

Lucas Machado Moschen

**Bayesian analysis of respondent-driven surveys
with outcome uncertainty**

Rio de Janeiro, Brazil

16 de junho de 2021

Lucas Machado Moschen

Bayesian analysis of respondent-driven surveys with outcome uncertainty

Monograph Project presented to the School
of Mathematics Applied (FGV) as a partial
requirement for continuing the monograph
work.

Getulio Vargas Foundation – FGV

School of Applied Mathematics

Undergraduate Course in Applied Mathematics

Rio de Janeiro, Brazil

16 de junho de 2021

Contents

1	INTRODUCTION	3
1.1	Respondent-driven sampling	4
1.2	Prevalence estimation with imperfect tests	4
1.3	Bayesian statistics	5
2	JUSTIFICATION	6
3	OBJECTIVES	7
3.1	Main	7
3.2	Specifics	7
4	METHODOLOGY	8
5	PRELIMINARY RESULTS	9
6	SCHEDULE	10
	Final considerations	11
	BIBLIOGRAPHY	12

1 Introduction

This work proposes to study the survey method Respondent-Driven Sampling (RDS), a chain-referral method with the objective of sampling from hard-to-reach populations when necessary to estimate the prevalence of some binary condition from this population. The modeling also accounts for sensibility and sensitivity since the imperfection of the detection tests.

Hidden or hard-to-reach populations have two main features: no sampling frame exists, given that their size and boundaries are unknown, and there are privacy concerns because the subjects are stigmatized or have illegal behavior (HECKATHORN, 1997). Fear of exposition or prosecution complicates the enumeration of the populations and the learning about them. Moreover, if the occurrence frequency of the condition is low, there are high logistic costs involved. Some examples are heavy drug users, sex workers, homeless people, and men who have sex with men.

Researches have been done with the development of some methods to reach these populations, such as, for example, snowball sampling (GOODMAN, 1961), key important sampling (DEAUX; CALLAGHAN, 1985), and targeted sampling (WATTERS; BIER-NACKI, 1989). HECKATHORN introduced the Respondent-Driven Sampling (RDS) to fill some gaps from other methods he depicted in his work. In his proposed approach, the researchers select a handful of individuals from the target population and give them coupons to recruit their peers. The individuals receive a reward for being recruited and for recruiting, which creates a dual incentive system. After 1997, several papers studied this topic more deeply.

Following the sampling from the target population, a questionnaire or a disease test is conducted. This work considers binary outcomes. For instance, asking about smoking status or testing for HIV infections. However, the diagnoses are subject to measure error, and regard their accuracy is a vital step (REITSMA et al., 2005). In particular, we propose the joint use of sensitivity (the ability to detect the condition) and specificity (the ability to identify the absence of it).

Nevertheless, because of our lack of knowledge about nature itself, it is necessary to model the uncertainty of this process, and Bayesian Statistics is the indicated area of study. In the Bayesian view, the parameters are random variables, and the beliefs about them are updated given new data. The idea is to propagate uncertainty about the outcome through the network of contacts, which has its probability distribution.

The objective of this work is to analyze the network structure 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.

1.1 Respondent-driven sampling

RDS is commonly used to survey hidden or hard-to-reach populations when no sampling frame exists (HECKATHORN, 1997). In this approach, the researchers select some individuals, called *seeds* from the target population, and give them a fixed amount of *recruitment coupons* to recruit their peers. Each recipient of the coupons reclaims it in the study site, is interviewed, and receives more coupons to continue the recruitment. This process occurs until it reaches some criteria. The sampling is without replacement, so the participants cannot be recruited more than once. Moreover, the respondents inform their *network degree*.

The subjects receive a reward for being interviewed and for each recruitment which establishes a dual system incentive. The *primary incentive* is the *individual-sanction-based control*, so there is a reward for participating. The second one is the *group-mediated social control* that influences the participants seeking to induce others to comply. When social approval is important, recruitment can be even more efficient and cheaper. Moreover, the material incentive can be converted into symbolic by the individuals.

Descrever o modelo de Heckathorn

Descrever o modelo de Crawford

Discutir detalhes adicionais de outros artigos como aplicações

1.2 Prevalence estimation with imperfect tests

Consider a population of interest and a known condition, such as, for example, a disease or a binary behavior. It is important to understand the proportion of individuals in this population exposed at time t , called *prevalence*. In hidden networks, the population size is unknown prior to the study, what makes the proceeding to be hard. Suppose a diagnostic test is done to measure the presence or the absence of this condition in the individuals.

Mathematically, let $\theta \in (0, 1)$ be the prevalence of a condition and Y_i be an indicator function of the condition in the i^{th} individual. Assuming for simplicity that all tests are performed at time t , and the sample is $\{y_1, \dots, y_n\}$, the RDS estimator was proposed based largely on Markov chain theory and social network theory (HECKATHORN, 1997; HECKATHORN, 2002).

$$a = 1 \tag{1.1}$$

Citar (VOLZ; HECKATHORN, 2008) para o estimador RDS II.

1.3 Bayesian statistics

Breve discussão sobre o teorema de Bayes e estatística bayesiana

2 Justification

There are two justifications for the importance of this work. First, hidden populations are often omitted from national representative surveys since they do not have fixed addresses or fear prosecution. However, the individuals can have a greater risk of drug abuse or having sexually transmitted infections. This combination creates an environment of aid absence from the government to these people. The second reason is mathematical. This topic has lots of gaps in Statistics that deserve attention. The correct sampling probabilities for the recruited members under RDS are hard to obtain since not all links and nodes are observed, constituting missing data ([CRAWFORD, 2016](#)). In this fertile area, regression approaches to prevalence estimation taking the network structure can be built ([BASTOS et al., 2012](#)) and are still in development.

3 Objectives

3.1 Main

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.

3.2 Specifics

- a) Bibliography review including possible mathematical formulations of RDS, logistic regression when the outcome has uncertainty (imperfect tests) and applications in hidden or hard-to-reach populations;
- b) Problem description in mathematical terms considering the uncertainty in the network (missing data in the RDS) and the diagnose. Uncertainty propagation;
- c) Bayesian methods and prior calibration. Prior predictive checking on the positive test probability when there are weak priors of the regression parameters;
- d) Efficient implementation using statistical packages, as *rstanarm* ([Stan Development Team , 2021](#)) and *INLA* ([RUE; MARTINO; CHOPIN, 2009](#)). Comparison between MCMC and Laplace approximation;
- e) Analysis of RDS epidemiological studies.

4 Methodology

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*, given the simple connection to data processing and statistics.

Formal studies

In order to help the learning about the foundations, two subjects from the PhD in Mathematical Modelling at EMAP will be taken: Bayesian Statistics and Network Science. The first one ended in June, while the second will be finished on September.

5 Preliminary results

6 Schedule

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Final considerations

Sed consequat tellus et tortor. Ut tempor laoreet quam. Nullam id wisi a libero tristique semper. Nullam nisl massa, rutrum ut, egestas semper, mollis id, leo. Nulla ac massa eu risus blandit mattis. Mauris ut nunc. In hac habitasse platea dictumst. Aliquam eget tortor. Quisque dapibus pede in erat. Nunc enim. In dui nulla, commodo at, consectetur nec, malesuada nec, elit. Aliquam ornare tellus eu urna. Sed nec metus. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas.

Phasellus id magna. Duis malesuada interdum arcu. Integer metus. Morbi pulvinar pellentesque mi. Suspendisse sed est eu magna molestie egestas. Quisque mi lorem, pulvinar eget, egestas quis, luctus at, ante. Proin auctor vehicula purus. Fusce ac nisl aliquam ante hendrerit pellentesque. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Morbi wisi. Etiam arcu mauris, facilisis sed, eleifend non, nonummy ut, pede. Cras ut lacus tempor metus mollis placerat. Vivamus eu tortor vel metus interdum malesuada.

Sed eleifend, eros sit amet faucibus elementum, urna sapien consectetur mauris, quis egestas leo justo non risus. Morbi non felis ac libero vulputate fringilla. Mauris libero eros, lacinia non, sodales quis, dapibus porttitor, pede. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Morbi dapibus mauris condimentum nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Etiam sit amet erat. Nulla varius. Etiam tincidunt dui vitae turpis. Donec leo. Morbi vulputate convallis est. Integer aliquet. Pellentesque aliquet sodales urna.

Bibliography

BASTOS, L. S. et al. *Binary regression analysis with network structure of respondent-driven sampling data*. 2012. Citado na página 6.

CRAWFORD, F. W. The graphical structure of respondent-driven sampling. *Sociological Methodology*, v. 46, n. 1, p. 187–211, 2016. Disponível em: <https://doi.org/10.1177/0081175016641713>. Citado na página 6.

DEAUX, E.; CALLAGHAN, J. W. Key informant versus self-report estimates of health-risk behavior. *Evaluation Review*, v. 9, n. 3, p. 365–368, 1985. Disponível em: <https://doi.org/10.1177/0193841X8500900308>. Citado na página 3.

GOODMAN, L. A. Snowball Sampling. *The Annals of Mathematical Statistics*, Institute of Mathematical Statistics, v. 32, n. 1, p. 148–170, 1961. Disponível em: <https://doi.org/10.1214/aoms/1177705148>. Citado na página 3.

HECKATHORN, D. D. Respondent-driven sampling: A new approach to the study of hidden populations. *Social Problems*, [Oxford University Press, Society for the Study of Social Problems], v. 44, n. 2, p. 174–199, 1997. Disponível em: <http://www.jstor.org/stable/3096941>. Citado 2 vezes nas páginas 3 and 4.

HECKATHORN, D. D. Respondent-driven sampling ii: deriving valid population estimates from chain-referral samples of hidden populations. *Social problems*, Oxford University Press, v. 49, n. 1, p. 11–34, 2002. Citado na página 4.

REITSMA, J. B. et al. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of clinical epidemiology*, Elsevier, v. 58, n. 10, p. 982–990, 2005. Citado na página 3.

RUE, H.; MARTINO, S.; CHOPIN, N. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, Wiley Online Library, v. 71, n. 2, p. 319–392, 2009. Citado na página 7.

Stan Development Team . *Rstanarm Developer Notes*. 2021. Rstanarm website. Disponível em: <https://mc-stan.org/rstanarm/dev-notes/index.html>. Acesso em: 16 jun 2021. Citado na página 7.

VOLZ, E.; HECKATHORN, D. D. Probability based estimation theory for respondent driven sampling. *Journal of official statistics*, Statistics Sweden (SCB), v. 24, n. 1, p. 79, 2008. Citado na página 5.

WATTERS, J. K.; BIERNACKI, P. Targeted sampling: Options for the study of hidden populations. *Social Problems*, Oxford University Press, Society for the Study of Social Problems, v. 36, n. 4, p. 416–430, 1989. Disponível em: <http://www.jstor.org/stable/800824>. Citado na página 3.