A Computational Analysis of Snowball Sampling for the Estimation of Means

João Gabriel Malaguti

Alinne de Carvalho Veiga

Letícia de Carvalho Giannella

Snowball sampling is a non-probabilistic sampling method, widely used in the social sciences both for lacking the requirement of a frame, unlike many probabilistic methods, and for its ability in reaching hard-to-reach populations (NOY, 2008; HECKATHORN, 2007; GOODMAN, 1961), such as drug users (FRANK & SNIJDERS, 1994), the homeless (CLATTS, DAVIS & ATILLASOY, 1995) and queer people (BUENTING, 1992).

Formally, we can define snowball sampling thus: an initial sample is selected from a finite population. A referral request for other people belonging to the population of interest is made to all the individuals in the sample. All of the referrals that are not already part of the sample form the first wave (or first stage). For every individual of the first wave, the request is repeated and so on, until the stop condition set by the researcher is met: this can be the number of waves, the sample size or something qualitative, such as when responses present little variation. The method gets its name because, similar to a snowball rolling down a hill, the sample also increases in size over time (THOMPSON, 2014; GOODMAN, 1961).

Due to it being a non-probabilistic method, formal equations for the standard error of the mean do not exist, making analyses more complex. While there is a class of estimation methods (called respondent-driven sampling estimators or RDS estimators) that claim to be able to estimate the standard errors (HECKATHORN, 2011), these methods rely on a number of very heavy assumptions that are hardly met in practice (WEJNERT, 2009).

This study bypasses these issues by making use of computational statistics, in particular Monte Carlo simulations, which allows for approximate estimations based on different controlled scenarios in order to improve the understanding of the sampling method. Monte Carlo simulations are methods based on generating different independent samples to obtain approximate answers to stochastic problems, with their precision being related to the number of samples generated (iterations), such that the higher is the number of samples, the more precise the estimates are (BRANDIMARTE, 2014).

For this, random graphs (also called Erdős-Rényi graphs or Gilbert graphs) were used to propose/simulate populations. Graphs are generic mathematical models used to represent relationships (edges) between elements (nodes) of various types, such as flights between airports, sales and purchases across different enterprises, and, more famously, social networks (BARABÁSI, 2016).

Gilbert graphs are defined by the amount of nodes *N* and the probability of connection *p* between them, with the number of connections of the graph following a *Binomial(N, p)*. By manipulating the values for *N* and *p*, we can generate populations of different sizes and connection densities to better model real networks. However, real networks generally have two qualities: (i) they are sparse, that is, have an amount of connections much smaller than the maximum amount possible; and (ii) they have hubs, nodes with many more connections than the average, such as for example, a popular individual. While Gilbert graphs may fulfil the first condition, depending on the value of *p*, it cannot generate hubs (BARABÁSI, 2016).

In order to better mimic real applications of snowball sampling, in which individuals do not indicate all people they know belong to the population of interest, we defined a probability of indication ($\boldsymbol{pi}$) that varies from person to person. Conceptually, this probability of indication is influenced by multiple factors, some inherent, like personality, and others relational, like the level of trust with the field agents. This mathematical abstraction is unmeasurable in reality but allows for more sophisticated studies.

Alongside populations created by using random graphs, a variable of interest $y \sim Exponential\ (\lambda = 1/13)$ was appended to these, meant to mimic a generic income/expense variable, which are also non-negative and asymmetric. This variable has no association with the graphs' configurations.

Using R (R Core Team, 2023), along with the *igraph* (CSARDI et al., 2006), *doParallel*, *parallel* and *foreach* (WESTON et al., 2020) libraries, we outlined two studies as well as a convergence check for the estimates. Both studies use the same *y* variable, conjugated with different graphs, as well as the same desired sample size ($n = 650$), number of nodes ($N = 10000$) and number of iterations ($A = 10000$).

Each iteration functions similarly, on either study: (i) we take snowball samples of every population defined for the study; (ii) we calculate the mean and standard error (assuming simple random sampling without replacement) and store the values on an array; and (iii) completed all iterations, we calculate the Monte Carlo estimates using the previously mentioned array.

The first study explores the effect of the connection density, manipulating the probability of connection while maintaining the same number of nodes, creating a set of population graphs with probability of indication

$pi \sim Beta(9,3)$ and probability of connection p = { 0.001, 0.012, 0.023, 0.034, 0.045, 0.056, 0.067, 0.078, 0.089, 0.1}.

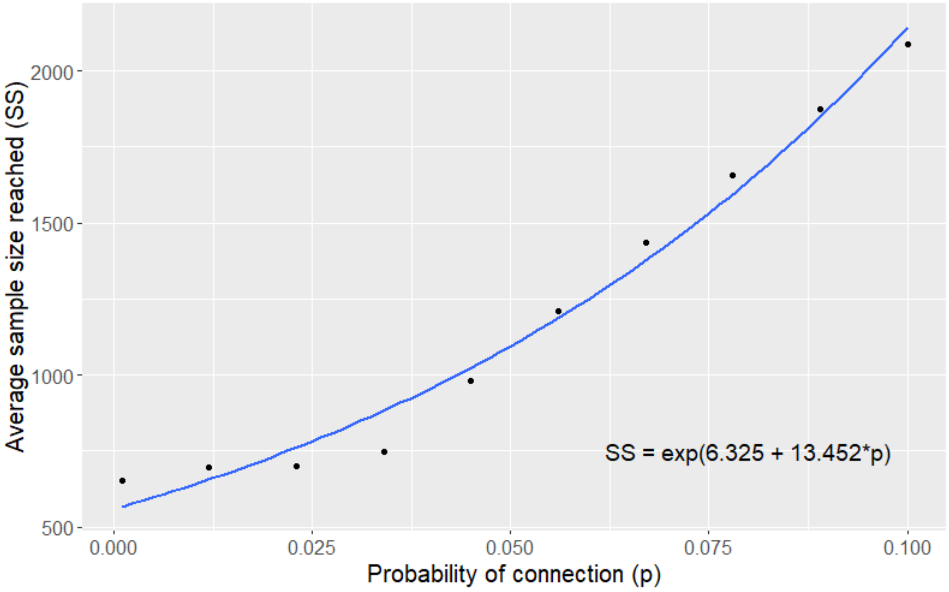
The second study focuses on the probability of indication, using the values outlined in Table 1 and a graph population with 10000 nodes and probability of connection p=0.01.

Table 1 – Parameters used for the second simulation study

Average (pi)	α	β
0.1	1/3	3
0.2	3/4	
0.3	9/7	
0.4	2	
0.5	3	
0.6	4.5	
0.7	7	
0.8	12	
0.9	27	

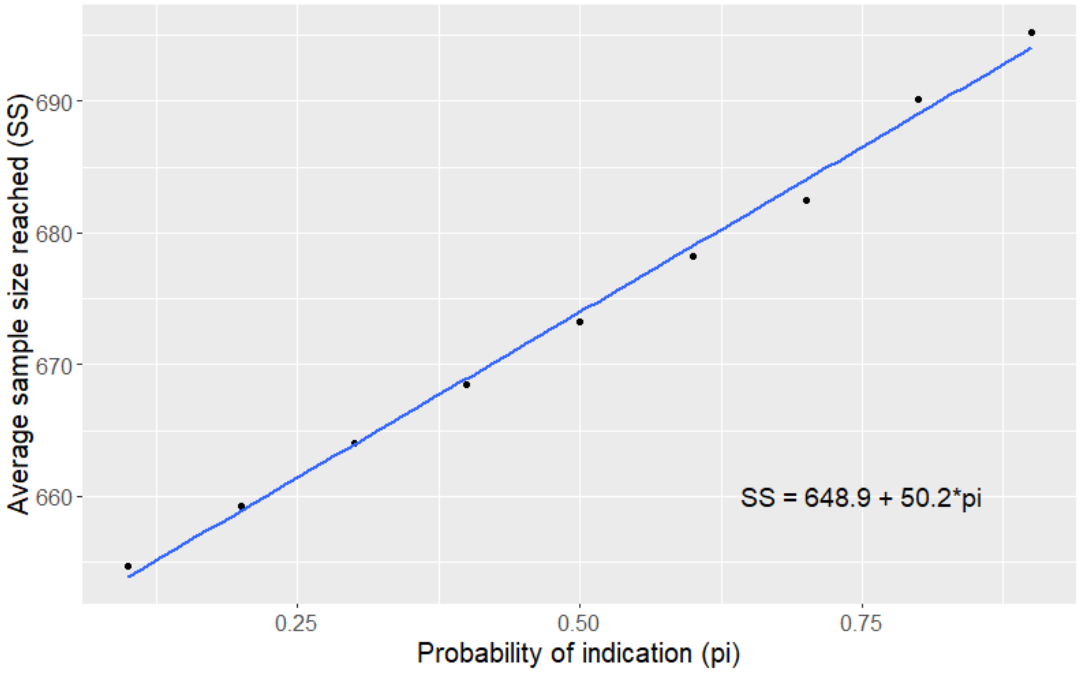
We found that, for this scenario, the connection density affects the average sample size exponentially, with the average sample size reached being modelled by the equation found in Figure 1. This behaviour indicates that the more connected a population is, the better will be the capacity of reaching the desired sample size. However, the sample size is limited by the amount of connections, which is something that the researchers may not be able to alter in the short-term.

Figure 1 –Average sample size reached (*SS*) by probability of connection (*p*)



The effect of the probability of indication on the average sample size, meanwhile, behaves linearly instead of exponentially (Figure 2). Though less influential in the sample size reached, the probability of indication can be altered in the short-term by for example, offering incentives for indicating and guaranteeing anonymity.

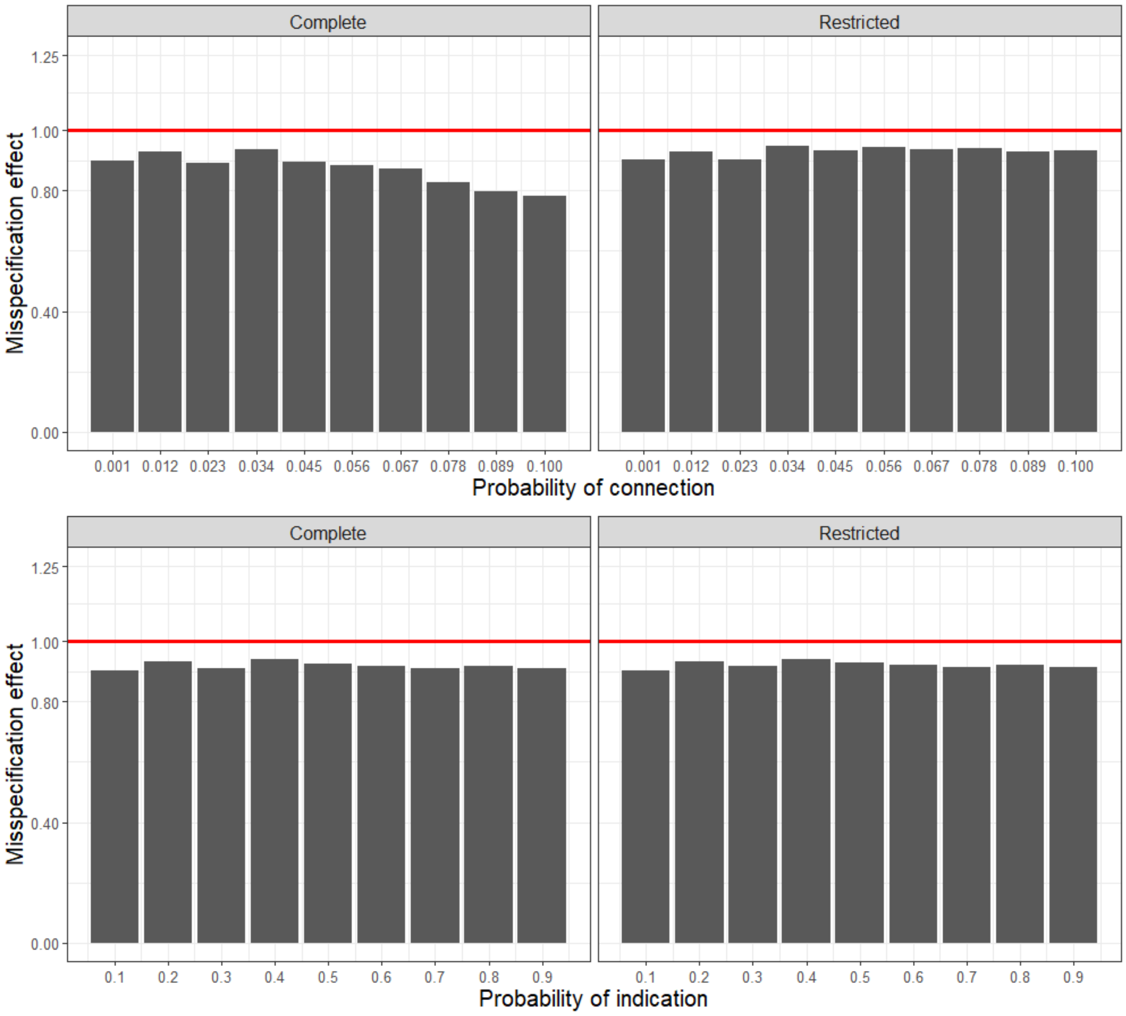
Figure 2 –Average sample size reached (*SS*) by probability of indication (*pi*)



Using the estimates for standard error of the mean (both assuming simple random sampling and Monte Carlo estimation), we also calculated Skinner's misspecification effect (*meff*), a measure which allows us to quantify the error committed when assuming simple random sampling for the estimation. We separated the cases into two groups with one using whichever sample size was reached, labelled "complete", and another using at most the desired sample size chosen previously (650 individuals), which was labelled "restricted", to allow for comparisons due to the difference between sample sizes.

For both studies, all values of *meff* were below 1 (Figure 3), meaning that the bias committed was negative. It also means that the formula for the standard error of the mean assuming simple random sampling consistently overestimates the standard error of the mean and therefore could be taken as an upper bound for the standard error under snowball sampling for this particular scenario.

Figure 3 –Misspecification effects by probability of connection (*p*) and probability of indication (*pi*) for different sample types



This is only a pilot study focusing on estimating the mean of a continuous variable that is unrelated to how a social network organizes itself, built upon graphs that, while sparse, do not present hubs. Therefore, these results are not generalizable into common usage and further studies are required.

Some studies are underway considering other aspects not contemplated by this pilot, such as the effect of altering the size of the initial sample, the inclusion of a probability of response (including a scenario of independence and a scenario of association with the probability of indication) and estimating proportions. We also seek to study scenarios where the variable of interest is related to how the network distributes itself and usage of more complex graphs that better represent real networks, like Barabási-Albert graphs.

Bibliography

BARABÁSI, A.-L. Network Science. Cambridge: Cambridge University Press, 2016.

BRANDIMARTE, P. Handbook in monte carlo simulation: applications in financial engineering, risk management, and economics. John Wiley & Sons, Hoboken, New Jersey, 2014.

BUENTING, J. A. Health life-styles of lesbian and heterosexual women. Health Care for Women International, Taylor & Francis, v. 13, n. 2, p. 165–171, 1992.

CLATTS, M. C.; DAVIS, W. R.; ATILLASOY, A. Hitting a moving target: The use of ethnographic methods in the development of sampling strategies for the evaluation of aids outreach programs for homeless youth in new york city. NIDA Research Monograph, v. 157, p. 117–135, 1995.

CSARDI, G. *et al.* The igraph software package for complex network research. InterJournal, complex systems, v. 1695, n. 5, p. 1–9, 2006.

FRANK, O.; SNIJDERS, T. Estimating the size of hidden populations using snowball sampling. Journal of Official Statistics, v. 10, p. 53–53, 1994.

GOODMAN, L. A. Snowball sampling. The annals of mathematical statistics, JSTOR, p.148–170, 1961.

HECKATHORN, D. D. 6. Extensions of Respondent-Driven Sampling: Analyzing Continuous Variables and Controlling for Differential Recruitment. Sociological Methodology, v. 37, n. 1, p. 151–208, ago. 2007.

HECKATHORN, D. D. Comment: Snowball versus Respondent-Driven Sampling. Sociological Methodology, v. 41, n. 1, p. 355–366, ago. 2011. ISSN 0081-1750, 1467-9531.

NOY, C. Sampling Knowledge: The Hermeneutics of Snowball Sampling in Qualitative Research. International Journal of Social Research Methodology, v. 11, n. 4, p. 327–344, out. 2008. ISSN 1364-5579, 1464-5300.

THOMPSON, S. Link-tracing and respondent-driven sampling. In: Hard-to-survey populations. Cambridge: Cambridge University Press, 2014. p. 503–516.

WESTON, S. foreach: Provides Foreach Looping Construct, 2020. R package version 1.5.1. Available at: <https://CRAN.R-project.org/package=foreach>.

WEJNERT, C. 3. An Empirical Test of Respondent-Driven Sampling: Point Estimates, Variance, Degree Measures, and Out-of-Equilibrium Data. Sociological Methodology, v. 39, n. 1, p. 73–116, ago. 2009.