Mathematics on Diagnostic process

Rachael Caelie (Rocky) Aikens

7/13/2021

Intro

This document introduces a mathematical model for how diagnosis might work in a medical system, and then proves some properties of the data that might be generated by that system. In particular, we will consider a model for how doctors might incorporate demographic information about their patient in their decision to engage in diagnostic follow-up (e.g. ordering diagnostic tests, referral to a specialist.). We'll show that the data that is generated using this "Bayesian" diagnostic approach can lead to incorrect beliefs about disease risk when classic but ubiquitous analytical mistakes are made.

Set Up

We'll start with a simple set-up. We'll begin with an interest in the diagnosis of a certain disease. We'll suppose that for this particular disease, there is a reliable method of diagnostic followup which can determine whether the patient has the disease (i.e. clear diagnostic criteria that can be assessed with a physical exam, a diagnostic test, or referal to a specialist.). Some clinician (say, a primary care provider) acts as a gatekeeper to decide whether this diagnostic follow-up should be done. A patient who is selected for follow-up and found to have the disease is said to be "diagnosed" with the disease. All other patients are said to be "undiagnosed." [1]

We'll define two groups of patients, which we'll call group 0 and group 1. These groups may be defined based on demographics (age group, sex, race, etc.) or based on some other pre-existing risk-factor (e.g. smoking status). Let r_0 and r_1 be the rates of disease in these groups, respectively, and let p_0 and p_1 be the probability that an individual in group 0 or 1 will be selected for diagnostic followup.

We'll also introduce the ratio, L, defined by:

$$p_1 = Lp_0$$

.

Inessence, this is the ratio of the rates at which individuals in each group are selected for follow-up. For example, if $L = 1$, the rates of follow-up in each group are at parity (background characteristic has no bearing on follow-up). $L < 1$ means that group 1 has a lower rate of follow-up than group 0. $L > 1$ means group 0 has the lower rate of follow-up.

The naive analyst

We'll also define the **naive analyst**. The naive analyst will always assume that individuals who recieved no diagnostic followup did not have the disease. Another way of saying this is that they assume all diagnosed

[1] Note that this is a bit of a simplification, since sometimes doctors may believe that the underlying disease state is so obvious that follow-up is not necessary. I don't expect that incorporating this nuance into the model would change the results much...

people have the disease and all undiagnosed people do not. This is a classic error in epidemiology: it is a failure to correct for misclassification. Unfortunately, it is also a ubiquitous mistake in biomedical science. There is a hidden selective process determining who is diagnosed and who is not, and the naive analyst fails to account for it.

Case 1: Underlying risk is equivalent

Suppose that underlying rates of disease in the two groups is actually the same: $r_0 = r_1 = r$. We'll imagine that an analyst is in the loop trying to estimate the underlying disease risk in the each group, and we'll denote those estimates $\hat{r}_0$ and $\hat{r}_1$, respectively.

Proposition 1.2: Suppose the rates of disease in the two groups are at parity ($r_0 = r_1 = r$), and a naive analyst tries to estimate the conditional disease risks, $\hat{r}_0$ and $\hat{r}_1$. Then,

$$E[\hat{r}_0] = p_0 r$$

,

$$E[\hat{r}_1] = p_1 r = L p_0 r$$

Thus

$$E[\hat{r}_1] = L E[\hat{r}_0]$$

.

Proof: The proof is somewhat immediate. $E[\hat{r}_0] = P(\text{group 0 individual diagnosed}) = p_0 r$. Likewise