11 Imperfect diagnosis: sensitivity and specificity

The goal of a diagnostic is to answer a question. Does my patient have COVID-19? Does this subject have antibodies to Epstein-Barr virus? Am I pregnant? This week we'll explore how we think about imperfect tests—tests that sometimes answer "yes" when the real answer is "no," and vice versa—and what kinds of mathematics we need to consider for our models to adjust for those imperfections.

11.1 Sensitivity and specificity

A mathematician or computer scientist might consider a diagnostic test as a binary classifier. A binary classifier takes information of some kind and produces a classification: 0 or 1. For our purposes, we'll think about 1 as positive (disease), and 0 as negative (no disease).

Sometimes the classifier gets the answer wrong. One possibility is that the true state is 1, but the classifier says 0. The other possibility is that the true state is 0 and the classifier says 1. Because there are two possible ways that the classifier could be wrong, we're going to generally avoid talking about accuracy, and instead introduce two works that talk about those two types of errors in particular.

When the classifier says 0, but the true state is 1, we call this a **false negative**. When the classifier says 1, and the true state is 1, then the classifier was able to sense the signal correctly, and we call this a **true positive.** If we imagine that the classifier is used over and over on situations where the true state is 1, then we either get a true positive or a false negative, and this means that

$$\Pr(\text{true positive}) = 1 - \Pr(\text{false negative})$$

And, because we associate the true positives with the ability of the test to sense the thing we're hoping to detect, we call the probability of true positives the **sensitivity**, se,

$$\begin{aligned} \text{se} &\equiv \Pr(\text{test is positive } \mid \text{subject is positive}) \\ &= \Pr(\text{true positive}) \\ &= 1 - \Pr(\text{false negative}) \ . \end{aligned} \tag{1}$$

When the classifier says 1, but the true state is 0, we call this a **false positive**. When the classifier says 0, and the true state is 0, then the classifier was able to correctly get that negative right, and we call this a **true negative.** If we imagine that the classifier is used over and over on situations where the true state is 0, then we either get a true negative or a false positive, and this means that

$$\Pr(\text{true negative}) = 1 - \Pr(\text{false positive})$$

And, because we associate the true negatives with the ability of the test to only detec the specific thing we're hoping to detect, we call the probability of true positives the **specificity**, sp,

$$\begin{aligned}\text{sp} &\equiv \Pr(\text{test is negative } \mid \text{subject is negative}) \\ &= \Pr(\text{true negative}) \\ &= 1 - \Pr(\text{false positive}) \ . \end{aligned} \tag{2}$$

What does this mean for data? Well, suppose that we have a population in which the prevalence is θ. In other words, if we were to choose someone randomly from the population, Pr(positive) $= \theta$. So what's the probability that our test comes back positive? Either (i) the person was positive and we got a true positive, or (ii) the person was negative and we got a false positive.[1] Therefore,

$$\Pr(\text{test positive}) = \theta\,\text{se} + (1-\theta)(1-\text{sp}) \tag{3}$$

What's the probability that the test comes back negative? You can write this down in one of two ways. Either

$$\Pr(\text{test negative}) = 1 - \Pr(\text{test positive}) = 1 - \theta\,\text{se} - (1-\theta)(1-\text{sp}) \tag{4}$$

or you could write out the two ways that one tests negative: either one is negative and the test is a true negative, or one is positive and the test is a false negative

$$\Pr(\text{test negative}) = (1-\theta)\text{sp} + \theta\,(1-\text{se}) \ . \tag{5}$$

Using algebra, you can easily show that Eq. (4) and (5) are equivalent.

An example can help us see how imperfect sensitivity and specificity can affect data. Suppose that we test 100 people and $\theta = 0.2$. If our test were perfect, the number of positives X should be, on average, 20, and drawn at random,

$$X \sim \text{Binomial}(100, 0.2) \ .$$

But what if our test has se $= 0.99$ and sp $= 0.95$? A 99% sensitivity and 92% specificity sound pretty good. However,

$$\Pr(\text{test positive}) = (0.2)(0.99) + (1-0.2)(1-0.92) = 0.262$$

and thus, in our actual data,

$$X \sim \text{Binomial}(100, 0.262) \ .$$

This means that, on average, were we to repeat our sampling many times, instead of estimating θ to be 0.2, we'd get 0.262, an error of $+31\%$.

[1] Recall the heuristic for independent probabilities: "or" is addition, while "and" is multiplication.

11.2 Positive predictive value (PPV)

What value does a test provide us when it says someone is positive? In other words, what is the probability that someone is positive, given that they test positive? Here, we can use Bayes rule,[2]

$$\begin{aligned}\Pr(\text{truly positive} \mid \text{tested positive}) &= \frac{\Pr(\text{tested positive} \mid \text{truly positive}) \times \Pr(\text{truly positive})}{\Pr(\text{tested positive})} \\ &= \frac{\text{se}\ \theta}{\text{se}\ \theta + (1-\text{sp})(1-\theta)} \\ &= \frac{\text{TP}}{\text{TP}+\text{FP}}\ , \end{aligned} \tag{6}$$

where TP and FP are true positives and false positives, respectively. This last line affords us a nice interpretation of PPV: the probability that one is *actually* positive is the proportion of all positives (TP+FP) that are truly positive (TP).

Example: Suppose that we are testing Ralphie for a rare buffalo disease. It affects 1% of buffalo. Our test is 90% sensitive and 99% specific. Ralphie tests positive. What is the probability that Ralphie has the rare buffalo disease?

The answer is 0.476.[3] In other words, even though the test seems pretty good, a positive result means that we go from 1% probability (the **pre-test probability**), to just 47.6% (the **post-test probability**). On the one hand, this means that the overwhelming majority of test-positive buffalo are actually still negative. On the other hand, the probability increased over eightfold due the test.

11.3 Negative predictive value (NPV)

What value does a test provide us when it says someone is negative? Here, just as with PPV, we can use Bayes rule,

$$\begin{aligned}\Pr(\text{truly negative} \mid \text{tested negative}) &= \frac{\Pr(\text{tested negative} \mid \text{truly negative}) \times \Pr(\text{truly negative})}{\Pr(\text{tested negative})} \\ &= \frac{\text{sp}(1-\theta)}{\text{sp}(1-\theta) + (1-\text{se})\ \theta} \\ &= \frac{\text{TN}}{\text{TN}+\text{FN}}\ , \end{aligned} \tag{7}$$

where TN and FN are true and false negative results, respectively. Using the same reasoning as PPV, the NPV is the proportion of negative results that are true negatives.

[2] $P(A \mid B) = \frac{P(B|A)P(A)}{P(B)}$

[3] Can you confirm this calculation using Eq. (6)?

Example (continued): What about our test of Ralphie? What if she tested *negative* for the disease in the previous example, using the same test? The pre-test probability that she was negative was 99%, while the post-test probability is now 99.898%. Put differently, her probability of having the disease went from 1% to 0.1%, decreasing by an order of magnitude.[4]

11.4 Adjusting prevalence calculations for an imperfect test

In a population with true prevalence θ, a perfect test would yield positive results with probability θ. However, an imperfect test would yield positive results with probability $\theta\,\text{se} + (1-\theta)(1-\text{sp})$. How can we estimate the true value of θ if we know that our test is imperfect, *and* we know se and sp?

Let the probability that a test comes back positive be ϕ, i.e.

$$\phi \equiv \theta\,\text{se} + (1-\theta)(1-\text{sp}) \ .$$

While it may be nice to have ϕ as a function of θ, what we'd really like is to estimate ϕ from our data, and then work out what θ is. In other words, we'll solve for θ,

$$\theta = \frac{\phi - (1-\text{sp})}{\text{se} + \text{sp} - 1} \ .$$

Given n tests, a total of n_{pos} of which came back positive, our MLE of ϕ is

$$\hat{\phi} = \frac{n_{\text{pos}}}{n} \ ,$$

and so we may plug this into our equation for θ to get

$$\hat{\theta} = \frac{\frac{n_{\text{pos}}}{n} - (1-\text{sp})}{\text{se} + \text{sp} - 1} \ .$$

As a warning, we should remember that $\frac{n_{\text{pos}}}{n}$ may take on values as low as 0 or as high as 1, simply by chance. After all, $n_{\text{pos}} \sim \text{Binom}(n,\phi)$ is a random variable. As a result, $\hat{\theta}$, which *should* always take on values between 0 and 1, may sometimes take on values that are negative or greater than one! When might this occur?

To gain some intuition, consider a scenario where the true prevalence is 1%, the false-positive rate for a test is also 1%, and the test is 100% sensitive. If we test 101 people, approximately 1 should be a true positive and approximately 1 should be a false positive. In other words, if everything goes according to *expectation*, we should get around 2 positive tests. If, by chance, we get just 1 positive test or even 0 positive tests, our estimate for θ will be negative!

[4] Once more, can you confirm these calculations using Eq. (7)?

11.5 How do we know sensitivity and specificity in the first place?

We must remember that sensitivity and specificity are, themselves, estimate from data. In other words, se and sp are not handed down to us from an oracle, but instead they must be sussed out from studies. We will briefly explore some of the ways in which we calibrate a test, learn its imperfections, and, in some cases, choose them according to our needs or goals.

One of the most common ways in which positive and negative results are determined by a test is via a cutoff c. Suppose that the test produces real values of some output, such that sample i produces output x_i. If $x_i > c$, then we call the sample positive, and if $x_i < c$, we call it negative. For example, an ELISA (enzyme-linked immunosorbent assay) is a common assay used in serological studies. It produces an optical density (OD), which the analyst must then decide as being positive or negative for the antibodies of interest. When the optical density is above the cutoff, the sample is positive, and when the optical density is below the cutoff, the sample is negative.

How do we choose the cutoff c? Here, we need **calibration samples** which we know to be positive and others which we know to be negative. These are called **positive and negative controls**. Suppose that we run a total of p positive controls through the assay and get a set of values $y_1, y_2, \ldots, y_p$ as output. Then suppose that we run a total of n negative controls through the assay and get a set of values $z_1, z_2, \ldots, z_n$.

- If there is a value c such that $y_i > c \; \forall \; i$ and $z_j > c \; \forall \; j$, the we can choose c to be any value between $\min_i y_i$ and $\max_j z_j$.[5] In other words, if the two sets of calibration samples are entirely non-overlapping, then any cutoff between the positive samples and the negative samples will provide perfect sensitivity and specificity *on the calibration samples*.
- If instead the distributions for y and z are overlapping, then no clean cutoff can be found. This occurs when the largest output from a negative control is larger than the smallest output from a positive control. In this scenario, one must choose a cutoff c depending on the sensitivity and specificity that one desires:
  - If we choose a very low cutoff, such that $c < y_i \; \forall \; i$, then the assay will *always* correctly classify the positive samples as positive, yet it may incorrectly classify some negatives as positives. Thus, this choice will give us se $= 1$ but sp < 1. (Perfect sensitivity, imperfect specificity.)
  - If we choose a very high cutoff, such that $c > z_i \; \forall \; j$, then the assay will *always* correctly classify the negative samples as negative, yet it may incorrectly classify some positives as negative. Thus this choice will give us sp $= 1$ but se < 1. (Perfect specificity, imperfect sensitivity.)
  - If we choose a cutoff in the middle, we may have imperfect sensitivity *and* imperfect specificity, yet the test may meet our needs nevertheless by balancing the goals of reducing false positives and false negatives alike.

[5] Recall that the symbol $\forall$ means "for all" and is a convenient mathematical shorthand.

If we get to choose c, when might we prioritize sensitivity over specificity, and vice versa? One way to answer this question is to consider how detrimental a false negative or a false positive would be. For instance, an engineer developing a "diagnostic test" for a mechanical problem with an aircraft engine will prefer false positives to false negatives, as a false negative could be catastrophic.

One alternative is to consider the so-called **Youden Index**, which is $J(c) = \text{se}(c) + \text{sp}(c) - 1$. This index quantifies the sum of sensitivity and specificity (minus one) for a particular choice of c, and often leads to Youden optimum of choosing the value of c that maximized $J(c)$. Those familiar with machine learning and the receiver-operator curve may note that $J(c)$ represents the vertical distance between the ROC and the diagonal.