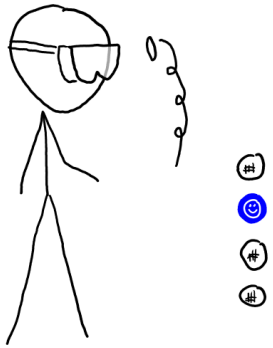
I tossed four coins and got one head.
What is it reasonable to infer about the probability of heads (call it θ)?



- "The maximum likelihood estimator is $\hat{\theta} = 25\%$, thus the true probability of heads is 25%"
  (hence if I tossed millions more coins that's the fraction of heads I'd see)

unjustified!

- "All we know for certain is that $0 < \theta < 1$"

logical, but useless!

- Let's use a random variable to express our beliefs about Θ.
  Thus, we're proposing a joint model $\Pr_{\Theta,X}$.

Θ unknown parameter

We might choose $\Theta \sim U[0,1]$ to express ignorance about Θ.

X probability model for the data

For this problem, a reasonable model is $X \sim \text{Bin}(4, \Theta)$

Bayes's rule tells us how to update our beliefs in the light of data: simply use the conditional distribution, $(\Theta|X = x)$.

$\Theta \sim U[0,1]$

$X \sim \mathrm{Bin}(4, \Theta)$

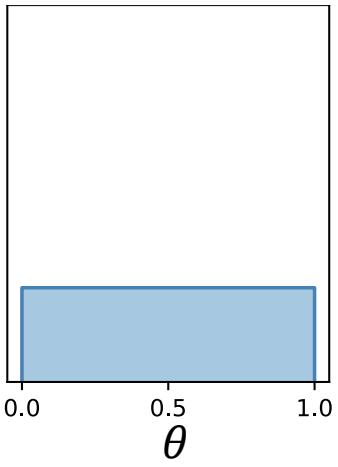
$\Pr_\Theta(\theta)$ is called the prior.

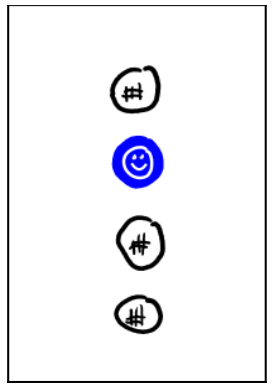
It expresses our beliefs prior to having seen the data.

$\Pr_\Theta(\theta|X = x)$ is called the posterior.

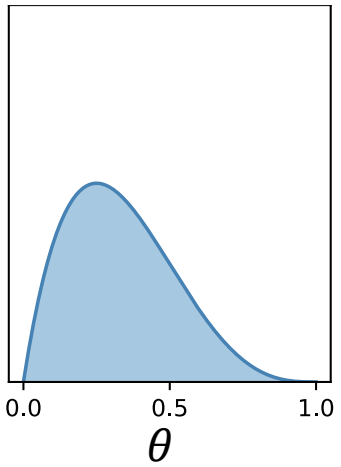
It expresses our beliefs about Θ *in the light of the data.*

$\Pr_\Theta(\theta)$



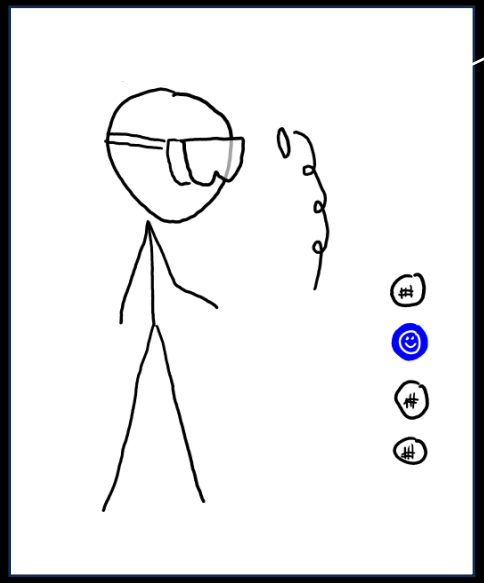


$\Pr_\Theta(\theta|X = 1)$



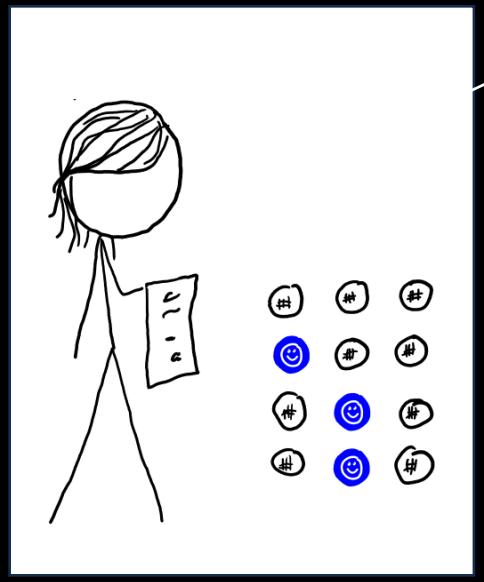
A PARADOX ABOUT THE MEANING OF PROBABILITY

I thought “probability” measured uncertainty. Surely there’s a difference in uncertainty between these two cases?!



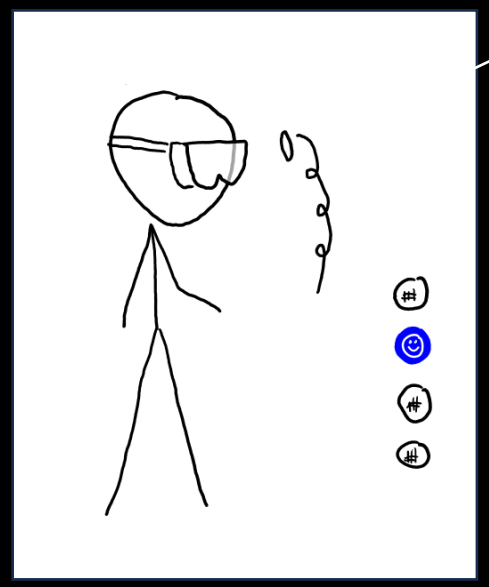
I tossed four coins
and got one head.

Using a Bin(n, θ) model, I estimate
the probability of heads is $\hat{\theta} = 25\%$



I tossed twelve coins
and got three heads.

Using a Bin(n, θ) model, I estimate
the probability of heads is $\hat{\theta} = 25\%$



I tossed four coins and got one head.

Using a Bin(n, θ) model, I estimate the probability of heads is $\hat{\theta} = 25\%$

QUESTION. If we toss 100 more coins, how many heads do you predict we'd see?

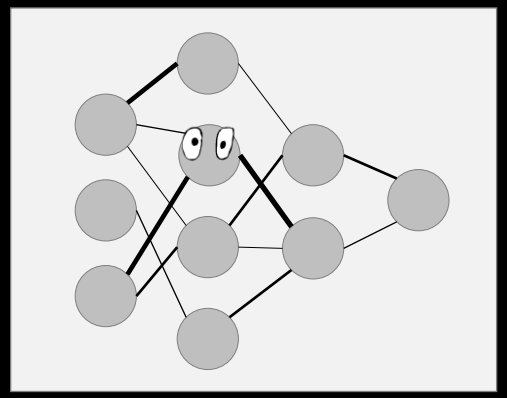
There are two types of uncertainty in our prediction:

EPISTEMIC UNCERTAINTY

- We don't know θ exactly

ALEATORIC UNCERTAINTY

- Even if we knew θ exactly, we still wouldn't know the exact number of heads we'll get



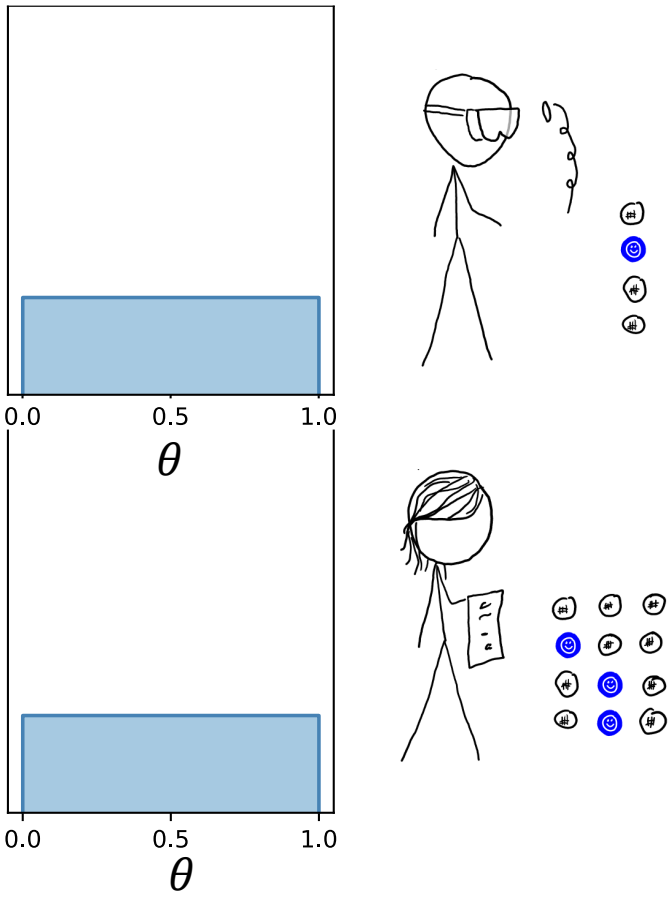
"This is a 40mph speed limit, with probability 98%."

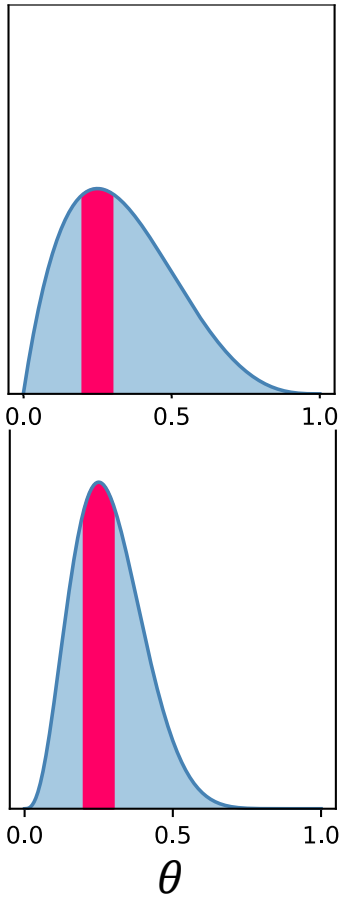
Does it mean 98%±0.01%
or 98%±63% ?

Neural networks classifiers report aleatoric *probabilities*, but they don't tell us their epistemic *confidence*.

In Bayesianism, the posterior distribution *is exactly* our epistemic belief about the parameter. It depends on the amount of evidence we've seen.

prior belief $\Pr_\Theta(\theta)$ + data x → posterior belief $\Pr_\Theta(\theta|X=x)$



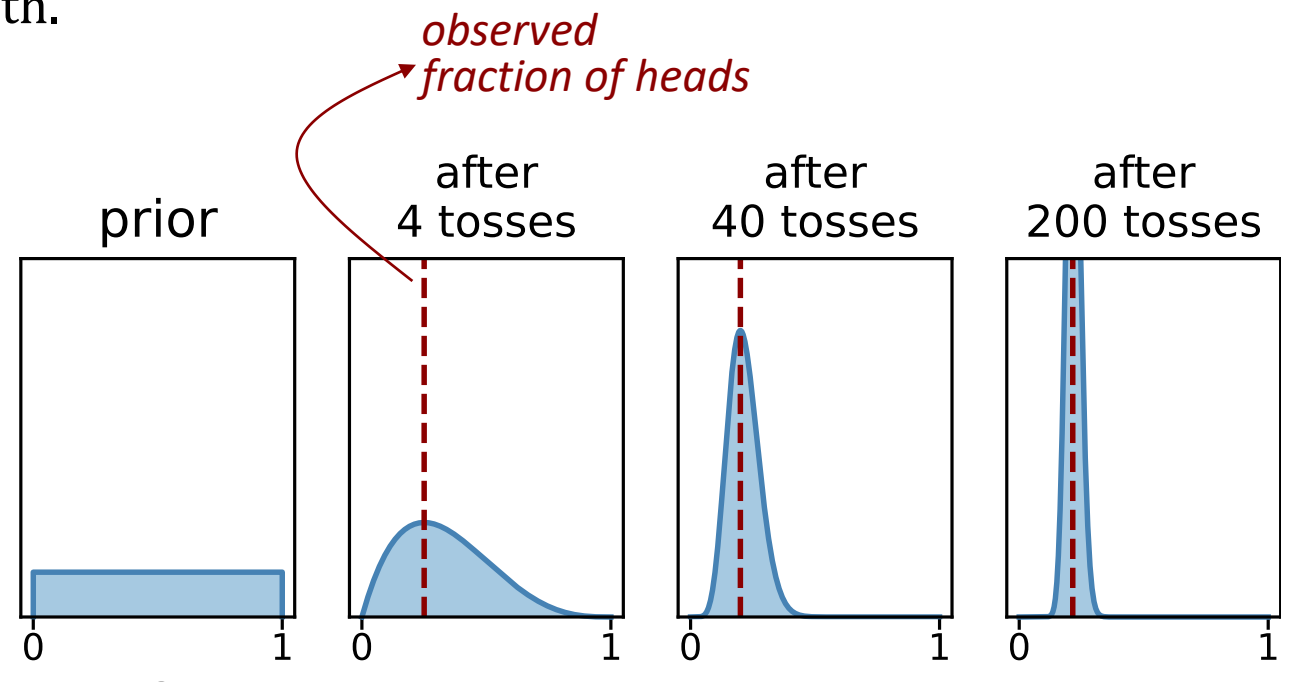


$\mathbb{P}(\Theta \in [.2, .3] \mid \text{data}) = 21\%$

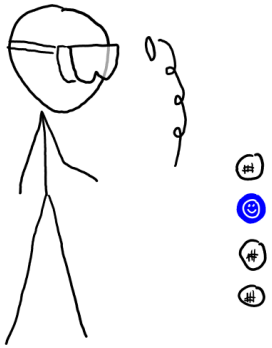
$\mathbb{P}(\Theta \in [.2, .3] \mid \text{data}) = 33\%$

A tighter posterior distribution for Θ means we are more confident about its value.

Typically, the more data we have, the closer our posterior gets to the truth.



We *must* have a prior belief about every unknown parameter. We *must* choose it before seeing the dataset in question.



$$\Theta \sim U[0,1]$$

$$\downarrow$$

$$X \sim \mathrm{Bin}(4,\Theta)$$

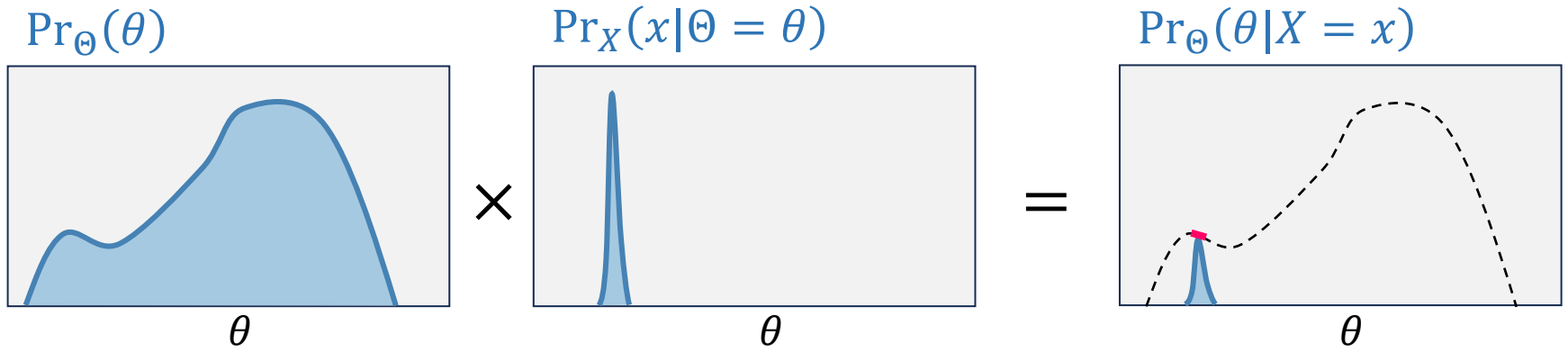
But where does the prior come from?

It comes from what you know already — it's how you can integrate your existing knowledge into your modelling.

If you don't have a prior to start off with, you have no business even thinking about the experiment!

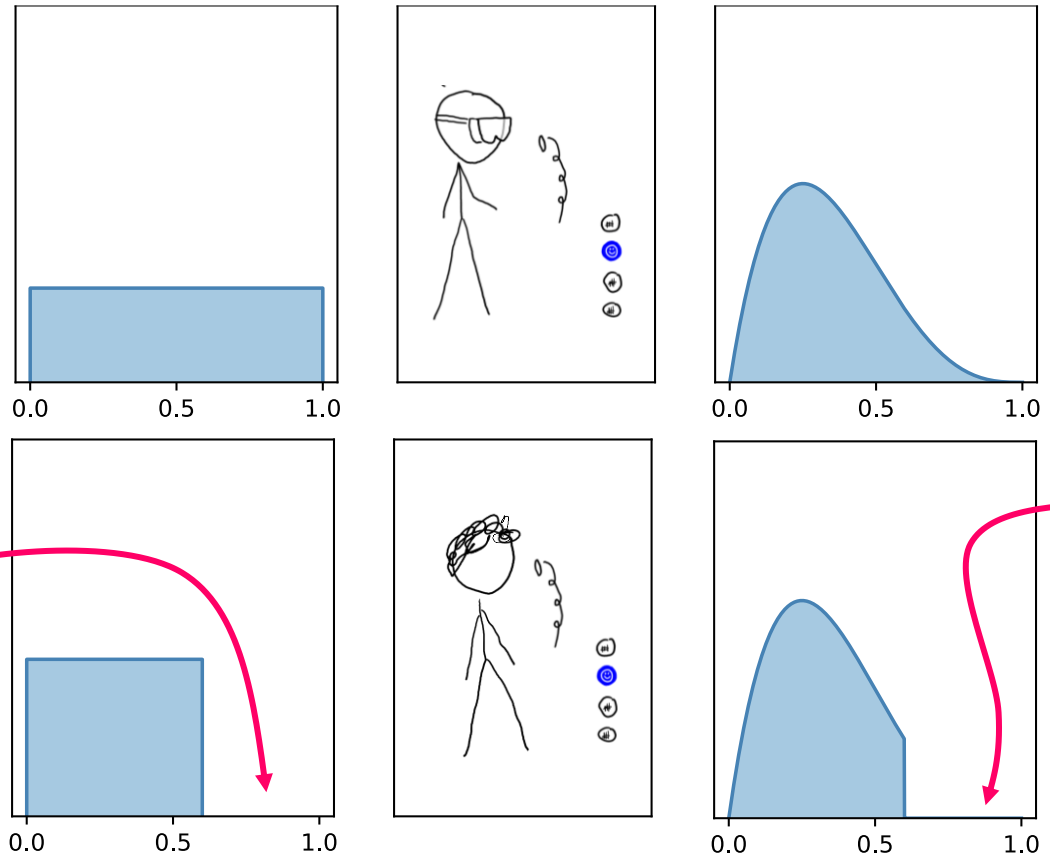
Often, with lots of data, the prior doesn't make much difference.

$$\Pr_{\Theta}(\theta|X = x) \;=\; \kappa \Pr_{\Theta}(\theta) \Pr_{X}(x|\Theta = \theta)$$



You are entitled to your own personal prior beliefs.
They are entirely your choice.

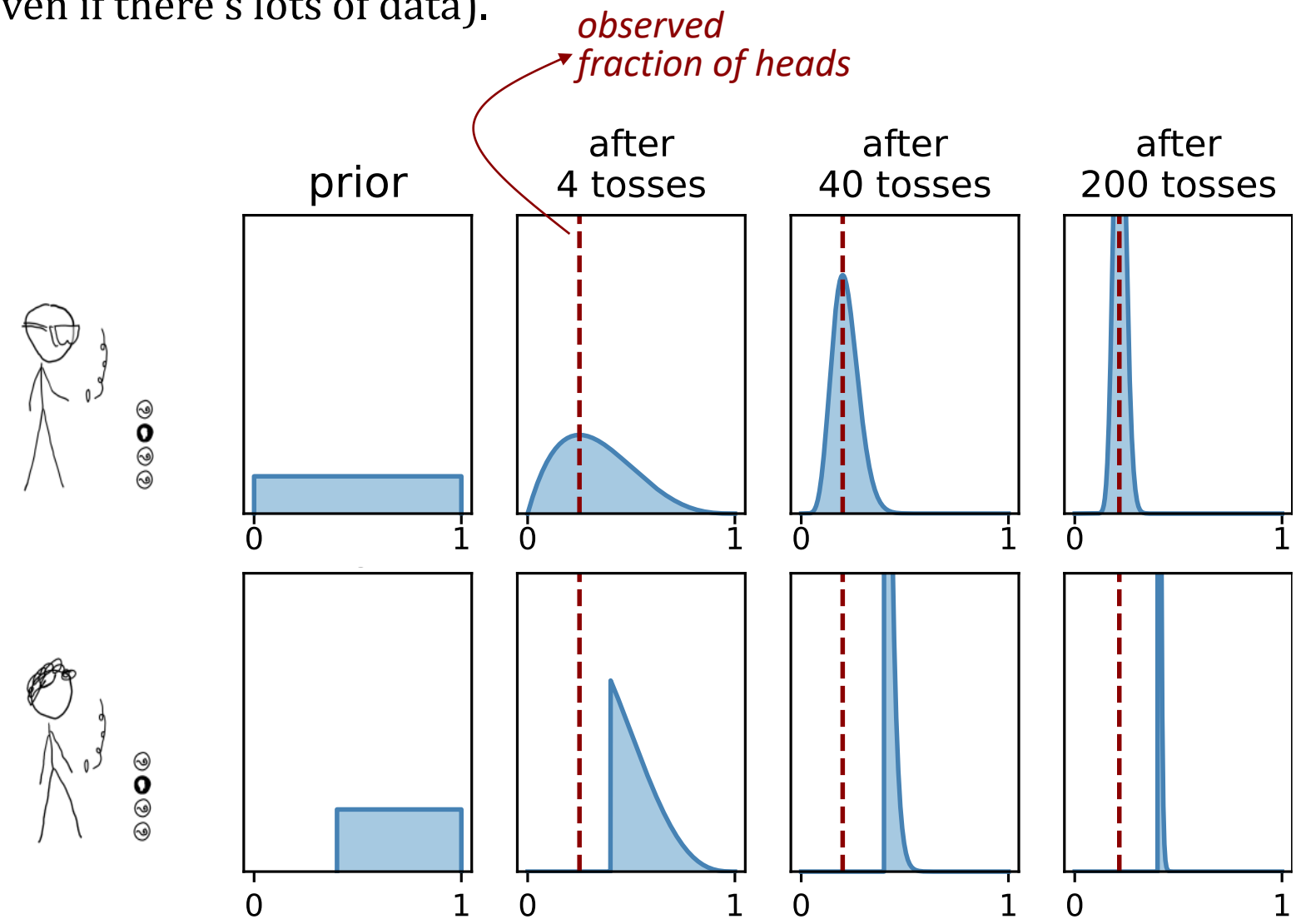
$\Pr_\Theta(\theta)$ $\Pr_\Theta(\theta|X=x) = \kappa \Pr_\Theta(\theta) \Pr_X(x|\Theta=\theta)$



Preconception that $\theta > 0.6$ is impossible

The preconception is unshakeable

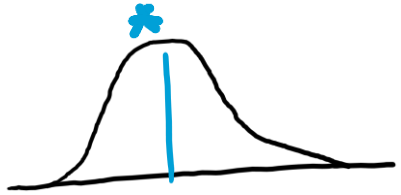
If your prior is extreme, it will be reflected in your posterior (even if there's lots of data).



How should we report the posterior distribution?

Prior distribution for Θ

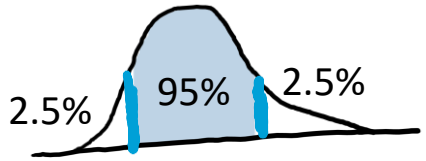
Posterior distribution for Θ



We could report the *posterior mean.*

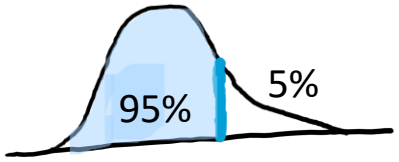
We could report the point with highest likelihood, the *MAP* or *maximum a-posteriori* estimate.

> Example (Laplace smoothing).
> We counted x successful outcomes from n trials.
> Using the model $X \sim \mathrm{Bin}(n, \Theta)$, and the prior $\Theta \sim U[0,1]$,
> the posterior mean of Θ is $(x+1)/(n+2)$.



We could report a *95% confidence interval* `[lo,hi]` such that

$$\mathbb{P}(\Theta < \mathtt{lo} \mid \text{data}) = 2.5\%$$
$$\mathbb{P}(\Theta > \mathtt{hi} \mid \text{data}) = 2.5\%$$

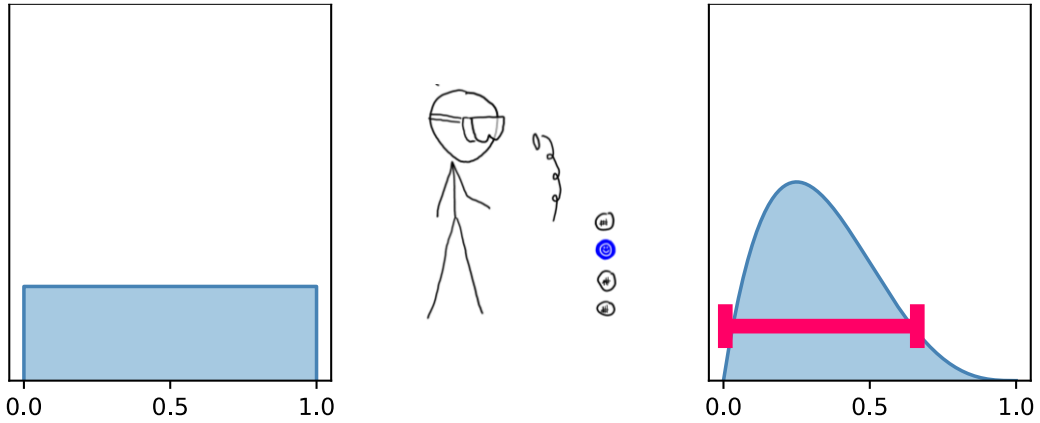


or indeed any other 95% confidence interval e.g.

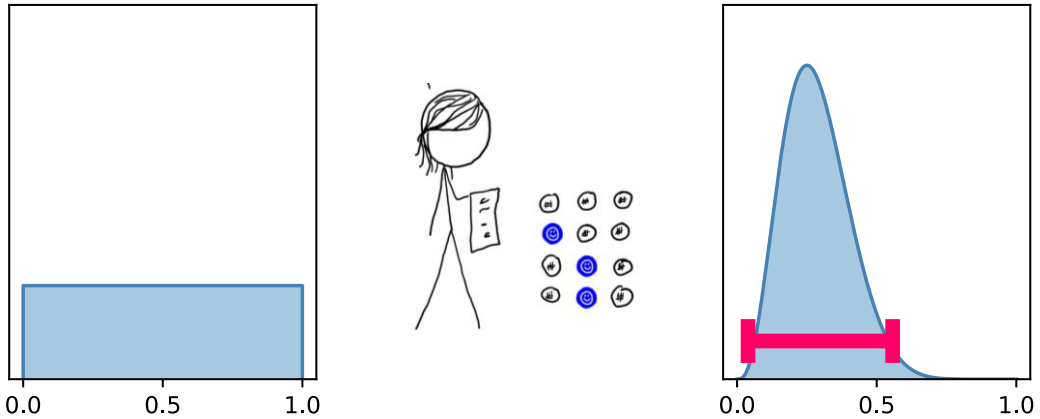
$$\mathtt{lo} = -\infty$$
$$\mathbb{P}(\Theta > \mathtt{hi} \mid \text{data}) = 5\%$$

(though these only really work well for continuous Θ,
as for discrete Θ we might not be able to hit the probabilities exactly)

prior belief $\Pr_\Theta(\theta)$ + data x → posterior belief $\Pr_\Theta(\theta|X=x)$



I estimate the probability of heads is 25%, and my 95% confidence interval is [3%, 72%]



I estimate the probability of heads is 25%, and my 95% confidence interval is [12%, 51%]

Rationalist community - Wikipedia

https://en.wikipedia.org/wiki/Rationalist_community

WIKIPEDIA
The Free Encyclopedia

Donate Create account Log in

Rationalist community

1 language

Article Talk Read Edit View history Tools

From Wikipedia, the free encyclopedia

This article is about the 21st-century movement. For the philosophical concept, see Rationalism. For other uses, see Rationalism (disambiguation).

The **rationalist community** is a 21st-century movement that formed around a group of internet blogs, primarily LessWrong and *Astral Codex Ten* (formerly known as *Slate Star Codex*). The movement initially gained prominence in the San Francisco Bay Area. Its biases. Common interests include probability, effectiv from artificial general intelligence.

The borders of the rationalist community are blurry ar groups.[1] Members who diverge from typical rationali "post-rationalist" (also known as "ingroup" and "TPOT adjacent".[3]

Description [edit]

Rationality [edit]

Rationalists define rationality to include epistemic rati instrumental rationality — acting in a way to achieve

The rationalists are concerned with applying science to Bayesian inference.[6] According to Ellen Huet, the unbiased, even when the conclusions are scary".[7]

Bayesianism is an epistemology, i.e. a theory of knowledge and evidence, that passes all these “smell tests”.

Surely it’s how any rational person should think!

To apply Bayes's rule, first write out your probability model $\Theta \longrightarrow X$ and state the distributions of Θ and of $(X;\Theta)$. Then,

ALGEBRAIC BAYES

1. Write out $\Pr_\Theta(\theta)$
2. Use the formula
   $\Pr_\Theta(\theta|X=x) = \kappa \Pr_\Theta(\theta)\Pr_X(x|\Theta=\theta)$
   then find κ to make this integrate to 1

… but these are usually intractable

This lets us calculate confidence intervals:

$$\mathbb{P}(\Theta \in \text{range}|X=x) = \int_{\theta\in\text{range}} \Pr_\Theta(\theta|X=x)\,d\theta$$

One way to do

COMPUTATIONAL BAYES

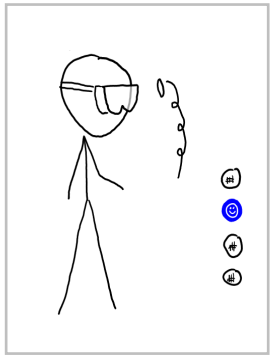
1. Generate a sample $(\theta_1, \dots, \theta_n)$ from Θ
2. Compute weights
   $w_i = \Pr_X(x|\Theta=\theta_i)$,
   then rescale weights to sum to one

$$\mathbb{P}(\Theta \in \text{range}|X=x) \approx \sum_{i=1}^{n} w_i 1_{\theta_i\in\text{range}}$$

It's more elegant to use the generalized version

$$\mathbb{E}[h(\Theta)|X=x] \approx \Sigma_i w_i h(\theta_i)$$

Example



I got $x = 1$ head out of $n = 4$ coin tosses. I propose the probability model $X \sim \text{Bin}(n, \Theta)$. I don't know Θ, so I'll treat it as a random variable, $\Theta \sim U[0,1]$.

Plot the distribution of $(\Theta|X = x)$.

Likelihood of the data: $X \sim Bin(n, \theta)$ $\quad Pr_x(x \mid \Theta=\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} = 4\theta(1-\theta)^3$

1. Generate a sample $(\theta_1, \ldots, \theta_n)$ from Θ:

```
Θ ~ U[0,1]
θsamp = np.random.uniform(0, 1, size = 100000)
```

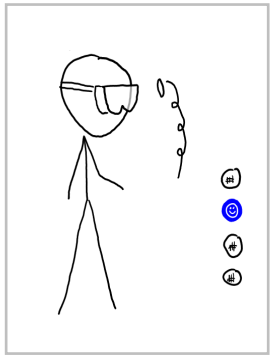
2. Compute weights $w_i = \Pr_X(x|\Theta = \theta_i)$, then rescale weights to sum to one:

```
w = 4 * (θsamp**1) * ((1-θsamp)**3)
w = w/npsum(w)
```

Reason about $(\Theta|X = x)$ indirectly, using

$$\mathbb{E}[h(\Theta)|X = x] \approx \Sigma_i w_i h(\theta_i)$$

Example



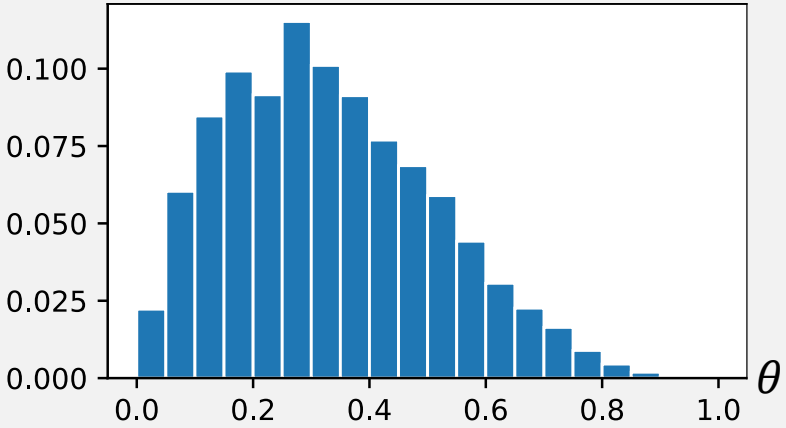
I got $x = 1$ head out of $n = 4$ coin tosses. I propose the probability model $X \sim \text{Bin}(n, \Theta)$. I don't know Θ, so I'll treat it as a random variable, $\Theta \sim U[0,1]$.

Plot the distribution of $(\Theta|X = x)$.

Reason about $(\Theta|X = x)$ indirectly, using

$$\mathbb{E}[h(\Theta)|X = x] \approx \Sigma_i w_i h(\theta_i)$$

Let's plot a histogram:
we'll split the range of θ into bins,
and for each θ-bin, we'll draw a bar
of height $\mathbb{P}(\theta \in \text{bin} \mid X = x)$



$$\mathbb{P}(\Theta \in \text{bin} \mid \text{data})$$

$$= \mathbb{E}(1_{\Theta \in \text{bin}} \mid \text{data}) = \mathbb{E}[h(\Theta) \mid \text{data}] \quad \text{where } h(\Theta) = 1_{\Theta \in \text{bin}}$$

using the indicator/probability trick

$$\approx \sum_i w_i h(\Theta_i)$$

by computational Bayes approximation

$$= \sum_i w_i 1_{\Theta_i \in \text{bin}} = \sum_{i:\Theta_i \in \text{bin}} w_i$$

unwrapping the indicator function

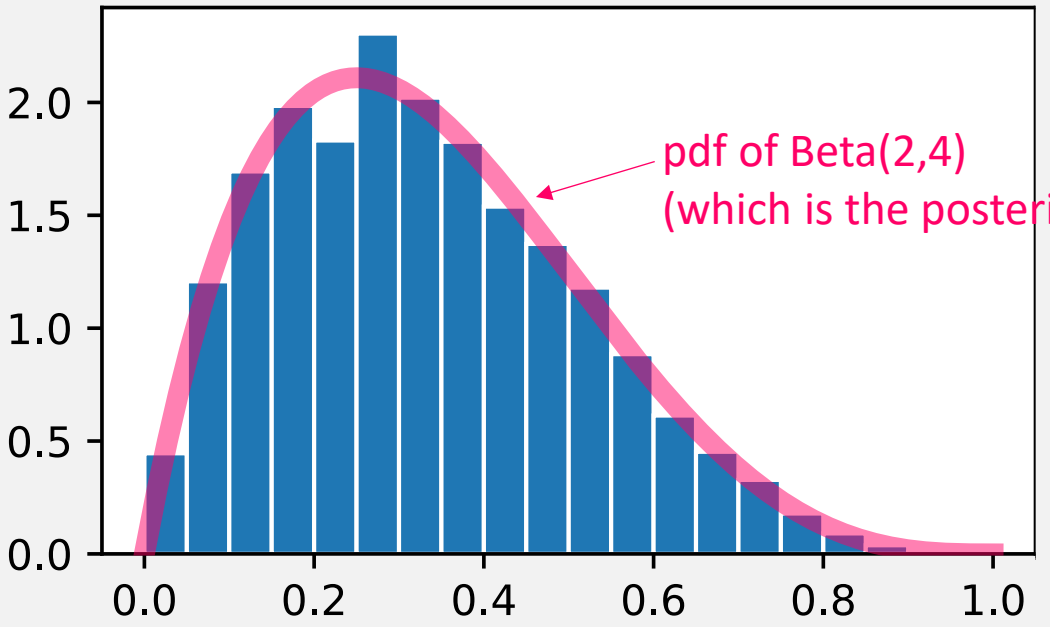
In code, to plot all bins: `plt.hist(θsamp, weights=w)`

For continuous random variables, I prefer to scale the bar heights so that the total *area* is 1. This is known as a *density histogram.*

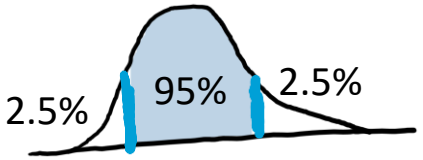
$$\text{bar height} = \frac{\mathbb{P}(\Theta \in \text{bin}|\text{data})}{\text{bin width}}$$

It means the plot is directly comparable to a pdf.

```
plt.hist(θsamp, weights=w) density=True)
```



How do we compute confidence intervals from a weighted sample?



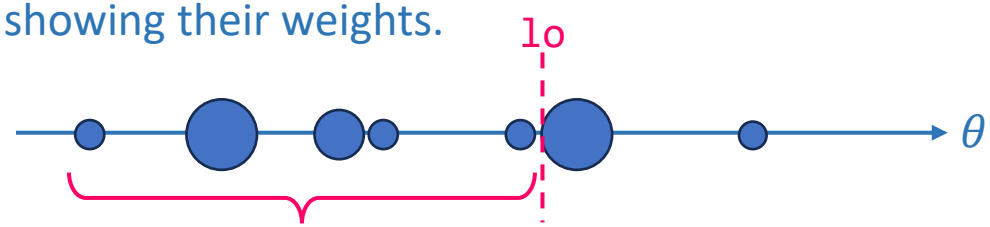
We'd like to report a *95% confidence interval* `[lo,hi]` such that

$$\mathbb{P}(\Theta < \texttt{lo} \mid \text{data}) = 2.5\%$$
$$\mathbb{P}(\Theta > \texttt{hi} \mid \text{data}) = 2.5\%$$

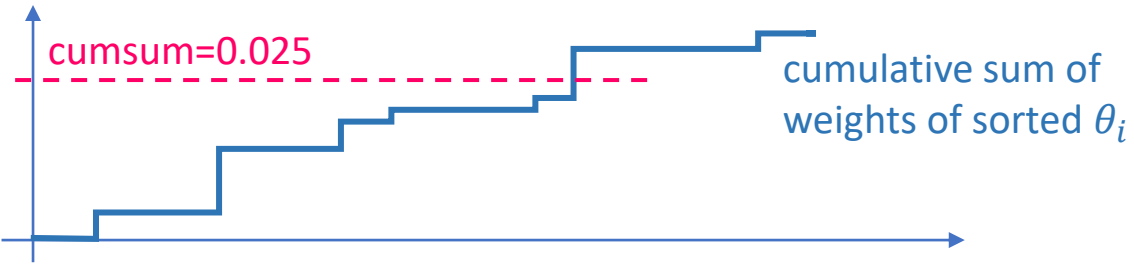
We can find `lo` via a computational Bayes estimate (and `hi` similarly):

$$\mathbb{P}(\Theta < \texttt{lo} \mid \text{data}) \approx \sum_i w_i \, 1_{\theta_i < \texttt{lo}}$$

Consider a dot-plot of the θ_i, showing their weights.



We want to choose `lo` so that the sum of weights for these θ_i is 0.025

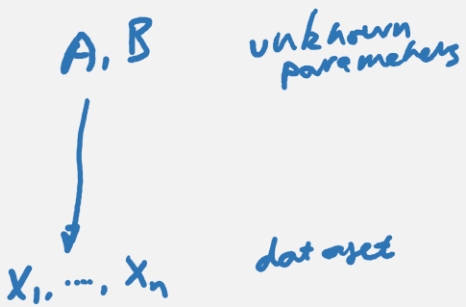



```
θsamp, w = …
i = np.argsort(θsamp)
θsamp, w = θsamp[i], w[i]
F = np.cumsum(w)
lo = θsamp[F<0.025][-1]
```

Exercise 8.3.2 (Multiple unknowns)

We have a dataset $[x_1, \dots, x_n]$. We propose to model it as independent samples from $U[A, A+B]$, where A and B are unknown parameters.

Using $A \sim \text{Exp}(0.5)$ and $B \sim \text{Exp}(1.0)$ as prior distributions for the unknown parameters, find the distribution of $(B|\text{data})$.



$$\Pr(x_1, \dots x_n \mid A=a, B=b) = \frac{1}{b^n} 1_{\text{all } x_i \in [a, a+b]}$$

$$= \frac{1}{b^n} 1_{a \le \min x_i} 1_{\max x_i \le a+b}$$

1. Generate a sample $(\theta_1, \dots, \theta_n)$ from Θ

2. Compute weights $w_i = \Pr_X(x|\Theta = \theta_i)$, then rescale weights to sum to one

Reason about $(\Theta|X = x)$ indirectly

```
# The dataset
x = [2, 3, 2.1, 2.4, 3.14, 1.8]

# Step 1
asamp = np.random.exponential(scale=1/0.5, size=100000)
bsamp = np.random.exponential(scale=1/1.0, size=100000)
#absamp = zip(asamp, bsamp)

# Step 2
w = 1/bsamp**len(x) * np.where((asamp <= min(x)) & (max(x) <= asamp+bsamp), 1, 0)
w = w / np.sum(w)

# Plot the posterior distribution of B
plt.hist(bsamp, weights=w, density=True)
```

Exercise 8.3.2 (Multiple unknowns)

We have a dataset $[x_1, \ldots, x_n]$. We propose to model it as independent samples from $U[A, A+B]$, where A and B are unknown parameters.

Using $A \sim \text{Exp}(0.5)$ and $B \sim \text{Exp}(1.0)$ as prior distributions for the unknown parameters, find the distribution of $(B|\text{data})$.

TIP. If n is large, we can run into underflow problems when we compute $\Pr(x_1, \ldots, x_n|\text{params})$ directly.

Fix: be clever about rescaling the weights, using the log-sum-exp trick (exercise 8.3.4).

TIP. First find the joint posterior distribution for *all* the unknown parameters. Then, just report on the readout we're interested in.

In maths,

$$\Pr_B(b|\text{data}) = \int_a \Pr_{A,B}(a, b|\text{data})\, da$$

Computationally,

1. Generate samples (a_i, b_i) from the joint prior
2. Compute a weight w_i for each pair
3. Plot a weighted histogram of just the b_i

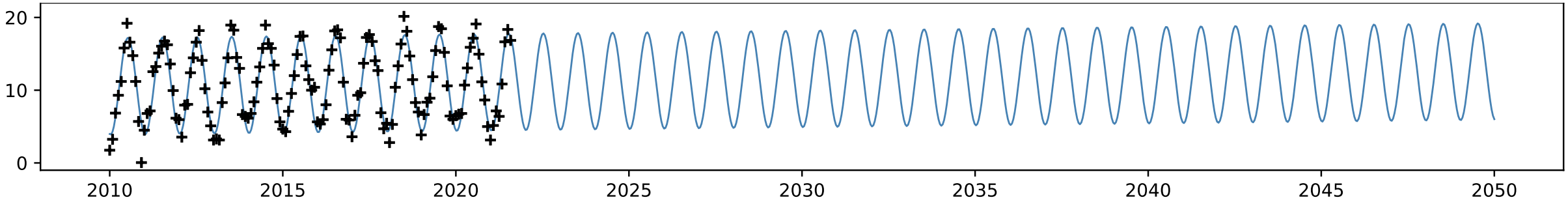
This is called *marginalization*.

Why does computational Bayes work?

It’s a simple argument —
see §6.2 of lecture notes.

Consider the dataset of monthly average temperatures in Cambridge.

Proposed model: $\text{Temp} \sim \alpha + \beta \sin\big(2\pi(t+\phi)\big) + \gamma(t-2000) + N(0,\sigma^2)$



If we fit this model we get the maximum likelihood estimate $\hat{\gamma} = 0.027$ °C/year.

How **confident** are we about this value?

Climate confidence challenge.

Find a 95% confidence interval for γ, for Cambridge from 1985 to the present. (Use your own priors for the unknowns.)

Please submit your answer on Moodle.