Chapter 5

Bayes’ Rule

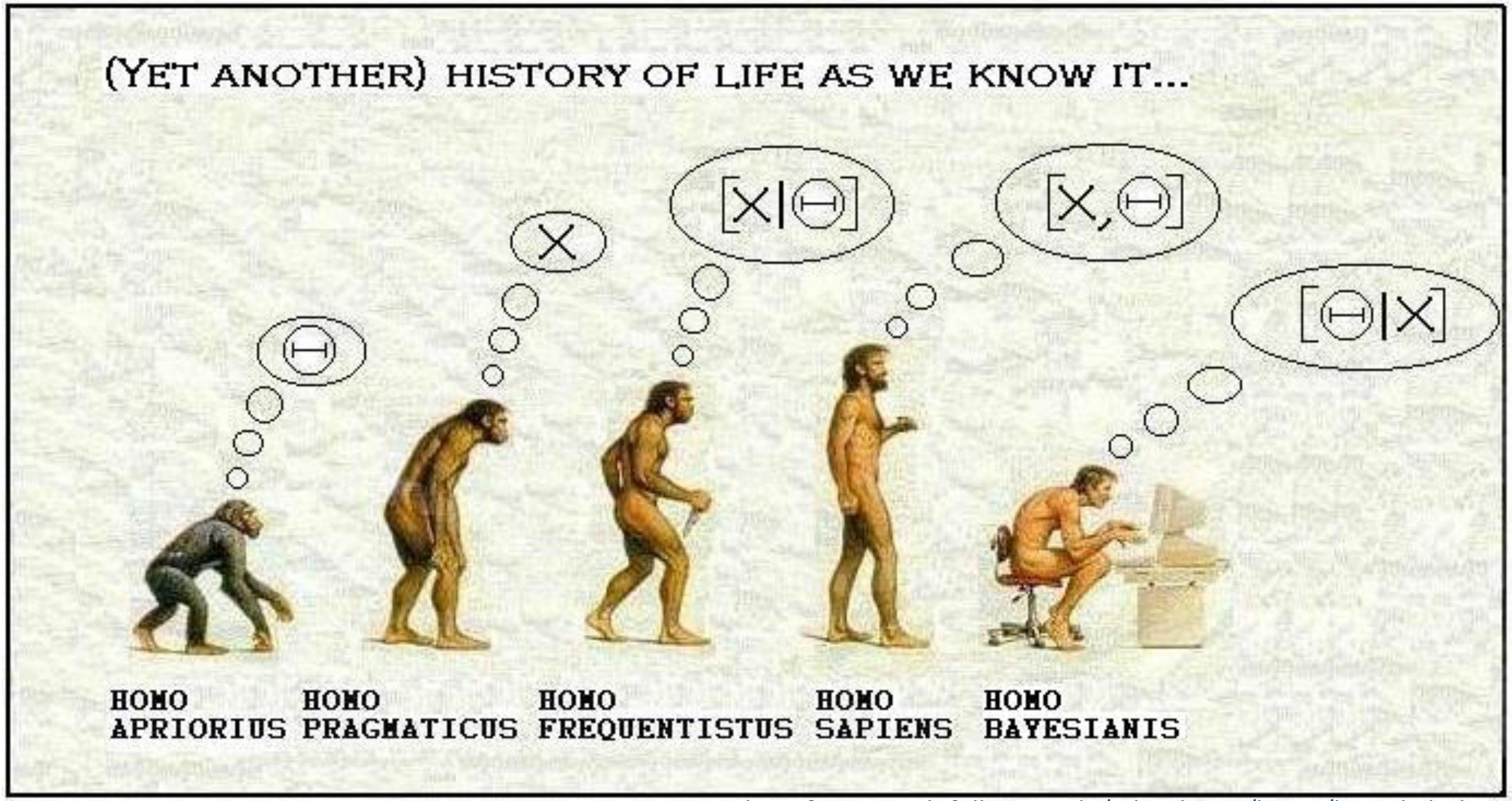


Photo from: psych.fullerton.edu/mbirnbaum/bayes/bayeslinks.htm

Why Bayes' Rule?

- What is the probability that it is cloudy?

1. We start with prior credibility allocated over two possible states of the sky: cloudy or sunny.
2. Then we took into account some other data, namely, that it is raining or that people are wearing sunglasses.

Bayes' rule is merely the mathematical relation between the prior allocation of credibility and the posterior reallocation of credibility conditional on data

Derived from definitions of conditional probability

Conditional probability:

$$p(c|r) = \frac{p(r, c)}{p(r)}$$

the probability of c given r is the probability that they happen together relative to the probability that r happens at all

Re-write the denominator in terms of :

$$p(c|r) = \frac{p(r|c)\, p(c)}{\sum_{c*} p(r|c*)\, p(c*)}$$

the c in the numerator is a specific fixed value, whereas the c* in the denominator is a variable that takes on all possible values

Bayes' rule intuited from a two-way discrete table (cont.)

- The act of spatial attention, when expressed in algebra, yields Bayes' rule

	Hair color				
Eye color	**Black**	**Brunette**	**Red**	**Blond**	**Marginal (Eye color)**
Brown	0.11	0.20	0.04	0.01	0.37
Blue	0.03	0.14	0.03	0.16	0.36
Hazel	0.03	0.09	0.02	0.02	0.16
Green	0.01	0.05	0.02	0.03	0.11
Marginal (hair color)	0.18	0.48	0.12	0.21	1.0

Prior: beliefs about hair color before knowing eye color

Posterior: beliefs about hair color given the observed eye color

	Hair color				
Eye color	**Black**	**Brunette**	**Red**	**Blond**	**Marginal (Eye color)**
Blue	$0.03/0.36 = 0.08$	$0.14/0.36 = 0.39$	$0.03/0.36 = 0.08$	$0.16/0.36 = 0.45$	$0.36/0.36 = 1.0$

Bayes' rule intuited from a disease diagnosis

	Disease		
Test result	**$\theta = \ddot\frown$ (present)**	**$\theta = \ddot\smile$ (absent)**	**Marginal (test result)**
$T = +$	$p(+\mid\ddot\frown)\,p(\ddot\frown) = 0.99 \cdot 0.001$	$p(+\mid\ddot\smile)\,p(\ddot\smile) = 0.05 \cdot (1 - 0.001)$	$\sum_\theta p(+\mid\theta)\,p(\theta)$
$T = -$	$p(-\mid\ddot\frown)\,p(\ddot\frown) = (1 - 0.99) \cdot 0.001$	$p(-\mid\ddot\smile)\,p(\ddot\smile) = (1 - 0.05) \cdot (1 - 0.001)$	$\sum_\theta p(-\mid\theta)\,p(\theta)$
Marginal (disease)	$p(\ddot\frown) = 0.001$	$p(\ddot\smile) = 1 - 0.001$	1.0

$$
\begin{aligned}
p(\theta = \ddot\frown \mid T = +) &= \frac{p(T = + \mid \theta = \ddot\frown)\,p(\theta = \ddot\frown)}{\sum_\theta p(T = + \mid \theta)\,p(\theta)} \\
&= \frac{0.99 \cdot 0.001}{0.99 \cdot 0.001 + 0.05 \cdot (1 - 0.001)} \\
&= 0.019
\end{aligned}
$$

Applied to parameters and data

Model of data:

p(data values | parameters values)
along with the prior, p(parameters values)

We use Bayes' rule to convert that to:

P ($parametersvalues$ | $datavalues$)

Data	Model parameter			Marginal
	...	θ value	...	
$\vdots$		$\vdots$		$\vdots$
D value	$\cdots$	$p(D,\theta) = p(D\|\theta)\,p(\theta)$	$\cdots$	$p(D) = \sum_{\theta^*} p(D\|\theta^*)\,p(\theta^*)$
$\vdots$		$\vdots$		$\vdots$
Marginal	...	$p(\theta)$	...	

Applied to parameters and data (cont.)

$$\underbrace{p(\theta|D)}_{\text{posterior}} = \underbrace{p(D|\theta)}_{\text{likelihood}} \underbrace{p(\theta)}_{\text{prior}} / \underbrace{p(D)}_{\text{evidence}}$$

The denominator is:

$$p(D) = \sum_{\theta*} p(D|\theta^*)p(\theta^*)$$

For continuous variables:

$$p(D) = \int \mathrm{d}\theta^* p(D|\theta^*)p(\theta^*)$$

Data-order invariance

$$p(\theta|D', D) = \frac{p(D', D|\theta)\, p(\theta)}{\sum_{\theta*} p(D', D|\theta*)\, p(\theta*)}$$

Bayes' rule

$$= \frac{p(D'|\theta)p(D|\theta)\, p(\theta)}{\sum_{\theta*} p(D'|\theta*)p(D|\theta*)\, p(\theta*)}$$

by assumption of independence

$$= \frac{p(D|\theta)p(D'|\theta)\, p(\theta)}{\sum_{\theta*} p(D|\theta*)p(D'|\theta*)\, p(\theta*)}$$

multiplication is commutative

$$= p(\theta|D, D')$$

Bayes' rule

Complete examples: estimating bias in a coin

- Identifying the type of data: ‘Head’ – ‘Tail’
- Creating a descriptive model with meaningful parameters
- We have: (is the outcome of a flip)

probability of heads:

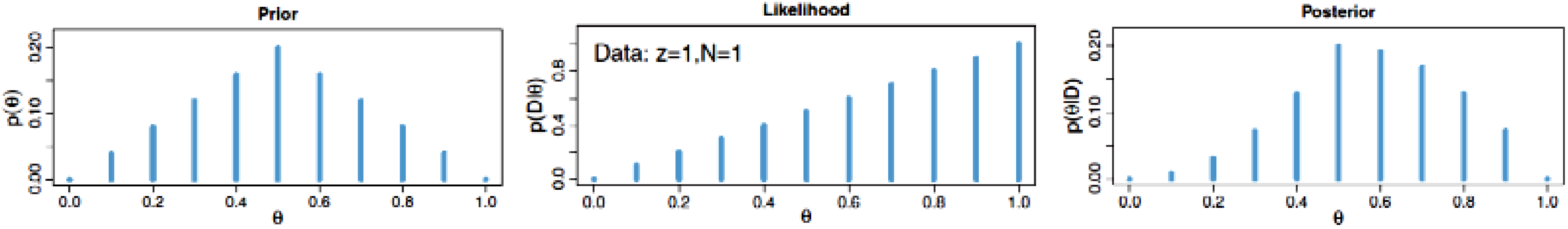
probability of tails:

combined probabilities of heads and tails:

(likelihood function)

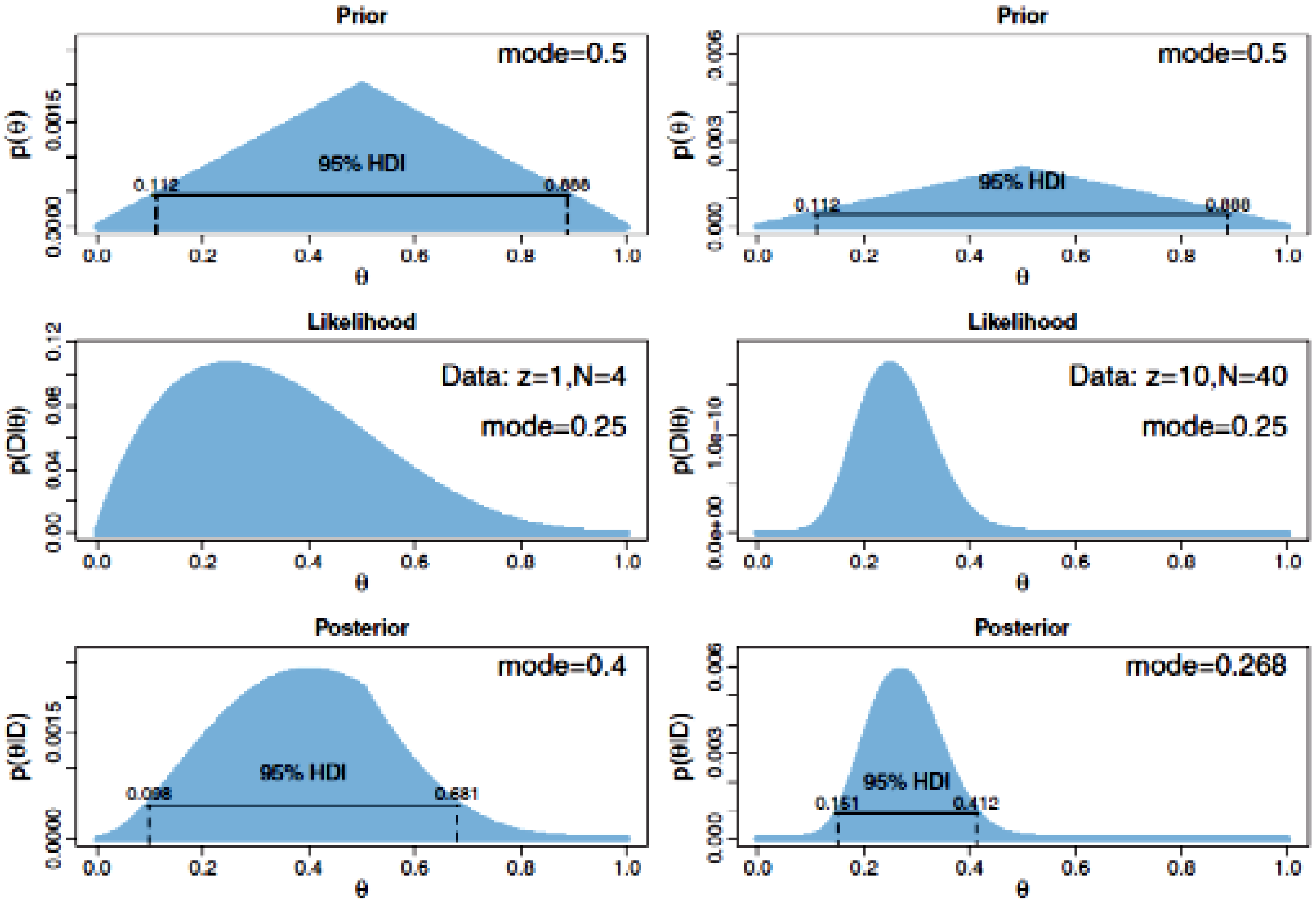
Complete examples: estimating bias in a coin (cont.)

- Establishing a prior distribution over the parameter values ($\theta = 0.0$, $\theta = 0.1$, $\theta = 0.2$, $\theta = 0.3$, and so forth up to $\theta = 1.0$)
- Collecting the data and applying Bayes' rule to re-allocate credibility across the possible parameter values



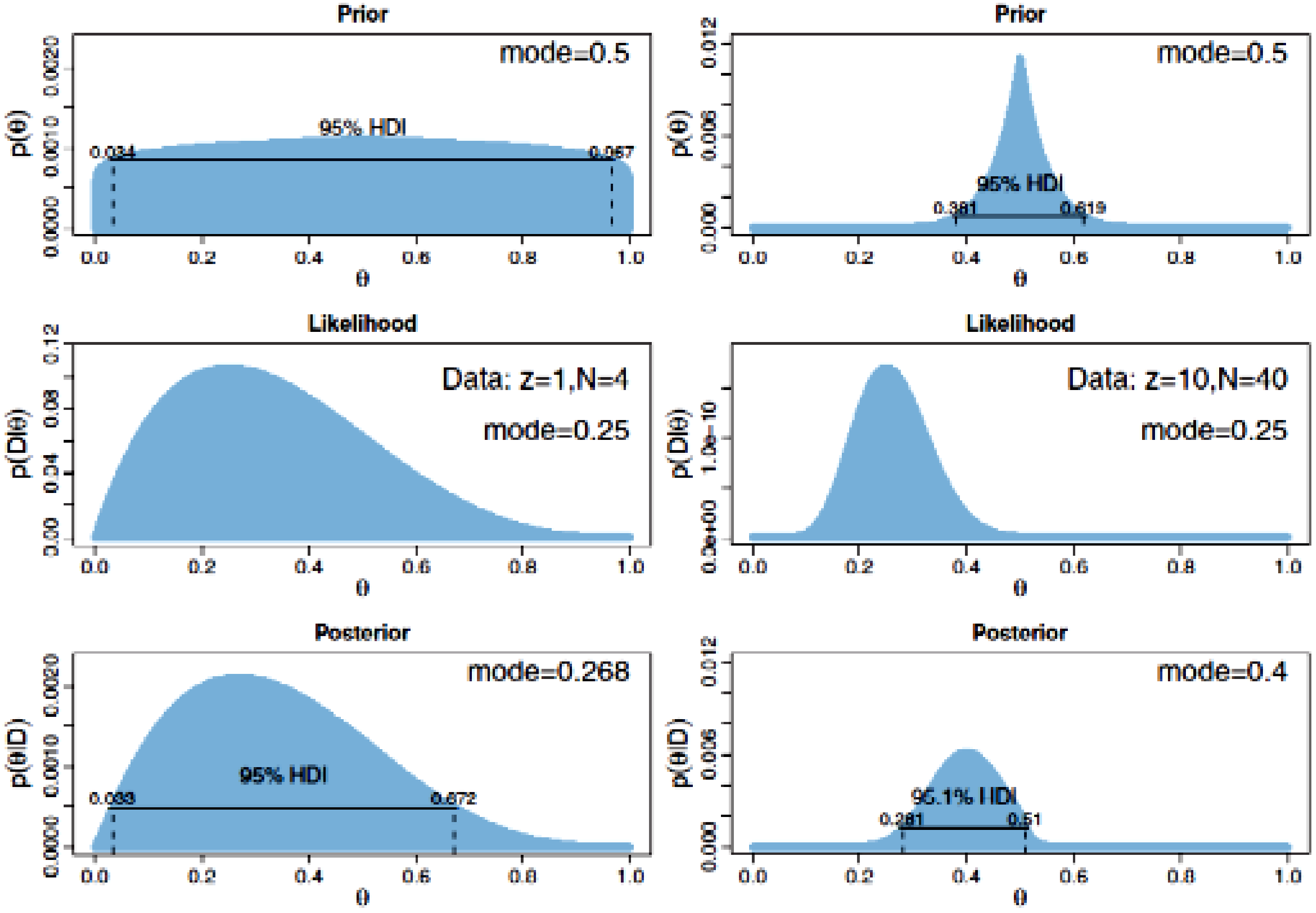
Influence of sample size on the posterior

The posterior model is between the model of the prior distribution and the model of the likelihood function, but it is closer to the likelihood model for larger sample sizes.



Influence of the prior on the posterior

With a strongly informed prior, it takes a lot of novel contrary data to budge beliefs away from the prior. But with a weakly informed prior, it takes relatively little data to shift the peak of the posterior distribution toward the data.



Why bayesian inference can be difficult

- For continuous parameters, the evidence (a.k.a. marginal likelihood) can be impossible to solve analytically
- Simple likelihood function and conjugate prior is not applicable for all of the problems

Suggestions:

- modern computer methods
- variational approximation
- numerical approximation of the integral (not work for models with many parameters)