Uncertainty-based analysis for non-probability samples

3rd Workshop on Methodologies for Official Statistics - Session *Data, data science and official statistics*

Pier Luigi Conti[1]
Daniela Marella[2]

[1]Dipartimento di Scienze Statistiche - Sapienza Università di Roma
[2]Dipartimento di Scienze Sociali ed Economiche - Sapienza Università di Roma

December 3, 2024

Nonprobability samples - 1

- Probability sampling: a (usually non-informative) sampling design is constructed on the basis of design variables known for all population units.
- Each population unit possesses a *known*, positive probability of being selected (inclusion probability).
- If population units have different inclusion probabilities, there is actually *selection bias*.
- Selection bias can be removed by *weighting* sampled units. A major role is played by the Inverse Probability Weighting (IPW) principle, consisting in giving each unit a weight equal to the reciprocal of its inclusion probability.

Nonprobability samples - 2

- *Nonprobability samples*: involves a certain degree of arbitrariness in the unit selection process.
- Inclusion probabilities are *unknown*
- It is not generally possible to remove selection bias through the IPW principle.
- The (unknown) selection process is frequently *selective* w.r.t. the target population: *inclusion probabilities may depend on the character of interest*.
- Consequence: estimates constructed through non-probability samples may be severely biased.

Symbols used - 1

- $\mathcal{U}_N = \{u_1, \ldots, u_N\}$ (finite) population of N units.
- A, B two independent samples.
- A *probability sample*, drawn according to a known, non-informative design.
- B *non-probability sample*.
- $\mathcal{Y}$ *study variable*, taking value y_i over unit u_i.
- $\boldsymbol{\mathcal{X}} = (\mathcal{X}_1, \ldots, \mathcal{X}_p)^T$ vector of p *auxiliary variables*, taking value $\mathbf{x}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{ip})$ over unit u_i.
- The values $(y_i, \mathbf{x}_i)$ are observed on B.
- The values $\mathbf{x}_i$ are observed on A.

Symbols used - 2

- D_i sample membership indicator of unit u_i for sample A.
- $\pi_i^A = E[D_i]$ inclusion probability of unit u_i (*known* for all units is A).
- δ_i sample membership indicator of unit u_i for sample B.
- $p_N(y, \mathbf{x}) = \sum \mathbb{I}_{(\mathbf{x}_i = \mathbf{x})} \mathbb{I}_{(y_i = y)} / N$ joint population probability mass function (ppmf).
- $p_N(y) = \sum \mathbb{I}_{(y_i = y)} / N$, $p_N(\mathbf{x}) = \sum \mathbb{I}_{(\mathbf{x}_i = \mathbf{x})} / N$ marginal ppmfs.
- $p_N(y|\mathbf{x}) = p_N(y, \mathbf{x}) / p_N(\mathbf{x})$ conditional ppmf.
- $p(y, \mathbf{x})$, $p(y)$, $p(\mathbf{x})$, $p(y|\mathbf{x})$ superpopulation joint, marginal, and conditional probability functions (spmf).

Main approaches in the presence of a reference survey - 1

- *Propensity score adjustment*. The probability of being included in the non-probability sample B is estimated from sample A through the covariates $\mathcal{X}$ (pseudo-inclusion probabilities). Information on $\mathcal{X}$ in sample B can be calibrated with that estimated from the probability sample A (Kott (2006), Disogra (2011)). After having estimated inclusion probabilities, design-based inference can be used for point estimates.
- *Mass imputation*. Models are fitted to the non-probability sample B to predict the response variable $\mathcal{Y}$ for units in the reference survey A; cfr., among the others, Kim (2021).

Basic assumptions - 1

- $(\delta_i,\ Y_i,\ \mathbf{X}_i)$ *i.i.d.* r.v.s $(i = 1, \dots, N$.
- Consequence: $(Y_i,\ \mathbf{X}_i)$ *i.i.d.* r.v.s $(i = 1, \dots, N$.
- Y is a *discrete* r.v., taking values y^j, $j = 1, \dots, J$.
- $\mathbf{X}$ is a *discrete* r.v., taking values $\mathbf{x}^h$, $h = 1, \dots, H$.
- δ_is are observed in sample A (assumption similar to Kim and Wang (2019), Marella (2023).

Absence of identifiability - 1

Main quantities of interest (to be estimated): joint and conditional spmfs

$$p(y, \mathbf{x}), \;\; p(y|\mathbf{x}) = \frac{p(y, \mathbf{x})}{p(\mathbf{x})}. \tag{1}$$

Since the estimation of $p(y, \mathbf{x})$ is essentially equivalent to the estimation of $p(y|\mathbf{x})$, in the sequel we will focus on the latter.

Absence of identifiability - 2

The conditional spmf $p(y|\mathbf{x})$ can be written as

$$
\begin{aligned}
p(y|\mathbf{x}) &= p(\delta = 1|\mathbf{x}) \times \underbrace{p(y|\mathbf{x},\, \delta = 1)}_{\textit{sample distribution}} \\
&+ p(\delta = 0|\mathbf{x}) \times \underbrace{p(y|\mathbf{x},\, \delta = 0)}_{\textit{sample complement distribution}}
\end{aligned}
$$

with

$$
\begin{aligned}
p(y|\mathbf{x},\, \delta = 1) &= \frac{p(\delta = 1|y,\, \mathbf{x})}{p(\delta = 1|\mathbf{x})} p(y|\mathbf{x}) \\
p(y|\mathbf{x},\, \delta = 0) &= \frac{p(\delta = 0|y,\, \mathbf{x})}{p(\delta = 0|\mathbf{x})} p(y|\mathbf{x})
\end{aligned}
$$

Absence of identifiability - 3

- $p(\delta = 1|\mathbf{x})$ is identifiable (estimable) from sample A, because δ_is are observed for units in A.
- $p(\delta = 0|\mathbf{x})$ is identifiable (estimable) from sample A, because δ_is are observed for units in A.
- $p(y|\mathbf{x}, \delta = 1)$ is identifiable (estimable) from sample B.
- $p(y|\mathbf{x}, \delta = 0)$ is not identifiable (estimable).

$p(y|\mathbf{x}, \delta = 0)$ is identifiable if

Remark 1: The non identifiability of the spmf $p(y|\mathbf{x})$ comes from the uncertainty on the selection mechanism having generated sample B.

Remark 2: If $p(\delta = 1|y, \mathbf{x}) = p(\delta = 1|\mathbf{x})$ (*non-informative* selection mechanism for B), then $p(y|\mathbf{x})$, $p(y|\mathbf{x}, \delta = 1)$, $p(y|\mathbf{x}, \delta = 0)$ would *coincide*. In this case we actually get identifiability.

Absence of identifiability - 4

Identification region for $p(y|x)$

$$H[p(y|\mathbf{x})] = \{p(\delta = 1|\mathbf{x})p(y|\mathbf{x}, \delta = 1) + p(\delta = 0|\mathbf{x})\gamma, \gamma \in \Gamma_{\mathbf{x}y}\}$$

where

$$\Gamma_{\mathbf{x}y} = \left\{ p(\delta = 1|\mathbf{x}, y) : p(\delta = 1|\mathbf{x}) = \sum_{y} p(\delta = 1|\mathbf{x}, y)p(y|\mathbf{x}) \right\}.$$

can be interpreted as the *class of all possible sampling designs that could have generated the non-probability sample B*.

Uncertainty and its measure - 1

- The non-identifiability of the spmf $p(y|\mathbf{x})$ comes from the *uncertainty* on the selection mechanism having generated B.
- The larger the class of plausible sampling designs $\Gamma_{\mathbf{x}y}$, the larger the class of plausible spmf for $Y|\mathbf{X}$ (*i.e.* $H[p(y|\mathbf{x})]$) and the larger the uncertainty on the data generating model $p(y|\mathbf{x})$.

Problem: How to measure the uncertainty for $p(y|\mathbf{x})$?

Uncertainty and its measure - 2

	$\delta = 0$	$\delta = 1$	Total
y^1	$p(y^1, \delta = 0\|\mathbf{x})$	$p(y^1, \delta = 1\|\mathbf{x})$	$p(y^1\|\mathbf{x})$
y^2	$p(y^2, \delta = 0\|\mathbf{x})$	$p(y^2, \delta = 1\|\mathbf{x})$	$p(y^2\|\mathbf{x})$
$\dots$	$\dots$	$\dots$	$\dots$
y^J	$p(y^J, \delta = 0\|\mathbf{x})$	$p(y^J, \delta = 1\|\mathbf{x})$	$p(y^J\|\mathbf{x})$
Total	$p(\delta = 0\|\mathbf{x})$	$p(\delta = 1\|\mathbf{x})$	1

Table: Contingency table of $(\mathcal{Y}, \delta)|\mathbf{x}$

Remark: $p(y^j|\mathbf{x})$ and $p(y^j, \delta = 0|\mathbf{x})$ are *unknown*.

Uncertainty and its measure - 3

Basic inequality

$$p(y^j, \delta = 1|\mathbf{x}) \leq p(y^j|\mathbf{x}) \leq 1 - \sum_{\substack{t=1 \\ t \neq j}}^{J} p(y^t, \delta = 1|\mathbf{x}) = p(\delta = 1|\mathbf{x}) - p(y^j, \delta = 1|\mathbf{x})$$

Uncertainty on $p(y|\mathbf{x})$ can be measured as the size of the interval $p(y^j|\mathbf{x})$ lies in.

Uncertainty and its measure - 4

Measure of uncertainty for $p(y^j|\mathbf{x})$

$$U(p(y^j|\mathbf{x})) = 1 - \sum_{t=1}^{J} p(y^t, \delta = 1|\mathbf{x}) = 1 - p(\delta = 1|\mathbf{x}).$$

- $0 \leq U(p(y^j|\mathbf{x})) \leq 1$.
- $U(p(y^j|\mathbf{x})) = 1$ if $p(\delta = 1|\mathbf{x}) = 0$ (no sample data available).
- $U(p(y^j|\mathbf{x})) = 0$ if $p(\delta = 1|\mathbf{x}) = 1$ (all the units in the population are sampled).

Uncertainty and its measure - 5

Measure of uncertainty for the conditional spmf $p(y|\mathbf{x})$

$$U(p(y|\mathbf{x})) = \frac{J}{\sum_{j=1}}^{J} U(p(y^j|\mathbf{x})) = 1 - p(\delta = 1|\mathbf{x}).$$

Uncertainty for the marginal probability spmf $p(y)$

$$U(p(y)) = \sum_{\mathbf{x}} p(\mathbf{x}) U^x(p(y|\mathbf{x})) = \sum_{\mathbf{x}} p(\mathbf{x})(1 - p(\delta = 1|\mathbf{x})) = 1 - p(\delta = 1).$$

Estimation of uncertainty measure - 1

Estimation of inclusion probability for sample B

- Conditional inclusion probabilities

$$\widehat{p}(\delta = 1|\boldsymbol{x}) = \frac{\sum_{i=1}^{N} \frac{1}{\pi_i^A} I_{(\boldsymbol{x}_i = \boldsymbol{x})} \delta_i D_i}{\sum_{i=1}^{N} \frac{1}{\pi_i^A} I_{(\boldsymbol{x}_i = \boldsymbol{x})} D_i}.$$

- Unconditional inclusion probabilities

$$\widehat{p}(\delta = 1) = \frac{\sum_{i=1}^{N} \frac{1}{\pi_i^A} \delta_i}{\sum_{i=1}^{N} \frac{1}{\pi_i^A}}. \quad (2)$$

Estimation of uncertainty measure - 2

Estimation of uncertainty measure

- Conditional uncertainty measure

$$\widehat{U}^x(p(y|\mathbf{x})) = 1 - \widehat{p}(\delta = 1|\mathbf{x}).$$

- Unconditional uncertainty measure

$$\widehat{U}(p(y)) = \sum_{x} \widehat{p}(\mathbf{x})(1 - \widehat{p}(\delta = 1|\mathbf{x})) = 1 - \widehat{p}(\delta = 1), \tag{3}$$

where weights $p(\mathbf{x})$ are estimated from the probability sample A *via* the Hájek estimator

$$\widehat{p}(\mathbf{x}) = \frac{\sum_{i=1}^{N} \frac{1}{\pi_i^A} I_{(\mathbf{x}_i = x)}}{\sum_{i=1}^{N} \frac{1}{\pi_i^A}}.$$

Properties: consistency, asymptotic normality,...

Reducing uncertainty though extra-samples information - 1

- Extra-sample information, when available, make it tighter the bounds on $p(y|\boldsymbol{x})$.
- In this way, the corresponding uncertainty is *reduced*.

Kinds of extra-sample information considered.

(i) Auxiliary information on the informative sampling design picking B.

(ii) Auxiliary information on the conditional spmf $p(y|\boldsymbol{x})$.

Reducing uncertainty though extra-samples information - 2

Auxiliary information on the sampling design selecting B
Conditionally on $\boldsymbol{x}$, extra-sample information expressed by inequality is considered.

(i-1) $p(y^j|\delta = 1, \boldsymbol{x}) \leq p(y^j, |\boldsymbol{x})$. In sample B the probability that $y = y^j$ is smaller than in the population, for $j = 1, \dots, h$ with $h \leq J$;

(i-2) $p(y^j|\delta = 1, \boldsymbol{x}) \geq p(y^j, |\boldsymbol{x})$. In sample B the probability that $y = y^j$ is larger than in the population, for $j = 1, \dots, h$ with $h \leq J$.

Reducing uncertainty though extra-samples information - 3

Auxiliary information on the distribution of $\mathcal{Y}|\boldsymbol{\mathcal{X}}$

Conditionally on $\boldsymbol{x}$, partial information of the distribution of $\mathcal{Y}|\boldsymbol{\mathcal{X}}$, in form of inequalities, is considered.

(ii-1) Preliminary estimates for some of the J parameters $p(y^j|\boldsymbol{x})$ are available.

(ii-2) A range of plausible estimates for some of the J parameters $p(y^j|\boldsymbol{x})$ are available.

Such an extra-sample information could be obtained from previous surveys, or from a small-scale pilot survey, or could be elicited experts on the topic of interest.

Reducing uncertainty though extra-samples information - 4

Effect of the extra-sample information

- The main effect of the above extra-sample information consists in *tightening* the bounds for $p(y^j|\boldsymbol{x})$:

$$l_j(\boldsymbol{x}) \leq p(y^j|\boldsymbol{x}) \leq u_j(\boldsymbol{x}).$$

- In this way, *uncertainty reduces to*

$$U^c(p(y^j|\boldsymbol{x})) = u_j(\boldsymbol{x}) - l_j(\boldsymbol{x}).$$

- The bounds $u_j(\boldsymbol{x})$, $l_j(\boldsymbol{x})$ are identifiable, and can be estimated on the basis of samples A, B.
- Uncertainty can be estimated, as well. Estimates are consistent, asymptotically normally distributed, ...

Estimation error - 1

For the sake of simplicity, let us confine ourselves to the estimation of the conditional probability $p(y^j|\boldsymbol{x})$. Similar considerations hold for the whole distribution $p(\cdot|\boldsymbol{x})$, of for unconditional probabilities, or other.

- $\widehat{u}_j(\boldsymbol{x})$, $\widehat{l}_j(\boldsymbol{x})$ sample(s) estimates of $u_j(\boldsymbol{x})$, $l_j(\boldsymbol{x})$, respectively.
- Each $\widehat{p}(y^j|\boldsymbol{x})$ in between $\widehat{u}_j(\boldsymbol{x})$ and $\widehat{l}_j(\boldsymbol{x})$ is a legitimate estimate of the *true* $p(y^j|\boldsymbol{x})$.
- Under wide regularity conditions, $\widehat{p}(y^j|\boldsymbol{x})$ tends in probability to some $p^*(y^j|\boldsymbol{x})$ in between $l_j(\boldsymbol{x})$ and $u_j(\boldsymbol{x})$. The same also holds for expectation: $E[\widehat{p}(y^j|\boldsymbol{x})] \rightarrow p^*(y^j|\boldsymbol{x})$.
- In general, $p^*(y^j|\boldsymbol{x}) \neq p(y^j|\boldsymbol{x})$.

Estimation error - 2

Estimation error

$$\widehat{p}(y^j|\mathbf{x}) - p(y^j|\mathbf{x}) = \underbrace{\left(\widehat{p}(y^j|\mathbf{x}) - p^*(y^j|\mathbf{x})\right)}_{\textit{decreases to } 0 \textit{ as the sample sizes increase}} + \underbrace{\left(p^*(y^j|\mathbf{x}) - p(y^j|\mathbf{x})\right)}_{\textit{does not decrease to } 0 \textit{ as the sample sizes increase}} .$$

Estimation error - 3

$$
\begin{aligned}
MSE(\widehat{p}(y^j|\mathbf{x})) &= E\left[\left(\widehat{p}(y^j|\mathbf{x}) - p(y^j|\mathbf{x})\right)^2\right] \\
&= V\left(\widehat{p}(y^j|\mathbf{x})\right) + \left(p^*(y^j|\mathbf{x}) - p(y^j|\mathbf{x})\right)^2 .
\end{aligned}
$$

- The variance term $V\left(\widehat{p}(y^j|\mathbf{x})\right)$ decreases to 0 as the sample sizes increase.
- The bias term $\left(p^*(y^j|\mathbf{x}) - p(y^j|\mathbf{x})\right)^2$ does not decrease to 0 as the sample sizes increase. However, it can be upper-bounded by squared uncertainty

$$
\left(p^*(y^j|\mathbf{x}) - p(y^j|\mathbf{x})\right)^2 \leq U^c(p(y^j|\mathbf{x}))^2
$$

where the upper bound on the r.h.s. can be estimated *via* sample data.