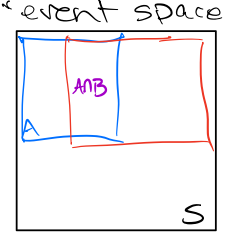


What do you know about model fitting? What techniques do we use?

Today we'll talk about a Bayesian approach to model fitting.

Conditional probability: probability of event A "given" event B (i.e. assuming event B has occurred)

$P(A|B)$



$P(A) = \frac{\text{area}(A)}{\text{area}(S)} = 1/4$   
 $P(B) = \frac{\text{area}(B)}{\text{area}(S)} = 3/8$   
 $P(A|B) = \frac{\text{area}(A \cap B)}{\text{area}(B)} = 1/3$   
 $P(B|A) = \frac{\text{area}(A \cap B)}{\text{area}(A)} = 1/2$

let's think about  $P(A \cap B)$ ...

$P(A \cap B) = \frac{\text{area}(A \cap B)}{\text{area}(S)}$ , by same logic as before.

By algebra, it's also true that

$$P(A \cap B) = \underbrace{\frac{\text{area}(A \cap B)}{\text{area}(B)}}_{P(A|B)} \cdot \underbrace{\frac{\text{area}(B)}{\text{area}(S)}}_{P(B)} = \underbrace{\frac{\text{area}(A \cap B)}{\text{area}(A)}}_{P(B|A)} \cdot \underbrace{\frac{\text{area}(A)}{\text{area}(S)}}_{P(A)}$$

therefore, it is true that

$$P(A|B)P(B) = P(B|A)P(A)$$

If we rearrange a bit...

$$\underbrace{P(A|B)}_{\text{Bayes' Rule}} = \frac{P(B|A)P(A)}{P(B)}$$

We often write as

$$\underbrace{P(H|E)}_{\text{posterior}} = \frac{\overbrace{P(E|H)P(H)}^{\text{likelihood prior}}}{\underbrace{P(E)}_{\text{evidence}}}$$

$H \rightarrow$  hypothesis

$E =$  evidence

With these labelings, we can think of Bayes' Rule as a way to "update" our *prior* belief regarding the truth of some hypothesis in light of some evidence we observe. But this all feels pretty abstract right now, so let's talk examples.

The classic one (and one that was very timely almost exactly three years ago) is regarding accuracy of tests for diseases. In this case, we have two hypotheses: a patient has a disease or does not. We also have two pieces of evidence that we can observe: a positive test result or a negative one. Thus, our "event space" has four possibilities:

	+ test	- test
sick	true positive	false negative
healthy	false positive	true negative

In medical parlance, tests are characterized by their **sensitivity** and **specificity**:

$$\text{sensitivity} = \frac{\text{true positives}}{\text{all sick}} = P(+ | \text{sick})$$

$$\text{specificity} = \frac{\text{true negatives}}{\text{all not sick}} = P(- | \text{healthy})$$

But for you, as a patient taking a test, you're probably more interested in a different question:

**If you test positive, what is the chance that you're actually sick?**

We can use Bayes' Rule to figure this out, but we'll find that there's one more crucial piece of information we'll need...

$$P(\text{sick} | +) = \frac{\overbrace{P(+ | \text{sick})}^{\text{likelihood (sensitivity)}} \overbrace{P(\text{sick})}^{\text{prior (prevalence)}}}{\underbrace{P(+)}_{\text{evidence (positivity rate)}}}$$

(what we care about)

(There are some IID assumptions buried underneath all of this, but let's not get too sidetracked...)

First, going back to our definition of conditional probabilities and considering our event space, we can see that the evidence term (positivity rate in this case) can be computed as follows:

$$P(+) = P(+ | \text{sick}) P(\text{sick}) + P(+ | \text{healthy}) P(\text{healthy})$$

And this type of construction is actually quite general: if we want to compute the probability of some event A that can be split by a set of *disjoint* events {B}, we can add up each sub-chunk as:

$$P(A) = \sum_i P(A | B_i) P(B_i)$$

Anyway, we now see that we can rewrite the answer to our question as:

$$P(\text{sick} | +) = \frac{\text{sensitivity} \cdot \text{prevalence}}{\text{sensitivity} \cdot \text{prevalence} + (1 - \text{specificity}) \cdot (1 - \text{prevalence})}$$

Let's try this out for some actual numbers. Typical values for sensitivity and specificity of a rapid COVID test as of when I just Googled it were 45.4% (yikes) and 99.8% respectively. (why might these be so different?)

Let's presume a prevalence of 5%. If you get a positive test, what are the chances you have COVID?

$$\frac{0.454 \cdot 0.05}{0.454 \cdot 0.05 + 0.002 \cdot .95} \approx 92.3\%$$

That's probably not too far off from what your intuition might have said. But let's consider a case with much better sensitivity and a much rarer disease. We'll let sensitivity and specificity both be 98% but assume the prevalence of the disease is only 0.5%...now what do we find?

$$\frac{0.98 \cdot 0.005}{0.98 \cdot 0.005 + 0.02 \cdot 0.995} \approx 19.8\%$$

That's likely a lot lower than you would have guessed for such good tests! It shows how unintuitive the effect of the prevalence can be. We can also frame this differently though, in terms of how much our belief has strengthened as a result of the test result. In the first case, 92.3/5 is about a factor of 18.5, whereas in the second, it's nearly 40!

This also emphasizes a bit the "philosophical" distinction between Bayesian and frequentist statistics – in a frequentist framework, probability necessarily refers to what fraction of the time something will happen, while in the Bayesian framework, probability represents our degree of belief in the truth of a statement. The distinction can be subtle, and in many cases we can frame something both ways, but this is the sort of thing your statistician friends might fight about after a few drinks...and does also have some bearing on Bayesian parameter estimation, which is what we're working up to here.

[Anyway, let's divert over to notebook to explore the dependence of our expression above on some of the parameters in a more systematic way...]

Okay, so this all seems cool, but what does it have to do with model fitting/parameter estimation? I'm so glad you asked! We can make some tweaks/"upgrades" to the approach we've described to adapt it quite well to this purpose!

The first one is that instead of simple true/false hypotheses like in our disease testing example, we consider a much larger number of hypotheses corresponding to all possible values/combinations of values for the parameter/parameters we wish to fit.

Second, we need to be able to incorporate multiple pieces of evidence, since in this case the evidence will correspond to measured data. This is not so hard – we can just plug in the posterior from one Bayesian update as the prior for the next step. Since multiplication is commutative, it won't even matter what order we feed in the data in!

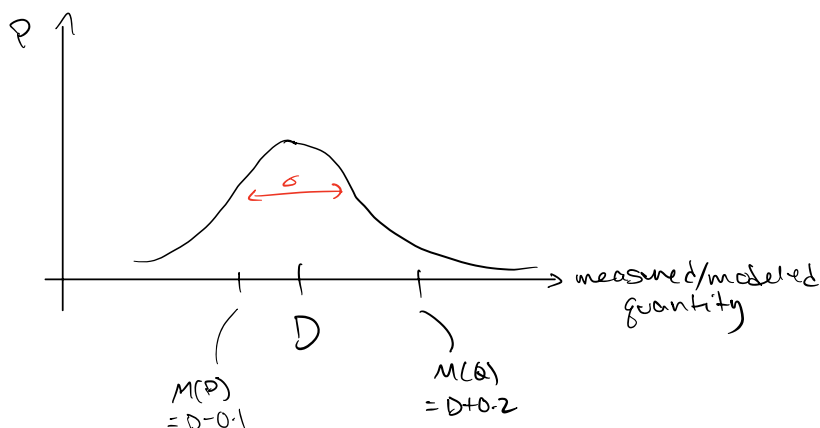
$$P(H|E_1) = \frac{P(E_1|H)P(H)}{P(E_1)} \Rightarrow P(H|E_1, E_2) = \frac{P(E_2|H)P(H|E_1)}{P(E_2)}$$

$$P(H|\{E_1, E_2, \dots, E_n\}) = \frac{P(E_n|H)P(H|\{E_1, E_2, \dots, E_{n-1}\})}{P(E_n)} \quad \leftarrow \dots$$

So we've mentioned that our hypotheses are values of parameters and our evidence is measurements. So how do we compute the likelihood and evidence terms then? Wow, you ask really good leading questions! Let's get into it!

To compute the likelihood, we need a way to compute what the chances of making the measurement we did would be, if the parameters of the model we're trying to fit took on some set of values. So we need a model, to start with. But we *also* need a notion of uncertainty. Why? Consider the case where you want to evaluate the likelihood of some parameter values  $P$  and  $Q$  given a measured value of  $D$ . We plug  $P$  and  $Q$  into our model  $M$  and get values of  $D-0.1$  and  $D+0.2$ , respectively. How should we assign likelihoods?

We probably have some intuition that  $P$  should be more likely than  $Q$ , but how much more likely? Well, if we know the scale of noise associated with our experiment, we can calculate:



The (here presumed Gaussian) experimental noise is precisely the right scale for this variance, because it is exactly what tells us how likely it is that we might have just measured some other value instead!

Okay, so last is the evidence term. We might at first think we need to do that complicated-looking sum from above, involving all the separate likelihoods and all that. And we could! But we don't have to. Because our posterior must be a valid probability distribution, we know that it has to integrate to 1. Therefore, we can treat the evidence term as simply a normalization constant and neglect explicitly computing it. Yay!

We have all the pieces now. Let's put them together first in a relatively simple problem. Imagine (contrived though it may be) that we have some data from someone launching projectiles from the surface of an alien planet (whose atmosphere provides negligible air resistance). Given measurements of their heights over time, we would like to assess the likelihood (or the strength of our *belief*) of the values of the initial launch velocity and the gravitational constant on this planet.

Our model here is then a simple analytical expression from our introductory physics class:

$$M(t; \underbrace{v_0, g}_{\text{parameters}}) = v_0 t + \frac{1}{2} g t^2$$

↙ experimental "condition"      ↘ parameters

If we make a measurement of  $y_i \pm \Delta y_i$  at time  $t_i$  and assume normal uncertainty:

$$P(y) \propto \exp\left(-\frac{(y - y_i)^2}{2 \cdot \Delta y_i^2}\right)$$

↙ fcn of  $y$       → we can ignore prefactor b/c normalization!

Since our model in this case is analytical, we can actually write out the full expression directly!

$$P(v_0, g | y(t_i) = y_i \pm \Delta y_i) \propto \exp\left(-\frac{(M(t_i; v_0, g) - y_i)^2}{2 \cdot \Delta y_i^2}\right) \cdot \text{prior}(v_0, g)$$

In our case, we'll assume a uniform prior, which means that that term effectively drops out as well since it will be the same for all velocities and g values. To the notebook!