Beta distribution

Peter Büchel

HSLU W

SA: Week 11

Bayesian Inference: Beta Distribution as Prior Distribution

- Goal of Bayesian inference: Change probabilities with additional relevant information
- Conditional probability and Bayes' theorem: Appropriate tools for adjusting these probabilities

Example

- Circular dartboard divided into 20 equal sections: Labeled 1 to 20
- Frank hits each of the 20 sections equally likely
- He hits dartboard all the time
- Probability that a dart thrown by Frank lands in section i:

$$P(i) = \frac{1}{20}$$

- Same for all sections
- A friend of Frank tells him that he did not hit the 20
- What is the probability that Frank hit the 5?

- Due to this information: Only sections 1 to 19 remain possible
- No preference for Frank to hit one of these regions: Probability 1/19
- Mathematically: Information means that a certain section is now no longer possible
- Original probability of tossing a 20 is subsequently distributed among remaining possible sections
- According to definition of conditional probability:

$$P(5 \mid \text{not } 20) = \frac{P(5 \cap \text{not } 20)}{P(\text{not } 20)} = \frac{P(5)}{P(\text{not } 20)} = \frac{1/20}{19/20} = \frac{1}{19}$$

- Corresponds exactly to our intuitive result
- $P(5)$: *Prior* probability
- Probability that 5 is hit *before* receiving additional information
- $P(5 \mid \text{not } 20)$: *Posterior* probability
- Probability that 5 is hit, but *with* additional information that 20 was not hit

In General

- Denote data by D
- θ: Probability of tossing a coin (later be some parameter of a probability distribution)
- Goal: Draw inferences about parameter θ from data D
- Use Bayes' theorem:

$$P(\theta \mid D) = \frac{P(D \mid \theta) \cdot P(\theta)}{P(D)}$$

- Notation:
  - $P(\theta \mid D)$: Posterior distribution
  - $P(D \mid \theta)$: Likelihood function
  - $P(\theta)$: Prior distribution
  - $P(D)$: *Evidence* or marginal probability

Evidence $P(D)$

- Calculate evidence $P(D)$ with law total probability
- If prior distributions are discrete as in Example last week, then:

$$\begin{aligned} P(D) &= P(D \mid \theta_0)P(\theta_0) + \ldots + P(D \mid \theta_{10})P(\theta_{10}) \\ &= \sum_{i=0}^{10} P(D \mid \theta_i)P(\theta_i) \end{aligned}$$

- In general: Discretised range of θ into n values, then:

$$\begin{aligned} P(D) &= P(D \mid \theta_0)P(\theta_0) + \ldots + P(D \mid \theta_n)P(\theta_n) \\ &= \sum_{i=1}^{n} P(D \mid \theta_i)P(\theta_i) \end{aligned}$$

- Use *probability density functions* instead of probabilities and sums become *integrals*
- For evidence:

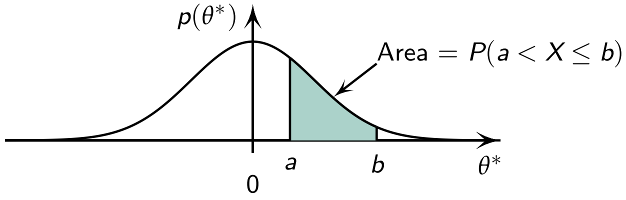
$$P(D) = \int P(D \mid \theta^*)p(\theta^*)\, d\theta^*$$

- Rolling a dice:

$$P(D) = \int_0^1 P(D \mid \theta^*)p(\theta^*)\, d\theta^*$$

Remarks

- To distinguish probabilities from probability densities: Denote probability densities by $p(\cdot)$, probabilities by $P(\cdot)$
- For us: Integrals essentially areas under a probability density curves
- Probability density $p(\theta^*)$ satisfies following properties:
  - $p(\theta) \geq 0$ for all $\theta \in \mathbb{R}$
  - Total area under curve is 1
  - Areas under probability density curves correspond to probabilities:



- Problems with the last two integrals: Generally no solution
- Problems with integrals in denominator of Bayes' theorem were the reason why Bayesian statistic were not really applicable for even mildly difficult problem
- With modern techniques (MCMC) and computer power: Circumvent integrals entirely

Ways Around Difficulty of Integrating

- Choose prior distribution so that integral is easily integrable
- Apply numerical method (MCMC): Completely bypasses integral in denominator of Bayes' theorem
- Start with the first

The likelihood function: Bernoulli distribution for coin toss

- Start with *one* coin toss: Probability of toss θ for H
- Corresponding probability $1-\theta$ for T
- Denote by $y = 1$ that H was tossed
- Correspondingly, $y = 0$ denotes that T was tossed
- Hence

$$P(y = 1 \mid \theta) = \theta \qquad \text{and} \qquad P(y = 0 \mid \theta) = 1 - \theta$$

- Probability distribution:

$$P(y = 1 \mid \theta) + P(y = 0 \mid \theta) = \theta + (1 - \theta) = 1$$

- Called: *Bernoulli distribution*

Multiple coin tosses

- Flip same coin multiple times
- Label outcome of i-th flip as y_i
- Set of outcomes: $\{y_i\}$
- Example: Toss H, T and T again then:

$$\{1, 0, 0\}$$

- Assume: Outcomes of random experiments are independent
- Means: Probability of overall outcome is product of probabilities of individual outcomes
- Example: Outcome in two coin tosses is T and H
- If the two events are independent, then;

$$P(\{0, 1\} \mid \theta) = P(0 \mid \theta) \cdot P(1 \mid \theta)$$

- Because of independency of experiment:

$$\{1, 0, 0\} = \{0, 1, 0\} = \{0, 0, 1\}$$

Example

- Tossed three H and two T
- Denote number of H by z and N total number of tosses
- Then:

$$z = 3 \qquad \text{and} \qquad N = 5 \qquad \text{and} \qquad N - z = 2$$

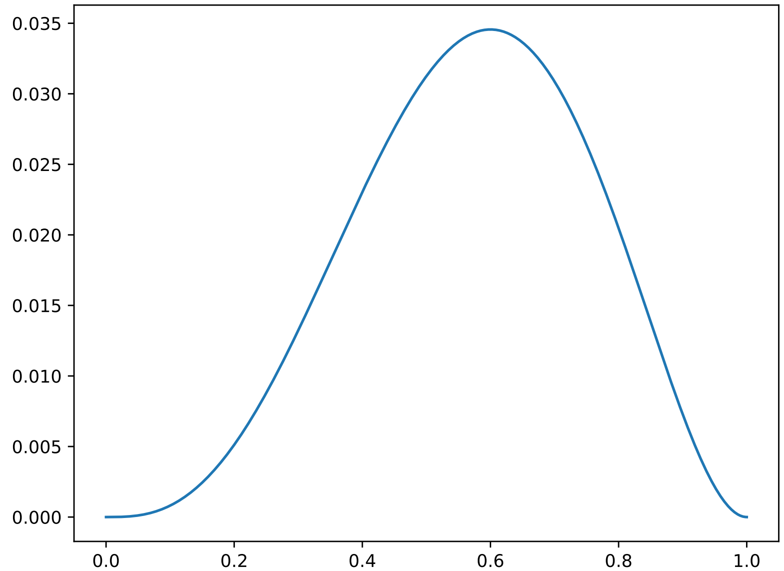
- Want to determine probability:

$$P(\{1, 1, 1, 0, 0\} \mid \theta)$$

- Because of independency

$$\begin{aligned} P(\{1, 0, 1, 1, 0\} \mid \theta) &= P(1 \mid \theta) \cdot P(0 \mid \theta) \cdot P(1 \mid \theta) \cdot P(1 \mid \theta) \cdot P(0 \mid \theta) \\ &= \theta \cdot (1 - \theta) \cdot \theta \cdot \theta \cdot (1 - \theta) \\ &= \theta^3 (1 - \theta)^2 \\ &= \theta^z (1 - \theta)^{N - z} \end{aligned}$$

- Function of θ:



In General

- This last example is easily generalised

Likelihood function for coin tosses

If N denotes number of tosses, z number of H and $N-z$ number of T, then:

$$P(\{y_i\} \mid \theta) = \theta^z (1-\theta)^{N-z} \tag{1}$$

- Formula is useful for applications of Bayes' theorem to large data sets
- Equation (1): Bernoulli likelihood function for multiple tosses
- Data $\{y_i\}$ given, θ variable
- Equation (1): Function in θ

Description of Probabilities: Beta Distribution

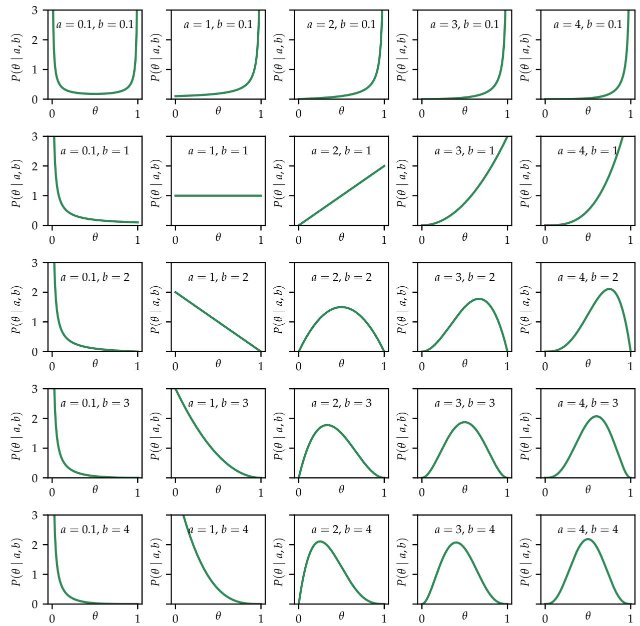
- Given Bernoulli likelihood function
- Need description of prior distribution to us Bayes' theorem
- Need mathematical description of the prior distribution of probabilities
- Need mathematical formula that describes prior distribution for each value of parameter θ on interval $[0, 1]$

- In principle: Any probability density function defined on the interval $[0, 1]$ possible
- Intend to use Bayes' theorem:

$$p(\theta \mid D) = \frac{P(D \mid \theta) \cdot p(\theta)}{P(D)}$$

- Can be shown: If prior distribution is a so-called *Beta distribution* and given Bernoulli likelihood function (1), posterior distribution is another Beta distribution
- Beta distribution: Continuous distribution which is defined on the interval $[0, 1]$ and has two parameters a and b

$p(\theta \mid a, b)$ as Function of θ for Certain Values of a and b



- Note: When b constant and when a becomes larger (from left to right), distribution "moves" to the right along θ
- When b becomes larger with constant a (from top to bottom), distribution moves to the left along θ
- Beta distribution becomes narrower when a and b become larger, i.e., $a + b$ becomes larger
- $a + b$ large: More certainty about where values of θ are concentrated than when $a + b$ small
- Variables a and b: *Form parameters* of beta distribution because they determine its shape
- Although Figure has mainly integer values of a and b, shape parameters can have any positive real value

Specifying a Beta Distribution (Prior)

- Wish to specify beta distribution that describes our prior belief over θ
- Choose a and b such that beta distribution corresponds to our prior belief
- Later: Properties of a and b that allow us to choose a and b cleverly
- But before: Little overview of Figure

Example

- Suppose know nothing about coin, i.e., each θ is "equally likely"
- Choose $a = 1$ and $b = 1$: Beta distribution uniformly distributed (all values of θ are "equally likely")
- If we think that coin is probably fair, but are not quite sure about this, might choose $a = 4$ and $b = 4$
- Because beta distribution reaches its maximum at $\theta = 0.5$ for $a = b$ with higher or lower values of θ also being moderately likely
- If *emphatically* convinced that coin is fair: Might choose $a = b = 100$

- Question: How *exactly* to choose parameters a and b
- Does it make a difference whether to choose $a = b = 5$ instead of $a = b = 4$?
- It *does* make a difference, but not a big one (later)
- However, it *does* make a significant difference whether

$$a = b = 1 \qquad \text{or} \qquad a = b = 10 \qquad \text{or} \qquad a = b = 100$$

Central Tendency

- Often think of prior belief in terms of central tendency and certainty
  - Where are most probable θ'?
  - Spread of θ's

Example

- What's the probability that a randomly selected person is left-handed?
- Based on everyday experience, it is perhaps 10 %, or 5 %, or 15 %
- Everyday experience is different for all people, of course, but proportions just mentioned are plausible and all make sense as prior probabilities
- For example, 90 % is *not* plausible
- Need to determine a and b that express belief

- If not very sure about this probability: Choose a and b rather small
- If uncertainty large: Choose $a + b$ rather small, so that distribution becomes relatively wide
- If uncertainty small: Choose $a + b$ rather large, so that distribution becomes relatively narrow

Example

- Consider coin minted by SNB that shows H after a toss
- Probability of coin showing H should be close to 50 %
- Consistent with assumption that National Bank is trustworthy
- Then choose $a = b$ with $a + b$ large

Central Tendency

- Our goal: Transform a prior belief, which is expressed in terms of tendency and uncertainty, into corresponding parameter values a and b in the beta distribution
- Useful: Express central tendency and spread of beta distr. by a and b
- It turns out: Mean of the Beta$(\theta \mid a, b)$ distribution is given by

$$\mu = \frac{a}{a+b}$$

- Mode (θ-value where distribution takes its maximum) for $a, b > 1$:

$$\omega = \frac{a-1}{a+b-2}$$

- Spread of beta distribution is related to "concentration":

$$\kappa = a + b$$

- See from Figure slide 20 that beta distribution becomes narrower or more concentrated as κ becomes larger

- Solve equations for μ and κ in terms of a and b
- For a and b in terms of mean μ and concentration κ:

$$a = \mu\kappa \qquad \text{and} \qquad b = (1-\mu)\kappa$$

- For a and b in terms of mode ω and concentration κ:

$$a = \omega(\kappa-2)+1 \qquad \text{and} \qquad b = (1-\omega)(\kappa-2)+1 \quad \text{for} \quad \kappa > 2$$

Example

- Want to create beta distribution whose mode is $\omega = 0.80$, and concentration $\kappa = 12$
- Using equations above to obtain corresponding shape parameters:

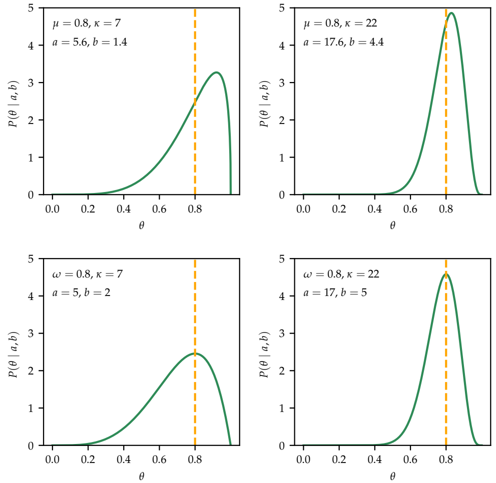
$$a = \omega(\kappa-2)+1 = 0.8 \cdot (12-2) = 9$$

- And:

$$b = (1-\omega)(\kappa-2)+1 = (1-0.8)(12-2)+1 = 3$$

- Mode preferable to mean, especially for very skewed distributions
- Mean of a skewed distribution lies in direction of longer tail

- Figure:



- Bottom panels: Plots of beta distributions with mode is $\theta = 0.8$
- Beta distribution whose mean μ is 0.8: Top panels
- Modes are clearly to the right of mean and not at $\theta = 0.8$

Posterior Beta

- It turns out: Posterior distribution is a again beta distribution for prior beta distribution and Bernoulli likelihood function

Posterior beta

If prior distribution is beta distribution $\text{Beta}(\theta \mid a, b)$, and data show z heads in N tosses, then posterior distribution is again beta distribution:

$$p(\theta \mid z, N) = \text{Beta}(\theta \mid z + a, N - z + b)$$

- Equation is crucial:

$$\underbrace{\text{Beta}(\theta \mid a, b)}_{\text{Prior}} \quad \rightarrow \quad \underbrace{\text{Beta}(\theta \mid z + a, N - z + b)}_{\text{Posterior}}$$

- Simplicity of this updating formula is one of the beauties of mathematical approach to Bayes' inference
- Unfortunately, this approach almost never works

Example

- Suppose prior distribution is

$$\text{Beta}(\theta \mid a = 1, b = 1)$$

- Corresponds to uniform distribution
- Flip coin once and observe H: $N = 1, z = 1, N - z = 0$
- Posterior distribution:

$$\text{Beta}(\theta \mid a + z, b + N - z) = \text{Beta}(\theta \mid 2, 1)$$

- Second row and third column in Figure slide 20
- From this graph: Probability of tossing heads has increased - starting from uniform distribution

- Flip coin again and observe T
- Posterior distribution:

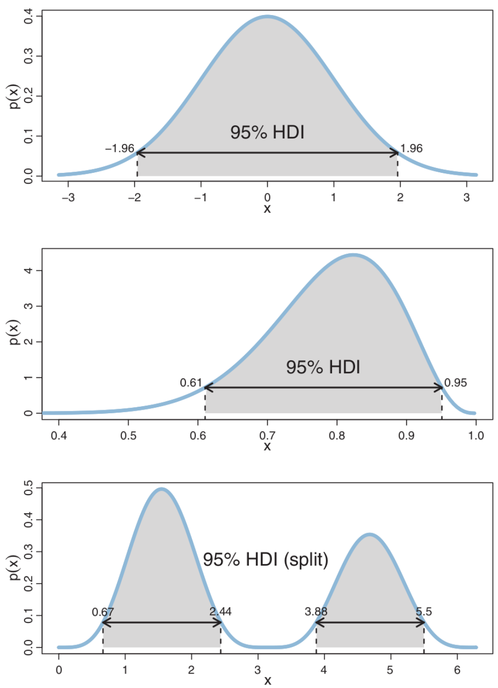
$$\text{Beta}(\theta \mid 2, 2)$$

- Third row and third column of Figure slide 20
- Probability that it could be a fair coin increases
- However: Spread is very large, i.e. value of κ is small
- This process continues for any set of data
- If initial prior is a beta distribution, then posterior is also always a beta distribution

Highest Density Interval (HDI)

- To summarise a distribution: Use *highest density interval* (HDI)
- HDI: Indicates which points of a distribution are most credible and which represent largest part of the distribution, e.g. 95 %
- Any point within HDI has higher credibility than any point outside
- Figure next slide: Three example distributions with HDI's

- Figure:



- Upper panel: Normal distribution with expected value 0 and standard deviation 1
- Since this normal distribution is symmetric about zero, the 95 %-HDI extends from -1.96 to $+1.96$
- Area under curve between these limits shaded grey: Area of 0.95
- Furthermore, probability density of any x within these boundaries has higher probability density than any x outside

- Central panel: Skewed distribution - together with the 95 %-HDI
- According to definition of HDI: Grey shaded area under curve between 95 %-HDI boundaries has area of 0.95
- Probability density for any x inside these boundaries is higher than any x outside
- Note: Area under density curve in left tail (left of lower HDI boundary) is larger than area in right tail
- HDI does not necessarily produce equal area tails outside HDI

- Lower panel: Fanciful bimodal probability density function
- In many realistic applications: Multi-modal distributions like this do not occur, but useful to illustrate definition of HDI
- In this case: HDI is divided into two sub-intervals, one for each mode of distribution
- Grey shaded area under density curve within 95 %-HDI boundaries: Total area of 0.95, and each x within these boundaries has a higher probability density than any x outside

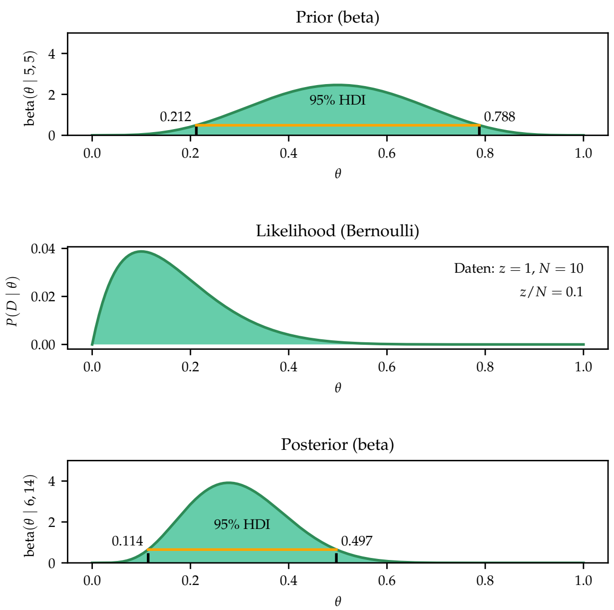
- If distribution refers to probabilities of values, then *width* of HDI is another way to measure uncertainty of beliefs
- If HDI wide: Belief is uncertain
- If HDI narrow: Belief relatively certain because know more precisely *where* 95 % of most likely values are

Remarks

- *Convention*: 95 % of most probable values is chosen for HDI
- Could also have chosen 94 % or 90 % or even 50 %
- Choice seems arbitrary, and to some extent it is, there is agreement among researchers on how to choose this percentage reasonably
- HDI has *nothing* to do with confidence interval

Example

- Figure:



- HDI: Most credible parameter values of distribution, which cover majority of distribution
- HDI summarises the posterior distribution in interval that covers most of the distribution, say 95 %
- Any point inside the HDI has higher credibility than any point outside

Example

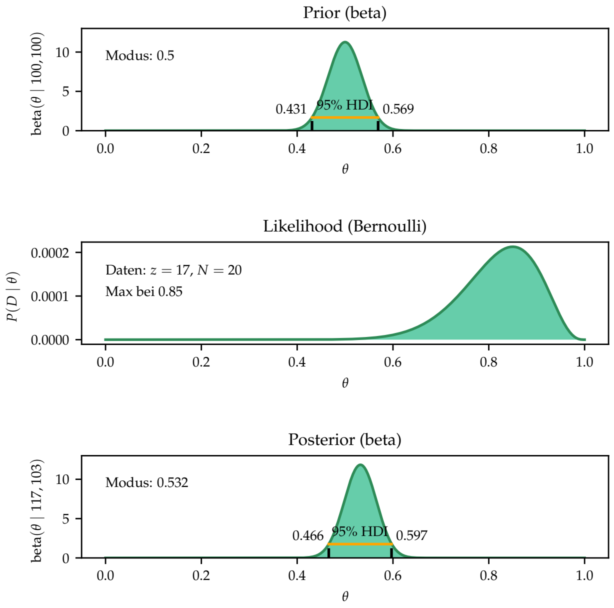
- Let θ be a measure in percent of population's preference for a candidate A over a candidate B
- For poll: Want estimate with 95 %-HDI that width does not exceed 5 %
- Representative poll: Candidate A ahead of candidate B, with a 95 %-HDI of $[0.25, 5.25]$
- Very sure that A will win election, since more than 95 % of the most likely values indicate that A will win
- If HDI is $[-1, 4]$, then candidate B still has a small but realistic probability of winning election, since a small fraction of the 95 % most likely values indicate win for B

Example

- Following three examples: Influence of the prior and likelihood function on posterior
- Note: Likelihood function the same: $N = 20$ and $z = 17$

Example

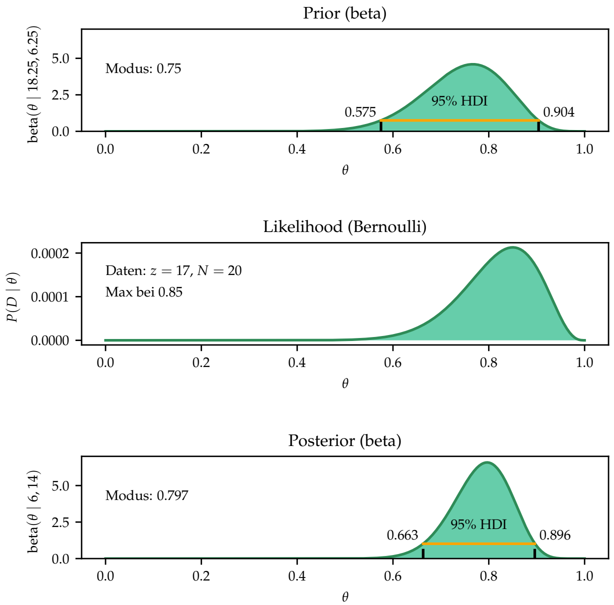
- Figure:



- Very sure that $\theta = 0.5$ since $a + b = 200$ large
- Compared to prior distribution, relatively little data was collected, so likelihood function has very little effect on posterior distribution
- Mode moves from 0.5 to 0.532
- Note: HDI from the posterior has become smaller
- With additional information: Obtained more certainty about most probable values of θ

Example

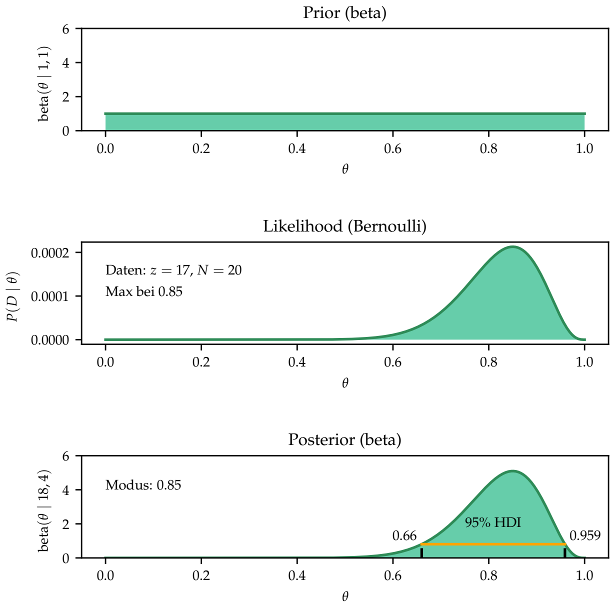
- Figure:



- Prior distribution and likelihood function are similar
- Thus the posterior distribution is also similar to the two
- Data confirm our prior belief
- Again, HDI becames smaller with additional information

Example

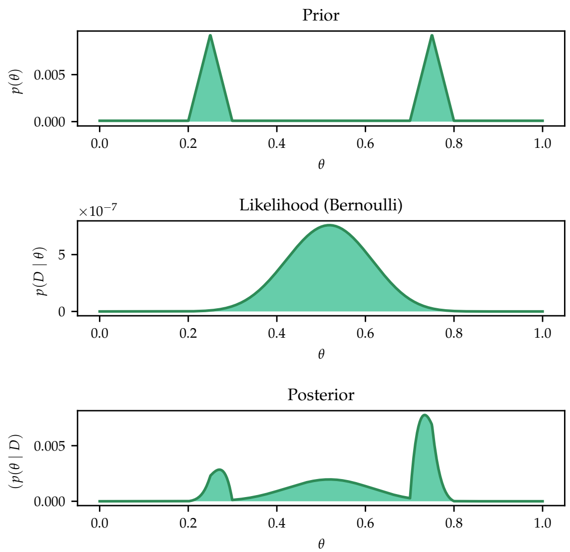
- Figure:



- Prior distribution is uniformly distributed
- Posterior distribution depends exclusively on likelihood function

Example

- Figure:



- Prior distribution does not correspond to a beta distribution
- Cannot apply our procedure either
- Use different techniques

Influence of prior on posterior distribution

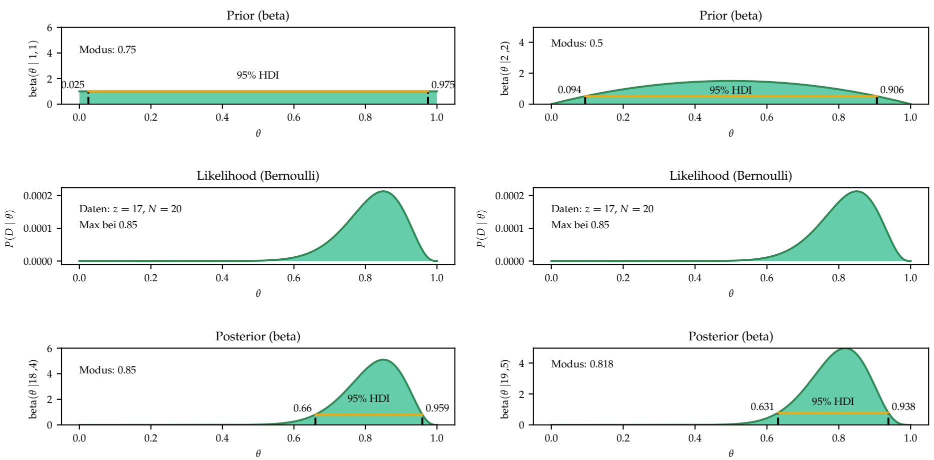
- Referred several times to importance of choice of prior distribution
- One criticism of Bayes inference: Seemingly subjective choice of prior

Example

- Fair coin
- Prior distribution: $a = b$
- Take data from previous examples: $z = 17$ and $N = 20$
- Compare HDIs of . prior and posterior distributions
- Assume: Know very little about coin
- Then small $a + b = 2a$

Example

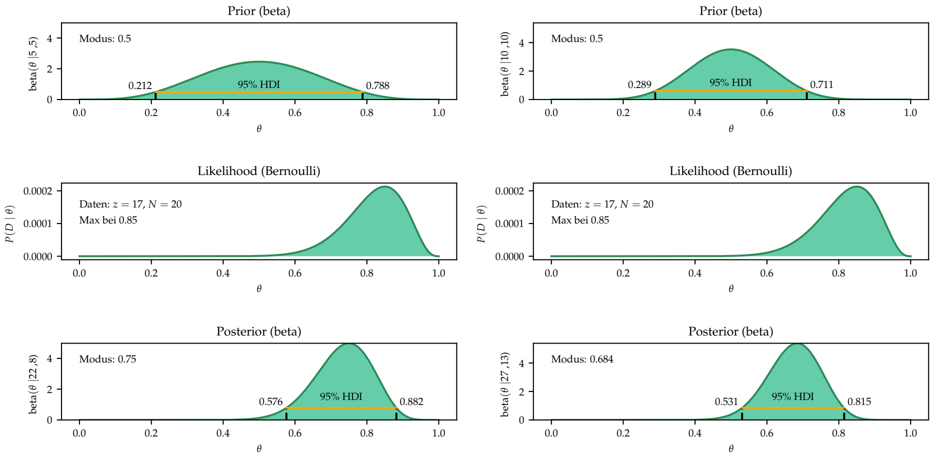
- Figure:



- Figure: Uniform distribution on the left and $a = b = 2$ for the prior on the right
- HDI's of two prior distributions very wide: 0.95 and 0.802
- HDI's of posterior distributions very similar: $[0.66, 0.959]$ and $[0.631, 0.938]$ each with a width of about 0.3
- Mode also very similar for both posterior distributions: 0.85 and 0.818
- Practical relevance: Does not matter whether choice is

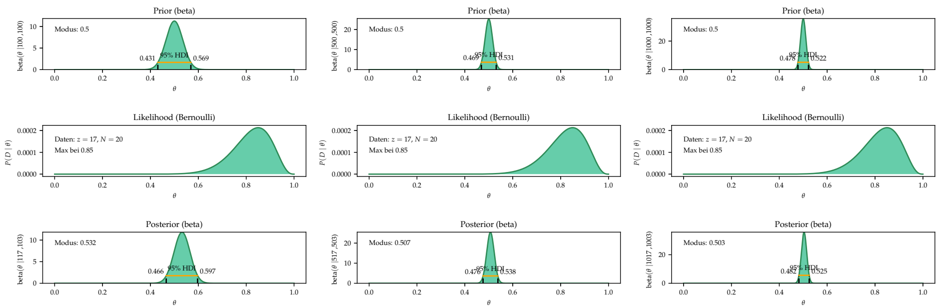
$$a = b = 1 \qquad \text{or} \qquad a = b = 2$$

- Figure:



- Two examples of not so large uncertainty: $a = b = 5$ and $a = b = 10$
- HDI's of the prior and posterior distributions are very similar, as is the mode of the posterior distribution
- Practical relevance: There does not seem to be much difference

- Figure:



- Three examples of large certainty with $a = b = 100, 500, 1000$
- HDI's of prior and posterior distributions respectively are very similar, as is the mode of posterior distribution
- Practical relevance: There does not seem to be much difference

Conclusion

- Practical relevance: Does not matter whether choice $a = b = 10$ or $a = b = 15$ for prior distribution
- However: It does matter whether

$$a = b = 1 \qquad \text{or} \qquad a = b = 10 \qquad \text{or} \qquad a = b = 100$$

- This choice is *not* arbitrary
- If we have no idea about coin: Choose $a = b = 1$
- If coin *looks* very symmetrical: Might choose $a = b = 10$
- If we have examined the coin more closely and find that it is very symmetrical, we may choose $a = b = 100$ or even larger values