Bayesian Methods for Data Science (DATS 6450 - 11)

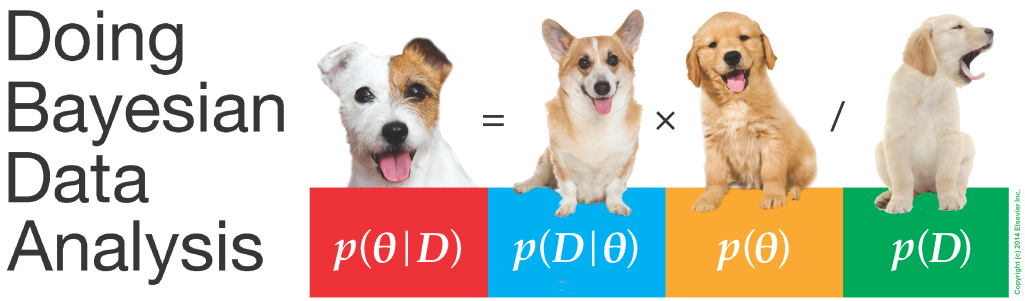
Bayes' Rule

Yuxiao Huang

Data Science, Columbian College of Arts & Sciences
George Washington University
yuxiaohuang@gwu.edu

September 11, 2019

Reference



Picture courtesy of the book website

- This set of slices is an excerpt of the book by Professor John K. Kruschke, with some trivial changes by the creator of the slides
- Please find the reference to and website of the book below:
  - Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier
  - `https://sites.google.com/site/doingbayesiandataanalysis/`

Overview

1. Bayes' rule
2. Applied to parameters and data
3. Complete examples: estimating bias in a coin
4. Why bayesian inference can be difficult

Bayes' rule

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

Example: reallocating posterior credibility

- Suppose you have some prior belief about cloudy weather, $P(\text{cloudy})$
  - suppose you have some data showing it is raining outside
    - now you have a posterior belief about the weather being cloudy, given the data, $p(\text{cloudy}|\text{raining})$
    - **Q:** how does the data change your belief?

Example: reallocating posterior credibility

- Suppose you have some prior belief about cloudy weather, $P(\text{cloudy})$
  - suppose you have some data showing it is raining outside
    - now you have a posterior belief about the weather being cloudy, given the data, $p(\text{cloudy}|\text{raining})$
    - **Q:** how does the data change your belief?
    - **A**:

$$p(\text{cloudy}) < p(\text{cloudy}|\text{raining})$$

  - suppose, instead, you have some data showing everyone outside is wearing sunglasses
    - now you have a posterior belief about the weather being cloudy, given the data, $p(\text{cloudy}|\text{sunglasses})$
    - **Q:** how does the data change your belief?

Example: reallocating posterior credibility

- Suppose you have some prior belief about cloudy weather, $P(\text{cloudy})$
  - suppose you have some data showing it is raining outside
    - now you have a posterior belief about the weather being cloudy, given the data, $p(\text{cloudy}|\text{raining})$
    - **Q:** how does the data change your belief?
    - **A**:

$$p(\text{cloudy}) < p(\text{cloudy}|\text{raining})$$

  - suppose, instead, you have some data showing everyone outside is wearing sunglasses
    - now you have a posterior belief about the weather being cloudy, given the data, $p(\text{cloudy}|\text{sunglasses})$
    - **Q:** how does the data change your belief?
    - **A:**

$$p(\text{cloudy}) > p(\text{cloudy}|\text{sunglasses})$$

Derived from definitions of conditional probability

- **Q:** Recall, based on the definition of conditional probability, what are $p(\theta|y)$ and $p(y|\theta)$?

Derived from definitions of conditional probability

- **Q:** Recall, based on the definition of conditional probability, what are $p(\theta|y)$ and $p(y|\theta)$?
- **A:**

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} \tag{1}$$

$$p(y|\theta) = \frac{p(\theta, y)}{p(\theta)} \tag{2}$$

- We can write eq.(2) as

$$p(\theta, y) = p(y|\theta)p(\theta) \tag{3}$$

- Replace $p(\theta, y)$ in eq.(1) with that in eq.(3):

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \tag{4}$$

- This is the Bayes' rule

Derived from definitions of conditional probability

- The denominator in Bayes' rule (eq.(4)) can be factorized as

$$p(y) = \begin{cases} \sum_\theta p(y|\theta)p(\theta), & \text{when } \theta \text{ is discrete} \\ \int_\theta \mathrm{d}\theta \; p(y|\theta)p(\theta), & \text{when } \theta \text{ is continuous} \end{cases}$$

- Based on the factorization of $p(y)$, we can write $p(\theta|y)$ as

$$p(\theta|y) = \begin{cases} \frac{p(y|\theta)p(\theta)}{\sum_\theta p(y|\theta)p(\theta)}, & \text{when } \theta \text{ is discrete} \\ \frac{p(y|\theta)p(\theta)}{\int_\theta \mathrm{d}\theta \; p(y|\theta)p(\theta)}, & \text{when } \theta \text{ is continuous} \end{cases}$$

- This is how we usually use the Bayes' rule

Applied to parameters and data

- Bayes' rule can be used to estimate the parameters based on the data
- Let θ be the parameters and D the data, then Bayes' rule can be written as

$$p(\theta|D) = p(D|\theta)p(\theta)/p(D)$$

- The factors of Bayes' rule have specific names:

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

- The evidence is also called the marginal likelihood

Example: disease diagnosis

- Suppose the prior probability of having a disease (θ) is 0.001:

$$p(\theta = 1) = 0.001$$

- There is a test (y) for the disease that has a 99% hit rate, which means that if a person has the disease, then the test result is positive 99% of the time:

$$p(y = 1|\theta = 1) = 0.99$$

- The test has a false alarm rate of 5%:

$$p(y = 1|\theta = 0) = 0.05$$

- **Q:** Suppose the test result is positive. What is the probability of having the disease?

Example: disease diagnosis

- **Q:** What do we already know and what do we want to know?

Example: disease diagnosis

- **Q:** What do we already know and what do we want to know?
- **A:**
  - we already know the prior probability ($p(\theta)$) and likelihood ($p(y|\theta)$)
  - we want to know the posterior probability ($p(\theta|y)$)
- **Q:** Can you solve this problem using Bayes' rule?

Example: disease diagnosis

- **Q:** What do we already know and what do we want to know?
- **A:**
  - we already know the prior probability ($p(\theta)$) and likelihood ($p(y|\theta)$)
  - we want to know the posterior probability ($p(\theta|y)$)
- **Q:** Can you solve this problem using Bayes' rule?
- **A:**

$$
\begin{aligned}
p(\theta = 1|y = 1) &= \frac{p(y = 1|\theta = 1)p(\theta = 1)}{\sum_\theta p(y = 1|\theta)p(\theta)} \\
&= \frac{p(y = 1|\theta = 1)p(\theta = 1)}{p(y = 1|\theta = 1)p(\theta = 1) + p(y = 1|\theta = 0)p(\theta = 0)} \\
&= \frac{0.99 \times 0.001}{0.99 \times 0.001 + 0.05 \times (1 - 0.001)} \\
&= 0.019
\end{aligned}
$$

Data order invariance

- Suppose we first observe data D then D'. Bayes' rule gives us

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

- Now suppose we observe data in a reversed order: first D' then D. Bayes' rule gives us

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

- **Q:** Does our final belief depend on the order of the data? In other words, does the following equation hold?

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

Data order invariance

- Suppose we first observe data D then D'. Bayes' rule gives us

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

- Now suppose we observe data in a reversed order: first D' then D. Bayes' rule gives us

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

- **Q:** Does our final belief depend on the order of the data? In other words, does the following equation hold?

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

- **A:** Our final belief *does not* depend on the order of the data, when the data are independent:

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

Complete examples: estimating bias in a coin

- **Q:** Recall, what are the five steps in bayesian inference?

Complete examples: estimating bias in a coin

- **Q:** Recall, what are the five steps in bayesian inference?
- **A:**
  - identify the data relevant to the research
  - define a model for the data
  - specify a prior distribution on the parameters
  - infer the posterior distribution of the parameters
  - check whether the posterior distribution fits the data well
- Let us follow the steps to estimate the bias in a coin

Step 1: identify the data

- The data consist of heads and tails
- We will use $y = 1$ to denote heads, and $y = 0$ to denote tails
- Note that heads and tails are categorical, not numerical

Step 2: define a model

- In this example we use bernoulli distribution
- The probability mass function of this model is

$$p(y|\theta) = \theta^y (1-\theta)^{1-y}$$

- In other words:

$$p(y|\theta) = \begin{cases} \theta, & \text{when } y = 1 \\ 1-\theta, & \text{when } y = 0 \end{cases}$$

Step 2: define a model

- Now we extend the model from one flip to multiple flips
- **Q:** Given a sequence of N outcomess, $D = y_1, \ldots, y_N$, what is the probability mass function of having z heads?

Step 2: define a model

- Now we extend the model from one flip to multiple flips
- **Q:** Given a sequence of N outcomess, $D = y_1, \ldots, y_N$, what is the probability mass function of having z heads?
- **A:**

$$
\begin{aligned}
p(D|\theta) &= \prod_i p(y_i|\theta) \\
&= \prod_i \theta^{y_i}(1-\theta)^{(1-y_i)} \\
&= \theta^{\sum_i y_i}(1-\theta)^{\sum_i (1-y_i)} \\
&= \theta^{\#\text{heads}}(1-\theta)^{\#\text{tails}} \\
&= \theta^{z}(1-\theta)^{N-z}
\end{aligned}
$$

Step 3: specify a prior

- For this example, we assume that there are only a few discrete values of the parameter θ
- The top panel of Figure 5.1 (see next page) shows the prior distribution of θ

Figure 5.1

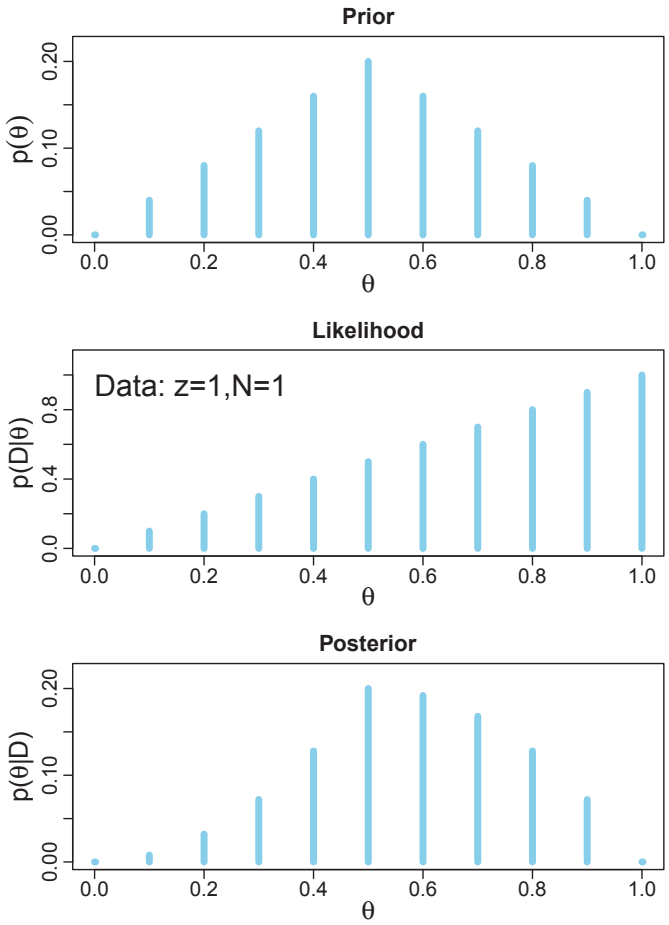


Figure 5.1: Bayes' rule applied to estimating the bias of a coin. There are discrete candidate values of θ. At each value of θ, the posterior is computed as prior times likelihood, normalized. In the data, denoted D, the number of heads is z and the number of flips is N. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Step 4: bayesian inference

- Suppose we flip the coin once and get head, that is

$$z = 1 \quad \text{with} \quad N = 1$$

- **Q:** Now we have the prior and likelihood, how can be obtain the posterior distribution?

Step 4: bayesian inference

- Suppose we flip the coin once and get head, that is

$$z = 1 \quad \text{with} \quad N = 1$$

- **Q:** Now we have the prior and likelihood, how can be obtain the posterior distribution?
- **A:**

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

- The middle and bottom panels of Figure 5.1 (see next page) show the likelihood and posterior distribution
- **Q:** Why the posterior is not symmetrical?

Figure 5.1

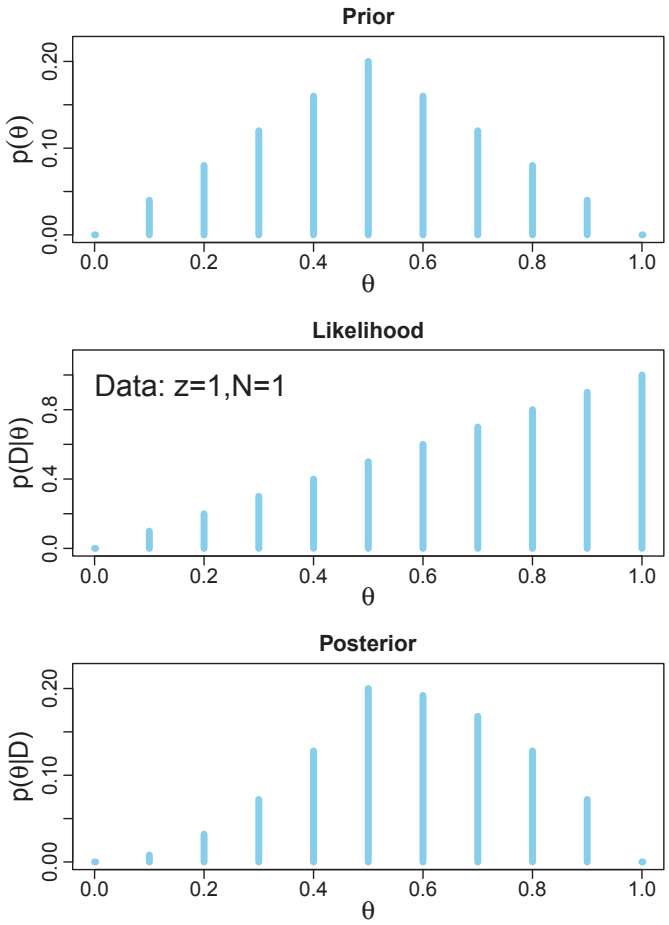


Figure 5.1: Bayes' rule applied to estimating the bias of a coin. There are discrete candidate values of θ. At each value of θ, the posterior is computed as prior times likelihood, normalized. In the data, denoted D, the number of heads is z and the number of flips is N. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Step 4: bayesian inference

- Suppose we flip the coin once and get head, that is

$$z = 1 \quad \text{with} \quad N = 1$$

- **Q:** Now we have the prior and likelihood, how can be obtain the posterior distribution?
- **A:**

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

- The middle and bottom panels of Figure 5.1 (see next page) show the likelihood and posterior distribution
- **Q:** Why the posterior is not symmetrical?
- **A:** Because the likelihood is not symmetrical

Influence of sample size on the posterior

- Figure 5.2 (see next page) shows two cases with the same prior ($p(\theta)$), the same proportion of heads (z/N), but different sample size (N):
  - in case 1, $N = 4$
  - in case 2, $N = 40$
- In the second case (with $N = 40$), the posterior is strongly influenced by the likelihood
- In the first case (with $N = 4$), the influence of the likelihood on the posterior is less strong (as shown by the residual of the prior in the posterior)
- Further, the width of HDI in the second case is much narrower than that in the first case. **Q:** What does this mean?

Influence of sample size on the posterior

- Figure 5.2 (see next page) shows two cases with the same prior ($p(\theta)$), the same proportion of heads (z/N), but different sample size (N):
  - in case 1, $N = 4$
  - in case 2, $N = 40$
- In the second case (with $N = 40$), the posterior is strongly influenced by the likelihood
- In the first case (with $N = 4$), the influence of the likelihood on the posterior is less strong (as shown by the residual of the prior in the posterior)
- Further, the width of HDI in the second case is much narrower than that in the first case. **Q:** What does this mean?
- **A:** larger sample size yields greater certainty in the posterior

Figure 5.2

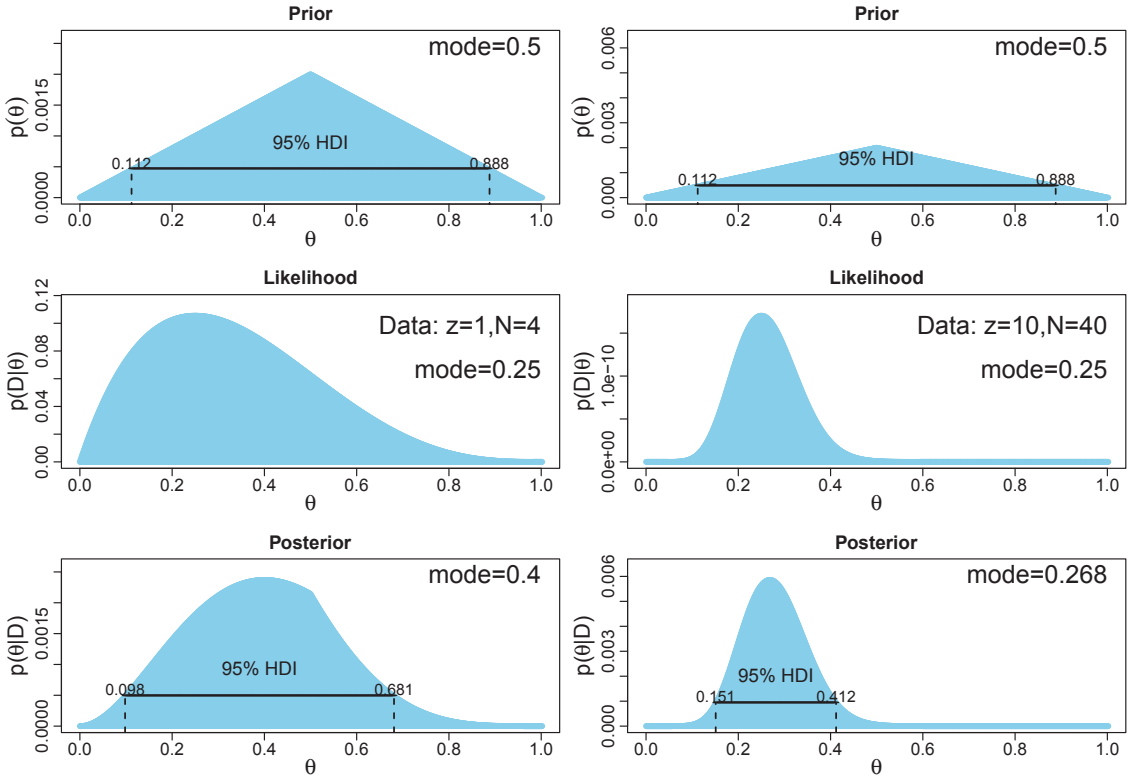


Figure 5.2: The two columns show different sample sizes with the same proportion of heads. The prior is the same in both columns but plotted on a different vertical scale. The influence of the prior is overwhelmed by larger samples, in that the peak of the posterior is closer to the peak of the likelihood function. Notice also that the posterior HDI is narrower for the larger sample. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Influence of the prior on the posterior

- The upper left panel of Figure 5.3 (see next page) shows a flat prior
- Although the sample size is small ($N = 4$), the posterior is strongly influenced by the likelihood (as shown in the middle and bottom panels)
- Conversely, the upper right panel shows a highly concentrated prior
- Although the sample size is relatively large ($N = 40$), the posterior is strongly influenced by the prior (as shown in the middle and bottom panels)

Figure 5.3

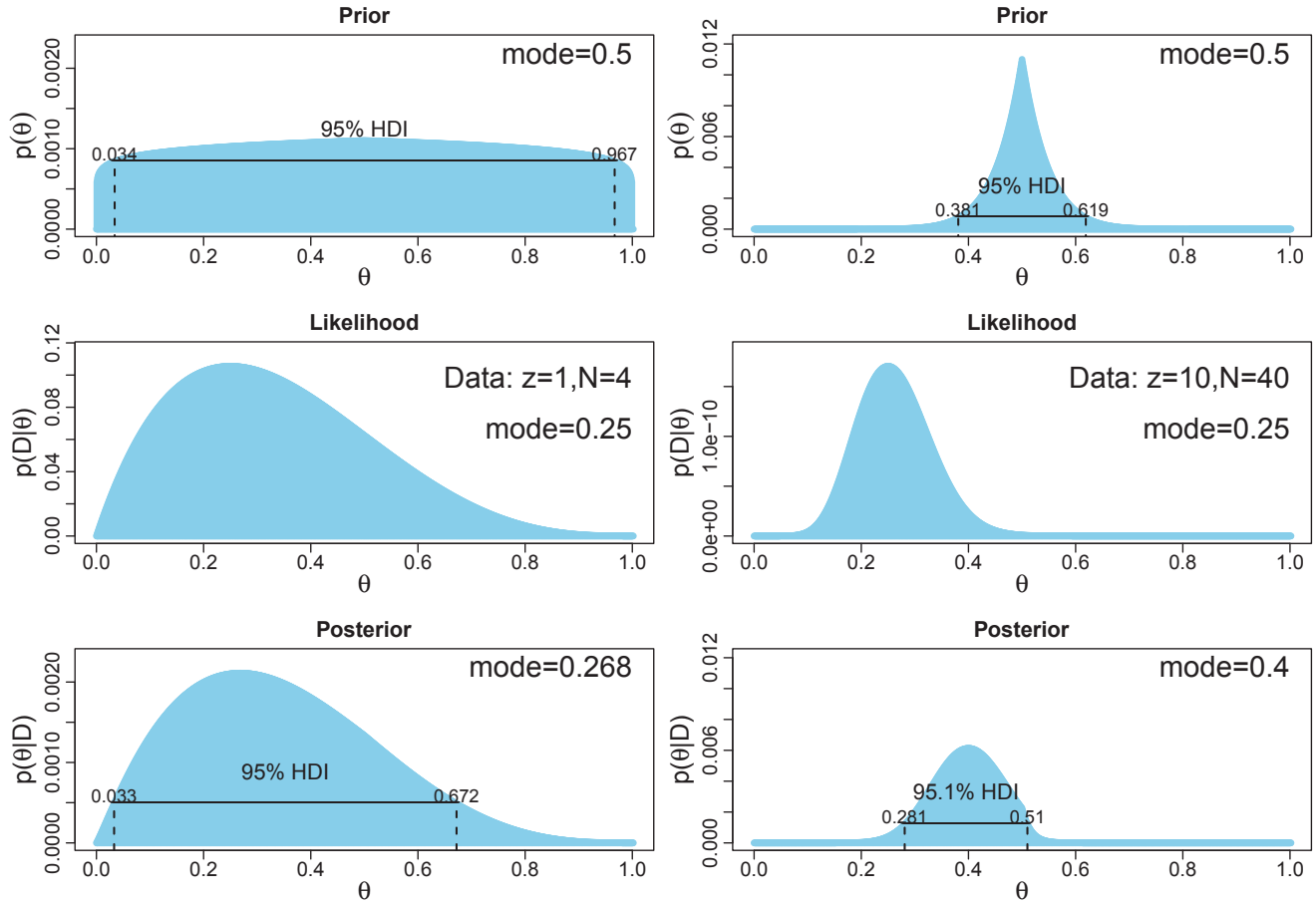


Figure 5.3: The left side is the same small sample as the left side of Figure 5.2 but with a flatter prior. The right side is the same larger sample as the right side of Figure 5.2 but with a sharper prior. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Why bayesian inference can be difficult

- Bayes' rule involves computing the evidence, $p(D)$
- It is challenging when the parameters are continuous. **Q:** Why?

Why bayesian inference can be difficult

- Bayes' rule involves computing the evidence, $p(D)$
- It is challenging when the parameters are continuous. **Q:** Why?
- **A:** Because we need to calculate the integral:

$$p(D) = \int_{\theta} \mathrm{d}\theta \; p(D|\theta)p(\theta)$$

- There are three ways to do this
  - formal analysis: impossible for some integrals
  - grid approximation: impossible with a moderately large parameter space
  - Markov Chain Monto Carlo: allowed bayesian inference to be practical