Unit 6: Bayesian Statistics

1. Basics of Bayesian Inference

4/22/2020

Recap from last time

1. What you mean by "probability" has implications for what statistical tools you should use

2. Bayesian probability conceives of probability as *subjective* rather *objective*. That means you can talk about probability of beliefs rather than of data.

3. This is an active area of research in statistics, and the solutions are less tidy (but also probably less wrong)
   than the models we have used so far

Key ideas

1. Likelihood ratios give us a way to compare models
   (the step function is approximating this)
2. Bayesian inference naturally encodes a preference for simpler models through posterior averaging
3. We can infer the values of unknown parameters in a way that reflects both the data and our prior beliefs

A reminder of Bayes' Rule

Likelihood

(What the data say)

Prior Probability

(What you used to believe)

Bayes' Rule:

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

Posterior Probability

(What you should believe now)

Deriving Bayes' Rule

$$P(A \,\&\, B) = \mathrm{P}(A|B)P(B)$$

$$P(A \,\&\, B) = \mathrm{P}(B|A)P(A)$$

Definition of joint probability

$$\mathrm{P}(B|A)P(A) = \mathrm{P}(A|B)P(B)$$

Transitive property

$$\mathrm{P}(B|A) = \frac{\mathrm{P}(A|B)P(B)}{P(A)}$$

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

Bayesian Inference for Coin Flips

HHTHT

HHHHH

What process produced these sequences?

Slides adapted from a tutorial by Josh Tenenbaum

What are hypotheses?

Hypotheses *H* refer to processes that could have generated the data *D*. for each hypothesis H_i, $P(D|H_i)$ 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 processes by which *D* could be generated

$$D = HHTHT$$

- Fair coin, $P(H) = 0.5$
- Biased coin with $P(H) = p$
- Several different coins and a rule about when to flip which,
- etc...

Statistical models

Comparing Hypotheses

1. Two simple hypotheses:
   $\mathbf{H_1}$: Fair coin — $p(H) = .5$ vs.
   $\mathbf{H_2}$: Always heads — $p(H) = 1$

2. Simple vs. complex hypothesis
   $\mathbf{H_1}$: Fair coin — $p(H) = .5$ vs.
   $\mathbf{H_2}$: Biased coin — $p(H) = p$

3. Infinitely many hypotheses
   $\mathbf{H_i}$: Biased coin — $p(H_i) = p_i$

Comparing simple hypotheses

1. Two simple hypotheses:
   H$_1$: Fair coin — *p(H)* = .5 vs.
   H$_2$: Always heads — *p(H)* = 1

Bayes' Rule:

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

Ratio Form

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

Bayes Rule in Odds Form

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

D: data

H_1, H_2: models

$P(H_1 | D)$: posterior probability H_1 generated the data

$P(D | H_1)$: likelihood of data under model H_1

$P(H_1)$: prior probability H_1 generated the data

Odds for two simple hypotheses

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

D: HHTHT

H_1: "fair coin" vs. H_2: "always heads"

$P(D|H_1) = 1/2^5$ $P(D|H_2) = 0$

$P(H_1) = 999/1000$ $P(H_2) = 1/1000$

$P(H_1|D) \,/\, P(H_2|D) = \textit{infinity}$

Odds for two simple hypotheses

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

D: HHHHH

H_1: "fair coin" vs. H_2: "always heads"

$P(D|H_1) = 1/2^5$

$P(D|H_2) = 1$

$P(H_1) = 999/1000$

$P(H_2) = 1/1000$

$P(H_1|D) / P(H_2|D) \approx 30$

Odds for two simple hypotheses

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

D: HHHHHHHHHH

H_1: “fair coin” vs. H_2: “always heads”

$P(D | H_1) = 1/2^{10}$ $P(D | H_2) = 1$

$P(H_1) = 999/1000$ $P(H_2) = 1/1000$

$P(H_1 | D) \,/\, P(H_2 | D) \approx 1$

Comparing simple and complex hypotheses

2. Simple vs. complex hypothesis

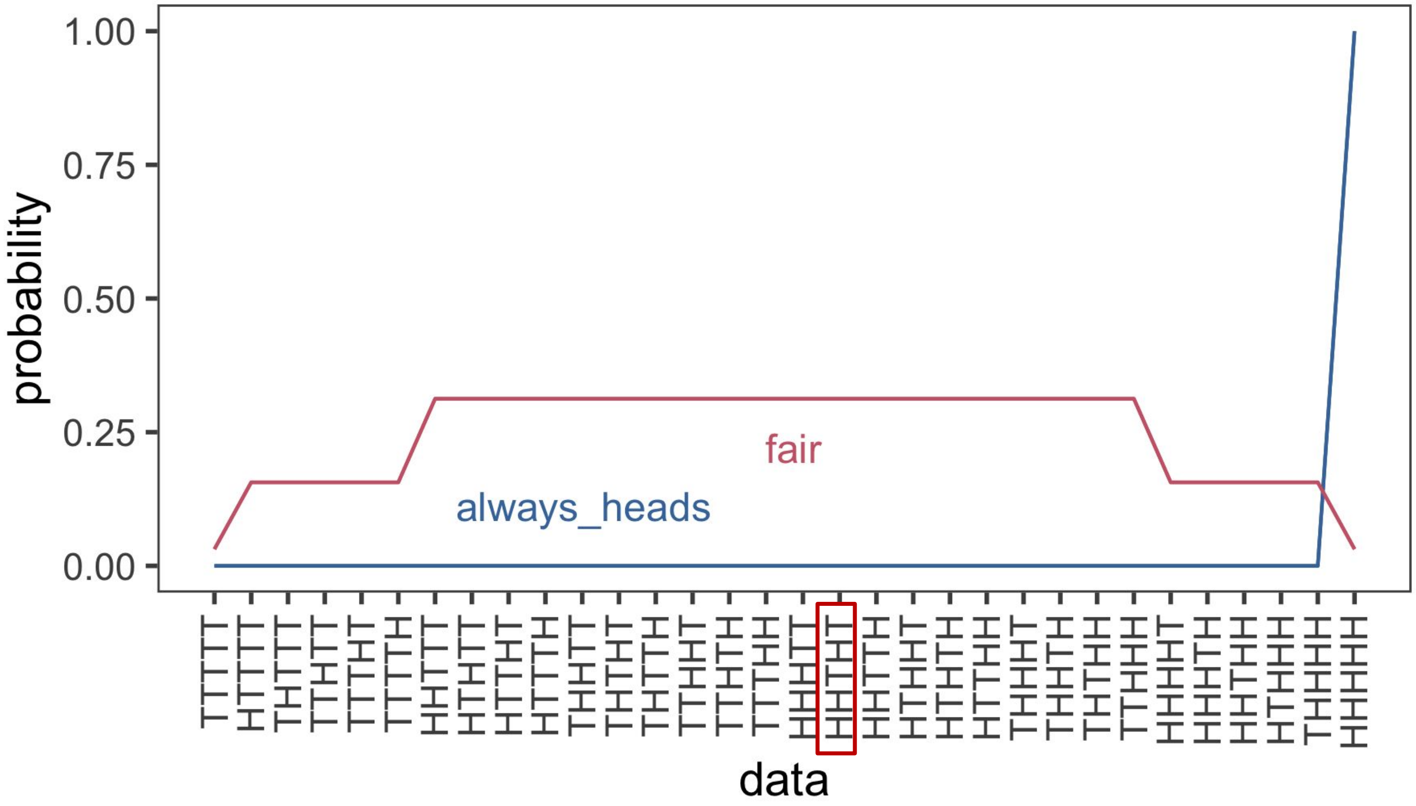
   H_1: Fair coin — $p(H) = .5$ vs.

   H_2: Biased coin — $p(H) = p$

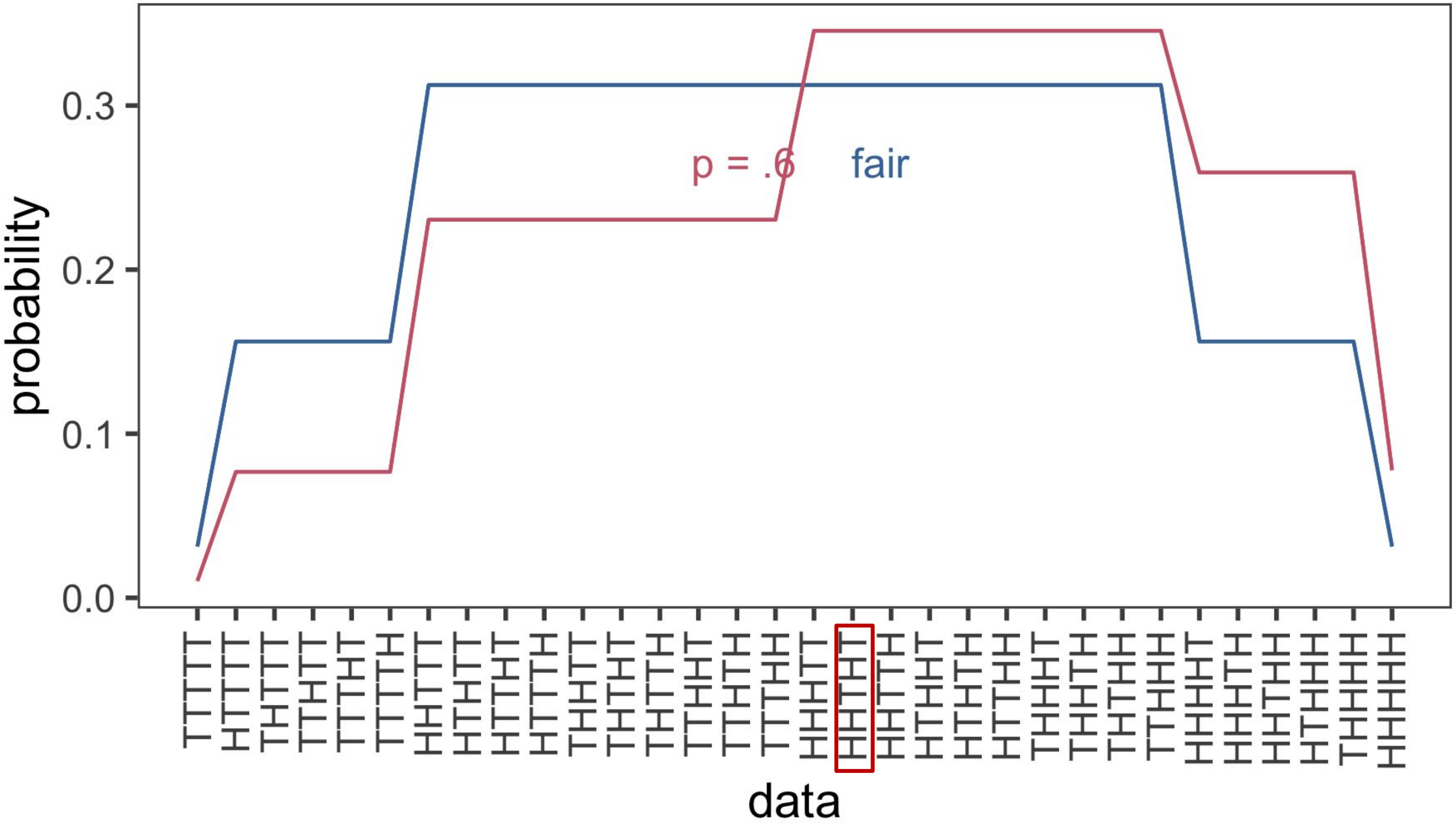
H_2: $P(H) = p$ is more complex than **H_1**:$P(H) = 0.5$ in two ways:

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 like than if P(H) = 0.5

Comparing simple hypotheses



Comparing simple and complex hypotheses



Comparing simple and complex hypotheses

2. Simple vs. complex hypothesis
   H_1: Fair coin — $p(H) = .5$ vs.
   H_2: Biased coin — $p(H) = p$

H_2: $P(H) = p$ is more complex than **H_1**:$P(H) = 0.5$ 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 like than if P(H) = 0.5

How do we deal with this?

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

Comparing simple and complex hypotheses

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

H_1: $p(H) = .5$ vs. H_2: $p(H) = p$

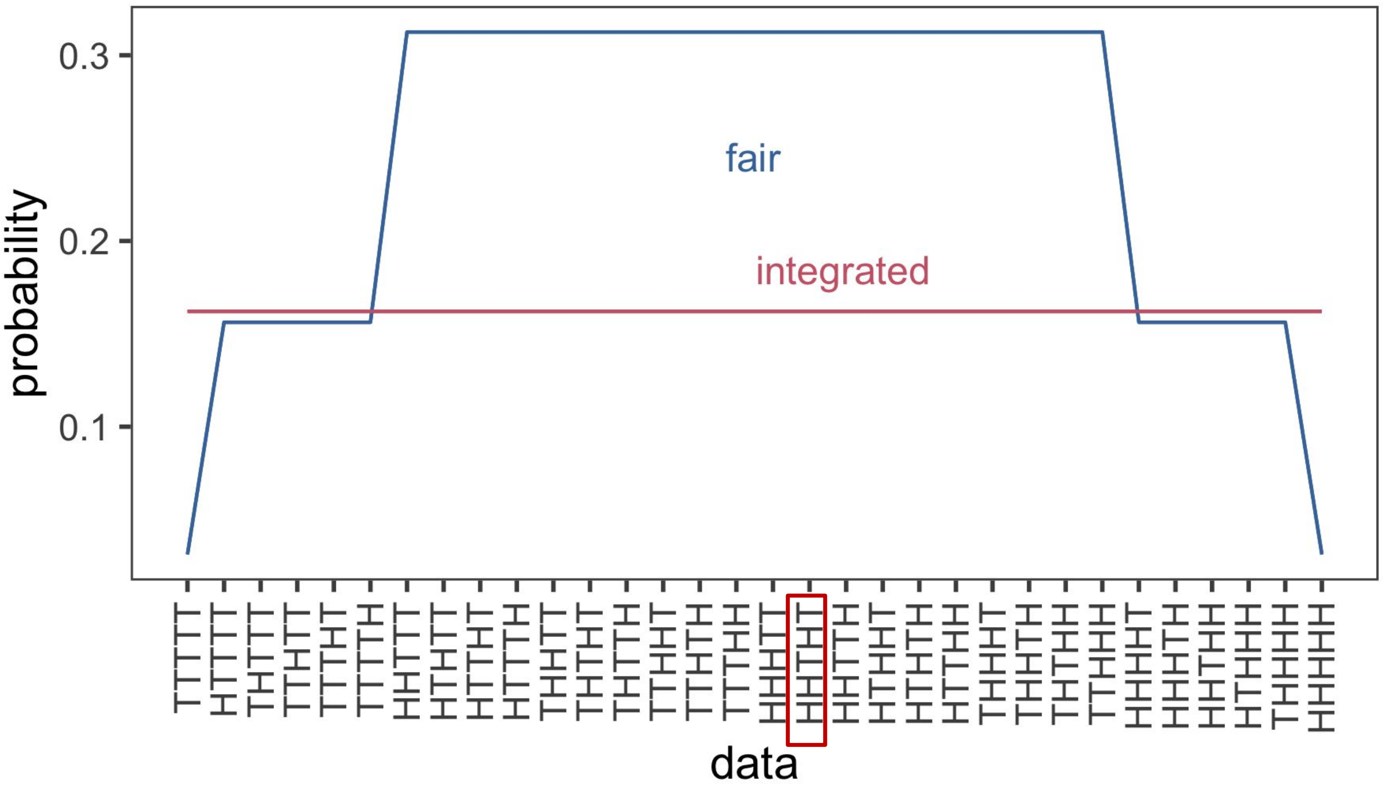
Computing $P(D|H_1)$ is easy: $P(D|H_1) = 1/2^N$

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

$$\mathrm{P}(D|H_2) = \int_0^1 \mathrm{P}(D|p)\,\underbrace{\mathrm{P}(p|H_2)}_{}dp$$

Prior on p

Assuming that every *p* is equally likely apriori

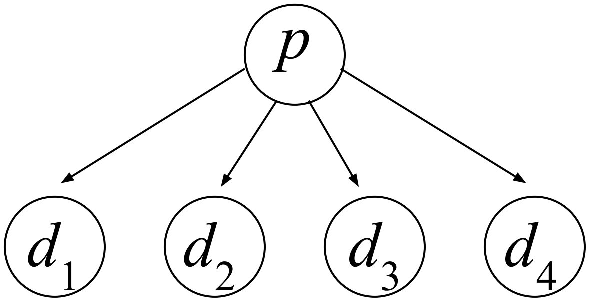


Comparing infinitely many hypotheses

2. Infinitely many hypotheses

   H$_i$: Biased coin — $p(H_i) = p_i$

Assume the data are generated from a model:



$$P(\mathrm{H}) = p$$

Picking a likelihood and prior

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

$$P(D \mid p) = p^{N_\mathrm{H}} (1\text{-}p)^{N_\mathrm{T}}$$

This gives us a likelihood.

But 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 because you have 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.

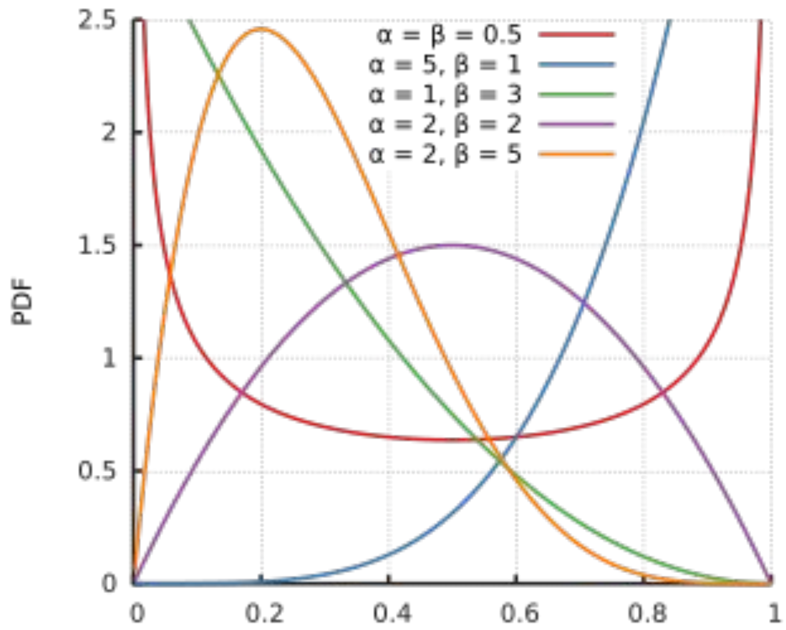
Or 500 heads and 500 tails, etc.

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 change your posterior beliefs about the coin weight.

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

Key ideas

1. Likelihood ratios give us a way to compare models
(the step function is approximating this)

2. Bayesian inference naturally encodes a preference for simpler models through posterior averaging

3. We can infer the values of unknown parameters in a way that reflects both the data and our prior beliefs