Innovations in the Recruitment of Respondent Driven Samples for Improved Inference to Hidden Populations

Ashton M. Verdery, M. Giovanna Merli, James Moody, Jeff Smith

Abstract

Respondent driven sampling (RDS) is a data collection approach for hidden and rare populations which relies on respondents' social networks to recruit participants and uses post-sample weighting procedures to obtain population representative estimates. RDS's ease of recruiting study participants quickly and cost-effectively comes at the cost of multiple assumptions about the sampling process which have been shown to be violated in practice with resulting large biases. Whereas prior work to improve the ability of RDS to produce valid and precise estimates of the characteristics of a hidden population has focused on the development of new statistical estimators, we test innovations to the RDS sample recruitment process itself. Using simulated sampling and empirically informed social networks of a hidden population of female sex workers in China, we examine whether modifications to the sampling procedure robustly improve RDS inference across a range of contexts.

Extended Abstract

Populations are "hidden" or "hard-to-reach" if they are effectively impossible to sample using conventional survey methods that require predefined sampling frames (Heckathorn 1997). Examples include population groups of key interest to contemporary public health research: intravenous drug users, female sex workers (FSWs) and men who have sex with men, who may be at high risk of acquiring or transmitting sexually transmitted infections (STI) including HIV/AIDS. The extent of public health problems within these hidden populations, and their impacts on the health dynamics of the general population, are difficult to discern, however, since traditional observational schemes—from direct observation to clinic-based inquiries to snowball sampling—lack a basis for inferring representation.

Respondent driven sampling (RDS) has emerged as one of the dominant methods for surveying hidden populations in a manner which can yield population representative estimates (White et al. 2012). Hundreds of studies have been conducted using RDS (Malekinejad 2008), and the U.S. National Institutes of Health have invested more than $150 million in grants listing RDS as a keyword (Verdery et al. 2014a). RDS samples are recruited through peer referral and a dual incentive structure – where respondents are compensated for both participating and recruiting new participants. Researchers then apply one of several statistical estimators to the sample data to account for the sampling design and biases introduced by the referral process (Gile and Handcock 2010; Tomas and Gile 2011).

Several methodological evaluations can be found in the literature. Taken together, they offer the following conclusions: a) RDS makes stringent assumptions about the population being sampled and the sampling process that are not met in practice; b) failure to meet these assumptions – especially the assumption that individuals recruit from their peers at random without respect to peers' attributes – leads to substantial biases in sample estimates of the population mean; c) even when RDS's assumptions are met, its estimates of the population mean remain highly uncertain because they can exhibit large sample to sample variance and because current variance estimation techniques have large biases (Goel and Salganik 2009; Goel and Salganik 2010; Gile and Handcock 2010; Lu et al. 2012; McCreesh et al. 2012; Merli et al. 2014; Verdery et al. 2014a). Yet, the RDS method is in widespread use. Public health practitioners praise it because it is a quick, cost-effective, confidential method to recruit samples of historically hidden but epidemiologically relevant populations and because it claims to provide a probabilistic framework for inference from the RDS sample to the hidden population (Platt et al. 2006; lots of others). This suggests that, in spite of its methodological weaknesses, RDS as a sampling strategy is here to stay.

A number of researchers have proposed new statistical approaches to improve estimation and inference from data collected via the traditional RDS protocols (Gile 2011; Gile and Handcock 2011; Gile, Johnston and Salganik 2014). Less attention has been devoted to altering the RDS sample recruitment process. Those who have considered changing this process have either proposed abandoning the respondent driven nature of the recruitment and directing the sampling process to fully explore the underlying network, e.g., Mouw and Verdery (2012), or they have invoked supplemental data collection to validate RDS assumptions (e.g., Yamanis et al. 2013) or have developed more robust statistical estimators (Lu 2013) which take advantage of the collection of supplemental data without changing the basic recruitment dynamics.

In this paper, we rely on simulation methods to test innovations to the conventional RDS sample recruitment process. We propose novel strategies for sample recruitment and test their efficacy within a simulation-evaluation framework. We parameterize our simulations with data from multiple sources: eight synthetic population social networks grounded in data on female sex workers in China from the PLACE-RDS Comparison Study (PRCS; Weir et al. 2012; Merli et al. 2014). We simulate two types of innovations to the RDS sample recruitment strategy and test their robustness to violations of RDS assumptions about referral biases, which prior work has identified as being particularly important (Yamanis et al. 2013; Gile and Handcock 2010). We first simulate a set of cases where RDS recruitment chains on a social network do not exhibit referral bias. Next, to characterize degrees of sensitivity to referral bias, we simulate sets of additional chains which exhibit increasingly large recruitment biases. This is a departure from prior literature which has tended to consider exclusively scenarios with and without referral bias. We supplement this analysis by locating empirically observed recruitment bias as measured with the unique data collected in the PRCS along these sensitivity curves to characterize where empirical samples tend to fall in this simulated framework. With this characterization in place, we next explore how RDS sample recruitment innovations shift the sensitivity curves. The innovations we test are as follows:

1) "Re-seeding" the sample after a set number of sampling units have been recruited. We achieve this by presenting additional coupons and recruitment opportunities to respondents at the time of their follow-up interviews where they come to collect the incentive payments for successfully recruiting additional survey participants. In re-seeding, we prioritize offering additional coupons to respondents who have recruited peers that differ from themselves along key dimensions or from the composition of the sample collected to that point as assessed by examining their set of recruits.
2) Differentially incentivizing recruitment according to geographic or social distance. Observations of social networks typically find strong clustering along a number of social and spatial dimensions (McPherson et al. 2001). We examine whether providing respondents extra incentives for recruiting others who differ substantially in these dimensions affects the efficacy of the sample. To do this, we examine a range of price sensitivities which accounts for the possibility that some respondents are not responsive to additional incentives.

In our simulation results, we evaluate the sensitivity of three key RDS statistical estimators – the naïve/unadjusted sample mean (Naïve), the RDS2/Volz-Heckathorn estimator (RDS2-VH) proposed by Volz and Heckathorn (2008), and the Linked Ego Networks (RDS1-

LEN) estimator proposed by Lu (2013). Recent work shows the RDS1-LEN estimator greatly outperforms all other RDS estimators in terms of inference, while the naïve sample mean performs second best (Verdery et al. 2014b); we include the RDS2-VH estimator because it is most often used in practice. Following Verdery et al. (2014b), we focus on distributional properties of these estimators and examine both biasedness of the resultant samples in terms of mean deviation from a population parameter and the sampling variance or design effects generated by each of these estimators.

Our goal is to produce general rules, adaptable to other hidden populations with different referral dynamics than female sex workers in China, and to offer guidance on the types of additional RDS data collection modules that could be collected to drive feasible adaptations of the RDS sample recruitment procedures to reach optimal population coverage so as to reduce biases and sampling variance using current estimators. These innovations, especially if presented in the form of guidelines for RDS practitioners, are critical for preserving and improving this effective sampling strategy for hidden populations.

References

Gile KJ, Handcock MS. Respondent-Driven Sampling: An Assessment of Current Methodology. *Sociol Methodol*. 2010;40(1):285–327. doi:10.1111/j.1467-9531.2010.01223.x.

Gile KJ. Improved Inference for Respondent-Driven Sampling Data With Application to HIV Prevalence Estimation. *J Am Stat Assoc*. 2011;106(493):135-146. doi:10.1198/jasa.2011.ap09475.

Gile, K. J., Johnston, L. G., & Salganik, M. J. (2014). Diagnostics for respondent-driven sampling. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*.

Gile, Krista J., and Mark S. Handcock. "Network model-assisted inference from respondent-driven sampling data." *arXiv preprint arXiv:1108.0298* (2011).

Goel S, Salganik MJ. Assessing respondent-driven sampling. *Proc Natl Acad Sci*. 2010;107(15):6743-6747. doi:10.1073/pnas.1000261107.

Goel S, Salganik MJ. Respondent-driven sampling as Markov chain Monte Carlo. *Stat Med*. 2009;28(17):2202–2229. doi:10.1002/sim.3613.

Heckathorn DD. Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations. *Soc Probl*. 1997;44(2):174-199. doi:10.2307/3096941.

Lu X, Bengtsson L, Britton T, et al. The sensitivity of respondent-driven sampling. *J R Stat Soc Ser A Stat Soc*. 2012;175(1):191-216. doi:10.1111/j.1467-985X.2011.00711.x.

Lu X. Linked Ego Networks: Improving estimate reliability and validity with respondent-driven sampling. *Soc Netw*. 2013;35(4):669-685. doi:10.1016/j.socnet.2013.10.001.

Malekinejad M, Johnston LG, Kendall C, Kerr LRFS, Rifkin MR, Rutherford GW. Using Respondent-Driven Sampling Methodology for HIV Biological and Behavioral Surveillance in International Settings: A Systematic Review. *AIDS Behav*. 2008;12(1):105-130. doi:10.1007/s10461-008-9421-1.

McCreesh, Nicky, Simon Frost, Janet Seeley, Joseph Katongole, Matilda Ndagire Tarsh, Richard Ndunguse, Fatima Jichi et al. "Evaluation of respondent-driven sampling." *Epidemiology (Cambridge, Mass.)* 23, no. 1 (2012): 138.

McPherson, Miller, Lynn Smith-Lovin, and James M. Cook. "Birds of a feather: Homophily in social networks." *Annual review of sociology* (2001): 415-444.

Merli MG, Moody J, Smith J, Li J, Weir S, Chen X. Challenges to Recruiting Population Representative Samples of Female Sex Workers in China Using Respondent Driven Sampling. *Soc Sci Med*. 2014. doi:10.1016/j.socscimed.2014.04.022.

Mouw T, Verdery AM. Network Sampling with Memory A Proposal for More Efficient Sampling from Social Networks. *Sociol Methodol*. 2012;42(1):206-256. doi:10.1177/0081175012461248.

Platt L, Wall M, Rhodes T, et al. Methods to Recruit Hard-to-Reach Groups: Comparing Two Chain Referral Sampling Methods of Recruiting Injecting Drug Users Across Nine Studies in Russia and Estonia. *J Urban Health*. 2006;83(1):39-53. doi:10.1007/s11524-006-9101-2.

Tomas A, Gile KJ. The effect of differential recruitment, non-response and non-recruitment on estimators for respondent-driven sampling. *Electron J Stat*. 2011;5:899-934. doi:10.1214/11-EJS630.

Verdery AM, Merli MG, Moody J, Smith J, Fisher J. *Assement of Multiple RDS Estimators Under Real and Ideal Recruitment Scenarios*.; 2014b. Under Review.

Verdery AM, Mouw T, Bauldry S, Mucha PJ. *Network Structure and Biased Variance Estimation in Respondent Driven Sampling*.; 2014a. Available at: http://arxiv.org/abs/1309.5109. Accessed September 25, 2013.

Volz E, Heckathorn DD. Probability based estimation theory for respondent driven sampling. *J Off Stat*. 2008;24(1):79.

Weir SS, Merli MG, Li J, et al. A comparison of respondent-driven and venue-based sampling of female sex workers in Liuzhou, China. *Sex Transm Infect*. 2012;88(Suppl 2):i95-i101. doi:10.1136/sextrans-2012-050638.

White RG, Lansky A, Goel S, et al. Respondent driven sampling—where we are and where should we be going? *Sex Transm Infect*. 2012;88(6):397-399. doi:10.1136/sextrans-2012-050703.

Yamanis TJ, Merli MG, Neely WW, et al. An Empirical Analysis of the Impact of Recruitment Patterns on RDS Estimates among a Socially Ordered Population of Female Sex Workers in China. *Sociol Methods Res*. 2013. doi:10.1177/0049124113494576.