Lucas Machado Moschen

Bayesian analysis of respondent-driven surveys with outcome uncertainty

Rio de Janeiro, Brazil

28 de junho de 2021

Lucas Machado Moschen

Bayesian analysis of respondent-driven surveys with outcome uncertainty

Bachelor dissertation project presented to the School of Applied Mathematics (FGV/EMAp) as a partial requirement for continuing the dissertation work.

Advisor: Prof. Luiz Max Carvalho

Getulio Vargas Foundation – FGV

School of Applied Mathematics

Undergraduate Course in Applied Mathematics

Rio de Janeiro, Brazil

28 de junho de 2021

Contents

1 INTRODUCTION . . . 3

1.1 Respondent-driven sampling . . . 4

1.2 Prevalence estimation with imperfect tests . . . 5

1.3 Bayesian statistics . . . 6

2 JUSTIFICATION . . . 7

3 OBJECTIVES . . . 8

3.1 Main . . . 8

3.2 Specifics . . . 8

4 METHODOLOGY . . . 9

5 PRELIMINARY RESULTS . . . 10

6 SCHEDULE . . . 13

References . . . 14

1 Introduction

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.

Research has been carried out 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 (HECKATHORN, 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). One common way to do this is to measure jointly *sensitivity* and *specificity*. The former is the ability to detect the condition, while the latter 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 paradigm, 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.

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. 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 of their peers 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 to induce others to comply to get the reward for the recruitment. When social approval is important, recruitment can be even more efficient and cheaper, since material incentive can be converted into symbolic by the individuals. In summary, accepting to be recruited will have a material incentive for both and a symbolic incentive for the recruited, since theirs peers also participated.

In a survey, questions about ethnicity, location (not necessarily fixed), gender, and religion, create possible (finite) states in which each participant is. Using 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 explicitly, 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 the apparent prevalence:

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

However, this estimator has two problems in this context: it assumes a perfect diagnostic test, which is often incorrect, and the samples in RDS are not independent by definition (network structure).

The first problem in (1.1) was tackled several times in the literature, such as (MCINTURFF et al., 2004). The diagnose accuracy can be measured in many ways and the most considered is the joint analysis of the *sensitivity* (γ_s) and the *specificity* (γ_e).

Definition 1.2.1 (Specificity)**.** It is the probability of a negative test conditioned on the absence of the disease (true negative).

Definition 1.2.2 (Sensitivity)**.** It is the probability of a positive test conditioned on the presence of the disease (true positive).

Let p be the probability of a positive test. Then, by Law of Total Probability:

$$p = \theta\gamma_s + (1 - \theta)(1 - \gamma_e). \tag{1.2}$$

Assuming the tests are conditionally independent given the presence or the absence of the disease in each individual, the number of positive tests X has binomial distribution with success probability p. In chapter 5 we present preliminary model that accounts this. Regression approaches can be also carried with a link function in θ. One important additional problem is to consider the correlation between γ_s and γ_e.

The second problem was a study object in (HECKATHORN, 1997, 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

$$\hat{\theta}^{RDSII} = \frac{\sum_{i=1}^{n} y_i \delta_i^{-1}}{\sum_{i=1}^{n} \delta_i^{-1}}, \tag{1.3}$$

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

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, Reverend 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 measuring the error in evaluating the parameter;
- 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 and cannot be reached, or they fear prosecution after showing themselves. 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 diagnostic. 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

Consider the following model (GELMAN; CARPENTER, 2020):

$$
\begin{aligned}
y &\sim \text{Binomial}(n, p), \\
p &= \theta\gamma_s + (1-\theta)(1-\gamma_e),
\end{aligned}
$$

such that y is the number of positive tests in a population of size n. In a Bayesian paradigm, a prior $\pi(\theta, \gamma_e, \gamma_s)$ must be specified. For instance, $\pi(\theta, \gamma_e, \gamma_s) = \pi(\theta)\pi(\gamma_e, \gamma_s)$ and $\theta \sim \text{Beta}(\alpha_\theta, \beta_\theta)$, in which α_θ and β_θ are positive hyperparameters. Since the three parameters θ, γ_e, and γ_s are not jointly identifiable only from y, prior information on γ_e and γ_s need be added. For this,

$$
\begin{aligned}
y_{negative} &\sim \text{Binomial}(n_{\gamma_e}, \gamma_e), \\
y_{positive} &\sim \text{Binomial}(n_{\gamma_s}, \gamma_s),
\end{aligned}
$$

such that $y_{negative}$ are negative tests on known negative subjects (by a gold standard for example) and $y_{positive}$ are positive tests on known positive. When considering separated experiments for specificity and sensitivity, there is no information about their correlation, which is the case for our model. Then we define the the prior distributions

$$
\begin{aligned}
\gamma_e &\sim \text{Beta}(a_e, b_e), \\
\gamma_s &\sim \text{Beta}(a_s, b_s), \\
\theta &\sim \text{Beta}(a_\theta, b_\theta).
\end{aligned}
$$

Using data from (BENNETT; STEYVERS, 2020) about COVID-19 seroprevalence in Santa Clara:

$$
\begin{aligned}
y/n &= 50/3330, \\
y_{negative}/n_{\gamma_e} &= 399/401, \\
y_{positive}/n_{\gamma_s} &= 103/122,
\end{aligned}
$$

we fit the model and obtain the results showed in Figure 1. All the codes were done in *Stan* and *PyStan*.

Other approach considers more than one study about specificity and sensitivity. A *hierarchical partial pooling* model for these studies can be done in the following way:

$$
\begin{aligned}
\text{logit}(\gamma_s^j) &\sim \text{Normal}(\mu_{\gamma_s}, \sigma_{\gamma_s}), \\
\text{logit}(\gamma_e^j) &\sim \text{Normal}(\mu_{\gamma_e}, \sigma_{\gamma_e}),
\end{aligned}
$$

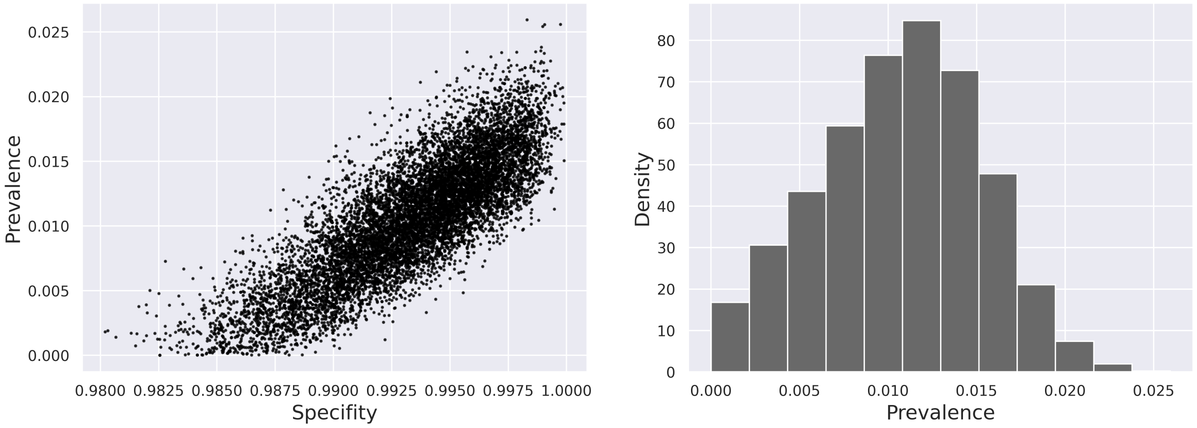

Figure 1 – Scatter plot of posterior simulations of prevalence against specificity and histogram of posterior simulations of the prevalence.

for $1 \leq j \leq K$ studies, such that the first study is the considered one. Partial pooling because the parameters can be sampled from the same distribution. Hierarchical because the parameters of this distribution have its one prior distributions. For instance,

$$\begin{aligned}
\mu_{\gamma_s} &\sim N(0, 10), \\
\mu_{\gamma_e} &\sim N(0, 10), \\
\sigma_{\gamma_s} &\sim N^+(0, 1), \text{ and} \\
\sigma_{\gamma_e} &\sim N^+(0, 1),
\end{aligned}$$

where $N^+(a, b)$ is the truncated normal distribution in $[0, +\infty)$. All the codes available at Github repository[1].

Finally, we studied a joint distribution for specificity and sensitivity, a possible bivariate beta distribution built in (OLKIN; TRIKALINOS, 2015). This distribution is derived from a Dirichlet distribution of order four. Let $U = (U[1], ..., U[4]) \sim \text{Dirichlet}(\boldsymbol{\alpha})$, where $\boldsymbol{\alpha} \in \mathbb{R}^4_+$. Therefore, defining $X = U[1] + U[2]$ and $Y = U[1] + U[3]$, we will have that (X, Y) has a well-defined probability distribution in $[0, 1] \times [0, 1]$ such that X and Y have marginally beta distributions, and they have correlation in all space. Depending on the definition of $\boldsymbol{\alpha}$, the correlation between the variables range from -1 and 1. Figure 2 shows some examples of this construction.

[1] `https://github.com/lucasmoschen/rds-bayesian-analysis`

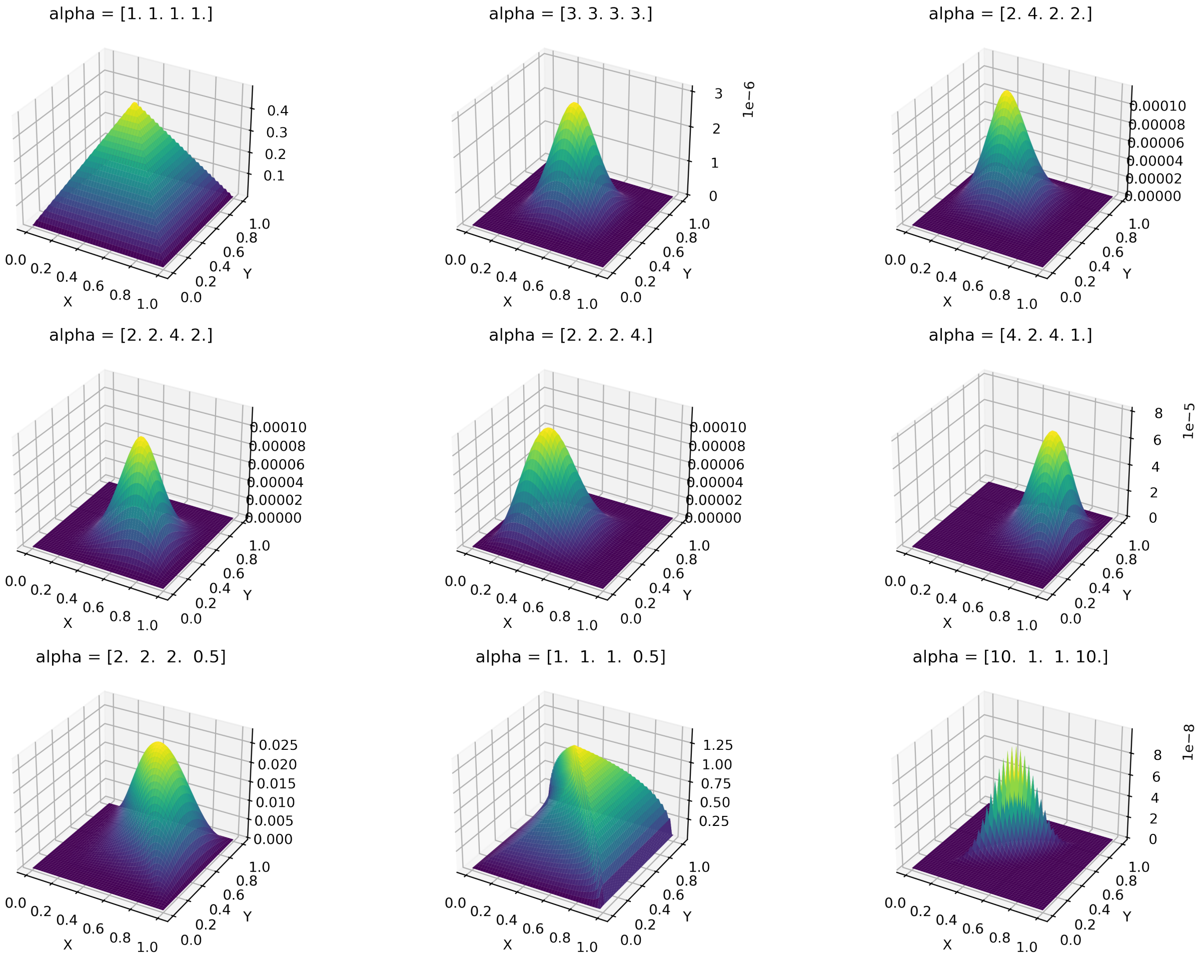

Figure 2 – Different choices of α and the joint distribution of the variables X and Y.

6 Schedule

A tentative schedule is depicted in Figure 3.

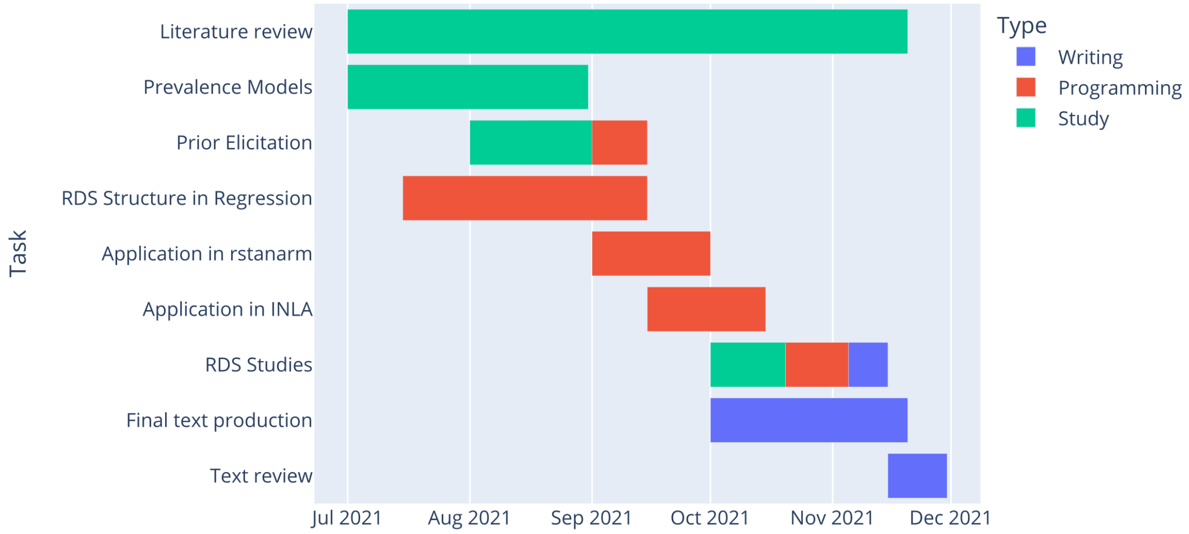

Figure 3 – Tentative schedule for the project during the second semester of 2021.

References

ALBUQUERQUE, Elizabeth Maciel de et al. **Avaliação da técnica de amostragem respondent-driven sampling na estimação de prevalências de doenças transmissíveis em populações organizadas em redes complexas**. 2009.
PhD thesis – ENSP.

BASTOS, Leonardo S. et al. **Binary regression analysis with network structure of respondent-driven sampling data**. [S.l.: s.n.], 2012. arXiv: 1206.5681 [stat.AP].

BENNETT, Stephen T; STEYVERS, Mark. Estimating COVID-19 antibody seroprevalence in Santa Clara County, California. A re-analysis of Bendavid et al. **MedRxiv**, Cold Spring Harbor Laboratory Press, 2020.

CRAWFORD, Forrest W; WU, Jiacheng; HEIMER, Robert. 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.

CRAWFORD, Forrest W. The Graphical Structure of Respondent-driven Sampling. **Sociological Methodology**, v. 46, n. 1, p. 187–211, 2016. Available from: <https://doi.org/10.1177/0081175016641713>.

DEAUX, Edward; CALLAGHAN, John W. Key Informant Versus Self-Report Estimates of Health-Risk Behavior. **Evaluation Review**, v. 9, n. 3, p. 365–368, 1985. Available from: <https://doi.org/10.1177/0193841X8500900308>.

GELMAN, Andrew; CARPENTER, Bob. Bayesian analysis of tests with unknown specificity and sensitivity. **Journal of the Royal Statistical Society: Series C (Applied Statistics)**, Wiley Online Library, v. 69, n. 5, p. 1269–1283, 2020.

GOODMAN, Leo A. Snowball Sampling. **The Annals of Mathematical Statistics**, Institute of Mathematical Statistics, v. 32, n. 1, p. 148–170, 1961. Available from: <https://doi.org/10.1214/aoms/1177705148>.

HECKATHORN, Douglas 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.

________. 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. Available from: <http://www.jstor.org/stable/3096941>.

MCINTURFF, Pat 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.

OLKIN, Ingram; TRIKALINOS, Thomas A. Constructions for a bivariate beta distribution. **Statistics & Probability Letters**, Elsevier, v. 96, p. 54–60, 2015.

REITSMA, Johannes 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.

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

RUE, Håvard; MARTINO, Sara; CHOPIN, Nicolas. 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.

STAN DEVELOPMENT TEAM. **Rstanarm Developer Notes**. [S.l.: s.n.], 2021. Rstanarm website. Available from: <https://mc-stan.org/rstanarm/dev-notes/index.html>. Visited on: 16 June 2021.

STATISTICAT, LLC. LaplacesDemon: A Complete Environment for Bayesian Inference within R. **R Package version**, v. 17, p. 2016, 2016.

VOLZ, Erik; HECKATHORN, Douglas D. Probability based estimation theory for respondent driven sampling. **Journal of Official Statistics**, Statistics Sweden (SCB), v. 24, n. 1, p. 79, 2008.

WATTERS, John K.; BIERNACKI, Patrick. 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. Available from: <http://www.jstor.org/stable/800824>.