Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *doi: 10.18637/jss.v000.i00*

The chords R Package- A Principled Approach to Respondent Driven Sampling

Jonathan D. Rosenblatt
Ben Gurion University

Yakir Berchenko
Gertner Institute for Epidemiology and Health Policy Research

Simon D. Frost
Cambridge University

Abstract

The abstract of the article.

Keywords: RDS, estimation, counting-process, R.

1. Introduction

As the name suggests, Respondent Driven Sampling (RDS) is a framework for sampling by chain-referral. RDS is a bundle of a sampling mechanism and analysis methods, most common in the study of marginalized populations which do not lend themselves to simple sampling (Heckathorn 1997, 2002).

In RDS seeds are selected – usually by convenience – from the target population, and given coupons. They use these coupons to recruit others, who themselves become recruiters. Recruits are given an incentive, usually money, for taking part in the survey, and also for recruiting others. This process continues in recruitment waves until the survey is stopped, usually when a target sample size is reached.

With the above sampling mechanism, highly connected individuals will be overrepresented in the sample. If the attribute of interest is correlated with an individual's degree, as is often the case (e.g. HIV), naïveestimates will be biased towards the state of the highly connected subgroups. An unbiased Horowitz-Thompson-type estimator (Horvitz and Thompson 1952) would require the knowledge of frequency of each degree. Clearly, the frequency of each degree will also be biased towards higher degrees, and thus cannot be recovered from the knowledge of individuals' degrees alone. The common remedy to this matter is the inverse-degree weighting heuristic (Crawford, Wu, and Heimer 2015; Guntuboyina, Barbour, and Heimer 2012).

In Berchenko, Rosenblatt, and Frost (2013) we proposed a generative model for RDS. The model is based on the idea that RDS spreads like an epidemic, and we can thus borrow epidemiological generative models. In particular SIR [TODO: add citation], for likelihood based inference. Having assumed a generative process, we can now estimate degree frequencies, introduce covariates, check the goodness of fit, and discuss the model's assumptions.

The details of the assumed generative model can be found in Berchenko *et al.* (2013), but the essentials are now detailed for completeness.

Denote N_k the unknown population frequency of degree k, i.e., the number of individuals with k "friends". Denote by x_t the degree of the respondent recruited at time t. Our task is to estimate $\hat{N}_1, \hat{N}_2, \ldots$, based on a sample of $x_{t_1}, \ldots, x_{t_\tau}$. Denote $\lambda_{k,t}$, the probability of recruiting an individual with degree k at time t. We assume the following generative multivariate counting model:

$$x \tag{1}$$

2. Work Flow

3. Some Technicalities

4. Conclusion

5. Future Work

References

Berchenko Y, Rosenblatt J, Frost SDW (2013). "Modeling and Analysing Respondent Driven Sampling as a Counting Process." *arXiv:1304.3505*. 1304.3505.

Crawford FW, Wu J, Heimer R (2015). "Hidden population size estimation from respondent-driven sampling: a network approach." *arXiv preprint*.

Guntuboyina A, Barbour R, Heimer R (2012). "On the impossibility of constructing good population mean estimators in a realistic Respondent Driven Sampling model." *arXiv preprint*.

Heckathorn D (1997). "Respondent-driven sampling: a new approach to the study of hidden populations." *Social Problems*, **44**(2), 174–199.

Heckathorn D (2002). "Respondent-driven sampling II: deriving valid population estimates from chain-referral samples of hidden populations." *Social Problems*, **49**(1), 11–34.

Horvitz DG, Thompson DJ (1952). "A generalization of sampling without replacement from a finite universe." *Journal of the American Statistical Association*, **47**(260), 663–685.

Affiliation:

Jonathan D. Rosenblatt
Department of Industrial Engineering and Management
Faculty of Engineering
Ben Gurion University of the Negev
P.O. 653, Beer Sheva, 8410501
E-mail: johnros@bgu.ac.il
URL: http://www.john-ros.com/

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
MMMMMM YYYY, Volume VV, Issue II *Submitted:* yyyy-mm-dd
doi:10.18637/jss.v000.i00 *Accepted:* yyyy-mm-dd