Prevalence estimation

Lucas Machado Moschen

School of Applied Mathematics,
Fundação Getulio Vargas

July 22, 2021

1 Introduction

A key question for epidemiologists and public health authorities is about the proportion of individuals exposed to the disease at time t. This quantity can be measured periodically, and the evolution shows how the transmission is going on. For instance, if after a year the proportion grew 50% it would be worrisome. We call it prevalence. High prevalence of a disease within a population might mean that there is a high incidence of it or prolonged survival without cure.

This report is the initial model 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

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.

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 i^{th} subject has at least one coupon before the j^{th} 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 et al. (2018), regression Bastos et al. (2012), communicable disease prevalence estimation Albuquerque et al. (2009), among others.

2 Preliminary definitions

Suppose we have a sample indexed by i. Let Y_i be the indicator function of the i^{th} individual's exposure to the disease, and T_i indicating whether the test in the i^{th} individual is positive at time t. Suppose that $\{Y_i\}$ and $\{T_i\}$ 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, $T_i = Y_i$ for every 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 pontual prevalence at time t. Being 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 parameter of the model has a probability distribution that incorporates the researcher's uncertainty about it.

Assumption 2. For each individual, we observe k features that are possible risk factors represented by the vector $\boldsymbol{x}_i \in \mathbb{R}^{k+1}$ of the i^{th} individual. The first component of $\boldsymbol{x}_i$ is 1 to handle the intercept term.

Assumption 3. Suppose that each individual has a probability θ_i of having been exposed to the disease that dependes on the prevalence θ and $\boldsymbol{x}_i$, not necessarily linearly. We therefore will have the probability of positive test p_i in the i^{th} individual. Therefore, the sequences $\{Y_i\}$ and $\{T_i\}$ are not identically distributed anymore.

Assumption 4. 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 with different stages.

3.1 Perfect tests

The first model only considers the risk factors $\boldsymbol{x}_i$.

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

where $g(\cdot)$ is a link function. Note that $\theta_i = p_i$ for all i in this case. The parameter $\beta \in \mathbb{R}^{k+1}$ is the risk effects. For Bayesian inference, priors on β and θ must be included. We use $\beta \sim \text{Normal}(\mu, \Sigma)$ and $\theta \sim \text{Beta}(\alpha^p, \beta^p)$, where $\mu \in \mathbb{R}^{k+1}$, $\Sigma \in \mathbb{R}^{(k+1)\times(k+1)}$ positive-definite matrix, $\alpha^p \in \mathbb{R}_{++}$, and $\beta^p \in \mathbb{R}_{++}$ are fixed hyperparameters.

3.1.1 Identifiability

3.2 Imperfect tests

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

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

where $\alpha^p, \alpha^s, \alpha^e, \beta^p, \beta^s, \beta^e \in \mathbb{R}_{++}$ 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. Ideally, we should build a model for it and include the uncertainty associated. Since it is not possible (for a while), we treat it as a random effect for each individual. It is modeled using Conditionally autoregressive

(CAR) in the Gaussian case. Therefore:

$$
\begin{aligned}
T_i &\sim \text{Bernoulli}(p_i) \\
p_i &= \gamma_s \theta_i + \gamma_e (1 - \theta_i), \\
\theta_i &= \theta + \nu_i + \omega_i \\
g(\nu_i) &= \boldsymbol{x}_i \beta, \\
\omega_i | \{\omega_j\}_{j \neq i}, d, \tau &\sim \text{Normal}\left(\frac{1}{d + n_i} \sum_{i \sim j} \omega_j, \frac{1}{(n_i + d)\tau} \right) \\
\beta &\sim \text{Normal}(\mu, \Sigma), \\
\theta &\sim \text{Beta}(\alpha^p, \beta^p) \\
\gamma_s &\sim \text{Beta}(\alpha^s, \beta^s), \\
\gamma_e &\sim \text{Beta}(\alpha^e, \beta^e), \\
d &\sim \\
\tau &\sim \text{Normal}^+(0, \sigma_\tau^2)
\end{aligned}
\tag{6}
$$

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.

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.

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.

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