STATISTICS IN TRANSITION new series, June 2023
Vol. 24, No. 3, pp. 31–37, https://doi.org/10.59170/stattrans-2023-032
Received – 25.05.2023; accepted – 31.05.2023

Discussion of "Probability vs. Nonprobability Sampling: From the Birth of Survey Sampling to the Present Day" by Graham Kalton

Julie Gershunskaya[1], **Partha Lahiri**[2]

In this excellent overview of the history of probability and nonprobability sampling from the end of the nineteenth century to the present day, Professor Graham Kalton outlines the essence of past endeavors that helped to define philosophical approaches and stimulate the development of survey sampling methodologies. From the beginning, there was an understanding that a sample should, in some ways, resemble the population under study. In Kiær's ideas of "representative sampling" and Neyman's invention of probability-based approach, the prime concern of survey sampling has been to properly plan for representing characteristics of the finite population. Poststratification and other calibration methods were developed for the same important goal of better representation.

Professor Kalton's paper underscores growing interest in the use of nonprobability surveys. With recent proliferation of computers and the internet, wealth of data becomes available to researchers. However, "opportunistic" information collected with present-day capabilities usually is not purposely planned or controlled by survey statisticians. No matter how big such a nonprobability sample could be, it may inaccurately reflect the finite population of interest, thus presenting a substantial risk of an estimation bias.

Below, we discuss several recent papers that propose ways to incorporate nonprobability surveys to produce estimates for both large and small areas. Specifically, we will consider two situations often encountered in practice. In the first situation, a nonprobability sample contains the outcome variable of interest, and the main task is to reduce the selection bias with the help of a reference probability sample that does not contain the outcome variable of interest. In the second situation, a probability sample contains the outcome variable of interest, but there is little or no sample available to produce granular level estimates. For such a small area estimation problem, we consider a case when we have access to a large nonprobability sample that does not contain the outcome variable but contains some related auxiliary variables also present in the probability sample. In both situations, researchers have discussed statistical data integration techniques in which a reference probability sample is combined with a nonprobability sample in an effort to overcome deficiencies associated with both probability and nonprobability samples.

[1]U.S. Bureau of Labor Statistics, 2 Massachusetts Ave NE Washington, DC 20212, USA, E-mail: Gershunskaya.Julie@bls.gov. ORCID: https://orcid.org/0000-0002-0096-186X.
[2]University of Maryland, College Park, MD 20742. USA. E-mail:plahiri@umd.edu. ORCID: https://orcid.org/0000-0002-7103-545X.

One way to account for the selection bias of a nonprobability sample is by estimating the sample inclusion probabilities, given available covariates. Then, the inverse values of estimated inclusion probabilities are used, in a similar manner as the usual probability sample selection weights, to obtain estimates of target quantities. Several approaches to estimation of nonprobability sample inclusion probabilities (or propensity scores) have been considered in the literature. Recent papers by Chen et al. (2020), Wang et al. (2021), and Savitsky et al. (2022) propose ways to estimate these probabilities based on combining nonprobability and probability samples. Kim J. and K. Morikawa (2023) propose an empirical likelihood based approach under a different setting. To save space, we will not discuss their approach. We now review three statistical data integration methods.

The approaches concern with the estimation of probabilities $\pi_{ci}(x_i) = P\{c_i = 1|x_i\}$ to be included into the nonprobability sample S_c, for units $i = 1,\ldots,n_c$, where c_i is the inclusion indicator of unit i taking on the value of 1 if unit i is included into the nonprobability sample, and 0 otherwise; x_i is a vector of known covariates for unit i; n_c is the total number of units in sample S_c. The problem, of course, is that we cannot estimate π_{ci} based on the set of units in nonprobability sample S_c alone, because $c_i = 1$ for all i in S_c. The probabilities are estimated by combining set S_c with a probability sample S_r. Due to its role in this approach, the probability sample here is also called "reference sample".

Assuming both nonprobability and probability samples are selected from the same finite population P, Chen et al. (2020) write a log-likelihood, over units in P, for the Bernoulli variable c_i :

$$\ell_1(\boldsymbol{\theta}) = \sum_{i\in P}\{c_i \log[\pi_{ci}(x_i,\boldsymbol{\theta})] + (1-c_i)\log[1-\pi_{ci}(x_i,\boldsymbol{\theta})]\}, \tag{1}$$

where $\boldsymbol{\theta}$ is the parameter vector in a logistic regression model for π_{ci}.

Since finite population units are not observed, Chen et al. (2020) employ a clever trick and re-group the sum in (1) by presenting it as a sum of two parts: part 1 involves the sum over the nonprobability sample units and part 2 is the sum over the whole finite population:

$$\ell_1(\boldsymbol{\theta}) = \sum_{i\in S_c}\log\left[\frac{\pi_{ci}(x_i,\boldsymbol{\theta})}{1-\pi_{ci}(x_i,\boldsymbol{\theta})}\right] + \sum_{i\in P}\log[1-\pi_{ci}(x_i,\boldsymbol{\theta})]. \tag{2}$$

Units in part 1 of the log-likelihood in (2) are observed; for part 2, Chen et al. (2020) employ the pseudo-likelihood approach by replacing the sum over the finite population with its probability sample based estimate:

$$\hat{\ell}_1(\boldsymbol{\theta}) = \sum_{i\in S_c}\log\left[\frac{\pi_{ci}(x_i,\boldsymbol{\theta})}{1-\pi_{ci}(x_i,\boldsymbol{\theta})}\right] + \sum_{i\in S_r} w_{ri}\log[1-\pi_{ci}(x_i,\boldsymbol{\theta})], \tag{3}$$

where weights $w_{ri} = 1/\pi_{ri}$ are inverse values of the reference sample inclusion probabilities π_{ri}. Estimates are obtained by solving respective pseudo-likelihood based estimating equations.

One shortcoming of the Chen et al. (2020) approach is that their Bernoulli likelihood is formulated with respect to an unobserved indicator variable. Although the regrouping

employed in (2) helps to find a solution, results obtained by Wang et al. (2021) indicate that it is relatively inefficient, especially when the nonprobability sample size is much larger than the probability sample size.

Wang et al. (2021) formulate their likelihood for an *observed* indicator variable and thus their method is different from the approach of Chen et al. (2020). To elaborate, Wang et al. (2021) introduce an imaginary construct consisting of two parts: they *stack* together nonprobability sample S_c (part 1) and finite population P (part 2). Since nonprobability sample units belong to the finite population, they appear in the stacked set twice. Let indicator variable $\delta_i = 1$ if unit i belongs to part 1, and $\delta_i = 0$ if i belongs to part 2 of the stacked set; the probabilities of being in part 1 of the stacked set are denoted by $\pi_{\delta i}(x_i) = P\{\delta_i = 1|x_i\}$. Wang et al. (2021) assume the following Bernoulli likelihood for observed variable δ_i:

$$\ell_2(\tilde{\boldsymbol{\theta}}) = \sum_{i \in S_c} \log\left[\pi_{\delta i}(x_i, \tilde{\boldsymbol{\theta}})\right] + \sum_{i \in P} \log\left[1 - \pi_{\delta i}\left(x_i, \tilde{\boldsymbol{\theta}}\right)\right], \tag{4}$$

where $\tilde{\boldsymbol{\theta}}$ is the parameter vector in a logistic regression model for $\pi_{\delta i}$. Since the finite population is not available, they apply the following pseudo-likelihood approach:

$$\hat{\ell}_2(\tilde{\boldsymbol{\theta}}) = \sum_{i \in S_c} \log\left[\pi_{\delta i}(x_i, \tilde{\boldsymbol{\theta}})\right] + \sum_{i \in S_r} w_{ri} \log\left[1 - \pi_{\delta i}\left(x_i, \tilde{\boldsymbol{\theta}}\right)\right]. \tag{5}$$

Existing ready-to-use software can be used to obtain estimates of $\pi_{\delta i}$. However, the actual goal is to find probabilities π_{ci} rather than probabilities $\pi_{\delta i}$. Wang et al. (2021) propose a two-step approach, where at the second step, they find π_{ci} by employing the following identity:

$$\pi_{\delta i} = \frac{\pi_{ci}}{1 + \pi_{ci}}. \tag{6}$$

Savitsky et al. (2022) use an exact likelihood for the estimation of inclusion probabilities π_{ci}, rather than a pseudo-likelihood based estimation. They propose to stack together nonprobability, S_c, and probability, S_r, samples. In this stacked set, S, indicator variable z_i takes the value of 1 if unit i belongs to the nonprobability sample (part 1), and 0 if unit i belongs to the probability sample (part 2). In this construction, if there is an overlap between the two samples, S_c and S_r, then the overlapping units are included into stacked set S twice: once as a part of the nonprobability sample (with $z_i = 1$) and once as a part of the reference probability sample (with $z_i = 0$). We do not need to know which units overlap or whether there are any overlapping units. The authors use first principles to prove the following relationship between probabilities $\pi_{zi}(x_i) = P\{z_i = 1|x_i\}$ of being in part 1 of the stacked set and the sample inclusion probabilities π_{ci} and π_{ri}:

$$\pi_{zi} = \frac{\pi_{ci}}{\pi_{ri} + \pi_{ci}}. \tag{7}$$

A similar expression (7) was derived by Elliott (2009) and Elliott and Valliant (2017) under the assumption of non-overlapping nonprobability and probability samples. The derivation given in Savitsky et al. (2022) does not require this assumption.

To obtain estimates of π_{ci} from the combined sample, Beresovsky (2019) proposed to parameterize probabilities $\pi_{ci} = \pi_{ci}(x_i, \boldsymbol{\theta})$, as in Chen et al. (2020), and employ identity (7) to present π_{zi} as a composite function of $\boldsymbol{\theta}$; that is, $\pi_{zi} = \pi_{zi}(\pi_{ci}(x_i, \boldsymbol{\theta})) = \pi_{ci}(x_i, \boldsymbol{\theta})/(\pi_{ri} + \pi_{ci}(x_i, \boldsymbol{\theta}))$.

The log-likelihood for observed Bernoulli variable z_i is given by

$$\ell_3(\boldsymbol{\theta}) = \sum_{i \in S_c} \log\left[\pi_{zi}(\pi_{ci}(x_i, \boldsymbol{\theta}))\right] + \sum_{i \in S_r} \log\left[1 - \pi_{zi}\left(\pi_{ci}(x_i, \boldsymbol{\theta})\right)\right]. \tag{8}$$

Since the log-likelihood *implicitly* includes a logistic regression model formulation for probabilities π_{ci}, Beresovsky (2019) labeled the proposed approach Implicit Logistic Regression (ILR). For the maximum likelihood estimation (MLE), the score equations are obtained from (8) by taking the derivatives, with respect to $\boldsymbol{\theta}$, of the composite function $\pi_{zi} = \pi_{zi}(\pi_{ci}(\boldsymbol{\theta}))$. This way, the estimates of π_{ci} are obtained directly from (8) in a single step. Savitsky et al. (2022) parameterized the likelihood, as in (8), and used the Bayesian estimation technique to fit the model.

Note that to implement the ILR approach, the reference sample inclusion probabilities π_{ri} have to be known for all units in the combined set. This is not a limitation for many probability surveys. As discussed in Elliott and Valliant (2017), if probabilities π_{ri} cannot be determined exactly for units in the nonprobability sample, they can be estimated using a regression model. Savitsky et al. (2022) used Bayesian computations to simultaneously estimate π_{ri} and π_{ci} for nonprobability sample units, given available covariates x_i.

It must be noted that the estimation method of Wang et al. (2021) can be similarly modified to avoid the two-step estimation procedure: a logistic regression model could be formulated for inclusion probabilities π_{ci}, while probabilities $\pi_{\delta i}$ in (6) could be viewed as a composite function, $\pi_{\delta i} = \pi_{\delta i}(\pi_{ci}(x_i, \boldsymbol{\theta})) = \pi_{ci}(x_i, \boldsymbol{\theta})/(1 + \pi_{ci}(x_i, \boldsymbol{\theta}))$. This approach is expected to be more efficient. Moreover, it avoids π_{ci} estimates greater than 1 that could occur when the estimation is performed in two steps. Once modified this way, preliminary simulations indicate that Wang et al. (2021) formulation would produce more efficient estimates than the Chen et al. (2020) counterpart, unless in a rare situation where the whole finite population rather than only a reference sample is available.

Simulations show that the exact likelihood method based on formulation of Savitsky et al. (2022) and Beresovsky (2019) performs better than the pseudo-likelihood based alternatives. In the usual situation where the reference probability sample fraction is small, the relative benefits of the exact likelihood approach are even more pronounced.

The existence of a well-designed probability reference sample plays a crucial role in the efforts to reduce the selection bias of a nonprobability sample. Importantly, an ongoing research indicates that the quality of estimates of the nonprobability sample inclusion probabilities is better if there is a good overlap in domains constructed using covariates from both samples. This observation harks back to problems appearing in traditional poststratification methods and to the notion of "representative sampling." Since survey practitioners usually do not have control over the planning or collection of the emerging multitude of nonrandom opportunistic samples, efforts should be directed to developing and maintaining comprehensive probability samples that include sets of good quality covariates. Beaumont et al. (2023)

proposed several model selection methods in application of the modeling nonprobability sample inclusion probabilities.

We now turn our attention to the second data integration situation involving small area estimation, a topic Professor Kalton touched on. This is a problem of great interest for making public policies, fund allocation and regional planning. Small area estimation programs already exist in some national statistical organizations such as the Small Area Income and Poverty Estimates (SAIPE) program of the US Census Bureau (Bell et al., 2016) and Chilean government system (Casas-Cordero Valencia et al., 2016.) The importance placed in the United Nations Sustainable Development Goals (SDG) for disaggregated level statistics is expected to increase the demand for such programs in various national statistical offices worldwide. Standard small area estimation methods generally use statistical models (e.g., mixed models) that combine probability sample data with administrative or census data containing auxiliary variables correlated with the outcome variable of interest. For a review of different small area models and methods, see Jiang and Lahiri (2006), Rao and Molina (2015), Ghosh (2020), and others.

A key to success in small area estimation is to find relevant auxiliary variables not only in the probability sample survey but also in the supplementary big databases. Use of a big probability or nonprobability sample survey could be useful here as surveys typically contain a large number of auxiliary variables that are also available in the probability sample survey. In the context of small area estimation, Sen and Lahiri (2023) considered a statistical data integration technique in which a small probability survey containing the outcome variable of interest is statistically linked with a much bigger probability sample, which does not contain the outcome variable but contains many auxiliary variables also present in the smaller sample. They essentially fitted a mixed model to the smaller probability sample that connects the outcome variable to a set of auxiliary variables and then imputed the outcome variable for all units of the bigger probability sample using the fitted model and auxiliary variables. Finally, they suggested to produce small area estimates using survey weights and imputed values of the outcome variable contained in the bigger probability sample survey. As discussed in their paper, such a method can be used even if the bigger sample is a nonprobability survey using weights constructed by methods such as the ones described earlier.

The development of new approaches demonstrates how the methods of survey estimation continue to evolve by taking into the future the best from fruitful theoretical and methodological developments of the past. As Professor Kalton highlights, we will increasingly encounter data sources that are not produced by standard probability sample designs. Statisticians will find ways to respond to new challenges, as is reflected in the following amusing quote:

> ...D.J. Finney once wrote about the statistician whose client comes in and says, “Here is my mountain of trash. Find the gems that lie therein." Finney’s advice was to not throw him out of the office but to attempt to find out what he considers "gems". After all, if the trained statistician does not help, he will find someone who will....(source: David Salsburg, ASA Connect Discussion)

Of course, nonprobability samples should not be viewed as a "mountain of trash." Indeed, they can contain a lot of relevant information for producing necessary estimates. It is just that one needs to explore different innovative ways to use information contained in nonprobability samples. In the United States federal statistical system, the need to innovate for combining information from multiple sources has been emphasized in the National Academies of Sciences and Medicine (2017) report on Innovations in Federal Statistics. As discussed, statisticians have been already engaged in suggesting new ideas, such as statistical data integration, to extract information out of multiple non-traditional databases. In coming years, statisticians will be increasingly occupied with finding solutions for obtaining useful information from non-traditional data sources. This is indeed an exciting time for survey statisticians.

References

Beaumont, J.-F., K. Bosa, A. Brennan, J. Charlebois, and K. Chu (2023). Handling nonprobability samples through inverse probability weighting with an application to statistics canada's crowdsourcing data. Survey Methodology (accepted in 2023 and expected to appear in 2024).

Bell, W. R., W. W. Basel, and J. J. Maples (2016). *An overview of the US Census Bureau's small area income and poverty estimates program*, pp. 349–378. Wiley Online Library.

Beresovsky, V. (2019). On application of a response propensity model to estimation from web samples. In ResearchGate.

Casas-Cordero Valencia, C., J. Encina, and P. Lahiri (2016). *Poverty mapping for the Chilean Comunas*, pp. 379–404. Wiley Online Library.

Chen, Y., P. Li, and C. Wu (2020). Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association 115*(532), 2011–2021.

Elliott, M. R. (2009). Combining data from probability and non-probability samples using pseudo-weights. *Survey Practice 2*, 813–845.

Elliott, M. R. and R. Valliant (2017). Inference for Nonprobability Samples. *Statistical Science 32*(2), 249 – 264.

Ghosh, M. (2020). Small area estimation: Its evolution in five decades. *Statistics in Transition New Series, Special Issue on Statistical Data Integration*, 1–67.

Jiang, J. and P. Lahiri (2006). Mixed model prediction and small area estimation, editor's invited discussion paper. *Test 15*, 1–96.

Kim J. and K. Morikawa (2023). An empirical likelihood approach to reduce selectionbias in voluntary samples. *Calcutta Statistical Association Bulletin 35 (to appear)*.

National Academies of Sciences, E. and Medicine (2017). *Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy*. Washington, DC: The National Academies Press.

Rao, J. N. K. and I. Molina (2015). *Small Area Estimation, 2nd Edition*. Wiley.

Savitsky, T. D., M. R. Williams, J. Gershunskaya, V. Beresovsky, and N. G. Johnson (2022). Methods for combining probability and nonprobability samples under unknown overlaps. https://doi.org/10.48550/arXiv.2208.14541.

Sen, A. and P. Lahiri (2023). Estimation of finite population proportions for small areas: a statistical data integration approach. https://doi.org/10.48550/arXiv.2305.12336.

Wang, L., R. Valliant, and Y. Li (2021). Adjusted logistic propensity weighting methods for population inference using nonprobability volunteer-based epidemiologic cohorts. *Stat Med. 40*(4), 5237–5250.