Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Comments on “Handling non-probability samples through inverse probability weighting with an application to Statistics Canada’s crowdsourcing data”

by Changbao Wu

Release date: June 25, 2024

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at infostats@statcan.gc.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

• Statistical Information Service	1-800-263-1136
• National telecommunications device for the hearing impaired	1-800-363-7629
• Fax line	1-514-283-9350

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology, June 2024
Vol. 50, No. 1, pp. 117-121
Statistics Canada, Catalogue No. 12-001-X

Comments on "Handling non-probability samples through inverse probability weighting with an application to Statistics Canada's crowdsourcing data"

Changbao Wu[1]

Abstract

We provide comparisons among three parametric methods for the estimation of participation probabilities and some brief comments on homogeneous groups and post-stratification.

Key Words: Inverse probability weighting; Participation probability; Pooled sample; Poststratification; Pseudo maximum likelihood; Reference probability sample.

Beaumont, Bosa, Brennan, Charlebois and Chu (2024) provided a thorough examination of inverse probability weighting methods for non-probability samples, including parametric approaches and tree-based classification methods, with a major focus on variable selection. This is an important research topic, as the demands on using non-probability samples in applied fields and official statistics have been steadily increasing in recent years. The current paper is a timely contribution to investigating and comparing different methodologies using a real world dataset. I would like to thank the Guest Editor, Dr. Partha Lahiri, for the invitation and I appreciate the opportunity for a short discussion. In Section 1, I provide comparisons among three parametric methods for the estimation of participation probabilities and for inverse probability weighting. I also provide some brief comments on the use of homogeneous groups for post-stratification in Section 2.

1. The methods of Chen, Li and Wu (2020), Valliant and Dever (2011), and Wang, Valliant and Li (2021)

These are three parametric methods frequently cited in recent literature on inverse probability weighting (IPW) using estimated participation probabilities for non-probability samples. There are conceptual differences among the three methods as well as similarities in numeric results when sample sizes are small relative to the population size.

The foundation for IPW estimators has been built under probability sampling with the Horvitz-Thompson estimator. Follow the notation of Beaumont et al. (2024), let $U = \{1, 2, \ldots, N\}$ be the finite population of size N. Let d denote the probability sampling design for selecting a probability sample s. Under the probability sampling design, the sample inclusion indicator variable $\delta_k = I(k \in s)$ is defined for every unit k in the target population U, i.e., for $k = 1, 2, \ldots, N$, where $I(\cdot)$ is the indicator function, and the sampling inclusion probabilities $\pi_k = P(\delta_k = 1 \mid U) = P(k \in s \mid U)$ can be computed based on the given

1. Changbao Wu, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo ON N2L 3G1. E-mail: cbwu@uwaterloo.ca.

sampling design, d. The joint distribution of $(\delta_1, \delta_2, \ldots, \delta_N)$ under repeated sampling characterizes the sampling design features, and the Horvitz-Thompson estimator $\hat{\theta}_{\text{HT}} = \sum_{k \in s} y_k / \pi_k$ for the population total $\theta = \sum_{k \in U} y_k$ is uniquely design-unbiased among a class of linear estimators. The establishment of this fundamental result in probability sampling involves (i) the positivity assumption, i.e., $\pi_k > 0$ for all k in U, so that $\hat{\theta}_{\text{HT}}$ can be re-written as $\hat{\theta}_{\text{HT}} = \sum_{k \in U} \delta_k y_k / \pi_k$; and (ii) the sample inclusion indicator δ_k is independent of y_k and $E_d(\delta_k \mid U) = \pi_k$, where the expectation E_d is with respect to the sampling design d.

Let s_{NP} be a non-probability sample of size n_{NP} from the population U. Let $\{(y_k, \mathbf{x}_k), k \in s_{\text{NP}}\}$ be the non-probability sample dataset. Once again, the sample participation indicator $\delta_k = I(k \in s_{\text{NP}})$ is defined for every unit k in the target population U, i.e., for $k = 1, 2, \ldots, N$. Unlike probability sampling, the participation probabilities $p_k = P(\delta_k = 1 \mid U)$ for the non-probability sample s_{NP} are unknown and hence need to be estimated, which requires assumptions on the form of p_k and an assumed model, denoted as q, for the participation mechanism. The model q leads to specifications of the joint distribution of $(\delta_1, \delta_2, \ldots, \delta_N)$. There are two commonly assumed components for model q: (i) the participation probabilities satisfy $p_k = P(\delta_k = 1 \mid \mathbf{x}_k, y_k) = P(\delta_k \mid \mathbf{x}_k) > 0$, $i = 1, 2, \ldots, N$; and (ii) the sample inculsion indicators $\delta_1, \delta_2, \ldots, \delta_N$ are conditionally independent given $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$.

It should be emphasized that IPW estimators for non-probability samples need to be constructed and evaluated under the assumed model q on participation mechanism and under the induced distribution on $(\delta_1, \delta_2, \ldots, \delta_N)$. It is where conceptual differences among the three methods can clearly be identified. The method of Chen et al. (2020) for estimating p_k is based on the full likelihood function $\prod_{k=1}^{N} p_k^{\delta_k} (1 - p_k)^{1 - \delta_k}$. With a pre-specified parametric form $p_k = p(\mathbf{x}_k, \boldsymbol{\alpha})$, the maximum pseudo likelihood estimator $\hat{\boldsymbol{\alpha}}$ is derived and assessed under an assumed parametric model q on $(\delta_1, \delta_2, \ldots, \delta_N)$ as well as the sampling design, d, for the reference probability sample s_P. The IPW estimator $\hat{\theta}_{\text{IPW}} = \sum_{k \in s_{\text{NP}}} y_k / \hat{p}_k$, where $\hat{p}_k = p(\mathbf{x}_k, \hat{\boldsymbol{\alpha}})$, is consistent for θ under the joint randomization of model q and the sampling design d. The results are rigorously established with no restrictions on the "sampling fraction" n_{NP} / N or the parametric form for $p_k = p(\mathbf{x}_k, \boldsymbol{\alpha})$, whether it follows from a logistic regression model or any other models suitable for a binary response variable.

The paper by Valliant and Dever (2011) was the first serious attempt in addressing estimation of participation probabilities under the two-sample setup as described in Beaumont et al. (2024). It inspired several followup papers, including Chen et al. (2020) and Wang et al. (2021). The proposed method was based on fitting a survey weighted logistic regression model to the pooled sample $s_{\text{NP}} \cup s_P$ with the "response variable" defined as $D_k = 1$ if $k \in s_{\text{NP}}$ and $D_k = 0$ if $k \in s_P$, for $k \in s_{\text{NP}} \cup s_P$, assuming there are no overlaps between s_{NP} and s_P. It is apparent that the D_i's are defined with the given s_{NP} and s_P and are conceptually different from the sample participation indicators $(\delta_1, \delta_2, \ldots, \delta_N)$. Under an assumed parametric model ξ on the D_k's with $\pi(\mathbf{x}_k, \boldsymbol{\gamma}) = P(D_k = 1 \mid s_{\text{NP}} \cup s_P)$, the "theoretical IPW estimator" $\hat{\theta} = \sum_{k \in s_{\text{NP}}} y_k / \pi(\mathbf{x}_k, \boldsymbol{\gamma})$ should be evaluated first against model ξ, leading to $E_\xi(\hat{\theta} \mid s_{\text{NP}}, s_P) = E_\xi \left\{ \sum_{k \in s_{\text{NP}} \cup s_P} D_k y_k / \pi(\mathbf{x}_k, \boldsymbol{\gamma}) \mid s_{\text{NP}}, s_P \right\} = \sum_{k \in s_{\text{NP}} \cup s_P} y_k$. In creating a set of weights for the pooled sample

$s_{\text{NP}} \cup s_P$ for "survey-weighted" logistic regression analysis on the D_k's and without any prior knowledge of how s_{NP} was selected, Valliant and Dever (2011) simply assigned "1" to each $k \in s_{\text{NP}}$, which essentially assumes that units in s_{NP} are exchangeable with respect to model q. The IPW estimator of Valliant and Dever (2011), when assessed under model q for $(\delta_1, \delta_2, \ldots, \delta_N)$, is inconsistent unless s_{NP} is a simple random sample from the population as shown by Chen et al. (2020).

The more recent paper by Wang et al. (2021) adapted a strategy which is on the opposite direction of Valliant and Dever (2011). Instead of pooling the two samples together, the authors first created an enlarged population $s_{\text{NP}}^* \cup U$, where s_{NP}^* consists of the same set of units in $s_{\text{NP}} \subset U$ but these units are viewed differently in the union of s_{NP}^* and U. The authors defined the indicator variable $R_k = 1$ if $k \in s_{\text{NP}}^*$ and $R_k = 0$ if $k \in U$, and further defined the probability $\pi_k = P(R_k = 1 \mid s_{\text{NP}}^* \cup U) = P(k \in s_{\text{NP}}^* \mid k \in s_{\text{NP}}^* \cup U)$, for all $k \in s_{\text{NP}}^* \cup U$. I had some real difficulty in putting this formulation into a suitable conceptual framework, since the indicator variable R_k depends on $s_{\text{NP}} = \{i \mid i \in U \text{ and } \delta_i = 1\}$, which depends on the full vector of sample inclusion indicators $(\delta_1, \delta_2, \ldots, \delta_N)$. It is unclear what kind of probability model is behind $P(\cdot)$ in defining $\pi_k = P(R_k = 1 \mid s_{\text{NP}}^* \cup U)$. This further led to my struggle in understanding the arguments behind the identity $p_k = P(\delta_k = 1 \mid U) = \pi_k / (1 - \pi_k)$ (Wang et al., 2021, page 5241, equation (9)). The identity implies that a logistic regression model on the R_k's would lead to a model on the δ_k's with the log-link function, a potential source of concerns when the sampling fraction n_{NP}/N is large.

It turns out that the three methods produce similar numeric results for the estimated participation probabilities when the sampling fraction n_{NP}/N is small. This can be explained by checking computational details for each of the methods. Under a parametric model $p_k = p(\mathbf{x}_k, \boldsymbol{\alpha})$ and the assumption of conditional independence of $(\delta_1, \delta_2, \ldots, \delta_N)$ given $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$, the full log-likelihood function for α is given by

$$\ell(\boldsymbol{\alpha}) = \sum_{k \in s_{\text{NP}}} \log\{p(\mathbf{x}_k, \boldsymbol{\alpha})\} + \sum_{k \in U - s_{\text{NP}}} \log\{1 - p(\mathbf{x}_k, \boldsymbol{\alpha})\}. \tag{1.1}$$

The second term on the right hand side of (1.1), denoted as

$$L_2 = \sum_{k \in U - s_{\text{NP}}} \log\{1 - p(\mathbf{x}_k, \boldsymbol{\alpha})\},$$

is not computable based on the non-probability sample s_{NP} since it involves $\mathbf{x}_k$ for all k not in the sample s_{NP}. The three methods, although conceptually distinctive, differ computationally only in terms of how the term L_2 is handled.

Let $\{(\mathbf{x}_k, d_k), k \in s_P\}$ be the reference probability sample dataset, where the d_k's are the survey weights for s_P. Treating L_2 as a total of a population of size $N - n_{\text{NP}}$, the method of Valliant and Dever (2011) is equivalent to estimating L_2 by

$$\hat{L}_2^{(1)} = \sum_{k \in s_P} w_k \log\{1 - p(\mathbf{x}_k, \boldsymbol{\alpha})\},$$

where $w_k = d_k(\hat{N} - n_{\text{NP}})/\hat{N}$ are the rescaled weights satisfying $\sum_{k \in s_P} w_k = \hat{N} - n_{\text{NP}}$ and $\hat{N} = \sum_{k \in s_P} d_k$ is the estimated population size for U. Noting that L_2 can be re-written as $L_2 = \sum_{k \in U} \log\{1 - p(\mathbf{x}_k, \boldsymbol{\alpha})\} - \sum_{k \in s_{\text{NP}}} \log\{1 - p(\mathbf{x}_k, \boldsymbol{\alpha})\}$, the method of Chen et al. (2020) replaces L_2 by

$$\hat{L}_2^{(2)} = \sum_{k \in s_P} d_k \log\{1 - p(\mathbf{x}_k, \boldsymbol{\alpha})\} - \sum_{k \in s_{\text{NP}}} \log\{1 - p(\mathbf{x}_k, \boldsymbol{\alpha})\},$$

which is design-unbiased (or design-consistent, depending on whether the d_k's are the basic design weights or calibrated/adjusted weights) regardless of the sampling fraction n_{NP}/N. The method of Wang et al. (2021) for estimating $\pi_k = P(R_k = 1 \mid s_{\text{NP}} \cup U)$ amounts to replacing L_2 by

$$\hat{L}_2^{(3)} = \sum_{k \in s_P} d_k \log\{1 - p(\mathbf{x}_k, \boldsymbol{\alpha})\}.$$

This clearly overshoots the target since $\hat{L}_2^{(3)}$ is an estimate for $\sum_{k \in U} \log\{1 - p(\mathbf{x}_k, \boldsymbol{\alpha})\}$ (or "undershoots" if we consider the fact that $\log\{1 - p(\mathbf{x}_k, \boldsymbol{\alpha})\} < 0$ for all k). However, the use of $\hat{L}_2^{(3)}$ to replace L_2 in $\ell(\boldsymbol{\alpha})$ results in a log-likelihood function that resembles a hypothetical scenario where the sample s_{NP} is taken from a larger population $s_{\text{NP}} \cup U$. The resulting π_k needs to be adjusted to make it closer to the actual participation probability p_k.

It is apparent that the three quantities $\hat{L}_2^{(1)}$, $\hat{L}_2^{(2)}$ and $\hat{L}_2^{(3)}$ do not differ too much when the sampling fraction n_{NP}/N is small, leading to similar estimated participation probabilities under these scenarios. The final adjustment step from the method of Wang et al. (2021), i.e., $p_k = \pi_k/(1 - \pi_k)$, also gives similar results, since we typically have $\pi_k = O(n_{\text{NP}}/N)$, which implies $p_k/\pi_k = 1 + o(1)$ when $n_{\text{NP}}/N = o(1)$.

2. Homogeneous groups and post-stratification

In practice, auxiliary variables which are included in probability or non-probability surveys are often categorical or ordinal, especially for surveys on human populations where basic information on demographic variables and social-economic indicators is routinely collected. When relevant variables for characterizing the participation mechanism are discrete, the IPW estimator is equivalent to a post-stratified estimator; see, for instance, Section 5 of Wu (2022) for a detailed discussion. A post-stratified (IPW) estimator uses uniform participation probabilities within the same post-stratum, which effectively removes the impact of extreme values of estimated participation probabilities often appearing with a parametric model when there are continuous auxiliary variables, and the estimator has a simple and easy-to-use form.

There are two major challenges, however, in forming homogeneous groups as post-strata. The first is the large number of initial groups when there are many discrete auxiliary variables that are available in the datasets. The variable selection methods discussed in Beaumont et al. (2024) have the potential to be practically useful in reducing the number of groups for the final IPW estimator. The second are scenarios where there are a large number of mixed discrete and continuous auxiliary variables. There exist methodologies developed in the missing data and causal inference literature that could be adapted for non-probability samples. Section 5 of Wu (2022) contains a brief discussion on rank-based methods. The rank-based method described in Beaumont et al. (2024) has similar spirits. This is an important topic that requires further research.

Variable selection using AIC or other similar criteria requires a true likelihood function. Beaumont et al. (2024) demonstrated the usefulness of the pseudo likelihood function of Chen et al. (2020) in the context of variable selection. I am excited to see such development and look forward to seeing more research effort in that direction.

Acknowledgements

This research was supported by grants from the Natural Sciences and Engineering Research Council of Canada and the Canadian Statistical Sciences Institute.

References

Beaumont, J.-F., Bosa, K., Brennan, A., Charlebois, J. and Chu, K. (2024). Handling non-probability samples through inverse probability weighting with an application to Statistics Canada's crowdsourcing data. *Survey Methodology*, 50, 1, 77-106. Paper available at http://www.statcan.gc.ca/pub/12-001-x/2024001/article/00004-eng.pdf.

Chen, Y., Li, P. and Wu, C. (2020). Doubly robust inference with non-probability survey samples. *Journal of the American Statistical Association*, 115, 2011-2021.

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, 2, 283-311. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2022002/article/00002-eng.pdf.