



A dual-frame approach for estimation with respondent-driven samples

Chien-Min Huang¹ · F. Jay Breidt²

Received: 30 September 2022 / Accepted: 11 March 2023
© Sapienza Università di Roma 2023

Abstract

Respondent-driven sampling (RDS) is an increasingly common method for surveying rare, hidden, or otherwise hard-to-reach populations. Instead of formal probability sampling from a well-defined frame, RDS relies on respondents themselves to recruit additional participants through their own social networks. By necessity, RDS is often initiated with a small, non-random sample. Standard RDS estimators have been developed under strong assumptions on the diffusion of sampling through the network over multiple waves of recruitment. We consider an alternative setting in which these assumptions are not met, and instead a large probability sample is used to initiate RDS and only a few waves of recruitment take place. In this setting, we develop dual-frame estimators that use both known inclusion probabilities from the initial sampling design and estimated inclusion probabilities from RDS, treated as a nonprobability sample. In a simulation study using network data from the Project 90 study, our dual-frame estimators perform well relative to standard RDS alternatives, across a wide range of recruitment behaviors. We propose a simple variance estimator that yields stable estimates and reasonable confidence interval coverage. Finally, we apply our dual-frame estimators to a real RDS study of smoking behavior among lesbian, gay, bisexual, and transgender (LGBT) adults.

Keywords Inverse propensity estimator · Network sampling · Nonprobability sample · Project 90

1 Introduction

1.1 Background on RDS

Sampling from rare, hidden, or otherwise hard-to-reach populations is challenging because screening costs can be very high and securing trust of potential respondents may be difficult. Respondent-driven sampling (RDS) is an increasingly common method for surveying such

✉ Chien-Min Huang
Chien-Min.Huang@colostate.edu

¹ Department of Statistics, Colorado State University, Fort Collins, CO, USA

² Department of Statistics and Data Science, NORC at the University of Chicago, Chicago, IL, USA

populations [13]. Instead of formal probability sampling from a well-defined frame, RDS relies on respondents themselves to recruit additional participants through their own social networks, alleviating both the screening and trust issues. In practice, respondents are given a limited number of “coupons” that they can use to recruit acquaintances.

By necessity, RDS is often initiated with a small, non-random sample. Inference for data from RDS then relies on modeling assumptions about the structure of the network and the process by which respondents recruit within their networks. There is extensive literature on inference for RDS; see [14] for an overview. We consider three standard RDS estimators in this paper. Each relies on the network degree, d_k , which is the number of acquaintances in the network of individual k . The Salganik–Heckathorn estimator (SH, also known as RDS-I) is specifically designed to estimate binary proportions under RDS [26]. For a population consisting of two groups A and B , the SH estimate for the population proportion in group A is

$$\hat{\mu}_A^{\text{SH}} = \frac{\hat{d}_B \hat{C}_{BA}}{\hat{d}_A \hat{C}_{AB} + \hat{d}_B \hat{C}_{BA}},$$

where \hat{d}_A, \hat{d}_B are the estimated average degrees in group A and group B , and $\hat{C}_{AB}, \hat{C}_{BA}$ are the estimated probabilities of cross-group recruitment. [25] provides a variance estimation approach using bootstrap that accounts for the dependence within the sample.

The more general estimator in Volz and Heckathorn [28] (VH, also known as RDS-II) has the form of a Hansen–Hurwitz estimator [12], with probabilities proportional to d_k . The estimate for the mean is

$$\hat{\mu}_y^{\text{VH}} = \frac{\sum_{k \in s} d_k^{-1} y_k}{\sum_{k \in s} d_k^{-1}},$$

where d_k is the degree of element k . Volz and Heckathorn derived a variance estimator from Hansen–Hurwitz theory that attempts to account for the dependencies in the RDS sample. VH makes strong assumptions that the network is fully connected and symmetric and that the recruitment process evolves as a random walk on the network, with seeds chosen with probability proportional to degree, recruiters given one coupon, recruiters choosing a single acquaintance at random, and sampling conducted with replacement.

Under stronger assumptions about the structure of the network, Gile [7] was able to weaken assumptions about the recruitment process, including the single coupon and the with-replacement assumptions, resulting in the successive sampling (SS) estimator. The estimator is derived iteratively, alternating between estimation of the population degree distribution using the current estimates of inclusion probabilities and estimation of the inclusion probabilities using the current estimated degree distribution. At convergence, the resulting estimator for the population mean is

$$\hat{\mu}_y^{\text{SS}} = \frac{\sum_{k \in s} y_k / \hat{\pi}(d_k)}{\sum_{k \in s} 1 / \hat{\pi}(d_k)},$$

where $\hat{\pi}(d_k)$ is the estimated inclusion probability with degree d_k . For variance estimation, Gile [7] introduced a bootstrap procedure.

In practice, only a few waves of recruitment may take place in RDS [20, 22], due to network limitations, recruiting failures, or time constraints for the study. Even if many seeds are used in order to generate some longer recruitment chains, the resulting sample is unlikely to meet the assumptions needed for proper inference with standard RDS estimators.

In some cases, it is feasible to select a large number of seeds as a probability sample representative of the population of interest. In this case, RDS is used to augment sample size and it is less critical to generate long chains that traverse the population network. Michaels

et al. [21] described an attempt to apply RDS to augment the number of respondents in a study of smoking behavior among lesbian, gay, bisexual, and transgender (LGBT) adults. To our knowledge, this is the first publication using a probability sample from a representative national panel as the starting seeds. In this paper, we develop novel methods for combining a probability sample of seeds and a nonprobability respondent-driven sample of recruits. We study the methods via simulation with sample sizes similar to [21] and apply the methods to data from the LGBT smoking behavior study [21].

1.2 Background on probability and nonprobability methods

We consider the setting in which RDS seeds are selected as a probability sample with known inclusion probabilities, and treat those being recruited via RDS as a nonprobability sample since we do not know the probability of the recruitment process. There is a growing literature on the combination of probability and nonprobability samples as survey costs increase and response rates decrease. In our context, the study variables of interest and the auxiliary variables are observed in both samples. Much of the literature in combining probability and nonprobability samples assumes study variables of interest are observed in the nonprobability sample only, while auxiliary variables are observed in both samples. But modifying such a method to include study variables also observed in the probability sample is usually straightforward, so we include such methods in our brief review.

An approach that uses bivariate Fay-Herriot models from small area estimation to combine domain-level point estimates from probability and nonprobability samples is described in [6]. [18] constructs a post-stratified estimator with two post-strata: the nonprobability sample and its complement, from which the probability sample is selected (probability sample elements that overlap with the nonprobability sample are excluded). The Bayesian approach for combining probability and nonprobability samples [24, 29] uses the nonprobability sample to provide a prior for estimates from the probability sample. Sample matching and mass imputation approaches [2, 17, 23, 30] implicitly or explicitly construct a model by regressing study variables on covariates in the nonprobability sample, then use the fitted model and observed covariates to predict the study variables on the probability sample. Since the probability sample is representative of the population, the predictions can be appropriately weighted (it is thus sometimes useful to think of the weights from the probability sample as being imputed to the nonprobability sample). The inverse weighting approach or quasi-randomization approach [2, 4, 19] estimates the propensities for the nonprobability sample by combining the probability and nonprobability samples based on the missing at random assumption. Doubly-robust estimators combine an estimated propensity and a regression model for the study variable [2, 19, 27] and are consistent if either model is correctly specified.

[16] (Chapter 3) proposed a class of dual-frame estimators for combining probability samples with nonprobability samples selected via expert judgment in the context of an application to a fisheries survey. In simulation studies across a range of population and judgment characteristics, the dual-frame approach yields stable estimates and reasonable confidence interval coverage, and the strategy that combines probability and judgment sampling dominates the classic strategy of pure probability sampling with known design weights. Because of the good performance of the dual-frame estimation technique in the judgment sampling application, particularly its empirical robustness, we adapt the method for inference in RDS with probability samples of seeds.

1.3 Overview of paper

We introduce notation and pseudolikelihood estimation of inclusion probabilities for the nonprobability sample in Sect. 2.1. The dual-frame estimator is introduced in Sect. 2.2 and variance estimation is discussed in Sect. 2.3. Our approach treats recruits as a Poisson sample from the complement of the probability sample of seeds, and does not require assumptions about the network. In Sect. 3, we describe simulation results using network data from the Project 90 study, comparing our dual-frame estimators to standard RDS alternatives (SH, VH, SS) across a wide range of recruitment behaviors. In Sect. 4, we apply our dual-frame estimators to the study in [21] of smoking behavior among LGBT adults. A brief discussion follows in Sect. 5.

2 Methods

2.1 Estimation of inclusion probability

Consider a finite population $U = \{1, 2, \dots, N\}$ and let $s_A \subset U$ denote the probability sample of seeds, with known inclusion probabilities. Each seed is given a limited number of coupons to recruit acquaintances who are then interviewed. Each new respondent is in turn given coupons to recruit acquaintances. The individuals being recruited in the survey, denoted as s_B , are treated as a nonprobability sample since we do not know their inclusion probabilities. Study protocols require that no individual can be included twice (sampled and recruited, or recruited through more than one network). Hence, $s_A \cap s_B = \emptyset$. Denote the combined sample as $s = s_A \cup s_B$.

Instead of making assumptions about the population network and the recruiting process as it moves through the network, we will treat s_B as a Poisson sample with unknown inclusion probabilities to be estimated. Let n_A denote the number of seeds, n_B the number of individuals being recruited, and $n = n_A + n_B$ the combined sample size. The probability sample indicators are $I_k^A = 1$ if $k \in s_A$, $I_k^A = 0$ otherwise; similarly, the nonprobability sample indicators are $I_k^B = 1$ if $k \in s_B$, $I_k^B = 0$ otherwise. The first-order inclusion probability for s_A is $\pi_k^A = E[I_k^A] = \Pr[I_k^A = 1]$ satisfying $\pi_k^A > 0$ for all $k \in U$ and known for all $k \in s_A$. The first-order inclusion probability for the nonprobability sample is

$$\begin{aligned} \pi_k^B &= \Pr[k \in s_B \mid k \in s_A] \Pr[k \in s_A] + \Pr[k \in s_B \mid k \notin s_A] \Pr[k \notin s_A] \\ &= 0 + \rho_k(1 - \pi_k^A). \end{aligned} \quad (1)$$

Because of the recruitment, the ρ_k and π_k^B are unknown and not necessarily positive for all $k \in U$. The first-order inclusion probability for the combined sample $s = s_A \cup s_B$ is then

$$\pi_k = \pi_k^A + (1 - \pi_k^A)\rho_k, \quad (2)$$

which is strictly positive for all $k \in U$, because s_A is a probability sample.

We specify a parametric model, $\rho_k = \rho(\mathbf{x}_k, \boldsymbol{\theta})$ in (1), where $\boldsymbol{\theta}$ are the true unknown parameters and \mathbf{x}_k is a vector of known auxiliary variables, available in both the s_A and s_B sample. We estimate the parameters via a likelihood-based method. We assume Poisson sampling for s_B , under which the log-likelihood function is

$$\begin{aligned} \ln L(\theta) &= \sum_{k \in U \setminus s_A} I_k^B \ln\left(\frac{\rho_k}{1 - \rho_k}\right) + \sum_{k \in U \setminus s_A} \ln(1 - \rho_k) \\ &= \sum_{k \in U \setminus s_A} I_k^B \ln\left(\frac{\rho(\mathbf{x}_k, \theta)}{1 - \rho(\mathbf{x}_k, \theta)}\right) + \sum_{k \in U \setminus s_A} \ln(1 - \rho(\mathbf{x}_k, \theta)). \end{aligned}$$

However, the second term of the log-likelihood involves data not in s_A or s_B . Following [2], we replace the second term by the unbiased Horvitz-Thompson [15] estimator (from the s_A sample) of its expectation, and compute the estimate $\hat{\theta}$ by maximizing the pseudo log-likelihood

$$\sum_{k \in U \setminus s_A} I_k^B \ln\left(\frac{\rho(\mathbf{x}_k, \theta)}{1 - \rho(\mathbf{x}_k, \theta)}\right) + \sum_{k \in U} \ln(1 - \rho(\mathbf{x}_k, \theta))(1 - \pi_k^A) \frac{I_k^A}{\pi_k^A}$$

We further assume a logistic model, $\text{logit}(\rho(\mathbf{x}_k, \theta)) = \mathbf{x}_k^\top \theta$, for which the pseudo log-likelihood is

$$\ell(\theta) = \sum_{k \in U \setminus s_A} I_k^B \mathbf{x}_k^\top \theta - \sum_{k \in U} \ln\left\{1 + \exp(\mathbf{x}_k^\top \theta)\right\} (1 - \pi_k^A) \frac{I_k^A}{\pi_k^A}. \tag{3}$$

We plug in the estimated parameters $\hat{\theta}$ to obtain initial estimates, $\tilde{\rho}_k = \rho(\mathbf{x}_k, \hat{\theta})$. Ideally, the initial estimates $\tilde{\rho}_k$ would then be calibrated to the nonprobability sample size

$$n_B = \sum_{k \in U} \tilde{\rho}_k (1 - \pi_k^A), \tag{4}$$

but this is not feasible because we do not observe \mathbf{x}_k for $U \setminus (s_A \cup s_B)$. We therefore estimate the right hand side of (4) from the probability sample

$$\sum_{k \in s_A} \frac{\tilde{\rho}_k (1 - \pi_k^A)}{\pi_k^A}, \tag{5}$$

and find the constant α that minimizes

$$\left| n_B - \sum_{k \in s_A} \frac{\alpha \tilde{\rho}_k (1 - \pi_k^A)}{\pi_k^A} \right|$$

subject to the constraints $\alpha \tilde{\rho}_k \in [0, 1]$ for all $k \in s_A \cup s_B$. The final estimates are then obtained as $\hat{\rho}_k = \alpha \tilde{\rho}_k$.

2.2 Dual-frame estimator

If π_k from (2) were known for all $k \in s$, we could compute the unbiased Horvitz-Thompson [15] estimator $\sum_{k \in s} y_k \pi_k^{-1}$ of the total and the asymptotically unbiased Hájek [10] estimator $\sum_{k \in s} y_k \pi_k^{-1} \left(\sum_{k \in s} \pi_k^{-1}\right)^{-1}$ of the mean, which are dual-frame estimators based on the combined sample. The Hájek estimator is standard because it does not require the population size to be known (population size is often unknown in hard-to-reach populations) and it is typically more efficient even if population size is known. Since ρ_k is unknown, we plug in $\hat{\rho}_k$. Assuming π_k^A is known for all the units in the sample, we then construct the ‘‘Combined’’ estimators for the total and the mean of the variable of interest:

$$\widehat{T}_{y,\text{com}} = \sum_{k \in s} \frac{y_k}{\pi_k^A + (1 - \pi_k^A)\widehat{\rho}_k},$$

$$\widehat{\mu}_{y,\text{com}} = \frac{\sum_{k \in s} y_k / \{\pi_k^A + (1 - \pi_k^A)\widehat{\rho}_k\}}{\sum_{k \in s} 1 / \{\pi_k^A + (1 - \pi_k^A)\widehat{\rho}_k\}}. \quad (6)$$

If π_k^A is unknown for $k \in s_B$, it can be approximated, as in the example of Sect. 4.

Under the assumed Poisson sampling model for recruiting, these estimators are design mean-square consistent and asymptotically normal, as shown in ([16] Chapter 3). While we do not expect the Poisson assumption to hold in the RDS setting, we conjecture that it is a useful approximation under a variety of RDS selection mechanisms. We assess this conjecture via simulation in Sect. 3.

The Combined estimator might be less efficient if the nonprobability sample size (number of recruits) is large relative to the probability sample size (number of seeds). This setting may be more favorable to standard RDS estimation. We therefore considered a convex combination of our Combined dual-frame estimator and a classic RDS estimator. We chose the VH estimator for the convex combination because it is probably the most commonly used estimator in current RDS practice ([28]), and [8] states that the VH estimator performs better than the SH estimator ([26]). The resulting ‘‘Convex’’ estimators of the total and mean are then

$$\begin{aligned} \widehat{T}_{y,\text{convex}} &= \frac{n_A}{n_A + n_B} \sum_{k \in s} \frac{y_k}{\pi_k^A + (1 - \pi_k^A)\widehat{\rho}_k} + \frac{n_B}{n_A + n_B} \sum_{k \in s} \frac{N d_k^{-1} y_k}{\sum_{k \in s} d_k^{-1}} \\ &= \sum_{k \in s} \left[\frac{n_A}{n_A + n_B} \frac{1}{\pi_k^A + (1 - \pi_k^A)\widehat{\rho}_k} + \frac{n_B}{n_A + n_B} \frac{N d_k^{-1}}{\sum_{k \in s} d_k^{-1}} \right] y_k \\ &= \sum_{k \in s} \widehat{w}_k y_k \\ \widehat{\mu}_{y,\text{convex}} &= \frac{\sum_{k \in s} \widehat{w}_k y_k}{\sum_{k \in s} \widehat{w}_k}, \end{aligned} \quad (7)$$

where d_k is the network degree of individual k .

2.3 Variance estimation

Under the combined design, with general probability sampling for s_A and the assumed Poisson sampling for s_B , the variance of (6) can be approximated by Taylor expansion. The variance approximation is a function of first-order inclusion probabilities π_k^A , second-order inclusion probabilities $\pi_{k\ell}^A = E[I_k^A I_\ell^A]$, and the unknown ρ_k . If the sampling design for s_A is measurable (with $\pi_{k\ell}^A > 0$ for all $k, \ell \in U$), then it is easy to show that the combined design is also measurable, allowing unbiased variance estimation for totals and approximately unbiased variance estimation for (6) if the ρ_k were known. In practice, we plug in estimates $\widehat{\rho}_k$ for the unknown ρ_k , treating $\pi_k^A + (1 - \pi_k^A)\widehat{\rho}_k$ as the combined inclusion probability; and we use the standard with-replacement variance estimation approximation, which does not require the second-order inclusion probabilities. The with-replacement approximation is available in standard survey software.

Similarly, the variance and variance estimator of (7) can be approximated by Taylor expansion and

$$n_A \left\{ (n_A + n_B) \left(\pi_k^A + (1 - \pi_k^A) \widehat{\rho}_k \right) \right\}^{-1} + n_B N d_k^{-1} \left\{ (n_A + n_B) \left(\sum_{k \in S} d_k^{-1} \right) \right\}^{-1}$$

can be treated as the corresponding combined inclusion probability in standard survey software.

Methods to account for the variability due to parameter estimation in $\widehat{\rho}_k$ include replication approaches [27] and the estimating equations approach [2, 19]. In combining probability and nonprobability samples in settings similar to the current RDS application, we have considered replication methods (delete-a-group jackknife and balanced repeated replication) and the estimating equations approach in simulation experiments. While the estimating equations approach helped reduce the bias in the variance estimation and improved the coverage of nominal 95% confidence intervals, it did so at the expense of increased root mean squared error (RMSE) for the variance estimation and wider confidence intervals. The simple, with-replacement approximation variance estimation performed better in terms of RMSE than the replication and estimating equations approaches. Further work to improve variance estimation in the RDS context of this paper is ongoing.

3 Simulation

3.1 Simulation study design

We evaluate the proposed Combined estimator (6) and Convex estimator (7) and compare to three standard RDS estimators (SH, VH, and SS) using an artificial population constructed from the Project 90 study data. These data were collected between 1988 and 1992 in Colorado Springs, CO to study heterosexuals’ transmission of HIV, and have become a classic example of network data on a hidden population. Several published studies [1, 5, 9] have used Project 90 data to compare RDS estimators. As in the prior studies, we constructed an artificial population consisting of the network subset with the largest connected component, which includes 4430 individuals and 18,407 edges. The data include 13 binary attributes for each individual, such as sex worker, pimp, and drug dealer, with value 1 indicating that the individual has the attribute. Table 1 summarizes the population proportions for the 13 binary attributes.

Table 1 Project 90 population proportions for 13 binary attributes

Attribute	Proportion	Attribute	Proportion
Female	0.43	Retired	0.03
Sex worker	0.06	Housewife	0.06
Pimp	0.02	Disabled	0.04
Client	0.10	Unemployed	0.17
Drug dealer	0.08	Homeless	0.14
Drug cook	0.01	Nonwhite	0.26
Thief	0.03		

We consider two initial sample sizes (150 or 300) for the without-replacement probability sample of seeds. The initial probability sample is subject to nonresponse, completely at random, at one of three response rates (0.2, 0.3, or 0.4). From the responding seeds, RDS recruiting is then employed, with a target final sample size of 150 (seeds plus recruits) in every case. Seeds are each given three coupons. In the simulation, the recruiting stops when the target sample size is achieved. Recruits always agree to participate and respond. Recruits are selected without replacement, that is, the individuals already in the seeds or already being recruited cannot be recruited again, as would be the case in practice. At initial seed sample size of 300 and response rate of 0.4, the expected number of responding seeds is 120, so only 30 RDS recruits are needed, while at initial seed sample size of 150 and response rate of 0.2, the expected number of responding seeds is 30, and 120 RDS recruits are needed.

Seeds are selected via simple random sampling. We have also conducted simulations in which seeds are selected with probability proportional to degree (rarely feasible in practice). Results are qualitatively similar and are not shown here.

We consider eight different recruitment behaviors: (1) *random*, in which three acquaintances are recruited at random with equal probabilities, if possible; (2) *recruit fraction*, in which 0, 1, 2, or 3 acquaintances are recruited at random, with probabilities (1/6, 1/6, 1/6, 1/2); (3) *degree*, in which recruitment probabilities are proportional to the degrees of acquaintances; (4) *inverse degree*, in which recruitment probabilities are proportional to the inverse degrees of acquaintances; (5) *prefer female*, in which females must recruit female acquaintances, if possible, and males recruit males; (6) *prefer pimp*, in which pimps must recruit pimp acquaintances, if possible, and non-pimps recruit non-pimps; (7) *expert female*, in which everyone must recruit female acquaintances, if possible; and (8) *expert pimp*, in which everyone must recruit pimp acquaintances, if possible. In the existing literature, recruitment is assumed to be at random, but our approach allows for differential recruitment.

For all the recruitment behaviors, we estimate the inclusion probabilities using the model

$$\text{logit}(\rho_k) = \theta_0 + \theta_1 d_k.$$

The model is misspecified for all the simulated recruitment behaviors, though it is somewhat similar to (3) *degree*.

For a baseline comparison with RDS, we could consider using only the responding probability seeds for inference, ignoring the RDS recruits. Instead, we use an expanded probability sample with expected number of required contacts equal to that of the RDS sample. For example, if the probability sample response rate is 0.2, with 300 initial seeds and overall target of 150 respondents (seeds plus recruits), then we expect to contact 390 people, with $(0.2)(300) + 90 = 150$ respondents via RDS but only $(0.2)(300 + 90) = 78$ respondents via probability sampling. We use the expanded probability sample as the baseline for a “fair” comparison to RDS, though in practice it is often not possible to simply take a larger probability sample.

For each initial seed sample size and seed response rate, we draw 1000 probability samples of seeds. These probability samples are used to generate RDS recruits under each of the eight recruitment behaviors. For each combination of seed sample and recruit sample, we estimate the inclusion probabilities assuming Poisson sampling for the recruit sample and construct the Combined estimator (6) and Convex estimator (7), along with estimated variances and nominal 95% confidence intervals, for each of the 13 binary attributes. We also construct three classic estimators (SH, VH, SS) along with estimated variances and nominal 95% confidence intervals using the R package RDS [11]. Our baseline for comparison is the expanded simple random sample with expected number of contacts equal to RDS.

3.2 Simulation study results

We computed root mean squared error (RMSE) for all six estimators across all attributes, recruitment behaviors, and simulation settings. To summarize the simulation results, we ranked the six estimators from 1 (lowest RMSE) to 6 (highest RMSE) within each attribute and recruitment behavior and averaged the ranks within each probability sample response rate and initial seed sample size. Average ranks are shown in Table 2. The average rank of SH is nearly 6 in most cases, indicating that it is almost always the worst estimator. The VH and SS estimators have average ranks of approximately 5 and 4, respectively, so that they are nearly always worse than our proposed Combined and Convex estimators, as well as Baseline. All of the classic RDS estimators have lower average rank (improved performance) for 150 seeds than for 300 seeds, reflecting increased waves of RDS recruiting and better alignment with classic RDS assumptions. Our proposed Convex and Combined estimators generally have the lowest average ranks, with greatest improvements over Baseline at the lowest response rate. At the highest response rate, Baseline is competitive or better with our proposed estimators.

We now focus in detail on the case of 300 initial seeds, response rate 0.3, and final sample size 150, so that our simulation is comparable to the 264 initial seeds, response rate 0.34, and final sample size 140 for the LGBT sample with recruitment condition in Table 2 of [21]. We apply our methods to the application of [21] in Sect. 4. Results for the remaining simulation settings in Table 2 are similar to those presented here and are provided in detail in the supplemental material.

In our simulations, SH nearly always has the highest RMSE by a large margin. We remove it from further consideration to avoid distorting the plots. We summarize the competitive estimators SS, VH, Convex, and Combined, and compare to the Baseline estimator.

Figure 1 shows the bias for point estimates of the 13 attributes with eight different recruitment behaviors. Because some estimation targets are small proportions, we chose not to report relative biases, some of which would be very large. Though our proposed estimators exhibit some bias, the bias does not lead to significant undercoverage of confidence intervals and RMSEs tend to be better than Baseline.

Figure 2 shows boxplots of the RMSE ratios for the 13 attributes, with one set of boxplots for each of the eight recruitment behaviors. In the ratios, RMSE for the Baseline estimator is the numerator and RMSE for SS, VH, Convex, or Combined is the denominator. Higher values are better, with RMSE ratios equal to one (solid reference line) indicating that no efficiency

Table 2 Average rank of RMSE among baseline, SH, VH, SS, combined, and convex estimators (lower is better, average rank of 6 is worst possible) across all attributes and recruitment behaviors for each probability sample response rate and initial seed sample size

Response Rate	Initial Seeds	Estimator					
		Baseline	SH	VH	SS	Combined	Convex
0.2	300	2.94	5.89	4.71	3.72	1.74	1.99
	150	3.65	5.62	4.27	3.32	2.24	1.89
0.3	300	2.27	5.93	4.81	3.80	2.13	2.06
	150	2.63	5.86	4.67	3.69	1.92	2.22
0.4	300	2.09	6.00	4.82	3.82	2.46	1.82
	150	2.18	5.88	4.79	3.83	1.91	2.41

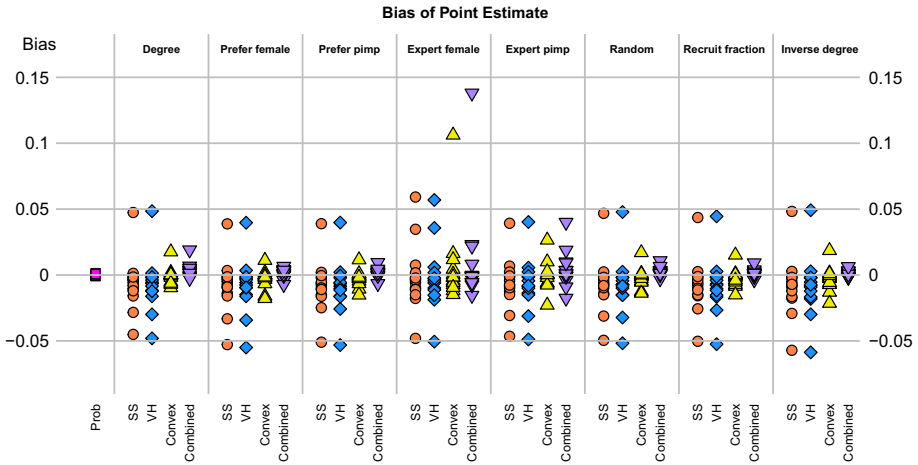


Fig. 1 Biases of point estimates of 13 attributes for each recruitment behavior, under simple random sampling of 300 initial seeds and seed response rate of 0.3. Each point corresponds to bias for one estimator type and one binary attribute. Results are based on 1000 simulated samples

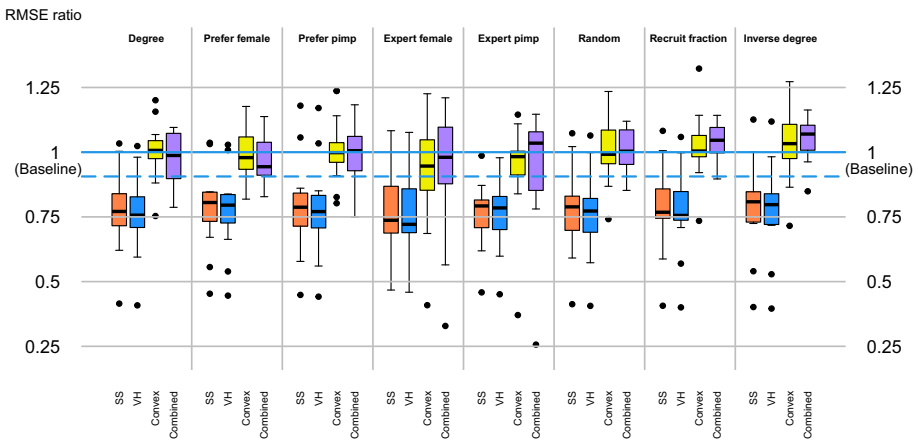


Fig. 2 Boxplots of RMSE ratios across all 13 attributes for each recruitment behavior and estimator type, under simple random sampling of 300 initial seeds and seed response rate of 0.3. In each ratio, RMSE of the Baseline estimator is the numerator and RMSE of SS, VH, Convex, or Combined is the denominator. Higher values are better. The solid reference line at one corresponds to RMSE equal to that of the Baseline estimator. The dashed reference line below one corresponds to the average RMSE (across binary characteristics) attained by ignoring the RDS recruits and using only the initial probability respondents. Results are based on 1000 simulated samples

was gained by RDS relative to probability sampling with the same expected number of contacts. An additional dashed reference line corresponds to the average RMSE ratio across binary characteristics for the probability sample of initial seeds only, ignoring the RDS recruits. Our proposed estimators almost always gain efficiency relative to ignoring RDS recruits, even with a misspecified probability mechanism and propensity model that needs to be estimated. Our proposed estimators dominate the classic RDS estimators and also tend to be better than the (often infeasible) Baseline estimator with the same expected number of contacts.

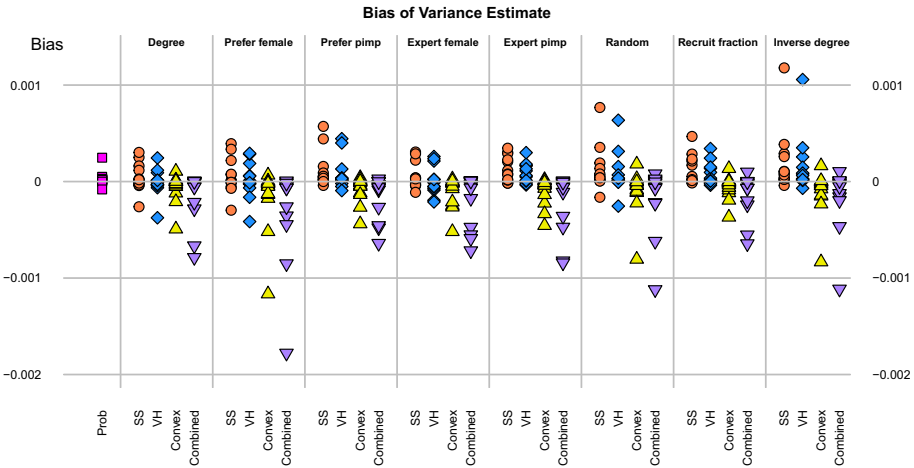


Fig. 3 Biases of variance estimates of 13 attributes for each recruitment behavior, under simple random sampling of 300 initial seeds and seed response rate of 0.3. Each point corresponds to bias for one estimator type and one binary attribute. Results are based on 1000 simulated samples

Figure 3 summarizes the bias of variance estimates across attributes and recruitment behaviors. The Baseline variance estimator is unbiased. The biases in nearly all cases are small and are comparable in magnitude between classic RDS bootstrap procedures and our simple approach described in Sect. 2.3. While we have some preliminary simulation results (not shown) for variance estimation that incorporates parameter estimation variability via estimating equations, we found that the slight reduction in bias was accompanied by an increase in variance and no improvement in confidence interval coverage.

Figure 4 summarizes results for estimation of variance as boxplots of relative RMSEs across the 13 attributes and eight recruitment behaviors. Each relative RMSE has the RMSE of the estimated variance in the numerator and the true variance (as approximated by Monte Carlo) in the denominator. Lower values are better. The relative RMSE of our proposed variance estimators is nearly always better than the variance estimators for SS and VH. Further, performance of our variance estimators is fairly constant across binary attributes and recruitment strategies, while SS and VH have considerably more variation.

Figure 5 summarizes the coverage of nominal 95% confidence intervals across the 13 attributes and eight recruitment behaviors. The Combined and Convex estimators generally have coverage that is closer to nominal and less variable than the coverage of SS and VH. Because many of the binary attributes are rare, confidence interval coverage even with the pure probability sample is below the nominal level.

4 Application

We illustrate our dual-frame methodology with an application of RDS in the sampling of US LGBT adults aged 18–55, as described in detail in [21]. This study, conducted by NORC at the University of Chicago in 2017, began with a probability sample selected from NORC’s AmeriSpeak® Panel. AmeriSpeak panelists are selected using rigorous design-based methods, beginning with a stratified multi-stage address-based sample of households from an enhanced version of the US Postal Service’s Computerized Delivery Sequence (CDS) file,

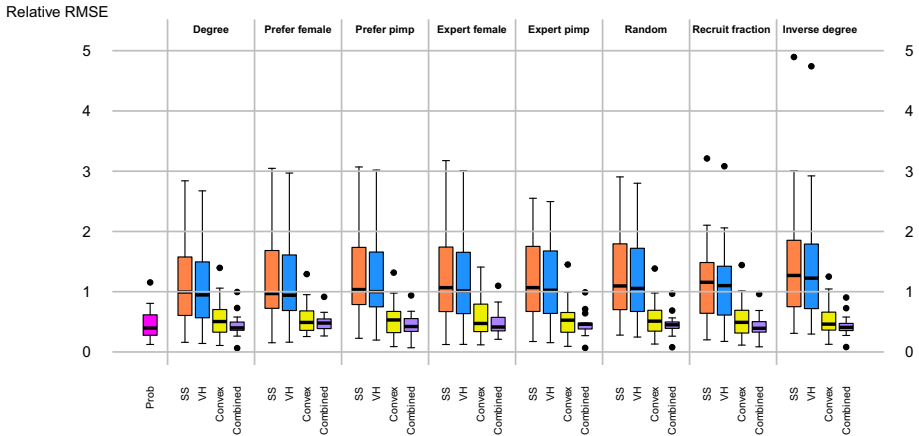


Fig. 4 Boxplots of relative RMSE of estimated variance across all 13 attributes for each recruitment behavior and estimator type, under simple random sampling of 300 initial seeds and seed response rate of 0.3. Each relative RMSE has RMSE of the variance estimator in the numerator and true variance (as approximated by Monte Carlo) is the denominator. Lower values are better. Results are based on 1000 simulated samples

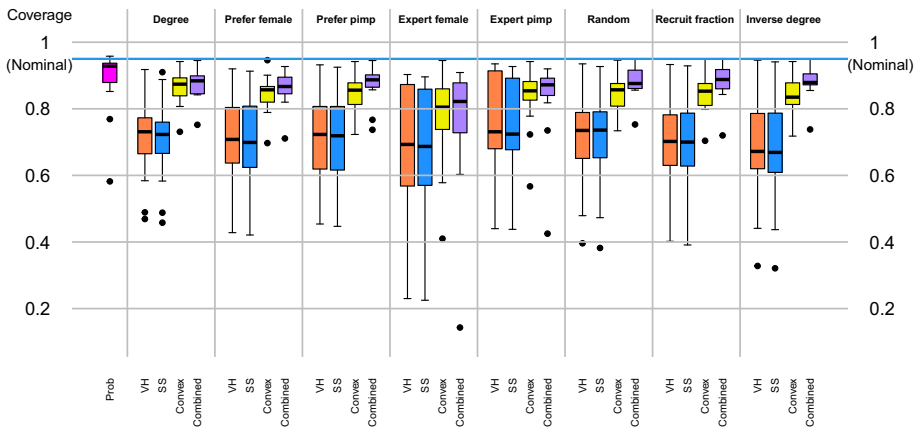


Fig. 5 Boxplots of confidence interval coverage across all 13 attributes for each recruitment behavior and estimator type, under simple random sampling of 300 initial seeds and seed response rate of 0.3. Horizontal reference line is at the nominal coverage level of 95%. Results are based on 1000 simulated samples

with additional listing of over 80,000 mostly rural households not available in the CDS. Both LGBT and non-LGBT panelists were selected in a stratified sample from the panel and assigned to one of two experimental conditions: either to **recruit** their LGBT friends and family into the study directly, or to **nominate** LGBT friends and family to be contacted by NORC. Referrals who completed the survey were then asked to recruit or nominate (depending on the assigned experimental condition of the seeds) their LGBT friends and family, with the study ending after four such rounds of referrals. In each round, respondents were allowed up to four referrals.

Base weights for all seeds (probability sample panelists) were computed as the product of the AmeriSpeak panel weight and the inverse of the selection probability from the panel to the seed sample. These base weights were adjusted for differential nonresponse using

standard weighting class adjustments and raked to population controls including age group, gender, and race/Hispanic ethnicity. Denote by s_A the set of all seed respondents (from either experimental condition) and let $\{w_k^A\}_{k \in s_A}$ denote the corresponding adjusted weights. We will treat s_A as the probability sample and $\{w_k^A\}$ as the design weights.

We are interested in domain D , the set of all US LGBT adults aged 18–55. Among all $n_A = |s_A| = 409$ random seeds who responded, there are $|s_A \cap D| = 182$ LGBT persons. Using only the probability sample and the design weights, we can make approximately unbiased estimates for the LGBT domain and construct valid confidence intervals via standard design-based techniques. These estimates will serve as the baseline for comparison to our dual-frame estimates, which incorporate the referrals as well as the seeds.

Let s_B denote the set of referral respondents, all of whom are LGBT. In this application, $n_B = |s_B| = 107$ referrals completed the survey.

We model the RDS referrals as a Poisson sample from $D \setminus s_A$ and modify (3) by restricting to D ,

$$\ell(\theta) = \sum_{k \in D \setminus s_A} I_k^B \mathbf{x}_k^\top \theta - \sum_{k \in D} \ln \left\{ 1 + \exp \left(\mathbf{x}_k^\top \theta \right) \right\} \left(w_k^A - 1 \right) I_k^A,$$

with $\mathbf{x}_k^\top = (1, d_k)$. In our application, 10 of the 289 LGBT respondents (5 seeds and 5 referrals) were missing degree. We used hot-deck imputation for these missing degree values within cells defined by LGBT status, race/ethnicity, and gender (missing degree among non-LGBT seeds does not affect our estimation for LGBT domain characteristics).

The pseudolikelihood parameter estimates are

$$(\hat{\theta}_0, \hat{\theta}_1) = (-10.87915, 0.0024068).$$

Using the initial pseudolikelihood estimates $\tilde{\rho}_k$, the estimated expected sample size in (5) is 103.68, a slight (3.1%) underestimate of the target $n_B = 107$. We then make the small calibration adjustment by choice of α , as described in Sect. 2.1, to obtain the final $\hat{\rho}_k$.

In this application, we do not have sufficient information about the design to compute the weights $\{w_k^A\}_{k \in s_B}$ for the RDS referrals s_B . These weights are needed to compute our combined dual-frame estimator. We statistically matched [2, 23, 30] each RDS referral to a random seed and assigned the weight and stratum of the seed to the referral. We used the StatMatch package [3] in R to conduct matching within classes defined by LGBT status and race/ethnicity, with Gower’s distance function used to choose “nearby” donors with respect to age, gender, and degree.

For each element in the combined LGBT sample, $k \in (s_A \cap D) \cup s_B$, we then computed the combined weight as

$$\hat{w}_k = w_k^A / \{1 + (w_k^A - 1)\hat{\rho}_k\}$$

where w_k^A is either from the original design or from matching.

We used both the original design weights and the combined dual-frame weights to compute estimates for 11 different survey items, seven of which are binary variables having to do with smoking behavior:

- Have you smoked at least 100 cigarettes in your life? (smoke100)
- In the past 30 days, have you ...
 - ...used e-cigarettes or other vaping products...? (ecig)
 - ...smoked regular cigars? (cigar)
 - ...smoked regular cigarillos? (cigarillo)
 - ...smoked little filtered cigars? (filtcigar)

- ...smoked marijuana or hashish? (*marijuana*)
- ...smoked a blunt? (*blunt*)

The remaining four items are binary questions about experience of discrimination:

- Have you ever experienced discrimination, been prevented from doing something, been hassled or made to feel inferior because of ...
 - ...your sex (that is, because you are male or female) (*discr_sex*)
 - ...your race, ethnicity or skin color (*discr_race*)
 - ...your sexual orientation (that is, because you are—or you are perceived to be by others—gay, lesbian, bisexual or straight) (*discr_orie*)
 - ...your gender identity (that is, because you are—or you are perceived to be by others—a man or a woman) (*discr_iden*)

Point estimates and approximate 95% confidence intervals (CIs) for each of the 11 items using only the probability sample and using the combined probability seeds with RDS referrals are shown in Fig. 6. In each case, the variance estimator used to construct the CI is the stratified with-replacement variance estimator available in standard survey software. For the dual-frame estimator, we use the same strata as in the original probability sample. This approach is justified in Appendix A, where we show that the combination of a stratified sample with a Poisson sample from its complement is a stratified sample with the original stratification.

In all cases except *discr_iden*, the dual-frame Combined point estimate is contained in the probability CI and is usually quite close to the probability point estimate, indicating that the RDS referrals have been successfully added to the probability sample without generating excess bias. In all but one case (*cigarillo*), the Combined standard error and CI width is smaller than the probability-only standard error and CI width: on average, the dual-frame values are 0.835 times the probability-only values. While this factor of 0.835 is estimated, it is comparable and slightly higher than $0.794 = \sqrt{182/(182 + 107)}$, the factor if the decrease were purely due to increasing sample size by 107 LGBT persons.

5 Discussion

We have considered respondent-driven sampling initialized with a relatively large probability sample of seeds and with relatively short recruitment chains, so that assumptions of standard RDS estimators are not met. Recent literature shows that such samples can occur in practice. For such samples, we propose a dual-frame estimation approach that treats the RDS recruits as a nonprobability sample with unknown inclusion probabilities, estimates the unknown inclusion probabilities using pseudolikelihood, combines the probability seeds with the nonprobability recruits using dual-frame methods, and produces point estimates and variance estimates using weighted estimation in standard survey software. In a limited simulation study with Project 90 network data, our proposed estimators perform well and dominate existing RDS estimators with respect to mean squared error and confidence interval coverage, for a range of initial sample sizes and response rates for random seeds, recruitment behaviors, and binary outcomes. The estimation approach yields sensible estimates with real data from an RDS study of LGBT smoking behavior, and appears to have promise in the setting considered here, in which classical assumptions of RDS estimation are not met.

Under some designs, it is possible to determine the inclusion probabilities under the seed sampling design for any individual (seed or recruit). This was not possible in our

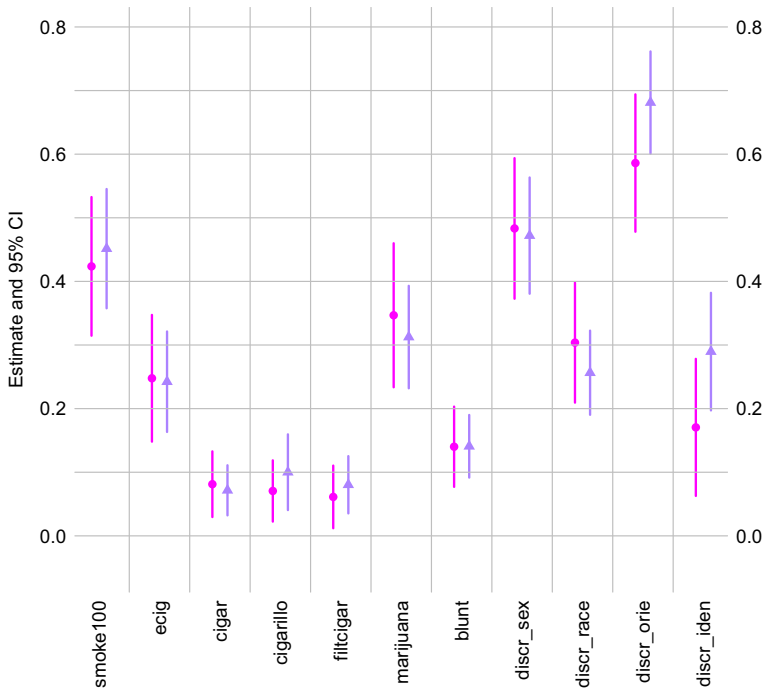


Fig. 6 Point estimates and approximate 95% confidence intervals of eleven binary characteristics from the NORC pilot study, using the probability-only sample of seeds (left, magenta with solid circle) and the dual-frame sample combining seeds and referrals (right, purple with solid triangle). See text for description of the eleven variables

empirical example, so we relied on sample matching to assign inclusion probabilities under the seed sampling design for the recruits. An interesting direction for further research would be to explore the implications of such matching for properties of the combined estimator. Comparisons to other approaches for combining probability and nonprobability samples would also be of interest. Finally, variance and confidence interval estimators that account for uncertainty due to parameter estimation while remaining robust to misspecification of the recruitment mechanism would be worthy of further exploration.

Supplementary information In the supplementary material, we include versions of Fig. 1 through Fig. 5 for all simulation settings described in Table 2.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s40300-023-00241-8>.

Declarations

Conflict of interest The authors declare no competing interests.

Appendix A

The combination of a stratified sample with a Poisson sample from its complement is a stratified sample with the original stratification. It suffices to show that the combined sample membership indicators are uncorrelated across strata, so that estimates from different strata are uncorrelated. Let elements k, ℓ belong to different strata under the original design, so that $\text{Cov}(I_k^A, I_\ell^A) = 0$. Then, because I_k^B and I_ℓ^B are conditionally independent under Poisson sampling from the complement of the A sample, we have

$$\begin{aligned} \text{Cov}(I_k, I_\ell) &= \text{Cov}\left(I_k^A + (1 - I_k^A)I_k^B, I_\ell^A + (1 - I_\ell^A)I_\ell^B\right) \\ &= \text{Cov}\left(\text{E}\left[I_k^A + (1 - I_k^A)I_k^B \mid s_A\right], \text{E}\left[I_\ell^A + (1 - I_\ell^A)I_\ell^B \mid s_A\right]\right) \\ &\quad + \text{E}\left[\text{Cov}\left(I_k^A + (1 - I_k^A)I_k^B, I_\ell^A + (1 - I_\ell^A)I_\ell^B \mid s_A\right)\right] \\ &= \text{Cov}\left(I_k^A + (1 - I_k^A)\rho_k, I_\ell^A + (1 - I_\ell^A)\rho_\ell\right) \\ &\quad + \text{E}\left[(1 - I_k^A)(1 - I_\ell^A)\text{Cov}\left(I_k^B, I_\ell^B \mid s_A\right)\right] \\ &= 0. \end{aligned}$$

References

- Baraff, A.J., McCormick, T.H., Raftery, A.E.: Estimating uncertainty in respondent-driven sampling using a tree bootstrap method. *Proc. Natl. Acad. Sci.* **113**(51), 14668–14673 (2016)
- Chen, Y., Li, P., Wu, C.: Doubly robust inference with nonprobability survey samples. *J. Am. Stat. Assoc.* **115**(532), 2011–2021 (2020)
- D’Orazio, M.: StatMatch: statistical matching or data fusion. (2022). R package version 1.4.1. <https://CRAN.R-project.org/package=StatMatch>
- Elliott, M.R., Valliant, R.: Inference for nonprobability samples. *Stat. Sci.* **32**(2), 249–264 (2017)
- Fellows, I.E.: Respondent-driven sampling and the homophily configuration graph. *Stat. Med.* **38**(1), 131–150 (2019)
- Ganesh, N., Pineau, V., Chakraborty, A., Dennis, J.M.: Combining probability and non-probability samples using small area estimation. In: *Proceedings of the Section on Survey Research Methods*, pp. 1657–1667. American Statistical Association, Alexandria, VA (2017)
- Gile, K.J.: Improved inference for respondent-driven sampling data with application to hiv prevalence estimation. *J. Am. Stat. Assoc.* **106**(493), 135–146 (2011)
- Gile, K.J., Handcock, M.S.: Respondent-driven sampling: an assessment of current methodology. *Sociol. Methodol.* **40**(1), 285–327 (2010)
- Goel, S., Salganik, M.J.: Assessing respondent-driven sampling. *Proc. Natl. Acad. Sci.* **107**(15), 6743–6747 (2010)
- Hájek, J.: Comment on an essay on the logical foundations of survey sampling, part one. *Found. Surv. Sample* **236** (1971)
- Handcock, M.S., Fellows, I.E., Gile, K.J.: RDS: respondent-driven sampling. Los Angeles, CA (2021). R package version 0.9-3. <https://CRAN.R-project.org/package=RDS>
- Hansen, M.H., Hurwitz, W.N.: On the theory of sampling from finite populations. *Ann. Math. Stat.* **14**(4), 333–362 (1943)
- Heckathorn, D.D.: Respondent-driven sampling: a new approach to the study of hidden populations. *Soc. Probl.* **44**(2), 174–199 (1997)
- Heckathorn, D.D., Cameron, C.J.: Network sampling: From snowball and multiplicity to respondent-driven sampling. *Ann. Rev. Sociol.* **43**, 101–119 (2017)
- Horvitz, D.G., Thompson, D.J.: A generalization of sampling without replacement from a finite universe. *J. Am. Stat. Assoc.* **47**(260), 663–685 (1952)
- Huang, C.-M.: Topics in estimation for messy surveys: imperfect matching and nonprobability sampling. Ph.D. dissertation, Colorado State University (2022)

17. Kim, J.K., Park, S., Chen, Y., Wu, C.: Combining non-probability and probability survey samples through mass imputation. *J. R. Stat. Soc. A Stat. Soc.* **184**(3), 941–963 (2021)
18. Kim, J.-K., Tam, S.-M.: Data integration by combining big data and survey sample data for finite population inference. *Int. Stat. Rev.* **89**(2), 382–401 (2021)
19. Kim, J.K., Wang, Z.: Sampling techniques for big data analysis. *Int. Stat. Rev.* **87**, 177–191 (2019)
20. Lee, S., Ong, A.R., Elliott, M.: Exploring mechanisms of recruitment and recruitment cooperation in respondent driven sampling. *J. Off. Stat.* **36**(2), 339 (2020)
21. Michaels, S., Pineau, V., Reimer, B., Ganesh, N., Dennis, J.M.: Test of a hybrid method of sampling the lgbt population: Web respondent driven sampling with seeds from a probability sample. *J. Off. Stat.* **35**(4), 731–752 (2019)
22. Middleton, D., Drabble, L.A., Krug, D., Karriker-Jaffe, K.J., Mericle, A.A., Hughes, T.L., Iachan, R., Trocki, K.F.: Challenges of virtual rds for recruitment of sexual minority women for a behavioral health study. *J. Surv. Stat. Methodol.* **10**(2), 466–488 (2022)
23. Rivers, D.: Paper Prepared for the 2007 Joint Statistical Meetings. Salt Lake City, UT (2007)
24. Sakshaug, J.W., Wiśniowski, A., Ruiz, D.A.P., Blom, A.G.: Supplementing small probability samples with nonprobability samples: a Bayesian approach. *J. Off. Stat.* **35**(3), 653–681 (2019)
25. Salganik, M.J.: Variance estimation, design effects, and sample size calculations for respondent-driven sampling. *J. Urban Health* **83**(7), 98–112 (2006)
26. Salganik, M.J., Heckathorn, D.D.: Sampling and estimation in hidden populations using respondent-driven sampling. *Sociol. Methodol.* **34**(1), 193–239 (2004)
27. Valliant, R.: Comparing alternatives for estimation from nonprobability samples. *J. Surv. Stat. Methodol.* **8**(2), 231–263 (2020)
28. Volz, E., Heckathorn, D.D.: Probability based estimation theory for respondent driven sampling. *J. Off. Stat.* **24**(1), 79–97 (2008)
29. Wiśniowski, A., Sakshaug, J.W., Perez Ruiz, D.A., Blom, A.G.: Integrating probability and nonprobability samples for survey inference. *J. Surv. Stat. Methodol.* **8**(1), 120–147 (2020)
30. Yang, M., Ganesh, N., Mulrow, E., Pineau, V.: Estimation methods for nonprobability samples with a companion probability sample. In: Proceedings of the Section on Survey Research Methods, pp. 1715–1723. American Statistical Association, Alexandria, VA (2018)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.