Journal of
AIDS & Clinical Research

Zhao, J AIDS Clin Res 2017, 8:2
DOI: 10.4172/2155-6113.1000667

Research Article

Open Access

Sampling and Estimation in Hidden Population Using Social Network

Yang Zhao*

Department of Mathematics and Statistics, University of Regina, Canada

Abstract

Characteristics of hidden populations (e.g. population of injection drug users) cannot be studied using standard sampling and estimation procedures. This article considers methods for estimating the population proportion of hidden population using social network. We compare the sampling and estimation technique of respondent-driven sampling with the simplified sampling procedure based on Markov-chain model and discusses the equivalence of these procedures. These procedures fail to provide formulae for estimating the variances of their estimators due to the complexities of their methods. We describe a simplified sampling procedure for collecting data on both the population and its social network, and provide a simple formula to estimate the population proportion efficiently. We further derive a formula to compute an estimate of the variance of the proposed estimator using the delta method. Simulation study is provided to illustrate the new sampling and estimation method.

Keywords: Hidden population; Markov-chain model; Population proportion; Respondent-driven sampling; Social network relationships.

Introduction

Special populations that cannot be studied using standard sampling and estimation procedures are called hidden populations. For example, the populations of injection drug users, men who have sex with men, illegal immigrants, and the homeless. Consistent estimation of the size of these populations are crucial for researchers and policy makers.

Salganik and Heckathorn [1] provide a comprehensive review of sampling and estimation methods for studying hidden populations, including targeted sampling (Watters and Biernacki [2]) and time-space sampling (Muhir et al. [3]). They mention that these methods often fail to provide accurate estimates of the true values. They further point out that a common drawback of most methods is that they fail to use the social network relationships in many hidden populations, that is the network of relationships among the real people in the population, for example the network of friendships. They propose a sampling and estimation method based on a snowball-type sampling design (Coleman [4]), called respondent-driven sampling, which makes use of the social network relationships in a hidden population to collect information from the population of interest such that unbiased estimations of the population characteristics are possible. However, the consistency of their estimator depends on the assumption that individuals are randomly recruited into the study. Other problems with the multi-wave or the snowball-type sampling designs are the costs in regards to both time and money.

To avoid using the multi-wave sampling procedure Zhao [5] introduces a Markov-chain model for estimating the social network relationships. It computes the long run transition probabilities based on the Markov theory, then estimate population proportion using the result of Salganik and Heckathorn [1]. However, none of Salganik and Heckathorn [1] and Zhao [5] provides formulae for estimating the variances of their estimators due to the complexities of their sampling and estimating procedures.

In this article we describe a simplified sampling procedure to collect information on both the population and its social network relationships simultaneously. We derive a consistent estimator of the population proportion of hidden population based on the simplified sampling design. Organization of the rest of the article is as follows. In Section 2 we briefly review the respondent-driven sampling method of Salganik and Heckathorn [1] and the Markov-chain model of Zhao [5], and discuss the equivalence of these two approaches. A simplified sampling and estimating procedure is described in Section 3. A formula for estimating the variance of the proposed estimator is derived. Section 4 provides simulation study to examine the small sample performance of the proposed method. Section 5 gives a brief discussion to conclude the results.

Estimation Methods Using Social Network

Respondent-driven sampling

The main idea of the respondent-driven sampling (Heckathorn [6] Salganik and Heckathorn [1]) is summarized as follows.

Assume that a population is divided into 2 groups A and B, and they are connected by social network relationships, say friendships. Let N_A and N_B be the total number of people in group A and B respectively, $P_A=N_A/(N_A+N_B)$ and $P_B=1-P_A$ be the finite population proportion for group A and B respectively. The object is to estimate P_A and P_B. Let D_{Ai} be the number of friendships of the i^{th} individual in group A. It's also called the degree of the i^{th} individual. The total number of friendships radiating from individuals in group A is

$$R_A=\sum_{i=1}^{N_A} D_{Ai} .$$

Define

$$\overline{D_A}=\frac{R_A}{N_A} \textit{ and } C_{CAB}=\frac{T_{AB}}{R_A} . \qquad (1)$$

Here $T_{AB}=T_{BA}$ is the total number of friendships radiating from members in group A to group B or vice versa. $\overline{D}_A$ is called the average degree of people in group A. Similarly, $\overline{D}_B$ and C_{BA} are defined as

***Corresponding author:** Yang Zhao, Department of Mathematics and Statistics, University of Regina, College West 307.14, Regina, SK, S4S 0A2, Canada; Tel: 306 585-4348; Fax: 306 585-4020; E-mail: zhaoyang@uregina.ca

Received February 07, 2017; **Accepted** February 20, 2017; **Published** February 27, 2017

Citation: Zhao Y (2017) Sampling and Estimation in Hidden Population Using Social Network. J AIDS Clin Res 8: 667. doi: 10.4172/2155-6113.1000667

above. Salganik and Heckathorn [1] show that

$$\mathrm{P}_{\mathrm{A}} = \frac{\overline{D}_B \mathrm{C}_{\mathrm{BA}}}{\overline{D}_A \mathrm{C}_{\mathrm{AB}} + \overline{D}_B \mathrm{C}_{\mathrm{BA}}}. \tag{2}$$

In practice, respondents are selected from the social network based on the respondent driven sampling design of Heckathorn [6], where a small number of initial seeds is selected first, then current seeds randomly recruit other friends into the sample, and the recruiting process continuous until the required sample size is reached. Let r_{AB} be the total number of recruitments from individuals in group A to individuals in group B, r_{AA} be the total number of recruitments from individuals in group A to other individuals in the same group, and the same for r_{BA} and r_{BB} . Based on the random recruitment assumption C_{AB} , C_{BA}, $\overline{D}_A$ and $\overline{D}_B$ can be consistently estimated by

$$\widehat{C}_{AB} = \frac{r_{AB}}{r_{AA}+r_{AB}}, \qquad \widehat{C}_{BA} = \frac{r_{BA}}{r_{BB}+r_{BA}}, \tag{3}$$

$$\overline{d}_A = \frac{n_A}{\sum_{i=1}^{n_A} \frac{1}{D_{Ai}}}, \quad \text{and} \quad \overline{d}_b = \frac{n_B}{\sum_{i=1}^{n_B} \frac{1}{D_{Bi}}},$$

respectively. Here n_A and n_B are the total numbers of individuals selected from groups A and B respectively. Then the population proportions can be estimated by substituting (3) to (2). Salganik and Heckathorn [1] show that these estimators are asymptotically unbiased regardless of how the initial seeds are selected.

The Markov-chain model

An important contribution of Zhao [5] is that they propose a one-wave sampling design to collection information about the population and its social network relationships. In this design selected individuals are required to recruit all their friends into the study, information on how many friendships they have in group A and group B is recorded respectively, and random recruitment assumption is not required. Furthermore they describe a Markov-chain model for the social network relationships. Instead of using groups A and B, they define 2 states, A and B, and it is assumed that each individual is either in state A or B but not both. Suppose individuals are selected using respondent-driven sampling design, and let P_{AB} be the probability that a randomly selected individual in state A will recruit an individual in state B, and P_{AA}=1−P_{AB} be the probability that a randomly selected individual in state A will recruit an individual in state A. Similarly P_{BA} and P_{BB} can be defined as above. Then the transition probability matrix for a first order Markov-chain model can be denoted as

$$P = \begin{pmatrix} P_{AA} & P_{AB} \\ P_{BA} & P_{BB} \end{pmatrix}.$$

Under the condition that P is an ergotic irreducible transition matrix, in the long run the probability that an individual in state A will be selected is

$$\pi_A = \frac{P_{BA}}{P_{BA}+P_{AB}}, \tag{4}$$

and π_B=1-π_A. Then to estimate the population proportion Zhao [5] recommends using the results of Salganik and Heckathorn [1] as

$$\mathrm{P}_{\mathrm{A}} = \frac{\overline{D}_B \pi_{\mathrm{A}}}{\overline{D}_A \pi_{\mathrm{B}} + \overline{D}_B \pi_{\mathrm{A}}} \quad \text{and} \quad P_B = 1 - P_A. \tag{5}$$

In practice to compute estimates of the population proportions P_A and P_B using the above formulae we need to estimate $\overline{D}_A$, $\overline{D}_B$, P_{AB} and P_{BA}. However, if we substitute (4) to (5) directly, we get

$$\mathrm{P}_{\mathrm{A}} = \frac{\overline{D}_B P_{BA}}{\overline{D}_A P_{AB} + \overline{D}_B P_{BA}}. \tag{6}$$

Comparing the estimators in (2) and (6), it is easy to see that P_{AB}, P_{BA} and C_{AB} ,C_{BA} are essentially measuring the same quantities in the two different models, and the two methods are therefore equivalent.

Neither Salganik and Heckathorn [1] nor Zhao [5] provide variance estimators for their estimators because of the complexities of their sampling and estimating techniques. Next we describe a simplified sampling and estimation procedure for estimating P_A and P_B, and the corresponding variances of the estimators.

A Simplified Sampling and Estimating Method

We consider the one-wave sampling design of Zhao [5]. Let A and B represent the two groups A and B in the same settings as Salganik and Heckathorn [1]. We defined new random variables Z_{Ai} and Z_{Bi} which represent the total number of friendships radiating from the i^{th} individual in group A to individuals in group B and the total number of friendships radiating from the i^{th} individual in group B to individuals in group A respectively. Here the within group friendships are ignored. We define

$$\overline{Z}_A = \frac{1}{N_A}\sum_{i=1}^{N_A} Z_{Ai} = \frac{1}{N_A} T_{AB} \quad \text{and} \quad \overline{Z}_B = \frac{1}{N_B}\sum_{i=1}^{N_B} Z_{Bi} = \frac{1}{N_B} T_{BA}, \tag{7}$$

they represent the average degree of associations from group A to group B and from group B to group A respectively. If we treat $\{Z_{Ai} : i = 1, \cdots, N_A\}$ and $\{Z_{Bi} : i = 1, \cdots, N_B\}$ as two sub-populations, then $\overline{Z}_A$ and $\overline{Z}_B$ are the corresponding sub-population means.

As T_{AB} = T_{BA} from (7) we can derive that

$$N_B = N_A \frac{\overline{Z}_A}{\overline{Z}_B}. \tag{8}$$

Substituting (8) to (9)

$$P_A = \frac{N_A}{N_A + N_B}, \tag{9}$$

we obtain

$$P_A = \frac{\overline{Z}_B}{\overline{Z}_A + \overline{Z}_B}, \text{ and } P_B = 1 - P_A. \tag{10}$$

Therefore consistent estimates of P_A and P_B can be obtained if both $\overline{Z}_A$ and $\overline{Z}_B$ can be estimated consistently. We know that $\overline{Z}_A$ and $\overline{Z}_B$ only contain the between group friendships and the within group friendships are completely ignored. The above result indicates (i) consistent estimation of the proportions P_A and P_B can be achieved using only the information of the between group friendships; and (ii) the one-wave (or two-wave) sampling design of Zhao [5] can be further simplified and for the individuals selected in the sample we only need to record the information on how many friendships they have in the other group.

In practice assume that a sample is drawn from a target population with two groups A and B. We will record the total number of friendships radiating to the other groups, Z_{Ai} or Z_{Bi}, for each individual selected from group A or B. Let $\overline{z}_A$ and $\overline{z}_B$ be the corresponding estimators of the sub-population means $\overline{Z}_A$ and $\overline{Z}_B$ respectively, then the proportions P_A and P_B can be estimated as

$$\hat{P}_A = \frac{\overline{z}_B}{\overline{z}_A + \overline{z}_B}, \qquad \hat{P}_B = 1 - \hat{P}_A, \tag{11}$$

and the variances can be estimated using the delta method as

$$\hat{V}ar(\hat{P}_A) = \hat{V}ar(\hat{P}_B) = \frac{\overline{z}_A^2 \hat{V}ar(\overline{z}_B) + \overline{z}_B^2 \hat{V}ar(\overline{z}_A) - 2\overline{z}_A \overline{z}_B \hat{C}ov(\overline{z}_A, \overline{z}_B)}{(\overline{z}_A + \overline{z}_B)^4}, \tag{12}$$

For example if $\overline{z}_A$ and $\overline{z}_B$ are the sample means of simple random samples selected from groups *A* and *B* independently, then we can compute an estimate of the variance as

$$\hat{V}ar(\hat{P}_A) = \hat{V}ar(\hat{P}_B) = \frac{\overline{z}_A^2 S^2{}_B / \mathrm{n_B} + \overline{z}_B^2 S^2{}_\mathrm{A} / \mathrm{n_A}}{(\overline{z}_A + \overline{z}_B)^4} \tag{13}$$

when the finite population correction factors $\frac{N_A - n_A}{N_A}$ and $\frac{N_B - n_B}{N_B}$ can be ignored. Here S_A^2 and S_B^2 are the sample variance for group A and B respectively.

In the appendix (Appendix 1) we show that our proposed estimators for P_A and P_B are equivalent to Salganik and Heckathorn's [1] estimators, however, they are much simplified which allow us to construct a formula to estimate their variances analytically.

Simulation Study

In this section we use simulation study to examine the small sample performance of the proposed sampling and estimation method. We consider the setting similar to that of Salganik and Heckathorn [1].

The numbers of friendships $D'_{Ai}s$ *and* $D'_{Bi}s$ are generated using exponential distribution with means μ_A and μ_B for groups *A* and *B* respectively, and $D'_{Ai}s$ *and* $D'_{Bi}s$ take the closest integer values. let *I* denote the interconnectedness, and $T_{AB}=T_{BA}=I \times min(R_A, R_B)$. We generate data for N_A=3, 000, N_B=7, 000, μ_A=20, μ_B=10, and *I*=0.6. We select simple random sample of size n_A and n_B from group *A* and *B* independently. Equations (11) and (13) are used to estimate P_A and P_B and the corresponding standard errors (*se.'s)*. Table 1 shows the results for estimation P_A based on 10, 000 replications for different sample sizes (n_A, n_B). We see that all the biases are close to 0, the means of *se.'s* are close to the empirical standard deviations (*sd.*), and the 95% coverage probabilities are close to the nominal value. The results indicate that the overall performance of the proposed method is acceptable for practical implementation.

(n_A, n_B)	**Bias**	**Mean of se.'s**	**95%CP**	**Empirical sd.**
(10,10)	-0.0003	0.0341	0.9326	0.0342
(15,15)	0.0005	0.0277	0.9335	0.0281
(20,20)	-0.0001	0.0244	0.9379	0.0246
(25,25)	$0.0^4 26$	0.0216	0.9408	0.0218
(30,30)	-0.0002	0.0197	0.9447	0.0199
(35,35)	-0.0001	0.0186	0.9478	0.0185

Note: $0.0^4 26$=0.000026.

Table 1: Estimation the population proportion (P_A) in small samples.

Discussion

This research describes a simplified sampling and estimation procedure for estimating the population proportion for hidden population. The new method makes significant improvements of Salganik and Heckathorn's [1] methodology by simplifying the formula of Salganik and Heckathorn's [1] estimator, and providing analytic formula for estimating the variance of the proposed estimator. The simplified estimator indicates that consistent estimate of the population proportion does not depend on the information of within group social network relationships, which allows us to further simplify the one-wave sampling procedure of Zhao [5]where the random recruitment assumption is not required.

The new sampling and estimation method is motivated by the initial idea of simplifying the sampling procedure of the respondent-driven sampling in Zhao [5]. They propose the one-wave sampling design where information on the social network relationships is observed completely for each individual selected in the sample and random recruitment is not required. We would expect that the social network relationships can be estimated more efficiently in the new sampling design. However, they fail to supply a new estimator to compute consistent estimates of population proportions, and they eventually use Salganik and Heckathorn's [1] estimator which is functionally complicated and analytic variance estimation is not available.

In applied statistics simple and efficient methods are always respectable. We hope that the proposed methods can be used to improve some studies in epidemiology and social problems.

Acknowledgement

This research was supported by grants from the Natural Sciences and Engineering Research Council of Canada (YZ).

References

1. Salganik MJ, Heckathorn DD (2004) Sampling and estimation in hidden populations using respondent-driven sampling. Sociological Methodology 34: 193-239.
2. Watters JK, Biernacki P (1989) Targeted sampling: Options for the study of hidden populations. Social Problems 36: 416-430.
3. Muhir FB, Lin LS, Stueve A, Miller RL, Ford WL, et al. (2001) A venue-based method for sampling hard-to-reach populations. Public Health Reports 116: 216-222.
4. Coleman JS (1958) Relational analysis: The study of social orgainzation with survey methods. Human Organization 17: 28-36.
5. Zhao Y (2011) Estimating the size of an injecting drug user population. World Journal of AIDS 1: 88-93.
6. Heckathorn DD (1997) Respondent-driven sampling: A new approach to the study of hidden populations. Social Problems 49: 11-34.