UCLA
UCLA Electronic Theses and Dissertations

Title
Testing a New Respondent Driven Sampling Estimator

Permalink
https://escholarship.org/uc/item/3c10q72x

Author
HONG, JI YEON

Publication Date
2016

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA

Los Angeles

Testing a New Respondent Driven Sampling Estimator

A thesis submitted in partial satisfaction

of the requirements for the degree

Master of Science in Statistics

by

Ji Yeon Hong

2016

© Copyright by

Ji Yeon Hong

2016

ABSTRACT OF THE THESIS

Testing a New Respondent Driven Sampling Estimator

by

Ji Yeon Hong

Master of Science in Statistics

University of California, Los Angeles, 2016

Professor Mark Stephen Handcock, Chair

The main purpose of this paper is to review and refine a new respondent driven sampling estimator developed by Ian Fellows and test this estimator by simulation. The previous estimators predicted the proportion of the interesting group appropriately only under strong assumptions such as small sample fraction and no seed bias. However, in reality, these assumptions do not always hold. Therefore we need to develop a new estimator which is not so sensitive to those assumptions. My study starts from the brief idea of Ian Fellows. I clarified and refined his theoretical idea and notation, and tested it under different conditions. By the simulation, I could verify that the new estimator predicts the true proportion of the interesting group better than previous estimators.

This thesis of Ji Yeon Hong is approved.

Hongquan Xu

Rick Paik Schoenberg

Mark Stephen Handcock, Committee Chair

University of California, Los Angeles

2016

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

Respondent driven sampling has been used for studying the underrepresented people such as HIV infected persons, drug users and minority ethnic groups. However there has always been an issue over how well the RDS estimators measure the true proportion of the underrepresented people in population. The previous RDS estimators provide unbiased estimates only when the sample fraction is relatively small compared with populations, when there is not a severe problem of seed bias or seed dependency, and when there is not a strong homophily within groups. However, when one of those strong assumptions is violated, the previous RDS estimators are not able to predict the proportion of underrepresented group properly.

Therefore, the main purpose of this paper is to test the new estimator provided by Ian Fellows (2014). He suggests estimating the average degrees of each group by Gile's (2011) successive sampling process. And he requires to estimate the transition probability from a certain group to another group through these estimated degrees. Since he claimed this idea through his brief notes, I need to clarify and refine his theoretical idea and the notations first. Then I will compare his new estimator with the previous estimators; RDS I, RDS II, and Gile's SS estimator. And by constructing the bootstrap confidence intervals, I will test the unbiasedness and efficiency of the new estimator.

CHAPTER 2

Review on Previous Estimators

Salganik and Heckathorn (2006) developed the RDS I estimate. Different from random sampling, sampled the respondent-driven sampling implies the dependency between recruiters and recruitees, and the frequency of transitions between different groups. So they created an estimator, including the elements which indicate the number of degrees from a certain group to another group.

Before I present the estimator, I need to set up some notations needed for the estimator. Let n be the number of a sample and $n_A$be the number of vertices which belong to group A in the sample. Let δ_i be the degree of individual i, and Let δ_A and δ_x be the average degrees of individual from group A and X respectively. And let δ_U be the average degree of total population. R_{AX} indicates the total number of recruitments from group A to group X. Then the RDS I can be expressed as follows:

$$P_A = \frac{n_A}{\delta_A}\left(\sum_X \frac{R_{AX} n_X}{R_{XA}\delta_X}\right)^{-1} \text{(Eq 2.1)}$$

Volz & Heckathorn (2008) developed the RDS II estimate to adjust the homophily problem when we use respondent driven sampling. RDS II can be expressed as follows:

$$P_A = (\frac{n_A}{\mathrm{n}})(\frac{\widehat{\delta_U}}{\widehat{\delta_A}}) \quad \text{(Eq 2.2)}$$

In the equation 2.2, the left part is estimation for the proportion of the sampled members from group A when we assume a standard random sampling. The right part is correction of network effects. If $\widehat{\delta_U}$ is greater than $\widehat{\delta_A}$, we are under-sampling individuals from group A so we inflate our estimate.

Gile (2011) presented a successive sampling process for respondent driven sampling estimators. From a fixed degree distribution of the sample, she estimates the population distribution of degrees. And by simulating m times of successive sampling, she estimates the inclusion probability of the vertices with different number of degrees respectively. From mapping the degree distribution to its corresponding inclusion probability, she estimates the proportion of a certain group in population via the generalized Horvitz-Thompson estimator.

CHAPTER 3

Reviewing and Clarifying Ian Fellows' Idea on New RDS Estimator

3.1 Large Sample Fraction Problem

Let $d_1, \dots, d_N$ be a set of fixed degrees for a population of size N, and $y_i \in \{0, 1\}$ be a binary outcome measure. Define $g_1 = \{i: y_i = 1\}$ and $g_0 = \{i: y_i = 0\}$. Set θ_1 as a transition probability from g_1 to g_0. And set θ_0 as a transition probability from g_0 to g_1. If we consider mutual relationship between g_1 and g_0, we can identify:

$$\theta_0 = \frac{\theta_1 \sum_{g_1} d_i}{\sum_{g_0} d_i} \quad \text{(Eq. 3.1.1)}$$

If we define $\overline{d_1} = \frac{1}{N_1}\sum_{g_1} d_i$ and $\overline{d_0} = \frac{1}{N_0}\sum_{g_0} d_i$, where N_1 and N_0 are the number of vertices in each group, we may rewrite this as:

$$\theta_0 = \theta_1 \frac{\overline{d_1} N_1}{\overline{d_0} N_0} \quad \text{(Eq. 3.1.2)}$$

Rearranging gives us

$$\bar{y} = \frac{N_1}{N} = \frac{\theta_0 \overline{d_0}}{\theta_1 \overline{d_1} + \theta_0 \overline{d_0}} \quad \text{(Eq. 3.1.3)}$$

This is similar to the previous estimators; RDS I by Salganik & Heckathorn; and RDS II by Volz and Heckathorn. We can rearrange the RDS I estimator as follows:

$$\widehat{P_A} = \frac{n_A}{\delta_A}\left(\sum_X \frac{R_{AX} n_X}{R_{XA} \delta_X}\right)^{-1} = \frac{n_A}{\delta_A}\left(\sum_X \frac{R_A \sigma_{AX} n_X}{R_X \sigma_{XA} \delta_X}\right)^{-1}$$

$$= \frac{n_A}{\delta_A}\left(\sum_X \frac{n_A \sigma_{AX} n_X}{n_X \sigma_{XA} \delta_X}\right)^{-1}$$

$$= \frac{n_A}{\delta_A}\left(\frac{n_A \sigma_{AX}}{\sigma_{XA}} \sum_X \delta_X^{-1}\right)^{-1}$$

$$= \frac{n_A}{\delta_A}\left(\frac{n_A \sigma_{AX}}{\sigma_{XA}} \frac{\sum_{i \in S \cap X} \delta_i^{-1}}{n_X}\right)^{-1}$$

$$= \frac{n_A}{\delta_A}\left(\frac{n_A \sigma_{AX}}{\sigma_{XA}} \frac{\frac{n_X}{\widehat{\delta_X}}}{n_X}\right)^{-1}$$

$$= \frac{\sigma_{XA} \widehat{\delta_X}}{\sigma_{AX} \delta_A} \quad \text{(Eq. 3.1.4)}$$

If we use Fellows' notation, we can translate the set A into g_1, and the set X into g_0. Therefore the equation 3.1.4 can be rewritten as:

$$\bar{y} = \frac{\widehat{\theta_0}\widehat{d_0}}{\widehat{\theta_1}\widehat{d_1}} \quad \text{(Eq. 3.1.5)}$$

Similarly, RDS II estimator can be rewritten as:

$$\bar{y} = (\frac{n_A}{\mathrm{n}})(\frac{\widehat{\delta_U}}{\widehat{\delta_A}})$$

$$= (\frac{n_1}{\mathrm{n}})(\frac{\hat{\hat{d}}}{\widehat{d_1}}) \quad \text{(Eq. 3.1.6)}$$

By comparing the equation 3.1.3 and the equation 3.1.5, we can notice that when we estimate the proportion of our interesting group by RDS I, we inflate the estimate by removing $\widehat{\theta_0}\widehat{d_0}$ from the denominator. If we compare equation 3.3 and 3.6, we can find that RDS II in equation 3.1.6 inflates the estimate when $\hat{\hat{d}}$ is greater than $\widehat{d_1}$, which means when we are under-sampling g_1. Given that RDS sampling is used for estimating the underrepresented group and given that homophily occurs frequently, RDS I and RDS II are approximate approaches for estimating the proportion of hard-to-reach groups. However, these estimators are biased if we compare them with the true value in equation 3.1.3, and the bias gets a significant problem in particular when the sample fraction is large or there is a seed-bias.

3.2 The Review on Ian Fellows' New RDS Estimator

One of the main purposes of this paper is to clarify Ian Fellows' theoretical idea on new estimator and refine some of the notations he used. In his brief note, he claims that we need to develop a new estimator, which can rigorously estimate the average degrees of g_1 and g_0 respectively. And he also suggests to estimate d_1 and d_0 by Gile's successive sampling (SS) process. From these estimated d_1 and d_0, we can estimate the transition probability θ_1 (from g_1 to g_0) and θ_0 (from g_0 to g_1) as well. Then we can treat an RDS sample as the time-ordered sample. Let t_1^i be a binary random variable, which indicates whether a vertex in g_1 is linked to g_0 or not at a time i; the time i indicates each wave of the RDS sample. Since the sample fraction is large, he develops the way to estimate the transition probability in non-sampled vertices. Let r_0 be the total number of edges originating from non-sampled members of g_0 to non-sampled members of g_1. And let r_1 be the total number of edges originating from non-sampled members of group g_1 to g_0. Finally, let n_{out1} be the number of non-sampled vertices, which belong to g_1, and let n_{out0} be the number of non-sampled vertices, which belong to g_0. Then we can write $P(t_1)$ as:

$$P(t_1) = \prod_{i=2}^{n} p\,(\,t_1^i \mid t_1^{i-1}\,) = \prod_{i=2}^{n}\ (\theta_1^{i-1})^{r_1}\ (1-\theta_1^{i-1})^{n_{out1}\widehat{d_{out1}}-r_1} \quad \text{(Eq. 3.2.1)}$$

From the equation 3.2.1, we can estimate θ_1, which maximizes $P(t_1)$ at time i. Let $p(t_1)$ be the probability of a binary random variable at time i, where the observed number of degrees

from g_1 to g_0 in sample is r_1. Then p(t_1) shows a binominal distribution function. For simplification, we can take the log of p(t_1):

$$\ln p(t_1) = r_1 \ln\theta_1 + (n_{out1}\widehat{\overline{d_{out1}}} - r_1) \ln(1-\theta_1) \quad \text{(Eq. 3.2.2)}$$

We can take a derivative of the equation 3.2.2 and set it as zero to maximize p(t_1):

$$\frac{d \ln p(t_1)}{d\theta_1} = \frac{r_1}{\theta_1} - \frac{n_{out1}\widehat{\overline{d_{out1}}} - r_1}{1-\theta_1} = 0 \quad \text{(Eq 3.2.3)}$$

If we solve the equation 3.2.3, we can estimate θ_1 as $\frac{r_1}{n_{out1}\widehat{\overline{dout_1}}}$. Similarly, we can estimate θ_0 as $\frac{r_0}{n_{out0}\widehat{\overline{d_{out0}}}}$.

We can rewrite this estimate as:

$$\theta_1^i = \frac{r_1^{i-1}}{u_1^{i-1} - z_1^{i-1}} \quad \text{(Eq 3.2.4)}$$

$$\theta_0^i = \frac{r_0^{i-1}}{u_0^{i-1} - z_0^{i-1}} \quad \text{(Eq 3.2.5)}$$

u_1^i is the total number of degrees originating from non-sampled vertices in group1, which is equal to $N_1\overline{d_1}$ - $\sum_{j\in s_1^{i-1}} d_j$, and u_0^i is the total number of degrees originating from non-sampled

vertices in group0, which is equal to $N_0\overline{d_0}$ - $\sum_{j\in s_0^{i-1}} d_j$. s_1^i and s_0^i are the sampled vertices in group1 and group0 respectively.

z_1^i is the total number of degrees originating from non-sampled vertices of group1 to the sample vertices in either group. z_0^i is the total number of degrees originating from non-sampled vertices of group0 to the sample vertices in either group. These can be estimated by

$$z_1^i = (1-\theta_1)\, u_1^{i-1} \frac{N_1\overline{d_1} - u_1^{i-1} - c_1^{i-1}}{N_1\overline{d_1}} + \theta_1 u_1^{i-1} \frac{N_0\overline{d_0} - u_0^{i-1} - c_0^{i-1}}{N_0\overline{d_0}} \text{ (Eq 3.2.6)}$$

$$z_0^i = (1-\theta_0)\, u_0^{i-1} \frac{N_0\overline{d_0} - u_0^{i-1} - c_0^{i-1}}{N_0\overline{d_0}} + \theta_0 u_0^{i-1} \frac{N_1\overline{d_1} - u_1^{i-1} - c_1^{i-1}}{N_1\overline{d_1}} \text{(Eq 3.2.7)}$$

As I mentioned above, r_1^i indicates the total number of edges originating from non-sampled members of group g_1 to g_0. r_0^i indicates the total number of edges originating from non-sampled members of group g_0 to g_1. These can be estimated by,

$$r_1^i = \theta_1 u_1^{i-1} \frac{u_0^{i-1}}{N_0\overline{d_0} - c_0^{i-1}} \text{ (Eq 3.2.8),}$$

$$r_0^i = \theta_0 u_0^{i-1} \frac{u_1^{i-1}}{N_1\overline{d_1} - c_1^{i-1}} \text{ (Eq 3.2.9),}$$

c_1^i is the sum of recruiter-recruitee edges incident on each node in group 1. c_0^i is the sum of recruiter-recruitee edges incident on each node in group 0.

To summarize, Ian Fellows' new estimator takes the following processes:

1) Set i to 1 for the first iteration, $\widehat{\theta_1^i}$ and $\widehat{\theta_0^i}$ to the sample fractions and $\hat{y}^0$ (initial value for the proportion of an interesting group) to the RDS I (SH) estimate.
2) Estimate $\widehat{\overline{d_1}}^i$ and $\widehat{\overline{d_0}}^i$ using Gile's successive sampling (SS) process with population sizes of (N-s) $\hat{y}^{i-1}$ and (N-s)(1-$\hat{y}^{i-1}$) respectively. (N is the number of population and s is the number of sample)
3) Using the degrees in step 2, estimate $\widehat{\theta_1^i}$
4) Estimate $\hat{y}^i = \frac{\widehat{\theta_0^i}\overline{d_0}^i}{\widehat{\theta_0^i}\overline{d_0}^i + \widehat{\theta_1^i}\overline{d_1}^i}$
5) If not converged, set i to i+1 and go to step 2.
6) Let the final estimate be $\hat{y}^i$.

CHAPTER 4

Data

In this research, I will use the two datasets in RDS package in R; fauxmadrona and fauxsycamore. These two datasets have different features regarding seed-dependency and sample fraction, which is good for testing the new estimator under different conditions. In both datasets, each vertex can be categorized into two groups; a group with a disease and a group without a disease. And we are interested in estimating the proportion of diseased persons.

As for fauxmadrona, the population is 1000 people in Seattle and the true proportion of diseased persons is 0.20. Fauxmadrona is the respondent driven sample with n=500, so the sample fraction is 50%. In fauxmadrona, out of 10 seeds, only two seeds are diseased persons, so the proportion of diseased persons in seeds is 20%. Therefore, we can say that there is no seed bias or seed dependency in fauxmadrona.

As for fauxsycamore, the population is 715 people in Oxford and the true proportion of diseased persons is 0.20. Fauxsycamore is the respondent driven sample with n=500 so the sample fraction is about 70%. In fauxsycamore, all 10 seeds are diseased persons, so the proportion of diseased persons in seds is 100%. Therefore, we can say that there is an extreme seed bias or seed dependency in fauxsycamore.

Table 4.1 shows a summary statistic on fauxmadrona and fauxsycamore.

Table 4.1: A Summary Statistic on Fauxmadrona and Fauxsycamore

Data	City	Seed Bias	Population Size	Sample Size	Sample Fraction	True Proportion of Diseased Persons
Fauxmadrona	Seattle	No Seed Bias	1000	500	50%	0.2
Fauxsycamore	Oxford	Extreme Seed Bias	715	500	70%	0.2

CHAPTER 5

Analysis

5.1 Comparison between Previous Estimators and the New Estimator

Using the process described in Chapter 3, I estimated the proportion of diseased persons in two datasets. The table 5.1 shows the comparison between the new estimator and the previous estimators. As for fauxmadrona, the New Estimate predicts the proportion of diseased persons accurately. Compared with RDS I, RDS II and Gile's SS estimator, the New Estimate's value is closer to the true value. The sample fraction of fauxmadrona and fauxsycamore is 50% and 70% respectively, which is not small, so that we can say that the new estimate works better under the large sample fraction condition than previous estimators. Furthermore, the new estimator predicts the true value better even under the extreme seed bias condition where all of the ten seeds are diseases people in fauxsycamore. From this result, we can say that our new estimator alleviates the problem of large sample fraction and seed bias compared with previous estimators. The difference between Gile's successive sampling and the New Estimator is that Gile's successive sampling calculates the inclusion probability corresponding to the degree distribution, while the New Estimator calculates the transition probability corresponding to the degree distribution. We can predict that the transition probability is more strongly affected by seed bias rather than inclusion probability, since the transition probability is directly influenced by the degree of seed bias, while the inclusion probability is directly influenced by the transition

probability. So the two steps are needed to estimate the inclusion probability, which makes it difficult to estimate the inclusion probability accurately without calculating the transition probability itself when there is an extreme seed bias. However, by estimating the transition probability directly, the new estimator succeeds in estimating the true value correctly.

Table 5.1: The Comparison between RDS Estimators

Data	True Value	RDS I (SH)	RDS II (VH)	Gile's SS	New Estimator
Fauxmadrona	0.20	0.1592	0.1644	0.1941	0.1961
Fauxsycamore	0.20	0.1087	0.1372	0.1814	0.1859

5.2 Bootstrapping for Respondent Driven Sampling

In respondent driven sampling, we need to use a special bootstrapping procedure, which is different from that in used for random sampling. Salganik (2006) developed a bootstrapping which can be used for respondent driven sampling, and verified that the coverage probability of confidence interval established from this process is higher. He claims that since the respondent driven sampling process generates dependencies in sample selection process, it is required to mimic such a procedure. Different from traditional bootstrapping which resample the observed samples randomly with replacement until the number of a replicate sample reaches to that of original sample, Salganik first divides the sample members into two groups based on how they were recruited. If a person is recruited from the person who belongs to group A, he or she is categorized as A_{rec}. And if a person is recruited from the person who belongs to group B, he or

she is categorized as B_{rec}. The first step of bootstrapping starts with resampling seeds randomly with a uniform selection probability. Then based on the group membership A_{rec} or B_{rec}, we can draw the persons with replacement. For example, if the chosen seed is recruited from group A, we draw with replacement from the set of sample members who were recruited by someone from group A. Then based on the group membership of the newly chosen persons, we sample the persons with replacement from either A_{rec} or B_{rec}. This process is repeated until the number of bootstrap sample reaches to that of original sample. By iterating this whole process with a certain number of time m, we can create m number of replicate samples. And from each of these samples, we can estimate the proportion of our interesting group; m number of estimates is established. Finally, we can construct bootstrap confidence interval from these m estiamtes.

5.3 Bootstrapping Result

I used a specific bootstrap procedure explained in 5.2, to compare the bootstrap confidence intervals of four different respondent driven sampling estiatmors; RDS I, RDS II, Giles'SS and the New Estimator. The table 5.3.1 shows the 95% bootstrap confidence intervals of fauxmadrona with 50 times replicates. The lengths of bootstrap confidence interval for all four estimators are not quite different from each other; it is about 0.13 to 0.14. However RDS I and RDS II do not include the true value 0.20 in their bootstrap confidence intervals. Gile's SS estimate contains the true value in its bootstrap confidence interval, and the mean value of estimates from bootstrapping is almost same as the true value. The New Estimate also contains the true value in its bootstrap confidence intervals. Therefore we can reconfirm that the New Estimator works better than RDS I and RDS II but not much better than Gile's SS estimator.

Table 5.3.1: The Comparison of Bootstrap Confidence Intervals for Fauxmadrona

	RDS I	RDS II	Gile's SS	New Estimator
Lower Bound of 95% Bootstrap CI	0.0808	0.0888	0.1268	0.1178
Mean	0.1592	0.1641	0.1945	0.1874
Upper Bound of 95% Bootstrap CI	0.2303	0.2395	0.2623	0.2561

The table 5.3.2 shows the 95% bootstrap confidence intervals of fauxsycamore with 50 times of iterations. Gile's SS estimator has the shortest length of bootstrap confidence interval, and RDS I and RDS II come next. The New Estimator shows the longest length of bootstrap confidence interval. Therefore, we may conclude that the New Estimator is an inefficient estimator for predicting true proportion. However, the RDS I and RDS II do not include true value 0.2408 in their confidence intervals. Gile's SS estimator contains the true value and the mean value from the bootstrapping is almost close to the true value. The New Estimator contains the true value in its bootstrapping confidence interval. So we can verify that the New Estimator works better than RDS I and RDS II even under large sample fraction and extreme seed bias problem. This result is noteworthy in terms that even under fauxmadrona, which has relatively smaller sample fraction, the confidence interval of RDS I and RDS II did not include true values.

From the table 5.3.1 and 5.3.2, we can conclude that the New Estimator provides us almost unbiased estimate even under the conditions of large sample fraction and extreme seed bias. The New Estimator estimates the true value more accurately than RDS I and RDS II, but

compared with the Gile's SS estimator, it is less efficient and a little bit more biased. However, we cannot generalize this result, since in this study, I only used the two specific data. There should be much more study on figuring out the relationship or the interaction between large sample fraction and seed bias problem. And in order to more accurately compare the different estimates, we also need to study the way to calculate the variance of this New Estimate. As I showed in the previous section 5.1, this study claims that the New Estimator is much less biased than RDS I and RDS II even under the extreme seed bias and large sample fraction problem, and sometimes it can work better than Gile's SS estimator.

Table 5.3.2: The Comparison of Bootstrap Confidence Intervals for Fauxsycamore

	RDS I	RDS II	Gile's SS	New Estimator
Lower Bound of 95% Bootstrap CI	0.0379	0.0718	0.1395	0.1065
Mean	0.1087	0.1455	0.2032	0.1786
Upper Bound of 95% Bootstrap CI	0.1802	0.2191	0.2670	0.2507

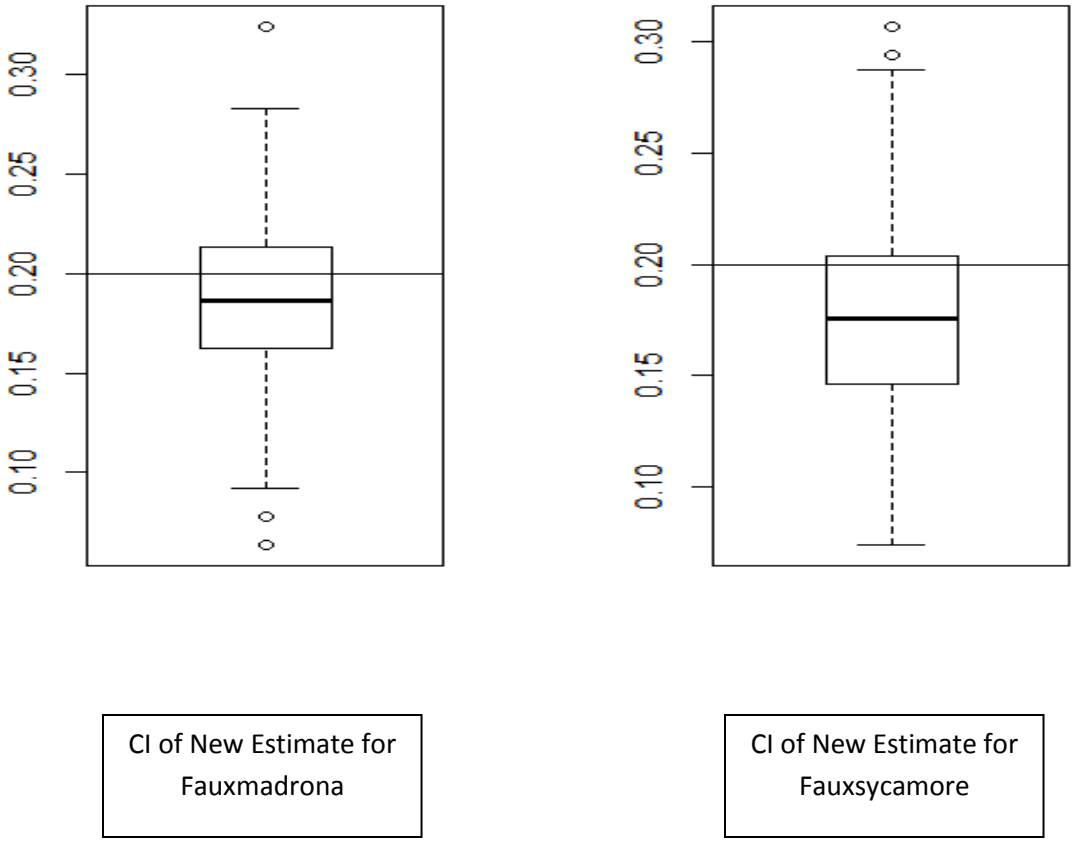

Figure 5.3.1: Boxplots of Bootstrap Confidence Interval for the New Estimate

CHAPTER 6

Discussion

By comparing the estimates of the New Estimator and the previous estimators and by comparing the bootstrap confidence intervals, we could find that the New Estimator predicts the true proportion of underrepresented group in population better than RDS I and RDS II even under the violation of some strong assumptions; large sample fraction and extreme seed bias.

Actually, this study is the opening research for establishing a New Estimator for respondent driven sampling. We need to test the estimator under diverse conditions. In this study, I only examined the goodness of the estimator under large sample fraction and extreme seed bias. And actually these two problems were intertwined with each other in my dataset. So we need to fix one of the conditions as well. Moreover, there are still several strong assumptions when we use the previous estimators, such as with-replacement sampling process and weak homophily. We should test the new estimator under the without-replacement sampling process and strong homophily condition. Furthermore, we also need to build up a theoretical variance of this new estimator to make this estimator as a rigorous one. There are tremendous tasks regarding this New Estimator from strengthening the mathematical theory to simulating it with a variety set of data. This paper just started the first step.

APPENDIX: R-code for Testing a New Estimator

```
########## Establishing a New Estimator

library(RDS)

Ian.estimate<-function(rds.data,init,

N=1000,number.ss.samples.per.iteration=500,number.ss.iterations=5)

{


   g1<-subset(rds.data,subset=(disease==1))

   classes1<-sort(unique(g1$degree))

   nums1<-table(g1$degree)

   weight1<-classes1/sum(g1$degree)

   nrow1=length(classes1)


   g0<-subset(rds.data,subset=(disease==0))

   classes0<-sort(unique(g0$degree))

   nums0<-table(g0$degree)
```

```
weight0<-classes0/sum(g0$degree)

nrow0=length(classes0)



theta<-number.ss.samples.per.iteration/N

theta1<-c(rep(0,number.ss.iterations))

theta0<-c(rep(0,number.ss.iterations))

yhat<-c(rep(0,number.ss.iterations))

prob1<-matrix(0,nrow1,number.ss.iterations)

prob0<-matrix(0,nrow0,number.ss.iterations)

d1<-c(rep(0,number.ss.iterations))

d0<-c(rep(0,number.ss.iterations))

N1<-c(rep(0,number.ss.iterations))

N0<-c(rep(0,number.ss.iterations))

c1<-c(rep(0,number.ss.iterations))

c0<-c(rep(0,number.ss.iterations))

s1<-c(rep(0,number.ss.iterations))

s0<-c(rep(0,number.ss.iterations))
```

```
u1<-c(rep(0,number.ss.iterations))

u0<-c(rep(0,number.ss.iterations))

z1<-c(rep(0,number.ss.iterations))

z0<-c(rep(0,number.ss.iterations))

r1<-c(rep(0,number.ss.iterations))

r0<-c(rep(0,number.ss.iterations))



theta1[1]<-theta

theta0[1]<-theta

yhat[1]<-init

prob1[,1]<-weight1 * sum(nums1/weight1) / ((N-number.ss.samples.per.iteration)*(yhat[1]))

prob0[,1]<-weight0 * sum(nums0/weight0) / ((N-number.ss.samples.per.iteration)*(1-yhat[1]))

d1[1]<-sum((nums1/prob1[,1])*classes1)/((N-number.ss.samples.per.iteration)*(yhat[1]))

d0[1]<-sum((nums0/prob0[,1])*classes0)/((N-number.ss.samples.per.iteration)*(1-yhat[1]))

N1[1]<-(N*theta*d1[1])/(theta*d1[1]+theta*d0[1])

N0[1]<-(N*theta*d0[1])/(theta*d1[1]+theta*d0[1])
```

```
    c1[1]<-nrow(rds.data[rds.data$wave==0&rds.data$disease==1,])

    c0[1]<-nrow(rds.data[rds.data$wave==0&rds.data$disease==0,])

    s1[1]<-0

    s0[1]<-0

    u1[1]<-N1[1]*d1[1]-s1[1]

    u0[1]<-N0[1]*d0[1]-s0[1]

    z1[1]<-((1-theta1[1])*u1[1]*(N1[1]*d1[1]-u1[1]-
c1[1])/(N1[1]*d1[1]))+theta1[1]*u1[1]*(N0[1]*d0[1]-u0[1]-c0[1])/(N0[1]*d0[1])

    z0[1]<-((1-theta0[1])*u0[1]*(N0[1]*d0[1]-u0[1]-
c0[1])/(N0[1]*d0[1]))+theta0[1]*u0[1]*(N1[1]*d1[1]-u1[1]-c1[1])/(N1[1]*d1[1])

    r1[1]<-theta1[1]*u1[1]*u0[1]/(N0[1]*d0[1]-c0[1])

    r0[1]<-theta0[1]*u0[1]*u1[1]/(N1[1]*d1[1]-c1[1])



    for(i in 2:number.ss.iterations){

     theta1[i]<-r1[i-1]/(u1[i-1]-z1[i-1])

     theta0[i]<-r0[i-1]/(u0[i-1]-z0[i-1])

     yhat[i]<-(theta0[i]*d0[i-1])/(theta0[i]*d0[i-1]+theta1[i]*d1[i-1])

     prob1[,i]<-weight1 * sum(nums1/weight1) / ((N-number.ss.samples.per.iteration)*(yhat[i]))
```

```
    prob0[,i]<-weight0 * sum(nums0/weight0) / ((N-number.ss.samples.per.iteration)*(1-yhat[i]))

    d1[i]<-sum((nums1/prob1[,i])*classes1)/((N-number.ss.samples.per.iteration)*(yhat[i]))

    d0[i]<-sum((nums0/prob0[,i])*classes0)/((N-number.ss.samples.per.iteration)*(1-yhat[i]))

    N1[i]<-(N*theta1[i]*d1[i])/(theta1[i]*d1[i]+theta0[i]*d0[i])

    N0[i]<-(N*theta0[i]*d0[i])/(theta1[i]*d1[i]+theta0[i]*d0[i])

    c1[i]<-c1[i-1]+sum(rds.data[rds.data$wave==i-
1&rds.data$disease==1&rds.data$rec.cat==1,]$degree)

    c0[i]<-c0[i-1]+sum(rds.data[rds.data$wave==i-
1&rds.data$disease==0&rds.data$rec.cat==0,]$degree)

    s1[i]<-s1[i-1]+sum(rds.data[rds.data$wave==i-1&rds.data$disease==1,]$degree)

    s0[i]<-s0[i-1]+sum(rds.data[rds.data$wave==i-1&rds.data$disease==0,]$degree)

    u1[i]<-N1[i]*d1[i]-s1[i]

    u0[i]<-N0[i]*d0[i]-s1[i]

    z1[i]<-((1-theta1[i])*u1[i]*(s1[i]-c1[i])/(N1[i]*d1[i])-c1[i])+theta1[i]*u1[i]*(s0[i]-
c0[i])/(N0[i]*d0[i]-c0[i])

    z0[i]<-((1-theta0[i])*u0[i]*(s0[i]-c0[i])/(N0[i]*d0[i])-c0[i])+theta0[i]*u0[i]*(s1[i]-
c1[i])/(N1[i]*d1[i]-c1[i])

    r1[i]<-theta1[i]*u1[i]*u0[i]/(N0[i]*d0[i]-c0[i])
```

```
    r0[i]<-theta0[i]*u0[i]*u1[i]/(N1[i]*d1[i]-c1[i])
    }
    return(yhat[5])
}

Ian.estimate(new.fauxmadrona,0.1592)
Ian.estimate(new.fauxsycamore,0.1087)

########## Creating Bootstrap Data and Calculating Bootstrap Intervals
recruiter.category<-function(rds.data,seed.count){
rc<-c(rep(0,nrow(rds.data)))
rc[1:seed.count]<-NA
for (i in (seed.count+1):nrow(rds.data)){
if (fauxmadrona[which(rds.data$id==rds.data$recruiter.id[i]),4]==1){
rc[i]<-1}
if (fauxmadrona[which(rds.data$id==rds.data$recruiter.id[i]),4]==0){
rc[i]<-0
```

```
}

}

return(rc)

}



rec.cat<-recruiter.category(fauxmadrona,10)

new.fauxmadrona<-cbind(fauxmadrona,rec.cat)

new.fauxmadrona<-as.rds.data.frame(new.fauxmadrona,network.size="degree")



rec.cat<-recruiter.category(fauxsycamore,10)

new.fauxsycamore<-cbind(fauxsycamore,rec.cat)

new.fauxsycamore<-as.rds.data.frame(new.fauxsycamore,network.size="degree")



resampling<-function(rds.data,n){

resamp.data<-rds.data

resamp.data[1,]<-rds.data[sample(nrow(rds.data),1),]

for (j in 2:n){
```

```
 if (resamp.data[j-1,4]==1){
 resamp.data[j,]<-rds.data[sample(which(rds.data$rec.cat==1),1),]
}
 if (resamp.data[j-1,4]==0){
 resamp.data[j,]<-rds.data[sample(which(rds.data$rec.cat==0),1),]
}
}
return(resamp.data)
}

resampling(new.fauxmadrona,500)
resampling(new.fauxsycamore,500)

boot.ci<-function(rds.data,N){
est<-rep(0,N)
init<-rep(0,N)
boot.data<-rep(list(rds.data),N)
```

```
for (i in 1:N){

boot.data[[i]]<-resampling(rds.data,500)

init[i]<-RDS.I.estimates(rds.data,"disease")$estimate[[2]]

est[i]<-Ian.estimate(boot.data[[i]],init[i])

}

boxplot(est)

abline(h=0.2)

list(quantile(est, probs = c(2.5,97.5)/100),mean(est))

}



par(mfrow=c(1,2))



boot.ci(new.fauxmadrona,500)

boot.ci(new.fauxsycamore,500)
```

REFERENCES

[1] Fellows, Ian E. (2014), A New Estimator for RDS, a manuscript.

[2] Gile, Krista J. (2011). Improved Inference for Respondent-Driven Sampling Data with Application to HIV Prevalence Estimation, *Journal of the American Statistical Association*, vol 106(493): pp.135-146.

[3] Gile , Krista J. and Hancock, Mark S. (2009). Respondent-Driven Sampling: An Assessment of Current Methodology. *Sociological Methodology*, vol 40(1): pp.285-327.

[4] Salganik, Matthew J. (2006). Variance Estimation, Design Effects, and Sample Size Calculations for Respondent-Driven Sampling. *Journal of Urban Health: Bulletin of the New York Academy of Medicine*, vol 83(7): pp. 98-112.

[5] Salganik, Matthew J. and Heckathorn, Douglas D. (2004). Sampling and Estimation in hidden populations using respondent-driven sampling. *Sociol Method*, vol 34: pp.193-239.

[6] Volz, Erik and Hecktahorn, Douglas D. (2008) Probability Based Estimation Theory for Respondent Driven Sampling, *Journal of Official Statistics*, vol 24 (1): pp. 79-97.