Prevalence estimation

Lucas Machado Moschen

School of Applied Mathematics,
Fundação Getulio Vargas

August 4, 2021

1 Introduction

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 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).

1.1 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 et al., 2018), regression (Bastos et al., 2012), communicable disease prevalence estimation (Albuquerque et al., 2009), among others.

2 Preliminary definitions

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

$$\gamma_e = \Pr(T = 0 | Y = 0). \tag{1}$$

Definition 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 $T = 1$:

$$\gamma_s = \Pr(T = 1 | Y = 1). \tag{2}$$

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

$$p = \gamma_s \theta + (1 - \gamma_e)(1 - \theta). \tag{3}$$

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.

Definition 2.3 (Link function)**.** A class of functions which maps a non-linear relationship to a linear one. Here we consider functions with domain $[0, 1]$. Examples include the logit and probit functions.

3 Prevalence model

Firstly, we make some assumptions to simplify the modeling:

Assumption 1. For a Bayesian modeling, we assume each model's parameter has a probability distribution that incorporates the researcher's uncertainty about it.

Assumption 2. For each individual, we observe k regressors that are possible risk factors represented by the vector $\boldsymbol{x}_i \in \mathbb{R}^k$ of the i^{th} individual. We assume that the probability θ_i of the i^{th} individual having been exposed to the disease dependes on the prevalence θ and $\boldsymbol{x}_i$. The probability of positive test in the i^{th} individual is denoted by p_i. Therefore, the sequences $\{Y_i\}_{i\in I}$ and $\{T_i\}_{i\in I}$ are not identically distributed anymore.

Assumption 3. Sensitivity and specificity have the same distribution for all individuals and it only depends on the test used to diagnose.

From above, we develop three different models.

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 $\boldsymbol{x}_i$.

$$
\begin{aligned}
T_i &\sim \text{Bernoulli}(\theta_i), \\
g(\theta_i) &= g(\theta) + \boldsymbol{x}_i^T \beta,
\end{aligned}
\tag{4}
$$

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 Experiments

https://github.com/lucasmoschen/rds-bayesian-analysis/blob/main/exercises/primary_model/model_experiments.ipynb

3.2 Imperfect tests

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

$$
\begin{aligned}
T_i &\sim \text{Bernoulli}(p_i) \\
p_i &= \gamma_s \theta_i + (1-\gamma_e)(1-\theta_i), \\
g(\theta_i) &= g(\theta) + \boldsymbol{x}_i^T \beta, \\
\beta &\sim \text{Normal}(\mu, \Sigma), \\
\theta &\sim \text{Beta}(a^p, b^p) \\
\gamma_s &\sim \text{Beta}(a^s, b^s), \\
\gamma_e &\sim \text{Beta}(a^e, b^e),
\end{aligned}
\tag{5}
$$

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 Imperfect test 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:

$$\begin{aligned}
T_i &\sim \text{Bernoulli}(p_i) \\
p_i &= \gamma_s \theta_i + (1-\gamma_e)(1-\theta_i), \\
g(\theta_i) &= g(\theta) + \boldsymbol{x}_i^T \beta + \omega_i, \\
\omega_i | \{\omega_j\}_{j \neq i}, \tau &\sim \text{Normal}\left(\rho \sum_j q_{ij}\omega_j, \tau^{-1}/\tilde{q}_{i+}\right) \\
\beta &\sim \text{Normal}(\mu, \Sigma), \\
\theta &\sim \text{Beta}(a^p, b^p) \\
\gamma_s &\sim \text{Beta}(a^s, b^s), \\
\gamma_e &\sim \text{Beta}(a^e, b^e), \\
\tau &\sim \text{Gamma}(a^\tau, b^\tau).
\end{aligned} \tag{6}$$

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

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

3.3.1 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. $\boldsymbol{s} = \text{tril}(QC)^T\mathbf{1} + C^T u$, 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 $\Pr(Q|w) \propto \exp[T(Q)^T w + V(Q)]$. With that, the model becomes

$$\begin{aligned}
T_i &\sim \text{Bernoulli}(p_i)\\
p_i &= \gamma_s\theta_i + (1-\gamma_e)(1-\theta_i),\\
g(\theta_i) &= g(\theta) + \boldsymbol{x}_i^T\beta + \omega_i,\\
\omega_i|\{\omega_j\}_{j\neq i},\tau &\sim \text{Normal}\left(\rho\sum_j q_{ij}\omega_j/q_{i+}, \tau^2/q_{i+}\right)\\
Q|w &\propto \exp[T(Q)^T w + V(Q)] \qquad (7)\\
\lambda &\sim \Gamma(a^\lambda, b^\lambda),\\
\beta &\sim \text{Normal}(\mu, \Sigma),\\
\theta &\sim \text{Beta}(a^p, b^p)\\
\gamma_s &\sim \text{Beta}(a^s, b^s),\\
\gamma_e &\sim \text{Beta}(a^e, b^e),\\
\tau &\sim \text{Normal}^+(0, \sigma_\tau^2).
\end{aligned}$$

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

4 Correlation analysis between specificity and sensitivity

In this section, we shall describe how to use the Bivariate Beta (Olkin and Trikalinos, 2015) to model the correlation between specificity and sensitivity.

4.1 Bivariate Beta construction

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 $\alpha_i > 0, i = 1, \dots, 4$ and $U_4 = 1 - U_1 + U_2 + U_3$. The joint density of U with respect to the Lebesgue measure is given by

$$f_U(u_1, u_2, u_3) = \frac{1}{B(\boldsymbol{\alpha})} u_1^{\alpha_1-1}u_2^{\alpha_2-1}u_3^{\alpha_3-1}(1-u_1-u_2-u_3)^{\alpha_4-1}, \qquad (8)$$

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 4.1. Let

$$X = U_1 + U_2 \text{ and } Y = U_1 + U_3. \qquad (9)$$

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

Proposition 1. *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.

$$\begin{aligned}
f_{U_1,U_2}(u_1, u_2) &= \int_{-\infty}^{\infty} f_U(u_1, u_2, u_3)\, du_3\\
&= \frac{1}{B(\boldsymbol{\alpha})}\int_0^1 u_1^{\alpha_1-1}u_2^{\alpha_2-1}u_3^{\alpha_3-1}(1-u_1-u_2-u_3)^{\alpha_4-1}\, du_3 \qquad (10)\\
&= \frac{1}{B(\boldsymbol{\alpha})} u_1^{\alpha_1-1}u_2^{\alpha_2-1}\int_0^1 u_3^{\alpha_3-1}(1-u_1-u_2-u_3)^{\alpha_4-1}\, du_3.
\end{aligned}$$

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

$$\begin{aligned} f_{U_1,U_2}(u_1,u_2) &= \frac{1}{B(\boldsymbol{\alpha})} u_1^{\alpha_1-1} u_2^{\alpha_2-1} \int_0^1 (1-u_1-u_2)^{\alpha_3-1} z^{\alpha_3-1} (1-u_1-u_2)^{\alpha_4} (1-z)^{\alpha_4-1}\, dz. \\ &= \frac{1}{B(\boldsymbol{\alpha})} u_1^{\alpha_1-1} u_2^{\alpha_2-1} (1-u_1-u_2)^{\alpha_3+\alpha_4-1} \int_0^1 z^{\alpha_3-1}(1-z)^{\alpha_4-1}\, dz. \\ &= \frac{1}{B(\boldsymbol{\alpha})} u_1^{\alpha_1-1} u_2^{\alpha_2-1} (1-u_1-u_2)^{\alpha_3+\alpha_4-1} \frac{\Gamma(\alpha_3)\Gamma(\alpha_4)}{\Gamma(\alpha_3+\alpha_4)} \\ &= \frac{1}{B(\alpha_1,\alpha_2,\alpha_3+\alpha_4)} u_1^{\alpha_1-1} u_2^{\alpha_2-1} (1-u_1-u_2)^{\alpha_3+\alpha_4-1}. \end{aligned} \tag{11}$$

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,

$$\begin{aligned} f_{U_1,X}(u_1,x) &= f(H^{-1}(u_1,x)) \left| \det \left[\left. \frac{dH^{-1}(v)}{dv} \right|_{v=(u_1,x)} \right] \right| \\ &= f(u_1, x-u_1) = \frac{1}{B(\alpha_1,\alpha_2,\alpha_3+\alpha_4)} u_1^{\alpha_1-1} (x-u_1)^{\alpha_2-1} (1-x)^{\alpha_3+\alpha_4-1}, \end{aligned} \tag{12}$$

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

$$\begin{aligned} f_X(x) &= \frac{1}{B(\alpha_1,\alpha_2,\alpha_3+\alpha_4)} \int_0^x u_1^{\alpha_1-1}(x-u_1)^{\alpha_2-1}(1-x)^{\alpha_3+\alpha_4-1}\, du_1 \\ &= \frac{1}{B(\alpha_1,\alpha_2,\alpha_3+\alpha_4)} (1-x)^{\alpha_3+\alpha_4-1} \int_0^x u_1^{\alpha_1-1}(x-u_1)^{\alpha_2-1}\, du_1. \\ &= \frac{1}{B(\alpha_1,\alpha_2,\alpha_3+\alpha_4)} (1-x)^{\alpha_3+\alpha_4-1} \int_0^x x^{\alpha_1-1}\left(\frac{u_1}{x}\right)^{\alpha_1-1} x^{\alpha_2-1}\left(1-\frac{u_1}{x}\right)^{\alpha_2-1} du_1. \end{aligned} \tag{13}$$

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

$$\begin{aligned} f_X(x) &= \frac{1}{B(\alpha_1,\alpha_2,\alpha_3+\alpha_4)} (1-x)^{\alpha_3+\alpha_4-1} x^{\alpha_1+\alpha_2-1} \int_0^1 u^{\alpha_1-1}(1-u)^{\alpha_2-1}\, du. \\ &= \frac{1}{B(\alpha_1,\alpha_2,\alpha_3+\alpha_4)} (1-x)^{\alpha_3+\alpha_4-1} x^{\alpha_1+\alpha_2-1} B(\alpha_1,\alpha_2) \\ &= \frac{1}{B(\alpha_1+\alpha_2,\alpha_3+\alpha_4)} (1-x)^{\alpha_3+\alpha_4-1} x^{\alpha_1+\alpha_2-1} \end{aligned} \tag{14}$$

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

$$f_{X,Y}(x,y) = \frac{1}{B(\boldsymbol{\alpha})} \int_\Omega u_1^{\alpha_1-1}(x-u_1)^{\alpha_2-1}(y-u_1)^{\alpha_3-1}(1-x-y+u_1)^{\alpha_4-1}\, du_1, \tag{15}$$

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,

$$\begin{aligned} f_{U_1,X,Y}(u_1, x, y) &= f_{U_1,U_2,U_3}(u_1, x - u_1, y - u_2) \\ &= \frac{1}{B(\boldsymbol{\alpha})} u_1^{\alpha_1 - 1}(x - u_1)^{\alpha_2 - 1}(y - u_1)^{\alpha_3 - 1}(1 - x - y + u_1)^{\alpha_4 - 1}, \end{aligned} \tag{16}$$

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

$$f_{X,Y}(x, y) = \frac{1}{B(\boldsymbol{\alpha})} \int_{\Omega} u_1^{\alpha_1 - 1}(x - u_1)^{\alpha_2 - 1}(y - u_1)^{\alpha_3 - 1}(1 - x - y + u_1)^{\alpha_4 - 1}\, du_1, \tag{17}$$

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

Proposition 3. *The covariance between X and Y is*

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

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

$$\mathrm{Cov}(U_i, U_j) = -\frac{\alpha_i\alpha_j}{\tilde{\alpha}^2(\tilde{\alpha} + 1)}, i, j = 1, ..., 4, i \neq j \tag{18}$$

and the variance of U_i is

$$\mathrm{Var}(U_i) = \frac{\alpha_i(\tilde{\alpha} - \alpha_i)}{\tilde{\alpha}^2(\tilde{\alpha} + 1)}, \tag{19}$$

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

$$\mathrm{Cov}(X, Y) = \mathrm{Cov}(U_1 + U_2, U_1 + U_3) = \frac{1}{\tilde{\alpha}^2(\tilde{\alpha} + 1)}(\alpha_1\alpha_4 - \alpha_2\alpha_3) \tag{20}$$

□

The main moments of X and Y are the following

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

The original paper has a mistake in page 6[1].

[1] https://www.wolframalpha.com/input/?i=simplify+%28a%28a%2B1%29+%2B+a*b+%2B+a*c+%2B+b*c%29%2F%28%28a%2Bb%2Bc%2Bd%29*%28a%2Bb%2Bc%2Bd%2B1%29%29+-+%28a+%2B+b%29*%28a%2Bc%29%2F%28a%2Bb%2Bc%2Bd%29%5E2+

4.2 Comments about integration

The density of (X, Y) with respect to the Lebesgue measure is $f_{X,Y}(x, y)$ as in equation (17). 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 $x = y$.

4.3 Specifying parameters $\boldsymbol{\alpha}$

Suppose that the researcher has knowledge about the main moments of X and Y, such that $\mathbb{E}(X) = m_1 \in (0,1), \mathbb{E}(Y) = m_2 \in (0,1), \mathrm{Var}(X) = v_1 \in (0, m_1)$, and $\mathrm{Var}(Y) = v_2 \in (0, m_2)$. Note that we will have a non-linear system with four equations and four unknown variables. After specifying these quantities, the correlation is implied by the solution. Hence, we want to solve the following

$$\begin{cases} m_1 = \dfrac{\alpha_1 + \alpha_2}{\tilde{\alpha}} \\ m_2 = \dfrac{\alpha_1 + \alpha_3}{\tilde{\alpha}} \\ v_1 = \dfrac{(\alpha_1 + \alpha_2)(\alpha_3 + \alpha_4)}{\tilde{\alpha}^2(\tilde{\alpha}+1)} = m_1 \dfrac{\alpha_3 + \alpha_4}{\tilde{\alpha}(\tilde{\alpha}+1)} \\ v_2 = \dfrac{(\alpha_1 + \alpha_3)(\alpha_2 + \alpha_4)}{\tilde{\alpha}^2(\tilde{\alpha}+1)} = m_2 \dfrac{\alpha_2 + \alpha_4}{\tilde{\alpha}(\tilde{\alpha}+1)}. \end{cases} \tag{21}$$

The first two equations of the system (21) 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:

$$\alpha_1 = \frac{m_1 + m_2 - 1}{1 - m_1}\alpha_3 + \frac{m_2}{1 - m_1}\alpha_4 \tag{22}$$

$$\alpha_2 = \frac{1 - m_2}{1 - m_1}\alpha_3 + \frac{m_1 - m_2}{1 - m_1}\alpha_4. \tag{23}$$

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 (21),

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

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 (21) as a function on of α_4. Then,

$$\Delta = \left(v_1(1 - m_1) + 2v_1\alpha_4 - m_1(1 - m_1)^2\right)^2 - 4v_1(\alpha_4 v_1(1 - m_1) - \alpha_4 m_1(1 - m_1)^2 + v_1\alpha_4^2),$$

and

$$\alpha_3 = \frac{1}{2v_1}\left(\left(m_1(1-m_1)^2 - v_1(1-m_1) - 2v_1\alpha_4\right) \pm \sqrt{\Delta}\right).$$

For now, it is hard to verify if $\alpha_3 > 0$ only when $\sqrt{\Delta}$ is summed. Therefore, we need to verify numerically. With these three expressions, from α_1, α_2, and α_3, the last equation from the system (21) indicates the solution α_4. Numerically, we reduced the dimension of the problem.

References

Albuquerque, E. M. d. et al. (2009). *Avaliação da técnica de amostragem respondent-driven sampling na estimação de prevalências de doenças transmissíveis em populações organizadas em redes complexas.* PhD thesis, ENSP.

Bastos, L. S., Pinho, A. A., Codeço, C., and Bastos, F. I. (2012). Binary regression analysis with network structure of respondent-driven sampling data.

Brook, D. (1964). On the distinction between the conditional probability and the joint probability approaches in the specification of nearest-neighbour systems. *Biometrika*, 51(3/4):481–483.

Crawford, F. W. (2016). The graphical structure of respondent-driven sampling. *Sociological Methodology*, 46(1):187–211.

Crawford, F. W., Wu, J., and Heimer, R. (2018). Hidden population size estimation from respondent-driven sampling: a network approach. *Journal of the American Statistical Association*, 113(522):755–766.

Heckathorn, D. D. (1997). Respondent-driven sampling: A new approach to the study of hidden populations. *Social Problems*, 44(2):174–199.

Lin, J. (2016). On the dirichlet distribution. *Mater's Report.*

Noordzij, M., Dekker, F. W., Zoccali, C., and Jager, K. J. (2010). Measures of disease frequency: prevalence and incidence. *Nephron Clinical Practice*, 115(1):c17–c20.

Olkin, I. and Trikalinos, T. A. (2015). Constructions for a bivariate beta distribution. *Statistics & Probability Letters*, 96:54–60.

Rothman, K. J., Greenland, S., Lash, T. L., et al. (2008). *Modern epidemiology*, volume 3. Wolters Kluwer Health/Lippincott Williams & Wilkins Philadelphia.