Bayesian Methods for Data Science (DATS 6450 - 11)

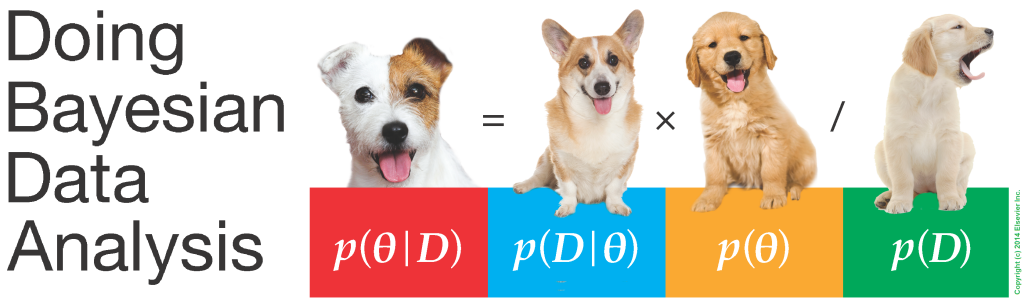
Inferring a Binomial Probability via Exact Mathematical Analysis

Yuxiao Huang

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

September 18, 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. The likelihood: bernoulli distribution
2. The prior: beta distribution
3. Examples

Bernoulli distribution

- **Q:** Recall, what is the probability mass function for the outcome of flipping a coin (i.e., the bernoulli distribution)?

Bernoulli distribution

- **Q:** Recall, what is the probability mass function for the outcome of flipping a coin (i.e., the bernoulli distribution)?
- **A:**

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

- **Q:** Recall, what is the probability mass function for a *specific* sequence of N outcomes including z heads?

Bernoulli distribution

- **Q:** Recall, what is the probability mass function for the outcome of flipping a coin (i.e., the bernoulli distribution)?
- **A:**

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

- **Q:** Recall, what is the probability mass function for a *specific* sequence of N outcomes including z heads?
- **A:**

$$p(D|\theta) = \theta^z(1-\theta)^{N-z}$$

Review of Holmes' example

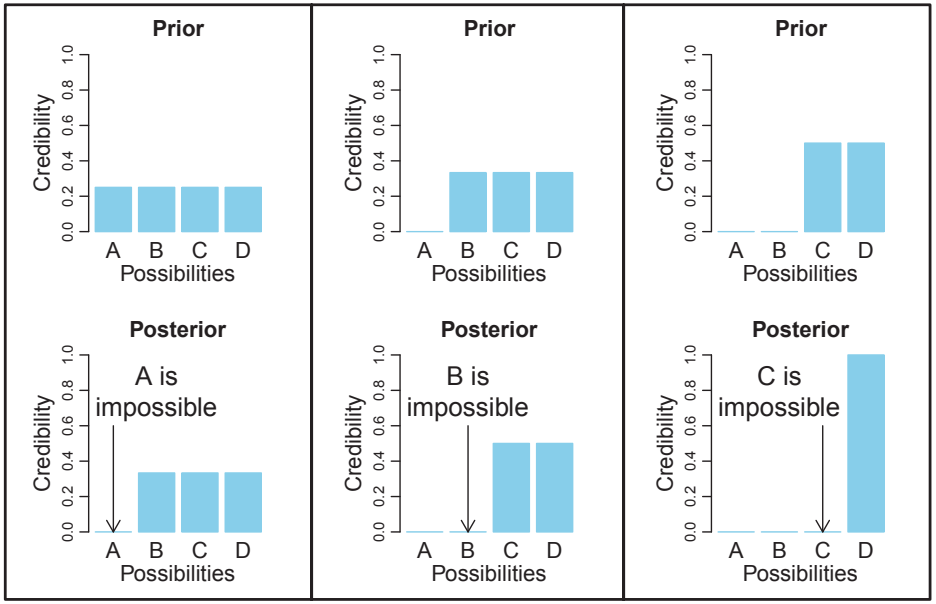


Figure 2.1: The upper-left graph shows the credibilities four possible causes for an outcome. The causes, labeled A, B, C and D, are mutually exclusive and exhaust all possibilities. The causes happen to be equally credible at the outset, hence all have prior credibility of 0.25. The lower-left graph shows the credibilities when one cause is learned to be impossible. The resulting posterior distribution is used as the prior distribution in the middle column, where another cause is learned to be impossible. The posterior distribution from the middle column is used as the prior distribution for the right column. The remaining possible cause is fully implicated by Bayesian re-allocation of credibility. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Conjugate prior

- Conjugate prior is a prior such that, the numerator of Bayes' rule, $p(D|\theta)p(\theta)$, has the same form as the prior, $p(\theta)$
- We call such prior the conjugate prior with respect to the likelihood
- **Q:** Why do we desire a conjugate prior?

Conjugate prior

- Conjugate prior is a prior such that, the numerator of Bayes' rule, $p(D|\theta)p(\theta)$, has the same form as the prior, $p(\theta)$
- We call such prior the conjugate prior with respect to the likelihood
- **Q:** Why do we desire a conjugate prior?
- **A:**
  - we need to update the posterior when new data are available
  - the new posterior has the following form

$$\text{likelihood} \times \text{old posterior}$$

  - for conjugate prior, the old posterior has the same form as the prior
  - thus the new posterior has the same form as

$$\text{likelihood} \times \text{prior},$$

  which still has the same form as the prior

Beta distribution

- The conjugate prior with respect to the bernoulli likelihood follows a beta distribution
- The probability density function of beta distribution is

$$\begin{aligned} p(\theta|a,b) &= \text{beta}(\theta|a,b) \\ &= \frac{\theta^{(a-1)}(1-\theta)^{(b-1)}}{B(a,b)} \end{aligned}$$

- Here, $B(a,b)$ is a normalizing constant (which ensures that the area under the beta density integrates to 1):

$$B(a,b) = \int_0^1 \mathrm{d}\theta \; \theta^{(a-1)}(1-\theta)^{(b-1)}$$

The parameters of beta distribution

- The a and b in the probability density function are the parameters of beta distribution
- Figure 6.1 (see next page) shows $p(\theta|a, b)$ as a function of θ for particular values of a and b:
  - when a gets bigger, the bulk of the distribution moves rightward
  - when b gets bigger, the bulk of the distribution moves leftward
  - when both a and b get bigger, the distribution gets narrower
- Here, a and b are called the shape parameters of the beta distribution

Figure 6.1

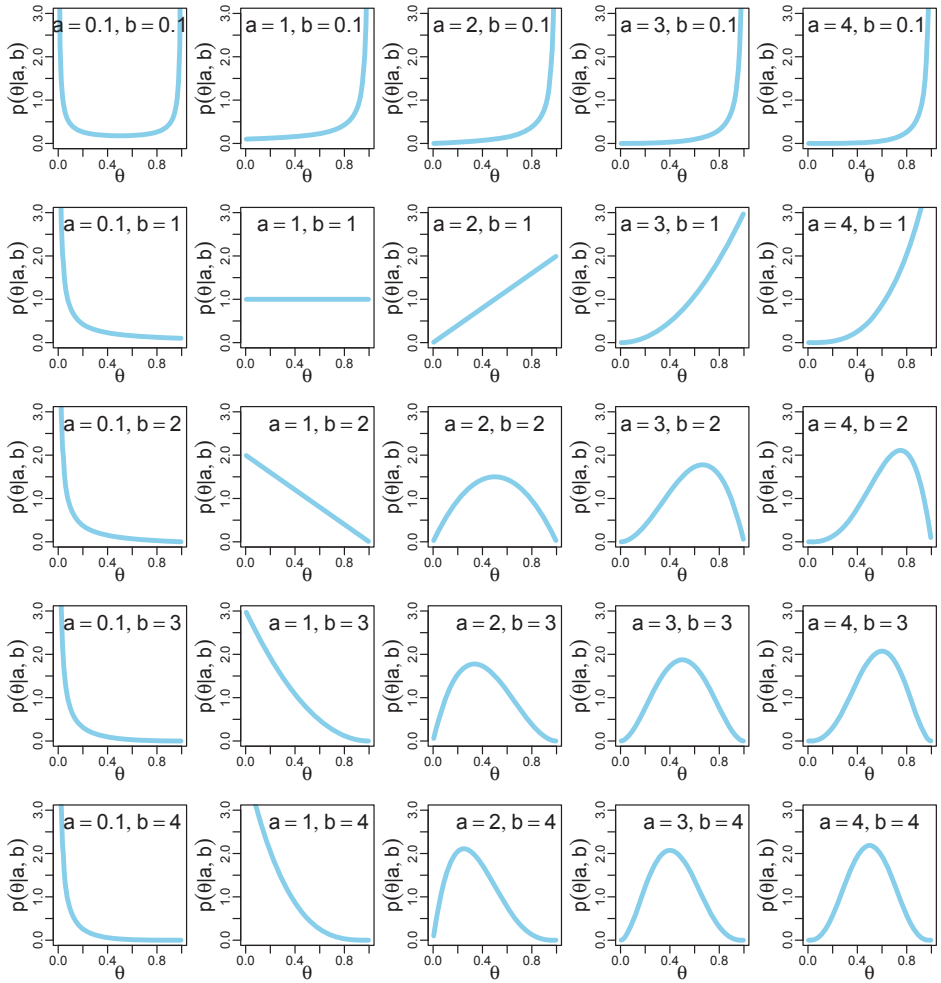


Figure 6.1: Examples of the beta distribution (Eqn. 6.1). The shape parameter a increases from left to right across the columns, while the shape parameter b increases from top to bottom across the rows. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Specifying a beta prior

- We can think of the shape of the beta distribution in terms of the central tendency and certainty about the central tendency
- Particularly, the concentration is

$$\kappa = a + b,$$

  which can be thought as the number of flips in the previously observed data
- The mean and mode of the beta distribution are

$$\mu = \frac{a}{a+b} \quad \text{and} \quad \omega = \frac{a-1}{a+b-2}$$

- As shown in Figure 6.2 (see next page), the mode can be more intuitive than the mean, especially for skewed distributions

Figure 6.2

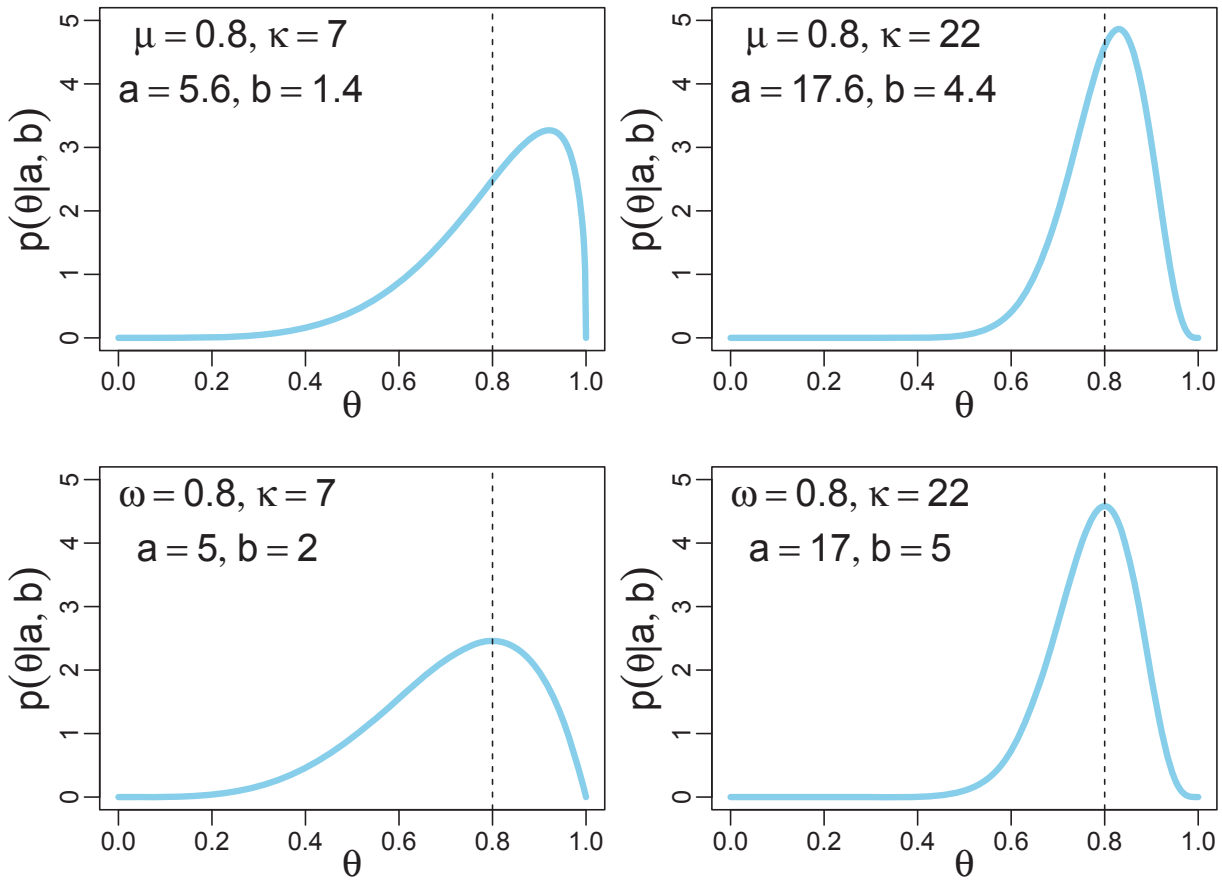


Figure 6.2: Beta distributions with a *mean* of $\mu = 0.8$ in the upper panels and a *mode* of $\omega = 0.8$ in the lower panels. Because the beta distribution is usually skewed, it can be more intuitive to think in terms of its mode instead of its mean. When κ is smaller, as in the left column, the beta distribution is wider than when κ is larger, as in the right column.

Specifying a beta prior

- Here, a and b can be determined by:
  - the mean μ and concentration κ

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

  - the mode ω and concentration κ

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

  - the mean μ and standard deviation σ

$$a = \mu\left(\frac{\mu(1-\mu)}{\sigma^2} - 1\right) \quad \text{and} \quad b = (1-\mu)\left(\frac{\mu(1-\mu)}{\sigma^2} - 1\right)$$

Specifying a beta prior

- In most applications, both a and b in the beta distribution are greater than 1
  - e.g., a fair coin
  - in such cases the concentration κ is greater than 2
  - thus it is most intuitive to use the mode to find a and b
- In some applications, either a or b is not greater than 1
  - e.g., a biased coin
  - in such cases the concentration κ may not be greater than 2
  - thus it is most intuitive to use the mean to find a and b

The posterior betas

- Suppose we have bernoulli likelihood and beta prior

$$p(z, N|\theta) = \theta^z(1-\theta)^{N-z}$$

$$p(\theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)}$$

- Based on Bayes' rule, the posterior is

$$p(\theta|z,N) = \frac{p(z,N|\theta)p(\theta)}{p(z,N)}$$

$$= \frac{\theta^z(1-\theta)^{N-z}\frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)}}{p(z,N)}$$

$$= \frac{\theta^{\big((z+a)-1\big)}(1-\theta)^{\big((N-z+b)-1\big)}}{B(z+a, N-z+b)}$$

The posterior betas

- The previous equation says that
  - if the likelihood is $\mathrm{bern}(z, N|\theta)$ and the prior is $\mathrm{beta}(\theta|a, b)$
  - then the posterior is $\mathrm{beta}(\theta|z + a, N - z + b)$
- Suppose:
  - the prior is $\mathrm{beta}(1, 1)$
  - we flip the coin once and observe head
- **Q:** What is the posterior?

The posterior betas

- The previous equation says that
  - if the likelihood is $\text{bern}(z, N|\theta)$ and the prior is $\text{beta}(\theta|a, b)$
  - then the posterior is $\text{beta}(\theta|z + a, N - z + b)$
- Suppose:
  - the prior is $\text{beta}(1, 1)$
  - we flip the coin once and observe head
- **Q:** What is the posterior?
- **A:** $\text{beta}(2, 1)$
- Now suppose:
  - we flip the coin again and observe tail
- **Q:** What is the posterior?

The posterior betas

- The previous equation says that
  - if the likelihood is $\text{bern}(z, N|\theta)$ and the prior is $\text{beta}(\theta|a, b)$
  - then the posterior is $\text{beta}(\theta|z + a, N - z + b)$
- Suppose:
  - the prior is $\text{beta}(1, 1)$
  - we flip the coin once and observe head
- **Q:** What is the posterior?
- **A:** $\text{beta}(2, 1)$
- Now suppose:
  - we flip the coin again and observe tail
- **Q:** What is the posterior?
- **A:** $\text{beta}(2, 2)$
- Figure 6.1 (see next page) shows the distributions

Figure 6.1

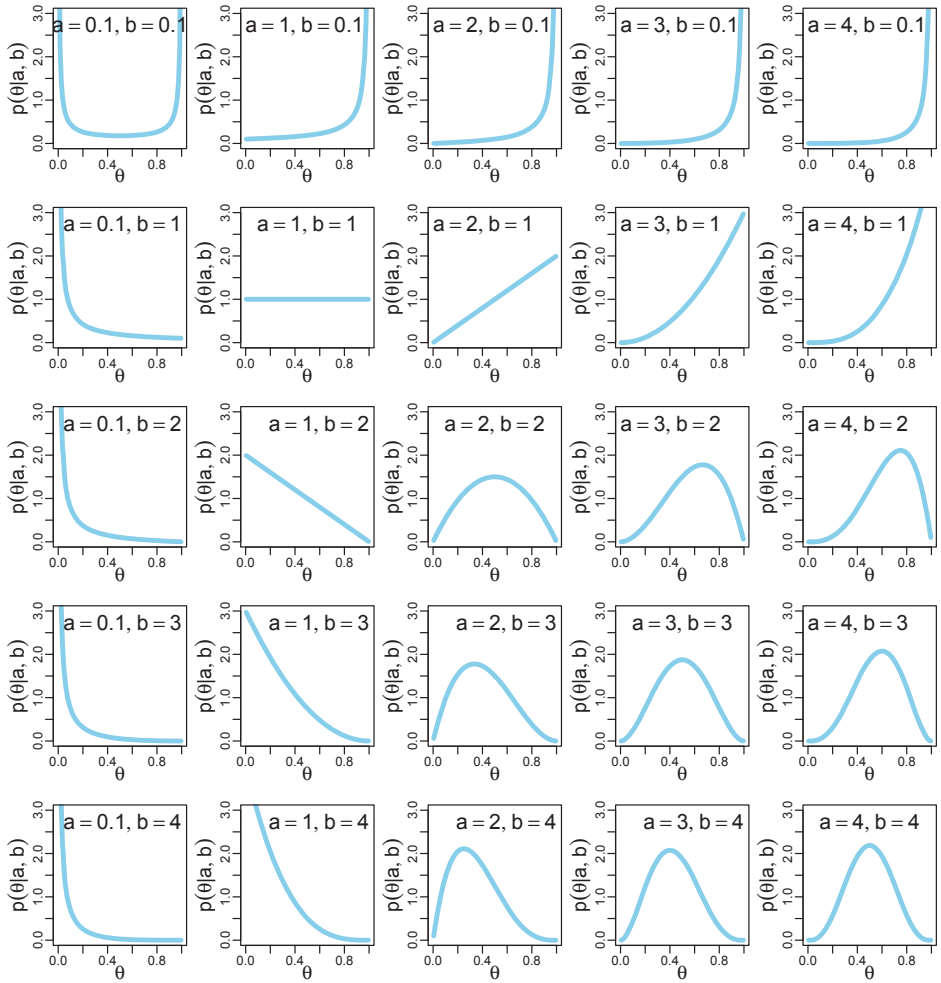


Figure 6.1: Examples of the beta distribution (Eqn. 6.1). The shape parameter a increases from left to right across the columns, while the shape parameter b increases from top to bottom across the rows. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Posterior is a compromise of prior and likelihood

- **Q:** Recall, when the likelihood is $\mathrm{bern}(z, N|\theta)$ and the prior is $\mathrm{beta}(\theta|a, b)$, what is the posterior?

Posterior is a compromise of prior and likelihood

- **Q:** Recall, when the likelihood is $\mathrm{bern}(z, N|\theta)$ and the prior is $\mathrm{beta}(\theta|a, b)$, what is the posterior?
- **A:**

$$\mathrm{beta}(\theta|z + a, N - z + b)$$

- **Q:** What is the mean of the posterior?

Posterior is a compromise of prior and likelihood

- **Q:** Recall, when the likelihood is $\mathrm{bern}(z, N|\theta)$ and the prior is $\mathrm{beta}(\theta|a, b)$, what is the posterior?
- **A:**

$$\mathrm{beta}(\theta|z + a, N - z + b)$$

- **Q:** What is the mean of the posterior?
- **A:**

$$\frac{z + a}{(z + a) + (N - z + b)} = \frac{z + a}{N + a + b}$$

Posterior is a compromise of prior and likelihood

- The posterior mean can be written as

$$\underbrace{\frac{z+a}{N+a+b}}_{\text{posterior}} = \underbrace{\frac{z}{N}}_{\text{data}} \underbrace{\frac{N}{N+a+b}}_{\text{weight}} + \underbrace{\frac{a}{a+b}}_{\text{prior}} \underbrace{\frac{a+b}{N+a+b}}_{\text{weight}}$$

- This says that the posterior mean is a weighted average of the prior mean and the data mean
  - when $N \uparrow$, data weight $\uparrow$, prior weight $\downarrow$
  - when $N \downarrow$, data weight $\downarrow$, prior weight $\uparrow$
- Figure 6.3 (see next page) shows the above decomposition of the posterior mean

Figure 6.3

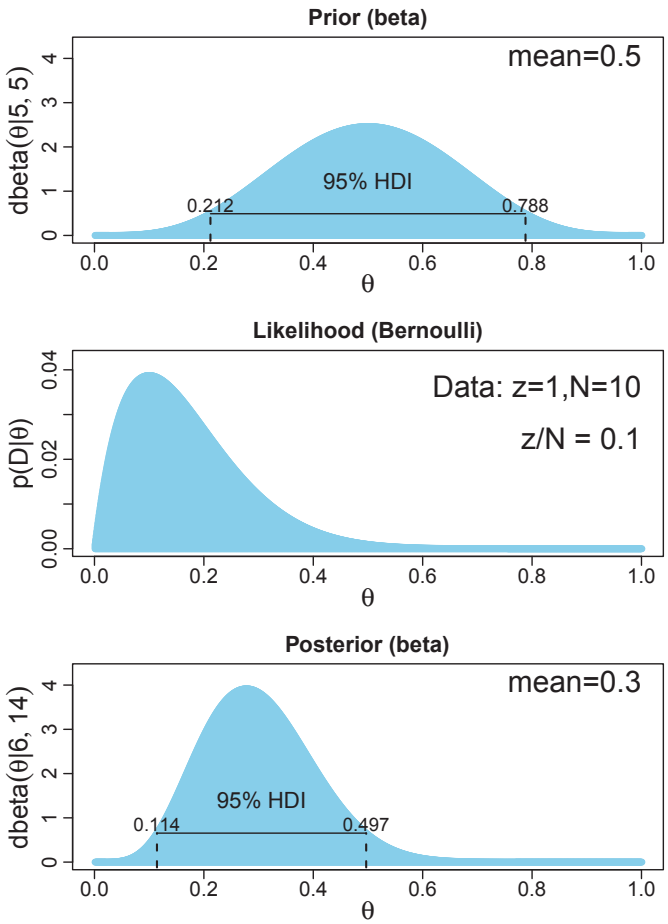


Figure 6.3: An illustration of Equation 6.9, showing that the mean of the posterior is a weighted combination of the mean of the prior and the proportion of heads in the data. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Prior knowledge expressed as a beta distribution

- Suppose we have three sets of beta priors

$$a = 250 \quad \text{and} \quad b = 250$$
$$a = 17.25 \quad \text{and} \quad b = 7.75$$
$$a = 1 \quad \text{and} \quad b = 1$$

- The top panels in Figure 6.4 (see next page) show how our confidence differs in these priors
- Suppose the bernoulli likelihood is

$$z = 17 \quad \text{and} \quad N = 20$$

- The bottom panels in the figure show how our confidence differs in the posteriors

Figure 6.4

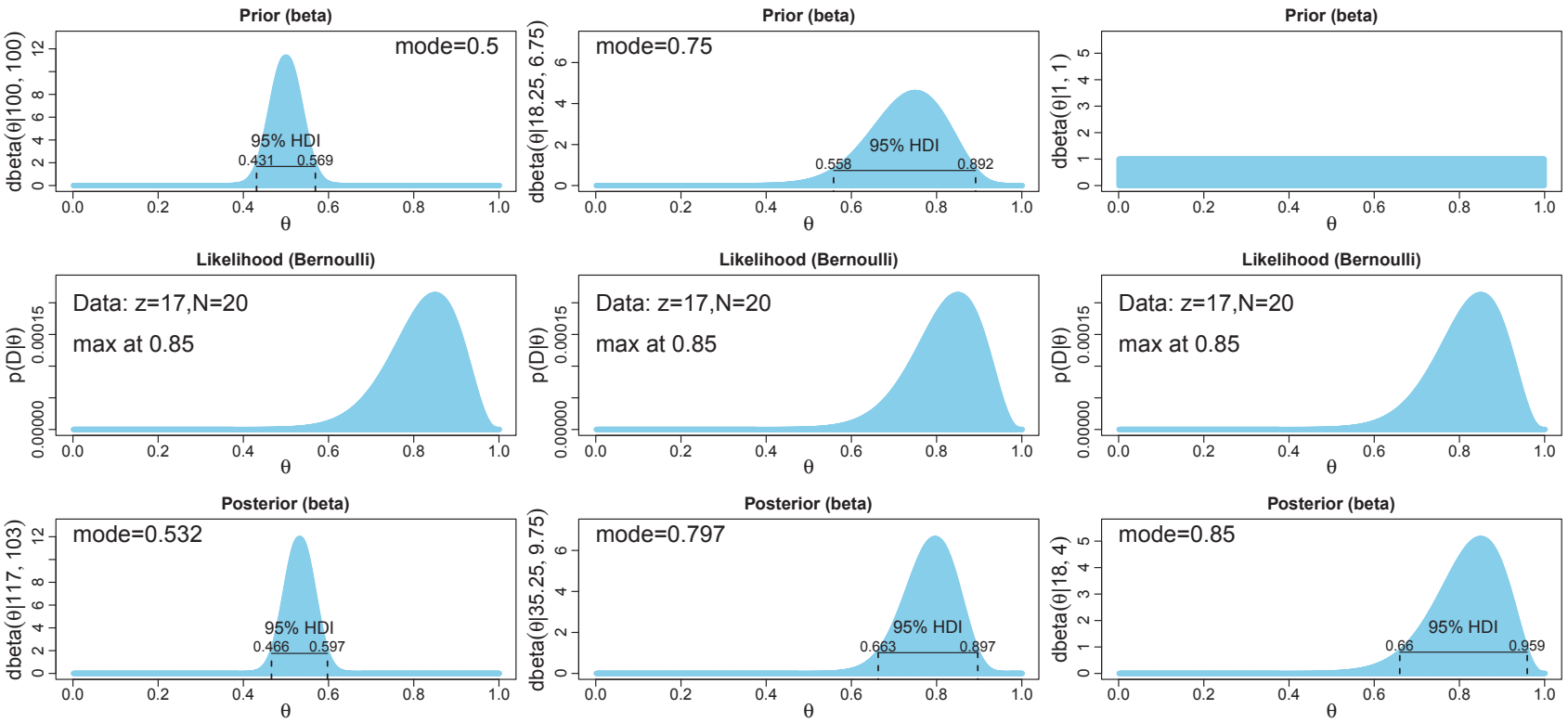


Figure 6.4: Examples of updating a beta prior distribution. The three columns show the same data with different priors. R code for this figure is described in Section 6.6. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Prior knowledge that cannot be expressed as a beta distribution

- Some priors cannot be expressed by beta distribution
- The top panel in Figure 6.5 (see next page) shows such a prior
- For such priors, we cannot use formal analysis to infer the posterior
- Instead, grid approximation (and Markov Chain Monte Carlo) can be used
- The bottom panel shows the posterior obtained by grid approximiation

Figure 6.5

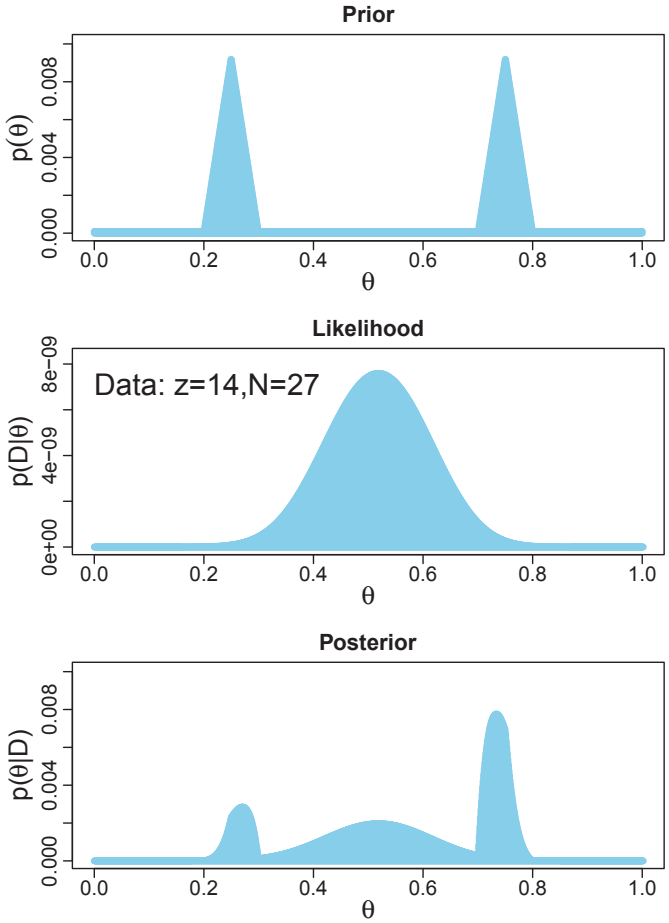


Figure 6.5: An example for which the prior distribution cannot be expressed by a beta distribution. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.