GLOBAL ACADEMIC RESEARCH INSTITUTE

COLOMBO, SRI LANKA

GARI International Journal of Multidisciplinary Research

ISSN 2659-2193

Volume: 05 | Issue: 02

On 31st October 2019

http://www.research.lk

Author: A.P.M Perera, K.P.A Ramanayake

University of Colombo, Sri Lanka

GARI Publisher | Education | Volume: 05 | Issue: 02

Article ID: IN/GARI/ICMD/2019/114 | Pages: 01-06 (06)

ISSN 2659-2193 | Edit: GARI Editorial Team

Received: 27.09.2019 | Publish: 31.10.2019

ASSESSING SALGANIK-HECKATHORN ESTIMATOE ON POPULATION CHARACTERISTICS

[1]A.P.M Perera, [2]K.P.A Ramanayake

University of Colombo, Sri Lanka

[1]gaayesha93@gmail.com, [2]asoka@stat.cmb.ac.lk

ABSTRACT

Hidden populations are known to be populations that do not have the preference to be discovered in the society. Researchers studying social sciences find these populations very attractive yet extremely difficult to access. Of all the methodologies proposed thus far Respondent Driven Sampling has the highest potential to address these populations. Yet the methodology consumes a greater deal of recourses both monetarily and man power which makes it difficult to do pilot studies in order to figure out the best parameters that should be used in the procedure. Salganik Heckathorn (SH) estimator is one of the acceptable estimates used for the population parameter estimation. Due to its simplicity many researchers favour to use SH estimator. Yet a considerable number of studies highlight the underperformance resultantly denigrating the estimator. This study tries to deflect these discredits by identifies the characteristics of the populations and the sampling combinations the estimator works best. It tries to overlay an open view on the estimator and assist the researchers to use the estimator in a way that would produce credible results.

Keyword: Respondent-Driven Sampling (RDS), Salganik-Heckathorn (SH) estimator, Simulation

INTRODUCTION

Drug users, Prostitutes, HIV infected people, etc. generally fall for the category of hidden populations. They play an important role in social corruption and thereby earns an urgent need of attention. Hidden property of these populations naturally doesn't place room for the existence of a sampling frame. Resultantly conducting statistical studies using probabilistic methodologies become questionable. Non-probabilistic sampling techniques such as convenient sampling, snowball sampling, judgemental sampling, etc. do allow to enter the pool of hidden populations yet does not allow to make credible statistical estimates. Of those introduced thus far Respondent Driven Sampling (RDS) has favourable characteristics to produce reliable estimates. SH estimator is one of the primitive estimators that could be used to estimate the population parameter when RDS is used as the sampling technique. Yet upon introduction of new estimators over the years the performance of SH estimator has been highly discredited. Regardless of these criticisms the popularity of the estimator remains due to its simplicity. The study done by Perera & Ramanayake (2019) proposes an algorithm to generate RDS populations.

Along with the assistance of this algorithm populations were simulated. Performance of SH estimator in each population and sampling technique were inspected with the aim to enlighten the usage of the SH estimator.

Respondent Driven Sampling (RDS)

Heckathorn (1997) first introduced the concept of RDS. From theoretical studies, it has proven that RDS consists of both probabilistic and non-probabilistic characteristics. The non-probabilistic characteristics enable to access these populations while probabilistic characteristics enable to do statistical analysis. The first few respondents are recruited by the researcher and referred to as 'Seeds'. Seeds get a reward for taking part in the study along with coupons for them to recruit individuals into the study. The first few individuals that get recruited by the seeds form the first wave. Recruits did by individuals in the first wave form the second wave. The process goes on until the desired sample size is met. Another important terminology used in RDS is 'degree' or 'network size' and refers to the entire number of individuals the respondent knows in the target population. Researchers has debated on the bias introduced by the mechanism used for seed selection. It can be proven that by adding more waves this bias could be mitigated. Theoretically, six waves would be enough to remove the bias introduced by seed selection (Magnani, et.al 2005).

Salganik-Heckathorn (SH) estimator

The SH estimator is one of the primitive estimators used in RDS. It essentially contemplates the referral pattern, network size, cross relation ties between subgroups of interest. SH estimator uses a two-stage estimation process. First data are used to make inferences about the network structure and use those inferences to make estimations (Wejnert, 2009). As the name itself implies it was introduced by Matthew Salganik and Douglas Heckathorn in 2004 as an attempt to produce asymptotically unbiased estimators for RDS. It should be noted that the estimator can only handle dichotomous response variables. The SH estimator wraps up with a load of assumptions as cited from the work done by Salganik & Heckathorn (2004). They are seeds are selected with proportionate to their degree, all ties reciprocate, that is two people in a tie knows each other, sampling is done with replacement, respondents are accurately aware of their network sizes with individuals having the characteristic of interest, recruiter randomly select peers into the study using the coupons (weak existence of homophily), a respondent receives one coupon and recruit one peer and network of the hidden population forms one connected population By following these assumptions, sampling occurs as a simple random walk. At the state of equilibrium, every respondent has a probability proportionate to their degree of being selected into the sample. Sampling is initiated, that is seeds are selected bearing the statement in mind. If a bias is introduced in the seed selection process, then sampling probabilities do not depend on the degree introducing bias to the final estimations as equilibrium is not met. In order to overcome this situation, studies must consist of long recruitment chains so that both the chain and the recruitment probabilities converge to an equilibrium (Fellows, 2018). It has been clearly stated by Salganik and Heckathorn, random seed selection equally produces asymptotically unbiased estimators for the population parameter. Yet a considerable number of studies put forward the inadequacy of the estimator when estimating the population parameter. This study tries to identify the combinations the estimator works well.

SIMULATIONS

The study makes use of two algorithms to simulate populations and to extract samples. 72 populations are generated by using the population simulation algorithm proposed by Perera & Ramanayake (2019) by changing the distributions for the degree (Skewed right, skewed left), proportion of response (p = 1/3, 1/2 and 2/3), population sizes (N=1000, 5000, 10000), association types (response variable associated with only the Characteristic variable (indicated as C), associated with the Characteristic variable and degree (indicated as CD), associated with only the degree (indicated as D), randomly allocated (indicated as R))

After simulating populations, samples are extracted using 4 combinations of sampling from each population. For this the sampling algorithm proposed by Pathirana & Ramanayake (2017) was used. The combinations are 10 seeds and 2 coupons, 10 seeds and five coupons, 25 seeds and 2 coupons, 25 seeds and 5 coupons

500 samples from each combination were generated in order to get a distribution of the estimates.

RESULTS

Note on the graphical representation:

Instead of using boxplots a circle is used to represent the estimated mean (centre of circle) and to be the relative variance (area of the circle) of the estimations for unambiguousness. Association of the response variable with other variables in a population is represented using colour codes as both degree and character (Blue Circle), degree only (Green Circle), character only (Orange Circle) and random (Red Circle).

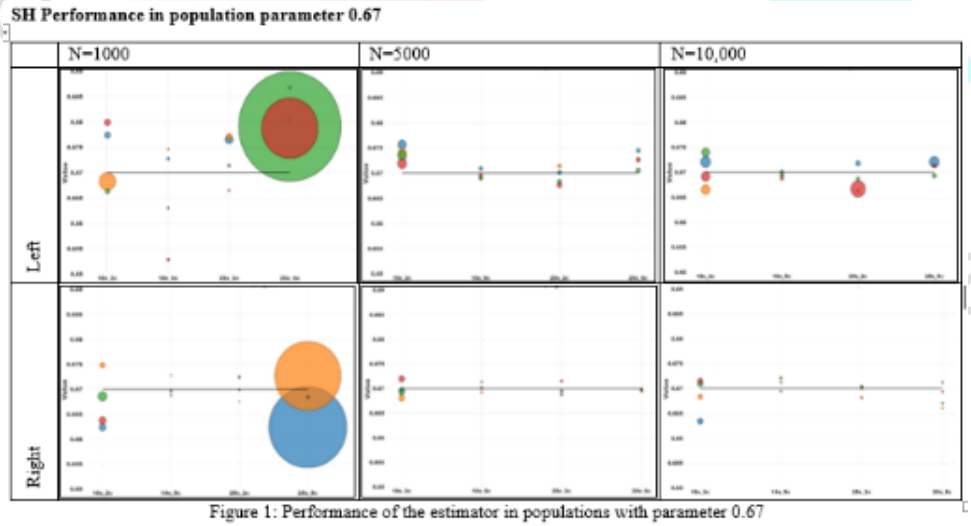

Figure 1: Performance of the estimator in populations with parameter 0.67

With respect to Figure 1, except for the case where the population size is 1000, in all other cases, the estimator works fine. Estimator works considerably well when a moderate amount of seeds and coupons was taken in all cases. When the population size is small and left skewed the best seed coupon combination would be 25 seeds and 2 coupons. In other cases, the SH estimator does not seem to perform pleasingly. A slight improvement in the performance could be seen when the populations are right skewed when compared with it's corresponding left skewed population. When the population is moderate that is around 5000 higher number of seeds and coupons seems to favour the estimation. Estimator seems to perform in a way opposite ways in the corresponding left and right skewed populations. It is clearly seen when the population size is small. In left skewed populations where the variance of the estimator is high has a lower variance when the estimations are done in the corresponding right skewed distribution.

SH Performance in population parameter 0.5

Estimator shows a significant underperformance when the populations are moderate sized and have a left-skewed network size distribution as shown in Figure 2. Yet in the contrary the estimator shows a satisfiable performance when the distribution is right skewed. Estimator works significantly well when the population size in large compared to other sized populations. In higher population sizes high number of seeds and coupons favours the performance of the estimator than in other instances. On the contrary for small populations, higher number of seeds and coupons would not be a good pick. From what is seen in Figure 2 it is best to go for a moderate amount of seeds and coupons especially the population size is small.

In instances where the response variable shows an association with both the degree and the character variable the estimator seems to perform well than in other instances.

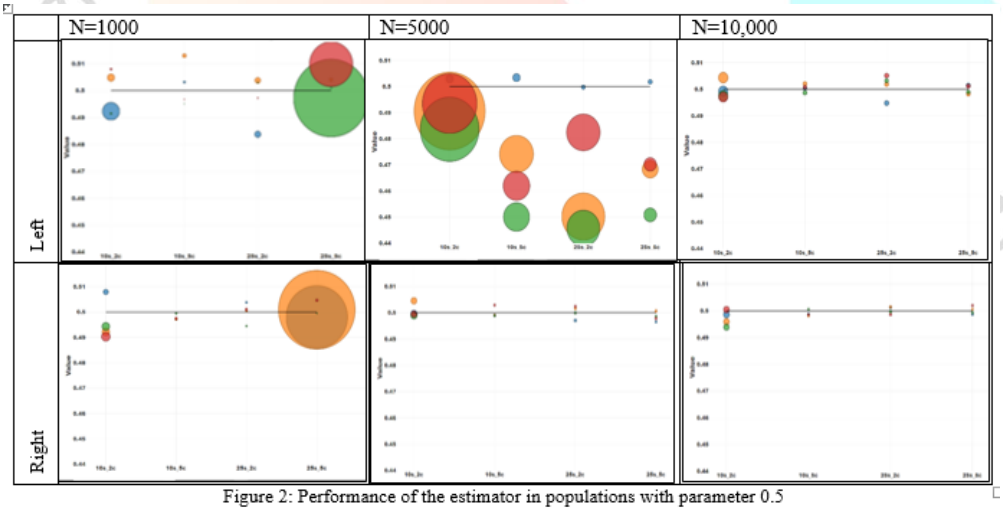

Figure 2: Performance of the estimator in populations with parameter 0.5

SH Performance in population parameter 0.33

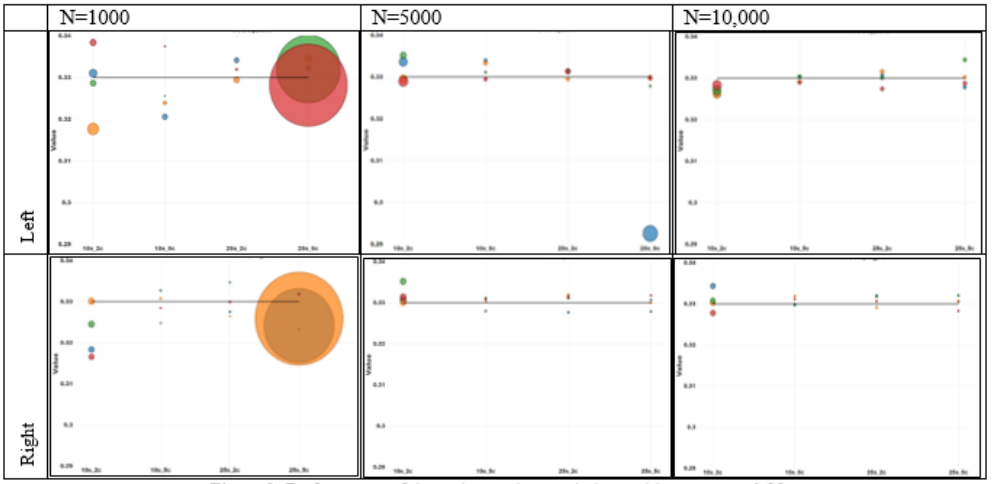

Figure 3: Performance of the estimator in populations with parameter 0.33

The pattern shown in Figure 1 show some similarities such as same variance distributions among populations, similar biases, etc. to those shown in Figure 3. Satisfactory performance could be seen in every seed and coupon combination except for the case in lower population size. Having a lower number of seeds and lower number of coupons does not facilitate for the performance of the estimator, neither does a larger number of seeds and coupons when the population size is small. A significant underperformance is seen when the response variable is proportionate to degree size and categorical variable when the population size is medium, left skewed and the number of coupons and seeds are high

CONCLUSTION

Despite of the many studies highlighting the underperformance of SH estimator, the study highlights situations where maximum performance could be seen. When the population size is very large the estimator performs well regardless of the sampling mechanism. Therefore, for larger populations, SH would be a good pick since the bias and the variance is small. At extreme ends, the estimators perform alike.

When the population size is small, a small number of seeds and coupons and a higher number of coupons and seeds should not be selected. A moderate amount of seed coupon would be preferable. A thorough analysis should be done on the seed coupon combination if to use SH as the estimator. It is best to avoid this estimator at these conditions.

When the population parameter is 0.5 a significant underperformance is seen when the population size is moderate, and the distribution is left-skewed. In such a situation using SH as the estimator may provide unreliable estimates.

It should be noted only 72 different populations are simulated. There are vast ways to extend this study to get a better understanding on the performance of the estimator. With the help of the developed algorithm by Perera and Ramanayake (2019) inspection of the performance of the estimator is way easier. This study

provides insight that SH estimator is not as bad as it is reputed to be. It shows evidence that there are instances where complete studies could be completely relied on this estimator. Yet the researcher needs to identify the properties of the target population before using this estimator since underperformance of the estimator also could be seen. It is always best to do a thorough pre-study before making the decision to use the estimator.

REFERENCES

Fellows, I. (2018). Respondent Driven Sampling and Homophily Configuration Graph. Statistics in Medicine, 1-20.

Magnani, R., Sabin, K., Saidel, T., & Heckathorn, D. (2005). Review of sampling hard-to-reach and hidden populations for HIV surveillance. AIDS, 67-72.

Pathirana, H., & Ramanayake, A. (2017). A Simulation Procedure in Choosing the Wave Length in Respondent Driven Sampling. Proceeding of the International Conference on Computational Modelling and Simulations, Colombo.

Perera, A., & Ramanayake, A. (2019). Assessing the effects of respondent driven sampling estimators on population characteristics. Proceeding of Asia International Conference on Multidisciplinary Research.

Salganik, M., & Heckathorn, D. (2004). Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling. Sociological Methodology, 34, 193 - 239.

Wejnert, C. (2009). An Empirical Test of Respondent Driven Sampling to Migrant Populations: Point Estimates, Variance, Degree Measures and Out of Equilibrium Data. Sociological Methodology, pp.39, 73-116.