# On several recent quasi-randomization approaches to estimation from non-probability samples

Vladislav Beresovsky, Terrance D. Savitsky, Julie Gershunskaya

FCSM Research & Policy Conference
October 22, 2024

Disclaimer: The views expressed are those of the authors and do not necessarily reflect the official policy or position of U.S. Bureau of Labor Statistics

# Why consider non-probability samples

- Probability based samples have long been an established way of conducting surveys
- **Problems with traditional probability-based surveys:** Lowering response rates, increased burden and cost of data collection
- **New opportunities:** Availability of data from variety of sources, related to the Internet, computers, etc.
  The demand for exploiting these resources is steadily growing.
- However, such "opportunistic" (non-probability based) data cannot be automatically regarded as representative, since this information is not based on a well designed random sample.
- Methods have been developed to account for potential selection bias

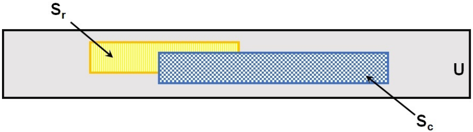
# Quasi-randomization approach

- Assume the existence of a latent mechanism that governs the non-probability sample selection.
- **Basic idea:** use information from available probability-based (*"reference"*) sample to uncover *latent probabilities to participate* in the non-probability survey
- Use these participation probabilities in estimation of target finite population quantities.
- We compare several methods for estimation of participation probabilities

# Setup and notation

$U$ target finite population of size $N$

$\mu = \sum_{i \in U} y_i / N$ target quantity



$S_c$ non-probability (*convenience*) sample

$(y_i, \mathbf{x}_i)$ observed on $S_c$

$I_{ci}$ inclusion indicator for $S_c$

$\pi_{ci} = P\{I_{ci} = 1 \mid i \in U\}$ (not known)

$S_r$ probability (*reference*) sample

$\mathbf{x}_i$ observed on $S_r$

$I_{ri}$ inclusion indicator for $S_r$,

$\pi_{ri} = P\{I_{ri} = 1 \mid i \in U\}$ (known)

| | $S_c$ | $S_r$ |
|---|---|---|
| $y_i$ | ✓ | |
| $\mathbf{x}_i$ | ✓ | ✓ |
| $I_{ci}$ | 1 | |
| $\pi_{ci}$ | ✗ | |
| $I_{ri}$ | | 1 |
| $\pi_{ri}$ | ✓ | ✓ |

We wish to estimate $\pi_{ci}$, then *Inverse Propensity Weighted (IPW)* estimator of population mean $\mu$ is

$$\hat{\mu} = \frac{\sum_{i \in S_c} y_i / \hat{\pi}_c}{\sum_{i \in S_c} 1 / \hat{\pi}_c}$$

# Pseudo-likelihood approach of **Chen, Li and Wu(2020)**

Consider $I_{ci} \sim \text{Bernoulli}(\pi_{ci})$ on population $U$:

$$\begin{aligned} \ell^{CLW}(\boldsymbol{\beta}) &= \sum_{i \in U} \left\{ I_{ci} \log\left[\pi_{ci}(\boldsymbol{\beta})\right] + (1 - I_{ci}) \log\left[1 - \pi_{ci}(\boldsymbol{\beta})\right] \right\} \\ &= \sum_{i \in S_c} \log\left[\frac{\pi_{ci}(\boldsymbol{\beta})}{1 - \pi_{ci}(\boldsymbol{\beta})}\right] + \sum_{i \in U} \log\left[1 - \pi_{ci}(\boldsymbol{\beta})\right], \end{aligned}$$

and $\text{logit}\left[\pi_{ci}(\boldsymbol{\beta})\right] = \boldsymbol{\beta}^{\boldsymbol{T}} \mathbf{x}_i$.

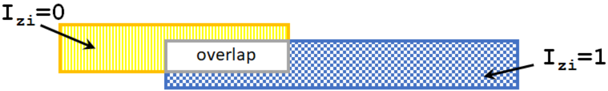
Since $U$ is not available, use pseudo-likelihood:

$$\hat{\ell}^{CLW}(\boldsymbol{\beta}) = \sum_{i \in S_c} \log\left[\frac{\pi_{ci}(\boldsymbol{\beta})}{1 - \pi_{ci}(\boldsymbol{\beta})}\right] + \sum_{i \in S_r} w_{ri} \log\left[1 - \pi_{ci}(\boldsymbol{\beta})\right],$$

where $w_{ri} = \pi_{ri}^{-1}$.

# Sample based approach (under negligible sampling overlap)

**Elliott (2009)**



Consider:

$\pi_{zi} = P\{I_{zi} = 1|\mathbf{x}_i\}$

on the pooled set

| | $S_c$ | $S_r$ | $S_c \cap S_r$ |
|---|---|---|---|
| $I_z$ | 1 | 0 | negligible |

Under "negligible" sampling overlap, approximate relationship holds:

$$\pi_{zi} \approx \frac{\pi_{ci}}{\pi_{ci} + \pi_{ri}}.$$

A two-step procedure:

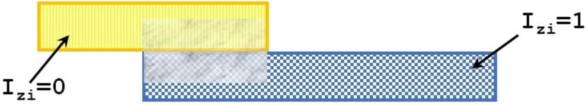
Step 1: Estimate $\pi_{zi}$ using standard methods

Step 2: Find $\pi_{ci}$ from $\pi_{zi} \approx \pi_{ci}/(\pi_{ci} + \pi_{ri})$

# Sample based approach (unknown overlap of any size)

**Savitsky, Williams, Gershunskaya and Beresovsky (2023)**

**Stacked sample:** $S = S_c + S_r$ (overlapping units appear in $S$ twice)



|       | $S_c$ | $S_r$ |
|-------|-------|-------|
| $I_z$ | 1     | 0     |

$\pi_{zi} = P\{I_{zi} = 1 | i \in S\}$ is probability to be in $S_c$ for units in stack $S$

*Key relationship for independent sampling probabilities (KRISP):*
Under stacked samples setup, assuming $S_c$ and $S_r$ are independently selected from $U$, relationship

$$\pi_{zi} = \frac{\pi_{ci}}{\pi_{ci} + \pi_{ri}}$$

✓ holds exactly,
✓ regardless of the size of sampling overlap.

# Implicit Logistic Regression (ILR)

**Beresovsky(2019)**:

The log-likelihood for observed Bernoulli variable $I_{zi}$ is

$$\ell^{ILR}(\boldsymbol{\beta}) = \sum_{i \in S_c} \log\left(\pi_{zi}[\pi_{ci}(\boldsymbol{\beta})]\right) + \sum_{i \in S_r} \log\left(1 - \pi_{zi}\left[\pi_{ci}(\boldsymbol{\beta})\right]\right),$$

where $\pi_{zi}$ can be treated as a *composite function*, based on KRISP,

$$\pi_{zi} = \pi_{zi}[\pi_{ci}(\boldsymbol{\beta})] = \frac{\pi_{ci}(\boldsymbol{\beta})}{\pi_{ri} + \pi_{ci}(\boldsymbol{\beta})}$$

and $\text{logit}\left[\pi_{ci}(\boldsymbol{\beta})\right] = \boldsymbol{\beta}^{\boldsymbol{T}}\mathbf{x}_i$.

Take derivatives wrt $\boldsymbol{\beta}$ and solve estimating equations, thus estimating $\pi_{ci}$ directly from the likelihood.

# Asymptotic variances of the estimates of $\mu$ and $\boldsymbol{\beta}$

$$\boxed{\begin{aligned}\text{Var}(\hat{\mu}) &\doteq \text{Var}[U(\mu)] - 2\boldsymbol{b}^T\text{Cov}[U(\mu), S(\boldsymbol{\beta})] + \boldsymbol{b}^T\text{Var}[S(\boldsymbol{\beta})]\boldsymbol{b},\\ \text{Var}(\hat{\boldsymbol{\beta}}) &\doteq \boldsymbol{H}^{-1}\text{Var}[S(\boldsymbol{\beta})]\boldsymbol{H}^{-1},\end{aligned}}$$

where $\boldsymbol{b} = S_{\boldsymbol{\beta}}^{-1}U_{\boldsymbol{\beta}}^T$, $U_{\boldsymbol{\beta}} = E[\partial U(\mu)/\partial\boldsymbol{\beta}^T]$, $S_{\boldsymbol{\beta}} = E[\partial S(\boldsymbol{\beta})/\partial\boldsymbol{\beta}]$, $\boldsymbol{H} = -S_{\boldsymbol{\beta}}$.

Both methods, CLW and ILR, are asymptotically equivalent:

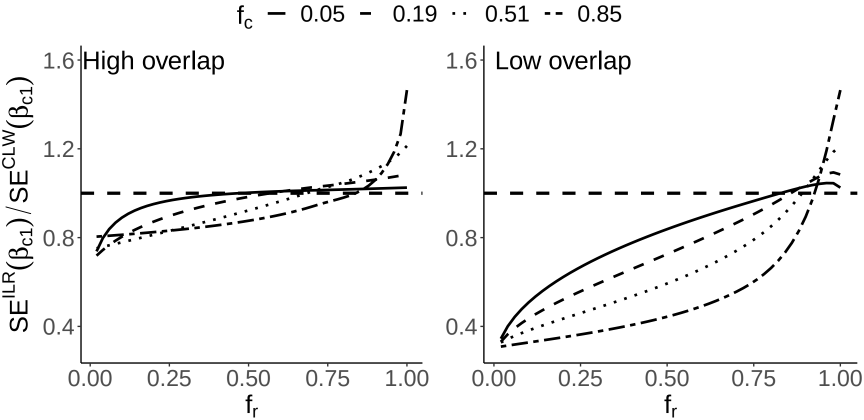
$$\text{Var}\begin{pmatrix}\hat{\mu}\\ \hat{\boldsymbol{\beta}}\end{pmatrix} \sim O\left(\frac{1}{n_r} + \frac{1}{n_c}\right) = O\left(\frac{1}{\min(n_r, n_c)}\right).$$

Note that

$$\text{Var}[S(\boldsymbol{\beta})] = \text{Var}[S_c(\boldsymbol{\beta})] + \text{Var}[S_r(\boldsymbol{\beta})].$$

The methods differ mostly due to contributions from $\text{Var}[S_r(\boldsymbol{\beta})]$.

# Relative efficiency of ILR and CLW estimates of propensity model parameter



# Classification Trees

We can use ILR and CLW methods with a classification trees algorithm.

We expect: a sample-based likelihood would be more efficient compared to a pseudo-likelihood, especially as the regression tree grows and its nodes are based on progressively smaller samples

For a given node $g$, find an optimal binary split based on a given covariate:

(1) find $(\pi_{cgL}, \pi_{cgR})$, estimates of probabilities in left and right branches, on a grid of possible splits;

(2) choose an optimal split based on an objective function

# Homogeneous groups (based on **CLW pseudo-likelihood**)

The algorithm splits data into homogeneous groups $g = 1, \ldots, G$, so that all units in a given group have the same probabilities $\pi_{cg}$.

Under (CLW) pseudo-likelihood:

$$\hat{\pi}_{cg} = \frac{n_{cg}}{\hat{N}_g}, \text{ where } \hat{N}_g = \sum_{i \in g} w_{ri}.$$

Estimated entropy impurity criteria:

$$\hat{I}^{CLW} = -\sum_{g=1}^{G} \frac{\hat{N}_g}{\hat{N}} \left[\hat{\pi}_{cg} \log(\hat{\pi}_{cg}) + (1 - \hat{\pi}_{cg}) \log(1 - \hat{\pi}_{cg})\right]$$

# Homogeneous groups (based on **stacked-samples setup**)

For sample-based (ILR) approach: no explicit expression for $\pi_{cg}$. It can be found as a solution of equation

$$\sum_{i \in S_g} \frac{\pi_{ri}}{\pi_{ri} + \pi_{cg}} = n_{rg},$$

where $S_g$ is the part of stacked sample $S$ belonging to group $g$.

Estimated entropy impurity criteria:

$$\hat{I}^{ILR} \qquad = - \sum_{g=1}^{G} \left[ \sum_{i \in S_{cg}} \log \hat{\pi}_{zi,g} + \sum_{i \in S_{rg}} \log(1 - \hat{\pi}_{zi,g}) \right],$$

where $\hat{\pi}_{zi,g} = \frac{\hat{\pi}_{cg}}{\pi_{ri} + \hat{\pi}_{cg}}$.

# Simulation example: setup

Suppose we already have a grown tree, up to some level $g$.
We focus on splitting node $g$ into two parts.

The setup is:

$N_g = 1,000$ is population size
Covariate: $x_{ig} \sim N(0,1)$
Study variable: $y_{ig} = 1 + x_{ig} + \epsilon_{ig}$, with $\epsilon_{ig} \sim N(0, 1.5^2)$

True *convenience* sample probabilities:

$\pi_{cg,L} = 0.80$ for $i$ with $x_{ig} <= 0$,
$\pi_{cg,R} = 0.20$ for $i$ with $x_{ig} > 0$.
Hence, convenience sample size is $n_{cg} \approx 500$.

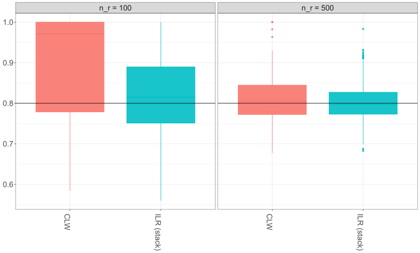
PPS design for *probability* sample $\pi_{rg,i} \propto x_{ig}$
Scenarios: $n_{rg} = 100, 500$

1000 simulation runs

Reference sample scenarios: $n_{rg} \in \{100, 500\}$

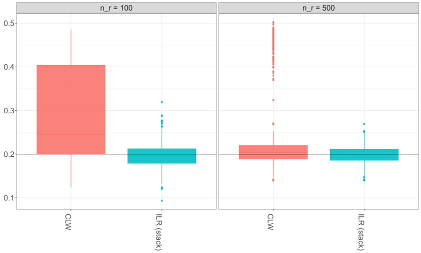
Estimators: CLW, ILR (stacked samples)

$n_r = 100$ $n_r = 500$



(a) $\pi_{cg,L} = 0.80$

$n_r = 100$ $n_r = 500$



(b) $\pi_{cg,R} = 0.20$

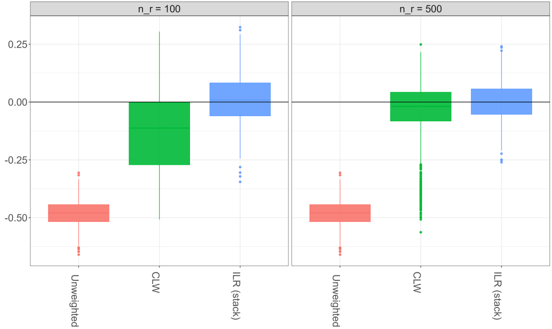
# Estimates of $\mu$

Estimators: Unweighted, CLW, ILR (stacked sample)

$$\hat{\mu} = \frac{\sum_{i\in S_c} y_i/\hat{\pi}_c}{\sum_{i\in S_c} 1/\hat{\pi}_c}$$

| | $n_r = 100$ | | $n_r = 500$ | |
|---|---|---|---|---|
| | Bias | MSE | Bias | MSE |
| Unweighted | -0.481 | 0.234 | -0.481 | 0.234 |
| CLW | -0.128 | 0.042 | -0.047 | 0.023 |
| ILR (stack) | 0.012 | 0.011 | 0.001 | 0.007 |

$n_r = 100$ $\qquad$ $n_r = 500$



# Summary

- The pseudo-likelihood CLW and stacked-samples based ILR approaches are **asymptotically equivalent**
- The stacked-samples ILR method is more efficient compared to CLW **under practically important scenarios** of:
  - a small reference sample and
  - a low overlap in covariates-defined domains (resulting in an insufficient representation of some population groups in either of the two samples)
- **Future research:** Linkage may lead to better estimates *if a good matching quality could be achieved*. However, thinking about the balance between cost/effort spent and the quality of record linkage, the question in context of estimation of participation probabilities is: *Do we really have to link?*

# CONTACT INFORMATION

Gershunskaya.Julie@bls.gov

Thank you!