Journal of Official Statistics
Vol. 10. No. 1, 1994, pp. 53–67
© Statistics Sweden

Estimating the Size of Hidden Populations Using Snowball Sampling

Ove Frank[1] *and Tom Snijders*[2]

Abstract: Snowball sampling is a term used for sampling procedures that allow the sampled units to provide information not only about themselves but also about other units. This might be advantageous when rare properties are of interest. This article illustrates snowball sample situations and discusses various modelling and estimation problems in this context. The problem of estimating the size of a population is discussed for both design-based and model-based approaches. An application to a study of heroin use is included. Simulation results are provided for comparing and evaluating various estimators.

Key words: Network sampling; random graphs; link-tracing designs.

1. Introduction

In order to estimate the number of cocaine users in a big city, standard sampling methods are very inefficient. This is due to the fact that small proportions cannot be estimated with sufficient accuracy without taking very large samples. A simple random sample of size n from a population of size N provides a sample proportion $\hat{p}$ that varies around the population proportion p with a standard deviation of $\sqrt{p(1-p)/n}$, so that for a small proportion p, a relative accuracy of 10% would require a sample size of about $400/p$, which could be very large. Such a large sample size would necessarily lead to superficial or cursory ways of interviewing. If the property investigated is generally considered to be socially sensitive or even taboo, then the interviews might result in severe underreporting.

A very small subpopulation or a subpopulation of individuals who are unwilling to disclose themselves will here be referred to as a hidden population. Very few members of a hidden population can usually be found by standard sampling methods. Often, however, there exists a contact pattern between the members of the hidden population, which means that they know or know of each other. If these contacts could be used for finding members of the hidden population, then new estimation problems arise because of the non-standard sampling procedure. Snowball sampling is a way of having initially sampled individuals lead you to other

[1] Department of Statistics, Stockholm University, S-106 91 Stockholm, Sweden.
[2] Department of Statistics and Measurement Theory, University of Groningen, The Netherlands.
Acknowledgements: Work for this paper was partially supported by the Swedish Council for Research in the Humanities and the Social Sciences (HSFR). The cooperation between the authors benefited from a travel grant from the Netherlands Organization for Scientific Research (NWO). One of the authors also acknowledges earlier work carried out under a contract with the research bureau Intraval.

members of the hidden population, which in turn could lead to further members, etc. Various statistical methods for snowball samples are investigated by Frank (1977, 1979). Some difficulties in statistical inference for snowball sampling are discussed by Kalton and Anderson (1986) and by Snijders (1992). A review of the literature on general link-tracing designs for investigating hidden populations is given by Spreen (1992). The reader is referred to this paper for further bibliographic references.

The hidden population with its contact pattern between members is naturally considered as a directed graph. The members of the hidden population are the vertices of the graph. The number of vertices, v, is estimated from sample information. Vertex i has an arc to vertex j if individual i, when questioned, would mention j as a member of the hidden population.

Assume that an initial sample of n members of the hidden population is available. Each of these individuals is supposed to name the other members they know of. Some are mentioned by several individuals; some of those mentioned are included in the initial sample and some are not. Those who are not in the initial sample and who are mentioned by at least one individual in the initial sample are said to belong to the first wave of the snowball sample created by the initial sample. Those who are not members of the initial sample or the first wave of the snowball sample but are mentioned by at least one member of the first wave are said to belong to the second wave of the snowball sample, etc. The snowball sample consists of the initial sample and all the waves successively found around it. A wave is final if its members do not mention any individuals that have not been previously mentioned. Snowball samples often are incomplete in the sense that the sampling stops (for obvious reasons) before the last wave.

The initial sample could be a simple random sample from a population containing the hidden population as a subpopulation. In practice a more convenient initial sample is usually obtained by site sampling, that is, by sampling certain sites where members of the hidden population are known to frequent. For instance, drug users could be initially sampled at certain bars, clubs, or police stations.

There is no frame of the hidden population as long as it is hidden, so random sampling procedures cannot be designed exclusively for the hidden population. However, that they cannot be designed does not mean that they cannot be used. A Bernoulli sampling design is a way of selecting individuals from a population according to a procedure that decides independently for each individual whether or not he or she should be selected for the sample. Such a procedure could work even if it is not run by the sampling investigator. In the context of sampling cocaine users, for instance, it is conceivable that the initial sample is created by the members of the hidden population that are in need of medical treatment or social support. A simple model for such self-generated initial samples could be Bernoulli samples with a common but unknown selection probability. More elaborate models could distinguish between different selection probabilities in different strata of the hidden population.

This paper illustrates how snowball sampling methods can be used to estimate hidden populations. Both design-based and model-based approaches are discussed. Section 2 introduces notation and terminology. Section 3 takes a simple model-based approach and develops estimators of the model parameters as well as of the size of the hidden population. Section 4 analyzes the precision of the estimators. In Section

5 a design-based approach is discussed and various estimators of the size of the hidden population are developed. A practical illustration is provided in Section 6, and simulation results which shed some light on the performance of the estimators are reported in Section 7. The paper closes with a discussion section.

2. Concepts of Snowball Sampling

Consider a directed graph on v vertices. The vertices are labeled by integers and the vertex set is denoted by $V = \{1, \ldots, v\}$. The arcs are ordered pairs (i,j) of vertices from V; if $i = j$, the arc is called a loop. The arc set W is a subset of V^2 containing all loops $\{(i,i) : i \in V\}$, for convenience. The initial sample S_0 is a subset of V. The initial sample and the arc set are represented by indicator variables $x = (x_i : i \in V)$ and $y = (y_{ij} : (i,j) \in V^2)$, respectively. Thus x_i is 1 or 0 according to whether or not vertex i is in the initial sample, and y_{ij} is 1 or 0 according to whether or not the graph contains an arc from i to j. The matrix y is the adjacency matrix of the graph, and the diagonal entries of y are all equal to 1.

Denote by A_j and B_j the subsets of vertices after and before vertex j, respectively. More precisely

$$A_j = \{i \in V : y_{ji} = 1\},$$

$$B_j = \{i \in V : y_{ij} = 1\}$$

so that A_j is indicated by row j of y, and B_j is indicated by column j of y. The sizes of A_j and B_j are called the out-degree and the in-degree of vertex j, and they are denoted by a_j and b_j, respectively. They can be obtained as the row and column sums of the adjacency matrix y

$$a_j = |A_j| = \sum_{i=1}^{v} y_{ji}, \quad b_j = |B_j| = \sum_{i=1}^{v} y_{ij}.$$

For any subset S of V we denote by $A(S)$ and $B(S)$ the subsets of vertices after and before any of the vertices in S, respectively, that is

$$A(S) = \bigcup_{j \in S} A_j, \quad B(S) = \bigcup_{j \in S} B_j.$$

The first wave of the snowball sample initiated by S_0 is given by $S_1 = A(S_0) \cap \bar{S}_0$. The second wave is given by $S_2 = A(S_1) \cap \bar{S}_0 \cap \bar{S}_1$, and so forth. The snowball initiated by S_0 is given by $S_0 \cup S_1 \cup \ldots \cup S_K$ where K is the number of waves of the snowball and S_{K+1} is the first empty set in the sequence $S_1, S_2, \ldots$

Figure 1 shows the adjacency matrix y of a graph on $v = 25$ vertices. The vertices have been ordered so that the first 5 vertices are the vertices in the initial sample S_0. The next 8 vertices are the vertices in the first wave S_1. Note that there are 13 vertices after S_0 but only 8 of these are in the

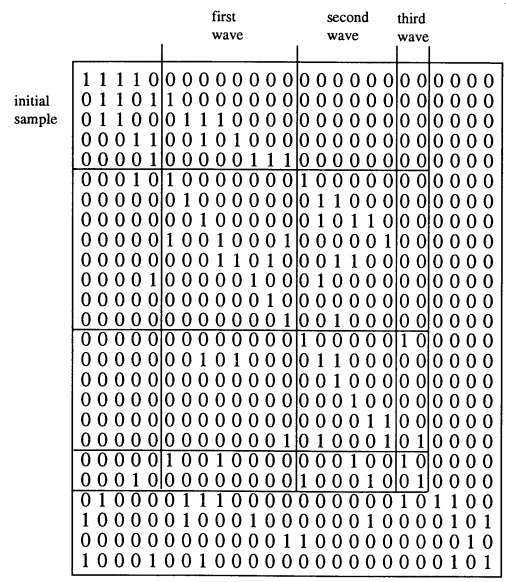

Fig. 1. Adjacency matrix of a graph on 25 vertices. The vertices have been ordered to simplify the illustration of a three-wave snowball sample of size 21 generated from an initial sample of size 5

complement of S_0. The next 6 vertices are the vertices in the second wave S_2. Note that there are 16 vertices after S_1 but 2 of these are in S_0 and 8 are in S_1. The next 2 vertices are the vertices in the third wave S_3. There are 8 vertices after S_3 but one of these is in S_0, 2 are in S_1, 3 are in S_2 and 2 are in S_3 so that S_4 is empty. Finally, the last 4 vertices are the vertices not reachable from S_3.

3. Model-Based Estimation

Assume that the initial sample S_0 is a Bernoulli subset of V with selection probability α. This means that the indicators $x_1, \ldots, x_v$ are independent identically distributed Bernoulli(α) variables. The initial sample size $n = |S_0|$ is then binomial (v, α).

Assume further that the arc set W is a Bernoulli subset of V^2 with selection probability 1 for the loops and selection probability β elsewhere. This means that the indicators (y_{ij}) are 1 on the diagonal and are independent identically distributed Bernoulli(β) variables off the diagonal. The parameters of the statistical model are v, α, and β. This paper concentrates on the estimation of v.

Let r be the number of nonloop arcs in the initial sample, that is, $r + n = |W \cap S_0^2|$. Conditionally on S_0, r is binomial $(n(n-1), \beta)$, and unconditionally r has a probability

$$\sum_{n=0}^{v} \binom{v}{n} \alpha^n (1-\alpha)^{v-n} \times \binom{n(n-1)}{r} \beta^r (1-\beta)^{n(n-1)-r}.$$

The conditional expected value of r is $\mathrm{E}(r|n) = n(n-1)\beta$, and the unconditional expected value is $\mathrm{E}r = v(v-1)\alpha^2\beta$. Let s be the number of arcs from the initial sample to the first wave of the snowball, that is, $s = |W \cap (S_0 \times S_1)|$. Conditionally on S_0, s is binomial $(n(v-n), \beta)$, and unconditionally s has a probability

$$\sum_{n=0}^{v} \binom{v}{n} \alpha^n (1-\alpha)^{v-n} \times \binom{n(v-n)}{s} \beta^s (1-\beta)^{n(v-n)-s}.$$

The conditional and unconditional expected values of s are $\mathrm{E}(s|n) = n(v-n)\beta$ and $\mathrm{E}s = v(v-1)\alpha(1-\alpha)\beta$. Now, conditionally on n, moment estimators of β and v can be obtained from the two equations

$$r = n(n-1)\beta$$
$$s = n(v-n)\beta$$

leading to

$$\hat{\beta}_1 = r/n(n-1)$$
$$\hat{v}_1 = [nr + (n-1)s]/r.$$

Unconditionally, moment estimators of α, β, v can be obtained from the three equations

$$n = v\alpha$$
$$r = v(v-1)\alpha^2\beta$$
$$s = v(v-1)\alpha(1-\alpha)\beta$$

leading to

$$\hat{\alpha}_2 = r/(r+s)$$
$$\hat{\beta}_2 = r(r+s)/n[(n-1)r + ns]$$
$$\hat{v}_2 = n(r+s)/r$$

generally provided that no denominators are zero.

There are other ways of getting moment estimators. The arc frequencies r and s are independent conditionally on n, and their sum $t = r + s$ is binomial $(n(v-1), \beta)$. Moreover, the size $m = |S_1|$ of the first wave of the snowball is conditionally binomial $(v-n,\ 1-(1-\beta)^n)$. The two

moment equations

$$t = n(v-1)\beta$$
$$m = (v-n)[1-(1-\beta)^n]$$

lead to an estimator of v satisfying

$$1 - \frac{m}{v-n} = \left[1 - \frac{t}{n(v-1)}\right]^n.$$

The solution to this equation will be denoted $\hat{v}_3$. It can be obtained by a straightforward iterative procedure. The corresponding β-estimator will be denoted $\hat{\beta}_3$. These moment estimators can also be obtained as maximum likelihood estimators, as we show next.

The maximum likelihood estimators will be derived conditionally on n. For $k = 0, \ldots, n$ let m_k be the number of individuals not in S_0 that are mentioned by exactly k members of S_0. Then

$$m = m_1 + m_2 + \ldots + m_n$$
$$s = m_1 + 2m_2 + \ldots + nm_n$$
$$m_0 = v - n - m$$

and $(m_1, \ldots, m_n)$ is multinomial $(v-n, p_1, \ldots, p_n)$ where

$$p_k = \binom{n}{k}\beta^k(1-\beta)^{n-k} \text{ for } k = 0, \ldots, n.$$

Now, r and $(m_1, \ldots, m_n)$ are sufficient and conditionally independent, and for these data the conditional likelihood is given by

$$L = \binom{n(n-1)}{r}\beta^r(1-\beta)^{n(n-1)-r} \times (v-n)! \prod_{k=0}^{n}(p_k^{m_k}/m_k!)$$
$$= \frac{(v-n)!}{(v-n-m)!}\beta^t(1-\beta)^{n(v-1)-t} \times \binom{n(n-1)}{r}\prod_{k=1}^{n}\frac{\binom{n}{k}^{m_k}}{m_k!}.$$

It follows that m and t are sufficient statistics, and the essential part of the likelihood is given by

$$g(\beta, v) = \frac{(v-n)!}{(v-n-m)!}\beta^t(1-\beta)^{n(v-1)-t}.$$

For any fixed v, this is maximized by $\beta = t/n(v-1)$, while for any fixed β, this is maximized by the integer part of the solution to the equation $g(\beta, v-1) = g(\beta, v)$ which is equivalent to $1 - m/(v-n) = (1-\beta)^n$. These equations are identical to the moment equations defining $\hat{v}_3$.

4. Distributions and Standard Errors

Conditionally on the size n of the initial sample, the size v of the hidden population is estimated by

$$\hat{v}_1 = n + s(n-1)/r$$

if the statistics r and s are used, and by $\hat{v}_3$ if the sufficient statistics t and m are used.

Conditionally on n, the statistics r and s are independent and have binomial distributions. The delta method, using a first order Taylor series, yields the following approximation to the asymptotic variance of $\hat{v}_1$

$$\text{Var}(\hat{v}_1|n) \approx (v-1)(v-n)(1-\beta)/\beta n(n-1)$$

which can be estimated by

$$\text{var } \hat{v}_1 = (n^2 - n - r)(n-1)s(s+r)/nr^3.$$

The estimated relative variance can be shown to be bounded as follows

$$\frac{\text{var } \hat{v}_1}{\hat{v}_1^2} \leqslant \frac{(n^2-n-r)s}{n(n-1)r(s+r)} < \frac{1}{r}.$$

Although rough, this bound does indicate the correct order of magnitude and can be helpful for the sample design.

Another way to approach the distribution of $\hat{v}_1$ is by conditioning not only on n but also on $t = r + s$. The conditional

distribution of s is hypergeometric with parameters $n(v-1)$, $n(n-1)$, and t. The estimation of v for this hypergeometric distribution is equivalent to the capture-recapture problem treated by Chapman (1951) who is cited by Johnson and Kotz (1969, pp. 146–147). Chapman (1951) suggests for the capture-recapture problem to use an approximation which in this case amounts to

$$\mathrm{E}\left\{\frac{1}{r+1}\middle| n,t\right\} \approx (vn-n+1)/(t+1)(n^2-n+1)$$

and which leads to the estimator $\hat{v} = n + (n-1)s/(r+1)$. The mathematical advantage of this estimator is that it has, in contrast to $\hat{v}_1$, finite mean and variance. From a practical point of view, however, this estimator is hardly different from $\hat{v}_1$ unless r is so small that both estimators have very large relative standard errors. Based on the hypergeometric distribution for r, it is also possible to construct exact tests and confidence intervals for v. We do not elaborate on this.

The joint distribution of t and m, necessary for the distribution of $\hat{v}_3$, is more difficult to handle. Since r and (s,m) are independent conditionally on n, the complicated part of the calculations involves the simultaneous distribution of s and m. Their marginal distributions are both binomial as noted above, binomial $(n(v-n),\beta)$ for s, and binomial $(v-n, 1-(1-\beta)^n)$ for m. The distribution of m conditional on s can be deduced by a combinatorial argument using the principle of inclusion and exclusion

$$P(m|n,s) = \frac{\binom{v-n}{m}}{\binom{n(v-n)}{s}} \times \sum_{k\geqslant 0}(-1)^k\binom{m}{k}\binom{n(m-k)}{s}.$$

This distribution is mentioned by Johnson and Kotz (1969, p. 252) in connection with restricted multinomial occupancy problems. The conditional distribution of s given m can be given as

$$P(s|n,m) = \left(\frac{\beta}{1-\beta}\right)^s\left(\frac{(1-\beta)^n}{1-(1-\beta)^n}\right)^m \times \sum_{k\geqslant 0}(-1)^k\binom{m}{k}\binom{n(m-k)}{s}.$$

Neither distribution lends itself to exact calculations of the moment properties of the estimator $\hat{v}_3$. An asymptotic approximation can be made under the following assumptions: n and v tend to infinity; the expected in- and out-degree, $\beta(v-1)$, tends to a positive finite limit λ; the initial sampling fraction n/v tends to 0 but n^2/v tends to infinity. Now m/n and t/n can be shown to converge in probability to λ, and the asymptotic variance of $\hat{v}_3$ can be shown to be equal to the asymptotic variance of the approximation $t(t+2n-2)/\ 2(t-m)$ which is

$$\mathrm{Var}(\hat{v}_3|n) \approx v^3/n^2\lambda(1+\lambda/2).$$

For the asymptotic marginal variance, n, may be replaced by αv, leading to

$$\mathrm{Var}\ \hat{v}_3 \approx v/\alpha^2\lambda(1+\lambda/2).$$

It can also be shown that a corresponding estimator is $\mathrm{var}\,\hat{v}_3 = (\hat{v}_3-n)^2/(t-m)$. The expression for $\mathrm{Var}\ \hat{v}_3$ can be compared with the asymptotic variance of $\hat{v}_1$ by substituting $\beta = \lambda/v$ and $n = \alpha v$, leading to

$$\mathrm{Var}\ \hat{v}_1 \approx v/\alpha^2\lambda.$$

This shows that the maximum likelihood estimator $\hat{v}_3$ has an asymptotic relative efficiency with respect to the moment estimator $\hat{v}_1$ which increases from 1 for $\lambda \to 0$ to infinity for $\lambda \to \infty$. The relative efficiency of $\hat{v}_3$ with respect to $\hat{v}_1$ can be estimated by $1 + t/2n$.

With the expressions given for $\mathrm{var}\,\hat{v}_1$ and $\mathrm{var}\,\hat{v}_3$, confidence intervals for v can be

constructed in the usual way (assuming approximate normality for the estimators). Thus, approximate 95% confidence intervals for v are given by

$$\hat{v}_1 \pm 2\sqrt{\operatorname{var} \hat{v}_1}$$
$$= n + (n-1)s/r$$
$$\pm 2\sqrt{(n^2 - n - r)(n-1)s(s+r)/nr^3}$$
$$\hat{v}_3 \pm 2\sqrt{\operatorname{var} \hat{v}_3} = \hat{v}_3 \pm 2(\hat{v}_3 - n)/\sqrt{t-m}.$$

5. Design-Based Estimation

This section is concerned with the estimation of the number of vertices of a fixed unknown directed graph. In other words, the arc indicators y_{ij} are not random but unknown. We call this design-based estimation because the population (digraph) is fixed and probability plays a role only via the sampling procedure; this term might be criticized by noting that the snowball sample is determined not only by chance, but also by the network structure, i.e., the arcs. The procedure for the initial sample is again a Bernoulli sample with unknown sampling probability α. In Section 5.1 the moment method is followed again, and the consistency of the produced estimator is investigated. In Section 5.2 a Horvitz-Thompson type estimator is derived. Section 5.3 presents jackknife estimators for the variances of the estimators for v.

5.1. *Moment estimators*

For the moment method, the statistics n, r, s, and m are used again. Since $x_1, \ldots, x_v$ are independent Bernoulli variables with parameter α, the expected values can easily be calculated

$$\mathrm{E}n = \mathrm{E}\sum_i x_i = v\alpha$$

$$\mathrm{E}r = \mathrm{E}\sum_{i \neq j} x_i x_j y_{ij} = (w - v)\alpha^2$$

where $w = |W|$ is the total number of arcs;

$$\mathrm{E}s = \mathrm{E}\sum_{i \neq j} x_i(1 - x_j)y_{ij}$$
$$= (w - v)\alpha(1 - \alpha)$$
$$\mathrm{E}m = \mathrm{E}\sum_j (\max_{i \in B_j} x_i - x_j)$$
$$= \mathrm{E}\sum_j \left(1 - x_j - \min_{i \in B_j}(1 - x_i)\right)$$
$$= v(1 - \alpha) - \sum_j (1 - \alpha)^{b}_{j}.$$

It can be concluded that the use of the statistics n, r, and s leads to estimators $\hat{\alpha}_2$ and $\hat{v}_2$ that were also found in the model-based approach. The statistic m can be used in combination with another statistic, denoted k, and defined as the number of vertices in the initial sample connected by at least one arc to another vertex in the initial sample

$$k = \sum_j x_j \max_{i \neq j} y_{ij} x_i.$$

Its expected value is

$$\mathrm{E}k = v\alpha - \alpha \sum_j (1 - \alpha)^{b_j - 1}.$$

This implies $\mathrm{E}k = \alpha E(k + m)$, which leads to

$$\hat{v}_4 = n(k + m)/k$$

as another moment estimator for v and $\hat{\alpha}_4 = k/(k + m)$ as the corresponding estimator for α. In analogy to the way $\hat{v}_1$ is related to $\hat{v}_2$, and motivated by simulation results (see Section 7), the formula for $\hat{v}_4$ can be slightly modified to give another estimator

$$\hat{v}_5 = [nk + (n-1)m]/k.$$

The consistency of these estimators is conveniently studied for an asymptotic situation that is compatible with the asymptotic situation considered at the end of Section 4. We assume that $v \to \infty$ and $\alpha \to 0$ in such a way that $\mathrm{E}n = v\alpha \to \infty$; moreover, we assume that all in- and

out-degrees are bounded: $a_i, b_i \leqslant M < \infty$. This implies that $(w - v)/v \leqslant M$. It can be shown that the assumptions imply that

$$\begin{aligned}\operatorname{Var} r &\leqslant 2\alpha^2(1-\alpha^2)w + 4M\alpha^3(1-\alpha)w \\ &= O(v\alpha^2)\end{aligned}$$

and

$$\begin{aligned}\operatorname{Var} s &\leqslant \alpha(1-\alpha)(1-\alpha+\alpha^2)w \\ &\quad + 2\alpha(1-\alpha)^3 Mw = O(v\alpha).\end{aligned}$$

This implies that $n/v\alpha$, $(r+s)/(w-v)\alpha$, and $r/(w-v)\alpha^2$ all converge in probability to 1 since the expectations are 1 and the variances tend to 0. It follows that $\hat{v}_2/v$ converges in probability to 1, so $\hat{v}_2$ is a consistent estimator.

Similarly, the asymptotic orders of $\operatorname{Var} m$ and $\operatorname{Var} k$ are given by

$$\begin{aligned}\operatorname{Var} m &\leqslant \alpha(1-\alpha)(w-v) \\ &\quad + \alpha(1-\alpha)vM^3 = O(v\alpha)\end{aligned}$$

and

$$\operatorname{Var} k \leqslant \alpha^2(w-v) + \alpha^2 vM^3 = O(v\alpha^2).$$

This implies that $m/(w-v)\alpha(1-\alpha)$ and $k/(w-v)\alpha^2$ both converge in probability to 1, and it follows that this also holds for $\hat{v}_4/v$ and $\hat{v}_5/v$. Thus it can be proved that $\hat{v}_1$, $\hat{v}_2$, $\hat{v}_4$, and $\hat{v}_5$ are consistent estimators. For the estimator $\hat{v}_3$ consistency is not guaranteed. Suppose that $(w-v)/v \to \lambda$ and denote the variance of the in-degrees by

$$\sigma^2 = \frac{1}{v}\sum_j \left(b_j - \frac{w}{v}\right)^2.$$

Then

$$\begin{aligned}\operatorname{Var}(t-m) &\leqslant \alpha(1-\alpha)w \\ &\quad + \alpha(1-\alpha)vM^3 = O(v\alpha)\end{aligned}$$

and $\hat{v}_3/v$ can be shown to converge in probability to $\lambda(\lambda+2)/[\lambda(\lambda+1)+\sigma^2]$. For arbitrary λ and σ^2, this limit can be anywhere between 0 and 2. Under the Bernoulli graph model, σ^2 tends to λ so that the limit is 1. It can be concluded that $\hat{v}_3$ is consistent under the Bernoulli graph model, but not under the model of an arbitrary fixed graph. In empirically found graphs, the variance of the in-degrees is often larger than the value that is expected under the Bernoulli model. A test of the Bernoulli model as a null hypothesis with the variance of the in-degrees as test statistic can be found in Snijders (1981).

5.2. *Horvitz-Thompson estimator*

If the initial sampling fraction α is known, (a Horvitz-Thompson estimator can be computed. This estimator is

$$\hat{v}_{\mathrm{HT}}(\alpha) = \sum_{j=1}^{v} \max_i x_i y_{ij}/\pi_j = \sum_{j \in S_0 \cup S_1} \pi_j^{-1}$$

where π_j is the inclusion probability

$$\begin{aligned}\pi_j &= P\{j \in S_0 \cup S_1\} = \mathrm{E} \max_i x_i y_{ij} \\ &= 1 - (1-\alpha)^{b_j}.\end{aligned}$$

In practice, α is not known; one of the estimators for α could be substituted. Since simulation results (see Section 7) suggest that $\hat{v}_1$ is slightly better than $\hat{v}_2$, and $\hat{v}_5$ is slightly better than $\hat{v}_4$, we restrict attention to $\hat{\alpha}_1 = n/\hat{v}_1$, $\hat{\alpha}_3 = n/\hat{v}_3$, and $\hat{\alpha}_5 = n/\hat{v}_5$. This yields the estimators $\hat{v}_6 = \hat{v}_{\mathrm{HT}}(\hat{\alpha}_1)$, $\hat{v}_7 = \hat{v}_{\mathrm{HT}}(\hat{\alpha}_3)$, and $\hat{v}_8 = \hat{v}_{\mathrm{HT}}(\hat{\alpha}_5)$.

For the estimators $\hat{v}_{\mathrm{HT}}(\alpha)$ and $\hat{v}_{\mathrm{HT}}(\hat{\alpha})$, it is necessary to observe the in-degrees of all sampled vertices. This can be difficult or even infeasible; we return later to discuss the observational requirements of the various estimators.

The variance of $\hat{v}_{\mathrm{HT}}(\alpha)$ depends on the joint inclusion probabilities

$$\pi_{ij} = P\{i,j \in S_0 \cup S_1\}.$$

It can be proved that

$$\begin{aligned}\pi_{ij} &= 1 - (1-\alpha)^{b_i} - (1-\alpha)^{b_j} \\ &\quad + (1-\alpha)^{b_i + b_j - b_{ij}}\end{aligned}$$

where b_{ij} is the number of vertices with arcs to both i and j

$$b_{ij} = \sum_{h=1}^{v} y_{hi} y_{hj}.$$

In order to calculate the Horvitz-Thompson or Yates-Grundy estimators for $\mathrm{Var}\, \hat{v}_{\mathrm{HT}}(\alpha)$, it is necessary to know b_{ij} for all i and j in the sample. This is a strong requirement on the observations; it practically implies that the in-neighbourhoods B_j are known for all $j \in S_0 \cup S_1$, which more or less implies a two-wave snowball sample. We do not pursue this but, instead, apply the jackknife principle to the initial sample in order to get standard errors for the various estimators.

5.3. *Jackknife standard errors*

The jackknife principle (explained, e.g., in Efron 1982) can be used to obtain standard errors for the derived estimators. Some computer simulations showed, however, that the standard way of applying the jackknife principle to the initial sample yields too large values for the standard errors. A more detailed analysis is necessary.

The most frequently applied jackknife method is based on excluding each one of the sample elements in turn. For an estimator $\hat{v}$, denote by $\hat{v}_{(i)}$ the corresponding estimator when vertex i is deleted from the initial sample and when all vertices j that are after i but not after any other vertices in the initial sample are deleted from the first wave of the snowball sample. (If vertex i is after some other vertex j in the initial sample, then in the data used to calculate $\hat{v}_{(i)}$, vertex i will show up in the first wave.) The standard version of the jackknife variance estimator for $\hat{v}$ is

$$\mathrm{var}_j^{(s)} \hat{v} = \frac{n-1}{n} \sum_{i=1}^{n} (\hat{v}_{(i)} - \hat{v}_{(\cdot)})^2$$

where

$$\hat{v}_{(\cdot)} = \frac{1}{n} \sum_{i=1}^{n} \hat{v}_{(i)}.$$

There are several approaches to argue why, in many situations, this makes sense as a variance estimator; see Efron (1982). One of these approaches is based on the observation that $\mathrm{var}^{(s)} j\hat{v}$ is unbiased if $\hat{v}$ is an average of n independent identically distributed (i.i.d.) random variables. More generally, it is an unbiased variance estimator if

$$\mathrm{var}\hat{v}_{(i)} = \frac{n}{n-1} \mathrm{var}\, \hat{v}$$

$$\rho(\hat{v}_{(i)}, \hat{v}_{(j)}) = \frac{n-2}{n-1} \text{ for } i \neq j.$$

These conditions are far from being satisfied in the case of the estimators $\hat{v}$ in this paper. Our estimators $\hat{v}$ are not at all similar to averages of n i.i.d. variables. The available data consist of the $n \times n$ adjacency matrix $Y_{00} = (y_{ij}; i,j \in S_0)$ for the initial sample together with the $n \times m$ adjacency matrix $Y_{01} = (y_{ij}; i \in S_0, j \in S_1)$ of arcs between initial sample and first wave (cf. Figure 1). Deleting vertex i from the initial sample has the following effect on these data matrices: row i *and* column i are deleted from Y_{00}; row i is deleted from Y_{01}; and there may be some changes of columns in Y_{01}. Matrix Y_{00} contains, conditionally on n, $n(n-1)$ nontrivial random variables (the diagonal elements are all 1). If the sampling probability α is small (as we have assumed), then $\mathrm{E}\{m|n\}$ is approximately proportional to n; this implies that, conditionally on n, the number of random variables in Y_{01} is approximately proportional to n^2. This suggests that the variances of our estimators $\hat{v}$ are inversely proportional, approximately, to $n(n-1)$ or n^2 rather than n. This is confirmed by the expressions found in Section 4: $\mathrm{Var}\,(\hat{v}_1|n)$ is approximately inversely proportional to $n(n-1)$, and $\mathrm{Var}\, \hat{v}_3$ to α^2.

It can be concluded that to propose a valid jackknife procedure we should not regard our estimators $\hat{v}$ as smooth functions of a sample of n i.i.d. random variables, but rather as functions of a numerical relationship between n vertices. More specifically, we regard the estimators $\hat{v}$ as analogous to the sample mean $\bar{z}$ of a square $n \times n$ matrix of random variables, where $z_{ii} \equiv 0$ and the z_{ij} for $i \neq j$ are i.i.d. random variables with variance σ^2. The sample mean is

$$\bar{z} = \frac{1}{n(n-1)} \sum_{i \neq j} z_{ij}.$$

The sample mean of all variables except those in the ith row and the ith column is denoted $\bar{z}_{(i)}$. We suggest that the relation between $\hat{v}$ and $\hat{v}_{(i)}$ is analogous to that between $\bar{z}$ and $\bar{z}_{(i)}$. For the z variables, it holds that

$$\operatorname{var} \bar{z}_{(i)} = \frac{n}{n-2} \operatorname{var} \bar{z}$$

$$\rho(\bar{z}_{(i)}, \bar{z}_{(j)}) = \frac{n-3}{n-1}.$$

This implies that an unbiased variance estimator for $\bar{z}$ is

$$\frac{n-2}{2n} \sum_{i=1}^{n} (\bar{z}_{(i)} - \bar{z}_{(\cdot)})^2$$

where

$$\bar{z}_{(\cdot)} = \frac{1}{n} \sum_{i=1}^{n} \bar{z}_{(i)}.$$

Accordingly, for an estimator $\hat{v}$ we propose the jackknife variance estimator

$$\operatorname{var}_J \hat{v} = \frac{n-2}{2n} \sum_{i=1}^{n} (\hat{v}_{(i)} - \hat{v}_{(\cdot)})^2.$$

Standard errors for $\hat{v}$ are obtained as $\{\operatorname{var}_J \hat{v}\}^{1/2}$. A further justification of these standard errors may be given by simulation experiments; see Section 7.

6. An Application and Some Practical Remarks

Practical work on these estimators was started by Snijders in the context of a study of cocaine use in Rotterdam. The results are presented in Bieleman, Diaz, Merlo, and Kaplan (1993). In this paper we refer to a study of heroin use in the town of Groningen conducted by the research bureau Intraval; see Intraval (1991). A snowball sample of heroin users was taken where the initial sample consisted of $n = 34$ persons. The respondents for this initial sample were found through contacts with social assistance agencies, medical doctors, and by visiting known meeting points of heroin users. The research bureau tried to obtain a more or less representative sample. The assumption of a Bernoulli sample for the initial sample is rather artificial. However, one may believe that the sampling method, although not probabilistic, yields results that are sufficiently close to those from a Bernoulli sample so that making this assumption will give estimates of v of the correct order of magnitude.

After an extensive interview in which the trustworthiness of the interviewer was made clear to the respondent, the respondent was asked to mention other heroin users in the town of Groningen. These "nominees" were identified with first name, nickname, profession, gender, and age categories. A data base was set up in which it was investigated which nominees could be identified with each other. This is not a straightforward activity since persons can be known by slightly different names, ages of nominees are often not known exactly by the respondents, and job categories also can be reported in different ways. Some choices made in this identification process were rather arbitrary, but they did not seriously affect the results.

With multiplicities, the number of nominations was $t = 311$, of which $r = 15$ were within the initial sample. The number of

nominees was 248, of which $k = 11$ were within the initial sample and the remaining $m = 237$ constituted the first wave. The result-ing estimates with their standard errors are

$$\hat{v}_1 = 685 \quad (171)_M \quad (140)_J$$
$$\hat{v}_3 = 662 \quad (73)_M \quad (63)_J$$
$$\hat{v}_5 = 745 \quad (153)_J$$

The M and J indicate, respectively, whether the standard errors are based on the Bernoulli digraph model or on the jackknife method.

From independent information, the police estimate the number of heroin addicts in Groningen at about 800. If this is likely to be an overestimate, then the estimates found are not unreasonable. If the police estimate is accurate we may conclude that the estimates are of the right order of magnitude, but somewhat low. The low estimates might result from the tendency of informal samples of social networks to overrepresent the "center" and underrepresent the "periphery" of the network, which results in too many nominations within the initial sample, and hence underestimation of v.

The difficulties associated with the identification of the nominees point to differences in the data requirements of the various estimators. For hidden populations characterized by forbidden activities or other social taboos, the information obtained about nominees will usually be restricted. Note that identification of nominees contained in the initial sample will be easier than identifying the others, since the interviews will lead to satisfactory knowledge about the respondents (i.e., those in the initial sample) but less will be known about noninterviewed persons in the first wave. This means that where t is a direct observation, the statistics r and k will be easily calculated but the statistic m will be harder to determine. This implies that $\hat{v}_1$ and $\hat{v}_2$ are more easily obtained than $\hat{v}_3$, $\hat{v}_4$, and $\hat{v}_5$. For the Horvitz-Thompson estimator the picture is bleaker. It requires the in-degrees of individuals in the initial sample and in the first wave. For many nonsymmetric social relations, in-degrees ("how many persons in the population would, if asked, mention you as another population member?") are virtually impossible to obtain reliably. For symmetric relationships in-degrees will be easier to obtain, but obtaining them requires interviews with the persons in the first wave, which is not necessary for the other statistics. As a consequence, $\hat{v}_6$, $\hat{v}_7$, and $\hat{v}_8$ will often be unavailable.

7. Some Simulation Results

The results obtained about consistency and standard errors of the estimators are rather crude asymptotic approximations. Some simulations have been carried out to get an impression of the finite sample/finite population performance of the estimators. For each of the population sizes $v = 100$ and $v = 1000$, three simulation experiments were performed. In each experiment, one digraph was generated from a stochastic model described below, and 1000 Bernoulli samples were generated from the vertices of the digraph; the averaged results for each experiment are presented. The three stochastic models differ in the expected values for σ^2, the variance of the in-degrees. In each of the models the expected in-degree (excluding the self-loop) was fixed at 5. The models are the following:

1. Constant in-degree: for each vertex i, five other vertices are chosen at random (without replacement) to have an arc going to i; here $\sigma^2 = 0$.
2. Bernoulli: all arcs are determined independently, and each ordered pair (i,j) with $i \neq j$ has a probability $5/(v-1)$ for an arc; here $\sigma^2 \approx 5$.

3. Two-block model: vertices are distinguished in two equal size groups and all arcs are determined independently; within the first group arcs have probability $12/(v-2)$, within the second group $6/(v-2)$, and between the two groups $1/v$; here $\sigma^2 \approx 7.25$.

The initial sampling probabilities were $\alpha = .2$ for $v = 100$, leading to $En = 20$, and $\alpha = .06$ for $v = 1000$, leading to $En = 60$. The associated values for the size of the first wave were $m \approx 50$ for $v = 100$ and $m \approx 240$ for $v = 1000$.

In the rare cases that a zero denominator occurred in any of the formulas for the estimators, the 0 was replaced by 1. In the rare cases that a value smaller than $n + m + 1$ was computed as an estimate of v, this value was replaced by $n + m + 1$.

All estimators presented in this paper have been considered in the simulations, but only the main results will be presented. For the two closely related estimators $\hat{v}_1$ and $\hat{v}_2$, and similarly for $\hat{v}_4$ and $\hat{v}_5$, it turned out that conditioning on n is preferable; $\hat{v}_1$ and $\hat{v}_5$ were consistently slightly better than $\hat{v}_2$ and $\hat{v}_4$. Of the Horvitz-Thompson type estimators, $\hat{v}_7$ was clearly better than $\hat{v}_6$ and $\hat{v}_8$. Therefore, results are presented only for the estimators $\hat{v}_1$, $\hat{v}_3$, $\hat{v}_5$ and $\hat{v}_7$. For comparison with $\hat{v}_7$, some results are also presented for the Horvitz-Thompson "estimator" $\hat{v}_{HT}(\alpha)$ although to calculate it the value of α is needed.

Table 1 gives the means and the root mean squared errors of the estimators. There seems to be, on average, a positive bias of the estimators. The ML-estimator (for Model 2) $\hat{v}_3$ is the estimator with the largest bias for Models 1 and 3. This is in accordance with the formulas for its asymptotic mean value, indicating that the bias of $\hat{v}_3$ is a decreasing function of σ^2. Furthermore, it is clear that $\hat{v}_3$ and $\hat{v}_7$ are much better (in terms of root mean squared error) than $\hat{v}_5$, and that the latter is slightly better than $\hat{v}_1$. When comparing $\hat{v}_7$ and $\hat{v}_{HT}(\alpha)$, it can be concluded that the loss due to having to estimate α is evident but modest, except for Model 1 and $v = 1000$.

Table 2 gives the average standard errors which can be compared with the root mean squared errors, and Table 3 gives the estimated coverage probabilities for the confidence intervals, which should be close to .95. The standard errors do not have a large bias as estimators for the root mean squared errors, except for Model 1 with $v = 1000$, where there appears to be a considerable underestimation. However, this underestimated variability of the estimators does not lead to too low coverage probabilities. It seems that, for Model 1 with $v = 1000$, the estimators have heavy-tailed distributions but the standard errors capture the variability in the bulk of the distribution (thus leading to satisfactory confidence intervals) and not the heavy tails.

In the preceding section it was remarked

Table 1. Means and root mean squared errors (between parentheses) for various estimators of v under various models

v	Model	$\hat{v}_1$	$\hat{v}_3$	$\hat{v}_5$	$\hat{v}_7$	$\hat{v}_{HT}$
100	1	105 (25)	111 (16)	101 (19)	108 (15)	100 (12)
100	2	106 (30)	103 (12)	102 (23)	103 (14)	100 (13)
100	3	106 (28)	97 (12)	101 (21)	100 (13)	99 (13)
1000	1	1074 (339)	1176 (238)	1066 (323)	1158 (231)	997 (107)
1000	2	1061 (300)	1017 (130)	1054 (282)	1017 (142)	1000 (115)
1000	3	1071 (294)	969 (123)	1058 (268)	979 (129)	994 (118)

Table 2. Average errors for various estimators of v under various models; SEM *denotes model-based standard error,* SEJ *jackknifed standard error, and* RMSE *root mean squared error*

v	Model	$\hat{v}_1$			$\hat{v}_3$			$\hat{v}_5$		$\hat{v}_7$	
		SEM	SEJ	RMSE	SEM	SEJ	RMSE	SEJ	RMSE	SEJ	RMSE
100	1	23	25	25	14	13	16	21	19	13	15
100	2	25	28	31	13	12	12	23	23	13	14
100	3	25	29	28	12	12	12	24	21	13	13
1000	1	273	302	339	166	165	238	289	323	165	231
1000	2	264	293	300	133	131	130	279	282	136	142
1000	3	271	302	294	125	126	123	286	268	130	129

that, of the estimators under consideration here, $\hat{v}_1$ has the least requirements of the data, followed by $\hat{v}_3$ and $\hat{v}_5$, while $\hat{v}_7$ poses much stronger requirements. The simulations show that $\hat{v}_3$ is much better than $\hat{v}_1$ (and $\hat{v}_5$), so the extra effort needed to collect the data is amply rewarded in this case. It is surprising that even for networks that are quite different from those obtained as outcomes of Bernoulli digraphs, $\hat{v}_3$ is much better than $\hat{v}_1$ and $\hat{v}_5$. The extra effort needed to collect the data for $\hat{v}_7$ is hardly rewarded when this estimator is compared to $\hat{v}_3$.

These simulation results suggest that $\hat{v}_3$ is to be recommended for practical use, possibly in combination with $\hat{v}_5$, and that (for all estimators considered) the confidence intervals based on either of the standard errors are indeed trustworthy. Further research is needed to investigate whether there exist types of networks, occurring in practice, where the bias of $\hat{v}_3$ is a source of concern.

8. Discussion

It may be concluded from this paper that it is possible to use a one-wave snowball sample for estimating a population size. In sociological applications a one-wave snowball sample can sometimes be obtained as a sample of "personal networks," that is, a sample of respondents who report on contacts with others. If the identities of all persons involved are observed, then the personal networks can be combined into a one-wave snowball sample.

The main doubt when applying the results of this paper will often be the validity of the assumption of a Bernoulli initial sample. This assumption may be

Table 3. Coverage relative frequencies for confidence intervals based on model-based (M) or jackknife (J) standard error for various estimators of v

v	Model	$\hat{v}_1$		$\hat{v}_3$		$\hat{v}_5$	$\hat{v}_7$
		M	J	M	J	J	J
100	1	.963	.955	.992	.955	.956	.968
100	2	.948	.959	.980	.963	.958	.949
100	3	.945	.962	.931	.924	.961	.933
1000	1	.965	.967	.941	.926	.962	.934
1000	2	.948	.956	.968	.957	.955	.945
1000	3	.953	.962	.932	.938	.960	.929

approximated to a reasonable extent by using several unrelated sources of contact with the hidden population, taking care of contacting initial sample members not in pairs or larger groups, etc., but the nature of hidden populations will mostly preclude a perfect Bernoulli sampling procedure. To the extent that the initial sample is not Bernoulli, the nature of the social network will often lead to overrepresentation of more central individuals (e.g., those with higher in- and out-degrees) at the expense of more peripheral population members. Since this will tend to increase nominations within the initial sample, a downward bias in the estimators will often be the result. The authors are planning further work where allowance is made for varying selection probabilities in the initial sample, e.g., stratified sampling.

The simulation results demonstrate a surprisingly good performance of the model-based maximum likelihood estimator $\hat{v}_3$ also in the case of networks that are very different from most digraphs produced by a Bernoulli digraph distribution. We propose that the estimators $\hat{v}_3$ and $\hat{v}_5$ be calculated, with their jackknife standard errors; if they do not contradict each other, then $\hat{v}_3$ can be chosen, otherwise it can be suspected that $\hat{v}_3$ has a considerable bias and $\hat{v}_5$ can be used. It must be noted that $\hat{v}_3$ and $\hat{v}_5$ pose stronger requirements concerning the data (identification of nominees in the first wave; see Section 6) than $\hat{v}_1$. There may be situations where the limited possibilities for identifying nominees restrict the research to estimator $\hat{v}_1$.

Intuition as well as the formulae for the estimators (note the important role played by the statistics r, k, and $t - m$) show that the precision of the estimators from the snowball sample depends strongly on the within-initial sample nominations. The sample design should be such that the number of these nominations is not too small; otherwise, very unstable estimates will be produced. This can be quantified by referring to the approximate formula

$$\operatorname{Var} \hat{v}_3 \approx v/\alpha^2\lambda(1 + \lambda/2).$$

Substituting $n \approx v\alpha$ yields

$$\frac{\operatorname{Var} \hat{v}_3}{v^2} \approx \frac{v}{n^2\lambda(1 + \lambda/2)}.$$

If a relative standard error of, e.g., 10% is required, this leads to an initial sample size

$$n \approx \left\{\frac{100v}{\lambda(1 + \lambda/2)}\right\}^{1/2}.$$

Recall that λ is the average degree in the social network. Many social networks are such that the average degree is not much higher than 10 or 12; this means that the initial sample size will have to be larger, in most cases, than the square root of the size of the hidden population.

9. References

Bieleman, B., Diaz, A., Merlo, G., and Kaplan, Ch.D. (1993). Lines Across Europe: Nature and Extent of Cocaine Use in Barcelone, Rotterdam, and Turin. Amsterdam: Swets and Zeitlinger.

Chapman, D.G. (1951). Some Properties of the Hypergeometric Distribution with Applications to Zoological Sample Censuses. University of California Publications in Statistics, 1, 131–159.

Efron, B. (1982). The Jackknife, the Bootstrap and Other Resampling Plans. Philadelphia: SIAM.

Frank, O. (1977). Survey Sampling in Graphs. Journal of Statistical Planning and Inference, 1, 235–264.

Frank, O. (1979). Estimation of Population Totals by Use of Snowball Samples. In Perspectives on Social Network Research, edited by P. Holland and S. Leinhardt, New York: Academic Press, 319–347.

Intraval (1991). Door regelen in de maat. (A study about reactions by drug users on measures taken by civil authorities and civilians.) (In Dutch.) Groningen/ Rotterdam: Intraval.

Johnson, N.L. and Kotz, S. (1969). Distributions in Statistics – 1: Discrete Distributions. New York: John Wiley.

Kalton, G. and Anderson, D.W. (1986). Sampling Rare Populations. Journal of the Royal Statistical Society, Ser. A 149, 65–82.

Snijders, T.A.B. (1981). The Degree Variance: An Index of Graph Heterogeneity. Social Networks, 3, 163–174.

Snijders, T.A.B. (1992). Estimation on the Basis of Snowball Samples: How to Weight? Bulletin de Méthodologie Sociologique, 36, 59–70.

Spreen, M. (1992). Rare Populations, Hidden Populations, and Link-Tracing Designs: What and Why? Bulletin de Méthodologie Sociologique, 36, 34–58.

Received August 1992
Revised September 1993