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).

2 Preliminary definitions

Suppose we have a sample indexed by i. Let Y_i be the indicator function of the i^{th} individual exposed to the disease, and T_i indicating whether the test in the i^{th} individual is positive. 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. For that, the evaluation of the diagnostic test must be regarded, and the following definitions are important:

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 it is impossible to have 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.

3 Prevalence model

Firstly, we make some assumptions to simplify the modeling:

Assumption 1. For a Bayesian modeling, we assume each parameter of interest of the model has a probability distribution to incorporate the uncertainty about it.

Assumption 2. Suppose for each individual we observe k features that can be possible risk factors, and $\boldsymbol{x}_i$ is the vector for the i^{th} individual.

Assumption 3. Suppose each individual has a probability θ_i of having been exposed to the disease and θ_i dependes on $\boldsymbol{x}_i$, not necessarily linearly. We therefore will have p_i the probability of positive test in the i^{th} individual.

Assumption 4. We assume 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 consider the risk factors $\boldsymbol{x}_i$.

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

where $g(\cdot)$ is a link function. These class of functions maps a non-linear relationship to a linear one. Examples include the logit and probit functions. For a Bayesian inference, priors on β must be included.

References

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