A Model with Representativeness and a Binary Covariate

Rachael Caelie (Rocky) Aikens

6/12/2020

A Model in which Diagnosis Patterns Differ by Baseline Covariates

We'll start with the same model as before, but now we suppose that the population we're studying is heterogeneous in terms of some baseline characteristic, X_i. Here, X_i is binary and about half the population has $X_i = 1$, half has $X_i = 0$. (This could be sex, naturally, but we might consider some other discrete group like race, some comorbidity, sexuality, etc.)

$$\begin{aligned} X_i &\sim_{iid} \text{Bernoulli}(0.5) \\ Z_i &\sim_{iid} \text{Bernoulli}(0.15) \\ R_i | Z_i = 1 &\sim_{iid} \text{Uniform}(0, 1) \end{aligned}$$

Where $R_i = 0$ whenever $Z_i = 0$.

Now let's consider that a person's probability of diagnosis depends on not only the representativeness of their disease but their baseline characteristics:

$$D_i \sim_{iid} \text{Bernoulli}\left(\phi(R_i, X_i)\right)$$

Without loss of generality, let's assume that people with $X_i = 0$ are more likely to be correctly diagnosed than people with $X_i = 1$. There are a couple reasons this could be

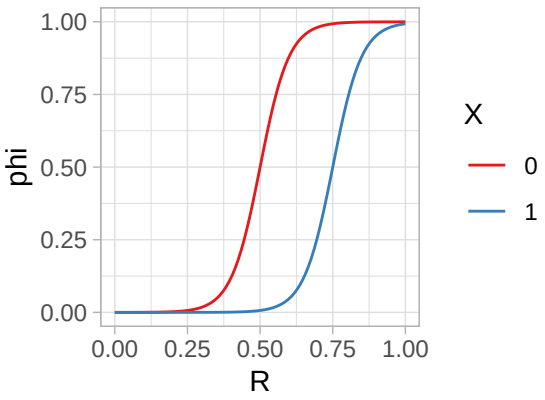
- 1. People with $X_i = 1$ have a different presentation with the disease than people with $X_i = 0$.
- 2. Doctors have a preconception that the disease is less common in people with $X_i = 1$.
- 3. People with $X_i = 1$ are less likely to see a doctor when they experience symptoms of this disease.

Explicitly, let's suppose:

$$\phi(R_i, X_i) = \frac{1}{1 + exp(-(\beta_0 + \beta_1 R_i + \beta_2 X_i))} - c_{x_i},$$

where again, $c_{x_i} = \frac{1}{1+exp(-(\beta_0+\beta_2 X_i))}$ is a corrective constant to ensure that $\phi(0, X_i) = 0$. Let Let $\beta_0 = -10$, $\beta_1 = 20$, and $\beta_2 = -5$.

Inessence, this means that people are diagnosed according to a sigmoid function, with people with $X_i = 1$ needing a higher representativeness to be diagnosed with the same probability as people with $X_i = 0$, shown below:


1.00
0.75
0.50
0.25
0.00
phi
0.00
0.25
0.50
0.75
1.00
R
X
0
1

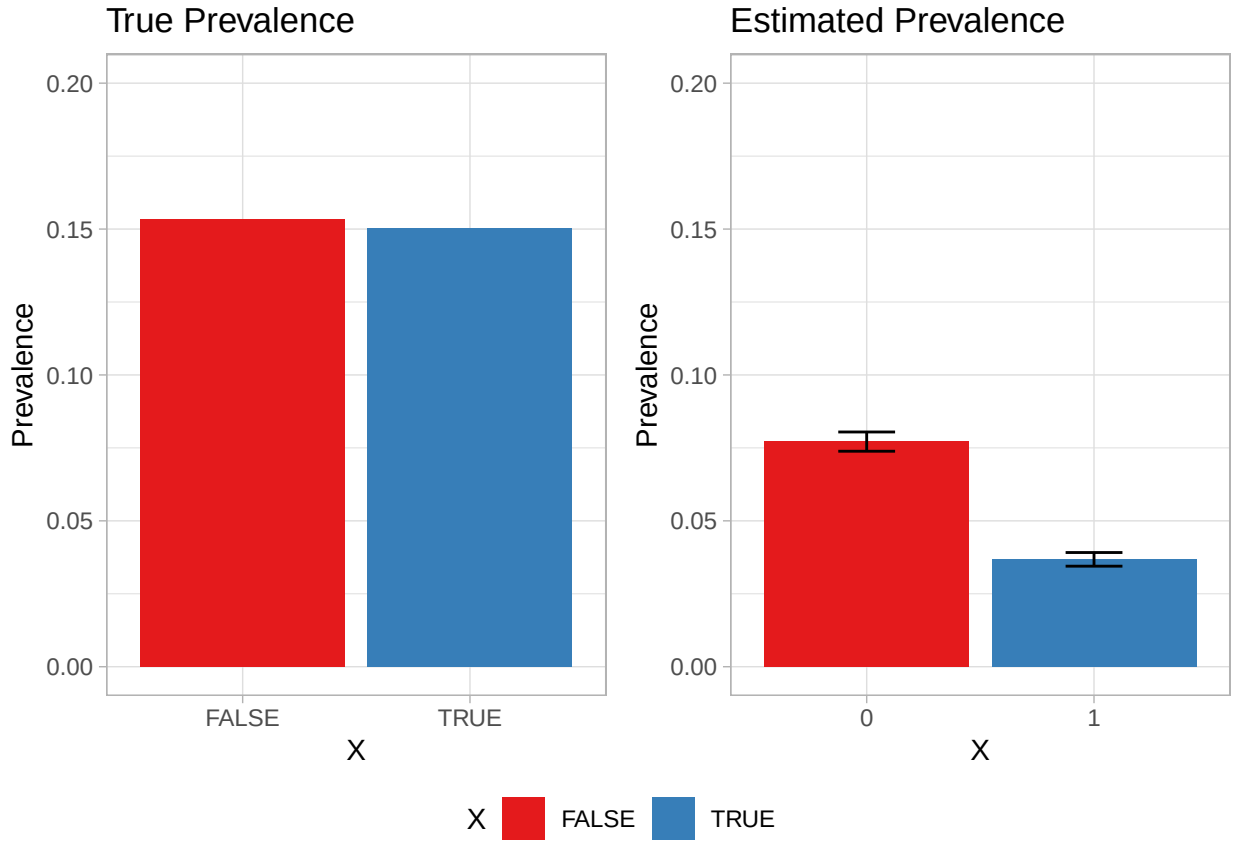
A Naive Study

Again, we collect a dataset of 5×10^4, this time recording the baseline characteristic $\{R_i, D_i, X_i\}_{i=1}^n$. Our naive researcher wants to understand:

1. What is the prevalence of the disease in each X group?
2. What is the distribution of representativeness of this disease in each X group?

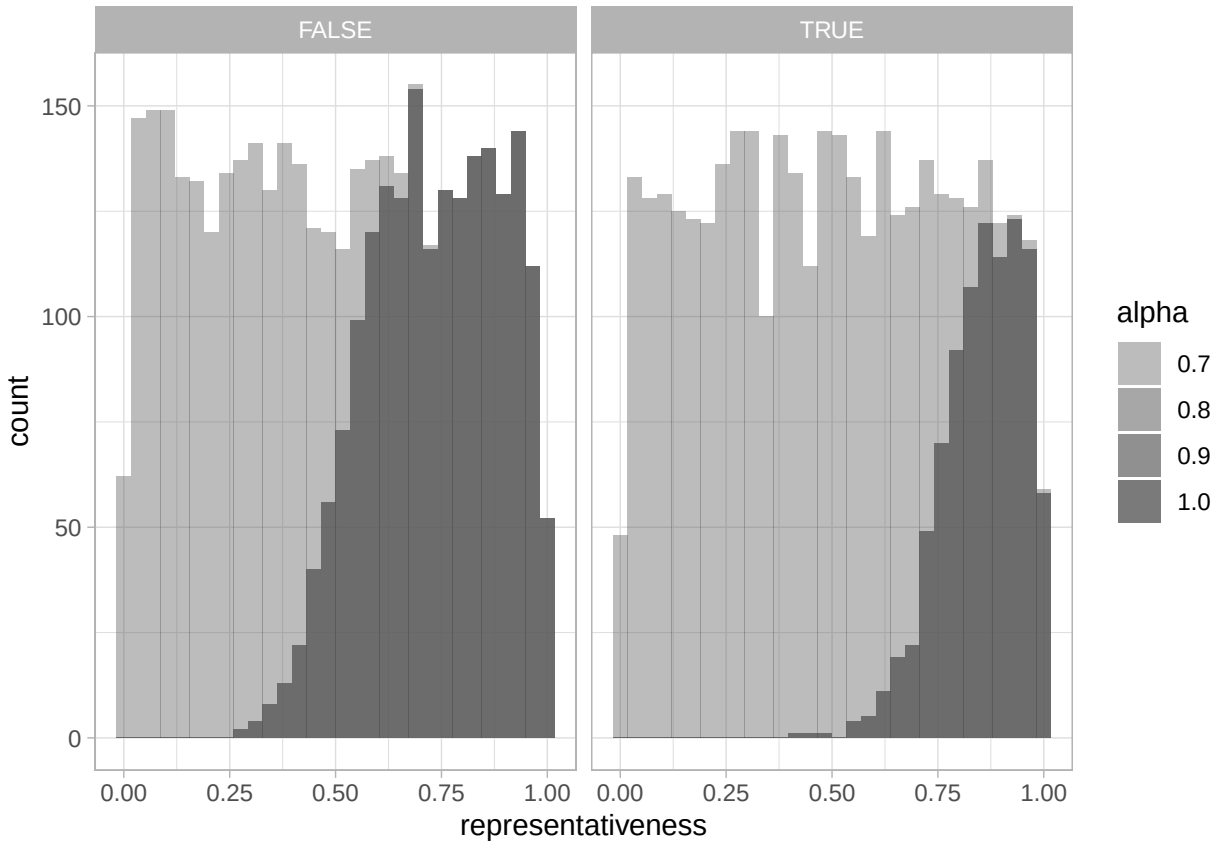
Prevalence

The plot below shows the true prevalence among groups (left) and the estimated prevalences with naive confidence intervals. The difference appears to be very statistically significant. However, the difference between the groups is not actually the fraction of people with the disease, but the fraction (and distribution) of people being diagnosed.



Representativeness

The plots below show the representativeness distributions of the diagnosed people in each group, with the hidden undiagnosed people shaded in grey. The naive researcher might conclude that this disease appears about half as ofen in people with $X_i = 1$, but when it does appear in this group, the disease is especially severe.



Overdiagnosis

The flipside of underdiagnosis is overdiagnosis. In this particular simulation set-up, the diagnostic function was so specific that nobody was incorrectly diagnosed with the disease who did not have it. In particular, the diagnostic function ϕ is normalized so that $\phi(0, X_i) = 0$ for any X_i.