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 $N := \sum_k N_k$, the total population size. Denote by x_t the degree of the respondent recruited at time t. Our task is to estimate $\hat{N}_1, \hat{N}_2, \dots$, based on a sample of $x_{t_1}, \dots, x_{t_\tau}$. Denote $\lambda_{k,t}$, the probability of recruiting an individual with degree k in the time interval $[t, t+\Delta]$. We assume the following generative multivariate counting model:

$$\lambda_{k,t} = \beta_k \frac{n_t}{N}(N_k - n_{k,t})\Delta t + o(\Delta t), \tag{1}$$

where β_k is some base recruitment rate, n_t is the number of recruiting individuals at time t, and $n_{k,t} := \sum_{s \leq t} I_{\{x_s = k\}}$ the number of recruited individuals of degree k. It thus follows that $\frac{I_t}{N}$ is the recruiting population proportion, and $(N_k - n_{k,t})$ the recruitment "potential" of degree k. In these terms, our model implies that the recruitment probability, in short enough time periods, is proportional to the recruiting team (a.k.a. the "snowball") and the recruiting potential. For more on this model, and relations to other existing models, see Berchenko *et al.* (2013).

Equipped with this generative model, we may state the estimation of $N_1, N_2, \dots$ as a likelihood maximization problem. Moreover, we show in Berchenko *et al.* (2013) that the maximum likelihood estimator (MLE) of $N_1, N_2, \dots$ is separable for the various k, and independent of β_k. This fact is used by Berchenko *et al.* (2013) for proof of the asymptotic properties of $\hat{N}_1, \hat{N}_2, \dots$, and in the **chords** package to accelerate the estimation task.

MLEs for $N_1, N_2, \dots$ in model (1) have an interesting finite sample property: They tend to return $\hat{N}_k = \infty$ with non null probabilities. This clearly calls for some regularization. The Bayesian framework, a celebrated regularization device, will not work in this problem. Informally speaking, this is because any shrinkage of ∞ is still ∞. In our package, we have implemented several regularization devices. Some very crude heuristics, and some less so. In particular, we generalized the ideas of Musa and Iannino (1991) and Osborne and Severini (2000) to our multivariate case. The former proposing a jacknife-type resampling scheme, and the latter an *integrated likelihood* approach.

The usage of the package, the MLE estimator, and the small-sample modifications, is demonstrated in the next section.

2. Work Flow

We start by demonstrating the workflow on some data supplied with the package. The data consists of [TODO: describe data]. For more details, see Salganik, Fazito, Bertoni, Abdo, Mello, and Bastos (2011).

At a high-level, the work-flow has the same functional-object-oriented flavour as **ggplot2**: All the data required for the estimation is contained in the `rds-class` objects, and so are the results. Like **ggplot2**, and unlike object-oriented, the estimation does not modify the object on which it operates. The output will thus be a new object, with the same data, and a new estimate.

A typical work-flow consists of: (a) casting the data into the RDS file format, (b) initializing an `rds-class` object, (c) compute initial estimator. (d) use a modified estimator to deal with infinite values.

We start by simulating 10 RDS samples. The population consists of 4 observed degree classes $(2, 5, 10, 20)$, with $1,000$ individuals per degree. Formally, $N_k = 1,000$ for $k = 2, 5, 10, 20$, so that $N = 4,000$.

```
R> library(magrittr)
R> library(chords)
R> dk <- c(2, 5, 10, 20) # four degree classes
R> true.dks <- rep(0,max(dk))
R> true.dks[dk] <- dk
R> true.Nks <- rep(0,max(dk))
R> true.Nks[dk] <- 1e3
R> beta <- 1
R> theta <-  0.1
R> true.log.bks <- rep(-Inf, max(dk))
R> true.log.bks[dk] <- theta*log(beta*dk)
R> sample.length <- 1e3 # the total sample size
R> nsims <- 1e1
R> simlist <- list()
R> for(i in 1:nsims){
+   simlist[[i]] <- makeRdsSample(
+     N.k =true.Nks ,
+     b.k = exp(true.log.bks),
+     sample.length = sample.length)
+ }
```

The file format adheres to the RDS conventions as described in http://www.respondentdrivensampling.org/reports/RDSAT_7.1-Manual_2012-11-25.pdf. In particular note the `NS1` variable which encodes the (self reported) degree of each recruitee, `refCoupNum` for the recruiter's coupon number, `coup1-coup3` for the coupons given upon recruitment, and `interviewDt` for the time of recruitment. The scale of `interviewDt` is immaterial since the initial time will be later subtracted, and any rescaling will merely change the time units of the baseline recruitment rate β_k. We now ready to call the `rds-object` constructor:

```
R> #rds.object<- initializeRdsObject(brazil)
```

`rds.object` will now hold all the information needed for the estimation, and an empty slot for future estimates. The default maximum likelihood estimates are now computed:

```
R> #rds.object <- Estimate.b.k(rds.object = rds.object)
```

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.

Musa JD, Iannino A (1991). “Estimating the Total Number of Software Failures Using an Exponential Model.” *SIGSOFT Softw. Eng. Notes*, **16**(3), 80–84. ISSN 0163-5948. doi:10.1145/127099.127123.

Osborne JA, Severini TA (2000). “Inference for Exponential Order Statistic Models Based on an Integrated Likelihood Function.” *Journal of the American Statistical Association*, **95**(452), 1220–1228. ISSN 0162-1459. doi:10.1080/01621459.2000.10474322.

Salganik MJ, Fazito D, Bertoni N, Abdo AH, Mello MB, Bastos FI (2011). “Assessing Network Scale-up Estimates for Groups Most at Risk of HIV/AIDS: Evidence From a Multiple-Method Study of Heavy Drug Users in Curitiba, Brazil.” *American Journal of Epidemiology*, **174**(10), 1190–1196. doi:10.1093/aje/kwr246.

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