

Design-Based Estimators for Snowball Sampling

Termeh Shafie *

Department of Statistics, Stockholm University
SE-106 91 Stockholm, Sweden

Abstract

Snowball sampling, where existing study subjects recruit further subjects from among their acquaintances, is a popular approach when sampling from hidden populations. Since people with many in-links are more likely to be selected, there will be a selection bias in the samples obtained. In order to eliminate this bias, the sample data must be weighted. However, the exact selection probabilities are unknown for snowball samples and need to be approximated in an appropriate way. This paper proposes different ways of approximating the selection probabilities and develops weighting techniques using the inverse of the selection probabilities. Some numerical examples for small graphs and simulations on larger networks are provided to compare the efficiency of the weighting techniques. The simulation results indicate that the suggested re-weighted estimators should be preferred to traditional estimators with equal sample weights for the initial snowball sampling waves.

1 Introduction

Standard sampling and estimation techniques require the sample selection to be done with known probabilities. However, for many populations of interest it is impractical or impossible to construct a sampling frame needed for the calculation of these probabilities. This could be due to the difficulty of locating members of the target population. These populations are referred to as hidden populations and are characterized by their lack of sampling frames and in some cases, also their strong privacy concerns. Examples of such populations are drug users, commercial prostitutes, illegal immigrants and the homeless. Kalton and Anderson (1986) discuss difficulties in statistical inference for rare or hidden populations and review different sampling procedures and their limitations.

One approach to sampling members of hard-to-reach populations while still obtaining unbiased estimates of population characteristics is through snowball sampling, where the initially sampled individuals will lead you to the other members of the hidden population, which in turn lead to other members and so on. Biernacki & Waldorf (1981) review problems and techniques of snowball sampling and applications of its

*termeh.shafie@stat.su.se

use can be found in Welch (1975) and Snow et al. (1981). Using this approach to include elements from the hidden population will lead to a sample selection bias. This occurs since those with many contacts are more likely to be included in the sample. Frank (1977, 1979) and Thompson and Frank(2000) consider statistical problems for snowball sampling and Thompson (2006) treats a special case of snowball sampling called walk sampling, where only one further vertex is selected at each sampling stage.

Since snowball samples are selected with unequal probabilities, the sample mean can no longer be used as a basis for making unbiased estimation about population characteristics. The degrees of the sampled elements will, at each sampling wave, determine the sample sizes obtained and if these degrees are correlated to the outcome of the study variable, we may get large biases in our estimations. To obtain an unbiased estimate, it is necessary to weight the sample data in some way. Typically these weights are inversely related to probabilities of selection.

In this paper, the probabilities of selection at each sampling wave are approximated in different ways and weighting techniques are applied to the sample data. The weighting techniques proposed are all based on the link information in the obtained sample, and substitute the equal sample weights in traditional estimators. Some simulations on larger networks are performed. The simulations consider two special cases and compare the efficiency of the proposed weighting techniques in terms of bias correction when making estimations using snowball samples of five waves.

2 Snowball Sampling

The snowballing process is as follows. It is done by first identifying a few members of the population, the initial sample, also referred to as starting seeds. A convenient way of finding the initial seeds is by site sampling. For instance, homeless could be initially sampled at a shelter.

The next step is then to ask each of the gathered seeds to identify other members of the population. Those who are not in the initial sample but mentioned by at least one individual in the initial sample, are part of the first wave of the snowball sample. Those who are neither members of the initial sample nor the first wave but mentioned by at least one member of the first wave, are said to belong to the second wave of the snowball sample, and so on. A wave is final if no new individuals are mentioned that have not been mentioned earlier or when a predefined wave number or sample size is reached.

2.1 Estimations with Snowball Samples

Some notations are introduced. Let U be a population with a known or unknown number of elements N . If the population is represented by a graph, the elements are the vertices and the contacts are the edges between the vertices. Each element is characterized by a real-valued property y_i which is unknown but observable if element i is sampled. Assuming that the population consists of drug-users, y_i may be quantities like the average amount of money spent on drugs per week or an indicator variable which equals 1 if the subject has a permanent residence. We are interested in the average quantity

$$\bar{y}_U = \sum_U \frac{y_i}{N}.$$

There exists no list or frame to sample from but the elements are connected by social relations. We pick an initial sample S_0 that we know of and these are questioned about y_i . Thereafter the elements of the initial sample are asked to give names and addresses of other members of the population whom they know of. We set $z_{ij} = 1$ if person i mentions person j . These are the edges between the elements of a graph. For simplicity, we assume that this relation is symmetric, i.e. $z_{ij} = z_{ji}$. In other words, if person i mentions person j , then person j will also mention person i . Thus the graphs considered are undirected. The adjacency matrix of the graph is defined as \mathbf{A} , where ij^{th} element of \mathbf{A} is z_{ij} . The degree for each vertex $i \in U$ is obtained by summing across the columns of each row in \mathbf{A} , i.e. $d_i = \sum_{j=1}^N z_{ji}$. Note that we assume that the graphs considered are all connected, i.e there are no isolated vertices.

The usual procedure is to stop the sampling after a fixed number of waves or when a sample of a sufficient size is reached. When sampling is done in snowball waves the chosen initial seed and its out-going edges will determine the obtained sample size. Traditionally, this is considered a sample and the population average, \bar{y}_U , is estimated by an unweighted sample average

$$\frac{\sum_S y_i}{\sum_S 1} = \frac{\sum_{i=1}^n y_i}{n},$$

where S is the obtained sample and n its size.

It is obvious that persons with many contacts, $d_i = \sum_U z_{ij}$, will have larger tendency to be included in the sample. If y_i is related to the number of social relations there may occur a large sample selection bias. However, since we observe the number of relations we may try to use a weighted estimator,

$$y_\omega^* = \sum_U \omega_i y_i, \quad (1)$$

ω_i are weights which can be computed from the sample, where $\omega_i = 0$ for all $i \notin n$, and satisfy $\sum \omega_i = 1$. Note that we assumed symmetry, i.e. $z_{ij} = z_{ji}$. If this assumption does not hold one has to ask about how many people person i believes would mention him/her.

It is obvious that the estimate y_ω^* is unbiased if and only if $E(\omega_i) = 1/N$ for all i , that is

$$E(y_\omega^*) = E\left(\sum_U \omega_i y_i\right) = \sum_U E(\omega_i) y_i = \sum_U \frac{y_i}{N} = \bar{y}_N. \quad (2)$$

3 Probability Related Weighting Techniques

Using equal sample weights when making estimations about population characteristics was shown to lead to biased results when snowball sampling. To adjust for this, sampled elements ought to be weighted by the reciprocal of their selection probabilities. In order to do this, the inclusion probabilities of the vertices in the population are approximated as shown below, and the weights are modified according to the approximations made. These modified weights will then be used in the estimator formulas (such as the average given in equation 1) and are referred to as design-based estimators.

3.1 Re-Weighting (RW)

Different re-weighting procedures (RW) are proposed here to evaluate whether or not the weights can be adjusted to approach $E(\omega_i) = 1/N$. These weighting techniques are

based on approximations of the sample selection probabilities by using the observed information about the relations between the elements in the sample. Using this information is rational since the probability of vertices in the population being included in the sample are correlated with their corresponding degrees. In other words, all weighting techniques presented here are based on the observed degrees in the sample.

Practically, these re-weightings can quite easily be implemented on social networks by asking each sampled unit to name their outgoing relations, even if we stop the sampling after that specific unit. By doing this we will obtain the degree of each sample element and approximating the sample selection probabilities needed for re-weighting.

In order to obtain

$$\sum \omega_i = 1,$$

all proposed weights are normalized according to

$$\omega_i = \frac{\vartheta_i}{\sum_S \vartheta_i},$$

where ϑ_i is defined for each of the four RW's in the following subsections.

3.1.1 RW1

The first proposed technique is performed by assuming that the inverse of the degrees are approximations of the inclusion probabilities for each vertex i in the population, i.e.

$$P(i \in S) \propto d_i \quad \text{for } i = 1, \dots, N.$$

Thus, the selected sample elements are weighted proportional to these probabilities and we have that

$$\vartheta_i = \frac{1}{d_i} \quad \text{for } i = 1, \dots, n.$$

Intuitively, these weights seem like a good and straightforward option. However, the degrees of the vertices in the initial sample will not affect their probabilities of selection. Taking this into consideration, RW2-RW4 are developed.

3.1.2 RW2

The second re-weighting technique, RW2, is similar to RW1 but with the difference that we change the weights of the starting elements for the reason mentioned above. This initial seed value is arbitrarily chosen and set proportional to 2. For the first wave we have the selection probability for the initial vertex proportional to $1/2$, and the selection probabilities for the remaining $(n - 1)$ draws proportional to $1/d_i$. Thus we have that

$$P(i \in S) \propto \begin{cases} 2 & \text{if vertex } i \text{ is a seed} \\ d_i & \text{if vertex } i \text{ is not a seed,} \end{cases}$$

and

$$\vartheta_i = \begin{cases} 1/2 & \text{if vertex } i \text{ is a seed} \\ 1/d_i & \text{if vertex } i \text{ is not a seed.} \end{cases}$$

3.1.3 RW3

The third re-weighting method approximates the sample selection probabilities with the inverse of degrees but with an unknown constant, c , added to the inclusion probabilities of the vertices in the sample. Thus we have that

$$P(i \in S) \propto d_i + c \quad \text{for } i = 1, \dots, N,$$

and

$$\vartheta_i = \frac{1}{d_i + c} \quad \text{for } i = 1, \dots, n.$$

Throughout this paper, we will use the constant value $c = 0.5$. Note also that when $c = 0$, RW3 coincides with RW1.

3.1.4 RW4

The last re-weighting technique presented here is somewhat different than the previous ones. Here, we will use the observed mean degrees of the sample to approximate the inclusion probabilities.

Assume the inclusion probability of the initial vertex is $(1/N)$ and the inclusion probabilities for the remaining $(n - 1)$ draws is equal to $d_i / \sum_i^N d_i$, where d_i is the degree of vertex i . Thus, the sample selection probabilities for each possible sample is approximately inversely proportional to

$$\left[\frac{1}{N} + (n - 1) \frac{d_i}{\sum_i^N d_i} \right].$$

After multiplying with $\sum d_i / (n - 1)$, and estimating the population mean degree $(\sum_i^N d_i / N)$ by the sample mean degree $(\sum_i^n d_i / n)$, we have that

$$\frac{1}{\vartheta_i} = \left[\frac{\bar{d}}{(n - 1)} + d_i \right], \quad \text{for } i = 1, \dots, n.$$

In this expression all terms can be calculated from the sample implying that no information about the population is needed.

4 Simulations

In this section, simulations are performed to evaluate how the different re-weightings (RW1-RW4) work for larger networks.

When the degrees d_i of all vertices $i \in U$ are determined, the networks can be generated by the algorithm presented in Shafie (2009), where the creation of a simple graph with only one type of undirected edges is described. It is only possible to construct such networks when a number of conditions are satisfied, one which is that no degree may be larger than half of the sum of degrees, or larger than $(N - 1)$.

In addition to earlier assumptions in this paper, we here assume that the vertices consist of two separate groups, denoted A and B. The degrees of the vertices in the two groups, d_A and d_B , are kept fixed and the snowballing from this network starts with one initial seed chosen randomly from the graph population. This initial seed is denoted by S_0 .

We consider the case when sampling is done in waves, that is, the sample sizes obtained at each wave depend on the degrees of the vertices selected at the previous wave. Thus, as our sampling procedure proceeds to the subsequent waves, the

probability of the elements in group A selecting those in group B (and vice versa) is proportional to the degrees of the vertices in the graph. The parameter of interest to estimate is the proportion of group A in the graph population, π_A , i.e. the population average of the study variable which is a dichotomous variable taking on the value 1 if the sampled element posses a specific property.

Further in this section, another simulation is run to illustrate the stopping rule bias mentioned in section 3.5. A two group network of $N = 100$ is generated in the same way as already mentioned. The proportion estimates of π_A are compared and it is shown that the biases differ depending on if we choose to sample in waves, or stop the snowballing after a fixed sample size n has been reached.

4.1 Case 1: Proportion Estimates when the two population groups are Equally Sized

We will consider two different simulation cases here and for both of these cases, snowball sampling will be performed in five waves. For the first case, networks of size $N = 100$ were simulated where the vertices in group A have degrees $d_A = 6$ and the degrees for the elements in group B is three times smaller, $d_B = 2$. Both groups are assumed to be equally sized implying that the parameter is equal to $\pi_A = 0.5$. From this network, snowball sampling was applied in 5 waves and the proportion of group A was estimated using the traditional estimator with equal sample weights and estimators with proposed RW weights.

Sample sizes obtained at each snowballing wave and the proportion estimates were averaged over 1000 repeated runs. The simulation results are plotted in Figure 4, where all four re-weighted proportion estimates are plotted with the traditional estimator, against both wave number and expected sample sizes. As seen from Figure 4, using the traditional estimator with equal sample weights will over-estimate π_A .

All the re-weighted estimators perform better than $P_{A,T}$ at the initial waves of the snowballing. For the first wave $P_{A,RW4}$ result in the best estimations. For wave two and three, $P_{A,RW3}$ give the best estimation results. After these waves, all of the re-weighted estimators fail in producing good estimation results and the traditional estimator should be preferred.

How well each of the estimator performs is dependent on the obtained sample sizes at each wave. When the sampling fractions increases, the traditional estimator becomes the better option, implying that re-weightings should be considered when snowballing less than 3 waves. Also, the standard errors decrease as the wave number increases. This is due to the larger variation in sample sizes and sample group compositions at the initial waves, where there are more relation options to new individuals not sampled in the previous waves.

For an estimation strategy (i.e. the combination of a sampling plan and an estimator), when the sample comprises the population, an estimator is said to be Cochran consistent if the value of the estimator is the same as the parameters of interest. As seen, none of the proposed estimators $P_{A,RW1} - P_{A,RW4}$ are Cochran consistent. In fact, the bias of these estimator increase as $n \rightarrow N$.

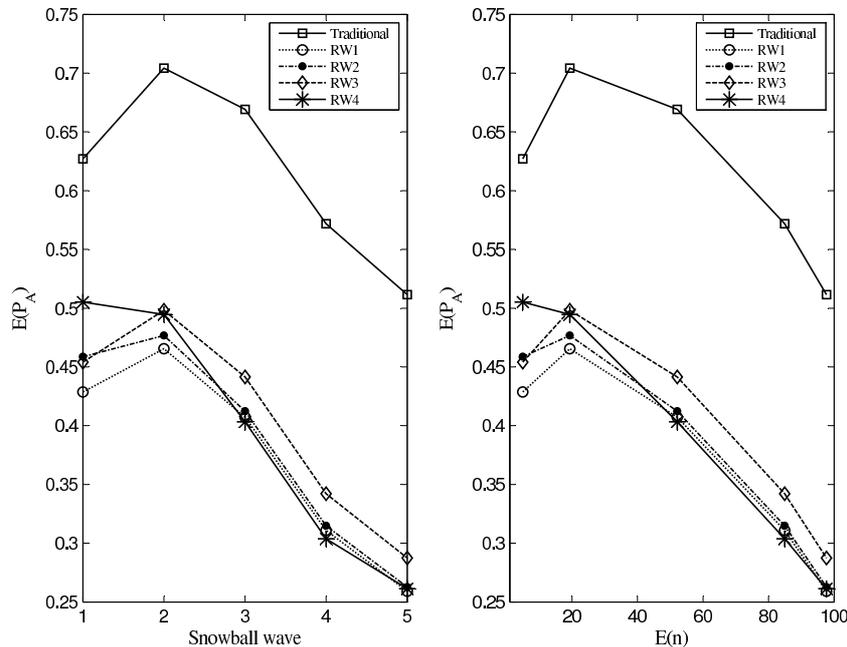


Figure 1: Case 1 proportion estimates of π_A averaged of 1000 repeated runs and plotted against wave number and expected sample sizes at each wave.

4.2 Case 2: Proportion Estimates when the two population groups are Un-Equally Sized

For the second simulation case, the network size remains the same ($N = 100$), but we drop the assumption about equally sized groups and instead assume that group A elements consist of only 20% of the population. Further, we assume a larger divergence between the degrees of those in group A and those in group B. The degrees are set to $d_A = 10$ and $d_B = 2$. As for the first simulation case, we use the traditional and the re-weighted estimators for estimating π_A . The results plotted in Figure 5. The results are consistent with those for Case 1. $P_{A,RW4}$ gives the smallest biases for the three first snowballing waves and for the subsequent waves, the traditional estimator with equal sample weights should be preferred.

Assume that we are interested in the estimation of another population proportion denoted Q . For instance, assume that the population consists of drug-users grouped after gender (A or B) and let Q be another binary study variable of heroin-users in the population of drug-users. The distribution of this variable over the graph could for instance be such that the majority of heroine-users are in group B, consistent with the second simulation case given here. The expected value of \hat{Q} is then a linear function of the estimated P_A ;

$$E(\hat{Q}) = P_A Q_A + (1 - P_A) Q_B.$$

As seen, if the selection bias is ignored in the estimations of P_A , they will reflect on other estimates made on population characteristics.

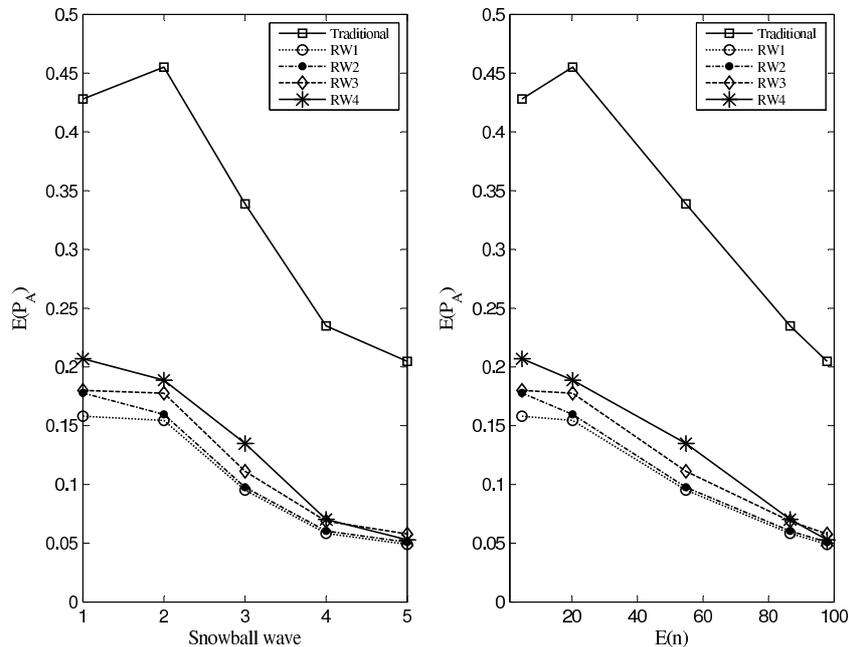


Figure 2: Case 2 proportion estimates of π_A averaged of 1000 repeated runs and plotted against wave number and expected sample sizes at each wave.

5 Conclusions

In this paper, four different ways of approximating the selection probabilities of snowball samples are presented using information about the degree of vertices in the obtained sample. These probabilities are then used to weight the sample data to eliminate the selection bias evident in the sampling procedure, i.e. the fact that people with many ingoing links are more likely to be sampled. These weights are inversely related to the probabilities of selection.

The weighting techniques are applied on snowball samples (performed in waves) from small graphs of only six vertices with varying degrees, but also on larger simulated two-group population networks with fixed degrees in each respective group. The results show that all re-weightings are to be preferred to equal sample weights, but only for the initial waves, where the selection bias is most visible. As the sampling fraction increase, the bias of the traditional estimator with equal weights decrease while the opposite occurs for the proposed estimators.

General conclusions about the re-weighting techniques can not be made since their performance is highly dependent on the graph size and structure. For the simulations made in this paper ($N = 100$), the fourth re-weighting, RW4, using the observed mean degrees of the samples obtained to estimate the inclusion probabilities, was shown to be preferable when estimating the group proportion of two group populations. In this paper, the mean degree of the graph was estimated using a straight mean of the observed degrees of the sampled elements.

To further evaluate the proposed weighting techniques and in order to get somewhat general results, larger simulations need to be performed. Also, one should con-

sider the MSE and variance. For small samples the traditional estimator may be the most preferable since the variance is smaller with equal sample weights.

We suggest a couple of topics for future research. One topic is how to estimate N by using information of how many already sampled elements are mentioned in the subsequent waves. Another topic is how to improve the estimates RW1-RW4 when N is known (e.g. with the use of finite population correction factors). And finally, a third topic is to combine these two topics.

In this paper, the approximations of the inclusion probabilities were all, as mentioned, based on the degrees of the sampled elements of the hidden population. These degree-orders are self-reported by the sampled subjects and may be inaccurate in real-life situations introducing bias in prevalence estimates. This emphasizes the need for more research on methods to measure the degree of each respondent more accurately and to study the robustness of these methods.

References

- Biernacki, P., and Waldorf, D. 1981. *Snowball sampling: Problems and Techniques of Chain Referral Sampling*. Sociological Methods and Research, 10(2):141–163.
- Frank, O. 1977. *Survey sampling in graphs*. Journal of Statistical Planning and Inference 1: 235–64.
- Frank, O. 1979. *Estimation of population totals by use of snowball samples*. P. Holland and S. Leinhardt (eds), Perspectives on Social Network Research. New York: Academic Press. Pp. 319–47.
- Kalton, G., and Anderson, D. 1986. *Sampling Rare Populations*. Journal of the Royal Statistical Society 149, No. 1, 65–82.
- Shafie, T. 2009. *Generating Flexible Networks*. Paper from upcoming dissertation.
- Snow, R. E., Hutcheson, J. D., Prather, J. E. 1981. *Using Reputational Sampling to Identify Residential Clusters of Minorities Dispersed in a Large Urban Region: Hispanics in Atlanta*. Georgia. Proc. of the Section on Survey Research Methods, American Statistician Association 101–106.
- Thompson, S.K. 2006 *Targeted Random Walk Designs*. Survey Methodology 32: 11–24.
- Thompson, S. and Frank, O. 2000 *Model-based estimation with Link-Tracing Sampling Designs*. Survey Methodology 26: 87–98.
- Welch, S. 1975. *Sampling by Referrals in a Dispersed Population*. Public Opinion Quarterly 39(2):237–45.