Modelling the variance of respondent-driven sampling using *Reddit*

Wilson, Samuel B.Sc. Student; Smith, Aaron, Ph.D

Introduction

Respondent-driven sampling (RDS) is a technique that can be used for estimating disease prevalence in hidden populations. Characteristically, RDS is:

- Used for populations that lack *sampling frames*
  - *EG – The HIV community*
- Collected via a *Snowballing* method, where current members of the sample recruit new ones
  - Problem: Members tend to recruit their *friends*

Using data collected from *Reddit,* an online social networking service, small "infected" sample sets will be simulated from a large population, creating the *Reddit* RDS model.

It can be difficult to estimate the variance of an RDS estimate for the whole population. For estimation, sample variances will be compared to past work and the *New York Jazz* RDS sample set.

Methodology

The information on *Reddit* is broken up into multiple 'subreddits', where users can post. Each of these 'subreddits' represent a segregated group in the RDS model. In taking author 'username' of the top 900 posts from five different 'subreddits', an adequate sample to represent a human population of diverse social and ethnic groups can be determined. When the same 'username' appears in multiple groups, this can be considered an 'edge' between the two.

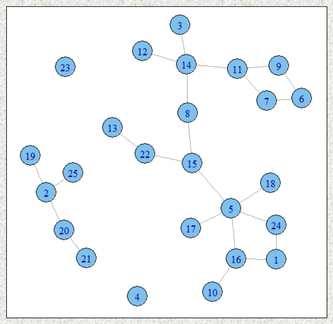

Figure 1: A Respondent-driven sample diagram. Each blue circle represents one recruit, and each line is an edge between them.

Sub A	Sub B	Edges (of 900)
1	2	28
1	3	3
1	4	9
1	5	5
2	3	2
2	4	15
2	5	11
3	4	11
3	5	1
4	5	10

Figure 2: The number of edges between each subreddit. This number is taken out of a possible 900 members.

Using this data, we are able to use mathematical models created for RDS sampling to determine the variance of the *Reddit* dataset, and the uniform normal distribution. This ratio:

$$D(G) = Var[Reddit] : Var[Uniform] \mid D(G) = Degree\ Factor$$

It is through analyzing changes in the degree factor at different 'bottlenecks' in our sample that an estimator for the degree factor will be obtained.

Results

Analysis on the *Reddit* sample produced results, with focus on the deviation of degree variance:

1. As the number of number of people one recruitee can recruit into the population increases, the deviation from the degree of variance increased.
2. The last recruitee recruited by one recruitee had a smaller deviation on their degree than the first person recruited.
3. The mean degree of a sample remained the same throughout all members. The mean degree of 100 was observed for all independent variables.

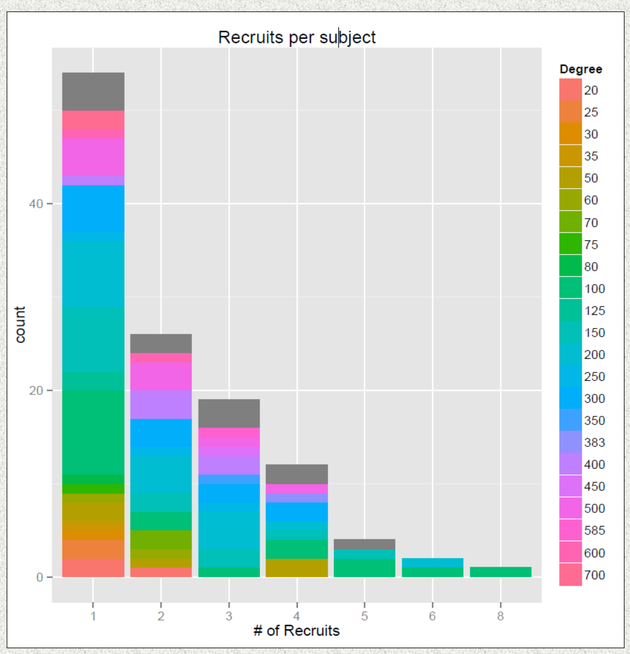

Figure 3: The gradient of degree for the *Reddit* RDS sample. A value of 100 indicates the mean degree of the sample. This sample is a measure of the *Reddit* data set, compressed to the size of the *Ney York Jazz ˜representation.* This allows for increased prediction accuracy and precision.

Conclusion

The *Reddit* Respondent-driven sample showed similar trends to the *New York Jazz* sample set. The *Reddit* data set has a much higher sample size (n = 4500) than the *New York Jazz* (n = 275).

This confirms that the *Reddit* RDS is an accurate model for common Respondent-driven sample populations.

The next step in this area of research is to attempt to overcome some of the error present due to bottlenecking.

- Catastrophic for large samples.

Acknowledgements & Contact Information

Thank you to those who have aided in this research question.

- Credit behind the mentorship of this project is due to *Dr. Aaron Smith*, Ph.D Mathematics, University of Ottawa.
- The University of Ottawa and Undergraduate Research Office for sponsoring the *Undergraduate Research Opportunity Program* Baccalaureate.

Contact Information:

- *Samuel Wilson*, Student B.Sc/B.ScE Biotechnology, 2018, Faculty of Science
- SWILS081@uottawa.ca

References

- **Goel, S., & Salganik, M. J. (2009). Respondent-driven sampling as Markov chain Monte Carlo. *Statistics in Medicine , 28* (17), 2202-2229.**
- **Heckathorn, D. (1997). Respondent driven sampling: A new approach to the study of hidden populations. *Social Problems , 44*, 174-199.**
- **Heckathorn, D., & Volz, E. (2008). Probability Based Estimation Theory for Respondent Driven Sampling. *Journal of Official Statistics , 24* (1), 79-97.**
- **Volz, E., Wejnert, C., Cameron, C., Spiller, M., Barash, V., Degani, I., et al. (2012). Respondent-Driven Sampling Analysis. *Cornell University* .**
- ***http://www.reddit.com/***

uOttawa