Lee, Sunghee. 2009. "Understanding Respondent Driven Sampling from a Total Survey Error Perspective." *Survey Practice* 2 (6). https://doi.org/10.29115/SP-2009-0029.

ARTICLES

Understanding Respondent Driven Sampling from a Total Survey Error Perspective

Sunghee Lee[1]

[1] UCLA Center for Health Policy Research and Department of Biostatistics

Keywords: survey practice

https://doi.org/10.29115/SP-2009-0029

Survey Practice

Vol. 2, Issue 6, 2009

Understanding Respondent Driven Sampling from a Total Survey Error Perspective

Challenges of scientific data collection with rare and hidden populations are well documented and understood (e.g., Sudman, Sirken, and Cowan 1988). For instance, the Latino gay, bisexual and transgender (GBT) population studied in Ramirez-Valles et al. (2005) is not only ethnically but also sexually rare – these people may be hidden in the general population and, therefore, elusive. These hidden and elusive populations are often associated with negative health behaviors and outcomes, yet studying such subpopulations is challenging. From the traditional probability sampling perspectives, building a sampling frame for such populations requires a great deal of resources, which is often regarded as ideal but impractical. To locate potential subjects of the hidden population, one will need to screen the sample drawn from the general population in order to identify and locate enough samples. The rarer the population is, the higher the cost of screening.

In order to approach this sampling issue, several methods utilizing the social networks of those populations were suggested recently as an alternative to the probability sampling methods. Among them is Respondent Driven Sampling (RDS), first introduced by Heckathorn (1997, 2002). RDS stems from the fact that, although hidden in the general population from outsiders' viewpoint, some hidden population units are often linked to other units of the same population, forming some types of networks. For instance, chances of a GBT person knowing another GBT person are much higher than a member of the general population knowing any GBT persons. Once researchers establish contacts with a hidden population unit (or node), they can exploit the unit's network links (or edges) to other units for sampling hidden populations. The first set of the hidden population unit may come from some type of convenient samples in RDS, such as a snowball sample. After collecting data with these first-wave participants (*seeds*), researchers ask the participants play a role of recruiters; they recruit those who quality for the study (*alters*) from their individual networks as the second wave of sample. After the second wave of data collection, this new set of participants recruits the next wave of participants. Recruitment waves continue until the desired sample size is achieved. There are monetary compensations for participating in the data

collection and recruitment. This recruitment is claimed to following the Markov process, achieving stationary probabilities where the final sample obtained from RDS becomes unbiased of the hidden population of interest.

By eliminating the screening procedure, RDS is substantially less expensive than traditional probability sampling methods. The economic nature of RDS has attracted much attention from the research community. In particular, the field of epidemiology and urban health appears to embrace this method as a valid tool to obtain representative sample data, as evidenced by a recent volume of *Journal of Urban Health* (Vol. 83, Supp. 1, 2006) dedicated to RDS. Other journals such as *American Journal of Public Health*, *Addictive Behaviors*, *AIDS and Behavior*, and *Journal of Acquired Immune Deficiency Syndromes* include studies using RDS.

The main issue perceived by survey researchers is that RDS is ultimately a nonprobability sampling method. When using a probability sample, the unbiasedness of sample estimates is mathematically guaranteed without requiring assumptions or utilizing models in estimation (e.g., Cochran 1977; Kish 1965). On the other hand, the sample selection is subjective under nonprobability sampling, which does not guarantee the elimination of selection biases.[1] Another issue is, unlike traditional probability sampling or adaptive sampling approaches (Thompson and Collins 2002), where the sample selection is controlled by researchers, RDS lets study participants control the selection process entirely, which makes the statistical inference difficult (Frost et al. 2006).

The best way to evaluate RDS is to compare its performance with that of traditional samples (Frost et al. 2006). Studies concluding RDS as valid and effective often do not include such verifications or use deficient methods for comparisons. A comparison between a probability-based stratified sample and an RDS sample by Martin, Wiley, and Osmond (2003) showed large discrepancies in estimates between two samples. When compared to the population data, an RDS sample also revealed critical issues (Wejnert and Heckathorn 2008). To date, there has been only one article published in *Journal of Official Statistics* read by survey researchers (Volz and Heckathorn 2008). Given the inconclusive findings and the lack of publication in the survey research literature, it is not surprising to find unfamiliarity with RDS among survey researchers who rely on traditional probability-based samples, despite its claimed attractive properties.

1 Semaan, Lauby, and Liebman (2002) added the 'representative' sample to the sampling distinction category. The representativeness that Semaan et al. argue, however, is still left up to the subjective judgment and is not the same as the unbiasedness of probability sampling as shown in the 1948 U.S. presidential election polls. In essence, the 'representative' sample category does not add much as it belongs to the non-probability sample.

Claimed unbiasedness of RDS is dependent on a set of explicit and implicit assumptions. For a better understanding of RDS, we will use the structure of total survey error (TSE), a fundamental framework in survey research, and examine the RDS assumptions. TSE acknowledges that survey data are subject to some amount of error from four sources: coverage, sampling, nonresponse and measurement error (Deming 1994; Groves 1989). As RDS is a sampling method, the unbiasedness in its data is only respect to sampling error. Additionally, the unique RDS sample recruitment process poses ample room for other errors. Below, we examine respective RDS assumptions and discuss how they are related to TSE. Because RDS is based on nonprobability sampling and TSE probability sampling, TSE components may not map onto RDS assumptions clearly.

Network structure

The hidden population network forms only one component. That is, there is a path between every person and every other person in the network of the hidden population. In order to meet this assumption, the network must be dense and consist of a single component.

Although the frame is not used in RDS, the social network determines the overall structure of the frame, because the RDS selection process itself exploits the network. Therefore, RDS coverage error is directly related to network structure assumptions. Networks with multiple components or with loosely linked units are likely to result in coverage error.

Since the composition of RDS frame is heavily dependent upon the performance of the recruiters, inaccuracy of recruiters' knowledge about the target population and ability to translate that into recruitment tasks is a key element for measurement error that can lead to selection bias. As shown in Lee, Mathiowetz, and Tourangeau (2007) what appears to be as simple as having a disability could be interpreted very differently from one person to another. Studies like Heckathorn and Jeffri (2003) leaving the interpretation of its target population (e.g., jazz musicians) up to the study participants may cause concerns. Therefore, whether researchers can ensure recruiters' understanding about the hidden population is critical.

Complete response

RDS assumes 100% response rates in two stages: 1) all sampled/recruited subjects participate, and 2) all study participants recruit their alters.

In the literature, studies using RDS often do not report their survey-level response rates, which makes the evaluation of this assumption difficult. At the recruitment level, Wejnert and Heckathorn (2008) and Martin, Wiley, and Osmond (2003) similarly report that close to 50% of the respondents did not recruit their alters. With the monetary incentives used to encourage participation, this is a surprisingly low figure.

A violation on the complete response assumption in RDS is the same as nonresponse in probability samples, which allows for nonresponse error. However, the nature of this assumption differs between RDS and probability sampling. First, in probability sampling, there is no alter recruitment, eliminating the second stage of nonresponse stated above. As response rates should consider all stages of data collection cumulatively, RDS response rates must be a multiplicative of the response status not only to the survey itself but to the recruitment, implying potentially very low overall response rates. Second, while statistical adjustments, such as weighting, are applied to alleviate nonresponse error in traditional probability sampling, RDS estimation does not include such adjustments (e.g., Volz and Heckathorn 2008).

Random recruitment

Recruiters are assumed to not use any judgment in selecting their alters, resulting in random recruitment. Any given unit within the network of a recruiter selected at the k^{th} wave has the equal selection probability into the $(k+1)^{th}$ wave.

The RDS sampling process carried out by recruiters introduces room for selection biases, because the selection processes are out of researchers' control. As hypothesized in Martin, Wiley, and Osmond (2003), a recruiter may use their own judgment and select those who may benefit from the study at a higher rate. Violations of random recruitment are closely related to sampling, coverage and measurement errors.

Equilibrium condition

As recruitment waves continue, the characteristics of recruited alters become independent of the seeds' characteristics, approaching to equilibrium at a geometric rate. This assumption is directly related to the Markov process where the future and past states become independent given the present state.

Equal Homophily

Homophily is the tendency of individuals in the hidden population to associate with those with similar traits. In RDS, the propensity of a member of one group recruiting someone from the same group is assumed the same as that of a member of another group.

Inadequate recruiter performance in sample selection process affects equilibrium and equal homophily, introducing coverage, measurement and sampling errors. Often RDS employs usage of monetary incentives through distributing coupons, which may attract certain groups more than others, violating equilibrium and equal homophily assumptions and resulting in measurement and nonresponse errors that may distort the sample selection mechanism.

When these assumptions are violated, bringing in systematic errors, the RDS sample-based estimates may be biased. It is, therefore, evident that the RDS assumptions resemble the TSE components. Evaluating RDS from traditional survey researchers' perspectives using TSE will not only provide crystallized understandings about RDS but also generate productive dialogues between survey researchers using probability samples and researchers studying hidden populations.

REFERENCES

Cochran, W.G. 1977. *Sampling Techniques*. 3rd ed. New York: Wiley.

Deming, W.E. 1994. "On Errors in Surveys." *American Sociological Review* 9 (4): 359–69.

Frost, S.D.W., K.C. Brouwer, M.A.F. Cruz, R. Ramos, M.E. Ramos, R.M. Lozada, C. Magis-Rodriguez, and S.A. Strathdee. 2006. "Respondent-Driven Sampling of Injection Drug Users in Two U.S.-Mexico Border Cities: Recruitment Dynamics and Impact on Estimates of HIV and Syphilis Prevalence." *Journal of Urban Health* 83 (Suppl 1): 83–97.

Groves, R.M. 1989. *Survey Errors and Survey Costs*. New York: Wiley.

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

———. 2002. "Respondent-Driven Sampling II: Deriving Valid Population Estimates from Chain-Referral Samples of Hidden Populations." *Social Problems* 49: 11–34.

Heckathorn, D.D., and J. Jeffri. 2003. "Social Network of Jazz Musicians." In *Changing the Beat: A Study of Worklife of Jazz Musicians, Vol III: Respondent-Driven Sampling: Survey Results by the Research Center for Arts and Culture, National Endowment For Arts Research Division Report #43*, 43:48–61. Washington, D.C.

Kish, L. 1965. *Survey Sampling*. New York: John Wiley & Sons.

Lee, S., N. Mathiowetz, and R. Tourangeau. 2007. "The Measurement and Mismeasurement of Disability." *Journal of Official Statistics* 23 (2): 163–84.

Martin, J.L., J. Wiley, and D. Osmond. 2003. "Social Networks and Unobserved Heterogeneity in Risk for AIDS." *Population Research and Policy Review* 22: 65–90.

Ramirez-Valles, J., D.D. Heckathorn, R. Vázquez, R. Diaz, 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): 403–8.

Semaan, S., J. Lauby, and J. Liebman. 2002. "Street and Network Sampling in Evaluation Studies of HIV Risk-Reduction Interventions." *AIDS Review* 4: 213–23.

Sudman, S., M.G. Sirken, and C.D. Cowan. 1988. "Sampling Rare and Elusive Populations." *Science* 240: 991–96.

Thompson, S.K., and L.A. Collins. 2002. "Adaptive Sampling in Research on Risk-Related Behaviors." *Drug and Alcohol Dependence* 68 (Suppl 1): 57–67.

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

Wejnert, C., and D.D. Heckathorn. 2008. "Web-Based Network Sampling: Efficiency and Efficacy of Respondent-Driven Sampling for Online Research." *Sociological Methods & Research* 37 (1): 105–34.