SURVEYS · SAMPLING

Sampling Rare and Hard-to-Reach Populations

Some groups are too small, too scattered or too hidden to be found efficiently with ordinary samples. This article explains why rare and hard-to-reach populations are difficult to survey and describes the main methods used to sample them.

ModQu research article · 1,182 words · 12 references (APA 7th edition)

Many important questions concern small groups: recent immigrants, people with a rare disease, owners of a particular type of business, people experiencing homelessness or people who use drugs. Surveying them is difficult. Some groups are simply **rare**, making up a small share of the population, so a general sample contains few of them. Others are **hard to reach** or **hidden**: no list of them exists, they are mobile, or they have reasons not to identify themselves (Sudman et al., 1988; Tourangeau et al., 2014). This article describes the main methods for sampling such groups, summarised in Figure 1.

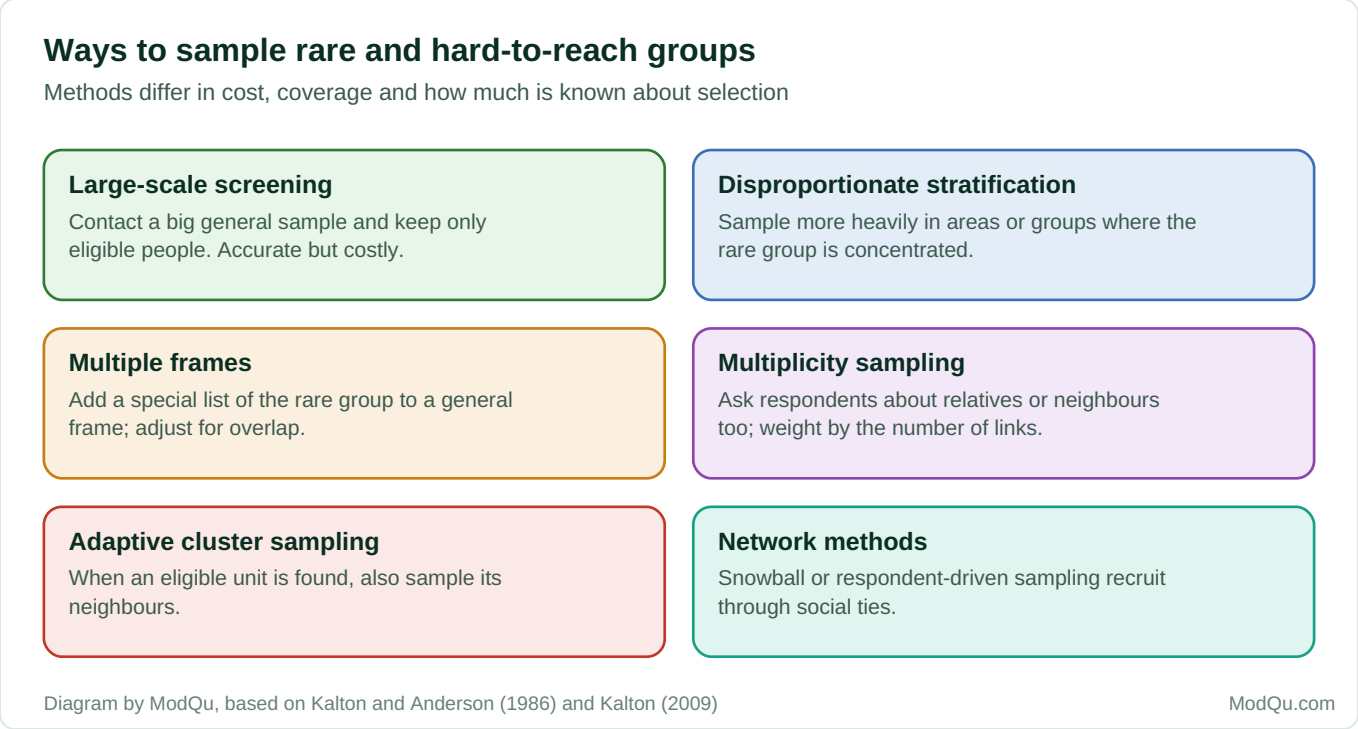


Figure 1. Methods for sampling rare and hard-to-reach groups. Diagram by ModQu, based on Kalton and Anderson (1986) and Kalton (2009).

Why these groups are hard to survey

Tourangeau et al. (2014) distinguish several reasons why a group may be hard to survey. Some groups are **hard to sample** because they are rare or no frame lists them. Some are **hard to identify** because membership is sensitive or stigmatised, so people may not reveal it in a screening interview. Some are **hard to find** because they move often or have no fixed address. Some are **hard to persuade**, because they distrust surveys or the organisations behind them. And some are **hard to interview** because of language, disability or other barriers. A single group can face several of these problems at once, and each calls for a different response: better frames for the first, careful wording and trust-building for the second and fourth, flexible fieldwork for the third, and translated or adapted questionnaires for the last.

These distinctions help researchers diagnose where their design is likely to fail. A survey of recent migrants may have an adequate frame but still miss people who do not speak the main language. A survey of people with a stigmatised health condition may find them through a patient list but still face reluctance to take part.

Screening

The most straightforward method is to draw a large general sample and use short **screening** questions to identify eligible members. Screening keeps the strengths of a probability sample, but its cost rises steeply as the group gets rarer. If 1 per cent of households are eligible, about 10,000 households must be screened to find 100. Screening by telephone or mail is cheaper than in person, and screening questions must be designed so that eligible people are not missed or discouraged from identifying themselves (Kalton & Anderson, 1986).

Disproportionate stratification

When the rare group is concentrated in some areas or parts of a frame, those parts can be sampled more heavily. For example, if a minority group lives mainly in certain neighbourhoods, sampling those neighbourhoods at a higher rate finds more members per screening contact. The gain depends on how concentrated the group is; if many members live elsewhere, sampling the low-density areas lightly still leaves them represented, but with larger weights (Kalton, 2009).

Multiple frames

Sometimes a partial list of the rare group exists, such as a membership register or a patient registry. Combining it with a general frame improves efficiency while the general frame covers people who are not on the list. People who appear on both frames have two chances of selection, and estimators for dual frame surveys adjust for this (Lohr & Rao, 2000).

Multiplicity sampling

In **multiplicity** or network sampling, respondents in a general sample report not only on themselves but on defined relatives or neighbours, for example their siblings. This finds more members of a rare group per interview. Each reported person is weighted by the number of respondents who could have reported them (Sirken, 1970).

Adaptive cluster sampling

When rare units are clustered, **adaptive cluster sampling** begins with an initial sample and, whenever an eligible unit is found, adds neighbouring units to the sample. This concentrates effort where the group is found, and estimators exist that keep the design unbiased (Thompson, 1990).

Network sampling: snowball and respondent-driven sampling

For hidden groups with no list and no geographic concentration, researchers often rely on social networks. In **snowball sampling**, initial participants recruit others they know (Goodman, 1961). It reaches people that other methods cannot, but the sample reflects the social ties of the starting points.

Respondent-driven sampling adds structure. Each participant receives a limited number of coupons to recruit others, recruitment chains continue for several waves, and participants report the size of their personal networks. Under certain assumptions, these data allow estimates that correct for the higher chance of selection of well-connected people (Heckathorn, 1997; Salganik & Heckathorn, 2004). Evaluations have shown that the assumptions are often hard to meet, and that estimates can have large uncertainty, so results should be interpreted with care (Gile & Handcock, 2010).

Online recruitment

Online advertising and targeted invitations offer a fast way to reach some rare groups, especially young people and those who share online communities. Studies have recruited hard-to-reach groups through social media quickly and at low cost (Ramo & Prochaska, 2012). These are nonprobability samples, however, and the people who respond to online invitations may differ from other members of the group.

Choosing a method

The choice depends on how rare the group is, whether a list or concentration exists, how sensitive membership is, the budget and whether population estimates or in-depth understanding are needed. Probability-based methods, such as screening, disproportionate stratification and multiple frames, are preferred when estimates of prevalence or representative descriptions are required. Network and online methods are often the only practical options for hidden groups, and they are valuable for understanding experiences, but their estimates rest on stronger assumptions (Kalton, 2009; Tourangeau et al., 2014). Combining methods, such as a probability screening sample supplemented by network recruitment, can balance cost and quality. Whatever the method, the report should state how members of the group were found, how many were screened to find them and which parts of the group the design is likely to miss, so that readers can judge how far the findings extend.

An example

A city wants to survey people who provide unpaid care to a family member with dementia, estimated at about 2 per cent of adults. Screening a general sample of 20,000 households by mail would find around 400 carers at substantial cost. The city adds a list of carers registered with support services as a second frame, recognising that registered carers are likely to be more connected to services, and adjusts for people on both frames. It also asks respondents in the general sample whether they know other carers, but uses this only to describe networks, not for weighting. The resulting sample supports estimates of the number of carers and comparisons between registered and unregistered carers (illustrative example).

Conclusion

Rare and hard-to-reach groups require special designs. Screening, disproportionate stratification, multiple frames, multiplicity and adaptive methods extend probability sampling to small groups at increasing cost. Network and online methods reach hidden groups that other approaches miss, at the price of stronger assumptions. Being clear about which method was used, and what it can support, is essential for producing reliable knowledge about the groups that are hardest to study (Sudman et al., 1988; Kalton & Anderson, 1986).

References

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

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

Kalton, G., & Anderson, D. W. (1986). Sampling rare populations. *Journal of the Royal Statistical Society: Series A (General), 149*(1), 65–82. https://doi.org/10.2307/2981886

Kalton, G. (2009). Methods for oversampling rare subpopulations in social surveys. *Survey Methodology, 35*(2), 125–141.

Lohr, S. L., & Rao, J. N. K. (2000). Inference from dual frame surveys. *Journal of the American Statistical Association, 95*(449), 271–280. https://doi.org/10.1080/01621459.2000.10473920

Ramo, D. E., & Prochaska, J. J. (2012). Broad reach and targeted recruitment using Facebook for an online survey of young adult substance use. *Journal of Medical Internet Research, 14*(1), e28. https://doi.org/10.2196/jmir.1878

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

Sirken, M. G. (1970). Household surveys with multiplicity. *Journal of the American Statistical Association*, *65*(329), 257–266. https://doi.org/10.1080/01621459.1970.10481077

Sudman, S., Sirken, M. G., & Cowan, C. D. (1988). Sampling rare and elusive populations. *Science*, *240*(4855), 991–996. https://doi.org/10.1126/science.240.4855.991

Thompson, S. K. (1990). Adaptive cluster sampling. *Journal of the American Statistical Association*, *85*(412), 1050–1059. https://doi.org/10.1080/01621459.1990.10474975

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