Estimation Methods for Nonprobability Samples with a Companion Probability Sample

Michael Yang, Nada Ganesh, Edward Mulrow, and Vicki Pineau
NORC at the University of Chicago, 4350 East-West Highway. 8th Floor, Bethesda, MD 20814

Abstract
Probability sampling has been the standard basis for inference from a sample to a target population. In the era of big data and increasing data collection costs, however, there has been growing demand for estimation methods to combine probability and nonprobability samples in order to improve the cost efficiency of survey estimation without loss of statistical accuracy (or perhaps even with improvements in statistical accuracy). An array of methods for combining probability and nonprobability samples are found in the literature, which we have classified into the following methodological groups: calibration, statistical matching, super-population modeling, and propensity-based weighting. In addition, NORC researchers have developed a hybrid calibration method that incorporates "borrowed strength" methods from small area estimation in order to explicitly account for bias associated with the nonprobability sample. We compare and contrast the nonprobability weights and estimates derived from all the methods from food allergies survey data, which were collected via both a probability sample and a nonprobability sample.

Key Words: probability sample, nonprobability sample, fit-for-purpose

1. Introduction

While probability sampling remains the gold standard for survey estimation, often the incidence rate for a study's target population is so low that complete sampling frames are not available, or probability sampling methods for surveying the target population are too expensive. Thus, there has been growing demand for methods that use nonprobability samples and methods that combine probability and nonprobability samples in order to improve the statistical and cost efficiency of survey estimation.

Nonprobability samples may provide a lower cost alternative to probability samples; however, the quality of the data is oftentimes low, and in particular estimates based on nonprobability samples may be biased. A well thought out approach to using nonprobability samples, alone or in conjunction with probability samples, should provide a way to assess the quality of the data and determine its fitness for use. This paper reports some preliminary results into our research about estimation methods based on both a probability and a nonprobability sample. Specifically, using data collected by NORC, this paper compares the distribution of the nonprobability sampling weights (all of which are effectively modeled weights) and the weighted estimates under five different estimation methods. Although the probability sample may be used in modeling the nonprobability sample weights, the comparisons reported here are based on the nonprobability sample alone. Further results based on the combination of probability and nonprobability samples will be reported in a subsequent paper.

2. Methods Investigated

Researchers have proposed and experimented with a range of estimation methods based on nonprobability samples for decades. More recently, there has been increased interest in estimation methods that use both probability and nonprobability samples. We conducted a literature review to identify and delineate methods reported in journals, workshops, and conferences. Our focus was on reported studies in recent years that

represent research developments and/or empirical results related to nonprobability sample estimation. We documented the underlying statistical models, the model-based estimators, and the properties of the estimators, to the extent possible.

We identified four general estimation methodologies for nonprobability samples:

- **Calibration:** Calibrate total estimates to population benchmarks.
- **Statistical Matching (Matching):** Statistically match nonprobability and probability samples.
- **Super-population Modeling (Modeling):** Use a super-population model to derive population estimates.
- **Propensity Weighting (Propensity):** Model the propensity to be included in a nonprobability sample.

A fifth method has been developed by NORC's AmeriSpeak® team (Ganesh, et al, 2017):

- **Borrowed Strength (Borrowed)**: Relies on methods commonly used in Small Area Estimation.

For this preliminary investigation, the final outcome of each of the five methods is a set of weights for the nonprobability sample. Therefore, the five methods represent five different approaches to estimating the weights for the nonprobability sample. Of the five methods, Statistical Matching, Borrowed Strength, and Propensity Weighting rely on the availability of a probability sample, while Calibration and Super-population Modeling do not[1]. However, the latter two methods do rely on known population totals for covariates that are used for calibration or the super-population model. We now briefly describe each of the five methods as implemented in this study.

2.1 Calibration

The Calibration method involves constructing the weights such that estimates of totals from the sample match population benchmarks. The weights are developed through iterative proportional fitting (e.g. raking). Our approach is based on DiSogra et al. (2011) and Fahimi et al. (2015), and involves additional raking variables that presumably differentiate the nonprobability sample from probability samples. When the nonprobability sample is an opt-in online sample, these additional "webographic" raking variables, such as "early adopter" variables, attempt to correct for the nonprobability sample's higher likelihood of including respondents who are more likely to be users of new products and technology.

The calibration weights for the nonprobability sample was developed via the following steps:

- Assign a weight of 1 to all nonprobability sample units.
- Rake the weights to known demographic control totals from the Current Population Survey (CPS). The demographic variables include age, gender, education, race/ethnicity, geography, and income.
- Rake the weights further to match webographic control totals.

The webographic control totals are estimated using NORC's AmeriSpeak® Panel.

2.2 Statistical Matching

The goal is to match the nonprobability sample units to the probability sample units. The matching is carried out through a nearest neighbor hot deck algorithm based on a distance measure. The matching process

[1] As a general estimation methodology, calibration often uses probability samples.

resembles imputation in the sense that a donor from the probability sample is matched to a recipient from the nonprobability sample based on a set of matching variables (Bethlehem, 2015). Statistical matching is carried out using the R *StatMatch* package *RANDwNND.hotdeck* function (D'Orazio, 2017). Each nonprobability sample unit is matched to one and only one probability sample unit under the following conditions:

- A match is made by finding a pool of probability sample units with the 20 closest distances to a nonprobability unit, and randomly selecting one unit from the pool.
- Distances are measured using Gower's dissimilarity measure, which can use both categorical and continuous variables in the dissimilarity calculation.
- The nonprobability unit assumes the weight of the matched probability unit. However, when a probability unit is matched to multiple nonprobability units, each matched nonprobability unit weight is the probability unit weight divided by the number of matches.

We used Gradient Boosting (D'Orazio, Di, and Scanu, 2006) to determine the set of matching variables for use in the Gower dissimilarity calculation. The final matching variable set includes 15 demographic and webographic variables.

2.3 Super-population Modeling

This is referred to as the super-population approach as discussed in Elliot and Valliant (2017). Other important references for modeling approaches include Dever and Valliant (2010, 2016). Under this approach, we fit a model for the dependent Y variable, and then use the model to predict the Y for the entire population. The underlying model adopted here is a linear one. Under this model, the predictor of the population total can be expressed as a weighted function of the observed Y variable. Note that this approach could also be used with a probability sample, as discussed in Valliant, Dorfman, and Royall (2000). The difference is that design-based inference is not an option for a nonprobability sample.

To introduce the modeling approach for a nonprobability sample, consider estimating a finite population total. The general idea in model-based estimation is to sum the responses for the sample cases and add to them the sum of predictions for nonsample cases. The key to unbiased estimates is that the variables to be analyzed for the sample and nonsample follow a common model and that this model can be discovered by analyzing the sample responses. When both the sample and nonsample units follow the same model, model parameters can be estimated from the sample and used to make predictions for the nonsample cases. An appropriate model usually includes covariates which are known for each individual sample case. The covariates are usually unknown for individual nonsample cases.

Suppose that the mean of a variable y_i follows a linear model:

$$E_M(y_i|X_i) = X_i'\boldsymbol{\beta},$$

where the subscript *M* denotes the expectation with respect to the model, X_i is a vector of *p* covariates for unit *i* and $\boldsymbol{\beta}$ is a parameter vector. Given a sample *s*, an estimator of the slope parameter is

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}_s'\boldsymbol{X}_s)^{-1}\boldsymbol{X}_s'y_s$$

A predictor of the y population total is:

$$\hat{t} = \sum_{i\in s} y_i + (\boldsymbol{t}_{Ux} - \boldsymbol{t}_{sx})'\widehat{\boldsymbol{\beta}}$$

where $\boldsymbol{t}_{Ux}$ and $\boldsymbol{t}_{sx}$ are vectors of X totals for the population and sample, respectively. The estimated total $\hat{t}$ can be written as the weighted sum of the observed y's where the weights are:

$$w_i = 1 + (\boldsymbol{t}_{Ux} - \boldsymbol{t}_{sx})'(\boldsymbol{X}_s'\boldsymbol{X}_s)^{-1}X_i$$

This estimator is equal to the GREG (Särndal, Swensson and Wretman, 1992) if the inverse selection probabilities in that estimator are all set to 1. Note that these weights depend only on the *x*'s and not on *y*. As a result, the same set of weights can be used for all estimates. A single set of weights will not be equally efficient for every *y*, but this situation is also true for design-based weights.

2.4 Propensity Weighting

This is the propensity weighting or quasi-randomization approach as discussed in Elliot and Valliant (2017). It requires the presence of a probability sample, called a reference sample, selected from the same population. Under this approach, we fit a regression model to estimate the inclusion probability of the nonprobability units, and then use the predicted probabilities to derive the nonprobability sample weights or pseudo weights. Here are the steps for developing the propensity weights:

- Concatenate the probability sample and the nonprobability sample;
- Create a dichotomous variable, *Y*, which is coded 1 for nonprobability sample units and 0 for probability sample units;
- Fit a logistic regression model with *Y* as the response variable;
- Use the predicted propensities as the estimated inclusion probabilities for the nonprobability sample units;
- Compute the nonprobability sample weights as the inverse of the predicted inclusion probabilities.

Predictor variables in the logistic regression model include demographic (e.g., age, gender, race and ethnicity, marital status), socioeconomic (e.g., education, income, employment) webographic, and some attitudinal/behavioral variables. The final model is validated through cross validation and by examining model diagnostic statistics.

2.5 Borrowed Strength

To borrow strength across domains, we used small area estimation methods to model domain-level (geographic, population subgroup) estimates from the probability and the nonprobability sample (Ganesh et al., 2017). Each model includes a set of covariates (X), domain-level random effects, and sampling errors. The nonprobability model also includes fixed and random bias terms. The nonprobability sample weights are developed via the following steps:

- A Bivariate Fay-Herriot model (Rao, 2003) is used to jointly model the domain-level point estimates of food allergy incidence (arcsine transformed) from the probability sample (y_d^P) and the nonprobability sample (y_d^{NP}) :

$$y_d^P = \mathbf{x}_d'\boldsymbol{\beta} + \nu_d + \varepsilon_d^P$$
$$y_d^{NP} = b + \alpha_d^{NP} + \mathbf{x}_d'\boldsymbol{\beta} + \nu_d + \varepsilon_d^{NP}$$

  - d is a demographic group (e.g. 18-34 year old, male, Hispanic).
  - $\mathbf{x}_d$ is a vector of covariates.
  - ν_d's are domain level random effects.
  - b is a fixed effect bias term associated with the nonprobability sample estimate.
  - α_d's are random effect bias terms associated with the nonprobability sample estimate.

- ε_d^P, ε_d^{NP} are the sampling errors associated with y_d^P, y_d^{NP}.
- Small area estimates for each domain were obtained using an Empirical Best Linear Unbiased Predictor (EBLUP).
- Nonprobability sample weights are derived such that nonprobability-based estimates (using the weights) match the small area estimates for each domain.

The subgroup domains are defined by cross-classifying a set of demographic variables:

- Age (18-34 years, 35-49 years, 50-64 years, 65+ years),
- Education (Some college or less, college graduate or higher),
- Race/Hispanic ethnicity (Hispanic, non-Hispanic Black, non-Hispanic All Other), and
- Gender (male, female)

The choice of domains was motivated by "sufficient" sample size for the probability sample, and also to capture the variation in the substantive estimates across domains.

3. Test Data

In order to test and compare these methods, we used the Food Allergy Survey data that NORC collected in 2016 on behalf of Northwestern University. The main focus of the Food Allergy Survey is to measure the adult and child prevalence of self-reported and doctor-diagnosed food allergies, both current and outgrown, allergy reactions, experiences in allergy treatments, events coinciding with development or outgrowing a food allergy, and perceived risks associated with food allergies. Additional topics for all respondents, regardless of having a food allergy or not, included chronic conditions, family history of food allergy and chronic conditions, and contact with animals.

The Food Allergy data were collected via both a probability sample and a nonprobability sample. The probability sample was selected from NORC's AmeriSpeak® Panel, a multistage probability sample selected from NORC's National Frame that represents the U.S. household population. The probability sample produced 7,218 completed surveys. The nonprobability sample was selected from SSI's nonprobability web panel, consisting of 33,331 completed surveys. Only the adult data were used for testing the methods of this paper.

Both the probability and nonprobability samples contain a large number of demographic and webographic variables. Demographic variables include: Age, Gender, Race/ethnicity, Education, Employment, Marital Status, Household income, Household size (including children), Home ownership, Household telephone service, and more. Webographic variables include household internet access among others.

4. Comparisons of Weights and Weighted Estimates

All five methods investigated are ultimately model-based, but the underlying model and the number and nature of the covariates differ across the methods. We now compare the resulting weights and the weighted estimates across these methods.

4.1 Weight Distributions

We used each method to derive pseudo weights for the nonprobability sample. To enable a reasonable comparison across the different methods, each set of weights was scaled so that the sum of the weights equaled the nonprobability sample size of 33,331. Figure 1 is a distribution comparison via boxplots of the weight distributions for the five methods.

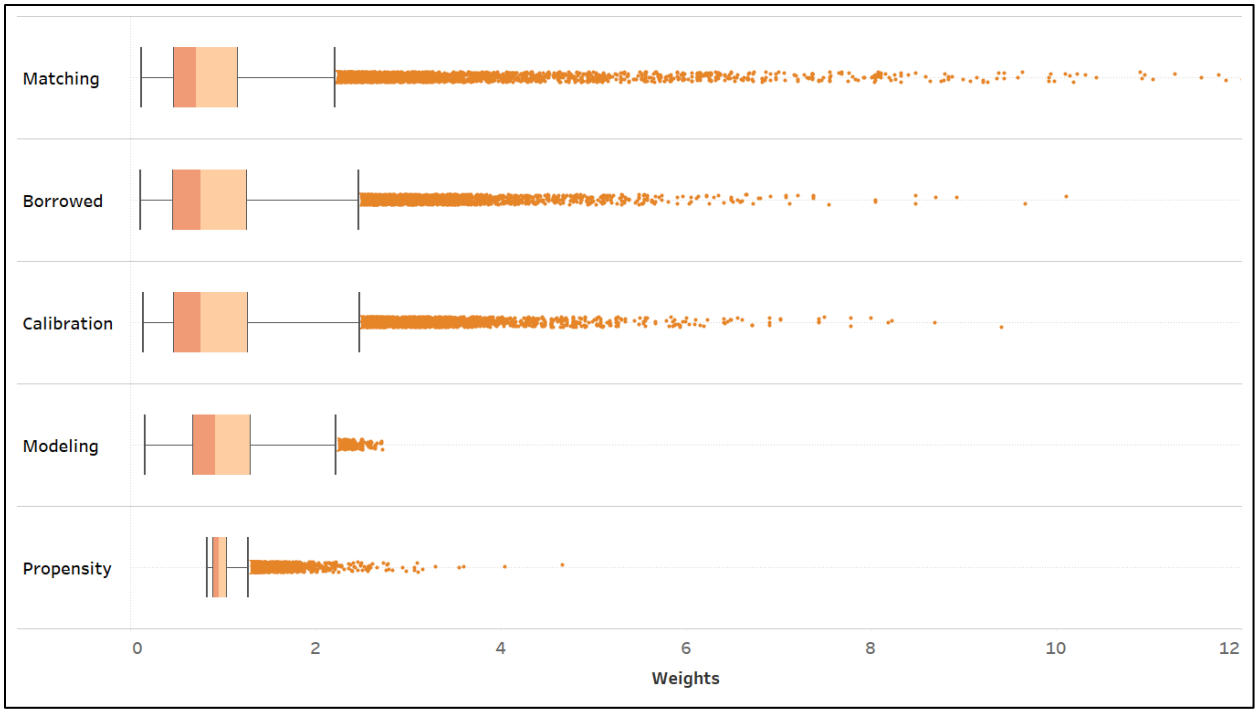

Figure 1: Nonprobability sample scaled weights distribution comparisons across methods via boxplots. The five methods are ordered by median weight. Matching, i.e. Statistical Matching, has 22 weights greater than 12, which are not shown in the plot. The maximum Statistical Matching weight has value 28.12.

Figure 1 shows that the distribution of the modeled nonprobability sample weights varies greatly across the five methods. The weights under the Statistical Matching method have the largest variance, followed by Borrowed Strength and Calibration. The Super-population Modeling and Propensity Weighting methods generate much less weight variations. The variations might be the results of different sets of covariates used. Methods that lead to greater weight variations will likely produce higher standard errors for population estimates compared to those that generate less weight variations.

4.2 Nonprobability Estimates of Key Variables

Using the weights derived from each method, we estimated the percentage of individuals with food allergies and other doctor diagnosed chronic conditions. Comparison results are shown in Table 1. Estimates based on the probability sample are also provided as a reference, along with the upper and lower 95% confidence bounds.

Table 1: Estimated Percentage of Individuals with Chronic Conditions by Method
Methods are ordered left to right, and survey variables top to bottom, by the estimate count outside the confidence bounds.

Chronic Condition Variable	Probability Sample Estimates			Nonprobability Sample Estimates				
Have you ever had…	LCB	Mean	UCB	Calibration	Propensity	Matching	Borrowed	Modeling
Doctor diagnosed Asthma	12.0	13.1	14.2	13.1	12.9	13.2	12.0	12.2
Doctor diagnosed Uricaria/chronic hives	0.8	1.0	1.3	1.0	1.0	0.9	0.8	0.8
Doctor diagnosed EoE	0.1	0.2	0.3	0.3	0.3	0.3	0.2	0.2
Doctor diagnosed Diabetes	8.9	9.6	10.4	10.6	10.1	9.7	10.1	10.1
Doctor diagnosed FPIES	0.1	0.2	0.3	0.3	0.3	0.4	0.2	0.3
Doctor diagnosed Eczema	6.7	7.5	8.3	7.0	6.7	6.7	6.4	6.4
Doctor diagnosed Insect sting allergy	3.8	4.4	5.0	4.0	3.9	3.8	3.6	3.5
Doctor diagnosed Latex allergy	2.4	2.8	3.2	2.4	2.6	2.4	2.1	2.2
Doctor diagnosed seasonal allergies	21.8	23.1	24.3	21.9	21.3	21.3	20.6	20.5
A food allergy during your lifetime	20.3	21.6	22.8	28.6	27.8	28.5	21.7	27.5
Doctor diagnosed Medication allergy	14.8	15.8	16.9	13.2	13.2	12.5	12.2	12.5
No doctor diagnosed chronic conditions	46.3	47.9	49.5	51.9	52.5	51.5	52.3	52.4
Doctor diagnosed Other condition	7.9	8.7	9.6	7.1	7.2	6.7	6.7	6.6

Outside confidence bounds

The weighted estimates of the 13 key survey variables are largely comparable across the methods. However, there are exceptions. About half of the weighted estimates from the nonprobability sample are outside the 95% confidence limits of the probability sample estimates (highlighted in Table 1). This does not necessarily mean that the difference is statistically significant as these comparisons do not take into account the variance associated with the nonprobability sample estimates. Some of these differences may well be due to sampling error.

Table 1 shows that, based on the number of estimates outside the confidence bounds, weighted estimates under Calibration, Propensity Weighting, and Statistical Matching tend to be closer to probability sample estimates. However, further research is needed to understand the relative performance of the methods under different circumstances.

5. Summary and Future Research

Estimation methods for nonprobability samples are needed to reduce the costs of survey statistics and take advantage of the growth of big data. In the absence of clear theoretical guidance, however, we may need to rely on empirical research to increase our knowledge about the properties of various nonprobability samples. We hope that the accumulation of such knowledge will help us to develop fit-for-purpose nonprobability estimation methods for different types of nonprobability samples and survey variables. For example, opt-in web panels may differ from the general populations in predictable ways and such knowledge should be taken into account in developing estimation methods for opt-in web panels. Empirical research may also help to identify the key covariates for different types of survey variables. For example, basic demographic and socioeconomic variables, the most accessible covariates, may be highly predictive of some variables, but they may be very insufficient for predicting many other variables.

In this study, we produced five sets of pseudo weights for a nonprobability sample based on five different methods that have been explored by practitioners in recent years. Some of these methods require the presence of a probability sample while others do not. All five methods are model-based and depend on the use of some covariates although the number and nature of the covariates differ across the methods. We have relied mostly on demographic and webographic variables as covariates in implementing these methods. Based on our comparative analysis, there is no apparent "best" choice, which is hardly surprising. While the distributions of the modeled weights are different across the methods, all the methods produce similar point estimates. Weighted estimates based on the nonprobability sample are frequently outside the confidence bounds of the estimates based on the probability sample, but the absolute differences tend to be quite small. In the absence of standard error estimates for the nonprobability sample, we were not able to perform rigorous significance tests at this preliminary stage. That will be one of the main goals for future research.

Estimation methods with smaller variance are generally preferable. With nonprobability samples, however, estimation bias may be a much greater concern, which suggests that methods that produce larger variances may be preferable in the absence of bias estimates. In large nonprobability samples bias is likely to be the most important source of error. A key advantage of the Borrowed Strength approach is its ability to produce bias estimates—the fixed effect bias term b and the random effect bias terms α_d^{NP}. Further research is needed to explore bias estimation methods with the other methods. For example, based on the matching results, the Statistical Matching method may help to generate a measure of bias associated with the nonprobability sample. We plan to explore this in future research. If a bias estimate is available, we can combine it with standard variance estimates to support mean squared error estimation. In any case, the use of Total Survey Error techniques may be warranted. In future research, we will test and compare the five methods with other data sources. We realize that some important response variables may be weakly correlated with the covariates. Therefore, we will also conduct more rigorous model evaluation in order to improve the underlying models.

Although this current research focuses on nonprobability samples, our main objective is to develop methods for combining probability and nonprobability samples. In principle, composite estimators can be used to combine probability and nonprobability samples. Under composite estimation, model-based estimates from the nonprobability sample will be combined with design-based estimates from the probability sample to derive the combined estimate.

Acknowledgements

The authors acknowledge the support provided by the following colleagues: Julia Batishev, Adrijo Chakraborty, Ying Li, Lin Liu, Qiao Ma, and Xian Tao.

References

Bethlehem J. (2015) "Solving the nonresponse problem with sample matching?" Social Science Computer Review, Vol. 34, Issue 1, pp. 59 – 77.

Dever, J. and Valliant, R. (2010). A comparison of variance estimators for poststratification to estimated control totals. Survey Methodology. 36 45–56.

Dever, J. and Valliant, R. (2016). GREG estimation with undercoverage and estimated controls. Journal of Survey Statistics and Methodology 4 289–318.

DiSogra, C., Cobb, C., Dennis, J.M. and Chan, E. 2011. Calibrating nonprobability Internet samples with probability samples using early adopter characteristics. Proceedings of the American Statistical Association, Section on Survey Research. Joint Statistical Meetings (JSM). Miami Beach, FL.

D'Orazio, M. (2017). StatMatch: Statistical Matching. R package version 1.2.5. https://CRAN.R-project.org/package=StatMatch

D'Orazio M., Di Zio M., Scanu M. (2006) Statistical matching: Theory and practice. Wiley, Chichester.
Elliot, M. R., Valliant, R. (2017). "Inference for Nonprobability Samples," Statistical Science 2017, Vol. 32, No. 2, 249–264.
Elliot, M. R., Valliant, R. (2017). Inference for Nonprobability Samples, Statistical Science 2017, Vol. 32, No. 2, 249–264
Fahimi, M., Barlas, F.M., Thomas, R.K. and Buttermore, N. (2015). "Scientific surveys based on incomplete sampling frames and high rates of nonresponse," Survey Practice, v.8 (5).
Fay, R.E., and Herriot, R.A. (1979). "Estimates of income for small places: An application of James-Stein procedures to Census data," Journal of the American Statistical Association, v. 74 (366), pp. 269-277.
Ganesh, N., Pineau, V., Chakraborty, A., Dennis, J.M., (2017). "Combining Probability and Non-Probability Samples Using Small Area Estimation." Joint Statistical Meetings 2017 Proceedings.
Rao, J.N.K. (2003). Small Area Estimation, John Wiley & Sons, Inc.
Sarndal, C. E., Swensson B. and Wretman, J. (1992). Model Assisted Survey Sampling. Springer, New York.
Valliant, R., Dorfman, A. H. and Royall, R. M. (2000). Finite Population Sampling and Inference: A Prediction Approach. Wiley, New York.
Ybarra, L.M.R. and Lohr, S.L. (2008). "Small area estimation when auxiliary information is measured with error," Biometrika, v. 95 (4), pp. 919-931.