Challenges and Strategies in Dealing with Non-Probability Survey Samples

Changbao Wu

Department of Statistics and Actuarial Science
University of Waterloo

December 6, 2023 ISTAT Workshop

Non-Probability Survey Samples

- What is a non-probability sample?

  A sample with unknown participation/inclusion/selection mechanisms and an unknown sampled population

- Examples of non-probability samples
  - Samples selected from web- or phone-panels
  - Volunteer based samples
  - Convenient samples
  - Incomplete administrative records
  - ... ...

Non-Probability Survey Samples

- Probability survey samples with large nonresponse rates are essentially non-probability samples
- Xiao-Li Meng: in the discussion of Wu (2022)

  There is no such thing as probability sample in real life!

- Responses from Wu (2022):

  For human populations, this is probably a defendable statement since any rigorous rules and precise procedures are almost surely as aspiration, not prescription.

  Probability samples, however, do exist in other fields such as business and establishment surveys, agricultural surveys, and natural resource inventory surveys.

Non-Probability Survey Samples

- We heard of people talking ...

  Non-probability samples are biased samples. They are difficult to handle.

- All non-iid samples are biased. Even probability samples are biased (unless it is a simple random sample).

- We are not worried about the biased nature of probability samples since the biases can be corrected by suitable weighting using the known sample inclusions probabilities.

  The HT Estimator!

  J.N.K. Rao (2005): The NHT estimator. (Narain, 1951; Horvitz and Thompson, 1952)

Non-Probability Survey Samples

- Three major challenges in dealing with non-probability samples:
  - the unknown sample participation/inclusion/selection mechanisms
  - the unknown sampled population
  - the dearth of auxiliary population information required for valid estimation and inference
- Where do we start? Assumptions, assumptions, ...

  All models are wrong, but some are useful. – George Box

The Two-Sample Framework

- The finite population $\mathcal{U} = \{1, 2, \cdots, N\}$ consists of N labelled units; associated with unit i are
  - auxiliary variables $\boldsymbol{x}_i$
  - study variable y_i (the variable of interest)

  The goal is to estimate $\mu_y = N^{-1} \sum_{i=1}^{N} y_i$ for the study variable y

- $\mathcal{S}_A$: A non-probability sample of size n_A from $\mathcal{U}$ with data

$$\{(\boldsymbol{x}_i, y_i), i \in \mathcal{S}_A\}$$

- An existing reference probability sample $\mathcal{S}_B$ containing information on $\boldsymbol{x}$ (but not on y) from the same target population

$$\{(\boldsymbol{x}_i, d_i^B), i \in \mathcal{S}_B\},$$

  where d_i^B are the design weights for the sample $\mathcal{S}_B$

Two Statistical Models

- A model q for participation probabilities (propensity scores)
  - Let $R_i = I(i \in \mathcal{S}_A)$ be the indicator variable for unit i being included in the non-probability sample $\mathcal{S}_A$
  - The participation probabilities (propensity scores) are defined as

$$\pi_i^A = P(R_i = 1 \mid \boldsymbol{x}_i, y_i), \quad i = 1, 2, \cdots, N$$

  - The model q determines the joint distribution of $\{(R_i, \boldsymbol{x}_i, y_i), i = 1, 2, \ldots, N\}$ over the target finite population

- A model ξ for the outcome regression of y given $\boldsymbol{x}$
  - The first two moments of the model

$$m_i = E_\xi(y_i \mid \boldsymbol{x}_i), \quad v_i = V_\xi(y_i \mid \boldsymbol{x}_i), \quad i = 1, 2, \ldots, N$$

  - A semiparametric model with specified form $m_i = m(\boldsymbol{x}_i, \boldsymbol{\beta})$
  - A linear regression model: $m_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$

Two Key Assumptions for the participation Mechanism

A1 The participation indicator R_i and the study variable y_i are independent given the set of covariates $\boldsymbol{x}_i$:

$$(R_i \perp\!\!\!\perp y_i) \mid \boldsymbol{x}_i$$

(The ignorability assumption: similar to "missing-at-random" (MAR) for missing data)

A2 All units have non-zero participation probabilities:

$$\pi_i^A > 0, \quad i = 1, 2, \cdots, N$$

(The positivity assumption)

A Data Integration Problem

- A non-probability sample with information on $(\boldsymbol{x}, y)$

$$\{(\boldsymbol{x}_i, y_i),\ i \in \mathcal{S}_A\}$$

- An existing reference probability sample with information on $\boldsymbol{x}$

$$\{(\boldsymbol{x}_i, d_i^B)\ i \in \mathcal{S}_B\}$$

- The requirement that $\boldsymbol{x}$ is observed for both $\mathcal{S}_A$ and $\mathcal{S}_B$ can be problematic
- Data integration for valid statistical inference:
  - Each of $\mathcal{S}_A$ and $\mathcal{S}_B$ alone does not lead to valid inference on μ_y
  - Combine information from $\mathcal{S}_A$ and $\mathcal{S}_B$ for valid inference on μ_y

Inverse Probability Weighted (IPW) Estimators

- Let $\hat{\pi}_i^A, i \in \mathcal{S}_A$ be the estimated participation probabilities
- The IPW estimator of μ_y is given by

$$\hat{\mu}_{yIPW} = \frac{1}{\hat{N}} \sum_{i \in \mathcal{S}_A} \frac{y_i}{\hat{\pi}_i^A}$$

  where $\hat{N} = \sum_{i \in \mathcal{S}_A} (\hat{\pi}_i^A)^{-1}$
- The IPW estimator is an application of the HT estimator and the Hájek estimator from survey sampling
- The performance of $\hat{\mu}_{yIPW}$ depends on the behaviour of the estimated participation probabilities $\hat{\pi}_i^A$

Methods for Estimating Participation Probabilities

- Parametric methods
  - The pooled sample method (Valliant and Dever, 2011)
  - The pseudo maximum likelihood method (Chen, Li and Wu, 2020)
  - The two-step method (Wang, Valliant and Li, 2021)
- Nonparametric methods (Wu, 2022)
- Tree-based methods (Chu and Beaumont, 2019)

The Method of Chen, Li and Wu (2020)

- Consider a parametric model $\pi_i^A = \pi(\boldsymbol{x}_i, \boldsymbol{\alpha})$
- An example: the logistic regression model

$$\pi(\boldsymbol{x}_i, \boldsymbol{\alpha}) = \frac{\exp(\boldsymbol{x}_i^T \boldsymbol{\alpha})}{1 + \exp(\boldsymbol{x}_i^T \boldsymbol{\alpha})} = 1 - \frac{1}{1 + \exp(\boldsymbol{x}_i^T \boldsymbol{\alpha})}$$

- The full-likelihood function

$$L(\boldsymbol{\alpha}) = \prod_{i=1}^{N} (\pi_i^A)^{R_i} (1 - \pi_i^A)^{1-R_i}$$

- The full log-likelihood function

$$\begin{aligned} \ell(\boldsymbol{\alpha}) &= \sum_{i=1}^{N} \left\{ R_i \log \pi_i^A + (1 - R_i) \log (1 - \pi_i^A) \right\} \\ &= \sum_{i \in \mathcal{S}_A} \log \left\{ \frac{\pi(\boldsymbol{x}_i, \boldsymbol{\alpha})}{1 - \pi(\boldsymbol{x}_i, \boldsymbol{\alpha})} \right\} + \sum_{i=1}^{N} \log \{1 - \pi(\boldsymbol{x}_i, \boldsymbol{\alpha})\} \end{aligned}$$

The Method of Chen, Li and Wu (2020)

- The pseudo log-likelihood function

$$\ell_1(\boldsymbol{\alpha}) = \sum_{i \in \mathcal{S}_A} \log \left\{ \frac{\pi(\boldsymbol{x}_i, \boldsymbol{\alpha})}{1 - \pi(\boldsymbol{x}_i, \boldsymbol{\alpha})} \right\} + \sum_{i \in \mathcal{S}_B} d_i^B \log \{1 - \pi(\boldsymbol{x}_i, \boldsymbol{\alpha})\}$$

- Under the probability sampling design, p, for sample $\mathcal{S}_B$:

$$E_p\left\{\ell_1(\boldsymbol{\alpha})\right\} = \ell(\boldsymbol{\alpha})$$

- The pseudo log-likelihood function $\ell_1(\boldsymbol{\alpha})$ is valid replacement of the true log-likelihood function $\ell(\boldsymbol{\alpha})$

The Method of Chen, Li and Wu (2020)

- The pseudo score functions, defined as $\boldsymbol{U}_1(\boldsymbol{\alpha}) = \partial \ell_1(\boldsymbol{\alpha})/\partial \boldsymbol{\alpha}$, are given by

$$\boldsymbol{U}_1(\boldsymbol{\alpha}) = \sum_{i \in \mathcal{S}_A} \frac{\pi_i'(\boldsymbol{\alpha})}{\pi(\boldsymbol{x}_i, \boldsymbol{\alpha})\{1 - \pi(\boldsymbol{x}_i, \boldsymbol{\alpha})\}} - \sum_{i \in \mathcal{S}_B} d_i^B \frac{\pi_i'(\boldsymbol{\alpha})}{1 - \pi(\boldsymbol{x}_i, \boldsymbol{\alpha})}$$

  where $\pi_i'(\boldsymbol{\alpha}) = \partial \pi(\boldsymbol{x}_i, \boldsymbol{\alpha})/\partial \boldsymbol{\alpha}$
- The pseudo score functions are unbiased under the joint randomization of the participation model p and the survey design p (for $\mathcal{S}_B$:

$$E_{qp}\{\boldsymbol{U}_1(\boldsymbol{\alpha}_0)\} = \mathbf{0}$$

  where $\boldsymbol{\alpha}_0$ is the true value of the model parameters $\boldsymbol{\alpha}$
- Score functions are optimal among all unbiased estimating functions (Godambe, 1960)

The Method of Valliant and Dever (2011)

- Consider the pooled sample: $\mathcal{S}_A \cup \mathcal{S}_B$
- Model $\{D_i, i \in \mathcal{S}_A \cup \mathcal{S}_B\}$ where

$$D_i = 1 \text{ if } i \in \mathcal{S}_A \,; \qquad D_i = 0 \text{ if } i \in \mathcal{S}_B$$

- Note: the participation model q does not lead to a meaningful model on the D_i's
- The full log-likelihood function

$$\ell(\boldsymbol{\alpha}) = \sum_{i \in \mathcal{S}_A} \log\{\pi(\boldsymbol{x}_i, \boldsymbol{\alpha})\} + \sum_{i \in \mathcal{U} \setminus \mathcal{S}_A} \log\{1 - \pi(\boldsymbol{x}_i, \boldsymbol{\alpha})\}$$

- Estimate $\sum_{i \in \mathcal{U} \setminus \mathcal{S}_A} \log\{1 - \pi(\boldsymbol{x}_i, \boldsymbol{\alpha})\}$ using data from $\mathcal{S}_B$

The Method of Valliant and Dever (2011)

- The objective function of Valliant and Dever (2011)

$$\ell_2(\boldsymbol{\alpha}) = \sum_{i \in \mathcal{S}_A} \log\{\pi(\boldsymbol{x}_i, \boldsymbol{\alpha})\} + \sum_{i \in \mathcal{S}_B} w_i \log\{1 - \pi(\boldsymbol{x}_i, \boldsymbol{\alpha})\}$$

  where w_i are re-scaled from d_i^B such that $\sum_{i \in \mathcal{S}_B} w_i = \hat{N}_B - n_A$ and $\hat{N}_B = \sum_{i \in \mathcal{S}_B} d_i^B$
- The functions $\boldsymbol{U}_2(\boldsymbol{\alpha}) = \partial \ell_2(\boldsymbol{\alpha}) / \partial \boldsymbol{\alpha}$ are given by

$$\boldsymbol{U}_2(\boldsymbol{\alpha}) = \sum_{i \in \mathcal{S}_A} \frac{\pi_i'(\boldsymbol{\alpha})}{\pi(\boldsymbol{x}_i, \boldsymbol{\alpha})} - \left(1 - \frac{n_A}{\hat{N}_B}\right) \sum_{i \in \mathcal{S}_B} d_i^B \frac{\pi_i'(\boldsymbol{\alpha})}{1 - \pi(\boldsymbol{x}_i, \boldsymbol{\alpha})}$$

- We only have $E_{qp}\{\boldsymbol{U}_2(\boldsymbol{\alpha}_0)\} \doteq \boldsymbol{0}$ under two scenarios
  - $\mathcal{S}_A$ is a simple random sample from the target population
  - The sampling fraction n_A/N is very small (i.e., $n_A/N = o(1)$)

The Method of Wang, Valliant and Li (2021)

- A method for correcting biases in Valliant and Dever (2011)
- Consider an augmented population: $\mathcal{S}_A^* \cup \mathcal{U}$
- Model $\{\delta_i, i \in \mathcal{S}_A^* \cup \mathcal{U}\}$ where

$$\delta_i = 1 \;\text{ if }\; i \in \mathcal{S}_A^* \;; \qquad \delta_i = 0 \;\text{ if }\; i \in \mathcal{U}$$

- The authors argue that $\pi_i^A = p_i/(1-p_i)$ where

$$\pi_i^A = P(i \in | \; \mathcal{U}) \quad \text{and} \quad p_i = P(i \in \mathcal{S}_A^* \mid \mathcal{S}_A^* \cup \mathcal{U})$$

- Note: the participation model q does not lead to a meaningful model on the δ_i's

The Method of Wang, Valliant and Li (2021)

- The objective function

$$\ell_3(\alpha) = \sum_{i \in \mathcal{S}_A} \log\left\{\frac{\pi(\boldsymbol{x}_i, \boldsymbol{\alpha})}{1 + \pi(\boldsymbol{x}_i, \boldsymbol{\alpha})}\right\} - \sum_{i \in \mathcal{S}_B} d_i^B \log\{1 + \pi(\boldsymbol{x}_i, \boldsymbol{\alpha})\}$$

- Note: $E_p\{\ell_3(\alpha)\} \neq \ell(\alpha)$, not a likelihood-based objective function
- The functions $\boldsymbol{U}_3(\boldsymbol{\alpha}) = \partial \ell_3(\boldsymbol{\alpha}) / \partial \boldsymbol{\alpha}$ are given by

$$\boldsymbol{U}_3(\boldsymbol{\alpha}) = \sum_{i \in \mathcal{S}_A} \frac{\pi_i'(\boldsymbol{\alpha})}{\pi(\boldsymbol{x}_i, \boldsymbol{\alpha})\{1 + \pi(\boldsymbol{x}_i, \boldsymbol{\alpha})\}} - \sum_{i \in \mathcal{S}_B} d_i^B \frac{\pi_i'(\boldsymbol{\alpha})}{1 + \pi(\boldsymbol{x}_i, \boldsymbol{\alpha})}$$

- The result $E_{qp}\{\boldsymbol{U}_3(\boldsymbol{\alpha}_0)\} = \boldsymbol{0}$ holds for general cases
- Wang, Valliant and Li (2021) can be viewed as a special case of estimating equations based methods, among them the score functions are optimal (Godambe, 1960)

Nonparametric Estimation of Participation Probabilities

- The participation probabilities

$$\pi_i^A = P(R_i = 1 \mid \boldsymbol{x}_i) = E_q(R_i \mid \boldsymbol{x}_i) = \pi(\boldsymbol{x}_i)$$

  are the conditional mean function of R given $\boldsymbol{x}$
- The "standard" Nadaraya-Watson kernel regression estimator of $\pi(\boldsymbol{x})$ is given by

$$\tilde{\pi}(\boldsymbol{x}) = \frac{\sum_{j=1}^{N} K_h(\boldsymbol{x} - \boldsymbol{x}_j) R_j}{\sum_{j=1}^{N} K_h(\boldsymbol{x} - \boldsymbol{x}_j)}$$

- The nonparametric kernel regression estimator of the propensity scores is given by (Yuan et al., 2023)

$$\hat{\pi}_i^A = \hat{\pi}(\boldsymbol{x}_i) = \frac{\sum_{j \in \mathcal{S}_A} K_h(\boldsymbol{x}_i - \boldsymbol{x}_j)}{\sum_{j \in \mathcal{S}_B} d_j^B K_h(\boldsymbol{x}_i - \boldsymbol{x}_j)}, \qquad i \in \mathcal{S}_A$$

Model-based Prediction (MP)

- Two "model-based prediction estimators" for $\mu_y = N^{-1}\sum_{i=1}^N y_i$

$$\tilde{\mu}_{yMP1} = \frac{1}{N}\sum_{i=1}^{N} \hat{m}_i\,, \qquad \tilde{\mu}_{yMP2} = \frac{1}{N}\Big\{\sum_{i\in\mathcal{S}_A}(y_i - \hat{m}_i) + \sum_{i=1}^{N}\hat{m}_i\Big\}$$

where $\hat{m}_i$ is an estimate for $m_i = E_\xi(y_i \mid \boldsymbol{x}_i)$

- Two "practical" model-based prediction estimators for μ_y

$$\hat{\mu}_{yMP1} = \frac{1}{N}\sum_{i\in\mathcal{S}_B} d_i^B\hat{m}_i\,, \qquad \hat{\mu}_{yMP2} = \frac{1}{N}\Big\{\sum_{i\in\mathcal{S}_A}(y_i - \hat{m}_i) + \sum_{i\in\mathcal{S}_B} d_i^B\hat{m}_i\Big\}$$

The so-called Mass-Imputation estimators (Kim et al., 2021)

- Under a linear model (with an intercept), we have

$$\hat{m}_i = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}} \quad \text{and} \quad \sum_{i\in\mathcal{S}_A}(y_i - \hat{m}_i) = 0$$

Doubly Robust (DR) Estimators

- The IPW estimators are a general tool for any y
- The MP estimators are y-specific, and require a model ξ on $y \mid \boldsymbol{x}$
- The "standard" doubly robust estimator of μ_y

$$\tilde{\mu}_{DR} = \frac{1}{N} \sum_{i \in \mathcal{S}_A} \frac{y_i - \hat{m}_i}{\hat{\pi}_i^A} + \frac{1}{N} \sum_{i=1}^{N} \hat{m}_i$$

- The doubly robust estimator of Chen et al. (2020)

$$\hat{\mu}_{DR2} = \frac{1}{\hat{N}^A} \sum_{i \in \mathcal{S}_A} \frac{y_i - \hat{m}_i}{\hat{\pi}_i^A} + \frac{1}{\hat{N}^B} \sum_{i \in \mathcal{S}_B} d_i^B \hat{m}_i$$

- The estimator $\hat{\mu}_{DR2}$ is consistent if one of the two models, q on $(R_i \mid \boldsymbol{x}_i)$ and ξ on $(y_i \mid \boldsymbol{x}_i)$, is correctly specified
- The concept of double robustness is rooted in model-assisted estimation in survey sampling (Cassel et al., 1976)

Calibration-based Methods

- The pseudo maximum likelihood estimator $\hat{\boldsymbol{\alpha}}$ for a parametric form $\pi_i^A = \pi(\boldsymbol{x}_i, \boldsymbol{\alpha})$ is the solution to the pseudo score equations
- Estimating equations based approach with the assumed parametric form $\pi_i^A = \pi(\boldsymbol{x}_i, \boldsymbol{\alpha})$: The estimator $\hat{\alpha}$ solves

$$\boldsymbol{G}(\boldsymbol{\alpha}) = \sum_{i \in \mathcal{S}_A} \frac{\boldsymbol{h}(\boldsymbol{x}_i, \boldsymbol{\alpha})}{\pi(\boldsymbol{x}_i, \boldsymbol{\alpha})} - \sum_{i \in \mathcal{S}_B} d_i^B \boldsymbol{h}(\boldsymbol{x}_i, \boldsymbol{\alpha}) = \boldsymbol{0}$$

  with a user-specified $\boldsymbol{h}(\boldsymbol{x}_i, \boldsymbol{\alpha})$
- The pseudo maximum likelihood method of Chen et al. (2020) corresponds to $\boldsymbol{h}(\boldsymbol{x}_i, \boldsymbol{\alpha}) = \pi_i'(\boldsymbol{\alpha})/\{1 - \pi(\boldsymbol{x}_i, \boldsymbol{\alpha})\}$
- The method of Wang et al. (2021) corresponds to $\boldsymbol{h}(\boldsymbol{x}_i, \boldsymbol{\alpha}) = \pi_i'(\boldsymbol{\alpha})/\{1 + \pi(\boldsymbol{x}_i, \boldsymbol{\alpha})\}$
- Consistency of estimating equations based estimator $\hat{\alpha}$ is (loosely) argued through $E_{qp}\{\boldsymbol{G}(\boldsymbol{\alpha}_0)\} = \boldsymbol{0}$

The Calibrated IPW Estimator

- The estimating functions based method becomes a calibration method if we choose $\boldsymbol{h}(\boldsymbol{x}_i, \boldsymbol{\alpha}) = \boldsymbol{x}_i$:

$$\sum_{i \in \mathcal{S}_A} \frac{\boldsymbol{x}_i}{\pi(\boldsymbol{x}_i, \boldsymbol{\alpha})} = \sum_{i \in \mathcal{S}_B} d_i^B \boldsymbol{x}_i \quad \left(\text{or} \quad \sum_{i=1}^{N} \boldsymbol{x}_i\right) \tag{1}$$

  where $\boldsymbol{x}$ and $\boldsymbol{\alpha}$ have the same dimensions

- The method leads to the so-called calibrated IPW estimator (Chen et al., 2020; Rao, 2021; Beaumont and Rao, 2021; Chen et al., 2023)

$$\hat{\mu}_{yIPW} = \frac{1}{\hat{N}} \sum_{i \in \mathcal{S}_A} \frac{y_i}{\hat{\pi}_i^A},$$

  where $\hat{\pi}_i^A = \pi(\boldsymbol{x}_i, \hat{\boldsymbol{\alpha}})$ and $\hat{\boldsymbol{\alpha}}$ solves calibration equations in (1)

The Calibrated IPW Estimator

- The calibrated IPW estimator is approximately model-unbiased under a linear regression model ξ with $m_i = E(y_i \mid \boldsymbol{x}_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}$:

$$
\begin{aligned}
E_{\xi p}\left\{\frac{1}{N}\sum_{i\in\mathcal{S}_A}\frac{y_i}{\pi(\boldsymbol{x}_i,\hat{\boldsymbol{\alpha}})}\right\} &= E_p\left\{\frac{1}{N}\sum_{i\in\mathcal{S}_A}\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}}{\pi(\boldsymbol{x}_i,\hat{\boldsymbol{\alpha}})}\right\} \\
&= E_p\left(\frac{1}{N}\sum_{i\in\mathcal{S}_B} d_i^B \boldsymbol{x}_i\right)^T \boldsymbol{\beta} \\
&= \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i^T\boldsymbol{\beta} = E_\xi(\mu_y)
\end{aligned}
$$

- The calibrated IPW estimator is doubly robust under a linear regression model
- The calibrated IPW estimator does not require the estimation of the regression coefficients $\boldsymbol{\beta}$

The Calibrated IPW Estimator

- The "standard" two-sample framework requires all auxiliary variables $\boldsymbol{x}$ be available in both $\mathcal{S}_A$ and $\mathcal{S}_B$
- A research problem:

  How to combine auxiliary information from two (or more) reference probability samples as well as information from census?

- The calibration-based approach, with $\hat{\alpha}$ solving

$$\sum_{i \in \mathcal{S}_A} \frac{\boldsymbol{x}_i}{\pi(\boldsymbol{x}_i, \boldsymbol{\alpha})} = \sum_{i \in \mathcal{S}_B} d_i^B \boldsymbol{x}_i \quad \left(\text{or} \quad \sum_{i=1}^{N} \boldsymbol{x}_i \right),$$

  allows components of the "population controls" $\sum_{i=1}^{N} \boldsymbol{x}_i$ to be estimated from different reference probability samples or from census

The Calibrated IPW Estimator

- Need an iterative procedure for solving

$$\boldsymbol{G}(\boldsymbol{\alpha}) = \sum_{i \in \mathcal{S}_A} \frac{\boldsymbol{x}_i}{\pi(\boldsymbol{x}_i, \boldsymbol{\alpha})} - \sum_{i \in \mathcal{S}_B} d_i^B \boldsymbol{x}_i = \boldsymbol{0}$$

- Assume $\pi(\boldsymbol{x}_i, \boldsymbol{\alpha}) = g(\boldsymbol{x}_i^T \boldsymbol{\alpha})$ for some monotone increasing smooth inverse link function $g(\cdot)$
- The "Hessian matrix" is given by

$$\boldsymbol{H}(\boldsymbol{\alpha}) = \frac{\partial}{\partial \boldsymbol{\alpha}} \boldsymbol{G}(\boldsymbol{\alpha}) = -\sum_{i \in \mathcal{S}_A} \frac{g'(\boldsymbol{x}_i^T \boldsymbol{\alpha})}{\{g(\boldsymbol{x}_i^T \boldsymbol{\alpha})\}^2} \boldsymbol{x}_i \boldsymbol{x}_i^T ,$$

- The matrix $\boldsymbol{H}(\boldsymbol{\alpha})$ is negative definite, as long as $\{\boldsymbol{x}_i, i \in \mathcal{S}_A\}$ is of full rank
- The Newton-Raphson procedure is guaranteed to converge

1. Settings and Assumptions
2. Estimation of Participation Probabilities
3. Calibration and Doubly Robust Estimation
4. **Poststratification**
5. Undercoverage
6. Additional Remarks

A Simple Scenario

- A major problem with IPW estimators: sensitive to small estimated participation probabilities
- Suppose $\boldsymbol{x} = (x_1, x_2)^T$, with x_1 having two levels and x_2 having three levels,
- There are a total of $K = 2 \times 3 = 6$ subpopulations defined by $\boldsymbol{x}$
- Within each subpopulation, the participation probabilities $\pi_i = P(i \in \mathcal{S}_A \mid \boldsymbol{x}_i) = \pi(\boldsymbol{x}_i)$ are a constant
- More generally, the components of $\boldsymbol{x}$ are all categorical or ordinal
- The $\mathcal{S}_A$ can be poststratified into $\mathcal{S}_A = \mathcal{S}_{A1} \cup \cdots \cup \mathcal{S}_{AK}$ corresponding to the cross-classification of sampled units using the combinations of levels of the $\boldsymbol{x}$ variables.
- Let n_k be the size of $\mathcal{S}_{Ak}$ and N_k be the size of the corresponding subpopulation

A Simple Scenario

- The participation probabilities

$$\pi_i^A = \pi(\boldsymbol{x}_i) = E_q(n_k)/N_k \quad \text{for} \quad k \in \mathcal{S}_{Ak}$$

- The estimated participation probabilities $\hat{\pi}_i^A = n_k/\hat{N}_k$ for $i \in \mathcal{S}_{Ak}$, where $\hat{N}_k$ is an estimate of N_k
- The IPW estimator $\hat{\mu}_{yIPW}$ reduces to the poststratified estimator

$$\hat{\mu}_{yPST} = \frac{1}{\hat{N}^A} \sum_{k=1}^{K} \sum_{i \in \mathcal{S}_{Ak}} \frac{y_i}{\hat{\pi}_i^A} = \sum_{k=1}^{K} \hat{W}_k \bar{y}_k$$

where $\bar{y}_k = n_k^{-1} \sum_{i \in \mathcal{S}_{Ak}} y_i$, $\hat{W}_k = \hat{N}_k/\hat{N}^A$ and $\hat{N}^A = \sum_{k=1}^{K} \hat{N}_k$

- Poststratify $\mathcal{S}_B$ based on $\boldsymbol{x}$: $\mathcal{S}_B = \mathcal{S}_{B1} \cup \cdots \cup \mathcal{S}_{BK}$
- Use $\hat{N}_k = \sum_{i \in \mathcal{S}_{Bk}} d_i^B$ and $\hat{N}^A = \sum_{i \in \mathcal{S}_B} d_i^B$

A General Procedure for Poststratification (Wu, 2022)

- The dimension of auxiliary variables $\boldsymbol{x}$ is not low and/or some components of $\boldsymbol{x}$ are continuous
- The first part of the procedure: Form homogeneous groups in $\mathcal{S}_A$ in terms of participation probabilities
  - Compute the initial $\hat{\pi}_i^A = \pi(\boldsymbol{x}_i, \hat{\alpha}), i \in \mathcal{S}_A$ based on an assumed parametric model, q.
  - Choose K such that $n_A = m_A K$, where m_A is an integer
  - Order the initial estimated participation probabilities

$$\hat{\pi}_{(1)}^A \leq \hat{\pi}_{(1)}^A \leq \cdots \leq \hat{\pi}_{(n_A)}^A$$

  - Let $\mathcal{S}_{A1}$ be the set of the first m_A units in the sequence, $\mathcal{S}_{A2}$ be the second m_A units in the sequence, and so on
  - The poststratified estimator of μ_y is computed as $\hat{\mu}_{yPST} = \sum_{k=1}^K \hat{W}_k \bar{y}_k$

A General Procedure for Poststratification (Wu, 2022)

- The second part of the procedure: Obtain the estimated stratum weights $\hat{W}_k$, $k = 1, 2, \cdots, K$ using the reference probability sample $\mathcal{S}_B$
  - Determine the strata boundaries as $b_k = \max\{\hat{\pi}_i^A : i \in \mathcal{S}_{Ak}\}$, $k = 1, 2, \cdots, K-1$, with $b_0 = 0$ and $b_K = 1$
  - Compute $\hat{\pi}_i = \pi(\boldsymbol{x}_i, \hat{\alpha})$, $i \in \mathcal{S}_B$.
  - Define $\mathcal{S}_{Bk} = \{i \mid i \in \mathcal{S}_B, b_{k-1} < \hat{\pi}_i \leq b_k\}$, $k = 1, 2, \cdots, K$.
  - Calculate $\hat{N}_k = \sum_{i \in \mathcal{S}_{Bk}} d_i^B$, $k = 1, 2, \cdots, K$.

  The estimated stratum weights $\hat{W}_k = \hat{N}_k / \hat{N}^B$, $\hat{N}^B = \sum_{i \in \mathcal{S}_B} d_i^B$
- The choice of K:
  - The balance between homogeneity of the units within each post-stratum (in terms of participation probabilities) and the stability of the poststratified estimator (in terms of the stratum sample sizes)
  - When n_A is small: $K = 5$
  - When n_A is not small: Choose $K \geq 5$ to ensure that $m_A \geq 30$

Undercoverage Problems

- Assumption **A1**: $(R_i \perp\!\!\!\perp y_i) \mid \boldsymbol{x}_i$
- Assumption **A1** may be reasonable if:

  All key factors and features that may characterize behaviours for participation in the survey are included in the sample data as part of the $\boldsymbol{x}$ variables for $\mathcal{S}_A$ (and are also available in the reference probability sample $\mathcal{S}_B$)

- Assumption **A2**: $\pi_i^A = P(R_i = 1 \mid \boldsymbol{x}_i, y_i) > 0$ for $i = 1, 2, \ldots, N$
- Violations of **A2** lead to undercoverage problems:

  If $\pi_i^A = 0$ for $i \in \mathcal{U}_0$, then the subpopulation $\mathcal{U}_0$ is not represented in any way by the sample $\mathcal{S}_A$.

Undercoverage Problems

- Violation of **A2** leads to invalid IPW-based estimation methods even if **A1** holds
- A basic result on inverse probability weighting for finite populations:

  The Horvitz-Thompson estimator

  $$\hat{\mu}_{yHT} = \frac{1}{N}\sum_{i\in\mathcal{S}}\frac{y_i}{\pi_i}$$

  is design-unbiased for μ_y if and only if $\pi_i > 0$ for all $i = 1, 2, \ldots, N$

Undercoverage Problems

- Violation of **A2** also leads to invalid model-based prediction methods even if **A1** holds
- Assumption **A1**, $(R_i \perp\!\!\!\perp y_i) \mid \boldsymbol{x}_i$, implies that

$$E_\xi(y_i \mid \boldsymbol{x}_i, R_i = 1) = E_\xi(y_i \mid \boldsymbol{x}_i) \tag{2}$$

  so the model parameters $\boldsymbol{\beta}$ in $m_i = E_\xi(y_i \mid \boldsymbol{x}_i) = m(\boldsymbol{x}_i, \boldsymbol{\beta})$ can be estimated using $\{(y_i, \boldsymbol{x}_i), i \in \mathcal{S}_A\}$ (with $R_i = 1$)
- However, equation (2) implicitly requires $P(R_i = 1) > 0$, which also requires $P(R_i = 1 \mid \boldsymbol{x}_i) > 0$

Undercoverage Problems

- The severity of undercoverage depending on

  (i) the size of the uncovered subpopulation $\mathcal{U}_0$
  (ii) the difference between $\mathcal{U}_0$ and the rest of the population

- Two possible scenarios of undercoverage (Chen et al., 2023):

  (i) stochastic undercoverage
  (ii) deterministic undercoverage

- The calibrated IPW estimator can be a useful tool for dealing with undercoverage if

  (i) a linear outcome regression model is suitable (no need to estimate β)
  (ii) population controls of auxiliary variables are reliable

- Post-stratification can also be a useful tool

Undercoverage - A Proposed Solution (Chen et al., 2023)

- Any full solutions to undercoverage problems require
  - A correct identification of

$$\mathcal{U}_0 = \{i \mid i \in \mathcal{U} \ \text{ and } \ \pi_i^A = 0\}, \qquad \mathcal{U}_1 = \{i \mid i \in \mathcal{U} \ \text{ and } \ \pi_i^A > 0\}$$

  - Additional information from $\mathcal{U}_1$ on y
- The concept of (an unspecified) accessibility function $\Phi(\boldsymbol{x})$, a convex function to equivalently define (through an unknown cut-off value, c) (Chen et al., 2023)

$$\mathcal{U}_0 = \{i \mid i \in \mathcal{U} \text{ and } \Phi(\boldsymbol{x}_i) \le c\}, \ \ \mathcal{U}_1 = \{i \mid i \in \mathcal{U} \text{ and } \Phi(\boldsymbol{x}_i) > c\}$$

- Identify $\mathcal{U}_0$ and $\mathcal{U}_1$ through a convex hull partition of $\mathcal{S}_B$

$$\mathcal{S}_B = \mathcal{S}_{B,0} \cup \mathcal{S}_{B,1}$$

- A new subsample from $\mathcal{S}_{B,0}$ with information on y

"Survey Design" for Non-Probability Samples

- Yes, "design" is part of a non-probability survey
- The first major design question: What types of auxiliary variables to be included for data collection

  Variables which might play a role in participation behaviour or have certain prediction power for the study variable need to be included. For human populations, demographic variables and social-economic indicators should be considered.

- The second major design question: What are the existing (large scale) probability survey samples from the same target population with information on auxiliary variables
- Quality and relevance of auxiliary variables are the keys to the success of a non-probability survey sample

Inferential Procedures: *Validity* vs *Efficiency*

- Statistical analysis with non-probability samples:
  - *Validity* refers to the consistency of the point estimators (or to a lesser extent: approximate unbiasedness)
  - *Efficiency* is measured by the asymptotic variance
  - *Validity* is the primary goal; *efficiency* is secondary
- Non-probability samples may have a very large sample size
- Large sample sizes are a double-edged sword:
  - When the inferential procedures are valid, large sample sizes lead to more efficient inference
  - When the estimators are biased, large sample sizes make the bias even more pronounced
  - Will a non-probability survey sample with a 80% sampling fraction always provide better estimation results than a small probability sample? (Meng, 2018)

Do We Still Need Probability Surveys?

- Non-probability samples do not fit into the traditional design-based or model-based inferential frameworks for probability survey samples
- Design-based theory for probability survey samples, however, plays a crucial role in the development of methodologies and strategies in dealing with non-probability samples
- The newfound role of probability survey samples:

  Valid and efficient statistical inference with non-probability samples requires auxiliary information from the target population. A few high quality national probability surveys with carefully designed survey variables can play a pivotal role in analysis of non-probability survey samples.

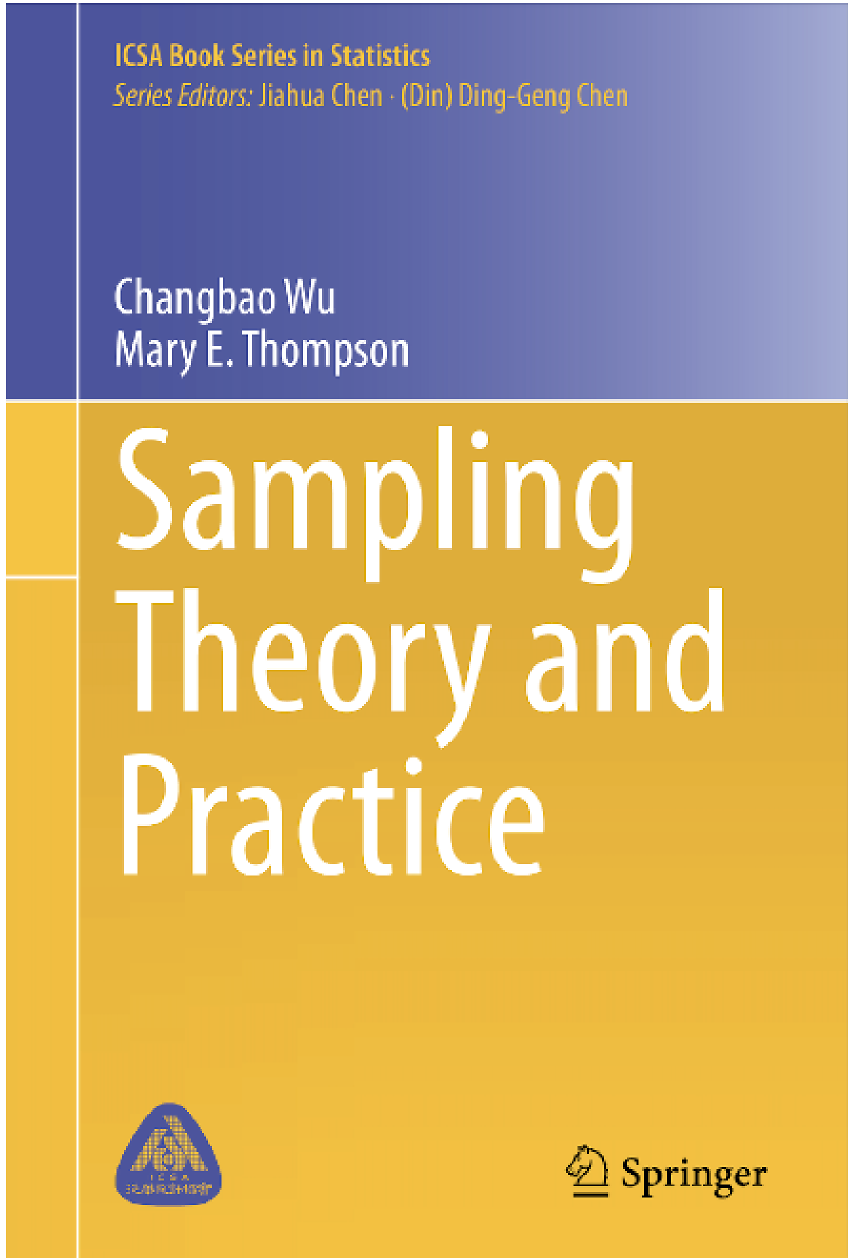
  – Wu, C. (2022)

Reference

- Beaumont, J.-F. and Rao, J.N.K. (2021). Pitfalls of Making Inferences from Non-probability Samples: Can Data Integration Through Probability Samples Provide Remedies? *The Survey Statistician*, **83**, 11–22.
- Cassel, C.M., Särndal, C.-E. and Wretman, J.H. (1976). Some Eesults on Generalized Difference Estimation and Generalized Regression Estimation for Finite Populations. *Biometrika*, 63, 615–620.
- Chu, K.C.K., and Beaumont, J.-F. (2019). The Use of Classification Trees to Reduce Selection Bias for a Non-probability Sample with Help from a Probability Sample. Proceedings of the Survey Methods Section of SSC.
- Chen, Y., Li, P. and Wu, C. (2020). Doubly Robust Inference with Non-probability Survey Samples. *JASA*, **115**, 2011–2021.
- Chen, Y., Li, P. and Wu, C. (2023). Dealing with Undercoverage for Non-probability Survey Samples. *Survey Methodology*, accepted.
- Kim, J. K., Park, S., Chen, Y. and Wu, C. (2021). Combining Non-probability and Probability Survey Samples Through Mass Imputation. *Journal of the Royal Statistical Society, Series A*, **184**, 941–963.

Reference

- Rao, J.N.K. (2005). Interplay Between Sample Survey Theory and Practice: An Appraisal. *Survey Methodology* **31**, 117–138.
- Rao, J. N. K. (2021). On Making Valid Inferences by Integrating Data from Surveys and Other Sources. *Sankhyā B*, **83**, 242–272.
- Valliant, R., and Dever, J.A. (2011). Estimating Propensity Adjustments for Volunteer Web Surveys. *Sociological Methods & Research*, **40**, 105–137.
- Wang, L., Valliant, R. and Li, Y. (2021). Adjusted logistic propensity weighting methods for population inference using nonprobability volunteer-based epidemiologic cohorts. *Statistics in Medicine*, **40**, 5237–5250.
- Wu, C. (2022). Statistical Inference with Non-probability Survey Samples (With Discussion). *Survey Methodology*, **48**, 283–311.
- Wu, C. and Thompson, M.E. (2020). *Sampling Theory and Practice*. Springer.
    - Chapter 9: Validity and efficiency for missing data analysis
    - Chapter 17: Analysis of non-probability survey samples


ICSA Book Series in Statistics
Series Editors: Jiahua Chen · (Din) Ding-Geng Chen
Changbao Wu
Mary E. Thompson
Sampling Theory and Practice
Springer