Lucas Machado Moschen

Bayesian analysis of respondent-driven surveys with outcome uncertainty

Rio de Janeiro, Brazil

18 de junho de 2021

Lucas Machado Moschen

Bayesian analysis of respondent-driven surveys with outcome uncertainty

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

Advisor: Prof. Luiz Max Fagundes de Carvalho

Getulio Vargas Foundation – FGV

School of Applied Mathematics

Undergraduate Course in Applied Mathematics

Rio de Janeiro, Brazil

18 de junho de 2021

Contents

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; BIERNACKI, 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.

In a survey, questions about ethnicity, location (not necessarily fixed), gender, and religion, create possible (finite) states in which each participant is. By statistical tests, one can verify the association between the recruiter and recruited responses. HECKATHORN models it as a Markov chain where the states are the possible answers, and the links are the recruitments. Considering an ergodic chain, an equilibrium mix of recruits will be attained when the number of waves goes to infinity, and it approaches the equilibrium at a geometric rate. Therefore, we obtain the distribution of the states posterior to enough waves. Posterior studies (HECKATHORN, 2002) explained how to access bias and other statistical considerations.

Besides considering only the states where the individual is located, (CRAWFORD, 2016) analyses the network structure given by RDS with a continuous-time model incorporating the recruitment time, the network degree, and the pattern of coupon use. This configuration enables the treatment of unobserved links and nodes as missing data. Let $G = (V, E)$ be an undirected graph representing the hidden population. The *recruitment graph* $G_R = (V_R, E_R)$ represents the recruited individuals and the recruitment edge. Given that each individual can be sampled only once, it is not possible to observe the *recruitment-induced subgraph*, that is the induced subgraph generated by V_R. Moreover, the *coupon matrix* C defined by $C_{ij} = 1$ if the ith subject has at least one coupon before the jth recruitment event, is also observed with the recruitment times. Assuming an exponential and independent distribution of the times, the likelihood can be written, and the distribution interpreted as an exponential random graph model.

These models allowed several applications in social sciences, epidemiology, and statistics, including hidden populations size estimation (CRAWFORD; WU; HEIMER, 2018), regression (BASTOS et al., 2012), communicable disease prevalence estimation (ALBUQUERQUE et al., 2009), among others.

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*. 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 (parameter of interest) of the condition and Y_i be an indicator function of the presence of the condition in the ith individual. Assuming for simplicity that all tests are performed at time t, and the sample is $\{y_1, ..., y_n\}$, the maximum likelihood estimator is

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} y_i. \tag{1.1}$$

However this estimator has two problems in this context: it is assumed perfect diagnostic test, what is often incorrect, and the samples in RDS are not independent by definition (network structure). The latter was a study object in (HECKATHORN, 1997; HECKATHORN, 2002) where the estimator was proposed based largely on Markov chain theory and social network theory. (VOLZ; HECKATHORN, 2008) improved it with the RDS II estimator considering the network degree

$$\hat{\theta}^{RDSII} = \frac{\sum_{i=1}^{n} y_i \delta_i^{-1}}{\sum_{i=1}^{n} \delta_i^{-1}}, \tag{1.2}$$

such that δ_i is the ith individual's degree. However, this is an area of research in progress.

The first problem in (1.1) was tackled several times in the literature, such as (MCINTURFF et al., 2004). A possible way to handle is to bring to the model the sensitivity (γ_s) and specificity (γ_e). Let π be the probability of a test comes positive. Then

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

Establish a link function in $(\pi, \gamma_s, \gamma_e)$, it is possible to use linear regression and prior distributions to the regressors $\boldsymbol{\beta}$. One important problem is to consider the correlation between γ_s and γ_e.

1.3 Bayesian statistics

There are two more common interpretations of probability and statistics: frequentist and Bayesian. While the frequentists define probability as the limit of a frequency in a large number of trials, the Bayesians represent an individual's degree of belief in a statement that is updated given new information. This philosophy allows assigning probabilities to any event, even if a random process is not defined (STATISTICAT, 2016).

In 1761, Reverent Thomas Bayes wrote for the first time the Bayes' formula relating the probability of a parameter after observing the data with the evidence (written through a likelihood function) and previous information about the parameter. Pierre Simon Laplace rediscovered this formula in 1773 (ROBERT, 2007), and this theory became more common in the 19th century. After some criticisms, a modern treatment considering Kolmogorov's axiomatization of the theory of probabilities started after Jeffreys in 1939. The recent development of new computational tools brought these ideas again.

Bayesian inference is composed by the following:

- A distribution for the parameters θ that quantifies the uncertainty about θ before data;
- A distribution of the data generation process given the parameter, such that, when it is seen as function of the parameter, is called likelihood function;
- When considering decision theory, a loss function indicating a measure of error;
- Posterior distribution of the parameter conditioned on the data. All inferences are based on this probability distribution.

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) Joint prior distribution for sensitivity and specificity of the diagnostic test;

e) Efficient implementation using statistical packages, as *rstanarm* (Stan Development Team , 2021) and *INLA* (RUE; MARTINO; CHOPIN, 2009). Comparison between MCMC and Laplace approximation;

f) 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 study:

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

A tentative schedule is depicted in Figure XXX.

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, consectetuer 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 consectetuer 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

ALBUQUERQUE, E. M. d. et al. *Avaliação da técnica de amostragem respondent-driven sampling na estimação de prevalências de doenças transmissíveis em populações organizadas em redes complexas.* Tese (Doutorado), 2009. Citado na página 5.

BASTOS, L. S. et al. *Binary regression analysis with network structure of respondent-driven sampling data.* 2012. Citado 2 vezes nas páginas 5 and 7.

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 2 vezes nas páginas 4 and 7.

CRAWFORD, F. W.; WU, J.; HEIMER, R. Hidden population size estimation from respondent-driven sampling: a network approach. *Journal of the American Statistical Association*, Taylor & Francis, v. 113, n. 522, p. 755–766, 2018. Citado na página 5.

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 3 vezes nas páginas 3, 4, and 5.

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 2 vezes nas páginas 4 and 5.

MCINTURFF, P. et al. Modelling risk when binary outcomes are subject to error. *Statistics in medicine*, Wiley Online Library, v. 23, n. 7, p. 1095–1109, 2004. Citado na página 5.

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.

ROBERT, C. *The Bayesian choice: from decision-theoretic foundations to computational implementation.* [S.l.]: Springer Science & Business Media, 2007. Citado na página 6.

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 8.

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 8.

STATISTICAT, L. Laplacesdemon: A complete environment for bayesian inference within r. *R Package version*, v. 17, p. 2016, 2016. Citado na página 6.

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.