NIH Public Access
Author Manuscript
Sociol Methodol. Author manuscript; available in PMC 2012 August 1.

Published in final edited form as:
Sociol Methodol. 2011 August 1; 41(1): 355–366. doi:10.1111/j.1467-9531.2011.01244.x.

SNOWBALL VERSUS RESPONDENT-DRIVEN SAMPLING

Douglas D. Heckathorn[*]
[*]Cornell University

Leo Goodman (2011) provided a useful service with his clarification of the differences among snowball sampling as originally introduced by Coleman (1958–1959) and Goodman (1961) as a means for studying the structure of social networks; snowball sampling as a convenience method for studying hard-to-reach populations (Biernacki and Waldorf 1981); and respondent-driven sampling (RDS), a sampling method with good estimability for studying hard-to-reach populations (Heckathorn 1997, 2002, 2007; Salganik and Heckathorn 2004; Volz and Heckathorn 2008).

This comment offers a clarification of a related set of issues. One is confusion between the latter form of snowball sampling, and RDS. A second is confusion resulting from multiple forms of the RDS estimator that derives from the incremental manner in which the method was developed. This comment summarizes the development of the method, distinguishing among seven forms of the estimator.

1. SNOWBALL SAMPLING

As described in Leo Goodman's (2011) comment, snowball sampling was developed by Coleman (1958–1959) and Goodman (1961) as a means for studying the structure of social networks. Several years after Coleman's and Goodman's development of snowball sampling, what was also termed snowball sampling emerged as a nonprobability approach to sampling design and inference in hard-to-reach, or equivalently, hidden populations. Sampling these populations is difficult because standard statistical sampling methods require a list of population members (i.e., a "sampling frame") from which the sample can be drawn. Yet for a hidden population, constructing the frame using methods such as household surveys is infeasible when the population is small relative to the general population, geographically dispersed, and when population membership involves stigma or the group has networks that are difficult for outsiders to penetrate (Sudman and Kalton 1986). Groups with these characteristics are relevant to research in many areas, including public health (e.g., drug users), public policy (e.g., illegal immigrants), and arts and culture (e.g., musicians).

This nonprobability form of snowball sampling became a widely employed method in qualitative research on hard-to-reach populations. In a review article, Biernacki and Waldorf (1981) observed that beginning with Becker's (1963) study of marijuana smokers; snowball sampling had become both a standard technique in qualitative research, and a topic in methodology textbooks. In this literature, "chain-referral sampling" served as the general term for a class of sampling methods including not only snowball sampling, but also link-tracing designs, Klovdahl's "random walk" approach, and others.

Chain-referral-sampling of a hidden population begins with a convenience sample of initial subjects, because if a random sample could be drawn, the population would not qualify as

Direct correspondence to Douglas D. Heckathorn, douglas.heckathorn@cornell.edu.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

hidden. These initial subjects serve as “seeds,” through which wave 1 subjects are recruited; wave 1 subjects in turn recruit wave 2 subjects; and the sample subsequently expands wave by wave like a snowball growing in size as it rolls down a hill.

In snowball studies, sample proportions are used to estimate terms such as HIV prevalence. In a typical but disappointingly common pattern, the introduction explains the limits of inference from a convenience sample, then results are presented objectively, and finally the discussion section then analyzes group differences as though these comparisons were based on a probability sample.

This transformation of the snowball method did not go without comment. Spreen (1992:41) noted that “the historical purpose that Coleman had in mind … of using a snowball design to study social structure changed into a total [sic] different purpose of using some kind of snowball design, namely as an expedient for locating members of a special population.” Thus, the transition from snowball sampling for studying network structure to snowball sampling as a convenience sampling method fit the needs of scholars whose exclusive concern was accessing hidden populations.

2. THE EMERGENCE OF RESPONDENT-DRIVEN SAMPLING

The use of snowball sampling in research on hidden populations created a widespread perception of snowball sampling in particular and chain-referral methods in general as convenience sampling methods. Erickson’s (1979) article on problems of inference from chain data from hidden populations expressed the common wisdom. As she states, the sample begins with a convenience sample with bias of unknown magnitude and unknown direction and this bias is then compounded in unknown ways as the sample expands from wave to wave. Hence, as applied to hidden populations, chain-referral samples are inherently limited to convenience samples.

The judgment that chain-referral sampling is a convenience method was challenged in a series of papers leading to the development of a new method for collecting and analyzing chain-referral data, Respondent-driven sampling (RDS). This method was initially developed as part of a NIH/NIDA-funded HIV prevention project in Connecticut (Heckathorn 1997), and has now emerged as a rapidly growing area of research with contributions from numerous independent groups of researchers (e.g., see Goel and Salganik 2009; Gile and Handcock 2010).

The method evolved incrementally in a series of papers which expanded and strengthened the method. Because a different form of the method was presented in each paper, the term RDS refers not to a single method, but to a series of methods that have as their common core an effort to convert chain-referral sampling into a sampling method of good estimability. Table1 summarizes this sequential development. It lists the seven estimators that have appeared in the peer-review literature. Designated E#1 to E#7, the table lists the information required to calculate each estimator, its limitations, and the distinctive contribution offered by each.

The initial paper (Heckathorn 1997) employed a Markov modeling of the peer recruitment process to show that, contrary to the conventional wisdom, bias from the convenience sample of initial subjects was progressively attenuated as the sample expanded wave by wave. This model employed data from peer recruitments to estimate the probability of recruitment across groups. These probabilities were organized into a recruitment matrix, specifying the probability of members of each group (e.g., Hispanics), recruiting members from their own and each other group (e.g., Whites, Blacks, Hispanics, and Others); probabilities that served as the “transition probabilities” of the Markov model. The model

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

showed that as the sample expanded wave by wave, it approached an equilibrium that was independent of the starting point— that is, it was independent of the convenience sample of seeds from which it began. The implication was that this sampling method could potentially become reliable if the number of waves was sufficiently large. That is, any selection of seeds would ultimately produce the same equilibrium sample composition. Hence, it does not matter if the initial sample is nonrandom, if the number of waves reaches a threshold value large enough to eliminate bias from the initial selection of seeds. Furthermore, the analysis showed that bias from the seeds was reduced at a geometric rather than an arithmetic rate, a feature that accelerates the reduction of bias.

Using the Markov equilibrium as the population estimator (E#1), the initial (Heckathorn 1997) RDS paper showed that the sample became self-weighting (see Theorem 3, p. 192) if groups had equal homophily, where homophily was as defined by either the in-group bias estimator in Coleman's (1958–1959) relational analysis, or equivalently, the in-breeding bias estimator in Rapport's (1979) biased network theory. Though independently developed, both approaches to defining homophily are based on the idea that structure in a network, i.e., homophily, occurs when affiliation patterns fit depart from random mixing, such that affiliations are formed in ways that reflect respondent characteristics such as SES, age, or religion.

This analysis demonstrated that, under certain conditions, population estimates derived from a chain-referral sample could become not merely reliable but also valid. However, the paper did not show how an unbiased estimate could be derived for cases where homophily was unequal.

These analyses (Heckathorn 1997) also revealed a significant constraint on the method:

> When inbreeding terms are very large, reflecting mutual isolation of the system's groups, the approach to equilibrium slows, e.g., when the terms exceed .99, scores of waves of recruitment may be required for equilibrium to be approximated. Therefore, given the practical limitations that constrain the number of waves of recruitment, this means that equilibrium will be reached only when inbreeding is not extreme. The implication is that when the boundaries separating groups are virtually impassable, RDS should be used to draw samples from within such groups, and not across them, even should inbreeding terms prove to be equal.

Such a possibility was illustrated in the paper. The analysis focused on a pair of cities (i.e., cities #2 and #3) with more than 99% recruitment from within each city and its adjoining area. To render the analyses tractable, the sample was divided into two subsamples, one for each city.

A subsequent paper (Heckathorn 2002) introduced new RDS population estimators (E#2 and E#3). 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,

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

the new estimators were 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.

Two versions of the estimator (E#2 and E#3) were introduced because of an issue arising when analyzing three-category and larger groups. Whereas in a dichotomous system, the estimator consisted of a single equation, nondichotomous systems were more complex. For calculating population estimates, using the reciprocity model required solving systems of linear equations that were overdetermined—for example, a four-category system required solving a system of seven linear equations with four unknowns. The problem was that, owing to stochastic variation and other factors, differing estimates would be generated based on which four equations were chosen to calculate the four unknowns. A standard approach to solving such systems is linear least squares, which uses a regression-like logic to reconcile differences among the equations (E#2). An alternative based on data smoothing was also introduced (E#3), and the paper concluded future research would be required to identify the best approach.

Subsequent analyses (Heckathorn 2007) showed that data smoothing (E#3) yielded greater statistical efficiency (i.e., about 20% narrower confidence intervals). The reason is that data smoothing pooled cross-group recruitment data to estimate a reduced number of parameters, so each estimate was based on a greater amount of data.

The paper (Heckathorn 2002:27–28) also showed how confidence intervals for RDS population estimates can be generated using a special form of bootstrapping in which subsamples are generated, not through random selection from the sample but from a process that takes into account differences in homophily across groups. That is, the bootstrap subsamples are generated, not through random selection from the sample but rather from generating simulated recruitment chains based on the set of transition probabilities specifying the probabilities of members of each group recruiting members of each other group. It was with the introduction of means for calculating confidence intervals that RDS approaches a probability sampling method. For a more detailed presentation and analysis of this approach, see Salganik (2006).

This paper also clarified the role of homophily in RDS analysis, as illustrated by Heckathorn (2002:28) in a figure showing that as homophily increases, standard error increases exponentially. This reinforces the necessity to subdivide samples at high-homophily breakpoints.

An additional paper (Salganik and Heckathorn 2004) introduced a further estimator (E#4), 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], so bias is minor in samples of substantial size.

Specification of the assumptions must be satisfied for the 2004 estimator (E#4) to yield asymptotically unbiased population estimates provided a theoretically grounded means for assessing bias in RDS samples, through tests to see if the assumptions were approximated. Six assumptions were required for the proof:

1. Respondents know one another as members of the target population, as is typical of groups such as drug users or musicians.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

2. The network of the target population forms a single component. This assumption is plausible if the network was created through a contact pattern, if it has small-world properties, or if its network sizes fit a power-law distribution.
3. Sampling occurs with replacement. Therefore, the sampling fraction must be small enough for a sampling-with-replacement model to be appropriate.
4. Respondents can accurately report their personal network size, i.e., the number of those they know who fit the requirements of the study such as drug injectors or jazz musicians (see Wejnert [2009] for an examination of this assumption).
5. Respondents recruit randomly from their personal networks. This assumption becomes more plausible when members of the target population have easy and nonthreatening access to the research sites.
6. Respondents recruit only a single recruit, so recruitment effectiveness is uniform across groups.

The first five assumptions provide guidance both on when RDS is a suitable method and on suitable research designs. The sixth assumption is frequently counterfactual, because it is common for some groups to recruit more effectively than others, but it was eliminated based on subsequent theoretic development (see the discussion of E#7 below).

A further development of the RDS method occurred in a Volz-Heckathorn (2008) paper that derived two RDS estimators based on network principles. The first of the two, termed RDS II (E#5), permitted analysis of continuous variables. A data smoothed version of this estimator (E#6) was also introduced (Volz and Heckathorn 2008:11), which yielded point estimates equivalent to the original Salganik and Heckathorn (E#4) estimator. Hence, two profoundly different logics, the reciprocity model underlying E#4, and the network approach underlying E#6, produced identical point estimates. This paper also went beyond the simulation-based boot-strap approach to variance estimation by proposing an analyticallyderived estimation procedure.

A further refinement of the RDS estimator (Heckathorn 2007) provided a means for controlling for bias from differential recruitment, in which some groups recruit more effectively than others. This was accomplished by dividing the RDS sampling weight into an individual-based component calculated from respondents’ self-reported network sizes, and a group-based component calculated from the recruitment matrix based on each group’s patterns of recruitment of their own and other groups. This eliminated the need for the sixth assumption in the Salganik-Heckathorn (2004) paper, in which each recruiter had only a single recruit. It also reduced constraints on RDS research design, permitting multistage designs to more effectively sample low-density sectors of social networks (Heckathorn 2007:188).

As the use of RDS expanded, confusion between snowball sampling and RDS became increasingly common, including both articles employing snowball sampling but calling it RDS, and articles saying that the development of RDS showed that snowball sampling was, after all, a form of probability sampling. In an effort to avoid such confusion, a recent article (Magnani et al. 2005: S71) listed the different features of the methods.

RDS has now been employed in more than 120 studies in dozens of countries (Malekinejad et al. 2008). The literature is now expanding in several directions, including sensitivity analyses to assess the effects of violations of its assumptions (Gile and Handcock 2010), development of new estimators which may prove more reliable and valid than current estimators, applying the method to a greater range of groups (Ramirez-Valles et al. 2005),

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

and addressing the ethical issues that arise when studying stigmatized and vulnerable populations (Semaan et al. 2009).

Acknowledgments

This research was supported by National Institutes of Health/NINR award 1R21NR10961. I thank Michael W. Spiller, Christopher J. Cameron, and Vladimir Barash for helpful comments and advice.

REFERENCES

Becker, Howard S. Outsiders: Studies in the Sociology of Deviance. New York: Macmillan; 1963.

Biernacki, Patrick; Waldorf, Dan. Snowball Sampling: Problems and Techniques of Chain Referral Sampling. Sociological Methods and Research. 1981; 10:141–163.

Coleman, James S. Relational Analysis: The Study of Social Organizations with Survey Methods. Human Organization. 1958–59; 17:28–36.

Erickson, Bonnie H. Some Problems of Inference from Chain Data. In: Schuessler, Karl F., editor. Sociological Methodology. Vol. vol. 10. Cambridge, MA: Blackwell Publishers; 1979. p. 276-302.

Gile, Krista J.; Handcock, Mark S. Respondent-Driven Sampling: An Assessment of Current Methodology. In: Liao, Tim Futing, editor. Sociological Methodology. Hoboken, NJ: Wiley-Blackwell; 2010. p. 285-327.

Goel, Sharad; Salganik, Matthew J. Respondent-driven sampling as Markov chain Monte Carlo. Statistics in Medicine. 28:2202–2229. 209. [PubMed: 19572381]

Goodman, Leo A. Snowball Sampling. Annals of Mathematical Statistics. 1961; 32:148–170.

Goodman, Leo A. On Respondent-Driven Sampling and Snowball Sampling in Hard-To-Reach Populations and Snowball Sampling Not in Hard-to-Reach Populations. In: Liao, Tim Futing, editor. Sociological Methodology. Vol. vol. 41. Hoboken, NJ: Wiley-Blackwell; Forthcoming

Heckathorn, Douglas D. Respondent-Driven Sampling: A New Approach to The Study of Hidden Populations. Social Problems. 1997; 44:174–199.

Heckathorn, Douglas D. Respondent-Driven Sampling II: Deriving Valid Population Estimates from Chain-Referral Samples of Hidden Populations. Social Problems. 2002; 49:11–34.

Heckathorn, Douglas D. Extensions of Respondent-Driven Sampling: Analyzing Continuous Variables and Controlling for Differential Recruitment. In: Xie, Yu, editor. Sociological Methodology. Vol. vol. 37. Boston, MA: Blackwell Publishing; 2007. p. 151-207.

Magnani, Robert; Sabin, Keith; Saidel, Tobi; Heckathorn, Douglas D. Review of Sampling Hard-to-Reach and Hidden Populations for HIV Surveillance. AIDS 2005. 2005; 19 Suppl2:S67–S72.

Malekinejad M, Johnston LG, Kendall C, Kerr LR, Rifkin MR, Rutherford GW. Using Respondent-Driven Sampling Methodology for HIV Biological and Behavioral Surveillance in International Settings: A Systematic Review. AIDS and Behavior. 2008; 12:105–130.

Ramirez-Valles, Jesus; Heckathorn, Douglas D.; Vázquez, Raquel; Diaz, Rafael M.; Campbell, Richard T. From Networks to Populations: The Development and Application of Respondent-Driven Sampling among IDUs and Latino Gay Men. AIDS and Behavior. 2005; 9(4):387–402. [PubMed: 16235135]

Rapoport, Anatol. A Probabilistic Approach to Networks. Social Networks. 1979; 2:1–18.

Salganik, Matthew J. Variance Estimation, Design Effects, and Sample Size Calculations for Respondent-Driven sampling. Journal of Urban Health. 2006; 83:i98–i112. [PubMed: 16937083]

Salganik, Matthew J.; Heckathorn, Douglas D. Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling. In: Stolzenberg, Ross M., editor. Sociological Methodology. Vol. vol 34. Boston, MA: Blackwell Publishing; 2004. p. 193-239.

Semaan, Salaam; Santibanez, Scott; Garfein, Richard S.; Heckathorn, Douglas D.; Des Jarlais, Don C. Ethical and Regulatory Considerations in HIV Prevention Studies Employing Respondent-Driven Sampling. International Journal of Drug Policy. 2009; 20(1):14–27. [PubMed: 18243679]

Spreen, Marinus. Rare Populations, Hidden Populations, and Link-Tracing Designs: What and Why? Bulletin de Méthodologie Sociologique. 1992; 36:34–58.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Sudman, Seymour; Kalton, Graham. New Developments in the Sampling of Special Populations. Annual Review of Sociology. 1986; 12:401–429.

Volz, Erik; Heckathorn, Douglas D. Probability-Based Estimation Theory for Respondent Driven Sampling. Journal of Official Statistics. 2008; 24:79–97.

Wejnert, Cyprian. An Empirical Test of Respondent-Driven Sampling: Point Estimates, Variance, Degree Measures, and Out-of-Equilibrium Data. In: Xie, Yu, editor. Sociological Methodology. Vol. vol. 39. Hoboken, NJ: Wiley-Blackwell; 2009. p. 73-116.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

TABLE 1

Evolution of RDS Population Estimators

RDS Estimator	Information Employed	Type of Estimator	Limitations	Variance Estimation	Distinctive Contribution
(E#1) Heckathorn 1997	Recruitment matrix	Markov equilibrium	Limited to nominal variables; no control for homophily differences	None	Shows that sample is self-weighting when homophily is uniform across groups
(E#2) Heckathorn 2002	Recruitment matrix, self-reported degrees	Reciprocity-model based estimator; linear least squares for nondyadic categories	Limited to nominal variables	Bootstrap	Controls for differences in degree and homophily across groups, with variance estimation
(E#3) Heckathorn 2002	Recruitment matrix, self-reported degrees	Reciprocity-based RDS estimator; data smoothing used for nondyadic categories	Limited to nominal variables	Bootstrap	Using data smoothing rather than linear least squares yields narrower confidence intervals
(E#4) Salganik and Heckathorn 2004	Recruitment matrix, self-reported degrees	Reciprocity-based RDS Estimator	Limited to nominal variables	Bootstrap	Proof that estimate is asymptotically unbiased, and better network estimation
(E#5) Volz and Heckathorn, 2008	Sample proportions, self-reported degrees	Network-based RDS estimator	Does not control for differential recruitment	Analytic variance estimation	New and analytically tractable estimator; permits analysis of continuous variables
(E#6) Volz and Heckathorn, 2008	Recruitment matrix, self-reported degrees	Network-based RDS estimator with data smoothing	Analyses limited to nominal variables	Analytic variance estimation	Demonstrates convergence between reciprocity and network-based estimators
(E#7) Heckathorn 2007	Recruitment matrix, self-reported degrees	Dual-component RDS estimator	Shares limitations inherent in RDS	Bootstrap	Permits analysis of continuous variables; controls for differential recruitment