Unit 2: Bayesian Learning

1. Bayesian Inference

10/1/2020

Bayesian inference

1. Bayesian probability is a way of thinking about probability as subjective belief.

2. We can use Bayesian inference to compare models of the world

3. Bayesian inference is a framework for learning about the world

Rules of probability

For any event A, let $P(A)$ be the probability of event A

1. $0 \leq P(A) \leq 1$
2. $P(A) + P(\sim A) = 1$
3. Events A and B are *independent* iff $P(A + B) = P(A)P(B)$

$\vdots$

But what is probability?

Why do we think that a coin is fair if $P(\text{heads}) = .5$

Classical probability

The probability of an event is the ratio of the number of cases favorable to it, to the number of all cases possible when nothing leads us to expect that any one of these cases should occur more than any other, which renders them, for us, equally possible.

Pierre-Simon Laplace (1812)

$P\left(\text{heads}\right) = .5$ because there are two outcomes, and nothing makes us think they are not equally likely

The problems with classical probability

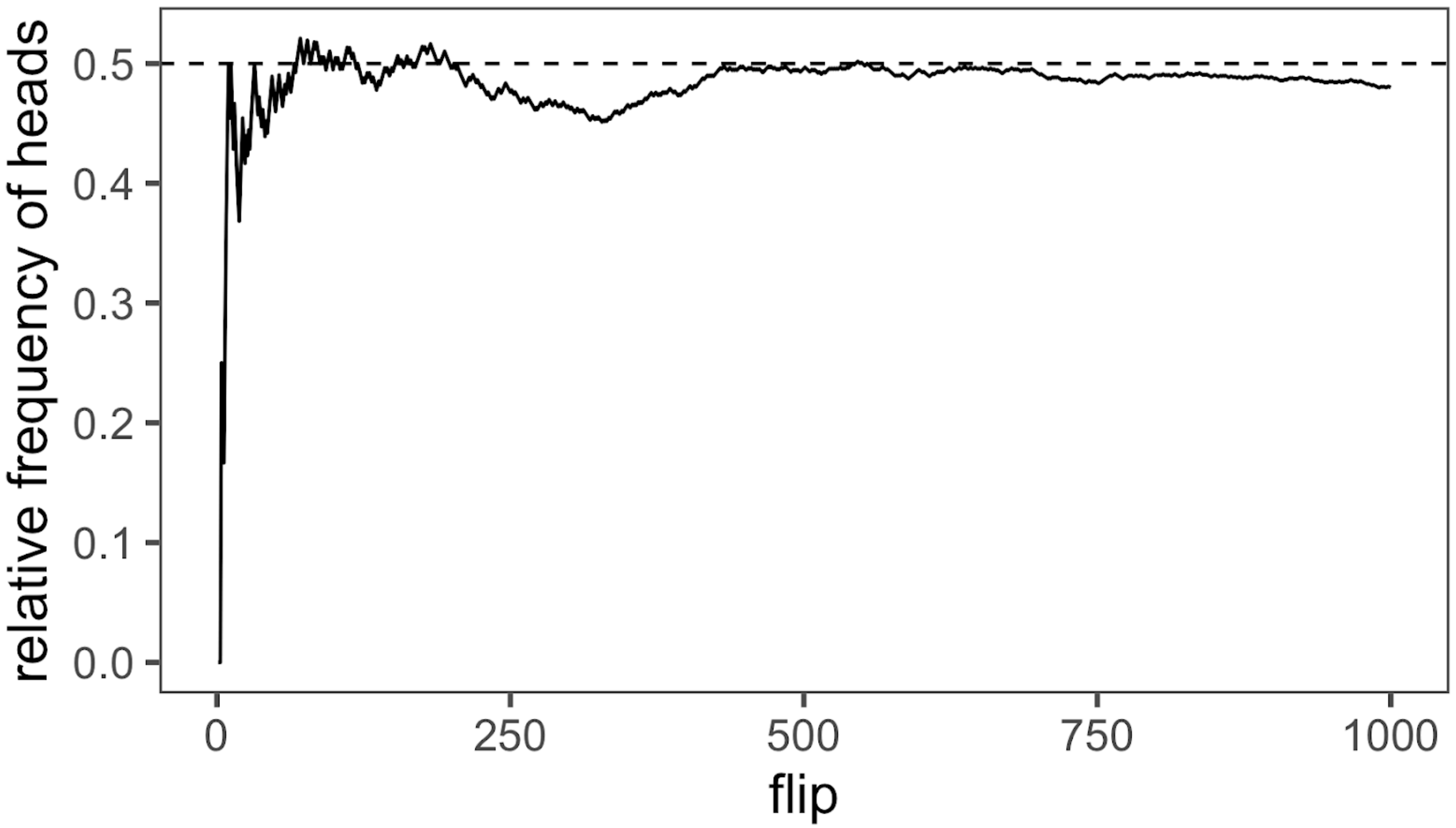
The probability of an event is the ratio of the number of cases favorable to it, to the number of all cases possible when nothing leads us to expect that any one of these cases should occur more than any other, which renders them, for us, equally possible.

Pierre-Simon Laplace (1812)

1. It's **circular**. A fair coin is defined a coin that is fair
2. It's **hard to generalize**. Often, hard to justify the principle of indifference. We'd like talk about cases where we don't know all the possible outcomes, where they aren't equally likely, etc. E.g. probability a bus comes on time.

Frequentist probability

The probability of an event is defined by the limit of its *relative frequency* over many trials of an experiment.



$P\left(\text{heads}\right) = .5$

because if you flip a coin over and over and over again for long enough, half of the flips will have come up heads.

The problems with frequentism

The probability of an event is defined by the limit of its *relative frequency* over many trials of an experiment.

But what about events that have never happened before and will never happen again?

E.g. Probability that the we will have in-person class in the spring

What about things that aren't "events"

E.g. Probability that Germ theory is correct?

Bayesian probability

Probability is subjective, it exists only in your mind.

What you mean when you talk about $P(A)$ is the strength of your belief that A will happen. Think of it as how much you would be willing to bet on A.

Further, your $P(A)$ can be different from my $P(A)$.

$P\left(\text{heads}\right) = .5$ because I expect it to come up heads 50% of the time based on my prior belief about the coin and my experience flipping it.

Reverend Thomas Bayes

Published posthumously by Price, and generalized into the from we use today by Laplace

But how should you form your beliefs?

In practice, we don't want to say you can have any old belief. We want to talk about the belief that a **rational** agent should have after observing some data

Prior probability
(What you used to believe)

Likelihood
(What the data say)

Bayes rule: $$P\left(H|D\right) = \frac{P\left(D|H\right)P(H)}{P(D)}$$

Posterior probability
(What you used to believe)

Deriving Bayes' rule

$$P(A \,\&\, B) = P\left(A|B\right) P(B)$$

Definition of joint probability

$$P(A \,\&\, B) = P\left(B|A\right) P(A)$$

$$P\left(B|A\right) P(A) = P\left(A|B\right) P(B)$$

Transitive property

$$P\left(B|A\right) = \frac{P\left(A|B\right) P(B)}{P(A)} \qquad P\left(H|D\right) = \frac{P\left(D|H\right) P(H)}{P(D)}$$

The problems with Bayesianism

Bayes rule gives you a way to compute how much you should believe in some hypothesis (posterior) if you know three things:

1. The likelihood of the data under that hypothesis
2. The prior probability of that hypothesis
3. The probability of the data

$$P\left(H|D\right) = \frac{P\left(D|H\right) P\left(H\right)}{P\left(D\right)}$$

Problem: We only know the likelihood (1)

Priors are the biggest problem with Bayesianism because priors are *subjective* (i.e. reasonable people can disagree about the right prior).

There are some techniques for dealing with this, but it's a real problem.

Still… priors matter!

Why priors matter

Suppose you wake up tomorrow feeling like you have a fever.

$$P\left(\text{fever} \mid \text{cold}\right) = .01$$

$$P\left(\text{fever} \mid \text{covid-19}\right) = .6$$

$$P\left(\text{fever} \mid \text{malaria}\right) = 1$$

(I made these numbers up)

Which of these ailments do you think you are most likely to have?

Probably covid-19, because $P(\text{covid-19}) \gg P\left(\text{malaria}\right)$

But note, you probably don't have a cold because $P\left(\text{fever} \mid \text{cold}\right)$ is low.

The problems with Bayesianism

Bayes rule gives you a way to compute how much you should believe in some hypothesis (posterior) if you know three things:

1. The likelihood of the data under that hypothesis
2. The prior probability of that hypothesis
3. The probability of the data

$$P\left(H|D\right) = \frac{P\left(D|H\right)P(H)}{P(D)}$$

Problem: We only know the likelihood (1)

You can't compute the probability of the data, but often you don't actually care about the posterior probability of the hypothesis H_1.

You only care whether it is more probable or less probable than some alternative hypothesis H_2

Classical probability

The probability of an event is the ratio of the number of cases favorable to it, to the number of all cases possible when nothing leads us to expect that any one of these cases should occur more than any other, which renders them, for us, equally possible.

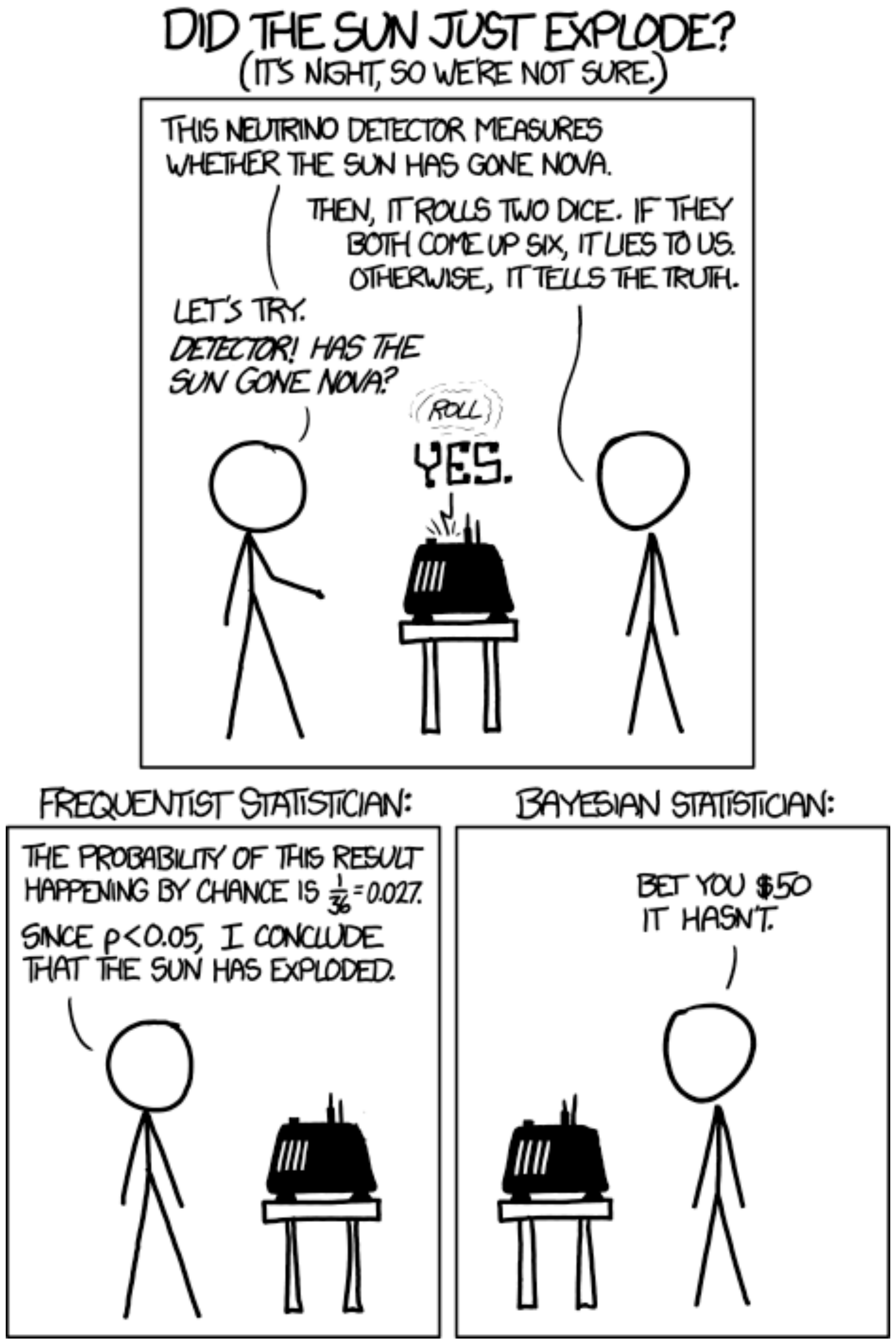
Pierre-Simon Laplace (1812)

1. It's **circular**. A fair coin is defined a coin that is fair
2. It's **hard to generalize**. Often, hard to justify the principle of indifference. We'd like talk about cases where we don't know all the possible outcomes, where they aren't equally likely, etc. E.g. probability a bus comes on time.

The relative probability of two hypotheses

$$\frac{P\left(H_1 \mid D\right)}{P\left(H_2 \mid D\right)} = \frac{\dfrac{P\left(D \mid H_1\right)P\left(H_1\right)}{P(D)}}{\dfrac{P\left(D \mid H_2\right)P\left(H_2\right)}{P(D)}}$$

Often you actually want to compare hypotheses



Null hypothesis testing draws inferences by rejecting the Null
(i.e. finding that you observed data that is unlikely under the null)

But sometimes the data are just unlikely!

Sometimes the data are even more unlikely under a reasonable alternative hypothesis.

Randall Munroe, XKCD
https://xkcd.com/1132/

Frequentism vs Bayesianism

In frequentism, probabilities are **objective**. They are properties of the world defined by the long-run outcomes of random process.

The **parameters** we want to estimate have some true exact value, and we can try to estimate them by talking about how future samples from the random process would look. $P\left(D|H\right)$

In Bayesianism, probabilities are **subjective**. They are properties of the mind of the experimenter.

What are estimating the parameters of hypotheses and not the world. We can talk about how much or how little certainty we have about the truth of our hypotheses. $P\left(H|D\right)$

Bayesian inference for coin flips

HHTHT

HHHHH

What process produced these sequences?

adapted slides by
Josh Tenenbaum

What are hypotheses?

Hypotheses H refer to processes that could have generated the data D.

For each hypothesis H_i, $P\left(D \mid H_i\right)$ is the probability of D being generated by the process identified by hypothesis H_i

Bayesian inference gives us a method for inferring a distribution of belief over these hypotheses, given that we observed data D

Hypotheses H are mutually exclusive: only one process could have generated D

Hypotheses for coin flips

Describe the process by which D could have been generated

$$D = HHTHT$$

- Fair coin $P(H) = .5$
- Biased coin with $P(H) = p$
- Several different coins and a rule about which to flip
- etc

Statistical Models

Comparing hypotheses

1. Two simple hypotheses:

   H_1 Fair Coin — $P(H) = .5$
   H_2 Always Heads — $P(H) = 1$

2. Simple vs complex hypothesis:

   H_1 Fair Coin — $P(H) = .5$
   H_2 Biased Coin — $P(H) = p$

3. Infinitely many hypotheses:

   H_i Biased coin — $P\left(H_i\right) = p_i$

Comparing two simple hypotheses

1. Two simple hypotheses:

H_1 Fair Coin — $P(H) = .5$
H_2 Always Heads — $P(H) = 1$

Bayes rule: $$P\left(H|D\right) = \frac{P\left(D|H\right)P(H)}{P(D)}$$

Ratio form: $$\frac{P\left(H_1|D\right)}{P\left(H_2|D\right)} = \frac{P\left(D|H_1\right)P\left(H_1\right)}{P\left(D|H_2\right)P\left(H_2\right)}$$

Bayes' rule in odds form

$$\frac{P\left(H_1 \mid D\right)}{P\left(H_2 \mid D\right)} = \frac{P\left(D \mid H_1\right) P\left(H_1\right)}{P\left(D \mid H_2\right) P\left(H_2\right)}$$

D: data

H_1, H_2 : models

$P\left(H_1 \mid D\right)$: posterior probability H_1 generated the data

$P\left(D \mid H_1\right)$: likelihood of data under model H_1

$P\left(H_1\right)$: prior probability H_1 generated the data

Odds for two simple hypotheses

$$\frac{P\left(H_1|D\right)}{P\left(H_2|D\right)} = \frac{P\left(D|H_1\right)P\left(H_1\right)}{P\left(D|H_2\right)P\left(H_2\right)}$$

$$D = HHTHT$$

H_1 : “fair coin”

$$P\left(D|H_1\right) = \frac{1}{2}^5$$

$$P\left(H_1\right) = \frac{999}{1000}$$

H_2 : “always heads”

$$P\left(D|H_1\right) = 0$$

$$P\left(H_2\right) = \frac{1}{1000}$$

$$\frac{P\left(H_1|D\right)}{P\left(H_2|D\right)} = \infty$$

Odds for two simple hypotheses

$$\frac{P\left(H_1|D\right)}{P\left(H_2|D\right)} = \frac{P\left(D|H_1\right)P\left(H_1\right)}{P\left(D|H_2\right)P\left(H_2\right)}$$

$$D = HHHHH$$

H_1 : "fair coin"

$$P\left(D|H_1\right) = \frac{1}{2}^{5}$$

$$P\left(H_1\right) = \frac{999}{1000}$$

H_2 : "always heads"

$$P\left(D|H_1\right) = 1$$

$$P\left(H_2\right) = \frac{1}{1000}$$

$$\frac{P\left(H_1|D\right)}{P\left(H_2|D\right)} \approx 30$$

Odds for two simple hypotheses

$$\frac{P\left(H_1|D\right)}{P\left(H_2|D\right)} = \frac{P\left(D|H_1\right)P\left(H_1\right)}{P\left(D|H_2\right)P\left(H_2\right)}$$

$$D = HHHHHHHHHH$$

H_1 : "fair coin"

$$P\left(D|H_1\right) = \frac{1}{2}^{10}$$

$$P\left(H_1\right) = \frac{999}{1000}$$

H_2 : "always heads"

$$P\left(D|H_2\right) = 1$$

$$P\left(H_2\right) = \frac{1}{1000}$$

$$\frac{P\left(H_1|D\right)}{P\left(H_2|D\right)} \approx 1$$

Comparing simple and complex hypotheses

2. Two simple hypotheses:

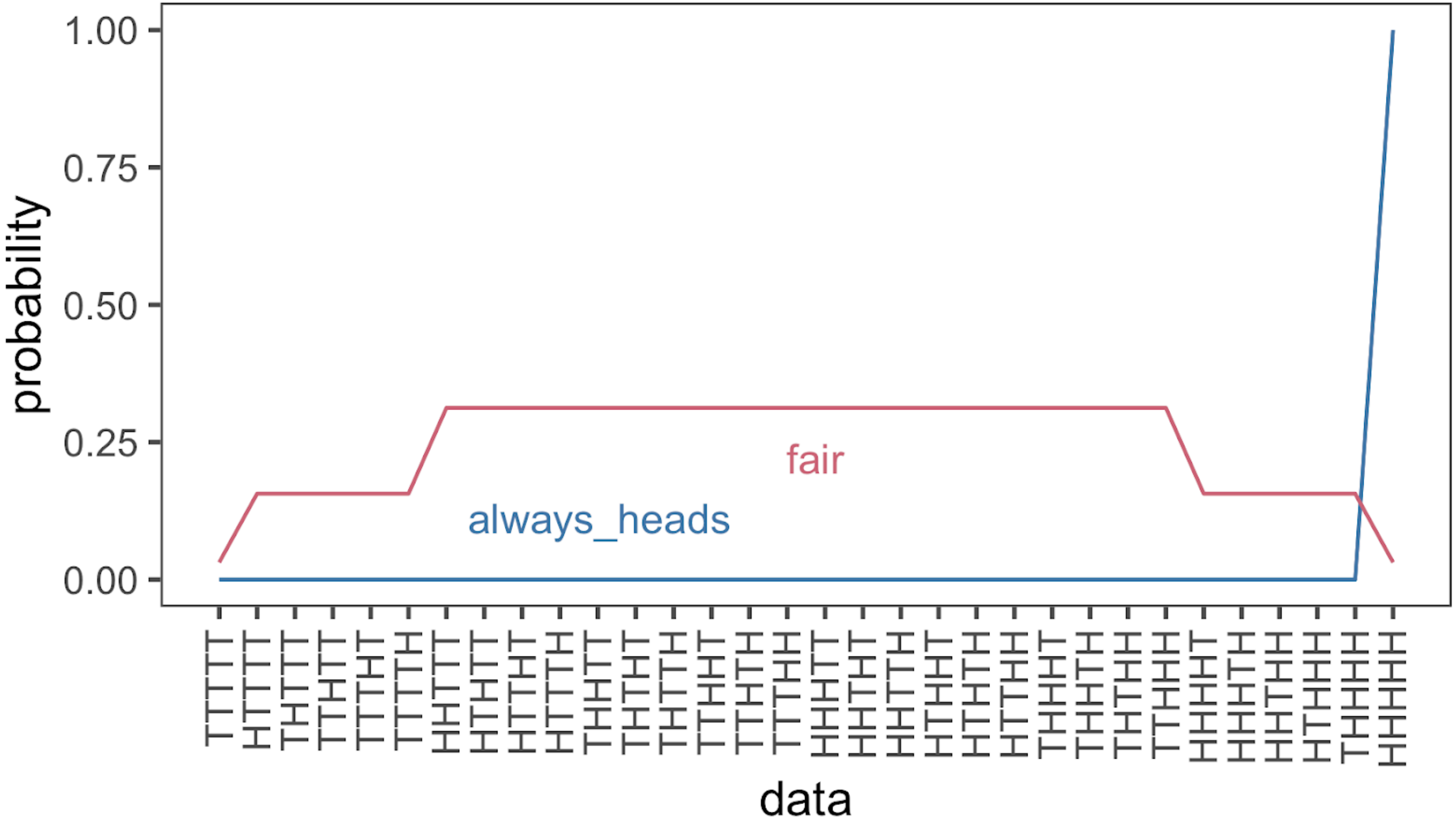
H_1 Fair Coin — $P(H) = .5$
H_2 Always Heads — $P(H) = p$

$H_2 : P(H) = p$ is more complex than $H_1 : P(H) = .5$ in two ways:

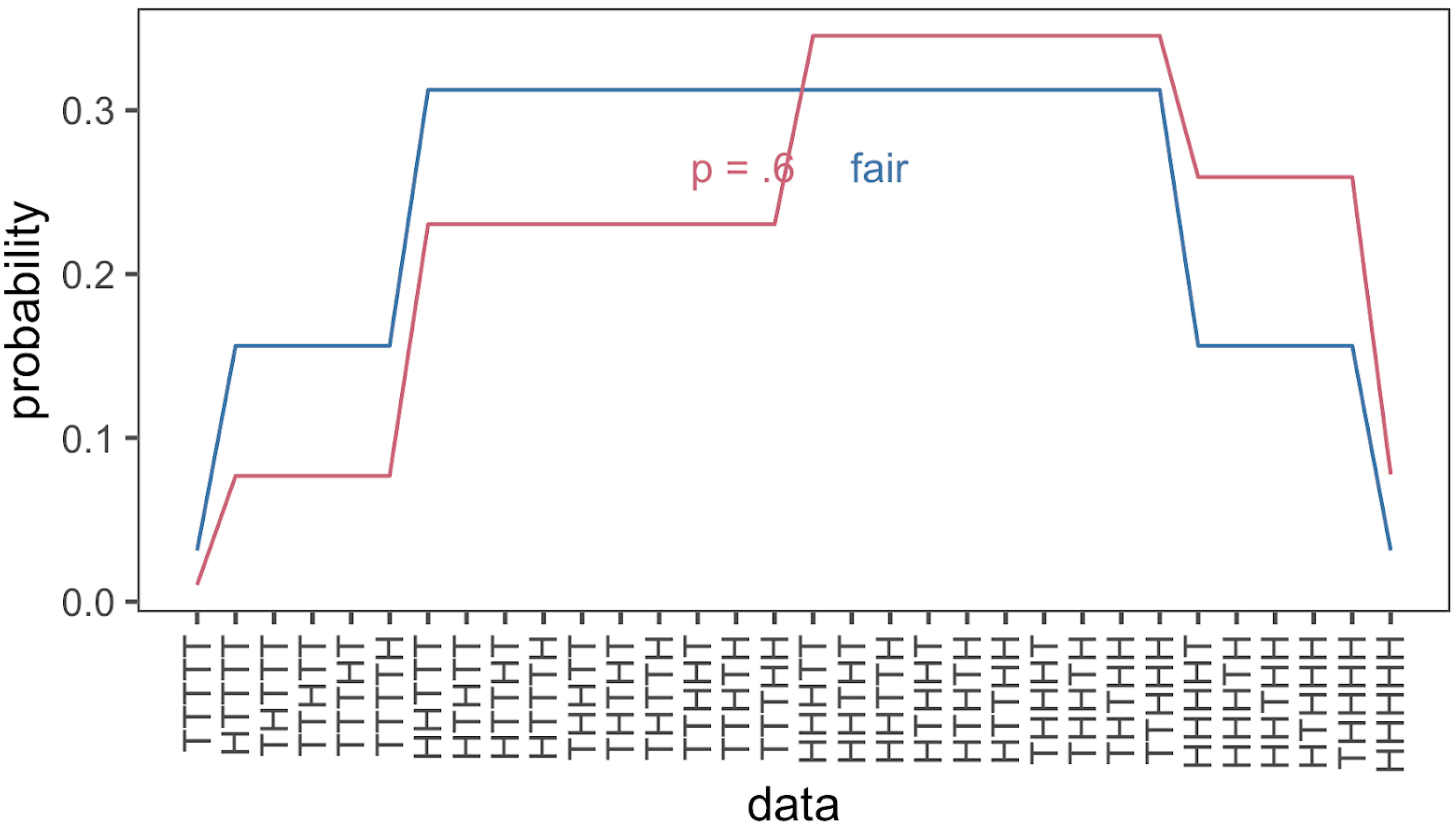
1. H_1 is a special case of H_2

2. for any observed data D,
   we can choose p such that D is more likely under H_2

Comparing simple hypotheses



Comparing simple and complex hypotheses



Comparing simple and complex hypotheses

2. Two simple hypotheses:

H_1 : Fair Coin — $P(H) = .5$
H_2 : Biased Coin — $P(H) = p$

$H_2 : P(H) = p$ is more complex than $H_1 : P(H) = .5$ in two ways:

1. H_1 is a special case of H_2
2. for any observed data D,
   we can choose p such that D is more likely under H_2

How do we deal with this?

1. Frequentist: hypothesis testing
2. Bayesian: falls out of rules of probability

Comparing simple and complex hypotheses

$$\frac{P(H_1|D)}{P(H_2|D)} = \frac{P(D|H_1)\,P(H_1)}{P(D|H_2)\,P(H_2)}$$

$$D = HHTHT$$

$H_1 : P(H) = .5$ $\qquad$ $H_2 : P(H) = p$

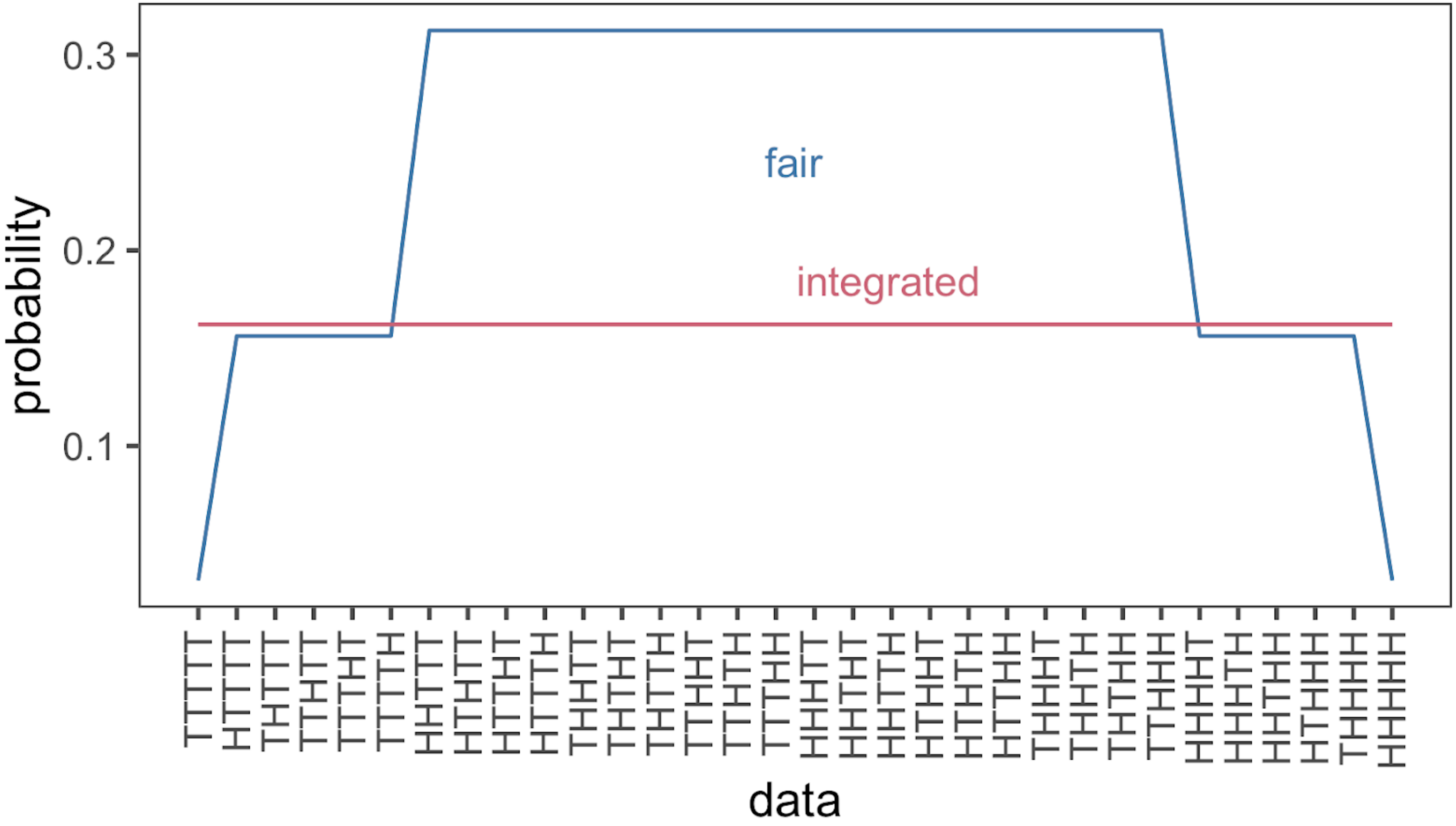
Computing $P(D|H_1)$ is easy: $P(D|H_1) = \frac{1}{2^N}$

We can compute $P(D|H_2)$ by averaging over p:

$$P(D|H_2) = \int_0^1 P(D|p)\,\underline{P(p|H_2)}$$

Prior on p

Assuming every p is equally likely apriori

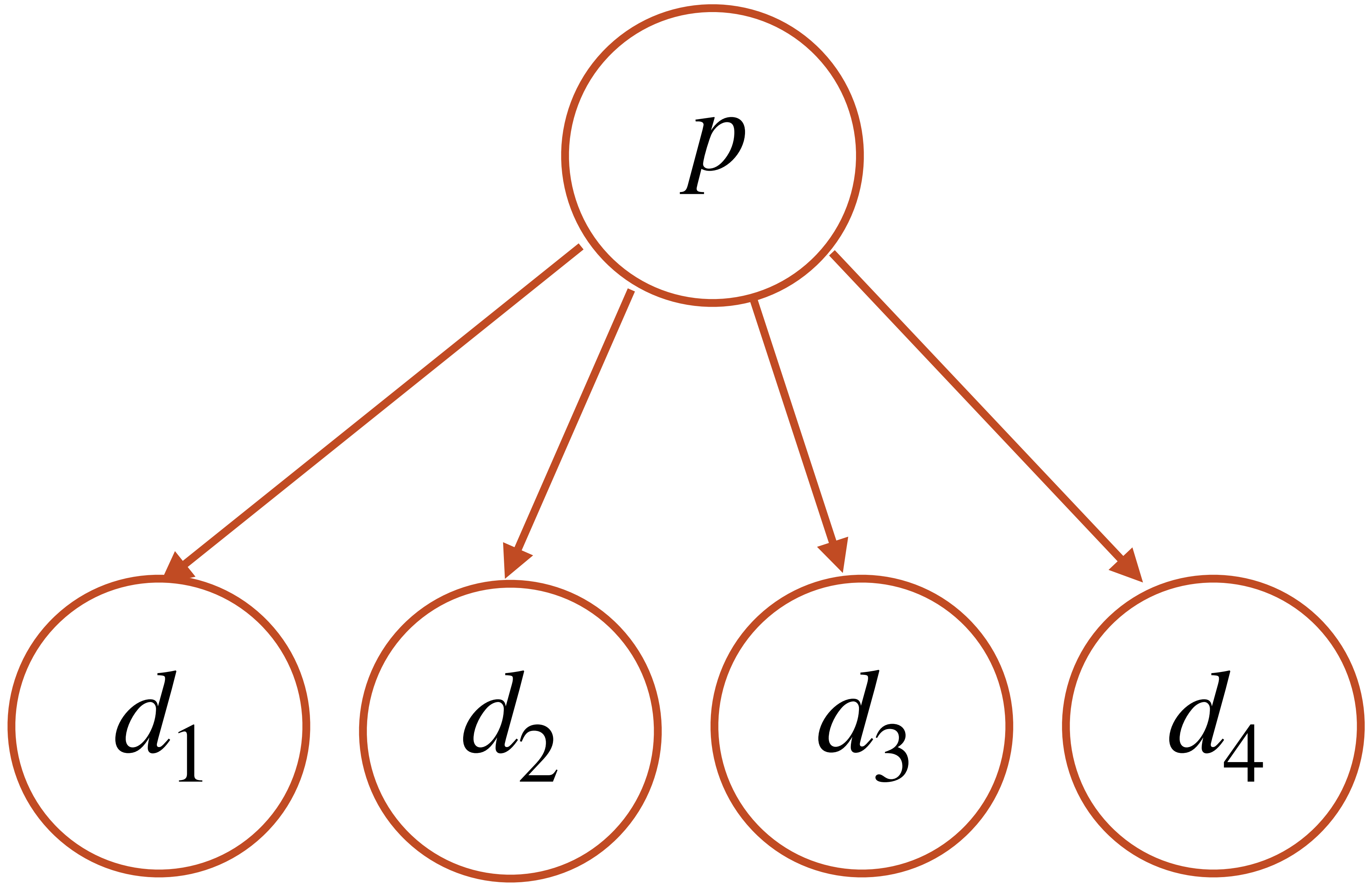


Comparing infinitely many hypotheses

3. Infinitely many hypotheses:

H_i : Biased coin — $P\left(H_i\right) = p_i$

Assume the data are generated from a model:



$P(H) = p$

Picking a likelihood and prior

For a coin with weight p, the likelihood of observing the data D is:

$$P\left(D\,|\,p\right) = p^{N_H}\left(1-p\right)^{N_T}$$

This gives a likelihood.

How do we pick a prior?

Comparing infinitely many hypotheses for coins

Suppose you flipped a coin 10 times and saw 5H and 5T

How likely do you think you are to see H on the next flip?

Probably 50/50, you've seen 5H and 5T

Suppose you flipped a coin 10 times and saw 4H and 6T

How likely do you think you are to see H on the next flip?

Probably closer to 50/50 than 40/60. Why? Prior knowledge

Imagining coin flips

One way of thinking about what you believed is that you are combining your previous experience of coin flips with the data D.

You could model this as seeing
e.g. 5 heads and 5 tails in the past.

Or 50 heads and 50 tails.

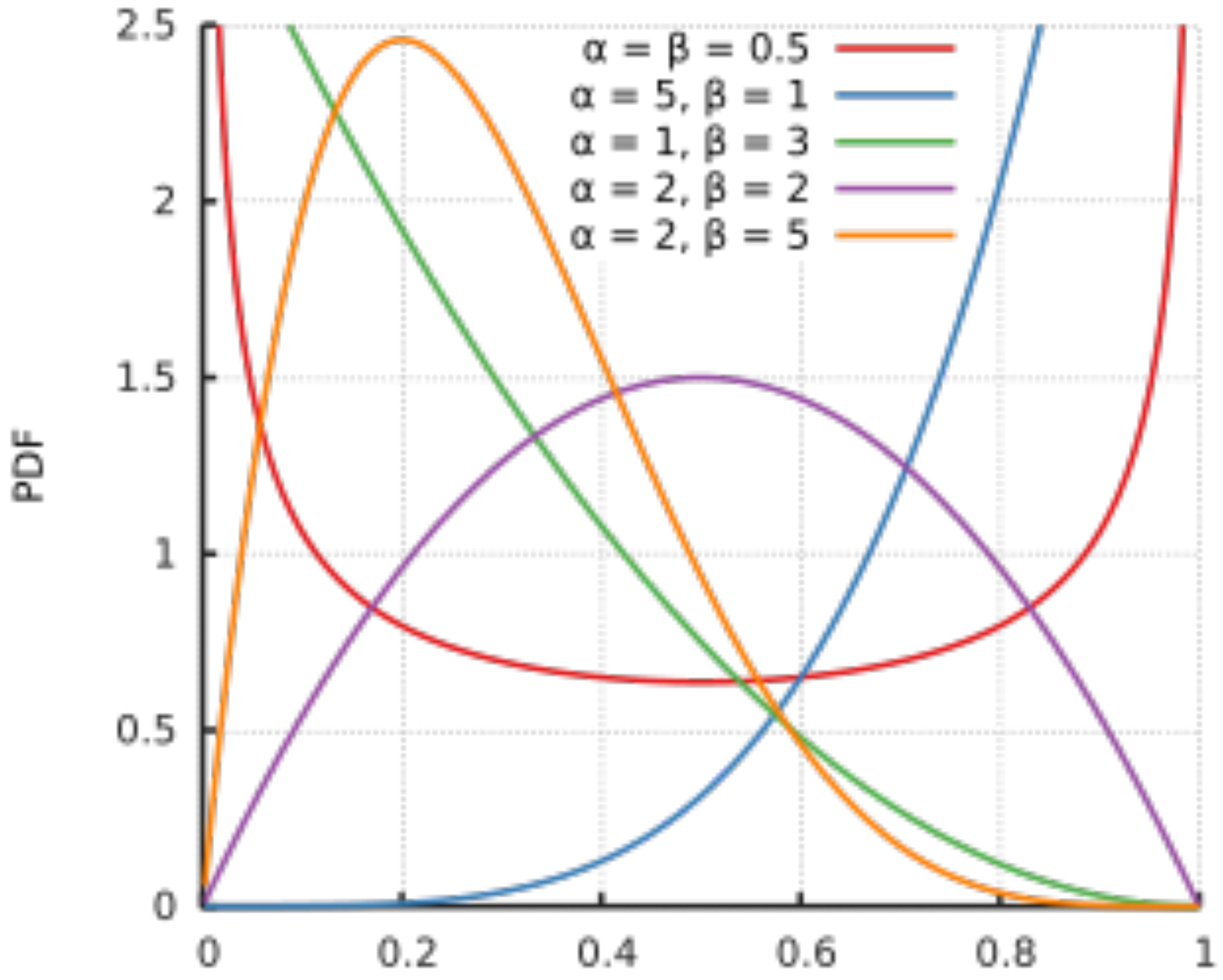
Or 500 heads and 500 tails, etc.

The more experience you have seen the less you should be moved by seeing the data D.

Formalizing imagined coin flips

These hypothetical coin flips can be modeled by a distribution called **Beta** which has two parameters: α and β.

Beta (α, β) encodes models seeing α heads and β tails in the past.



Wikipedia

What does this model predict?

Try this shiny app to explore how changing your prior (by changing α and β), and changing the data you observe, affect your posterior beliefs about the coin weight.

https://shiny.stat.ncsu.edu/jbpost2/BasicBayes/

Bayesian inference

1. Bayesian probability is a way of thinking about probability as subjective belief.

2. We can use Bayesian inference to compare models of the world

3. Bayesian inference is a framework for learning about the world