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 X_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 $\{X_i\}$ 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, $T_i = X_i$ 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 $X = 0$, the *specificity* γ_e is the probability of $T = 0$:

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

Definition 2.2 (Sensitivity)**.** Probability of a positive test correctly identified. In mathematical terms, conditioned on $X = 1$, the *sensitivity* γ_s is the probability of $T = 1$:

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

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:

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

□

3 Prevalence model

Assumption 1.

References

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