FUNDAÇÃO GETULIO VARGAS
SCHOOL OF APPLIED MATHEMATICS

LUCAS MACHADO MOSCHEN

BAYESIAN ANALYSIS OF RESPONDENT-DRIVEN SURVEYS WITH OUTCOME UNCERTAINTY

Rio de Janeiro
2021

Contents

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.

2 Theoretical background

Throughout this chapter, we shall describe the theoretical background taken under consideration for the developed models and analysis, including the Respondent-driven sampling (Section 2.1), the prevalence estimation problem (Section 2.2), Bayesian statistics (Section 2.3), and computational methods (Section 2.5) used in our research.

2.1 Respondent-driven sampling

Respondent-driven sampling (RDS) is commonly used to survey hidden or hard-to-reach populations when no sampling frame exists (HECKATHORN, 1997), which means there is no enumeration of the population, since size and boundaries are unknown. 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 some criteria is reached. The sampling is without replacement, so the participants cannot be recruited more than once. Moreover, the respondents inform how many subjects from the population they know.

The subjects receive a reward for being interviewed and for each recruitment of their peers which establishes a dual incentive system. 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.

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 edges, that is, $(i, j) \in E_R$ if, and only if, i recruited j. 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 (CRAWFORD, 2016).

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.

2.2 Prevalence estimation problem

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:

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} y_i. \tag{2.1}$$

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 (2.1) was tackled several times in the literature, such as (MCINTURFF et al., 2004). 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

$$\hat{\theta}^{RDSII} = \frac{\sum_{i=1}^{n} y_i\delta_i^{-1}}{\sum_{i=1}^{n} \delta_i^{-1}}, \tag{2.2}$$

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

Let I be a index set and Y_i be the indicator function of the i^{th} individual's exposure to the disease, and T_i indicating whether the test of the i^{th} individual is positive at time t. Suppose that $\{Y_i\}_{i\in I}$ and $\{T_i\}_{i\in I}$ are two independent and identically distributed random variables with $\Pr(X = 1) = \theta$ and $\Pr(T = 1) = p$. We say that θ is the prevalence and p is the apparent prevalence in the population.

If the test is perfect, then for every i, $T_i = Y_i$, and $\theta = p$ (with probability one when they are random variables). Unfortunately, this is not true in the real world, what makes important to regard the evaluation of the diagnostic, and the following definitions are used:

Definition 2.2.1 (Specificity)**.** Probability of a negative test correctly identified. In mathematical terms, conditioned on $Y = 0$, the *specificity* γ_e is the probability of $T = 0$:

$$\gamma_e = \Pr(T = 0|Y = 0). \tag{2.3}$$

Definition 2.2.2 (Sensitivity)**.** Probability of a positive test correctly identified. In mathematical terms, conditioned on $Y = 1$, the *sensitivity* γ_s is the probability of $T = 1$:

$$\gamma_s = \Pr(T = 1|Y = 1). \tag{2.4}$$

Theorem 1 (Relation between prevalence and apparent prevalence)**.** *These quantities are related by the following equation:*

$$p = \gamma_s\theta + (1 - \gamma_e)(1 - \theta). \tag{2.5}$$

Proof. This is a direct application of the definition of conditional probability and the countable additivity axiom of Probability:

$$\begin{aligned} p &= \Pr(T = 1) = \Pr(T = 1, Y = 1) + \Pr(T = 1, Y = 0) \\ &= \Pr(T = 1|Y = 1)\Pr(Y = 1) + \Pr(T = 1|Y = 0)\Pr(Y = 0) \\ &= \Pr(T = 1|Y = 1)\Pr(Y = 1) + (1 - \Pr(T = 0|Y = 0))(1 - \Pr(Y = 1)) \\ &= \gamma_s\theta + (1 - \gamma_e)(1 - \theta). \end{aligned}$$

□

The intuition behind this equation is pretty simple: the proportion of positive test counts the correct identified exposed individuals and the incorrect identified not exposed. Observe that if $\gamma_s = \gamma_e = 1$, we have the trivial case $p = \theta$. Moreover, if $\gamma_s = \gamma_e = 0.5$, we have that $p = 0.5$ and there is no information about θ.

Remark. Actually, we are interested in the prevalence at time t. When it is impossible to test every individual at the same time, we assume that all individuals remain exposed to the disease at time of the last tested individual.

2.3 Bayesian statistics

"Statistics should be considered an interpretation of natural phenomena, rather than an explanation." (ROBERT, 2007)

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.

A key quantity for epidemiologists and public health researchers is the proportion of individuals exposed to a disease at time t, which is called *prevalence*. When measured periodically, its evolution can identify potential causes of the infection and prevention and care methods (NOORDZIJ et al., 2010). The prevalence differs from *incidence* that measures the proportion of people who develop new disease during a specified period of time (ROTHMAN; GREENLAND; LASH, et al., 2008). Therefore, prevalence reflects both incidence and the duration of disease.

This report presents the initial models for my bachelor dissertation entitled "Bayesian analysis of respondent-driven surveys with outcome uncertainty", which proposes to study prevalence when the diagnostic tests are imperfect and the population is hidden, that is, there is no sampling frame for it (HECKATHORN, 1997).

2.4 Generalized linear models

Generalized linear models are an extension of classical linear models. Let $y \in \mathbb{R}^n$ be a realization of a random variable $Y : \Omega \to \mathbb{R}^n$ associated with a phenomena such that each component Y_i is independent of the others. The systematic process in modelling is the specification of the vector $\mu = \mathbb{E}[Y]$ through a small number of parameters $\beta_1, \dots, \beta_p$. The classical linear model assumes that $Y_i \overset{iid}{\sim} \text{Normal}(\mu_i, \sigma^2)$ and $\mu = X\beta$, where $X \in \mathbb{R}^{n \times p}$ is the data, where X_{ij} is the measure of the j-th covariate in the i-th individual.

The main generalization of this aspect is the introduction of the *link function*. This is a monotonic differentiable function g such that $\eta_i = g(\mu_i)$ and $\eta = X\beta$. Therefore the link function relates the linear predictor η to the expected value μ. The distribution of Y

may also come from another exponential family distribution Maybe explain or cite what is this?

Classical link functions when Y_i has Binomial distribution with parameter $0 < \mu < 1$ are

1. *logit*: $\eta = \log(\mu/(1-\mu))$ that represents the log odds of $Y_i = 1$.
2. *probit*: $\eta = \Phi^{-1}(\mu)$ where the $\Phi(\cdot)$ is the Normal cumulative distribution function;
3. *complementary log-log*: $\eta = \log(-\log(1-\mu))$.

2.5 Computational methods

2.5.1 Hamiltonian Monte Carlo

We follow (BETANCOURT, 2017). This method was developed in the late 1980s as Hybrid Monte Carlo to tackle calculations in Lattice Quantum Chromodynamics. Instead of moving in the parameter space randomly with uninformed jumps, the direction from the vector field given by the gradients are used to trace out a trajectory through the *typical set*, the region which has significant contribution to the expectations. However, if only the gradient was used, the trajectory would pull towards the mode of the distribution, so more geometric constraints are needed. In order to a satellite rotate around the Earth, we have to endow ir with enough momentum to counteract the gravitational field, turning the system into a conservative one.

First, we introduce auxiliary momentum parameters p_n (lift) of the same dimension from the parameter space $\Omega \subseteq \mathbb{R}^D$. Then q_n turns to (q_n, p_n), with the use the joint probability distribution $\pi(q,p) = \pi(p \mid q)\pi(q)$. Particularly, we use

$$\pi(q,p) = e^{-H(q,p)},$$

such that H is the *Hamiltonian*. Note that $H(q,p) = -\log \pi(p \mid q) - \log \pi(q) =: K(p,q) + V(q)$. We call K the kinetic energy, and V the potential energy. The vector field is generated by Hamilton's equations,

$$\frac{dq}{dt} = \frac{\partial H}{\partial p} = \frac{\partial K}{\partial p}$$
$$\frac{dp}{dt} = -\frac{\partial H}{\partial q} = -\frac{\partial K}{\partial q} - \frac{dV}{dq}.$$

Therefore, we are able to define the Hamiltonian flows $\phi_t : (p,q) \to (p,q), \forall t \in \mathbb{R}$.

2.5.1.1 Diagnostics

The importance of diagnosing. The potential problems that it can show.

- Divergent transitions;
- Transitions that hit the maximum tree depth;
- Low E-BFMI values;
- Low effective samples sizes;
- $\hat{R} \notin (0.95, 1.05)$.

3 Statistical modelling

Fisher (1922, p. 311) stated that the objective of statistics is to reduce the data since its volume is impossible to comprehend. In that sense, few quantities, generally parameters, should represent the whole phenomenon catching the most relevant information. Years later, Newman studied the theory of modelling which chan be divided in three aspects (LEHMANN, 2012, p. 161):

1. Models of complex phenomena are created by putting together simple building elements that the researcher is familiar with and can handle;
2. There are two types of models: the *explanatory models*, which will be focused on this work, and the *interpolatory formulae*.
3. An explanatory theory necessitates a thorough understanding of the problem's scientific context. In this regard, we investigated this kind of problem involving Respondent-driven sampling and prevalence estimation as introduced in Chapter 2.

In this chapter, we develop models that enclose these ideas building each block separately. For a Bayesian modelling, we assume that each parameter of the model has a probability distribution that incorporates the researcher's uncertainty about it. For each individual, we observe k covariates that are possible risk factors represented by the vector $\boldsymbol{x}_i \in \mathbb{R}^k$ of the i^{th} individual. We denote θ_i the probability of the i-th individual have been exposed to the disease that depends on the prevalence θ and $\boldsymbol{x}_i$. We also consider when it depends on a spatial random effect caused by the connections analysed by the RDS. The probability of positive test in the i^{th} individual is denoted by p_i.

Another important feature of the model is that sensitivity and specificity have the same distribution for all individuals and it only depends on the test used to diagnose. This is an assumption that must be analysed for the studied disease. For instance, COVID-19 tests have different sensibilities and specificities for symptomatic and asymptomatic individuals.

From above, we develop three different models: the first considers perfect tests, that is, $\gamma_s = \gamma_e = 1$ and no spatial random effect; the second considers imperfect tests, regarding γ_s and γ_e, but ignoring the RDS structure; and the third one has imperfect tests and RDS structure.

3.1 Perfect tests

The first model supposes the samples are independent and the test is perfect, which means that $\theta_i = p_i$ for all i. Therefore it only considers the risk factors $\boldsymbol{x}_i$.

$$\begin{aligned} T_i &\sim \text{Bernoulli}(\theta_i), \\ g(\theta_i) &= g(\theta) + \boldsymbol{x}_i^T \beta, \end{aligned} \tag{3.1}$$

where v^T denotes the transpose of v, and $g(\cdot)$ is a link function. The parameter $\beta \in \mathbb{R}^k$ is the risk effects. For Bayesian inference, priors on β and θ must be included. We use $\beta \sim \text{Normal}(\mu, \Sigma)$ and $\theta \sim \text{Beta}(a^p, b^p)$, where $\mu \in \mathbb{R}^k$, $\Sigma \in \mathbb{R}^{k \times k}$ symmetric positive-definite matrix, $a^p \in \mathbb{R}_{++}$, and $b^p \in \mathbb{R}_{++}$ are fixed hyperparameters.

Remark. If the risk factors are zero, i.e $\boldsymbol{x}_i = 0$, the probability of the i^{th} having been exposed is the prevalence θ, which means that in a population with no risk effects, the probability of a person has the disease is exactly the proportion in this population.

3.1.1 Identifiability

3.1.2 Toy example

3.2 Sensitivity and specificity

In this section, we describe a model to infer about sensitivity and specificity separately from the final model for prevalence estimation. This is interesting to experiment and analyse different prior specification approaches. The model is the following: suppose having a gold standard test for a disease and we want to estimate the sensitivity and specificity of another simpler and faster test. In a population, with the gold standard, it's impossible to differentiate the true negatives from the true positives - in this scenario, "true" means the decision of the gold standard. However not all diseases have a perfect gold standard. Therefore, we denote

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

such that $y_{negative}$ are negative tests on known negative subjects and $y_{positive}$ are positive tests on known positive. In a classic Bayesian analysis, we have to define a prior distribution for the parameters $(\gamma_e, \gamma_s) \sim \pi(\gamma_e, \gamma_s)$.

For this, we consider three different approaches:

1. Treating each parameter as independent with a beta distribution with pre-specified hyperparameters.

2. Hierarchical partial pooling, when dealing with more studies.

3. A Bivariate Beta (see Appendix A) distribution.

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*.

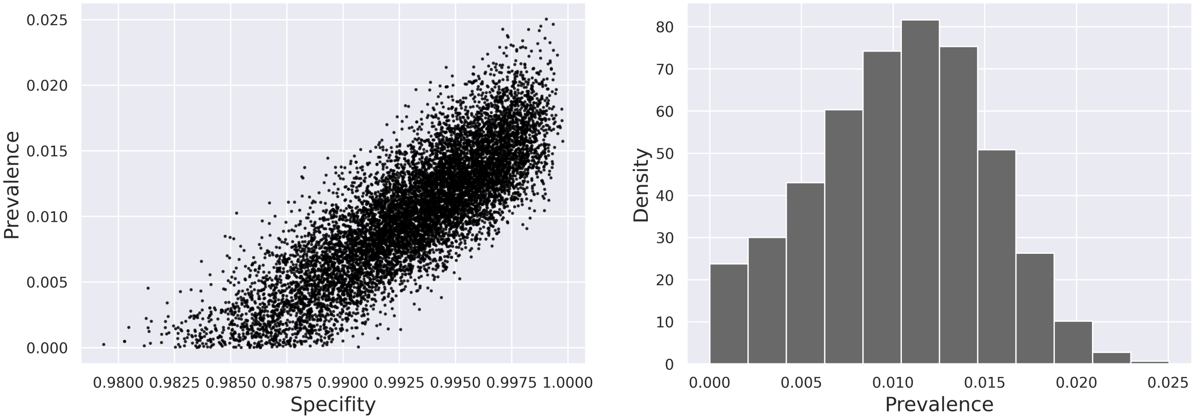


Figure 1 – Scatter plot of posterior simulations of prevalence against specificity and histogram of posterior simulations of the prevalence.

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:

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

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.

In this section, we shall describe how to use the Bivariate Beta (see Appendix A) to model the correlation between specificity and sensitivity.

3.2.1 Specifying parameters $\boldsymbol{\alpha}$

Suppose that the researcher has knowledge about the main moments of X and Y, such that $\mathbb{E}(X) = m_1 \in (0, 1), \mathbb{E}(Y) = m_2 \in (0, 1), \text{Var}(X) = v_1 \in (0, 1)$, and $\text{Var}(Y) = v_2 \in (0, 1)$. Notice that $v_1 + m_1^2 = \text{Var}(X_1) + \mathbb{E}[X_1]^2 = \mathbb{E}[X_1^2]$ and

$$\mathbb{E}[X_1^2] - \mathbb{E}[X_1] = \frac{(\alpha_1 + \alpha_2 + 1)(\alpha_1 + \alpha_2)}{(\tilde{\alpha} + 1)\tilde{\alpha}} - \frac{\alpha_1 + \alpha_2}{\tilde{\alpha}} = -\frac{(\alpha_1 + \alpha_2)(\alpha_3 + \alpha_4)}{\tilde{\alpha}(\tilde{\alpha} + 1)} < 0,$$

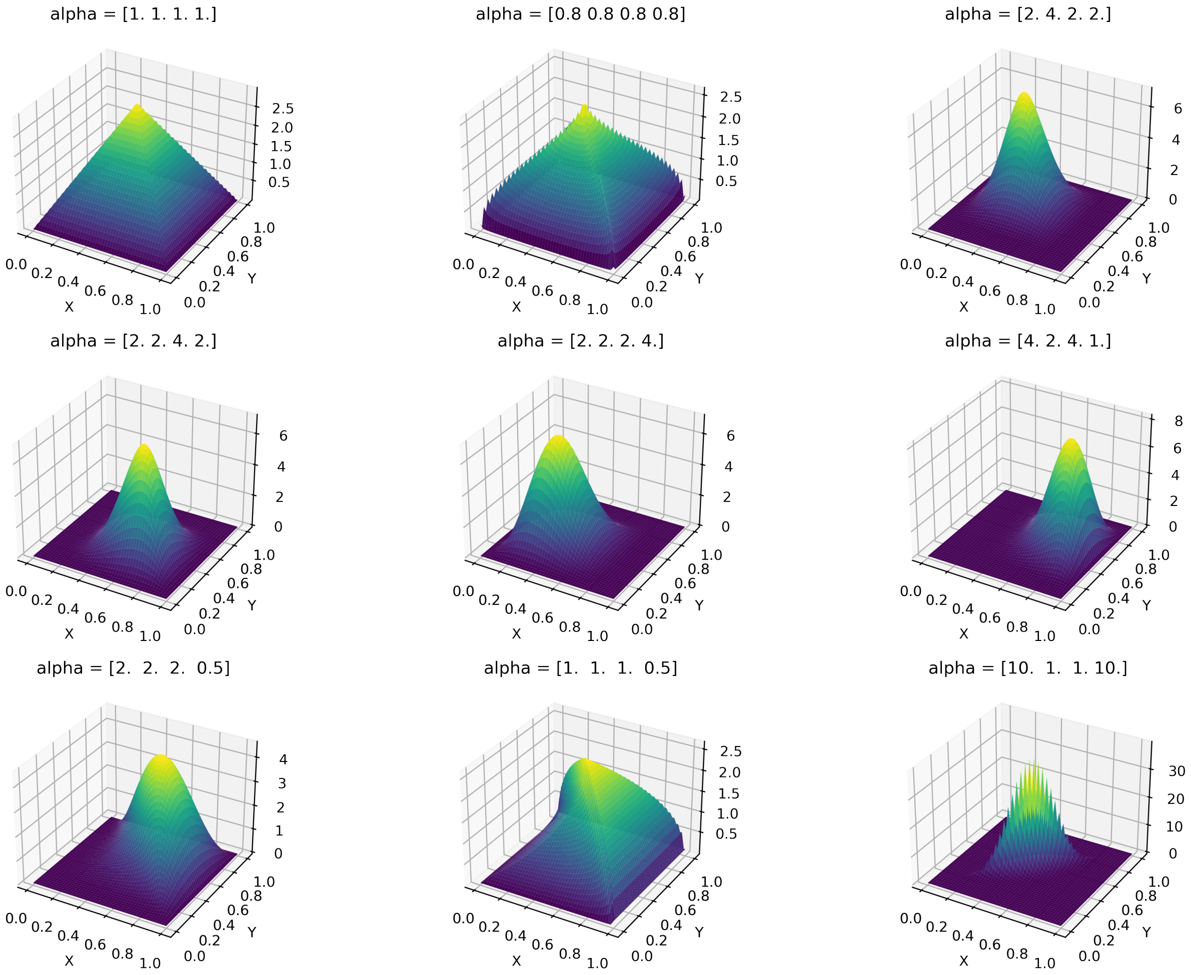
that is, $v_1 + m_1^2 - m_1 < 0 \implies v_1 < m_1 - m_1^2$ and similarly, $v_2 < m_2 - m_2^2$. After fixing these quantities, we will have a non-linear system with four equations and four unknown variables. Hence, we want to solve the following

$$\begin{cases}
m_1 = \dfrac{\alpha_1 + \alpha_2}{\tilde{\alpha}} \\
m_2 = \dfrac{\alpha_1 + \alpha_3}{\tilde{\alpha}} \\
v_1 = \dfrac{(\alpha_1 + \alpha_2)(\alpha_3 + \alpha_4)}{\tilde{\alpha}^2(\tilde{\alpha} + 1)} = m_1 \dfrac{\alpha_3 + \alpha_4}{\tilde{\alpha}(\tilde{\alpha} + 1)} \\
v_2 = \dfrac{(\alpha_1 + \alpha_3)(\alpha_2 + \alpha_4)}{\tilde{\alpha}^2(\tilde{\alpha} + 1)} = m_2 \dfrac{\alpha_2 + \alpha_4}{\tilde{\alpha}(\tilde{\alpha} + 1)}.
\end{cases} \tag{3.2}$$

Proposition 1. *System (3.2) has a solution if, and only if, the relation*

$$v_2 = \frac{(1 - m_2)\tilde{\alpha}}{\tilde{\alpha}(\tilde{\alpha} + 1)} = \frac{1 - m_2}{\frac{m_1 - m_1^2}{v_1}} = \frac{v_1(1 - m_2)}{m_1(1 - m_1)}, \tag{3.3}$$

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



alpha = [1. 1. 1. 1.]

alpha = [0.8 0.8 0.8 0.8]

alpha = [2. 4. 2. 2.]

alpha = [2. 2. 4. 2.]

alpha = [2. 2. 2. 4.]

alpha = [4. 2. 4. 1.]

alpha = [2. 2. 2. 0.5]

alpha = [1. 1. 1. 0.5]

alpha = [10. 1. 1. 10.]

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

is satisfied. When there is a solution, there will be infinitely many and they all lay in the ray

$$\mathcal{L} = \{(1, -1, -1, 1)\alpha_4 + k : \alpha_4 > 0\},$$

such that $k = ((m_1 + m_2 - 1)\tilde{\alpha}, (1 - m_2)\tilde{\alpha}, (1 - m_1)\tilde{\alpha}, 0)$*.*

Proof. The first two equations of the system (3.2) can be rewritten as a linear system:

$$\begin{aligned}(m_1 - 1)\alpha_1 + (m_1 - 1)\alpha_2 + m_1\alpha_3 + m_1\alpha_4 = 0\\ (m_2 - 1)\alpha_1 + m_2\alpha_2 + (m_2 - 1)\alpha_3 + m_2\alpha_4 = 0,\end{aligned}$$

which is equivalent to

$$\begin{aligned}\alpha_1 + \alpha_2 + \frac{m_1}{m_1 - 1}\alpha_3 + \frac{m_1}{m_1 - 1}\alpha_4 &= 0\\ \alpha_2 + \frac{1 - m_2}{m_1 - 1}\alpha_3 + \frac{m_1 - m_2}{m_1 - 1}\alpha_4 &= 0.\end{aligned}$$

Then, we can write α_1 and α_2 as functions of α_3 and α_4:

$$\alpha_1 = \frac{m_1 + m_2 - 1}{1 - m_1}\alpha_3 + \frac{m_2}{1 - m_1}\alpha_4 \tag{3.4}$$

$$\alpha_2 = \frac{1 - m_2}{1 - m_1}\alpha_3 + \frac{m_1 - m_2}{1 - m_1}\alpha_4. \tag{3.5}$$

With that expression, let $\alpha_1 = a_3\alpha_3 + a_4\alpha_4$ and $\alpha_2 = b_3\alpha_3 + b_4\alpha_4$. Denote $c_3 = a_3 + b_3 + 1$ and $c_4 = a_4 + b_4 + 1$. Then, consider the third equation of the system (3.2),

$$\begin{aligned}
&\frac{v_1}{m_1} = \frac{\alpha_3 + \alpha_4}{\tilde{\alpha}(\tilde{\alpha} + 1)} = \frac{\alpha_3 + \alpha_4}{(\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4)^2 + (\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4)} \\
&\implies \frac{v_1}{m_1}(\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4)^2 = \alpha_3 + \alpha_4 - \frac{v_1}{m_1}(\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4) \\
&\implies \frac{v_1}{m_1}(c_3\alpha_3 + c_4\alpha_4)^2 = \left(1 - \frac{v_1}{m_1}c_3\right)\alpha_3 + \left(1 - \frac{v_1}{m_1}c_4\right)\alpha_4 \\
&\implies \frac{v_1 c_3^2}{m_1}\alpha_3^2 + \left(\frac{2v_1c_3c_4\alpha_4 + v_1c_3}{m_1} - 1\right)\alpha_3 + \left(\frac{v_1c_4^2\alpha_4^2 + v_1c_4\alpha_4}{m_1} - \alpha_4\right) = 0 \\
&\implies v_1c_3^2\alpha_3^2 + (2v_1c_3c_4\alpha_4 + v_1c_3 - m_1)\alpha_3 + (v_1c_4^2\alpha_4^2 + v_1c_4\alpha_4 - m_1\alpha_4) = 0.
\end{aligned}$$

Using a Computer Algebra System (CAS) with the Python library SymPy, the above expression can be simplified as follows:

$$v_1\alpha_3^2 + \left(v_1(1 - m_1) + 2v_1\alpha_4 - m_1(1 - m_1)^2\right)\alpha_3 - \alpha_4 m_1(1 - m_1)^2 + \alpha_4 v_1(1 - m_1) + v_1\alpha_4^2 = 0.$$

This way, the solutions of the above equation are function of α_4. Therefore, after solving the equations, we can use the last equation of the system (3.2) as a function on of α_4. Let,

$$\Lambda = \left(v_1(1 - m_1) + v_1\alpha_4 - m_1(1 - m_1)^2\right).$$

Then,

$$\begin{aligned}
\Delta &= \left(v_1(1 - m_1) + 2v_1\alpha_4 - m_1(1 - m_1)^2\right)^2 - 4v_1(\alpha_4 v_1(1 - m_1) - \alpha_4 m_1(1 - m_1)^2 + v_1\alpha_4^2), \\
&= (\Lambda + v_1\alpha_4)^2 - 4v_1\alpha_4\Lambda \\
&= \Lambda^2 - 2\Lambda v_1\alpha_4 + (v_1\alpha_4)^2 \\
&= (\Lambda - v_1\alpha_4)^2 \\
&= \left(v_1(1 - m_1) - m_1(1 - m_1)^2\right)^2 \\
&= (1 - m_1)^2(v_1 + m_1^2 - m_1)^2.
\end{aligned}$$

Note that $v_1 + m_1^2 = \text{Var}(X_1) + \mathbb{E}[X_1]^2 = \mathbb{E}[X_1^2]$ and

$$\mathbb{E}[X_1^2] - \mathbb{E}[X_1] = \frac{(\alpha_1 + \alpha_2 + 1)(\alpha_1 + \alpha_2)}{(\tilde{\alpha} + 1)\tilde{\alpha}} - \frac{\alpha_1 + \alpha_2}{\tilde{\alpha}} = -\frac{(\alpha_1 + \alpha_2)(\alpha_3 + \alpha_4)}{\tilde{\alpha}(\tilde{\alpha} + 1)} < 0.$$

Therefore,

$$\sqrt{\Delta} = (1 - m_1)(m_1 - v_1 - m_1^2)$$

and

$$\begin{aligned}\alpha_3 &= \frac{1}{2v_1}\left(\left(m_1(1-m_1)^2 - v_1(1-m_1) - 2v_1\alpha_4\right) \pm (1-m_1)(m_1 - v_1 - m_1^2)\right)\\ &= -\alpha_4 + \frac{(1-m_1)(m_1-m_1^2-v_1) \pm (1-m_1)(m_1-v_1-m_1^2)}{2v_1}.\end{aligned}$$

When the sign is negative, we have that $\alpha_3 = -\alpha_4$, an impossible solution. Then,

$$\alpha_3 = \frac{(1-m_1)(m_1-m_1^2-v_1)}{v_1} - \alpha_4.$$

We summarize the expressions in function of α_4:

$$\begin{aligned}\alpha_3 &= \frac{(1-m_1)(m_1-m_1^2-v_1)}{v_1} - \alpha_4\\ \alpha_1 &= \frac{m_1+m_2-1}{1-m_1}\alpha_3 + \frac{m_2}{1-m_1}\alpha_4 = \frac{(m_1+m_2-1)(m_1-m_1^2-v_1)}{v_1} + \alpha_4\\ \alpha_2 &= \frac{1-m_2}{1-m_1}\alpha_3 + \frac{m_1-m_2}{1-m_1}\alpha_4 = \frac{(1-m_2)(m_1-m_1^2-v_1)}{v_1} - \alpha_4.\end{aligned}$$

From here, one can calculate that

$$\tilde{\alpha} = \frac{m_1 - m_1^2 - v_1}{v_1}.$$

Since $\alpha_2 + \alpha_4 = (1-m_2)\tilde{\alpha}$, we have that the last equation of the system (3.2) is given by (3.3), that is, the system (3.2) has a solution if and only if, equation (3.3) is satisfied. If it is, the solution is the ray

$$\mathcal{L} = \{(1,-1,-1,1)\alpha_4 + k : \alpha_4 > 0\},$$

such that $k = ((m_1+m_2-1)\tilde{\alpha}, (1-m_2)\tilde{\alpha}, (1-m_1)\tilde{\alpha}, 0)$.

□

Now change the fourth equation of (3.2) by:

$$\mathrm{Cor}(X,Y) = \frac{\alpha_1\alpha_4 - \alpha_2\alpha_3}{\sqrt{(\alpha_1+\alpha_2)(\alpha_3+\alpha_4)(\alpha_1+\alpha_3)(\alpha_2+\alpha_4)}} = \frac{\alpha_1\alpha_4 - \alpha_2\alpha_3}{\tilde{\alpha}^2\sqrt{m_1m_2(1-m_1)(1-m_2)}}$$

Supposing the expression for α_1, α_2 and α_3, that is, m_1, m_2 and v_1 are fixed, and supposing we fix $\rho = \mathrm{Cor}(X,Y)$, we can simplify the above expression (using a software) as follows:

$$\rho = \frac{1}{\tilde{\alpha}\sqrt{m_1m_2(1-m_1)(1-m_2)}}\alpha_4 - \sqrt{\frac{(1-m_1)(1-m_2)}{m_1m_2}},$$

which is linear on α_4, that is, for fixed values of m_1, m_2, v_1 and ρ, there is an unique α_4, and hence, α_1, α_2 and α_3 that satisfies system (3.2) with the fourth equation changed by the correlation.

3.3 Imperfect tests

This model includes the sensitivity and specificity of the diagnostic test.

$$\begin{aligned} T_i &\sim \text{Bernoulli}(p_i) \\ p_i &= \gamma_s\theta_i + (1-\gamma_e)(1-\theta_i), \\ g(\theta_i) &= g(\theta) + \boldsymbol{x}_i^T\beta, \\ \beta &\sim \text{Normal}(\mu, \Sigma), \\ \theta &\sim \text{Beta}(a^p, b^p) \\ \gamma_s &\sim \text{Beta}(a^s, b^s), \\ \gamma_e &\sim \text{Beta}(a^e, b^e), \end{aligned} \tag{3.6}$$

where $a^p, a^s, a^e, b^p, b^s, b^e \in \mathbb{R}_{++}$ are fixed hyperparameters. This model does not include prior knowledge about the correlation between specificity and sensitivity.

3.3.1 Toy example

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.

3.4 Imperfect tests and respondent-driven sampling

For now, we consider the network dependence induced by the RDS with no associated model. Therefore, we treat it as a random effect for each individual. Conditionally autoregressive (CAR) models in the Gaussian case are used. Let $[\tilde{Q}]_{ij} = \tilde{q}_{ij}$ be a fixed matrix which measures the distance between i and j, and $\tilde{q}_{i+} = \sum_j \tilde{q}_{ij}$. In general, we use

$$\tilde{q}_{ij} = \begin{cases} 1, & \text{if } i \text{ recruited } j \text{ or the contrary} \\ 0, & \text{otherwise.} \end{cases}$$

Next we define the scaled adjacency matrix $Q = D^{-1}\tilde{Q}$, such that D is a diagonal matrix with $D_{ii} = \tilde{q}_{i+}$. Finally let $|\rho| < 1$ be a parameter to controls the dependence between neighbors. Hence, we specify the model as follows:

$$\begin{aligned}
T_i &\sim \text{Bernoulli}(p_i) \\
p_i &= \gamma_s\theta_i + (1-\gamma_e)(1-\theta_i), \\
g(\theta_i) &= g(\theta) + \boldsymbol{x}_i^T\beta + \omega_i, \\
\omega_i|\{\omega_j\}_{j\neq i}, \tau &\sim \text{Normal}\left(\rho\sum_j q_{ij}\omega_j, \tau^{-1}/\tilde{q}_{i+}\right) \\
\beta &\sim \text{Normal}(\mu, \Sigma), \\
\theta &\sim \text{Beta}(a^p, b^p) \\
\gamma_s &\sim \text{Beta}(a^s, b^s), \\
\gamma_e &\sim \text{Beta}(a^e, b^e), \\
\tau &\sim \text{Gamma}(a^\tau, b^\tau).
\end{aligned} \tag{3.7}$$

By Brook's Lemma (BROOK, 1964), the joint distribution of ω can be specified as

$$\omega \sim \text{Normal}\left(0, \left[\tau(D-\rho\tilde{Q})\right]^{-1}\right).$$

3.4.1 Toy example

1. Between the model with the log odds of prevalence having a Gaussian prior distribution and the other with the prevalence having a Beta prior distribution, the latter was usually faster and without divergences. Therefore the preferable model is with the prevalence.

2. Non-centered distributions are really worst.

3. Comparison between parametrization of sigma and tau showed that they are similar in sight of time of execution, energy and divergences, among others diagnostics. However, the mean estimate of sigma is more controlled. The median estimate is very similar. This happens because there are a few very high samples for τ that will have high weight in the final result. Small samples for σ have less impact, despite having some.

4. I observed that the RDS structure worsens the funil aspect of the scatter plot. Although the sparsity may cause divergence problems, which makes sense, sparsity and tree structure appears to be a bad feature for the model. Apparently there is no connection with the number of connected components. Test with one and five had similar results (actually they were better in the five case, but the difference can be random noise).

3.4.2 Exponential Random Graph Model (ERGM)

RDS has the constraint of being without replacement. For that reason, we do not observe all links among the samples (CRAWFORD, 2016). Considering the model developed by Crawford, we can model the matrix Q as *Exponential Random Graph Model* (ERGM). Define the following

1. $\boldsymbol{s} = \text{tril}(QC)^T\mathbf{1} + C^Tu$, such that Q is the adjacency matrix of the recruited subjects, C is the *Coupon Matrix*, u the vector of the number of edge ends belonging to each vertex (in the order of recruitment) that are not connected to any other sampled vertex, and $\text{tril}(M)$ the lower triangle of M.
2. $T(Q) = -\lambda\boldsymbol{s}$, such that λ is the rate of the recruitment time.
3. $V(Q) = \sum_{k \text{ is not seed}} \log(\lambda\boldsymbol{s}_k)$
4. $w = (0, t_2 - t_1, ..., t_n - t_{n-1})$ is the vector of the waiting times between recruitments.

Therefore $\Pr(Q|w) \propto \exp[T(Q)^Tw + V(Q)]$. With that, the model becomes

$$\begin{aligned}
T_i &\sim \text{Bernoulli}(p_i) \\
p_i &= \gamma_s\theta_i + (1-\gamma_e)(1-\theta_i), \\
g(\theta_i) &= g(\theta) + \boldsymbol{x}_i^T\beta + \omega_i, \\
\omega_i|\{\omega_j\}_{j\neq i}, \tau &\sim \text{Normal}\left(\rho\sum_j q_{ij}\omega_j/q_{i+}, \tau^2/q_{i+}\right) \\
Q|w &\propto \exp[T(Q)^Tw + V(Q)] \\
\lambda &\sim \Gamma(a^\lambda, b^\lambda), \\
\beta &\sim \text{Normal}(\mu, \Sigma), \\
\theta &\sim \text{Beta}(a^p, b^p) \\
\gamma_s &\sim \text{Beta}(a^s, b^s), \\
\gamma_e &\sim \text{Beta}(a^e, b^e), \\
\tau &\sim \text{Normal}^+(0, \sigma_\tau^2).
\end{aligned} \tag{3.8}$$

The problem with this model is that we are assigning a posterior distribution for Q.

4 Discussion about prior distributions and sensitivity analysis

4.1 Prior analysis of sensitivity and specificity

4.2 Prior analysis on the parameter tau

4.3 Prior analysis on theta

5 Real data applications

6 Conclusion

Parte final do trabalho, apresenta as conclusões correspondentes aos objetivos ou hipóteses.

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.

BETANCOURT, Michael. A conceptual introduction to Hamiltonian Monte Carlo. **arXiv preprint arXiv:1701.02434**, 2017.

BROOK, D. On the distinction between the conditional probability and the joint probability approaches in the specification of nearest-neighbour systems. **Biometrika**, JSTOR, v. 51, n. 3/4, p. 481–483, 1964.

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>.

FISHER, Ronald A. On the mathematical foundations of theoretical statistics. **Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character**, The Royal Society London, v. 222, n. 594-604, p. 309–368, 1922.

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>.

LEHMANN, Eric L. Model specification: the views of Fisher and Neyman, and later developments. In: SELECTED Works of EL Lehmann. [S.l.]: Springer, 2012. P. 955–963.

LIN, Jiayu. On the dirichlet distribution. **Mater's Report**, Queen's University Kingston Ontario, Canada, 2016.

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.

NOORDZIJ, Marlies et al. Measures of disease frequency: prevalence and incidence. **Nephron Clinical Practice**, Karger Publishers, v. 115, n. 1, p. c17–c20, 2010.

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.

ROTHMAN, Kenneth J; GREENLAND, Sander; LASH, Timothy L, et al. **Modern epidemiology**. [S.l.]: Wolters Kluwer Health/Lippincott Williams & Wilkins Philadelphia, 2008. v. 3.

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>.

Appendix

APPENDIX A – Bivariate Beta distribution

Let $U = (U_1, U_2, U_3, U_4) \sim \text{Dirichlet}(\boldsymbol{\alpha})$, where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ with $\alpha_i > 0, i = 1, \ldots, 4$ and $U_4 = 1 - U_1 + U_2 + U_3$. The joint density of U with respect to the Lebesgue measure is given by

$$f_U(u_1, u_2, u_3) = \frac{1}{B(\boldsymbol{\alpha})} u_1^{\alpha_1 - 1} u_2^{\alpha_2 - 1} u_3^{\alpha_3 - 1} (1 - u_1 - u_2 - u_3)^{\alpha_4 - 1}, \tag{A.1}$$

when $u_i \in [0, 1], i = 1, 2, 3,$ $u_1 + u_2 + u_3 \leq 1$, and 0 otherwise. The normalizing constant is, for $v \in \mathbb{R}^n$,

$$B(v) = \frac{\prod_{i=1}^n \Gamma(v_i)}{\Gamma\left(\sum_{i=1}^n v_i\right)}.$$

Definition A.0.1. Let

$$X = U_1 + U_2 \text{ and } Y = U_1 + U_3. \tag{A.2}$$

The distribution of (X, Y) is *Bivariate Beta* with parameters $\boldsymbol{\alpha}$.

Proposition 2. *The marginal distribution of X is Beta with parameters $\alpha_1 + \alpha_2$ and $\alpha_3 + \alpha_4$. Similarly, the marginal distribution of Y is Beta with parameters $\alpha_1 + \alpha_3$ and $\alpha_2 + \alpha_4$.*

Proof. First we derive the probability density of (U_1, U_2) with respect to the Lebesgue measure.

$$\begin{aligned} f_{U_1,U_2}(u_1, u_2) &= \int_{-\infty}^{\infty} f_U(u_1, u_2, u_3)\, du_3 \\ &= \frac{1}{B(\boldsymbol{\alpha})} \int_0^1 u_1^{\alpha_1 - 1} u_2^{\alpha_2 - 1} u_3^{\alpha_3 - 1} (1 - u_1 - u_2 - u_3)^{\alpha_4 - 1}\, du_3 \\ &= \frac{1}{B(\boldsymbol{\alpha})} u_1^{\alpha_1 - 1} u_2^{\alpha_2 - 1} \int_0^1 u_3^{\alpha_3 - 1} (1 - u_1 - u_2 - u_3)^{\alpha_4 - 1}\, du_3. \end{aligned} \tag{A.3}$$

Let $u_3 = (1 - u_1 - u_2)z$. Then,

$$\begin{aligned} f_{U_1,U_2}(u_1, u_2) &= \frac{1}{B(\boldsymbol{\alpha})} u_1^{\alpha_1 - 1} u_2^{\alpha_2 - 1} \int_0^1 (1 - u_1 - u_2)^{\alpha_3 - 1} z^{\alpha_3 - 1} (1 - u_1 - u_2)^{\alpha_4} (1 - z)^{\alpha_4 - 1}\, dz. \\ &= \frac{1}{B(\boldsymbol{\alpha})} u_1^{\alpha_1 - 1} u_2^{\alpha_2 - 1} (1 - u_1 - u_2)^{\alpha_3 + \alpha_4 - 1} \int_0^1 z^{\alpha_3 - 1} (1 - z)^{\alpha_4 - 1}\, dz. \\ &= \frac{1}{B(\boldsymbol{\alpha})} u_1^{\alpha_1 - 1} u_2^{\alpha_2 - 1} (1 - u_1 - u_2)^{\alpha_3 + \alpha_4 - 1} \frac{\Gamma(\alpha_3)\Gamma(\alpha_4)}{\Gamma(\alpha_3 + \alpha_4)} \\ &= \frac{1}{B(\alpha_1, \alpha_2, \alpha_3 + \alpha_4)} u_1^{\alpha_1 - 1} u_2^{\alpha_2 - 1} (1 - u_1 - u_2)^{\alpha_3 + \alpha_4 - 1}. \end{aligned} \tag{A.4}$$

We conclude that

$$(U_1, U_2, 1 - U_1 - U_2) \sim \text{Dirichlet}(\alpha_1, \alpha_2, \alpha_3 + \alpha_4).$$

Define

$$H(v) = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} v, \text{ for } v \in \mathbb{R}^2.$$

Then $(U_1, X) = H(U_1, U_2)$ and $H(\cdot)$ is bijective and differentiable function. By the Change of Variable Formula,

$$\begin{aligned} f_{U_1,X}(u_1, x) &= f(H^{-1}(u_1, x)) \left| \det \left[\frac{dH^{-1}(v)}{dv} \bigg|_{v=(u_1,x)} \right] \right| \\ &= f(u_1, x - u_1) = \frac{1}{B(\alpha_1, \alpha_2, \alpha_3 + \alpha_4)} u_1^{\alpha_1 - 1}(x - u_1)^{\alpha_2 - 1}(1 - x)^{\alpha_3 + \alpha_4 - 1}, \end{aligned} \tag{A.5}$$

where (u_1, x) belongs to the triangle defined by the points (0,0), (0,1), and (1,1). The distribution of X for $x \in [0, 1]$ is

$$\begin{aligned} f_X(x) &= \frac{1}{B(\alpha_1, \alpha_2, \alpha_3 + \alpha_4)} \int_0^x u_1^{\alpha_1 - 1}(x - u_1)^{\alpha_2 - 1}(1 - x)^{\alpha_3 + \alpha_4 - 1} \, du_1 \\ &= \frac{1}{B(\alpha_1, \alpha_2, \alpha_3 + \alpha_4)} (1 - x)^{\alpha_3 + \alpha_4 - 1} \int_0^x u_1^{\alpha_1 - 1}(x - u_1)^{\alpha_2 - 1} \, du_1. \\ &= \frac{1}{B(\alpha_1, \alpha_2, \alpha_3 + \alpha_4)} (1 - x)^{\alpha_3 + \alpha_4 - 1} \int_0^x x^{\alpha_1 - 1} \left(\frac{u_1}{x} \right)^{\alpha_1 - 1} x^{\alpha_2 - 1} \left(1 - \frac{u_1}{x} \right)^{\alpha_2 - 1} \, du_1. \end{aligned} \tag{A.6}$$

Setting $u = u_1/x$ (if $x = 0, f_X(x) = 0$, then suppose $x > 0$), we have,

$$\begin{aligned} f_X(x) &= \frac{1}{B(\alpha_1, \alpha_2, \alpha_3 + \alpha_4)} (1 - x)^{\alpha_3 + \alpha_4 - 1} x^{\alpha_1 + \alpha_2 - 1} \int_0^1 u^{\alpha_1 - 1}(1 - u)^{\alpha_2 - 1} \, du. \\ &= \frac{1}{B(\alpha_1, \alpha_2, \alpha_3 + \alpha_4)} (1 - x)^{\alpha_3 + \alpha_4 - 1} x^{\alpha_1 + \alpha_2 - 1} B(\alpha_1, \alpha_2) \qquad \text{(A.7)} \\ &= \frac{1}{B(\alpha_1 + \alpha_2, \alpha_3 + \alpha_4)} (1 - x)^{\alpha_3 + \alpha_4 - 1} x^{\alpha_1 + \alpha_2 - 1} \end{aligned}$$

Therefore $X \sim \text{Beta}(\alpha_1 + \alpha_2, \alpha_3 + \alpha_4)$. Similarly $Y \sim \text{Beta}(\alpha_1 + \alpha_3, \alpha_2 + \alpha_4)$.

□

Proposition 3. *The joint density of (X, Y) with respect to the Lebesgue measure is given by*

$$f_{X,Y}(x, y) = \frac{1}{B(\boldsymbol{\alpha})} \int_\Omega u_1^{\alpha_1 - 1}(x - u_1)^{\alpha_2 - 1}(y - u_1)^{\alpha_3 - 1}(1 - x - y + u_1)^{\alpha_4 - 1} \, du_1, \tag{A.8}$$

where

$$\Omega = (\max(0, x + y - 1), \min(x, y)).$$

Proof. Note that

$$\begin{bmatrix} U_1 \\ X \\ Y \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} U_1 \\ U_2 \\ U_3 \end{bmatrix},$$

where the linear function is bijective and differentiable function, such that the determinant of the derivative is 1. By the Change of Variable Formula,

$$\begin{aligned} f_{U_1,X,Y}(u_1,x,y) &= f_{U_1,U_2,U_3}(u_1, x-u_1, y-u_2) \\ &= \frac{1}{B(\boldsymbol{\alpha})} u_1^{\alpha_1-1}(x-u_1)^{\alpha_2-1}(y-u_1)^{\alpha_3-1}(1-x-y+u_1)^{\alpha_4-1}, \end{aligned} \tag{A.9}$$

where $0 \leq u_1 \leq x, u_1 \leq y$, and $0 \leq 1-x-y+u_1$. Hence,

$$f_{X,Y}(x,y) = \frac{1}{B(\boldsymbol{\alpha})} \int_\Omega u_1^{\alpha_1-1}(x-u_1)^{\alpha_2-1}(y-u_1)^{\alpha_3-1}(1-x-y+u_1)^{\alpha_4-1}\, du_1, \tag{A.10}$$

such that $\Omega = \{u_1 : \max(0, x+y-1) < u_1 < \min(x,y)\}$. □

Proposition 4. *The covariance between X and Y is*

$$\operatorname{Cov}(X,Y) = \frac{1}{\tilde{\alpha}^2(\tilde{\alpha}+1)}(\alpha_1\alpha_4 - \alpha_2\alpha_3).$$

Proof. Let $\tilde{a} = \sum_i \alpha_i$. The covariance between U_i and U_j is (LIN, 2016)

$$\operatorname{Cov}(U_i,U_j) = -\frac{\alpha_i\alpha_j}{\tilde{\alpha}^2(\tilde{\alpha}+1)}, i,j = 1,...,4, i \neq j \tag{A.11}$$

and the variance of U_i is

$$\operatorname{Var}(U_i) = \frac{\alpha_i(\tilde{\alpha}-\alpha_i)}{\tilde{\alpha}^2(\tilde{\alpha}+1)}, \tag{A.12}$$

since $U_i \sim \text{Beta}(\alpha_i, \tilde{\alpha}-\alpha_i)$. Therefore

$$\operatorname{Cov}(X,Y) = \operatorname{Cov}(U_1+U_2, U_1+U_3) = \frac{1}{\tilde{\alpha}^2(\tilde{\alpha}+1)}(\alpha_1\alpha_4 - \alpha_2\alpha_3) \tag{A.13}$$

□

The main moments of X and Y are the following

$$\begin{aligned} \mathbb{E}(X) = \mathbb{E}(U_1+U_2) &= \frac{\alpha_1+\alpha_2}{\alpha_1+\alpha_2+\alpha_3+\alpha_4} \\ \mathbb{E}(Y) = \mathbb{E}(U_1+U_3) &= \frac{\alpha_1+\alpha_3}{\alpha_1+\alpha_2+\alpha_3+\alpha_4} \\ \operatorname{Var}(X) = \operatorname{Cov}(U_1+U_2, U_1+U_2) &= \frac{1}{\tilde{\alpha}^2(\tilde{\alpha}+1)}(\alpha_1+\alpha_2)(\alpha_3+\alpha_4) \\ \operatorname{Var}(Y) = \operatorname{Cov}(U_1+U_3, U_1+U_3) &= \frac{1}{\tilde{\alpha}^2(\tilde{\alpha}+1)}(\alpha_1+\alpha_3)(\alpha_2+\alpha_4) \\ \operatorname{Cor}(X,Y) = \frac{\operatorname{Cov}(X,Y)}{\sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}} &= \frac{\alpha_1\alpha_4 - \alpha_2\alpha_3}{\sqrt{(\alpha_1+\alpha_2)(\alpha_3+\alpha_4)(\alpha_1+\alpha_3)(\alpha_2+\alpha_4)}} \end{aligned}$$

Remark. The original paper has a mistake at page 6.

A.1 Comments about integration

The density of (X, Y) with respect to the Lebesgue measure is $f_{X,Y}(x, y)$ as in equation (A.10). Therefore it can be undefined in sets of null Lebesgue measure in $\mathbb{R}^2$. This section aims to find them to help writing the function properly. If $\alpha_i \geq 1$, $i = 1, ..., 4$, the integral is clearly well defined for every $x, y \in [0, 1]$. Let $0 < \alpha_2 = \alpha_3 = a \leq 0.5$ and $x = y < 0.5$. Then

$$\begin{aligned} f_{X,Y}(x,y) &= \frac{1}{B(\boldsymbol{\alpha})}\int_0^x u_1^{\alpha_1-1}(x-u_1)^{a-1}(x-u_1)^{a-1}(1-2x+u_1)^{\alpha_4-1}\,du_1 \\ &= \frac{1}{B(\boldsymbol{\alpha})}\int_0^{x/2} u_1^{\alpha_1-1}(x-u_1)^{2a-2}(1-2x+u_1)^{\alpha_4-1}\,du_1 + \\ &\quad + \frac{1}{B(\boldsymbol{\alpha})}\int_{x/2}^{x} u_1^{\alpha_1-1}(x-u_1)^{2a-2}(1-2x+u_1)^{\alpha_4-1}\,du_1 \end{aligned}$$

Note that the first integral is well defined and non-negative. If $\alpha_1 \geq 1$,

$$\begin{aligned} &\int_0^{x/2} u_1^{\alpha_1-1}(x-u_1)^{2a-2}(1-2x+u_1)^{\alpha_4-1}\,du_1 \\ &\qquad \leq \int_0^{x/2} \frac{x}{2}^{\alpha_1-1}\left(\frac{x}{2}\right)^{2a-2} \max\left(\left(1-\frac{3}{2}x\right)^{\alpha_4-1}, (1-2x)^{\alpha_4-1}\right) du_1 < +\infty. \end{aligned}$$

If $0 < \alpha_1 < 1$,

$$\begin{aligned} &\int_0^{x/2} u_1^{\alpha_1-1}(x-u_1)^{2a-2}(1-2x+u_1)^{\alpha_4-1}\,du_1 \\ &\qquad = \lim_{t\to 0^+}\int_t^{x/2} u_1^{\alpha_1-1}\left(\frac{x}{2}\right)^{2a-2} \max\left(\left(1-\frac{3}{2}x\right)^{\alpha_4-1}, (1-2x)^{\alpha_4-1}\right) du_1 \\ &\qquad = K(x)\lim_{t\to 0^+}\int_t^{x/2} u_1^{\alpha_1-1}\,du_1 \\ &\qquad = \frac{K(x)}{\alpha_1}\lim_{t\to 0^+}\left[\left(\frac{x}{2}\right)^{\alpha_1} - t^{\alpha_1}\right] < +\infty. \end{aligned}$$

where $K(x)$ is a function of x. Moreover, since the integrand is non-negative, so is the integral. On the other hand, the second integral is not defined:

$$\begin{aligned} &\int_{x/2}^{x} u_1^{\alpha_1-1}(x-u_1)^{2a-2}(1-2x+u_1)^{\alpha_4-1}\,du_1 \\ &\qquad \geq \int_{x/2}^{x} \min\left(\left(\frac{x}{2}\right)^{\alpha_1-1}, x^{\alpha_1-1}\right)(x-u_1)^{2a-2}\min\left(\left(1-\frac{3}{2}x\right)^{\alpha_4-1}, (1-x)^{\alpha_4-1}\right) du_1 \\ &\qquad = K'(x)\int_0^{x/2} v^{2a-2}\,dv \\ &\qquad = \begin{cases} \dfrac{K'(x)}{2a-1}\lim_{t\to 0^+}\left[(x/2)^{2a-1} - t^{2a-1}\right] & \text{if } a < 0.5 \\ K'(x)\lim_{t\to 0^+}\left[\log(x/2) - \log(t)\right] & \text{if } a = 0.5 \end{cases} \\ &\qquad \to +\infty. \end{aligned}$$

Based on this divergence, we conclude that if $0 < \alpha_2 = \alpha_3 \leq 0.5$ and $x = y < 0.5$, $f_{X,Y}(x, y)$ is not defined. Note that if $x = y \geq 0.5$, divergence problems still happens, since the problems appear when u_1 converges to x. Similar calculations show that if $x + y = 1$ and $0 < \alpha_1 = \alpha_4 \leq 0.5$, the density is also not defined. More generally, $f_{X,Y}(x, y)$ is not defined if

1. $\alpha_1 + \alpha_4 \leq 1$ and $x + y = 1$.

2. $\alpha_2 + \alpha_3 \leq 1$ and $x = y$.