Proceedings of Statistics Canada Symposium 2013
Producing reliable estimates from imperfect frames

# Respondent-Driven Sampling: Origins, Current Developments, and Alternative Estimators

Douglas D. Heckathorn, Christopher Cameron, and Yongren Shi[1]

## Abstract

Respondent-Driven Sampling (RDS) has become the method of choice in studies of hidden and hard-to-reach populations. RDS, which is a form of network sampling, provides a means for drawing probability samples of populations which cannot be effectively sampled using traditional population survey methods because they lack a sampling frame, or because these populations have social networks that are hard for outsiders to penetrate due to stigma or privacy concerns. Examples include populations of relevance to public health, such as drug users and sex trafficking victims; and on populations of relevance to arts and culture such as jazz musicians. This paper summarizes the methods and describes two recent developments: improved methods for variance estimation and controlling for bias resulting from large sampling fractions.

Key Words: Respondent-Driven Sampling; Hidden populations; Hard-to-reach populations.

## 1. Introduction

Hard-to-reach populations are difficult to sample because they lack a sampling frame (list of population members) and they are a small part of the general population who have privacy concerns or whose networks that are hard for outsiders to penetrate (*e.g.*, drug users, gay men). Consequently, it has not been possible to answer questions such as: How many Latino restaurant workers in Chicago are paid less than the minimum wage? How many drug users in Chicago are HIV positive? How many homeless people live in Houston? Additional groups are emerging that are "hard-to-reach" including undocumented workers, and those that use satellite, internet, or cell phones.

Due to the difficulty of sampling such populations, three less-than-ideal sampling methods were generally employed including: (1) Institutional Sampling draws probability samples of those with institutional contacts, but it excludes those who avoid those contacts and carry different demographic characteristics. For involuntary such as prisons or jails, these tend to be lower status individuals; and for voluntary institutions such as professional associations these tend to be higher status individuals; (2) Time/Location Sampling draws probability samples of those who are accessible in large public venues, but coverage is limited because it excludes those who shun those settings. For example, in drug studies, this tends to exclude many women and young people; (3) Chain Referral Sampling (Snowball-Type Methods) have greater coverage, because respondents are reached through their social networks, but they have been convenience rather than probability sampling methods. Hence the dilemma: a choice between statistical validity (*i.e.*, methods #1 and #2) and coverage (*i.e.*, method #3).

## 2. Origins

A series of papers originating with Heckathorn (1997) developed Respondent-driven sampling (RDS), a new method for collecting and analyzing chain-referral that challenged the prevailing view of chain-referral sampling as a non-probability convenience method. The RDS enterprise, in which data from chain-referral samples become the basis for statistically valid population estimates, is now an active research area pursued by numerous independent research teams, including Krista J. Gile and Mark S. Handcock from Hard-to-Reach Population Methods Research Group, Matthew Salganik from Princeton, scholars from Stockholm University, and a team from the National Center for Health Statistics (NCHS).

[1] Douglas D. Heckathorn, Christopher Cameron, and Yongren Shi, Cornell University.

Prior to the development of RDS, the conventional wisdom, as expressed by Erickson (1987) was that since the chain referral sample begins with non-randomly selected seeds, chain referral samples are tainted by an unknown selection bias. Chain referral could produce probability samples only if the seed selection bias could be determined. This created a paradox for researchers studying hidden populations: If the initial subjects who serve as "seeds" could be drawn randomly, the population would not be qualified as "hard-to-reach." Furthermore, according to Erickson, this unknown bias from the seeds is further multiplied in unknown ways when the sample expands from the seeds to wave #1, from wave #1 to wave #2, and so forth.

In contrast, the original RDS paper (Heckathorn 1997) modeled recruitment in a chain-referral sample as a first-order Markov process, where states correspond to relevant group characteristics, *e.g*., gender. Recruitment then involves either a state change (*e.g*., female to male recruitment) or a stable state (*e.g*., female to female recruitment). The article then drew on a theorem termed the "law of large numbers for regular Markov Chains," to show that if the sample expanded through enough waves, it would attain an equilibrium which was independent of the starting point; that is, independent of the initial subjects from which the sample began. The implication was that bias from the choice of seeds can be overcome if recruitment chains are sufficiently long. A further property of regular Markov chains holds that this equilibrium will be attained at a rapid (*i.e*., geometric) rate rather than a slow (*i.e*., algebraic) rate.

This initial demonstration showed that a chain-referral sample could be a reliable sampling method, because the same equilibrium sample would be drawn irrespective of the convenience sample of seeds from which sampling began. However, the Markov equilibrium generally did not correspond to a valid population estimator, so bias remained.

The second analytic component of the initial RDS paper showed the conditions under which a chain-referral sample would yield an unbiased sample. The analysis drew on Anatol Rapoport's (1979) "biased network theory" in which networks grow through a stochastic process. Though Rapoport's focus was on network structure, the initial RDS paper extended the analysis to chain-referral samples. This was shown through an original theorem (see Heckathorn 1997, Theorem 3, p. 192) showing that if groups had equal "in-breeding bias," as defined by biased network theory, or equivalently, equal "homophily", the sample would be unbiased. This showed that chain-referral samples could be not only reliable, but also valid. However, a significant limitation remained; the method continued to be biased when homophily was unequal.

## 3. Current developments

A subsequent paper (Heckathorn 2002) introduced a more refined RDS population estimator designed to yield unbiased samples when homophily was unequal. Drawing not only on the data from the recruitment matrix but also from self-reported network sizes, the estimators compensated both for differences in homophily across groups and for differences in the mean degree (*i.e*., personal network size) across groups. This was accomplished through what was termed the "reciprocity model." The essential idea was that respondents recruit acquaintances, friends, and relatives, so their relationships tend to be reciprocal. Therefore, the number of ties linking any two groups must be the same in both directions—for example, monogamous marriage is a reciprocal relationship, so for any two groups, X and Y, the number of Xs married to Ys must equal the number of Ys married to Xs. From this elemental network property, proportional group sizes can be calculated based on two types of network information, the proportion of cross-cutting ties between the groups and the relative sizes of each group's networks. Drawing the former information from the recruitment matrix, and the latter from the respondents' self-reported network size, the estimator was calculated based on these two types of network information (see Heckathorn 2002;22). Given that this controlled for the effects of differences in homophily and network size across groups, these estimators became validly applicable across the full range of RDS data sets in which these two network attributes are generally different. This paper also introduced a Bootstrap-based method for estimating the variance of the population estimates.

An additional paper (Salganik and Heckathorn 2004) introduced a further refinement of the RDS estimator, which employed a multiplicity approach to estimate relative group network sizes. It also introduced a proof that this RDS estimator is asymptotically unbiased when the assumptions of the method are met—that is, bias is only on the order of 1/[sample size]. Consequently, bias is minor in samples of substantial size if the assumptions are satisfied. A

further paper derived a RDS estimator using only network data (Volz and Heckathorn, 2008) and included an analytically derived measure of the estimator's variance. Finally, another paper reduced the number of assumptions on which the RDS estimation process is based, and introduced means for analyzing continuous variables (Heckathorn 2007).

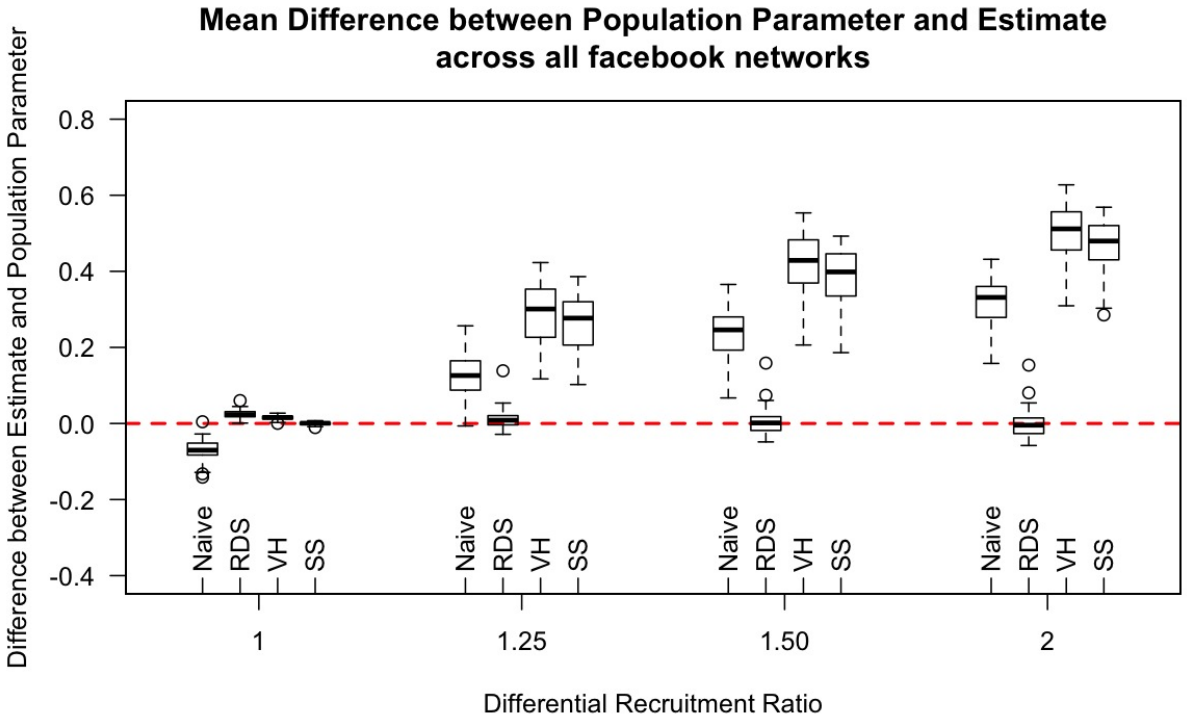
The initial reactions to RDS were mixed. Given the prevailing view that snowball-type methods were strictly limited to convenience sampling, some rejected the possibility of transforming chain-referral sampling into a statistically valid method. This included an effort by an OMB statistician to block a four-city NEA-funded RDS study of jazz musicians by suggesting that any such an endeavor had no more validity than trying to combine astrology with statistics, and that those connected to such an endeavor were jeopardizing their professional reputations. Others embraced the RDS method because of its ease of application and the lack of competing methods which would be both practical in application and sound analytically. These ease of acceptance of RDS also became greater with the development of methods employing a similar logic, especially Steven Thompson's (Thompson and Frank, 2000) development of adaptive sampling and link-tracing designs.

## 4. Alternative estimators

More recently there have been more than a dozen conferences devoted to RDS, RDS has been applied in several dozen countries (Johnston *et al*, 2008), literature has emerged evaluating the validity of the assumptions of RDS (*e.g*., see Wejnert, 2009), and a growing number of alternative RDS estimators have been proposed (*e.g*., see Gile 2011).

Evaluating methods for sampling hidden populations is difficult because, by definition, the population parameters are not known. One approach is to sample from a known population, *e.g*., see Wejnert's (2009) study of university undergraduates. Here we focus on a second approach, sampling from a population with a fully specified social network. Given such a network, simulated samples can be drawn repeatedly to measure the variances associated with population estimates, something which is generally impractical when sampling from a known population. Drawing simulated samples from a fully specified network also makes it possible to explore the effects of variations in recruitment behavior on the bias of estimates. We demonstrate this approach using a collection of Facebook network snapshots from 2005 (Traud *et. al*. 2011), selecting the 87 universities with at least 2,000 users to evaluate four alternative RDS estimators. We compare the performance of the naïve estimator (sample proportion), the Salganik-Heckathorn estimator (Salganik and Heckathorn 2004), the Volz-Heckathorn estimator (Volz and Heckathorn 2008), and Gile estimator (Gile 2011). The analyses focus on differential recruitment, a common form of bias in RDS data sets, in which subgroups have both differing homophily and recruitment effectiveness, and hence the more effectively recruiting group is oversampled. This form of bias is typical in studies of injection drug users (IDUs), when HIV positives respondents are homophilous, and also recruit greater numbers of peers than do HIV negatives (*e.g*., see Ramirez-Valles, 2005).

The figure compares the average error in estimates of the proportion of freshmen across four levels of differential recruitment activity (DF): groups recruit equally (DF=1), freshmen recruit about one quarter more than the non-freshmen (DF = 1.25), one half more (DF = 1.5), or twice as much as non-freshman (DF = 2). The sampling fraction is held constant at 30 percent. The vertical axis indicates the difference between the estimate and the population parameter. It is apparent by inspection that for three of the four estimates, bias increases substantially with differential recruitment activity. Bias is greatest for the Volz-Heckathorn (2008) estimate (*i.e*., RDS II), a feature resulting from the construction of the estimator; it ignores recruitment patterns and derives estimates only from the network sizes of respondents. Bias is less for Gile's "sequential" estimator, a feature which reflects this estimator's derivation from the Volz-Heckathorn estimator. The "naïve" estimator, which consists merely of the mean sample composition, has a slightly lower level of bias. Finally, bias is low for the Salganik-Heckathorn (2004) estimator.




## References


Erickson, B.H. (1979), "Some Problems of Inference from Chain Data", *Sociological Methodology*, 10:276-302.

Gile, K.J. (2011), "Improved Inference for Respondent-Driven Sampling Data with Application to HIV Prevalence Estimation", *Journal of the American Statistical Association*, 106(498): 135-146.

Gile, K.J. and M.S. Handcock (2010), "Respondent-Driven Sampling: An Assessment of Current Methodology", *Sociological Methodology*, 40:285-327.

Goel, D. and M.J. Salganik (2009), "Respondent-Driven Sampling as Markov Chain Monte Carlo", *Statistics in Medicine*, 28:2202-2229.

Heckathorn, D.D. (1997), "Respondent-Driven Sampling: A New Approach to The Study of Hidden Populations", *Social Problems*, 44: 174-99.

Heckathorn, D.D. (2002), "Respondent-Driven Sampling II: Deriving Valid Population Estimates from Chain-Referral Samples of Hidden Populations", *Social Problems*, 49: 11-34.

Heckathorn, D.D. (2007), "Extensions of Respondent-Driven Sampling: Analyzing Continuous Variables and Controlling for Differential Recruitment", *Sociological Methodology*, 37: 151-207.

Heckathorn, D.D. and J. Jeffri (2003), "Social Networks of Jazz Musicians" in Changing the Beat: A Study of the Worklife of Jazz Musicians, Volume III: Respondent-Driven Sampling: Survey Results by the Research Center for Arts and Culture, National Endowment for the Arts Research Division Report #43, Washington DC.

Johnston, L.G., Malekinejad, M., Kendall, C., Iuppa, I.M. and G.W. Rutherford (2008), "Implementation challenges to using respondent-driven sampling methodology for HIV biological and behavioral surveillance: field experiences in international settings", *AIDS and Behavior*, 12(4 Suppl):S131-141.

Kajubi, P., Kamya, M.R., Raymond, H.F., Chen, S., Rutherford, G.W., Mandel, J.S., and W. McFarland (2008), "Gay and bisexual men in Kampala, Uganda", *AIDS and Behavior*, 12(3):492-504.

Lansky, A., Drake, A., and H.T. Pham (2009), "HIV-Associated Behaviors among Injecting-Drug Users -- 23 Cities, United States, May 2005-February 2006", Morbidity and Mortality Weekly Report April 10, 2009, 58(13):329-332.

Ramirez-Valles, J., Heckathorn, D.D., Vázquez, R., Diaz, R.M. and R.T. Campbell (2005), "From Networks to Populations: The Development and Application of Respondent-Driven Sampling among IDUs and Latino Gay Men", *AIDS and Behavior*, 9(4):387-402.

Rapoport, A. (1979), "A Probabilistic Approach to Networks", *Social Networks*, 2:1-18.

Salganik, M.J. and D.D. Heckathorn (2004), "Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling," *Sociological Methodology*, 34:193-239.

Thompson, S.K. and O. Frank (2000), "Model-Based Estimation with Linktracing Sampling Designs", *Survey Methodology*, 26(1):87-98.

Traud, A.L., Mucha, P.J. and M.A. Porter (2011), "Social Structure of Facebook Networks", arXiv:1102.2166

Volz, E and D.D. Heckathorn (2008), "Probability Based Estimation Theory For Respondent Driven Sampling", *Journal of Official Statistics*, 24: 79-97.

Wejnert, C. (2009), "An Empirical Test of Respondent-Driven Sampling: Point Estimates, Variance, Degree Measures, and Out-of-Equilibrium Data", *Sociological Methodology*, 39: 73-116.