

Respondent driven-sampling

Bayesian analysis of respondent-driven survey with outcome uncertainty

Lucas Moschen

School of Applied Mathematics
Fundação Getulio Vargas

June 20, 2021

Table of Contents

- ① Introduction
- ② Justification
- ③ Objectives
- ④ Methodology
- ⑤ Preliminary results
- ⑥ Schedule

Table of Contents

- ① Introduction
- ② Justification
- ③ Objectives
- ④ Methodology
- ⑤ Preliminary results
- ⑥ Schedule

Hidden and hard-to-reach populations

- ▶ No sampling frame exists: size and boundaries of the population are unknown.
- ▶ Privacy concerns: stigmatized or illegal behavior.
- ▶ Fear of exposition or prosecution complicates the enumeration and learning about these populations.
- ▶ High logistic cost when the occurrence frequency is low.
- ▶ Examples: Heavy drug users, sex workers, homeless people, and men who have sex with men.

Respondent-driven sampling

- 1 The researchers select a handful of individuals from a target population who serve as *seeds*.
- 2 Each participant receives a fixed number of *recruitment coupons* and invite members of their own social network to participate in exchange of a reward.
- 3 The sampling is without replacement.
- 4 If the individual accepts to participate, they answer a questionnaire and inform the network degree. One important point is that the recruiter doesn't say the name of the other members, reducing the mask effect.

The RDS can be seen mathematically in two different approaches.

- ▶ **Stochastic process** [[Heckathorn, 1997](#)]

Each recruiter's social characteristics affect the characteristics of the recruits. There are a limited number of states that subjects can assume and the recruits are function of the recruiter characteristics.

- ▶ **Graphical structure** [[Crawford, 2016](#)]

A hidden population is an undirected graph, and we observe it partially in the *recruitment graph*, as also the coupon matrix and recruitment times. The unobserved graph is treated as *missing data* what can be interpreted as as Exponential Random Graph Model.

Prevalence estimation with imperfect tests

- ▶ Prevalence (θ) = Proportion of a disease (or condition) at time t ;
- ▶ *Specificity* (γ_s) and *Sensitivity* (γ_e) are (jointly) measures of the disease diagnose accuracy;
- ▶ If a sample $\{y_1, \dots, y_n\}$ is observed, its mean is the maximum likelihood estimator, but it supposes a perfect test;

- ▶ If π is the probability of a positive test,

$$\pi = \theta\gamma_s + (1 - \theta)(1 - \gamma_e),$$

and we can study θ regarding γ_s, γ_e .

- ▶ Observe also that for a network the independence of the sample is also not valid. Other estimators are better.

- ▶ Interpretation based on an individual's degree of belief in a statement;
- ▶ Bayes' formula relates the probability of a parameter after observing new data with evidence and previous information about it;
- ▶ It allows the quantification of uncertainty in a straightforward way: the process do not need to be random.

Table of Contents

- ① Introduction
- ② Justification
- ③ Objectives
- ④ Methodology
- ⑤ Preliminary results
- ⑥ Schedule

- ▶ Hidden populations are often omitted from national representative surveys and have higher risk of drug abuse or sexually transmitted infections.
- ▶ The topic has a lot of gaps in Statistics and regression approaches to prevalence estimation taking the network structure can be built [[Bastos et al., 2012](#)].

Table of Contents

- ① Introduction
- ② Justification
- ③ Objectives
- ④ Methodology
- ⑤ Preliminary results
- ⑥ Schedule

Main objective

The objective of this work is to analyze the network structure of RDS as a stochastic object, along with the sensibility and sensitivity. We also intend to apply this framework efficiently, comparing Monte Carlo algorithms and Laplace approximations.

- 1 Bibliography review;
- 2 Problem description in mathematical terms and uncertainty propagation;
- 3 Bayesian methods and prior calibration;
- 4 Joint prior distribution for sensitivity and specificity;
- 5 Efficient implementation using statistical packages, as *rstanarm* and *INLA*. Comparison between MCMC and Laplace approximation;
- 6 Analysis of RDS epidemiological studies.

Table of Contents

- ① Introduction
- ② Justification
- ③ Objectives
- ④ Methodology
- ⑤ Preliminary results
- ⑥ Schedule

Document research

The theoretical foundation will be through papers in the topics indicated in the introduction, RDS, bayesian statistics, and prevalence estimation through regression.

Technical resources

All the necessary programming will be done in the programming languages *Python* and *R*.

Formal study

Two subjects from the PhD in Mathematical Modelling at EMap will be taken: Bayesian Statistics and Network Science.

Table of Contents




- ① Introduction
- ② Justification
- ③ Objectives
- ④ Methodology
- ⑤ Preliminary results**
- ⑥ Schedule

Regression

Table of Contents

- ① Introduction
- ② Justification
- ③ Objectives
- ④ Methodology
- ⑤ Preliminary results
- ⑥ Schedule

References I

-  Bastos, L. S., Pinho, A. A., Codeço, C., and Bastos, F. I. (2012). Binary regression analysis with network structure of respondent-driven sampling data.
-  Crawford, F. W. (2016). The graphical structure of respondent-driven sampling. *Sociological Methodology*, 46(1):187–211.
-  Heckathorn, D. D. (1997). Respondent-driven sampling: A new approach to the study of hidden populations. *Social Problems*, 44(2):174–199.