SURVEYS · SAMPLING

Respondent-Driven Sampling and Network-Based Recruitment

Respondent-driven sampling recruits members of hidden populations through their social networks and uses information about those networks to estimate population characteristics. This article explains how it works, what it assumes and how far its results can be trusted.

ModQu research article · 1,061 words · 12 references (APA 7th edition)

Some populations cannot be sampled from a list. People who inject drugs, sex workers, undocumented migrants and members of informal occupational groups are rarely listed anywhere, and many have reasons to avoid being identified. Yet their members often know each other. **Network-based recruitment** uses these social ties to find respondents. The simplest form, snowball sampling, asks participants to name or recruit others (Goodman, 1961; Biernacki & Waldorf, 1981). **Respondent-driven sampling** (RDS) adds structure and statistical theory, with the aim of producing estimates for the hidden population rather than only a collection of cases (Heckathorn, 1997).

How RDS works

A study begins with a small number of **seeds**, members of the target population chosen by the researchers. Each seed completes the survey and receives a small number of coupons, usually three, to pass to other members of the population they know. People who bring a coupon to the study site, or use it online, become participants, complete the survey and receive coupons of their own. Recruitment continues in waves, as shown in Figure 1 (Heckathorn, 1997).

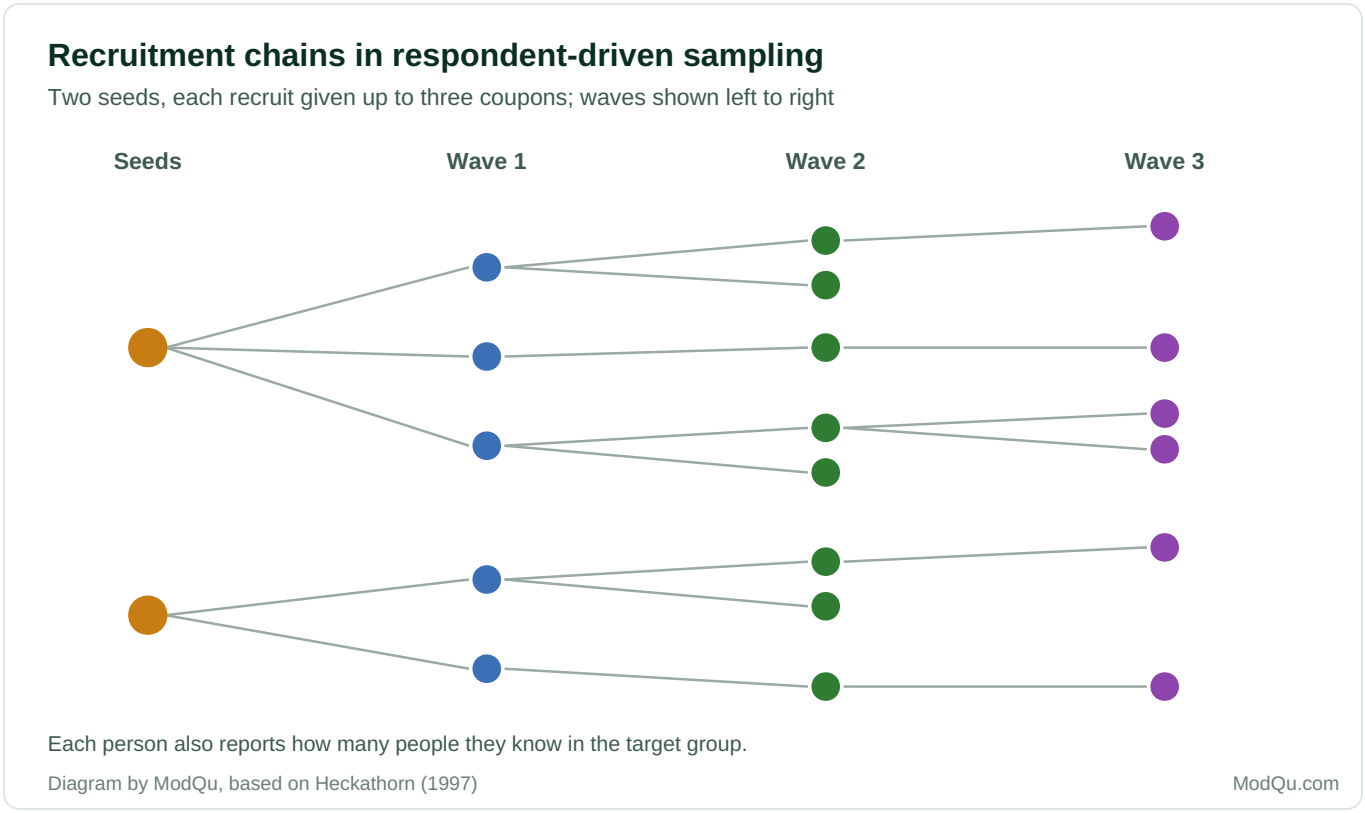


Figure 1. Recruitment chains in respondent-driven sampling. Diagram by ModQu, based on Heckathorn (1997).

Two features distinguish RDS from snowball sampling. First, participants recruit others themselves, rather than giving names to researchers, which protects privacy and lets the recruitment reach people researchers could not contact. Participants are usually paid both for taking part and for each successful recruitment. Second, the study records who recruited whom and asks each participant how many members of the target population they know, their **network size** (Heckathorn, 2002).

How estimates are made

People with larger networks are more likely to be recruited, because more people can pass them a coupon. RDS estimators correct for this by giving each respondent a weight that is inversely related to their reported network size (Volz & Heckathorn, 2008). The theory also suggests that, as recruitment continues over many waves, the composition of the sample becomes independent of the seeds, much as a random walk through a network eventually forgets where it started (Salganik & Heckathorn, 2004).

The assumptions

These estimates rely on several assumptions. The target population must form a connected network. Ties must be reciprocal, so that if one person can recruit another, the reverse is also possible. Participants must recruit at random from among their contacts, and must report their network sizes accurately. Sampling is assumed to be with replacement, and recruitment chains long enough to lose dependence on the seeds (Gile & Handcock, 2010).

In practice, these assumptions are often doubtful. People tend to recruit friends who are similar to them, report network sizes imprecisely and pass coupons to those they see most often. Recruitment chains may stop after a few waves.

How well does it work?

RDS has been widely used, particularly in public health surveillance of populations at risk of HIV in many countries (Malekinejad et al., 2008). Its statistical performance has been questioned. Simulations based on real networks found that RDS estimates could be much more variable than standard calculations suggest, with design effects far larger than those of conventional surveys (Goel & Salganik, 2010). Assessments of the methodology identified sensitivity to the choice of seeds and to violations of the assumptions, and new estimators have been proposed to reduce these problems (Gile & Handcock, 2010; Gile, 2011).

The conclusion of most methodologists is that RDS is a valuable way to reach hidden populations and to collect data that would otherwise not exist, but that its estimates should be treated with caution and reported with appropriately wide uncertainty.

An example

A city health department wants to estimate how many people who inject drugs have been tested for hepatitis C in the past year. No list of this population exists, and few would identify themselves in a household survey. The department recruits eight seeds through outreach workers, choosing people of different ages, sexes and neighbourhoods. Each participant receives three coupons and a small payment for each person they recruit who comes to the study office. Over ten weeks, recruitment runs to six or seven waves in most chains and produces 450 participants. Analysts check that the share tested for hepatitis C settles to a stable value in later waves, apply an estimator that weights by reported network size and report an estimate of 55 per cent with a wide confidence interval, noting that people who rarely use services may be under-represented if they are less often recruited (illustrative example).

Online and digital variants

RDS was designed for face-to-face fieldwork, with coupons handed over in person. Online versions now use unique links or codes that participants share by message or social media. These can reach dispersed populations quickly and reduce costs. They also bring new challenges: links can be posted publicly, which breaks the chain structure; it is harder to confirm that a recruit belongs to the target population; and people without internet access are missed. Studies using online RDS should limit the use of each code, verify eligibility and examine whether recruitment patterns differ from those expected in the population.

Good practice

- Choose diverse seeds, covering different subgroups and locations.
- Limit the number of coupons per person so that chains grow long rather than wide.
- Record recruitment links and ask carefully worded questions about network size.
- Check whether the sample composition stabilises across waves.
- Use several estimators and compare results.
- Report the study following agreed guidelines, such as the STROBE-RDS statement, which lists the information readers need to judge an RDS study (White et al., 2015).

Network recruitment beyond RDS

Network-based recruitment is also used informally in business and organisational research, for example to find specialists in a niche industry by asking each interviewee to suggest others. Such snowball samples are useful for reaching experts and gathering insight, but they are not designed for estimation, and their results reflect the networks of the first contacts (Biernacki & Waldorf, 1981). When the goal is to describe a rare group with known precision, probability methods such as screening should be considered first (Tourangeau et al., 2014).

Conclusion

Respondent-driven sampling turns social networks into a recruitment tool and offers a statistical framework for estimating characteristics of populations that cannot otherwise be sampled. Its assumptions are demanding and often only partly met, so its estimates carry more uncertainty than they appear to. Used carefully and reported transparently, it remains one of the few ways to produce systematic knowledge about hidden groups.

References

Biernacki, P., & Waldorf, D. (1981). Snowball sampling: Problems and techniques of chain referral sampling. *Sociological Methods & Research, 10*(2), 141–163. https://doi.org/10.1177/004912418101000205

Gile, K. J., & Handcock, M. S. (2010). Respondent-driven sampling: An assessment of current methodology. *Sociological Methodology, 40*(1), 285–327. https://doi.org/10.1111/j.1467-9531.2010.01223.x

Gile, K. J. (2011). Improved inference for respondent-driven sampling data with application to HIV prevalence estimation. *Journal of the American Statistical Association, 106*(493), 135–146. https://doi.org/10.1198/jasa.2011.ap09475

Goel, S., & Salganik, M. J. (2010). Assessing respondent-driven sampling. *Proceedings of the National Academy of Sciences, 107*(15), 6743–6747. https://doi.org/10.1073/pnas.1000261107

Goodman, L. A. (1961). Snowball sampling. *Annals of Mathematical Statistics, 32*(1), 148–170. https://doi.org/10.1214/aoms/1177705148

Heckathorn, D. D. (1997). Respondent-driven sampling: A new approach to the study of hidden populations. *Social Problems, 44*(2), 174–199. https://doi.org/10.2307/3096941

Heckathorn, D. D. (2002). Respondent-driven sampling II: Deriving valid population estimates from chain-referral samples of hidden populations. *Social Problems, 49*(1), 11–34. https://doi.org/10.1525/sp.2002.49.1.11

Malekinejad, M., Johnston, L. G., Kendall, C., Kerr, L. R. F. S., Rifkin, M. R., & Rutherford, G. W. (2008). Using respondent-driven sampling methodology for HIV biological and behavioral surveillance in international settings: A systematic review. *AIDS and Behavior, 12*(Suppl. 1), 105–130. https://doi.org/10.1007/s10461-008-9421-1

Salganik, M. J., & Heckathorn, D. D. (2004). Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological Methodology, 34*(1), 193–240. https://doi.org/10.1111/j.0081-1750.2004.00152.x

Tourangeau, R., Edwards, B., Johnson, T. P., Wolter, K. M., & Bates, N. (Eds.). (2014). *Hard-to-survey populations*. Cambridge University Press. https://doi.org/10.1017/CBO9781107628771

Volz, E., & Heckathorn, D. D. (2008). Probability based estimation theory for respondent driven sampling. *Journal of Official Statistics, 24*(1), 79–97.

White, R. G., Hakim, A. J., Salganik, M. J., Spiller, M. W., Johnston, L. G., Kerr, L., Kendall, C., Drake, A., Wilson, D., Orroth, K., Egger, M., & Hladik, W. (2015). Strengthening the reporting of observational studies in epidemiology for respondent-driven sampling studies: "STROBE-RDS" statement. *Journal of Clinical Epidemiology, 68*(12), 1463–1471. https://doi.org/10.1016/j.jclinepi.2015.04.002