o)}{\Pr(Z_i^{EV} = 0 | \boldsymbol{x}_i = \boldsymbol{x}_o)}$$

- Estimate odds $O$ in NPS
- Pseudo-weights are given by

$$w_{i,EV} \propto \frac{d_i}{O_i}$$

# Our assumptions

1. For all units in population, $\Pr(i \in \mathrm{NPS})$ and $\Pr(i \in PS)$ are nonzero
2. Auxiliary variables $\boldsymbol{X}$ govern inclusion mechanism of NPS
3. Inclusion weights $d$ for inclusion in PS are available or can be computed for PS and for NPS
4. Sampling fractions of PS and NPS are small so that they do not overlap
5. Inclusion in NPS and PS is independent after conditioning on $\boldsymbol{X}$
4. We can identify overlapping units in PS and NPS

# Our assumptions

- Our fourth assumption (“we can identify overlapping units in PS and NPS”) is useful for countries with registers on the population

# Proposed method

- Target population $U$ is divided into three non-overlapping subpopulations: $U = A \cup B \cup C$ with
  - $A = \{i: S_i^* = S_i = 1\}$, units in both NPS and PS
  - $B = \{i: S_i^* + S_i = 1\}$, units in either NPS or PS
  - $C = \{i: S_i^* + S_i = 0\}$, units in neither NPS nor PS
- Within subpopulation $B$, define
  - $Z_i = 1$ if $(S_i^*, S_i) = (1,0)$
  - $Z_i = 0$ if $(S_i^*, S_i) = (0,1)$

# How does it work?

- We look at non-overlapping part of NPS and PS
- This enables us to quantify differences between NPS and PS

# Proposed method: the “easy” part

- We have

$$\Pr(S_i = 1, i \in B | \boldsymbol{x}_i = \boldsymbol{x}_o)$$
$$= \Pr(S_i = 1 | \boldsymbol{x}_i = \boldsymbol{x}_o) \Pr(S_i^* = 0 | S_i = 1, \boldsymbol{x}_i = \boldsymbol{x}_o)$$

- By design, inclusion probability of unit $i$ in PS does not depend on $S_i^*$ after conditioning on $\boldsymbol{x}_i$, so we also have

$$\Pr(S_i^* = 1, i \in B | \boldsymbol{x}_i = \boldsymbol{x}_o)$$
$$= \Pr(S_i^* = 1 | \boldsymbol{x}_i = \boldsymbol{x}_o) \Pr(S_i = 0 | \boldsymbol{x}_i = \boldsymbol{x}_o)$$

# Proposed method: the “hard” part

- We can derive

$$\Pr(S_i^* = 1, i \in B | \boldsymbol{x}_i = \boldsymbol{x}_o)$$

$$= \Pr(S_i = 1, i \in B | \boldsymbol{x}_i = \boldsymbol{x}_o) \frac{\Pr(Z_i = 1 | \boldsymbol{x}_i = \boldsymbol{x}_o, i \in B)}{\Pr(Z_i = 0 | \boldsymbol{x}_i = \boldsymbol{x}_o, i \in B)}$$

# Proposed method: the “hard” part

- We can derive

$$\Pr(S_i^* = 1, i \in B | \boldsymbol{x}_i = \boldsymbol{x}_o)$$

$$= \Pr(S_i = 1, i \in B | \boldsymbol{x}_i = \boldsymbol{x}_o) \frac{\Pr(Z_i = 1 | \, \boldsymbol{x}_i = \boldsymbol{x}_o, i \in B)}{\Pr(Z_i = 0 | \, \boldsymbol{x}_i = \boldsymbol{x}_o, i \in B)}$$

# How does it work?

- This expression enables us to quantify differences between NPS and PS

# Proposed method: the “hard” part

$$O_i = \frac{\Pr(Z_i = 1 |\, \boldsymbol{x}_i = \boldsymbol{x}_o, i \in B)}{\Pr(Z_i = 0 |\, \boldsymbol{x}_i = \boldsymbol{x}_o, i \in B)}$$

- Can be estimated by any model that is suitable for binary class probability estimation, e.g. logistic regression or some machine learning method

# Dependent and independent samples

- Independent samples: inclusion in NPS is independent of inclusion in PS
- Dependent samples: inclusion in NPS depends on inclusion in PS

# Proposed method for independent samples

- In this case we have

$$\Pr(S_i = 1, i \in B | \boldsymbol{x}_i = \boldsymbol{x}_o)$$
$$= \Pr(S_i = 1 | \boldsymbol{x}_i = \boldsymbol{x}_o) \Pr(S_i^* = 0 | \boldsymbol{x}_i = \boldsymbol{x}_o)$$
$$= \Pr(S_i^* = 0 | \boldsymbol{x}_i = \boldsymbol{x}_o) / d_i$$

# Proposed method for independent samples

- Combining everything we found we get

$$\widehat{\Pr}(S_i^* = 1 | \boldsymbol{x}_i = \boldsymbol{x}_o) = \frac{\hat{O}_i}{\hat{O}_i + d_i - 1}$$

where $O_i$ is estimated using

$$\frac{\Pr(Z_i = 1 |\, \boldsymbol{x}_i = \boldsymbol{x}_o, i \in B)}{\Pr(Z_i = 0 |\, \boldsymbol{x}_i = \boldsymbol{x}_o, i \in B)}$$

- Pseudo-weights are given by

$$w_{i,ind} = \frac{1}{\widehat{\Pr}(S_i^* = 1 | \boldsymbol{x}_i = \boldsymbol{x}_o)} = 1 + \frac{d_i - 1}{\hat{O}_i}$$

# Proposed method for dependent samples

In this case we have

$$\Pr(S_i^* = 1|\boldsymbol{x}_i = \boldsymbol{x}_o)$$
$$= \frac{\Pr(S_i = 1|\boldsymbol{x}_i = \boldsymbol{x}_o)}{\Pr(S_i = 0|\boldsymbol{x}_i = \boldsymbol{x}_o)} O_i \Pr(S_i^* = 0|S_i = 1, \boldsymbol{x}_i = \boldsymbol{x}_o)$$

# How does it work?

- For dependent samples we also need to estimate probability of being included in NPS given being included in PS

# Proposed method for dependent samples

- $\Pr(S_i^* = 0 | S_i = 1, \boldsymbol{x}_i = \boldsymbol{x}_o)$ can be modeled on units in PS
- For example logistic regression can be used:

$$\log\left(\frac{\Pr(S_i^* = 1 | S_i = 1, \boldsymbol{x}_i = \boldsymbol{x}_o)}{\Pr(S_i^* = 0 | S_i = 1, \boldsymbol{x}_i = \boldsymbol{x}_o)}\right) = \beta^T \boldsymbol{x}_i$$

- This leads to

$$\widehat{\Pr}(S_i^* = 0 | S_i = 1, \boldsymbol{x}_i = \boldsymbol{x}_o) = 1/(1 + \exp(\hat{\beta}^T x_i)) \equiv \hat{L}_i$$

- Pseudo-weights are given by

$$w_{i,dep} = \frac{1}{\widehat{\Pr}(S_i^* = 1 | \boldsymbol{x}_i = \boldsymbol{x}_o)} = \frac{d_i - 1}{\hat{O}_i \hat{L}_i}$$

# Summary method for independent samples

1. Remove overlapping units in PS and NPS
2. Fit model using non-overlapping units with dependent variable $Z = 1$ for units from NPS, $Z = 0$ for units from PS

$$O_i = \frac{\Pr(Z_i = 1 | \boldsymbol{x}_i = \boldsymbol{x}_o, i \in B)}{\Pr(Z_i = 0 | \boldsymbol{x}_i = \boldsymbol{x}_o, i \in B)}$$

3. Estimate odds *O* in NPS and calculate pseudo-weights by

$$w_{i,ind} = 1 + \frac{d_i - 1}{O_i}$$

# Summary method for dependent samples

- Steps 1 and 2 are the same as for independent samples

3. In PS, fit the model with dependent variable $S^*$

$$\log\left(\frac{\Pr(S_i^* = 1 | S_i = 1, x_i = x_o)}{\Pr(S_i^* = 0 | S_i = 1, x_i = x_o)}\right) = \beta^T x_i$$

$$\Pr(S_i^* = 0 | S_i = 1, x_i = x_o) = 1/(1 + \exp(\beta^T x_i)) \equiv L_i$$

4. Estimate $O$ and $L$ in NPS and calculate pseudo-weights:

$$w_{i,dep} = \frac{d_i - 1}{O_i L_i}$$

# Variance estimation

- We use pseudo-population bootstrapping to estimate variance of estimators based on estimated pseudo-weights
  - Pseudo-population bootstrapping is suitable for many target parameters, estimation models, and sampling designs
- Pseudo-populations are created from NPS with $[w_i]$ copies and bootstrap samples are drawn from it
- Bootstrap variance gives quality measure for estimates based on estimated pseudo-weights

# How does it work?

- We need to take into account that NPS is selective and needs to be corrected by means of pseudo-weights
- Standard bootstrap does not do this
- Using NPS and pseudo-weights, we construct pseudo-populations that resemble true population
- We use those pseudo-populations to replicate process of drawing NPS and PS and using proposed method on NPS and PS
  - We do this many times and calculate variance over replications

# Pseudo-population bootstrap: part 1

1. Estimate $\hat{O}_i$, and weights of NPS ($w_i = w_{i,ind}$)
2. Normalize weights by $w_i N / \sum_{i \epsilon NP} w_i$ to obtain $\sum_{i \epsilon NP} w_i = N$
3. Used controlled rounding to round $w_i$ to its ceiling with probability $w_i - \lfloor w_i \rfloor$ and to its floor otherwise to obtain $[w_i]$ such that$\sum_{i \epsilon NP} [w_i] = N$
4. Create pseudo-population by copying unit $i$ $[w_i]$ times

# Pseudo-population bootstrap: part 2

5. Draw bootstrap probability sample ($S_P$) from pseudo-population according to design of PS, with inclusion probabilities $1/d_i$
6. Draw bootstrap nonprobability sample ($S_{NP}$) from pseudo-population with inclusion probabilities $1/w_i$
7. Remove overlapping units in $S_P$ and $S_{NP}$ and then estimate $\hat{O}_i$ to estimate weights and target parameter for pair $S_P$ and $S_{NP}$
8. Repeat steps 5-7 for $R$ times to acquire $R$ estimates
9. Compute bootstrap variance of the $R$ estimates

# Pseudo-population bootstrap: part 3

- Estimated population mean for bootstrap sample $r$

$$\hat{\theta}_r = \frac{\sum_{i \in NP} w_i y_i}{\sum_{i \in NP} w_i}$$

- Bootstrap estimate

$$\bar{\theta} = \frac{\sum_{r=1}^{R} \hat{\theta}_r}{R}$$

- Bootstrap variance

$$\widehat{\mathrm{Var}}(\hat{\theta}) = \frac{\sum_{r=1}^{R} \left(\hat{\theta}_r - \bar{\theta}\right)^2}{R - 1}$$

# Simulation study

- We evaluated our approach in a simulation study and compared it with several other approaches
  - Naïve estimator
  - Without removing overlap $w_{ind,2}$
  - Other methods: Elliott and Valliant (2017); Chen, Li, and Wu (2020); Valliant and Dever (2011); Wang, Valliant, and Li (2021); Kim and Wang (2019)

# Simulation study

- We used registered data from Dutch Online Kilometer Registration, which contains around 6.7 million records of privately owned cars in the Netherlands in 2012
- Variables include registration year of car, engine type of car, age of car owner, and target variable $y$ mileage of car
- Population of simulation study is simple random sample without replacement of size 100,000 drawn from register
- Target parameter is population mean of $y$: $\bar{y} = \ 11741.79$

# Simulation study

- Both PS and NPS are drawn by fixed-size unequal probability sampling without replacement
- When drawing two dependent samples, units already drawn in PS are given smaller chance to be included in NPS

# Evaluation measures

- To evaluate performance of considered methods over $M = 1{,}000$ independent replications we computed:
  - Relative bias (in percentages)

$$\frac{1}{M}\sum_{m=1}^{M}\frac{y_m - \bar{y}}{\bar{y}} \times 100\%$$

  - Root mean square error (RMSE)

$$\mathrm{RMSE} = \frac{1}{M}\left(\sum_{m=1}^{M}(y_m - \bar{y})^2\right)^{1/2}$$

# Relative bias

| | $f_{NP}$ | $f_P$ | Naive | $w_{ind}$ | $w_{dep}$ | $w_{ind,2}$ | EV | CLW | VD | WVL | KW |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Ind | 0.05 | 0.01 | 20.22 | 0.12 | 0.10 | 1.38 | 0.45 | **-0.05** | 2.07 | 0.48 | 20.22 |
| | | 0.10 | 20.21 | **-0.00** | -0.00 | 1.46 | 0.76 | -0.02 | 2.11 | 0.77 | -0.20 |
| | 0.30 | 0.01 | 15.36 | **-0.01** | 0.02 | 3.94 | 0.50 | -0.23 | 12.16 | 0.52 | -1.48 |
| | | 0.10 | 15.35 | 0.04 | 0.04 | 4.55 | 1.25 | 0.03 | 5.69 | 1.28 | **-0.01** |
| | 0.50 | 0.01 | 11.43 | -0.04 | **0.03** | 4.06 | 0.37 | -0.33 | -0.91 | 0.38 | -0.48 |
| | | 0.10 | 11.45 | 0.01 | 0.01 | 4.43 | 0.85 | **-0.00** | 5.30 | 0.87 | -0.01 |
| Dep | 0.05 | 0.01 | 20.23 | 1.15 | 0.30 | 1.34 | 0.41 | **-0.12** | 2.00 | 0.44 | 0.47 |
| | | 0.10 | 19.69 | 1.24 | 0.37 | 1.45 | 0.77 | **0.05** | 2.11 | 0.77 | -2.36 |
| | 0.30 | 0.01 | 15.31 | 2.88 | **0.16** | 3.95 | 0.53 | -0.17 | 12.13 | 0.55 | 15.30 |
| | | 0.10 | 14.72 | 3.19 | 0.58 | 4.38 | 1.23 | **0.25** | 5.48 | 1.24 | -6.06 |
| | 0.50 | 0.01 | 11.37 | 2.13 | **-0.21** | 4.04 | 0.37 | -0.30 | -0.46 | 0.39 | 7.81 |
| | | 0.10 | 10.78 | 2.17 | **0.18** | 4.16 | 0.82 | 0.19 | 4.97 | 0.83 | -9.23 |

# RMSE

| | $f_{NP}$ | $f_P$ | Naive | $w_{ind}$ | $w_{dep}$ | $w_{ind,2}$ | EV | CLW | VD | WVL | KW |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Ind | 0.05 | 0.01 | 2379 | **246** | 248 | 285 | 249 | 341 | 380 | 253 | 2379 |
| | | 0.10 | 2377 | **237** | 238 | 278 | 243 | 245 | 332 | 243 | 321 |
| | 0.30 | 0.01 | 1805 | **122** | 146 | 475 | 150 | 240 | 1851 | 157 | 845 |
| | | 0.10 | 1803 | **73** | 77 | 538 | 163 | 90 | 671 | 166 | 106 |
| | 0.50 | 0.01 | 1343 | **98** | 158 | 486 | 141 | 229 | 2581 | 149 | 294 |
| | | 0.10 | 1345 | **46** | 61 | 521 | 112 | 73 | 623 | 115 | 65 |
| Dep | 0.05 | 0.01 | 2379 | 282 | **258** | 292 | 259 | 335 | 373 | 264 | 3446 |
| | | 0.10 | 2317 | 262 | **228** | 275 | 240 | 241 | 330 | 241 | 664 |
| | 0.30 | 0.01 | 1799 | 356 | **138** | 476 | 149 | 236 | 1837 | 156 | 1798 |
| | | 0.10 | 1729 | 380 | 98 | 517 | 160 | **95** | 646 | 162 | 735 |
| | 0.50 | 0.01 | 1335 | 266 | **138** | 483 | 139 | 225 | 2415 | 148 | 1326 |
| | | 0.10 | 1266 | 259 | **57** | 491 | 108 | 73 | 585 | 110 | 1092 |

# Simulation for variances estimates

- We set number of pseudo-populations $D$ equal to 1 and ran pseudo-population bootstrap 500 times
- For independent samples we also set $D$ equal to 10 and ran pseudo-population bootstrap 50 times to illustrate effect of multiple pseudo-populations
- (Non)probability bootstrap samples drawn by fixed-size random systematic sampling given known inclusion probability for PS or estimated propensities for NPS
- True variances estimated by means of Monte Carlo simulation

# Variance estimation

- Relative Bias (%) of the estimated variances and the Coverage Rate of the confidence intervals (%) of the variance estimates

| | | Ind | $D = 1$ | Ind | $D = 10$ | Dep | $D = 1$ |
|---|---|---|---|---|---|---|---|
| $f_{NP}$ | $f_P$ | RB | CR | RB | CR | RB | CR |
| 0.05 | 0.01 | -7.81 | 94.10 | 10.20 | 94.02 | 398.00 | 96.80 |
| | 0.10 | 11.61 | 96.40 | -1.75 | 94.76 | -2.98 | 95.10 |
| 0.30 | 0.01 | 2.58 | 94.70 | 1.61 | 94.51 | -2.00 | 93.60 |
| | 0.10 | 0.86 | 95.00 | 1.54 | 94.53 | 4.16 | 87.60 |
| 0.50 | 0.01 | 3.06 | 95.40 | -2.44 | 93.83 | -5.24 | 93.60 |
| | 0.10 | -6.91 | 93.80 | -4.16 | 94.02 | -2.05 | 93.60 |

# When can we apply method?

- All four assumptions are important ones that many be violated in practice
  1. For all units in population, $\Pr(i \in \text{NPS})$ and $\Pr(i \in PS)$ are nonzero
  2. Auxiliary variables $\boldsymbol{X}$ govern inclusion mechanism of NPS
  3. Inclusion weights $d$ for inclusion in PS are available or can be computed for PS and for NPS
  4. We can identify overlapping units in PS and NPS

# References


- General overviews
  - Beresewicz, M., R. Lehtonen, F. Reis, L. Di Consiglio & M. Karlberg (2018), *An Overview of Methods for Treating Selectivity in Big Data Sources*. Statistical working papers, Eurostat.
  - Cornesse, C., A.G. Blom, D. Dutwin, J.A. Krosnick, E.D. de Leeuw, S. Legleye, J. Pasek, D. Pennay, B. Phillips, J.W. Sakshaug, B. Struminskaya, B. & A. Wenz A. (2020), A Review of Conceptual Approaches and Empirical Evidence on Probability and Nonprobability Sample Survey Research. *Journal of Survey Statistics and Methodology* **8**(1), 4-36.
  - Rao, J. N. K. (2020), On Making Valid Inferences by Integrating Data from Surveys and Other Sources. *Sankhya B* **83**, 242–272
  - Beaumont, J.-F. & J.N.K. Rao (2022), Pitfalls of Making Inferences from Non-Probability Samples: Can Data Integration Through Probability Samples Provide Remedies? *The Survey Statistician* **83**, 11-22.

# References


- Pseudo-weighting
  - Elliott, M.R. & R. Valliant (2017), Inference for Nonprobability Samples. *Statistical Science* **32**(2), 249-264.
  - Liu, A.-C., S. Scholtus & T. de Waal (2022), Correcting Selection Bias in Big Data by Pseudo Weighting. *Journal of Survey Statistics and Methodology*, https://doi.org/10.1093/jssam/smac029.