Malaysian Journal of Mathematical Sciences 13(S) April: 101–113 (2019)
Special Issue: 3rd International Conference on Mathematical Sciences and Statistics (ICMSS2018)

MALAYSIAN JOURNAL OF MATHEMATICAL SCIENCES

Journal homepage: http://einspem.upm.edu.my/journal

Respondent Driven Sampling Population Proportion Estimators

Bagheri, A. and Saadati, M. *[1]

[1] *National Population Studies and Comprehensive Management Institute, Iran*

E-mail: mahsa.saadati@psri.ac.ir
**Corresponding author*

ABSTRACT

Injection Drug Users (IDU) as hidden population and homeless, and children labor as rare populations are some of well-known examples of elusive populations. Unknown size of these populations is the most important challenge of research in this area. Respondent Driven Sampling (RDS) is a useful chain referral sampling method which has been applying to estimate proportion of these populations for the last two decades. These estimates are representative of whole target population and studies showed that they are asymptotically unbiased and more applicable comparing to the other chain referral sampling methods such as snowball sampling. The main aim of this article is to study three most applicable population proportion estimators, RDS-I, RDS-II, and RDS-SS. To make RDS-I estimator, respondent's degree and probability of selections should be considered. Hansen-Hurwitz estimator of respondents' degree is also used to create RDS-II estimator. Moreover, by assuming known population size and substituting respondents 'degree with simulated inclusion probability, RDS-SS estimator is calculated.

Keywords: Respondent Driven Sampling method, RDS-I estimator, RDS-II estimator, RDS-SS estimator.

1. Introduction

Any information about hidden populations, those are mostly dealing with behaviors which endanger the society health (Verma (2013)), could help improving public health. However, accurate information of these populations usually does not exist. Probability sampling methods such as simple random sampling are useless to achieve representative samples of these populations. These methods need sampling frame and the list of all population members, which often does not accessible for these kinds of populations. Some problems such as having small population size, difficulty of tracing and recognizing target population, and having sensitive behaviors make it impossible for researchers to determine their frame. An alternative method to reach hidden populations is institutional sampling. However, target population members in institutions are not random samples of these populations. So it is not possible to find precise and valid estimates of hidden populations by applying this sampling method (Watters and Cheng (1987)).

There are two most applicable sampling methods, targeted and time-space sampling, to study hidden populations in the literature. Watters and Biernacki (1989) pointed out that targeted sampling is a non-probability technique due to consisting unknown magnitude of sample selection biases. Though, in time-space sampling, time and location frames to sample target population members are identified. Units of this venue-time frame have selected probability and who ever from the target population that entering to the sample are interviewed. But, in this method, making statistical inferences is difficult because of inaccessibility to some venues (Stueve et al. (2001)).

Chain referral sampling methods are also applicable methods to study hidden populations and the most useful one which were first introduced by Goodman (1961) is snowball sampling. Comparing to probability sampling methods, these methods could cover the whole target population and do not counter coverage errors which results in biased estimates. However, the problem of these methods is due to statistical inferences of their estimates (Salganik and Heckathorn (2004)). These methods need initial samples, which known as seeds, and social networks. Selecting seeds non-randomly and facing more recruits from well-defined social networks comparing to isolated ones are two main problems of chain referral sampling methods. So, they are considered as non-probability or convenience sampling methods (Kalton (1983)).

To make unbiased estimates of hidden populations from chain referral sampling methods, Respondent Driven Sampling (RDS) method introduced by Heckathorn (1997, 2002) . Figure 1 presents the sampling procedure of this method which is collected according to chain referral sampling procedure. Seeds or initial samples are the first members of respondent driven samples which have been selected non-randomly from target hidden population and create the zero wave of sampling procedure. Comparing to the other target population members, the seeds are those with well-defined and strong social networks which have a considerable bigger social network size (degree). Each seeds have a specific number of coupons or quota to recruit him/her peer, someone with the same criteria as him/her seed. Thus, seeds make the first wave of this procedure by recruiting peers randomly from their social networks according to the number of their defined coupons. These newly member of RDS build second wave of this procedure by doing the same as their recruiters (seeds) and become new recruiters. This process continues till reaching the desired sample size which should be equilibrium sample that is a stable sample composition. In this way, when peer recruitment proceeds through a sufficiently large number of waves, the composition of the sample stabilizes, becoming independent of the seeds from which recruitment began and thereby overcoming any biases were created by sampling the nonrandom selected of seeds (Lu (2013)).

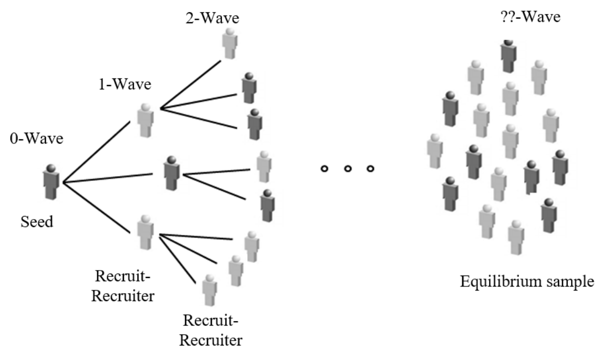

Figure 1: The sampling procedure of RDS method

There are two significant improvements in RDS method compared to the other non-random methods when sampling hidden populations. First, it uses dual incentives to impulse the respondents to recruit more peers into the research, and to improve response rate. Second, unbiased estimates can be obtained by RDS estimators, enabling researchers to draw conclusions for the entire target population from the RDS sample.

Many practices of RDS studies around the world proved that RDS method could be applied efficiently and effectively to study hidden and rare populations (Malekinejad et al. (2008); Johnston et al. (2008)). RDS has been applied to study a variety of hidden populations, such as HIV/AIDS-related high-risk populations (Malekinejad et al. (2008); Johnston et al. (2008)), jazz musicians (Heckathorn and Jeffri (2001, 2003)), visual artist (Jeffri et al. (2011)), regular nightlife users (Calafat et al. (2009)) , young people (Far et al. (2008); Kogan and Brody (2010)), homeless people (Gwadz et al. (2010)) , university students (Wejnert (2010)), migrant worker (Qiu et al. (2011); Jung (2012); Parker and Halpin (2011)), refugees (Burnham et al. (2012)), and immigrants (Castro and Yanez (2012); Song et al. (2012)).

There are also some reviews about RDS method which has been done by Bagheri and Saadati (2014), Bagheri (2015), Bagheri and Saadati (2015a,b), Bagheri and Saadati (2015c), Saadati and Bagheri (2015), Saadati and Bagheri (016a, 2016), Bagheri and Saadati (2017a,b). The main purpose of this article is to study most applicable method to estimate population proportion estimates of hidden populations which is one of the most important challenges of policy makers countering these populations. To achieve this aim, the sections of this article are arranged as follows. The assumptions which should be considered to make the population proportion estimates discussed in the next section. Section 3 is devoted to study three most referenced population proportion estimators for hidden populations. Section 4 presented remarkable conclusions of this study.

2. Matherials and Methods

To calculate population proportion estimates, mathematical models should be built to weight the sample and compensate for the fact of non-randomly selection of the sample. To do so, the following assumptions should be considered:

1. All recruiters in the target population are connected to each other and there is possibility to access someone from his/her own connections.
2. The ties between recruit and recruiters are reciprocal $(i \leftrightarrow j)$.
3. Replacement sampling is considered.
4. Recruiters can accurately report their personal network sizes.
5. Recruiters recruit randomly from their own personal networks.
6. It is assumed that only one coupon is used for recruiting peers though it could be expand to more than one coupon as well.

Given the above assumptions, the selected probability of each node (or structurally respondents) in the next wave when respondent i is selected in previous wave is equal to its degree inversed $\left(\frac{1}{d_i}\right)$. If the vector G represents the successive indices of nodes sampled by the random walk process which is a Markov process, G_k shows the index of the node sampled at the k^{th} wave, and $e_{ij} = 1$ shows the existence of a link between respondents i and j and $e_{ij} = 0$ otherwise, then:

$$Pr(G_{k+1} = j \mid G_k = i) = T_{ij} = \begin{cases} \frac{1}{d_i} & e_{ij} = 1 \\ 0 & e_{ij} = 0 \end{cases} \tag{1}$$

where $d_i = \sum_{j=1}^{N} e_{ij}$ is the degree of node i. Transition matrix also can be written as following:

$$Tr = \begin{bmatrix} 0 & \frac{e_{12}}{d_1} & \dots & \frac{e_{1N}}{d_1} \\ \frac{e_{21}}{d_2} & 0 & \dots & \frac{e_{2N}}{d_2} \\ \vdots & \vdots & \vdots & \\ \frac{e_{N1}}{d_N} & \frac{e_{N2}}{d_N} & \dots & 0 \end{bmatrix} \tag{2}$$

X is a vector $(x_1, x_2, \dots, x_n)$ which shows equilibrium state distribution for any RDS process such that $X^T Tr = X^T$. Considering mutual relationships between i and j respondents and $\sum_{i=1}^{N} x_i = 1$, the unique solution of equilibrium equation is

$$X = \left(\frac{d_1}{\sum_{j=1}^{N} d_j}, \frac{d_2}{\sum_{j=1}^{N} d_j}, \dots, \frac{d_N}{\sum_{j=1}^{N} d_j} \right)^T.$$

So, when a RDS sample reaches equilibrium, the probability that each node to be included in the sample is proportional to its degree. This conclusion is crucial, as it implies that, even collected samples in a non-random manner, RDS sample can be treated as a probability sample such that the inclusion probability of each subject in the sample can be approximated by its degree, which can be used as the sampling weight to generate population estimates.

3. Results and Discussion

In this section by considering the assumptions in Section 2, three most applicable estimators to investigate hidden population proportions will be discussed.

3.1 RDS-I estimator

RDS-I estimator is created based on reciprocal model (Salganik and Heckathorn (2004)).The main aim of this section is to introduce an estimator to estimate the proportion of the population that is made up of two groups of people, A and B (HIV-positive and HIV-negative). If N_A is the population size of group A and D_A is the average social network size or degree of group A; R_A, the total ties of all group A members could be defined as:

$$R_A = \sum_{i \in A} d_i = N_A.D_A \tag{3}$$

The probability of choosing a respondent in group A that randomly has relationship with a respondent in group B where T_{AB} is relation ties between groups A and B members is:

$$S_{AB} = \frac{T_{AB}}{R_A} \tag{4}$$

and also $S_{BA} = \frac{T_{BA}}{R_B}$. Considering mutual relationships between group A and B members results in $T_{AB} = T_{BA}$ and according to Equation (4), it can be resulted that $R_A.S_{AB} = R_B.S_{BA}$. Then Equation (2) results in $N_A.D_A.S_{AB} = N_B.D_B.S_{BA}$. By dividing this Equation to N, $P_A.D_A.S_{AB} = P_B.D_B.S_{BA}$. According to the following Equation, $1 = P_A + P_B$, the population proportion of groups A and B can be calculated as:

$$\begin{aligned} P_A &= \frac{D_B.S_{BA}}{D_A.S_{AB} + D_B.S_{BA}} \\ P_B &= \frac{D_A.S_{AB}}{D_A.S_{AB} + D_B.S_{BA}} \end{aligned} \tag{5}$$

In order to estimate Equation (5), first S_{AB} and S_{BA} must be estimated. Recruitments could be divided to four groups of r_{AA}, recruitment of a person in group A to another person in group A; r_{AB}, recruitment of a person in group A to another person in group B, and so on. Unbiased estimates for S_{AB} could be obtained by replacing r_{AB} instead of T_{AB} and $r_{AA} + r_{AB}$ to R_A in Equation (4). S_{BA} can also be estimated in the same way. In the next step, average degree of groups A and B should be obtained. There are two different approaches of estimating this average degree, considering degree distributions of sample, and population and also applying Hanson-Hurwitz estimators that result in the same estimate as following (Salganik and Heckathorn (2004)).

$$\hat{D}_A = \frac{n_A}{\sum_{i=1}^{n_A} \frac{1}{d_i}} \tag{6}$$

By subtituting the estimations of S_{AB}, S_{BA}, D_A and D_B in Equation (5), the estimations of P_A and P_B could be obtained.

3.1.1 Data smoothing approach

In the case of existing more than two groups in the population, the above process generates overdetermined equations because the number of unknown parameters is less than the number of equations. To study this procedure, three different groups, A, B and C could be considered, so the reciprocal model results in following:

$$\begin{aligned} P_A + P_B + P_C &= 1 \\ P_A.D_A.S_{AB} &= P_B.D_B.S_{BA} \\ P_A.D_A.S_{AC} &= P_C.D_C.S_{CA} \\ P_B.D_B.S_{BC} &= P_C.D_C.S_{CA} \end{aligned} \tag{7}$$

To solve the above equations, linear least squares may be applied, though another approach was proposed by Heckathorn (2007) which called data smoothing. The most important idea behind this method is considering equality of the number of respondents recruited by a group and the number of recruitments of that group. Moreover, reciprocal relations and randomly recruiting of respondents from their networks are assumed.

In this process, if $\hat{E}_A$ is the Markov equilibrium given the transition matrix, and RB is the total number of recruitments in the sample, $R_{AB} = S_{AB}\hat{E}_A RB$. In this way, the transformed recruitment matrix has the same original selection proportions among groups and equal sum of row and column. The next step is to make smoothed recruitment matrix R^{**} when its diagonal elements (R^{**}_{ii}) are equal to $S_{ii}\hat{E}_i RB$ and off-diagonal elements (R^{**}_{ij}) are equal to $\frac{S_{ij}\hat{E}_i RB + S_{ji}\hat{E}_i RB}{2}$ for i and $j = 1, \ldots, n$. The selection properties in Equation (5) are calculated according to R^{**}.So, the equations that cause over determination become redundant and the population proportion estimate can be calculated.

3.2 RDS-II estimator

By considering again two different groups in the sample U, group A with N_A size and group B with N_B size, probability of respondent i in the sample can be obtained by Hansen-Hurwitz estimator (Hansen and Hurwitz (1943), Cochran (1977). In this estimator respondents are weighted by the inverse of the sampling probability, then the proportion of individuals belonging to group

A is defined by (Volz and Heckathorn (2008)):

$$\hat{p} = \frac{\sum\limits_{i \in A \cap U} d_i^{-1}}{\sum\limits_{i \in U} d_i^{-1}} \tag{8}$$

The above estimator is called RDS-II estimator.

3.3 RDS-SS estimator

Gile (2011) developed RDS-SS estimator by considering known population size. This estimator has a similar form of RDS-II if instead of assuming the degree, the inclusion probability of each node is approximated by a series of simulated sampling from an estimated population degree distribution. This estimator can be calculated in the following steps:

Step 1:
if v_t is the number of respondents in the sample with degree i, probability mapping function when it is proportional to k can be calculated as:

$$f^0(k) = \frac{k}{N} \sum_t \frac{v_t}{t} \tag{9}$$

where $f^0(k) : k \to \pi$.

Step 2:
Population proportion degree could be estimated iteratively. The number of individuals with degree k in the population for $i = 1, \ldots, r$ where $f^0(k) = \frac{k}{N} \sum\limits_t \frac{v_t}{f^{i-1}(t)}$ could be estimated as:

$$N_k^i = N \frac{v_k}{f^{i-1}(k)} \tag{10}$$

In this process, the population size N is known. To estimate the inclusion probabilities for nodes from the population of N_k^i, M number of SS-samples of size n from this population should be simulated and then $f^i(k)$ could be calculated as:

$$f^i(k) \approx \frac{U_k + 1}{MN_k^i + 1} \tag{11}$$

where U_k is the total number of observed unites with degree k from the M number of SS-samples.

Step 3:
$f^r(k)$ after r iterations could be used as an approximation of inclusion probability for nodes of degree k. By substituting d_i with $f^r(d_i)$ in the RDS-II estimator, the following population proportion will be calculated:

$$\hat{p}_A = \frac{\sum\limits_{i \in A \cap U} f^r(d_i)^{-1}}{\sum\limits_{j \cap U} f^r(d_i)^{-1}} \tag{12}$$

It is recommended to use $M = 2000$ and $r = 3$ (Gile (2011)). The RDS-SS estimator has shown superior performance with simulations on networks of 1000 nodes with large fraction of sample sizes (over 50% of the network size).

4. Conclusions

Respondent driven sampling is just a two-decade known sampling method for studying hidden populations. These populations are those that are not easy to access but mostly important for policy makers. The main purpose of this study was reviewing different population proportion estimates. RDS-I estimator is made by considering respondent's degree and probability selections while RDS-II is created by respondent's degree Hansen-Hurwitz estimator. RDS-II estimator is easier form comparing to RDS-I. It is important mentioning that as long as data smoothing is applied, RDS-I and RDS-II estimators will coincide. RDS-SS estimator is also calculated by substituting respondent's degree with simulated inclusion probability when population size is assumed to be known. It is important noting that the RDS-SS estimator is dependent on the knowledge of the true population size, which is usually not known for hidden populations. A compromise would be to use this estimator as a sensitivity test method for checking the variation of estimate given a range of population sizes. Tomas and Gile (2011) quoted that to evaluate this estimator where more complex simulation settings were used, for example when RDS was implemented with differential recruitment and non-response rates, RDS-SS estimator failed to outperform other estimators under many situations.

Acknowledgements

The authors would like to acknowledge the following article extracted from a survey under the title of "design and application of respondent driven sampling in the study of hidden populations; design, management and implementation of

respondent driven sampling" which is supported by national population studies and comprehensive management institute in 2016 by the registered number of 11/24333.

References

Bagheri, A. (2015). Respondent driven sampling: A new approach to sampling rare and hidden diseases. *National Population Studies and Comprehensive Management Institute,Tehran, Iran. [In Persian].*, 11(4):753–61.

Bagheri, A. and Saadati, M. (2014). A survey on link-tracing sampling methods (social networks) and its application on demography. *National Population Studies and Comprehensive Management Institute,Tehran, Iran. [In Persian].*

Bagheri, A. and Saadati, M. (2015a). Exploring the effectiveness of chain referral methods in sampling hidden populations. *Indian Journal of Science and Technology*, 8(30):1–8.

Bagheri, A. and Saadati, M. (2015b). Respondent driven sampling statistical inferences to estimate demographical parameters. *National Population Studies and Comprehensive Management Institute,Tehran, Iran. [In Persian].*

Bagheri, A. and Saadati, M. (2015c). Sampling hard to reach populations: Comparing two chain referral approaches. *Proceedings of the 4th International Conference on Computer Science and Computational Mathematics, May 7-8, 2015, Langkawi, Malaysia.*

Bagheri, A. and Saadati, M. (2017a). Management and implementation sampling of injection drug users exposing high risk diseases. *Quarterly Journal of Research on Addiction*, 10(40):27–44.

Bagheri, A. and Saadati, M. (2017b). Population proportion estimator of respondent driven sampling for non-dichotomous variables, data smoothing approach. *Journal of Applied Sciences*, 17(4):196–203.

Burnham, G., Malik, S., Al-Shibli, A. S. D., and et al. (2012). Understanding the impact of conflict on health services in iraq: information from 401 iraqi refugee doctors in jordan. *International Journal of Health Planning and Management*, 27:e51–e64.

Calafat, A., Blay, N., Juan, M., Adrover, D., Bellis, M. A., and et al. (2009). Traffic risk behaviors at nightlife: Drinking, taking drugs, driving, and use of public transport by young people. *Traffic Injury Prevention*, 10:162–169.

Castro, M. C. and Yanez, S. Y. (2012). New forms of sampling for minority and hidden populations: respondent samples conducted in a south american immigrant population. *Universitas Psychologica*, 11:571–578.

Cochran, W. G. (1977). *Sampling techniques.* John Wiley & Sons, Inc, New York.

Far, A. C., Roig, D. A., Jerez, M. J., and Franzke, N. T. B. (2008). Relationship between alcohol, drug use and traffic accidents related to nightlife among a spanish youth sample in three autonomous communities in 2007. *Revista Espanola Salud Publica*, 82:323–331.

Gile, K. J. (2011). Improved inference for respondent -driven sampling data with application to hiv prevalence estimation. *Journal of the American Statistical Association*, 106:135–146.

Goodman, L. (1961). Snowball sampling. *The Annals of Mathematical Statistics*, 32:148–170.

Gwadz, M. V., Cleland, C., Quiles, R., Nish, D., Welch, J., and et al. (2010). Cdc hiv testing guidelines and the rapid and conventional testing practices of homeless youth. *Aids Education and Prevention*, 22:312–327.

Hansen, M. H. and Hurwitz, W. N. (1943). On the theory of sampling from finite populations. *The Annals of Mathematical Statistics*, 14:333–362.

Heckathorn, D. (1997). Respondent-driven sampling: A new approach to the study of hidden populations. *Social Problems*, 44:174–199.

Heckathorn, D. (2002). Respondent-driven sampling ii: Deriving valid population estimates from chain-referral samples of hidden populations. *Social Problems*, 49:11–34.

Heckathorn, D. D. (2007). Extensions of respondent -driven sampling: analyzing continuous variables and controlling for differential recruitment. *Sociological Methodology*, 37:151–208.

Heckathorn, D. D. and Jeffri, J. (2001). Finding the beat: Using respondent-driven sampling to study jazz musicians. *Poetics*, 28:307–329.

Heckathorn, D. D. and Jeffri, J. (2003). Social networks of jazz musicians. changing the beat: A study of the worklife of jazz musicians. *Respondent Driven Sampling: Survey Results by the Research Center for Arts and Culture.Washington DC: National Endowment for the Arts Research Division*, 3:48–61.

Jeffri, J., Heckathorn, D. D., and Spiller, M. W. (2011). Painting your life: A study of aging visual artists in new york city. *Poetics*, 39:19–43.

Johnston, L. G., Malekinejad, M., Kendall, C., Iuppa, I. M., and Rutherford, G. W. (2008). Implementation challenges to using respondent-driven sampling methodology for hiv biological and behavioral surveillance: Field experiences in international settingsy. *Aids and Behavior*, 12:S131–S141.

Jung, M. (2012). Immigrant workers' knowledge of hiv/aids and their sexual risk behaviors: A respondent -driven sampling survey in south korea. *Sexuality and Disability*, 30:199–208.

Kalton, G. (1983). *Introduction to Survey Sampling-Vol. 35 (Quantitative Applications in the Social Sciences*. SAGE Publications, California, ISBN-13: 978-0803921269,Pages: 96, 1st edition.

Kogan, S. M. and Brody, G. H. (2010). Linking parenting and informal mentor processes to depressive symptoms among rural african american young adult men. *Cultural Diversity & Ethnic Minority Psychology*, 16:299–306.

Lu, X. (2013). *Respondent-Driven Sampling Theory, Limitations & Improvements*. PhD thesis, From Department of Public Health Sciences,, Karolinska Institutet, Stockholm, Sweden, 1 edition.

Malekinejad, M., Johnston, L. G., Kendall, C., and et al. (2008). Using respondent -driven sampling methodology for hiv biological and behavioral surveillance in international settings: A systematic review. *Aids and Behavior*, 12:S105–S130.

Parker, P. and Halpin, B. (2011). Eastern european and baltic migrant workers'labour market performances in the republic of ireland. *Research West Review*, 1:1–26.

Qiu, P. Y., Caine, E., Yang, Y., Chen, Q., Li, J., and et al. (2011). Depression and associated factors in internal migrant workers in china. *Journal of Affective Disorders*, 134:198–207.

Saadati, M. and Bagheri, A. (2015). Sampling migrants by respondent driven sampling method. *Proceedings of the International Conference on Migration: Pattern, Consequences and Policies, October 27-28, 2015, Tehran, Iran.*

Saadati, M. and Bagheri, A. (2016a). Respondent driven sampling method comparing to the other sampling methods of hidden populations. *Iranian journal of Epidemiology*, 12(2):63–74.

Saadati, M. and Bagheri, A. (Winter 2016). Unbiased estimator of population proportion for hidden populations exposed to high-risk diseases. *Journal of Health System Research*, 12(4):63–74.

Salganik, M. J. and Heckathorn, D. D. (2004). Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological Methodology*, 34:193–240.

Song, E. Y., Leichliter, J. S., and et al (2012). The use of prescription medications obtained from non-medical sources among immigrant latinos in the rural southeastern u.s. *Journal of Health Care Poor Underserved*, 23:678–693.

Stueve, A., O'Donnell, L. N., Duran, R., and et al (2001). Time-space sampling in minority communities: Results with young latino men who have sex with men. *American Journal of Public Health*, 91:922–926.

Verma, V. (2013). *Sampling elusive populations: applications to studies of child labour*. Geneve: ILO.

Volz, E. and Heckathorn, D. D. (2008). Probability based estimation theory for respondent driven sampling. *Journal of Official Statistics*, 24:79–97.

Watters, J. K. and Biernacki, P. (1989). Targeted sampling:options for the study of hidden populations. *Social Problems*, 36:416–430.

Watters, J. K. and Cheng, Y. T. (1987). Hiv-1 (human immunodeficiency virus-type one) infection and risk among intravenous drug users in san francisco: Preliminary results and implications. contemp. *Drug Problems*, 14:397–410.

Wejnert, C. (2010). Social network analysis with respondent -driven sampling data: A study of racial integration on campus. *Social Networks*, 32:112–124.