RESEARCH ARTICLE

Probability Sampling Method for a Hidden Population Using Respondent-Driven Sampling: Simulation for Cancer Survivors

Minsoo Jung*

Abstract

When there is no sampling frame within a certain group or the group is concerned that making its population public would bring social stigma, we say the population is hidden. It is difficult to approach this kind of population survey-methodologically because the response rate is low and its members are not quite honest with their responses when probability sampling is used. The only alternative known to address the problems caused by previous methods such as snowball sampling is respondent-driven sampling (RDS), which was developed by Heckathorn and his colleagues. RDS is based on a Markov chain, and uses the social network information of the respondent. This characteristic allows for probability sampling when we survey a hidden population. We verified through computer simulation whether RDS can be used on a hidden population of cancer survivors. According to the simulation results of this thesis, the chain-referral sampling of RDS tends to minimize as the sample gets bigger, and it becomes stabilized as the wave progresses. Therefore, it shows that the final sample information can be completely independent from the initial seeds if a certain level of sample size is secured even if the initial seeds were selected through convenient sampling. Thus, RDS can be considered as an alternative which can improve upon both key informant sampling and ethnographic surveys, and it needs to be utilized for various cases domestically as well.

Keywords: Respondent-driven sampling - cancer survivors - social network - hidden population - simulation

Asian Pac J Cancer Prev, **16 (11)**, 4677-4683

Introduction

Parameter estimation can be conducted most precisely when the sample is obtained systematically through proportionally randomized sampling from a population. However, in some cases it is difficult to easily approach the population or to identify its characteristics. For example, groups such as injection drug users (IDUs), HIV infectees, LGBT (lesbian, gay, bisexual, and transgender) sexual minorities, or patients with incurable or intractable diseases which have a high possibility of causing stigma are known as hidden populations (Salganik and Heckathorn, 2004), and they have two characteristics. First, it is difficult to find out the exact size and the boundary of the population since there is no sampling frame. Second, there is a high possibility that the members of the population would not respond truthfully due to the fear of social stigma. The traditional survey method conducted on households has limitations in producing reliable samples from these hidden populations. This issue of sampling was raised in a recent study on cancer survivors as well (Xia and Gustafson, 2012; Jung, 2013; Yoshida et al., 2013; Cheung, 2014; Faghani et al., 2014). It is very important to obtain accurate samples from these populations for effective AIDS prevention and STD intervention.

There have been mainly three methods used for sampling a hidden population: snowball sampling or chain-referral sampling, key informant sampling, and target sampling. Snowball sampling is the most well-known method, and if the researcher's initial contact with the respondents is conducted in an ideal manner, the sample can be obtained by tracking subjects who are close to the researcher (Goodman, 1961). The researcher can expand the sample one after another, like rolling a snowball, as long as the respondents cooperate and consent. However, snowball sampling and these kinds of chain-referral sampling have some problems (Erickson, 1979: 299). First, the estimation on the individual becomes completely dependent on the characteristics of the initial seeds. Therefore, if a population's network is too broad or isolated, then the entire sample can be impacted. Second, chain-referral sampling involves obtaining samples from only those respondents who voluntarily agreed to participate in the survey due to the methodological characteristic, so it inevitably becomes selective sampling and could include a large degree of outlier. Third, chain-referral sampling such as snowball sampling can become a type of "masking" to intentionally avoid responding if the hidden population has strong norms to protect their peers or private interests. Fourth, chain-referral sampling systematically excludes all of the isolates who do not have any reference to sampling. Because of these potential biases, snowball sampling is categorized as a type of convenience sampling. Therefore, there have been many

*Department of Health Science, Dongduk Women's University, Seoul, South Korea *For correspondence: mins.jung@gmail.com*

skeptical perspectives arguing that sampling through an individual's chain-referrals cannot become proportionally randomized sampling.

In key informant sampling, which was proposed to overcome the limitations of snowball sampling, the information of respondents is collected in order to circumvent the respondent bias (Salganik and Heckathorn, 2004). When one attempts to sample a hidden population, for example, their health behavior patterns are identified through relevant occupations such as social workers, drug addiction counselors, civil servants in health departments, etc. This method helps to lessen the exaggeration of the pertinent group's problems and to collect data in a balanced manner. However, there are remaining issues. First, the information bias caused by experts of relevant occupations cannot be controlled and distortion may occur depending on external factors or the media. Second, the core information provided by experts in relevant occupations may be unclear or inaccurate. In addition, if the experts fall under similar systematic conditions, the institutional bias which is commonly shared by core information providers cannot be controlled.

Target sampling emerged as an alternative to the chain-referral sampling method and it is accompanied by two attributes (Watters and Biernacki, 1989). First, in this method, field investigators create a map of the target population. By doing so, the hidden population can be effectively reached and underestimated sampling-which can result when using the traditional approach-is prevented. Second, a predetermined number of samples are collected from the field based on ethnographic mapping. As the ethnographic mapping process becomes more accurate, the life of a hidden population can be better reflected; therefore, the validity and reliability of samples are enhanced. For example, if one can identify the locations of IDUs, where they obtain drugs, and how often they inject themselves, then more accurate samples that reflect the population well can be selected and collected. However, it takes considerable time for researchers to penetrate the target population and the accompanying risk is large. In addition, it is difficult to completely control subjectivity in a qualitative survey (Watters and Biernacki, 1989: 424-426). As a result of searching for other methods of chain-referral sampling, a methodology in social network is utilized. According to social network theory, even an individual who is isolated from the society can be reached within six (6) degrees (Killworth et al., 1998).

The task of improving chain-referral sampling by using the social network method was performed immediately (Frank and Snijder, 1994). After the initial seeds' diverse neighbors were selected, all the members of target populations known to the latter were listed. Here, random walk, which Klovdahl (1989) had applied to the analysis of network structures, was used. This method revealed the structural forms of networks where members of hidden populations are contacted by using snowball sampling in groups consisting only of single respondents (Spreen, 1992). Such a method was used and yielded effects in cases such as the analysis of cocaine users' network structures. Despite the elaboration of chain-referral sampling, however, certain problems remained. In particular, in relation to the sampling and analysis of hidden populations, the problem of extracting the initial respondents randomly was a serious one (Spreen, 1992: 49). This is because even when waves in chain-referral sampling increase, they invariably reflect biases in the initial sampling.

In this context, respondent-driven sampling (RDS) provided innovative methodology. This method was developed so that when sampling hidden populations based on the Markov chain theory and by using incentives based on social networks, the initial respondents' sociodemographic characteristics can yield effects similar to those of randomized proportional sampling independently of samples derived after undergoing certain levels of waves (Heckathorn et al., 1999). In addition, its effectiveness has been empirically demonstrated through AIDS prevention and intervention, Eastern Connecticut Health Outreach projects for drug users, and HIV testing and counseling programs (Heckathorn et al., 2001). The RDS method considerably resolves existing biases because the final samples are independent from the initial respondents' characteristics and are based on respondents' voluntarism. In addition, RDS uses as an incentive behavioral compliance, where a respondent receives a reward for prompting an appropriate individual in his or her social network to participate in the survey (Heckathorn, 1997). Such a method is based on normative sanctions against members of hidden populations. A method where one designates someone else in a hidden population for a reward signifies the receiving of a reward not for one's own participation but for a peer's participation and therefore activates social approval within that hidden population based on peer pressure (Heckathorn, 1997; Heckathorn, 2002). Consequently, such accompanying incentives promote sampling within hidden populations. When primary incentives are selected as instruments of control, it is difficult to derive genuine participation from the participants. However, accompanying incentives lead to the effect of monitoring peers' participation in surveys through the influence of peers (Heckathorn, 1997). Hidden populations often do not respond to material rewards. In such cases, through symbolic rewards, sampling based on the accompanying incentives of social networks turns into pressure to participate.

The properties and sampling procedure of RDS are as follows. First, the researcher recruits a small number of initial respondents, who will serve as the "seeds." Second, the researcher provides an economic incentive so that the initial respondents are inclined to introduce their peers after the survey. When other peers within the hidden population come to the survey location, they receive a material reward. Third, a double incentive is provided if the initial respondents participate in the survey. The incentive is provided for participating in the survey and for introducing their peers. Through this method, the initial respondents expand the sample in a chain-referral form. In order to prevent bias caused by respondents who know a wide range of people, the economic reward is only given to up to three referrals. Fourth, the referred respondent goes through a screening protocol in order to verify that he or she falls under the category of the hidden

population. For example, a referred peer goes through seven screening steps in order to identify whether he or she is a drug user. At the same time, repeated participation is prevented by collecting a minimum amount of identifiers of each individual during the recruitment process. Fifth, the researcher can utilize the social network information in every step of the wave process to reflect the parameters, which will create an incentive that will "steer" the direction of RDS sampling. This means the bonus that will be provided when a specific respondent is recruited for probabilistic sampling. For example, a five-dollar bonus can be provided if female drug users are underestimated. Sixth, the RDS is concluded when the targeted community is saturated.

RDS ultimately utilizes the social network information of the respondents from a hidden population to enable probabilistic sampling (Figure 1). This is distinguishable from snowball sampling, which is a type of convenience sampling. First of all, snowball sampling unilaterally expects the individual to participate in the survey; however, RDS utilizes a double incentive system. One is an incentive for the interviewee (i.e., primary incentive) and the other is for the other people who were recruited for the research (i.e., secondary incentive). This method incorporates a combination of an economic reward and a symbolic reward (for example, an opportunity to protect oneself and the others from critical epidemic). Second of all, unlike snowball sampling, RDS does not ask the participants who their peers are. This is very important when the internal control is strong for the hidden population. For example, if a drug user who was caught in a country where the social sanction on drug use is strong informs on another drug user, it breaks their informal norm, and he or she may become the target of revenge (Frank and Snijders, 1994: 62). In such a situation, RDS reduces masking, and it is because the peer is provided with an opportunity to make the decision about whether to participate or not. Third of all, chain-referral sampling, such as snowball sampling, is prone to bias when there is an individual who has a very wide network of people; however, there is no empirical basis for this assertion (Salganik and Heckathorn, 2004). In other words, no statistical conclusion was made with respect to the size of the social network of the individual and the bias of the population. Rather, RDS can effectively reduce such bias by giving a weight of which amount is determined by the reciprocal of the respondent's social network size.

The present study verifies the effectiveness of RDS, which was developed by Heckathorn and his colleagues, through computer simulation. Based on the Markov chain-referral theory and by using respondents' social network information, RDS makes possible probabilistic sampling when surveying hidden populations. However, although RDS is based on outstanding mathematical statistical theory, it needs to be verified repeatedly under diverse conditions. In particular, to control the sampling process effectively, it is necessary to increase the accuracy of estimation through computer simulation. In addition, even when the initial respondents have been extracted conveniently, it is necessary to demonstrate whether the final samples can actually be sampled completely independently of the initial respondents. We demonstrated through computer simulation whether RDS can be used in research on a hidden population.

Materials and Methods

Study design

This research aims to develop a simulation code for the RDS model, which is widely used in western countries for probabilistic sampling of hidden populations such as LGBT individuals, drug users, or cancer survivors, in order to create a framework to assess the effect of RDS when investigating a hidden population in Korea. To estimate a population from the social network information of the sample, we utilized the reciprocity model of Heckathorn (2002).

Probabilistic estimation of RDS

If the adjacency matrix in a respondent network is X, under the assumption that a direct edge exists from person i to person j, x, which is x_{ij} i.e., J for Jack i then either $x_{ij}=1$ or $x_{ij}=0$. Here, only a reciprocal relationship is considered; therefore, if $x_{ij}=1$, then $x_{ij}=1$. The degree of the adjacency relation of person i can be written as d_i, then it can be defined as $d_i=\sum_i x_{ij}$. The number of relationships that spread from the people in Group A is the sum of the degrees of all people from Group A, and it is defined as follows:

$$R_A=\sum_{i\in A} d_i=N_A D_A \qquad (1)$$

N_A, here, implies the number of people in Group A, and D_A is the average of the degrees of the people in Group A. When a relationship that begins from Group A is randomly selected, the probability of such a relationship to end with people from Group B can be defined as follows:

$$C_{A,B}=\frac{T_{AB}}{R_A} \qquad (2)$$

Here, T_{AB} is the number of networks with one person from Group A and one person from Group B. At this point, we only consider a reciprocal relationship; therefore,

$$R_A C_{A,B}= T_{AB} \qquad (3)$$
$$R_B C_{B,A}= T_{AB} \qquad (4)$$

is obtained from modification (2). The following modification is obtained from modifications (3) and (4), and the definition of R_A and R_B

$$N_A D_A C_{A,B}=N_B D_B C_{B,A} \qquad (5)$$

This modification (5) combines the attributes of each node and the attributes of their networks. If both sides are divided by N, which is the total size of the population, then the following (6) can be obtained.

$$PP_A D_A C_{A,B}= PP_B D_B C_{B,A} \qquad (6)$$
$$PP_A+PP_B=1 \qquad (7)$$

Here, PP_A and PP_B is the ratio of each population that

belongs to Group A and Group B. From modification (6) and modification (7), the following (8) can be obtained.

$$PP_A=\frac{D_BC_{B,A}}{D_AC_{A,B}+D_BC_{B,A}} \quad (8)$$

$$PP_B=\frac{D_AC_{B,A}}{D_AC_{A,B}+D_BC_{B,A}} \quad (9)$$

Modifications (8) and (9) demonstrate that the population ratio, PP_A and PP_B, can be restored only by using the knowledge of the network structure. This modification is valid for all network structures that are characterized by a reciprocal relationship.

Inductive process of approximate unbiased estimation of RDS

PP_A and PP_B can be obtained through a function of network-based estimators. For this, the following prerequisites are necessary. First, when each node recruits other nodes, it randomly recruits other nodes from the networks that it belongs to. Heckathorn (2002) presents empirical evidence that such a hypothesis is logical. Second, in the recruitment process, only sampling with replacement is taken into consideration. Third, networks of hidden populations theoretically have no isolated nodes and have at least one network. Fourth, early respondents are probabilistically sampled in proportion to their neighborhood relationships. In fact, those selected as early respondents from hidden populations are acquainted with the researcher, and relatively more known people like them tend to have relatively more neighbors in comparison with the population average. Under the four prerequisites mentioned above, if initial respondents are sampled in proportion to the number of their respective neighbors, the subsequent sampling procedure likewise will be sampled in proportion to the number of neighbors, and the probability for the neighbor of one node to be sampled at a particular point is probabilistically equal in the networks of the total population (Heckathorn, 2004).

Estimation of $C_{A,B}$ and $C_{B,A}$

Let us hypothesize that the set of networks in each sample is divided into four. In addition, let us define the number of neighbors connected from one person in population A to another person in population A as r_{AA}, and the number of neighbors connected from one person in population A to another person in population B as r_{AB}, respectively. Then estimates of $C_{A,B}$ and $C_{B,A}$ can be obtained as follows:

$$C_{A,B}=\frac{r_{AB}}{r_{AA}+r_{AB}} \quad (10)$$

$$C_{B,A}=\frac{r_{BA}}{r_{BB}+r_{BA}} \quad (11)$$

Estimation of D_A and D_B

Estimates of D_A and D_B can be obtained from the distribution of the neighborhood relationships of the samples or from the Horvitz-Thompson estimator, which is used frequently in the sample estimation theory. The estimators derived from these two equations have the same form. When n_A is the sample size of population A, estimates of D_A and D_B are obtained as follows:

$$\widehat{D_A}=\frac{n_A}{\sum_{i=1}^{n_A}\frac{1}{d_i}} \quad (12)$$

$$\widehat{D_B}=\frac{n_B}{\sum_{i=1}^{n_B}\frac{1}{d_i}} \quad (13)$$

The numerators and denominators in Equations (12) and (13) are equivalent to the Horvitz Thompson estimator, and because they are unbiased estimators, $\widehat{D_A}$ and $\widehat{D_B}$ in Equations (12) and (13) are approximate unbiased estimates.

Estimation of PP_A and PP_B

Because $\widehat{C_{A,B}}$ and $\widehat{C_{B,A}}$ have been obtained from Equations (10) and (11) and $\widehat{D_A}$ and $\widehat{D_B}$ have been obtained from Equations (12) and (13), respectively, it is now possible to estimate PP_A and PP_B. When Equations (10)-(13) are entered in Equations (8) and (9), the following equations are derived:

$$\widehat{PP_A}=\frac{\widehat{D_B}\widehat{C_{B,A}}}{\widehat{D_A}\widehat{C_{A,B}}+\widehat{D_B}\widehat{C_{B,A}}} \quad (14)$$

$$\widehat{PP_B}=\frac{\widehat{D_A}\widehat{C_{A,B}}}{\widehat{D_A}\widehat{C_{A,B}}+\widehat{D_B}\widehat{C_{B,A}}} \quad (15)$$

In the end, Equations (14) and (15) show that RDS can theoretically obtain probability samples of unbiased estimates. Based on such a process of deriving equations, through computer simulation, we show that the RDS estimator presents considerably accurate estimation. In addition, the degree of bias of these equations dramatically decreases in bias as the sample size increases in the n^{-1} order.

Computer simulation of RDS estimation

We conducted a computer simulation by using R ver. 3.1 (Lucent Technologies). The major command set of the simulation is as follows:

n=the number of nodes/subjects

ratio.A=ratio of group A from the population (0<ratio. A<1)

e=average number of edges in each node

p.in.net=possibility that an edge of each node belongs to the same group that consists of the previous node (p.in.net < 0.5 means that each node is connected to the other group of the social network; p.in.net=0.5 means that each node is independent from the initial node; and each node is connected with an internal peer group when p.in.net converges on 1)

pay=compensation level (0 ≤ pay ≤ 1; get more incentive when the value is closer to 1)

n.wave=establish how many waves enable the sampling procedure to continue

start.how=the way of setting initial seeds

(if start.how="rand," extract the sample randomly; if start.how="degree," extract the sample in proportion to the number of neighbors' edges; if start.how="fix," start sampling from the designated node in the result of start. node)

start.node=the beginning number of the node when start.how is "fix"

fix.seed=fixed or non-fixed randomized initial respondent

(if fix.seed=TRUE, then deduce fixed results; if fix.seed=FALSE, then deduce randomized results)

n.seed=the number of initial respondents if fix.seed=TRUE

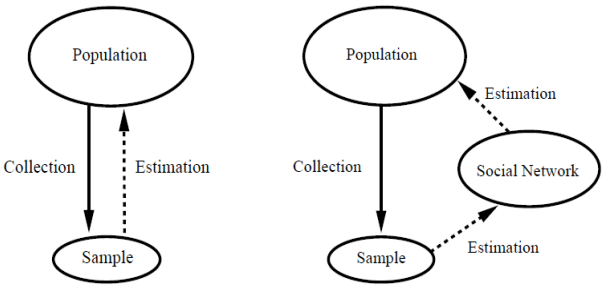
att=matrix that consists of both a series of columns and rows with attributes

(variance is fixed in 1)

Results

We obtained figures pertaining to the variations in the sampling wave path (Figure 2), ratio of group A in the population according to the wave propagation, ratio of group A in the sample, and group attribute value according to the sampling wave (Figure 3). R code was as follows: n=100; ratio.A=0.4; e=5; p.in.net=0.96; pay=1; n.wave=20; start.how="degree"; start.node=20; fix.seed=TRUE; n.seed=23456.

In Figure 1, blue and green refer to the nodes of groups A and B, respectively, and the black edges represent the neighborhood of each node. The red edges represent the RDS sampling wave propagation path in the population network. As configured in the simulation options, there are 100 nodes and 40% of them are in group A (ratio.A=0.4).



(a) Schematic of ordinary sampling and estimation

(b) Schematic of respondent-driven sampling

Figure 1. Comparison of General Sample Population and Respondent-Driven Sampling (RDS)

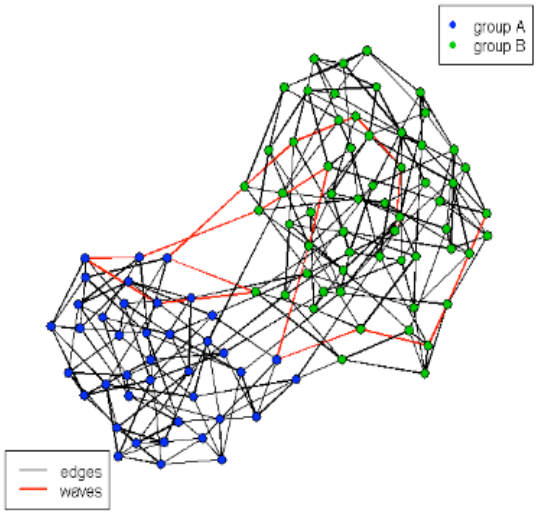


Figure 2. Results of RDS Simulation

An average of 5 neighborhoods exist between the nodes (e=5). Meanwhile, p.in.net has a value close to 1 so it does not have many connections with groups, and since n.wave=20, the progressed wave passed through a total of 20 neighborhoods. It can be observed that the provision of maximum incentive (pay=1) resulted in numerous movements of the wave between groups. If pay=0 is set, a closed network is established where the wave is progressed only within the group.

Figure 2 shows the ratio of group A in the population and the red dotted line refers to the ratio of group A in the population which is 0.4 (ratio.A=0.4). The blue dotted line refers to the real ratio of group A in the sample and the black line refers to the estimated value of group A in the sample. This estimated value used the estimator of Salganik and Heckathorn (2004). The simulation results showed that the estimated value of group A is similar

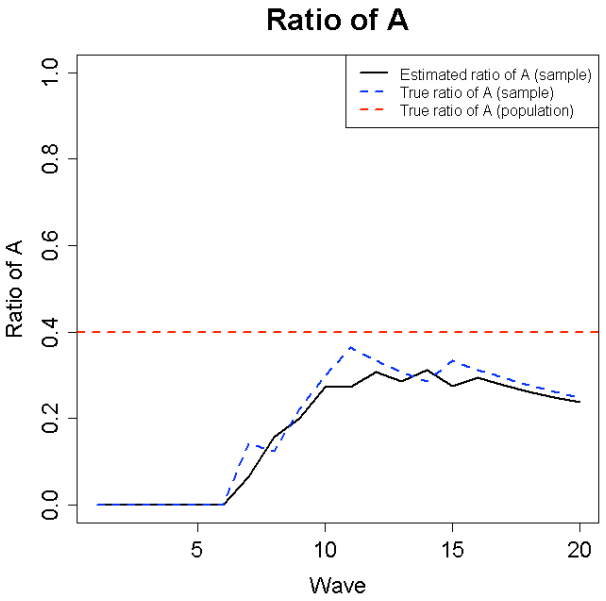


Figure 3. Ratio of A Group in Sampling Simulation

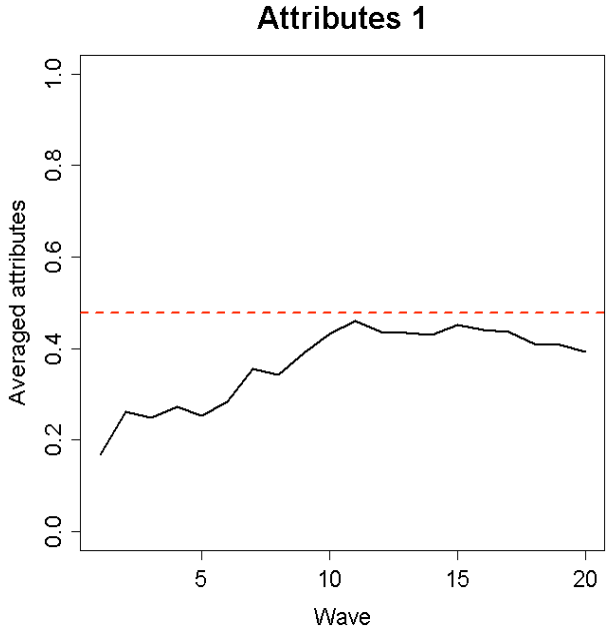


Figure 4. Increasing Patterns of Attribute Variables by Wave Growth in RDS Simulation

to the ratio of group A in the sample, and it was found that the estimated value of group A became an unbiased estimator, approximatively approaching the parameter as the wave increased.

Figure 3 shows how the three properties of the sample configured in the simulation changes according to the wave propagation process. The result is the standardization of the initial value for each property to a minimum of 0 and a maximum of 1. The red horizontal dotted line is the average value of each property in the entire population, and it can be observed that the average value of each property in the 10th wave is similar to the average value of the total population. This is in agreement with the tendency of group A to be similar with the ratio in the population. Consequently, the RDS probability sampling possibility explained by Heckathorn (2004) was proven in the simulation here as the group A ratio approaches the population value at the point in time where the composition of the properties in the sample becomes similar to the property composition of the population.

As described up to now, the unbiasedness of RDS is dependent on the proportional sampling assumption of the initial respondent's neighbors in the wave propagation process. This assumption is relatively satisfied in the field survey. However, this assumption is not perfect; thus, an analysis of its effects is necessary. RDS has the characteristic that the node selection chance in the w-th wave is only dependent on the node selection chance of the previous iteration or the (w-1)-th wave. The Markov chain is a universal approach method of modeling a case like this and the simulation of this study enabled RDS to execute the Markov chain.

Discussion

It was difficult to sample the characteristics of hidden populations such as drug addicts, homeless people, and LGBT individuals exactly through existing statistical sampling methods based on parameter. There was no typical sampling frame in these populations, and random dialing was very ineffective. It was possible to sample some drug addicts or homeless people from medical facilities or rehabilitation facilities, but they cannot represent the whole group. Therefore, target sampling or time-location sampling has been used as alternatives. However, these methods have limitations because they are not types of probabilistic sampling. The method of using a social network suggests a new possibility in the probabilistic sampling of hidden populations. RDS, which uses the information gained from the networks of individuals, enables cost-effective sampling and probabilistic estimation.

The principle of RDS is estimating a population based on the information gained from the sample about the social network. However, RDS needs some supplementations in order to be used in actual social surveys. First of all, RDS is not a type of random sampling, and the impact of the selection of initial seeds on the sampling process is not clear yet. However, a series of researches conducted by Douglas Heckathorn and his colleagues at Cornell University suggest that probabilistic sampling is possible through RDS, and that the estimation is asymptotically unbiased regardless of the respondents (Ramirez-Valles et al., 2005; Abdul-Quader et al., 2006; Iguchi et al., 2009; Semaan et al., 2009; Lansky et al., 2012).

We say that a population is hidden when there is no sampling frame or disclosing the members of the population may pose a potential threat to those members. It is difficult to approach those populations because the standard probabilistic sampling methods have low response rates and fail to obtain honest responses (Sudman et al., 1988). Existing methods such as snowball sampling, chain-referral sampling, target sampling, or key-informant sampling, which have been used to sample these groups, have showed several problems. This study demonstrated the effectiveness of RDS combined with chain-referral sampling and snowball sampling through a simulation which is different from the existing ways. Theoretical analysis of the Markov chain theory and structural incentive system shows that RDS could reduce the bias in sampling, which has been identified as a concern. According to the simulation of this study, we found that the final sample is independent from the initial seeds, even if a specific sampling was conveniently selected from the initial seed group like the existing chain-referral sampling.

RDS has limitations as well. RDS is suitable for sample groups which have the pattern of a network, like chain-referral sampling. The activities of the members need to create combining relationships between them such as a drug user purchasing drugs and sharing them with other addicts or engaging in high-risk sexual behaviors. Therefore, this method is not suitable for a national sample. The size of the territory in which the sample is valid depends on the physical distance of the social network. At the same time, it is impossible to sample a hidden population when no ties exist among the members. Meanwhile, the characteristic which defines the members of a population needs to be obvious in order to apply RDS. Also, peers should not be led to be included in the sample by a false incentive. Therefore, well-verified screening protocols are required. In addition, the network characteristic of respondents should be able to converge on the average of the population through the repetition of enough waves in the sampling process for RDS to be conducted ideally (Jung, 2012).

This study, which uses a new method for sampling a hidden population, has a number of implications. First, RDS seeks to address the stereotype that snowball sampling fails to overcome the bias of the initial seed. If the sampling process could reach equilibrium through enough waves, the final sample becomes independent from the initial seeds. Second, RDS can address the problem of chain-referral sampling of having a bias in favor of the cooperative respondents. An individual who refused to participate in a survey tends to participate in the survey more easily when his or her colleague asks them to do so, due to social pressure. Third, RDS addresses the phenomenon of the "mask" within the hidden population blocking probabilistic sampling. This is possible because an incentive system through a social network has a steering incentive which can sample both social outcasts and those who have many connections evenly. This study revealed

that RDS which uses the information gained through an individual's social network becomes independent form the initial seed before the fifth wave. Also, it could be used in surveys for hidden populations in Korea as well since it is possible to include various attribute variables to the model.

References

Abdul-Quader AS, Heckathorn DD, Sabin K, Saidel T (2006). Implementation and analysis of respondent driven sampling: Lessons learned from the field. *J Urban Health*, **83**, 1-5.

Cheung MR (2014). Surveying and optimizing the predictors for ependymoma specific survival using SEER data. *Asian Pac J Cancer Prev*, **15**, 867-70.

Erickson BH (1979). Some problems of inference from chain data. *Sociol Method*, **10**, 276-302.

Faghani S, Rahmani A, Parizad N, Mohajjel-Aghdam AR, Hassankhani H, Mohammadpoorasl A (2014). Social support and its predictors among Iranian cancer survivors. *Asian Pac J Cancer Prev*, **15**, 9767-71.

Frank O, Snijders TAB (1994). Estimating the size of hidden populations using snowball sampling. *J Offic Stat*, **10**, 53-67.

Goodman L (1961). Snowball sampling. *Ann Mathemat Stat*, **32**, 148-70.

Heckathorn DD (1997). Respondent-driven sampling: A new approach to the study of hidden populations. *Soc Problem*, **44**, 174-99.

Heckathorn DD (2002). Respondent-driven sampling II: Deriving valid population estimates from chain-referral samples of hidden populations. *Soc Problem*, **49**, 11-34.

Heckathorn DD (2011). Snowball versus respondent-driven sampling. *Sociol Method*, **41**, 355-66.

Heckathorn DD, Broadhead RS, Anthony DL, Weakliem DL (1999). AIDS and social networks: Prevention through network mobilization. *Sociol Focus*, **32**, 159 79.

Heckathorn DD, Broadhead RS, Sergeyev B (2001). A methodology for reducing respondent duplication and impersonation in samples of hidden populations. *J Drug Issue*, **31**, 543-64.

Iguchi MY, Ober AJ, Berry SH, et al (2009). Simultaneous recruitment of drug users and men who have sex with men in the United States and Russia using respondent-driven sampling: Sampling methods and implications. *J Urban Health*, **86**, 5-31.

Jung M (2012). Immigrant workers' knowledge of HIV/AIDS and their sexual risk behaviors: A respondent-driven sampling survey in South Korea. *Sexual Disabil*, **30**, 199-208.

Jung M (2013). Cancer control and the communication innovation in South Korea: Implications of cancer disparities. *Asian Pac J Cancer Prev*, **14**, 6121-7.

Killworth PD, McCarty C, Bernard HR, Shelley GA, Johnson EC (1998). Estimation of seroprevalence, rape, and homelessness in the United States using a social network approach. *Evaluat Rev*, **22**, 289-308.

Klovdahl A (1989). Urban social networks: Some methodological problems and possibilities. in Kochen M (ed). the small world (pp. 176-210). Norword, NJ: Ablex Publishing.

Lansky A, Drake A, Wejnert C, et al (2012). Assessing the assumptions of respondent-driven sampling in the national HIV Behavioral Surveillance System among injecting drug users. *Open AIDS J*, **6**, 77-82.

Ramirez-Valles J, Heckathorn DD, Vazquez R, Diaz RM, Campbell RT (2005). From networks to populations: The development and application of respondent-driven sampling among IDUs and Latino gay men. *AIDS Behav*, **9**, 387-402.

Salganik MJ, Heckathorn DD (2004). Sampling and estimation in hidden populations using respondent-driven sampling. In Stolzenberg RM (ed). Sociological Methodology (pp. 193 238). Boston, MA: Blackwell Publishing.

Semaan S, Santibanez S, Garfein RS, Heckathorn DD, Jarlais DC (2009). Ethical and regulatory considerations in HIV prevention studies employing respondent-driven sampling. *Int J Drug Policy*, **20**, 14-27.

Spreen M (1992). Rare populations, hidden populations, and link-tracing designs: What and why? *Bul Methodol Sociol*, **36**, 34-58.

Sudman S, Sirken MG, Cowan CD (1988). Sampling rare and elusive populations. *Science*, **240**, 991-6.

Watters JK, Biernacki P (1989). Targeted sampling: Options for the study of hidden populations. *Soc Problem*, **36**, 416-30.

Yoshida T, Nishijima Y, Hando K, Vilayvong S, Arounlangsy P, Fukuda T (2013). Primary study on providing a basic system for uterine cervical screening in a developing country: analysis of acceptability of self-sampling in Lao PDR. *Asian Pac J Cancer Prev*, **14**, 3029-35.

Xia M, Gustafson P (2012). A Bayesian method for estimating prevalence in the presence of a hidden sub-population. *Stat Med*, **31**, 2386-98.