A Comparison of Estimation Methods for Web-Based Respondent Driven Sampling

Kanru Xia[1], Vicki J Pineau[2], Nada Ganesh[3], Stuart Michaels[4], and Becky Reimer[5]
NORC at the University of Chicago

Abstract

Traditional probability-based sampling strategies are impractically expensive for surveying rare populations because of the large scale in-field screening required to find sufficient numbers of eligible persons. As a result, rare or hidden populations are often studied using versions of convenience samples, including non-probability Web panels, for which selection probabilities are unknown, design-based inference is not applicable, and study results not necessarily projectable to the target population. In an effort to improve methods for surveying rare or hidden population, NORC conducted a pilot study which tested a cost-effective alternative that combines probability sampling and Respondent-Driven Sampling (RDS) which we refer to as Web-based RDS. The initial, or seed, probability sample source for the pilot study is NORC's AmeriSpeak Panel®, which is a household, multi-client panel that uses the NORC National Frame to construct an address-based nationally representative sample panel of US households. The non-probability sample in the pilot study is generated from the seed sample using RDS methods in which respondents nominate/refer friends and family to take the survey. The target subpopulations for the pilot study are lesbian, gay, bisexual, and transgender (LGBT) Americans. We will present results from applying alternative estimation techniques to the Web-based RDS sample obtained in the pilot study, comparing the point estimates and their associated variances for health outcomes under each alternative to each other. NORC has been investigating various methods to combine probability and non-probability samples and of those a propensity approach and a small area estimation approach appear to be the most applicable to apply to the problem at hand. We will also compare estimates using an RDS estimator to the alternative estimators that combine probability and non-probability based samples.

Key Words: Respondent Driven Sampling, LGBT, Non-probability

1. Introduction

Surveying small or rare populations is important for policy making. When surveying rare populations, probability-based sampling has the advantage of providing a basis for formal statistical inference from a sample to the population but it also has its disadvantages. There is often no frame for the rare population, and researchers must sample a larger population and screen for the rare population. The screening cost can be impractically expensive due to the low eligibility and reluctance of members of the rare population to self-identify. To mitigate the cost, rare populations are often studied using non-probability samples, for which design-based statistical inference is not applicable, and study results

[1] Xia-kanru@norc.org; 55 E. Monroe St. 30th Fl., Chicago, IL 60603
[2] Pineau-vicki@norc.org; 4923 East Beach Drive, Oak Island, NC 28465
[3] Nada-Ganesh@norc.org; 4350 East-West Highway, Bethesda, MD 20814
[4] michaels-stuart@norc.org; 1155 East 60th Street, 2nd Floor, Chicago, IL 60637
[5] Reimer-Becky@norc.org; 55 E. Monroe St. 30th Fl., Chicago, IL 60603

might not be representative of the population (National Academies of Sciences 2018). Both probability and non-probability sampling have advantages and disadvantages.

In an effort to improve small population survey methods, NORC at the University of Chicago tested a cost-effective sampling alternative, referred to as Web-based respondent driven sampling (RDS), for the lesbian, gay, bisexual, and transgender (LGBT) population. Web-based RDS starts with an initial, or seed, sample of LGBT panelists from the NORC AmeriSpeak Panel®, which is a probability-based panel representative of the U.S. households.[6] From the seed sample, a non-probability sample is then constructed using RDS sampling, in which seed sample respondents nominate or refer friends and family to take the survey. This sampling method combines the strengths and advantages of both probability and non-probability sampling as well as the sample quality of AmeriSpeak®.

NORC at the University of Chicago has been investigating various estimation methods to combine probability and non-probability samples. In this paper, we investigated alternative estimation methods for Web-based RDS.

2. Methods

2.1 Data

In 2017, NORC at the University of Chicago conducted a Web-based RDS pilot study of the LGBT population aged 18 to 55. The sample was constructed using (A) panelists from AmeriSpeak as a seed sample, and (B) a RDS sample using generated from the seed sample. The seed sample drawn from the NORC AmeriSpeak Panel® included LGBT and non-LGBT panelists aged 18-55. The RDS referred sample included only LGBT respondents aged 18-55.

AmeriSpeak panelists are initially recruited for the panel using rigorous design-based sampling methods. US households are first sampled with a known, non-zero probability of selection from the NORC National Frame (an address-based sample) and then contacted by US mail, telephone interviewers, overnight express mailers, and field interviewers (face to face). It is important to note that the National Sample Frame from which the AmeriSpeak panel is constructed contains almost 3 million households, including over 80,000 rural households not available from the USPS Delivery Sequence File but identified by direct listing by field staff to ensure 99% coverage of U.S. households. AmeriSpeak recruited panelists fill out profile surveys on multiple topics and are asked whether they self-identify as LGBT.

For the pilot study, the probability based seed sample consisted of a random sample of AmeriSpeak LGBT panelists and a stratified random sample of the non-LGBT and unknown-LGBT-status panelists. A short web survey, focused on smoking behavior, was fielded to the sampled panelists. Each panelist that completed the survey was asked to refer LGBT friends and family to also participate in the pilot study. Referrals who completed the survey was also asked to refer their LGBT friends and family to complete the survey. The pilot included up to four rounds of referrals (Michaels). Table 1 shows the number of completes obtained in the pilot study by LGBT self-identification and by seed/referral status. For this paper, we exclude the non-LGBT seed sample and focus our estimation methods and analyses on the combined probability-based AmeriSpeak LGBT seed sample and the non-probability LGBT RDS sample.

[6] http://www.norc.org/Research/Capabilities/Pages/amerispeak.aspx

Table 1 Number of Completes from the Pilot Survey

	LGBT	Non-LGBT	Total
Seed	182	228	410
Referral	102	0	102
Total	284	228	512

2.2 Estimation Methods

We investigated three estimation methods using the pilot study survey data. The methods included an RDS estimation approach, a propensity approach, and a small area estimation approach.

The RDS estimation approach is a modified Voltz-Heckathorn (V-H) estimator. (Gile 2010) The V-H estimator treats the RDS sampling process as a random walk on the network connecting the target population. Each node on the network has a probability of selection proportional to the node's self-reported network size (reported size of friends and family). In this approach, the survey base weight is the inverse of the number of LGBT friends and family reported by each survey respondent. The base weights were raked to the population control totals, which were derived by combining 2016 National Health Interview Survey (NHIS) LGB population percentages and 2017 Current Population Survey (CPS) March supplement population totals. The demographics used in the raking included age group, gender, and race/Hispanic ethnicity.

The propensity approach fits a logistic regression model to estimate the inclusion probability of the nonprobability units (Yang 2018). The dependent variable is an indicator variable for non-probability (referral) sample vs. probability (seed) sample. The independent variables are respondent demographics and survey responses, excluding those compared in this paper. The final logistic model was chosen by stepwise selection and validated by cross-validation. In this approach, the survey base weight for the non-probability sample is the inverse of predicted probabilities from the logistic regression. The survey base weight for the probability sample is a product of AmeriSpeak® panel weight and the inverse of selection probability from the panel into the survey. Non-response adjustment was conducted for the probability sample using the weighting cell approach. The non-probability base weight and the probability non-response adjusted weights were separately raked to the same population control totals used in the RDS approach. Then, the raked probability and non-probability weights were combined based on the percent of respondents contributed from each sample source.

The small area estimation (SAE) approach models domain-level (geographic, population subgroup) estimates from the probability and the nonprobability sample to borrow strength across domains. A Bivariate Fay-Herriot model (Rao, 2003) was used to jointly model the domain-level point estimates of survey point estimates from the probability sample and the nonprobability sample. The model includes covariates, domain-level random effects, non-probability sample bias term, and sampling errors (Yang 2018). In this approach, the probability-sample-alone weight was calculated the same way as in the propensity approach, with the base weight constructed using the AmeriSpeak® panel weight and the inverse of selection probability into the seed sample. The base weight was then adjusted for differential non-response and raked to the population control totals. For the non-probability sample, the base weight was set to one and raked to the population control totals. We then generated weighted survey point estimates for the probability and non-probability samples by domains to use as dependent variables in SAE modeling. We defined domains by LGBT status (non-LGBT vs. LGBT), race/ethnicity (Non-Hispanic White vs. Others), and age group (18 to 34 vs. 35 to 55). Note that the non-LGBT probability sample was included in the modeling to increase the number of domains, from which we could borrow strength. The covariates are weighted point estimates of demographics (e.g., gender, marital status), socioeconomic (e.g., education, income, employment), and general health variables by domains from the NHIS and CPS surveys. We avoided smoking-behavior variables to simulate a more realistic scenario, in which researchers are unlikely to find covariates that exactly match the independent variable in subject matter. For each domain, we obtained small area estimates using an Empirical Best Linear Unbiased Predictor (EBLUP). Then the probability and non-probability raked weights were combined based on the percent of completes from each sample source and raked to the population control totals plus the modeled estimates obtained from the SAE model.

2.3 Comparing Methods

To compare the estimation methods, we calculated the root-mean-square-error (RMSE) of the weighted survey point estimates for each method for the following three survey estimated proportions:

- Ever smoked 100 or more cigarettes in life - Yes
- Currently smoking cigarettes (never smoked included in the denominator) – Every day or some days
- Used e-cigarettes or other vaping products in the last 30 days - Every day or some days

The RMSE is the square root of the sum of the squared bias and the squared standard error of a survey estimate. The bias is calculated as the difference between the weighted survey point estimate and the benchmark, which in this case, is the weighted survey estimate from the 2017 NHIS. Here, we disregard that the SAE model estimates are unbiased (under the model) and calculate the "bias" against benchmark as a way to compare the three estimation methods. To calculate standard errors (SE) for the RDS and propensity approaches, we used bootstrap variance estimation, in which multiple replicates of the sample are created by re-sampling (Wolter 2007). For the SAE approach, we used standard errors generated from the SAE models. For the 2017 NHIS, standard errors were calculated using the variance estimation strata and PSU provided in the public use data.

3. Results

Results in Tables 2 through 4 show that the propensity estimation approach has the lowest overall RMSE despite having the largest standard errors for all three survey estimates examined here. Table 2 presents results for "Ever Smoked 100+ Cigarettes", Table 3 presents results for "Currently Smoking Cigarettes", and Table 4 presents results for "Used E-cigarettes in the Last 30 days". The SAE approach has lower RMSE than the RDS approach for two survey estimates.

When we compare the RMSE for survey estimates by domain in Figures 1 through 3, the SAE approach has lower or similar RMSE as compared to the other two estimation approaches in most domains.

Table 2 Ever Smoked 100+ Cigarettes Overall Results

Approach	Estimate	SE	RMSE
RDS	51.5	4.23	14.53
Propensity	42.7	5.89	7.79

SAE	53.3	3.61	16.11
NHIS	37.6	7.14	

Table 3 Currently Smoking Cigarettes Overall Results

Approach	Estimate	SE	RMSE
RDS	29.2	3.49	9.12
Propensity	24.3	4.06	5.41
SAE	27.6	3.89	7.83
NHIS	20.8	5.03	

Table 4 Used E-cigarettes in the Last 30 Days

Approach	Estimate	SE	RMSE
RDS	20.7	2.79	8.15
Propensity	24.2	4.87	6.38
SAE	22.3	3.51	6.97
NHIS	28.4	5.17	

Figure 1 Ever Smoked 100+ Cigarettes by Domain RMSE

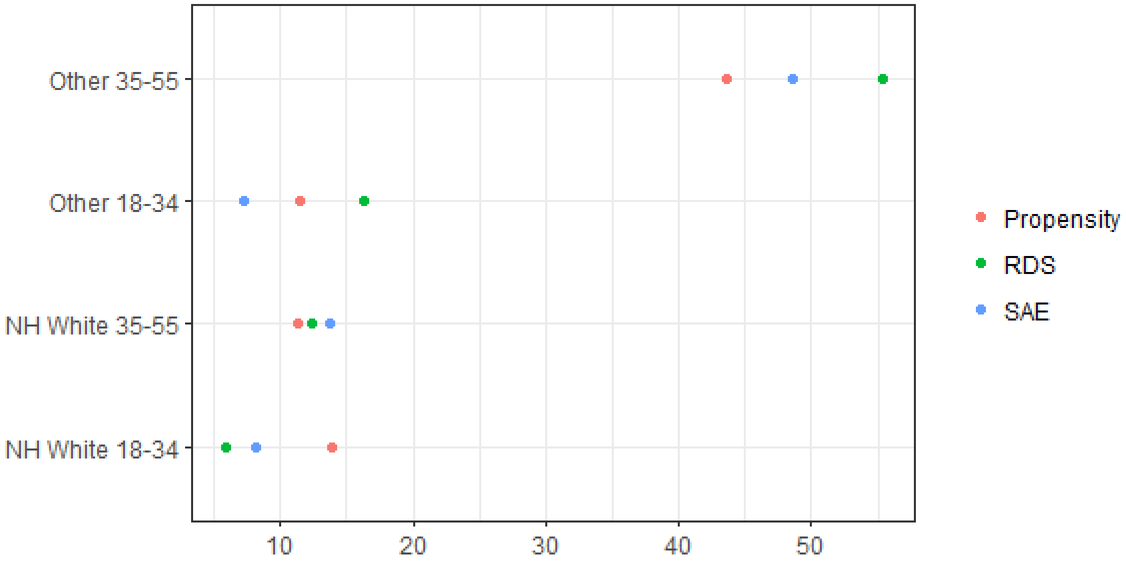


Figure 2 Currently Smoking Cigarettes by Domain RMSE

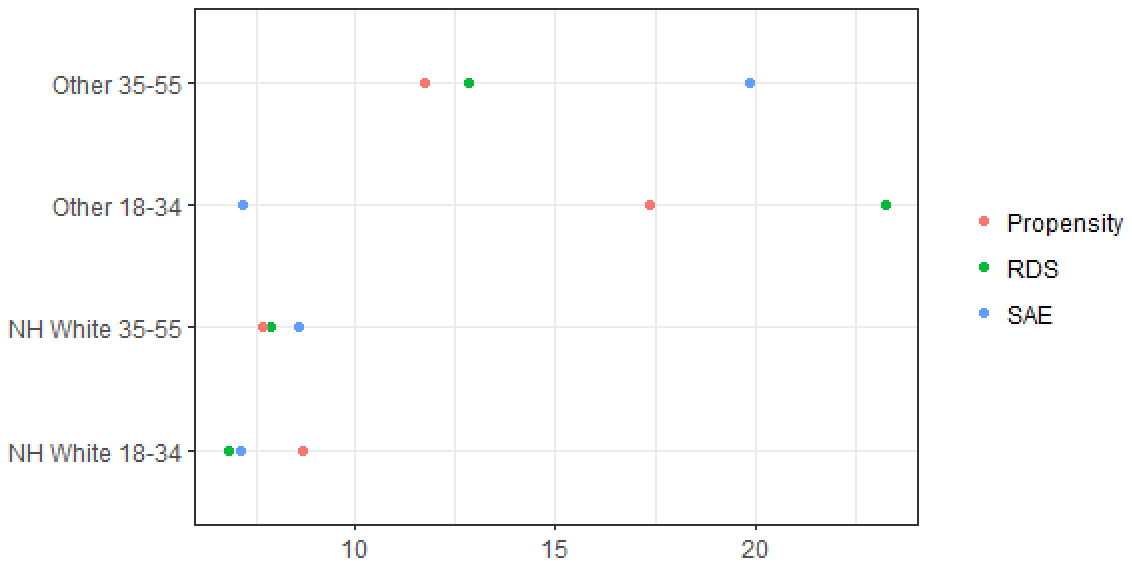
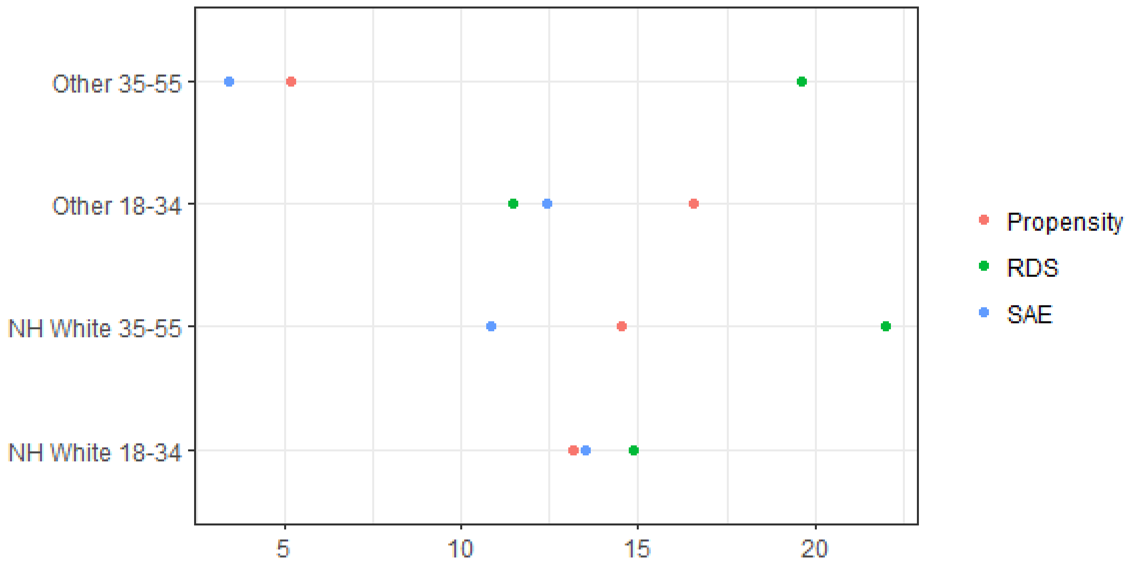


Figure 3 Used E-cigarettes in the Last 30 Days by Domain RMSE



4. Conclusion and Future Research

In this study, we compared three estimation methods for combining probability and non-probability samples in a Web-based RDS pilot study: an RDS estimation approach, a propensity approach, and SAE. When using the 2017 NHIS as a benchmark, the propensity approach has the lowest overall RMSE in all three smoking-behavior survey estimates. Its weighted point estimates are the closest to the benchmarks, and this helps to offset the largest SEs among the three estimation methods. When we examine estimates by domain, the SAE approach has the lowest average RMSE. This is expected, since a key purpose of SAE is to improve domain level estimates.

This research is subject to several limitations. First, the sample size is very small resulting in imprecise estimates and limiting the number of domains that could be assigned for the SAE approach. Second, the choice of benchmarks may have an impact on the findings. The NHIS public use data only included LGB status -- no transgender indicator was available. Thus, benchmarks estimates of LGBT are likely somewhat under-estimated. Also, NHIS is a general population survey with limited LGB sample size. One could argue that instead of using the NHIS control totals as benchmarks, we could have used the SAE modeled estimates as benchmarks, because they are theoretically unbiased.

Third, we only evaluated three survey estimates, since the pilot survey was short with a limited number of survey questions. Testing outcomes in other topic areas could lead to different results and conclusions. For future research, we would like to use data with larger sample sizes; evaluate survey variables not related to smoking behaviors; and compare additional estimation methods for combining probability and nonprobability samples, such as statistical matching.

References

Gile, K. J., and M. S. Handcock. "7. Respondent-Driven Sampling: An Assessment of Current Methodology." *Sociological methodology* 40, no. 1 (2010): 285-327.

Michaels, S. et al. “Test of a Hybrid Method of Sampling the LGBT Population: Web Respondent Driven Sampling with Seeds from a Probability Sample.” Forthcoming Journal of Official Statistics article.

National Academies of Sciences, Engineering, and Medicine. *Improving Health Research on Small Populations: Proceedings of a Workshop*. National Academies Press, 2018.

Rao, J.N.K. (2003). Small Area Estimation, John Wiley & Sons, Inc.

Wolter, K (2007). Introduction to Variance Estimation, Springer Science & Business Media.

Yang, M. et al. “Estimation Methods for Nonprobability Samples with a Companion Probability Sample.” JSM proceedings: 2018.