A Continuum of Hypotheses (aka "Beast Mode")

Are you getting a little bored with hypotheses that can only take on two mutually exclusive values (like H and $\sim H$ where H is Hamilton and $\sim H$ is Madison)?

Or even tired of hypotheses that can only take on a discrete and finite number of values (like $H1$ = potato, $H2$ = rock, and $H3$ = rotten turnip)?

So often in the real world, a hypothesis is a number (like the mass of the electron) that can take on an infinity of values. Let's tackle that!

How about we call the number in the model that we are doing experiments to learn about, a.

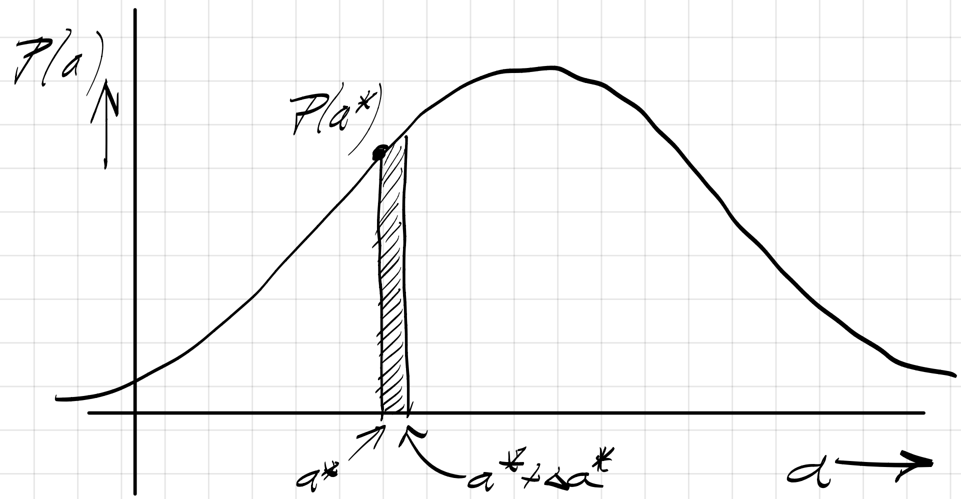
What is the probability that the value of a is a^*?

Trick question! The answer is 0 because a is a continuous variable and the probability that it has any exact value is 0.

What we really should be asking is what is the probability that a is between a^* and $a^* + \Delta a^*$. Provided Δa^* is small, this is

$P(a^*)\Delta a^*$

height width



What is going to inform us about $P(a)$? It is going to be some data (aka, some observations that we will abstractly write as just "data."

To summarize, what we should be asking about and what we wish we knew is

$$P(a^*|\text{data})\Delta a^*$$

Application of Bayes theorem

Bayes theorem comes from writing

$$P(a^* \cap \text{data})$$

two ways:

$$P(a^* \cap \text{data})\,\Delta a^* = P(a^*/\text{data})\,\Delta a^*\,P(\text{data})$$

and

$$P(a^* \cap \text{data})\,\Delta a^* = P(\text{data}/a^*)\,P(a^*)\,\Delta a^*$$

Set these equal:

$$P(a^*/\text{data})\,\Delta a^*\,P(\text{data}) = P(\text{data}/a^*)\,P(a^*)\,\Delta a^*$$

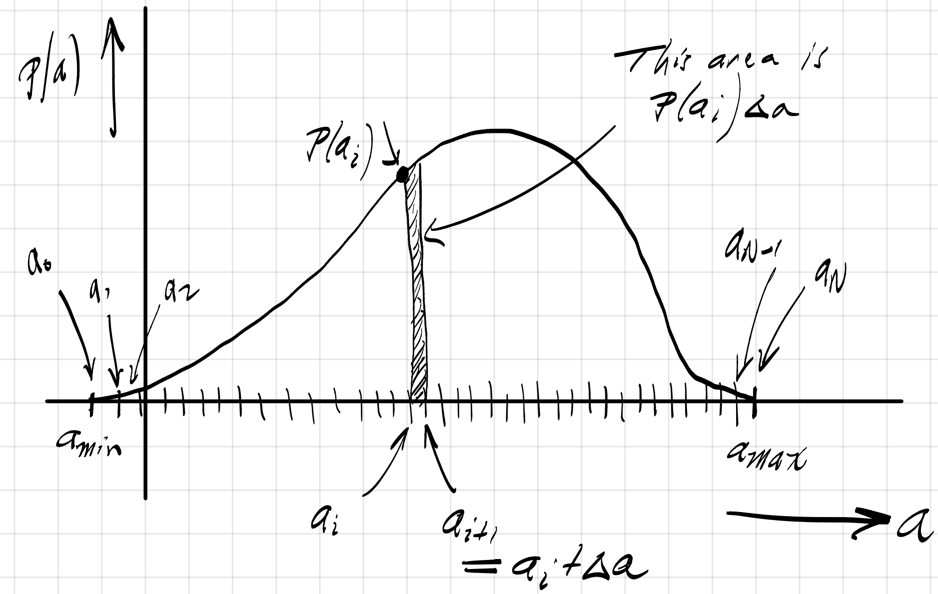
Solve for what we wish we knew:

$$P(a^*/\text{data})\,\Delta a^* = \frac{P(\text{data}/a^*)\,P(a^*)\,\Delta a^*}{P(\text{data})}$$

Now what!? The usual next step. We work on the denominator, $P(\text{data})$.

Just to make life (temporarily) a little easier, let's say that the conceivable values of a are between a_{min} and a_{max}. (Later you'll see it doesn't matter much what a_{min} and a_{max} are, and it could even be that $a_{min} = -\infty$ and $a_{max} = \infty$.)

Here is a graph that helps to visualize what we are about to do with the denominator:



We have imagined dividing up the region between a_{min} and a_{max} into N parts each of width Δa.

$$a_0 = a_{min}$$
$$a_1 = a_{min} + \Delta a$$
$$\vdots$$
$$a_i = a_{min} + i\,\Delta a$$
$$a_{i+1} = a_{min} + (i+1)\,\Delta a$$
$$\vdots$$
$$a_{N-1} = a_{min} + (N-1)\,\Delta a = a_{max} - \Delta a$$
$$a_N = a_{max}$$

$$\Delta a \equiv \frac{a_{max} - a_{min}}{N}$$

The probability that a is between a_i and $a_{i+1} = a_i + \Delta a$ is the shaded bar's area which is $P(a_i)\Delta a$ ← left-hand sum

(Ask yourself: why not $P(a_{i+1})\Delta a$ ← right-hand sum

or $\frac{P(a_i)+P(a_{i+1})}{2}\Delta a$? ← trapezoid approximation

Because it doesn't matter for small Δa!)

Now write $P(\text{data})$ as

$$\sum_{i=0}^{N-1} P(\text{data}|a_i)P(a_i)\Delta a$$

Again, we are using an approximation, but this approximation is about to become perfect as we let $N \to \infty$ and the sum will become an integral.

Now we take $N \to \infty$ while also making Δa smaller and smaller in such a way that we always maintain

$$\Delta a = \frac{a_{max} - a_{min}}{N}$$

Then the approximations get better and better, and instead of an approximation, we have the equality

$$P(\text{data}) = \lim_{N\to\infty} \sum_{i=0}^{N-1} \underbrace{P(\text{data}|a_i)P(a_i)}_{\text{Let me just call this } Q(a_i) \text{ for a moment}} \Delta a$$

$$P(\text{data}) = \lim_{N\to\infty} \sum_{i=0}^{N-1} Q(a_i)\Delta a$$

with $a_i = a_{min} + i\Delta a$ and $\Delta a = \frac{a_{max}-a_{min}}{N}$

This is the integral of Q!

$$P(\text{data}) = \int_{a_{min}}^{a_{max}} Q(a)\,da$$

$$= \int_{a_{min}}^{a_{max}} P(\text{data}|a)P(a)\,da$$

put back in that $Q(a) = P(\text{data}|a)P(a)$

Bringing it all together

Now that we have done all that work on the denominator, what have we learned? Another version of Bayes theorem, but this time for a continuum of hypothesis

$$P(a^*/\text{data})\Delta a^* = \frac{P(\text{data}/a^*)P(a^*)\Delta a^*}{P(\text{data})}$$

$$= \frac{P(\text{data}/a^*)P(a^*)\Delta a^*}{\int_{a_{min}}^{a_{max}} \underbrace{\underbrace{P(\text{data}/a)}_{\text{"the likelihood"}} \underbrace{P(a)}_{\text{"the prior"}}}_{\text{note that these show up in the numerator too}} da}$$

As usual, this won't get us far unless we have some guess for the prior.

The prior is no longer something we can tabulate, such as

$$P(H) = \frac{48}{98} \quad \text{and} \quad P(\sim H) = \frac{50}{98}.$$

Instead the prior is now a function of a: $P(a)$.

Comparison with Chapter 9 of Donovan and Mickey

Look at Chapter 9, p.122, Eq 9.31:

$$P(\theta/\text{data}) = \frac{P(\text{data}/\theta)\cdot P(\theta)}{\int P(\text{data}/\theta)\cdot P(\theta)\,d\theta}$$

Let's do three things: (1) The θ in the numerators on each side has nothing to do the θ in the integral. Let's call the θ in the numerators θ^*; (2) Let's multiply each side by some range $\Delta\theta^*$; (3) Let's acknowledge that the parameter θ might have some minimum and maximum value. Then Eq. 9.31 becomes

$$P(\theta^*/\text{data})\Delta\theta^* = \frac{P(\text{data}/\theta^*)\,P(\theta^*)\Delta\theta^*}{\int_{\theta_{min}}^{\theta_{max}} P(\text{data}/\theta)P(\theta)\,d\theta}$$

Compare with what I derived on the left. It is exactly the same except I called θ, a.