Estimating Prevalence

Rachael Caelie (Rocky) Aikens

6/10/2020

Estimating Prevalence

Suppose we observe a population of N individuals indexed by $i = 1...N$ and we want to estimate the prevalence of some disease. Let $D_i = 1$ if the person has the disease. The true prevalence in the sample is

$$\theta = \frac{1}{N} \sum_{i=1}^{N} D_i$$

A Perfect Diagnostic Test

Suppose for the moment that there is a perfect diagnostic test for this disease. Not everybody gets tested, but we can estimate the probability that someone will get tested. Let the probability that a person gets tested be π_i, and let $D_i^* = 1$ if the test comes back positive. A good estimator for the disease prevalence might be a weighted test-positive rate:

$$\hat{\theta} = \frac{1}{\sum_{i=1}^{N} \frac{1}{\pi_i}} \sum_{T_i=1} \frac{D_i^*}{\pi_i}$$

This is a pretty simple application of sampling weights. If the diagnostic test is perfect, then for everyone with $T_i = 1$ we know $D_i^* = 1$ if and only if $D_i = 1$. Basic sampling theorems show that this estimate is unbiased, provided everyone has a nonzero probability of being tested ($\pi_i > 0$ for all i).

An imperfect Diagnostic test.

Suppose the diagnostic test is not perfect anymore. Now our $\hat{\theta}$ with sampling weights will be biased because D^* does not necessarilty equal D_i for all tested individuals. But suppose we know the sensitivity and specificity of the test. We can correct our previous estimate,

$$\hat{\theta}$$

based on the sensitivity and specificity of the test:

$$\tilde{\theta} = \frac{\hat{\theta} - (1 - sp)}{se + sp - 1},$$

Assuming $se + sp > 1$. Now we have an unbiased estimate again, $\tilde{\theta}$.

What if sensitivity and specificity are unknown?

That's all well and good, but that estimation procedure describes a scenario that doesn't often happen in medicine. In reality, most tests - real ones like lab tests and metaphorical "tests" like calling a specialist or just going to the doctor in the first place - have unknown sensitivity and specificity. Often this is because there isn't a well agreed upon "ground truth" against which to measure sensitivity and specificity in the first place.

For example, sending someone to a specialist might be thought of as a diagnostic "test" to help improve their probability of a correct diagnosis for lupus, but we have no idea how much it helps, because there isn't an agreed upon definition for lupus in the first place. Without knowing the sensitivity and specificity of the diagnostic "test" of seeing a specialist, just understanding the sampling probabilities π_i for *who* gets sent to a specialist isn't sufficient to get an unbiased estimate of prevalence.

What a naive analyst might do

One mistake an analyst might make in this scenario is to use the total rate of positive tests:

$$\hat{\theta}^{BAD} = \frac{1}{N} \sum_{T_i=1}^{N} D_i^*$$

This is a biased estimate of prevalence because it doesn't take into account that not everyone was tested. Even if the test were perfect (perfect sensitivity and specificity), this would still be a bad estimate because it doesn't correct for the differing rates at which people are tested. A more cautious analyst might invoke sampling weights to get $\hat{\theta}$, but if the sensitivity and specificity of the test can't be corrected for, this estimate can still be very biased in either direction.