Bayesian Methods for Data Science (DATS 6450 - 11)

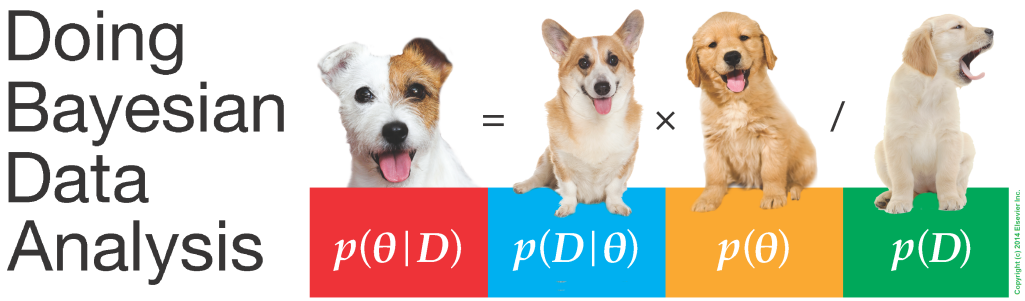
Hierarchical Models

Yuxiao Huang

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

October 23, 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

What are hierarchical models

- Consider a coin minted from a factory
- We denote the bias of the factory by ω, and the bias of the coin by θ
- Based on Bayes' rule, for any model the following condition holds:

$$p(\theta, \omega|D) \propto p(D|\theta, \omega)p(\theta, \omega)$$

- What is special to hierarchical models is that the terms on the right-hand side can be factored into a chain of dependencies:

$$\begin{aligned} p(\theta, \omega|D) &\propto p(D|\theta, \omega)p(\theta, \omega) \\ &= p(D|\theta)p(\theta|\omega)p(\omega) \end{aligned}$$

- The refactoring in the second line means that:
  - D is conditionally independent of ω, given θ
  - θ depends on ω

Why hierarchical models

- The idea is that the estimate of each individual parameter is simultaneously informed by data from all the other individuals
  - this is because all the individuals inform the higher-level parameters, which in turn constrain all the individual parameters
- Consider several coins minted from the same factory
  - if some coins are head-biased, then the factory could be head-biased, then the remaining coins could be head-biased:

$$\text{some coins} \rightarrow \text{factory} \rightarrow \text{the other coins}$$

- Besides, algorithms such as Gibbs sampling can take advantage of the dependencies in hierarchical models

A single coin from a single mint

- Consider flipping a single coin
- The likelihood is a bernoulli distribution

$$y_i \sim \text{dbern}(\theta)$$

- The prior is a beta distribution

$$\theta \sim \text{dbeta}(a, b)$$

- When representing the shape parameters, a and b, using the mode ω and concentration κ:

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

the prior can be written as

$$\theta \sim \text{dbeta}(\omega(\kappa - 2) + 1, (1 - \omega)(\kappa - 2) + 1)$$

Extend to hierarchical models

- Previously we treated both ω and κ as constant
- Now we still treat κ as a constant, denoted by K
- However:
  - we treat ω as another parameter to be estimated
  - we assume that the prior distribution of ω is a beta distribution

$$p(\omega) \sim \mathrm{dbeta}(A_\omega, B_\omega),$$

  where A_w and B_w are constants
  - figure 9.1 (see next page) shows the hierarchical model

Figure 9.1

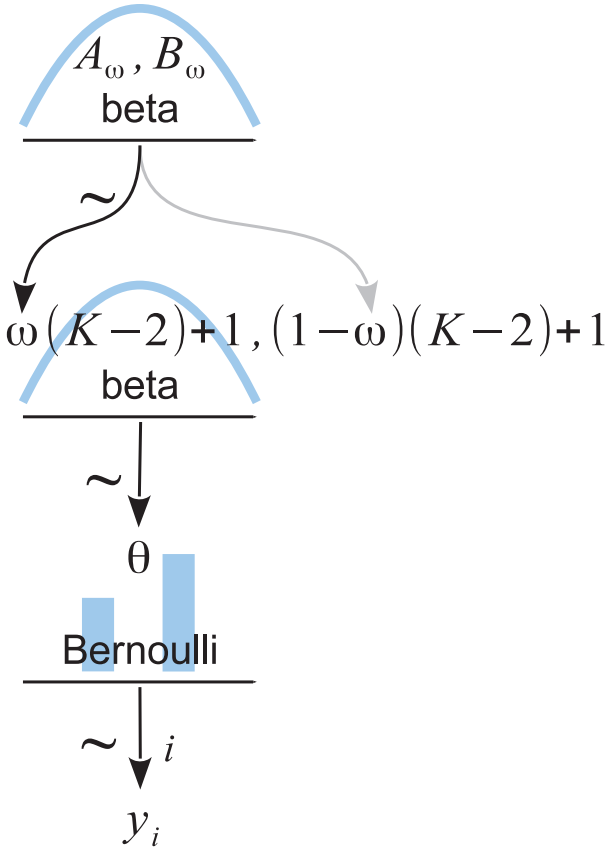


Figure 9.1: A model of hierarchical dependencies for data from a single coin. The chain of arrows illustrates the chain of dependencies in Equations 9.2, 9.4, and 9.5. (At the top of the diagram, the second instantiation of the arrow to ω is shown in grey instead of black merely to suggest that it is the same dependence already indicated by the first arrow in black.) Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Apply Bayes' rule to hierarchical models

- If we treat this situation as simply a case of two parameters, then Bayes' rule is merely

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

- However, when applying Bayes' rule to hierarchical model, we have

$$\begin{aligned} p(\theta, \omega|y) &= \frac{p(y|\theta, \omega)p(\theta, \omega)}{p(y)} \\ &= \frac{p(y|\theta)p(\theta|\omega)p(\omega)}{p(y)}, \end{aligned}$$

where the three distributions in the numerator are bernoulli, beta, and beta distributions, respectively

Posterior via grid approximation

- We cannot solve the previous posterior using formal analysis (since the form of the posterior is unknown)
- **Q:** What can we do?

Posterior via grid approximation

- We cannot solve the previous posterior using formal analysis (since the form of the posterior is unknown)
- **Q:** What can we do?
- **A:**
  - grid approximation: when there are limited number of domains
  - MCMC: otherwise
- Since here we only have parameters θ and ω that both have finite domains (interval $[0, 1]$), grid approximation is tractable
- Figure 9.2 (see next page) shows an example

Figure 9.2

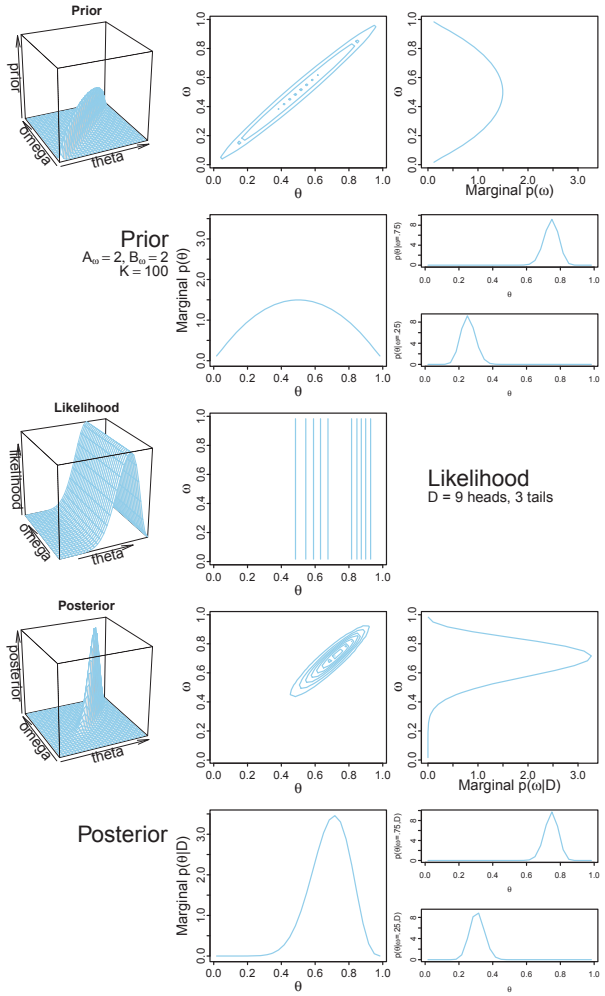


Figure 9.2: The prior has low certainty regarding ω, but high certainty regarding the dependence of θ on ω. The posterior shows that the distribution of ω has been altered noticeably by the data (see sideways plots of marginal $p(\omega)$), but the dependence of θ on ω has not been altered much (see small plots of $p(\theta|\omega)$). Compare with Figure 9.3, which uses the same data but a different prior. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Multiple coins from a single mint

- Assume that we have more than one coin created by the mint
- If each coin s has its own bias θ_s, then we use the data of coin s to estimate θ_s, and the data of all the coins to estimate ω
- The bias of coin s, θ_s, is distributed as

$$\theta_s \sim \mathrm{dbeta}(\omega(\kappa-2)+1, (1-\omega)(\kappa-2)+1)$$

- The data of coin s depend only on its bias θ_s, and is distributed as

$$y_{i|s} \sim \mathrm{dbern}(\theta_s),$$

where $y_{i|s}$ is the i^{th} datum of coin s

- The scenario is summarized in Figure 9.4 (see next page)

Figure 9.4

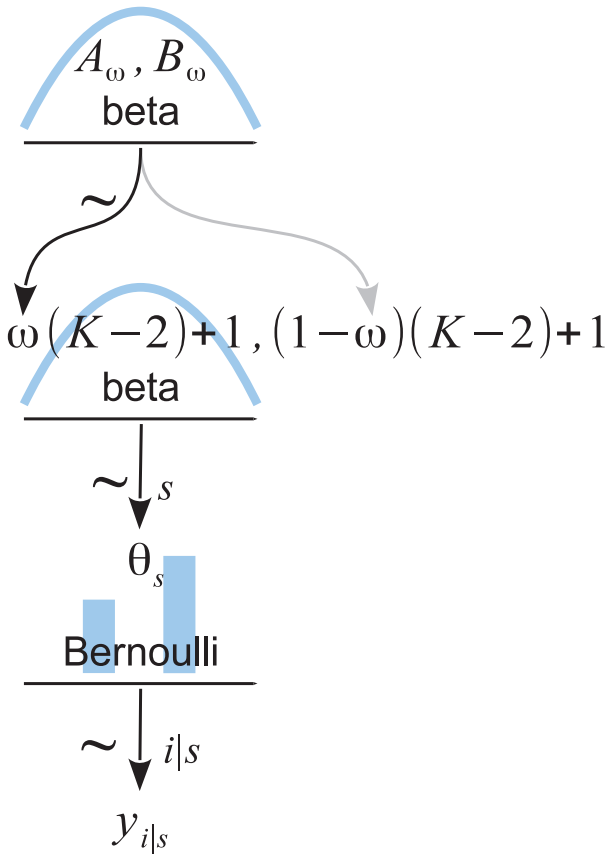


Figure 9.4: A model of hierarchical dependencies for data from several coins created independently from the same mint. A datum $y_{i|s}$, from the i^{th} flip of the s^{th} coin, depends on the value of the bias parameter θ_s for the coin. The values of θ_s depend on the value of the hyperparameter ω for the mint that created the coins. The ω parameter has a prior belief distributed as a beta distribution with shape parameters A_ω and B_ω.

Posterior via grid approximation

- Figures 9.5 and 9.6 (see next two pages) show grid approximations for two different choices of prior distributions
- The grid approximation displayed in the two figures used combs of only 50 points on each parameter (ω, θ_1, and θ_2)
- So the 3D grid has 50^3 points

Figure 9.5

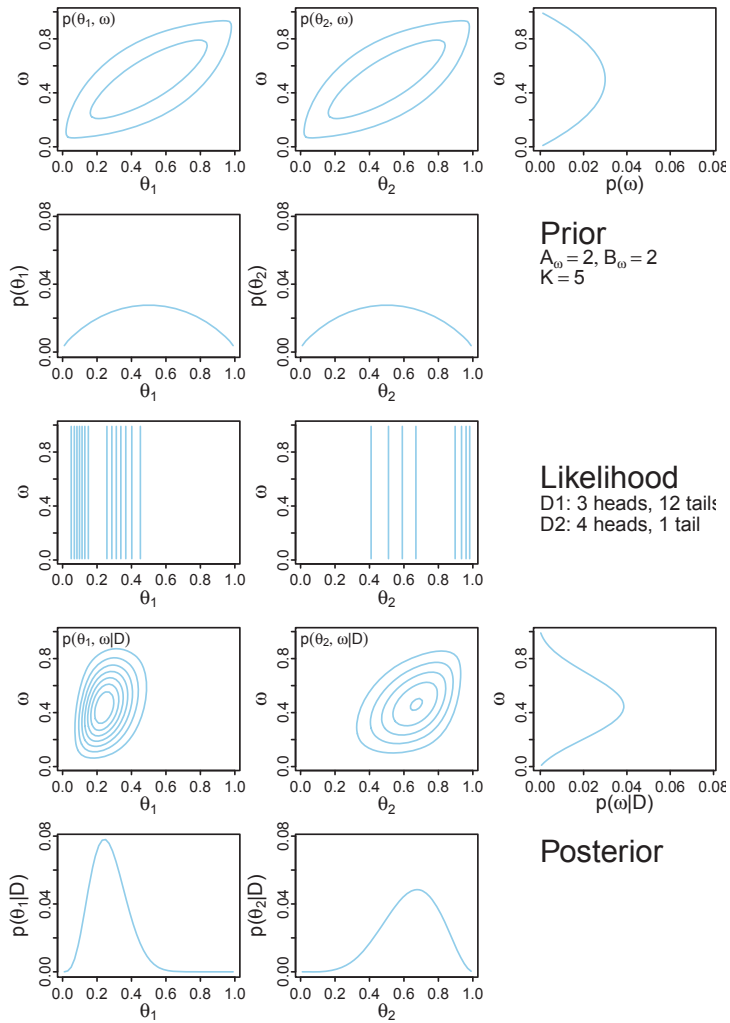


Figure 9.5: The prior imposes only a weak dependence of θ on μ (i.e., K is small), so the posteriors on θ_1 and θ_2 (bottom row) are only weakly influenced by each other's data. Compare with Figure 9.6, which uses the same data but a prior that has a strong dependence. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Figure 9.6

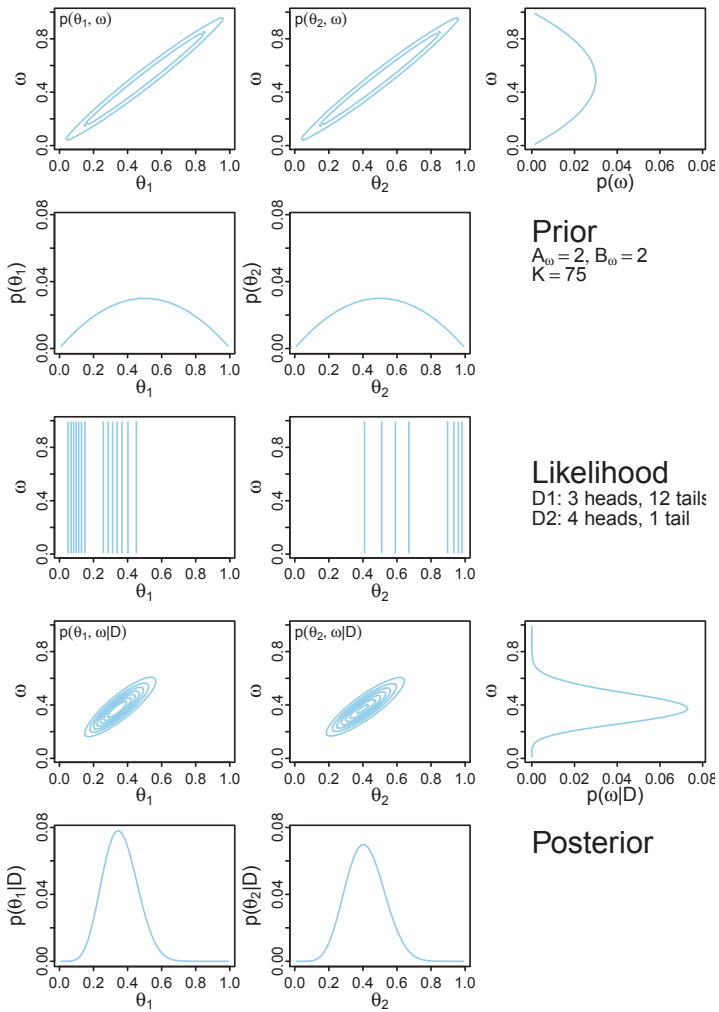


Figure 9.6: The prior imposes a strong dependency of θ on μ (i.e., K is large), so the posteriors on θ_1 and θ_2 (bottom row) are strongly influenced by each other's data, with θ_2 being pulled toward θ_1 because $N_1 > N_2$. Compare with Figure 9.5, which uses the same data but a prior that has a weak dependence.

A realistic model with MCMC

- In previous examples, we assume that the concentration parameter, κ, is constant, K
- Since the assumption does not hold in reality, here we treat κ as another parameter
- Because the value of $\kappa - 2$ must be non-negative, the prior distribution on $\kappa - 2$ must not allow negative values
- In this example, we assume $\kappa - 2$ is distributed as a gamma distribution
- The new hierarchical model is shown in Figure 9.7 (see next page)

Figure 9.7

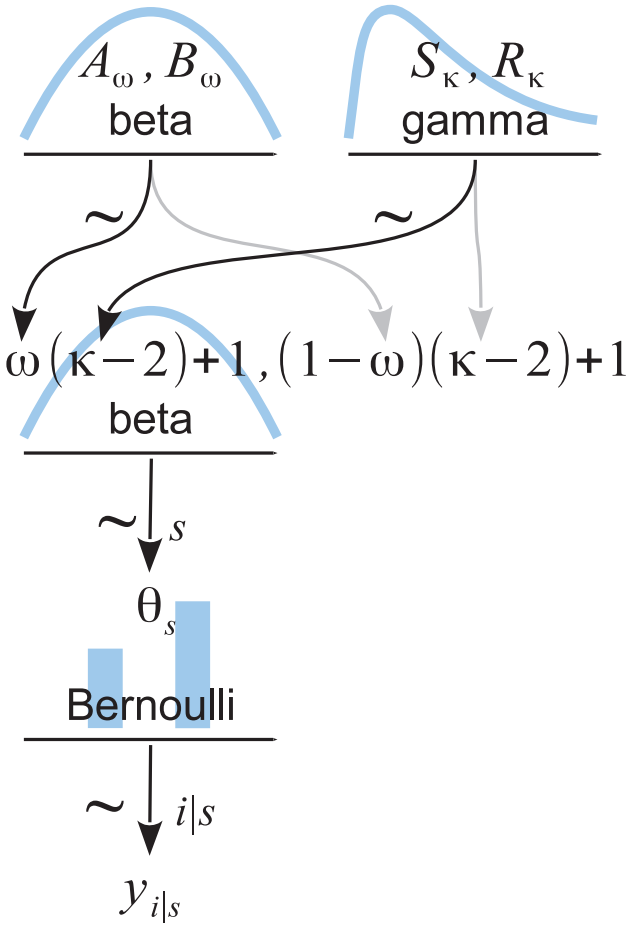


Figure 9.7: A model of hierarchical dependencies for data from several coins created independently from the same mint, with the characteristics of the mint parameterized by its mode ω and concentration κ. The value of $\kappa - 2$ has a prior distributed as a gamma density with shape and rate parameters of S_κ and R_κ. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Gamma distribution

- The $\text{gamma}(\kappa|s,r)$ distribution is a probability density for $\kappa \geq 0$, with two parameters that determine its exact form, called its shape s and rate r
- Figure 9.8 (see next page) shows examples of the gamma distribution with different values of the shape and rate parameters
- As we discussed previously, prior beliefs are most intuitively expressed in terms of the central tendency and a width of the distribution

Figure 9.8

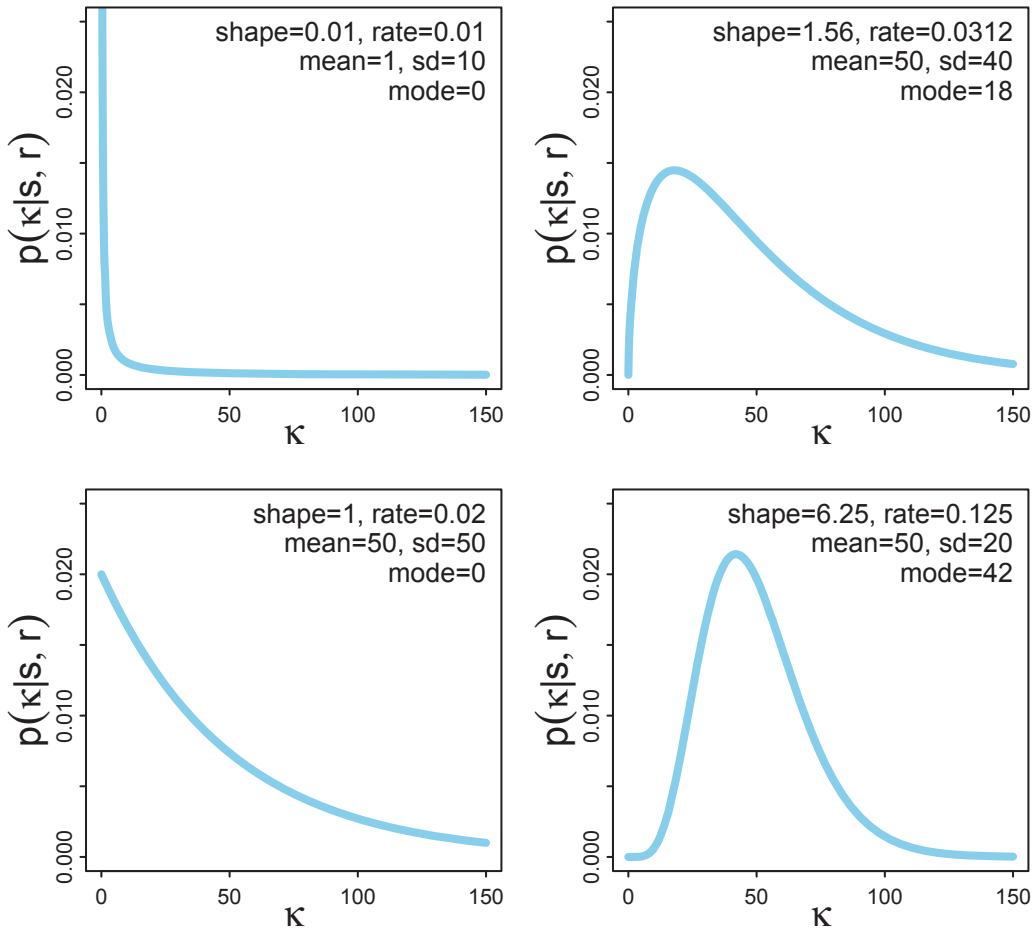


Figure 9.8: Examples of the gamma distribution. The vertical axis is $p(\kappa|s,r)$ where s is the shape and r is the rate, whose values are annotated in each panel.

Gamma distribution

- For gamma distribution:
  - the mean is:
$$\mu = \frac{s}{r}$$
  - the mode is:
$$\omega = \frac{s-1}{r} \quad \text{for } s > 1$$
  - the standard deviation is:
$$\sigma = \frac{\sqrt{s}}{r}$$
- The shape and rate can be represented using μ and σ, or ω and σ:

$$s = \frac{\mu^2}{\sigma^2} \quad \text{and} \quad r = \frac{\mu}{\sigma^2} \quad \text{for } \mu > 0$$

$$s = 1 + \omega r \quad \text{and} \quad r = \frac{\omega + \sqrt{\omega^2 + 4\sigma^2}}{2\sigma^2} \quad \text{for } \omega > 0$$

Example: therapeutic touch

- Rosa et al. (1998) investigated a key claim of practitioners of therapeutic touch, by checking whether practitioners can sense which of their hands is near another person's hand
- Figure 9.7 (see previous page) shows a hierarchical model for this example
- **Q:** What does the model mean?

Figure 9.7

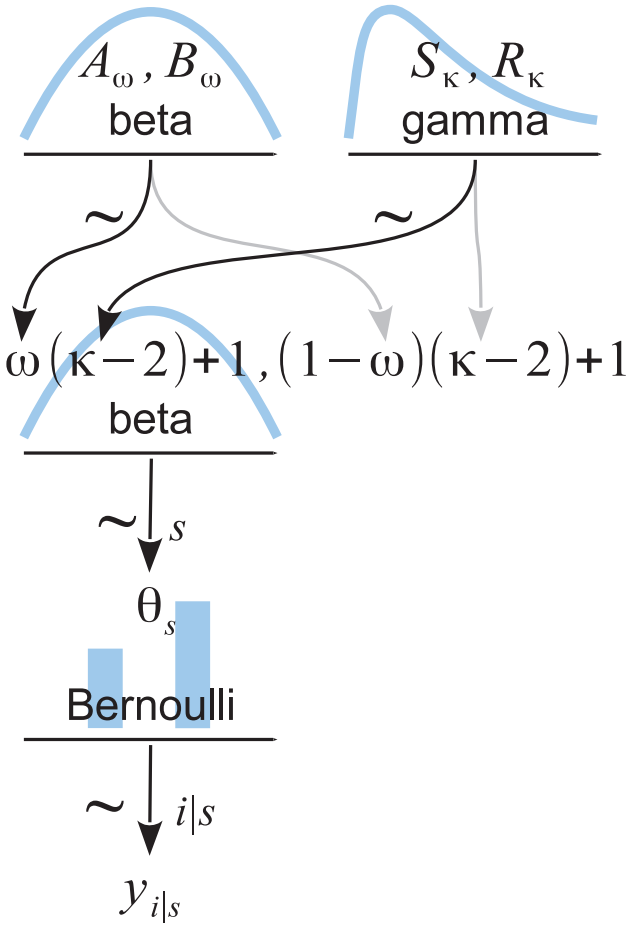


Figure 9.7: A model of hierarchical dependencies for data from several coins created independently from the same mint, with the characteristics of the mint parameterized by its mode ω and concentration κ. The value of $\kappa - 2$ has a prior distributed as a gamma density with shape and rate parameters of S_κ and R_κ. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Example: therapeutic touch

- Rosa et al. (1998) investigated a key claim of practitioners of therapeutic touch, by checking whether practitioners can sense which of their hands is near another person's hand
- Figure 9.7 (see previous page) shows a hierarchical model for this example
- **Q:** What does the model mean?
- **A:**
  - the result (success/failure) of each practitioner, $y_{i|s}$, is distributed as a bernoulli distribution with the practitioner's ability, θ_s, as the parameter
  - θ_s is distributed as a beta distribution with central tendency and concentration, ω and κ, as the parameters
  - ω and κ are distributed as beta and gamma distributions
- See `Jags-Ydich-XnomSsubj-MbernBetaOmegaKappa-Example.R` for details (available on the book website)
- The results are shown in Figures 9.10 and 9.11 (see next two pages)

Figure 9.10

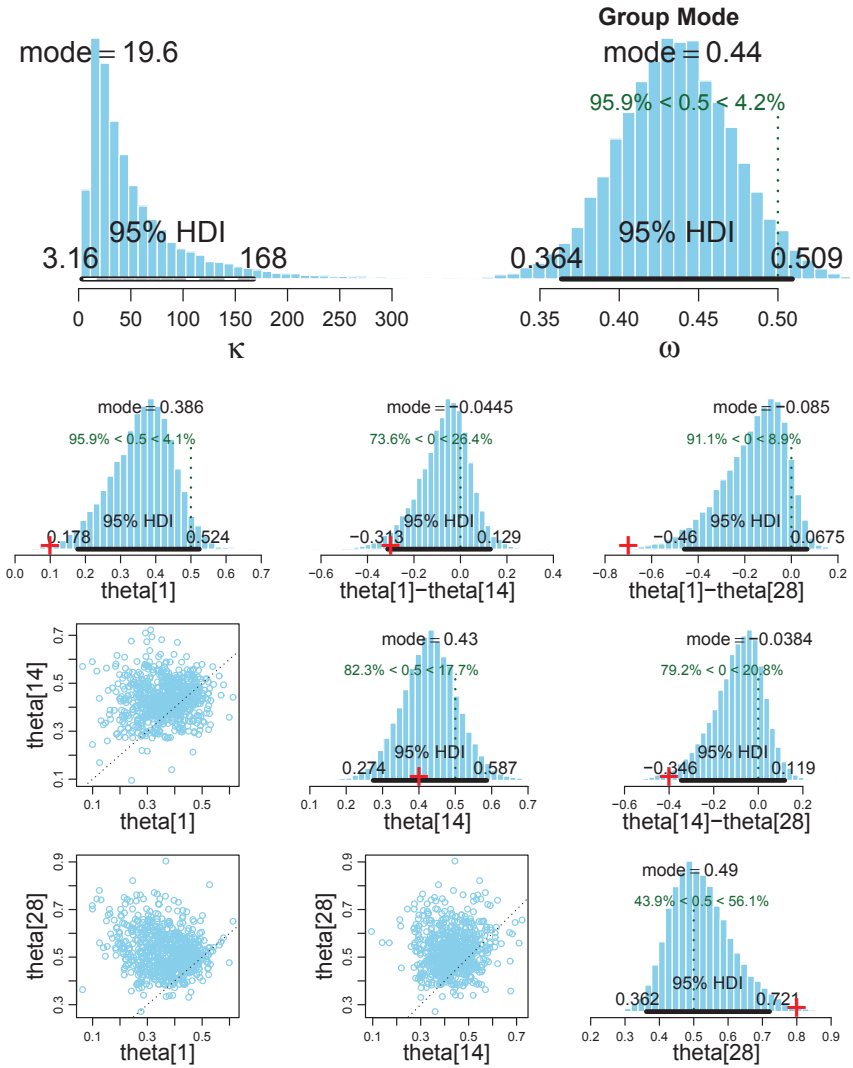


Figure 9.10: Marginal posterior distributions for the therapeutic touch data. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Figure 9.11

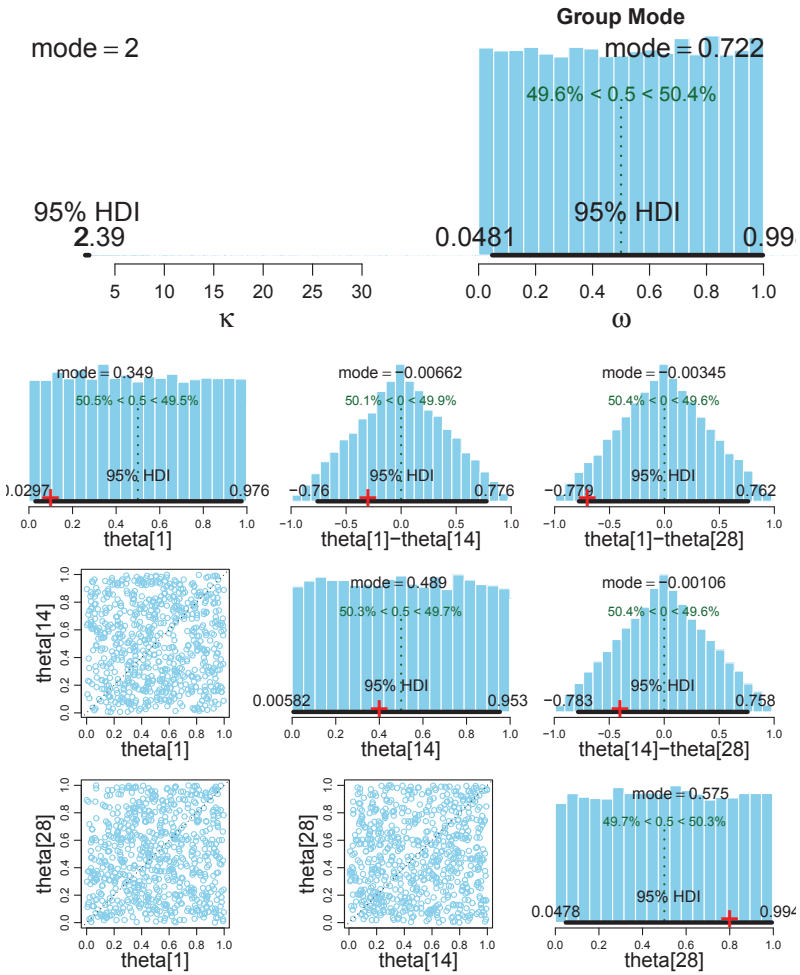


Figure 9.11: Marginal prior distributions for the therapeutic touch data. The upper-left panel, showing κ, does not plot well because it is a tall narrow peak near 2, with a long short tail extending far right. The estimated modal values of uniform distributions should be disregarded, as they are merely marking whatever random ripple happens to be a little higher than the other random ripples. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Speeding up JAGS

- **Q:** Recall, what can we do to speed up processing of JAGS?

Speeding up JAGS

- **Q:** Recall, what can we do to speed up processing of JAGS?
- **A:** Run the chains in parallel with the `runjags` package
- Another way is to replace evaluating a bernoulli distribution multiple times, with evaluating a binomial likelihood just once

$$p(z_s|\theta_s, N_s) = \binom{N_s}{z_s} \theta_s^{z_s} (1-\theta_s)^{(N_s - z_s)}$$

- See `Jags-Ydich-XnomSsubj-MbinomBetaOmegaKappa.R` for details (available on the book website)

Figure 9.10

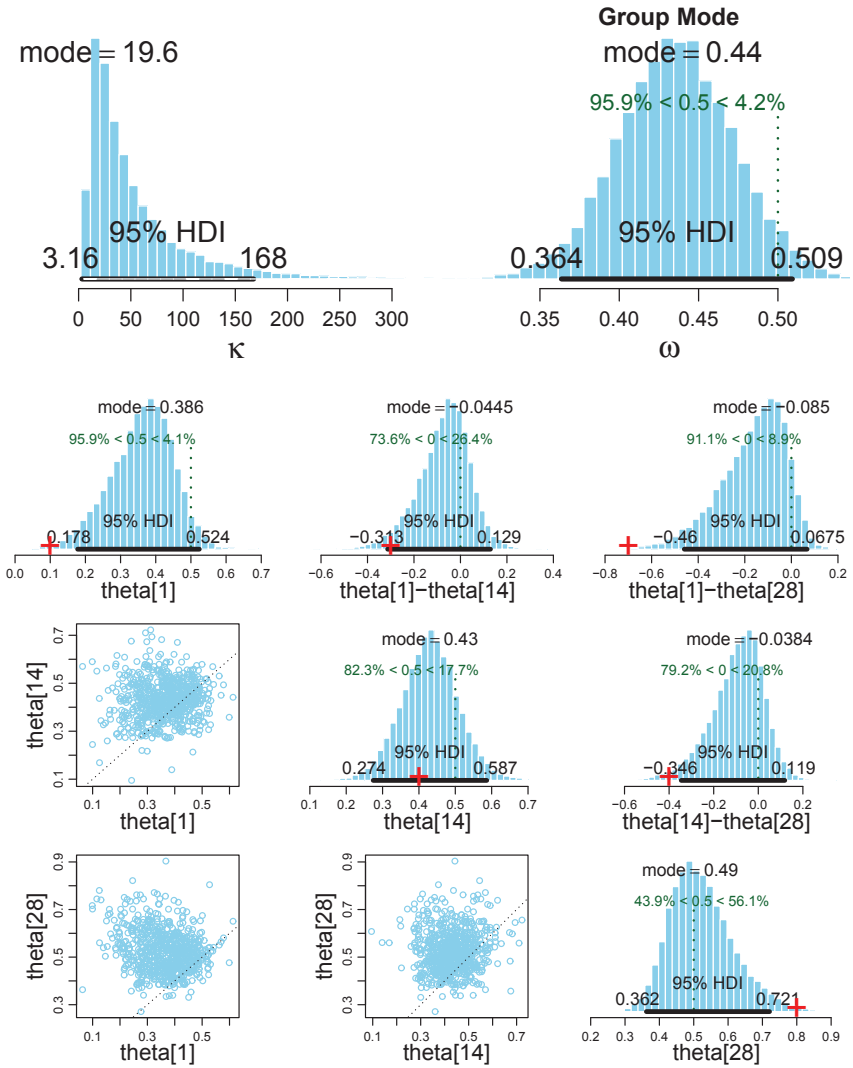


Figure 9.10: Marginal posterior distributions for the therapeutic touch data. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Shrinkage in hierarchical models

- In typical hierarchical models, the estimates of low-level parameters are pulled toward the modes of the higher-level distribution
  - when the higher-level distribution is unimodel, the low-level parameters are closer than they would be if there were not a higher-level distribution
  - when the higher-level distribution is multimodel, the parameters are further
- This phenomenon is called *shrinkage* of the estimates
- In Figure 9.10 (see previous page), the most credible values of individual-level biases, θ_s, are closer to the group-level mode, ω, than the individual proportions, $\frac{z_s}{N_s}$

Shrinkage in hierarchical models

- **Q:** Why does shrinkage happen?

Shrinkage in hierarchical models

- **Q:** Why does shrinkage happen?
- **A:**
  - it is because the estimate of each low-level parameter is influenced by two sources:
    - the subset of data that directly depend on the low-level parameter
    - the higher-level parameters on which the low-level parameter depends
  - the higher-level parameters are affected by all the data, and therefore the estimate of a low-level parameter is affected indirectly by all the data, via their influence on the higher-level parameters
- Shrinkage is (usually) desirable because the shrunken parameter estimates are less affected by random sampling noise than estimates derived without hierarchical structure

Example: Baseball batting abilities by position

- Consider professional baseball players who have different fielding positions (e.g., pitcher, catcher, and first base), with different specialized skills
- The goal is to estimate batting abilities for individual players, and for positions, and for the overarching group of professional players
- **Q:** What kind of hierarchical model you can think of?

Example: Baseball batting abilities by position

- Consider professional baseball players who have different fielding positions (e.g., pitcher, catcher, and first base), with different specialized skills
- The goal is to estimate batting abilities for individual players, and for positions, and for the overarching group of professional players
- **Q:** What kind of hierarchical model you can think of?
- **A:** One hierarchical model is shown in Figure 9.13 (see next page)
- For details see the following code (available on the book website):
  - `Jags-Ybinom-XnomSsubjCcat-MbinomBetaOmegaKappa.R`
  - `Jags-Ybinom-XnomSsubjCcat-MbinomBetaOmegaKappa-Example.R`
- The results are shown in Figures 9.14 to 9.17 (see the next 4 pages)

Figure 9.13

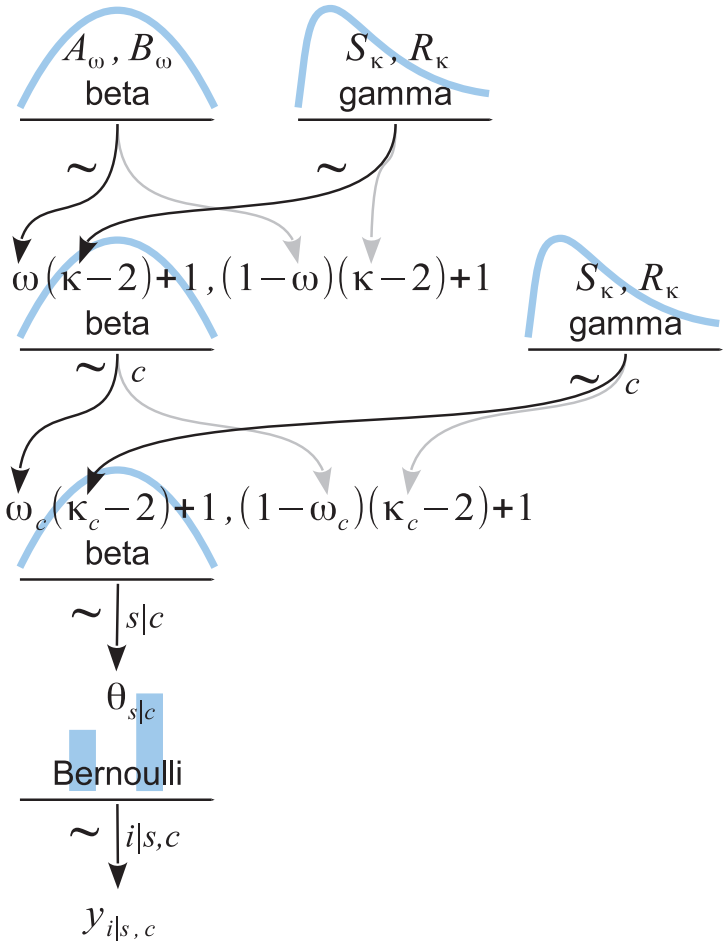


Figure 9.13: A model of hierarchical dependencies for data from several coins (indexed by subscript s) created by more than one category of mint (indexed by subscript c), with an overarching distribution across categories. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Figure 9.14

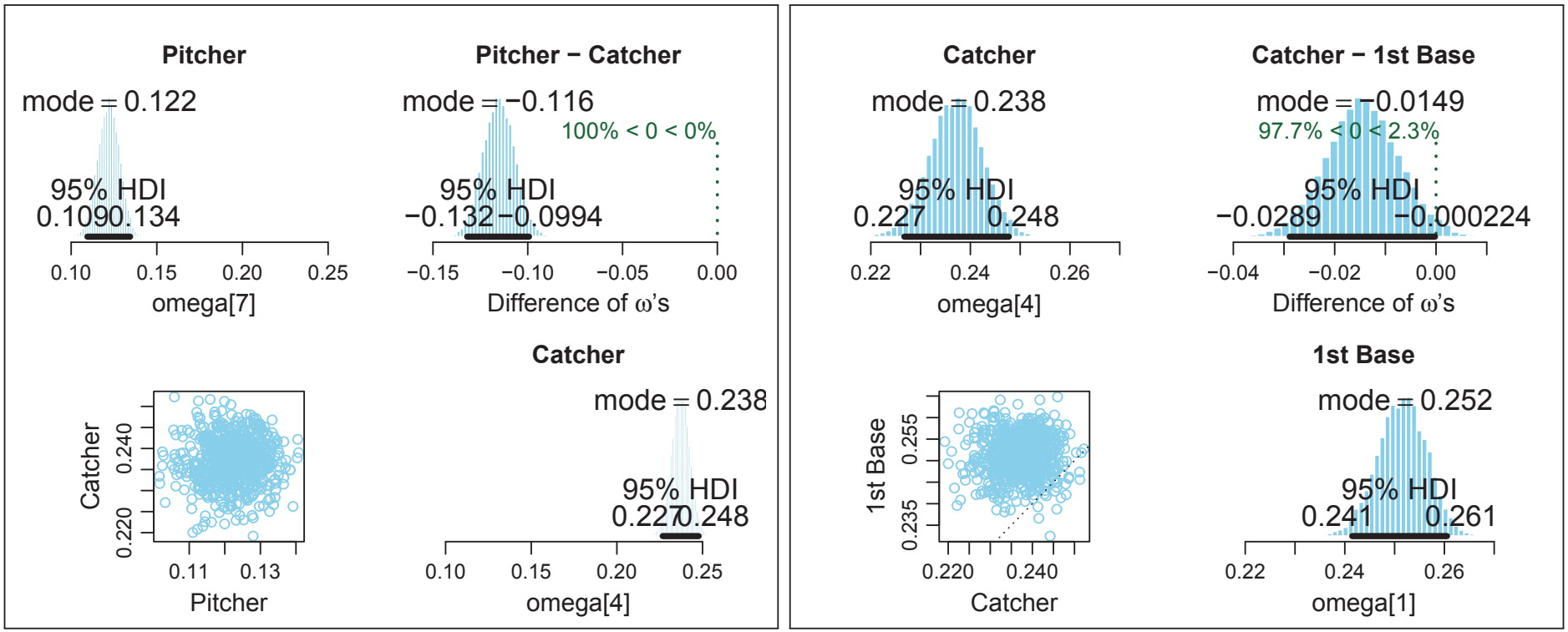


Figure 9.14: Marginal posterior distributions for baseball batting data. Left quartet shows that the pitchers have far lower batter abilities than the catchers. Right quartet shows that the catchers have marginally lower batting abilities than 1st-base men. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Figure 9.15

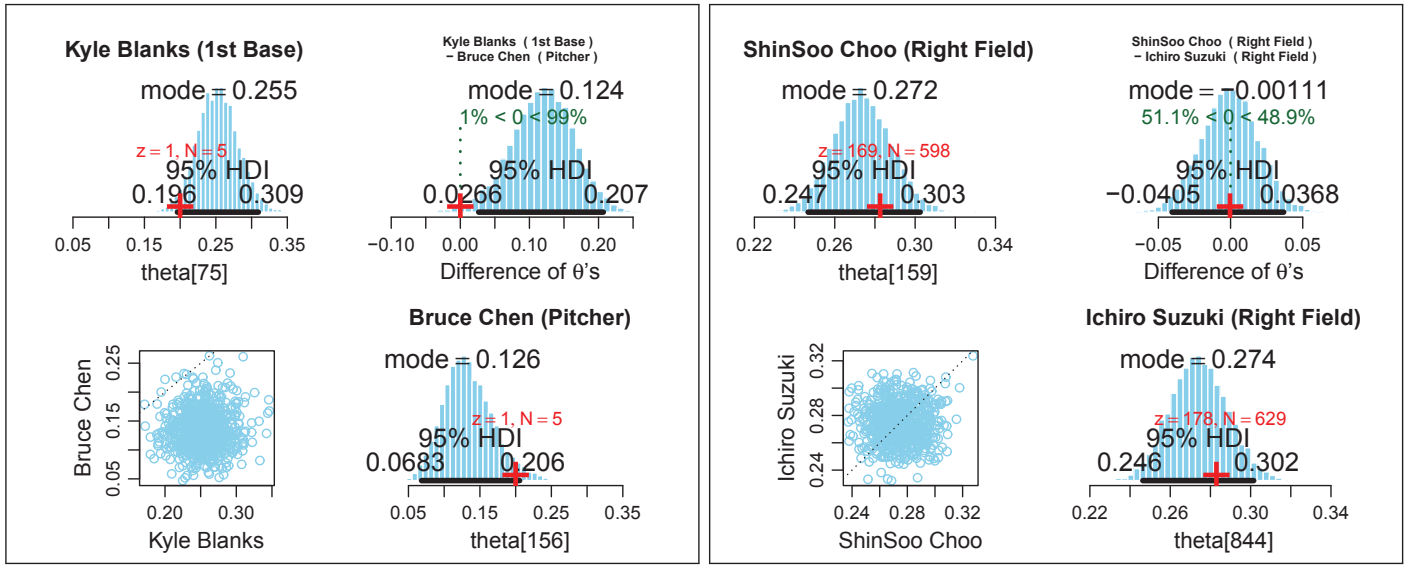


Figure 9.15: Marginal posterior distributions for baseball batting data. *Left quartet:* Individual estimates for two players with identical records of 1 hit in only 5 at-bats, but from two different positions. Although the batting records are identical, the estimated batting abilities are very different. *Right quartet:* Individual estimates for two right fielders with large numbers of at-bats. The posterior distributions of their individual performances have narrow HDI's compared with the left quartet, and are shrunken slightly toward the position-specific mode (which is about 0.247). The posterior distribution of their difference is essentially zero and the 95% HDI of the difference is very nearly contained within a ROPE from −0.04 to +0.04 (except for MCMC instability). Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Figure 9.16

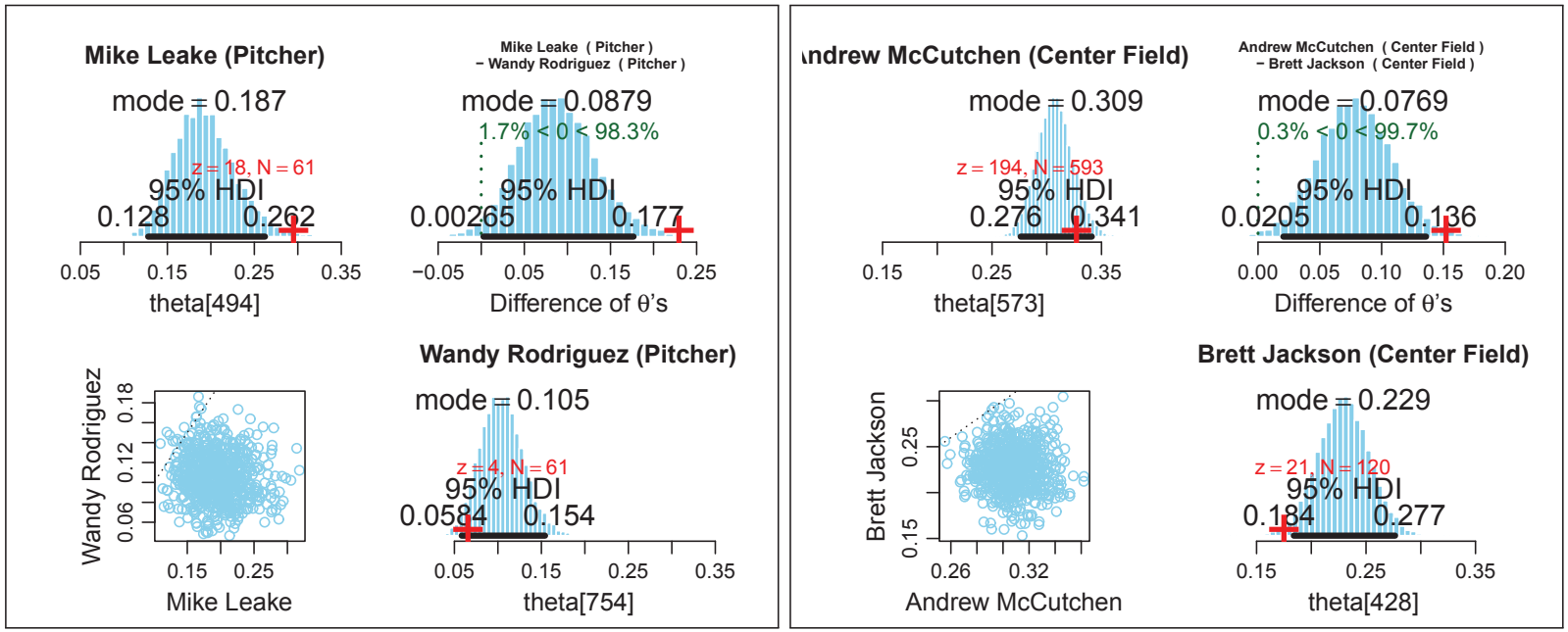


Figure 9.16: Marginal posterior distributions for baseball batting data. *Left quartet:* Two pitchers each with 61 at-bats but very different numbers of hits. Despite the difference in performance, shrinkage toward the position-specific mode leaves the posterior distribution of their difference marginally spanning zero. *Right quartet:* Two center fielders with very different batting averages, and moderately large at-bats. Despite some shrinkage toward the position-specific mode, the larger set of data makes the posterior distribution of their difference notably exclude zero.

Figure 9.17

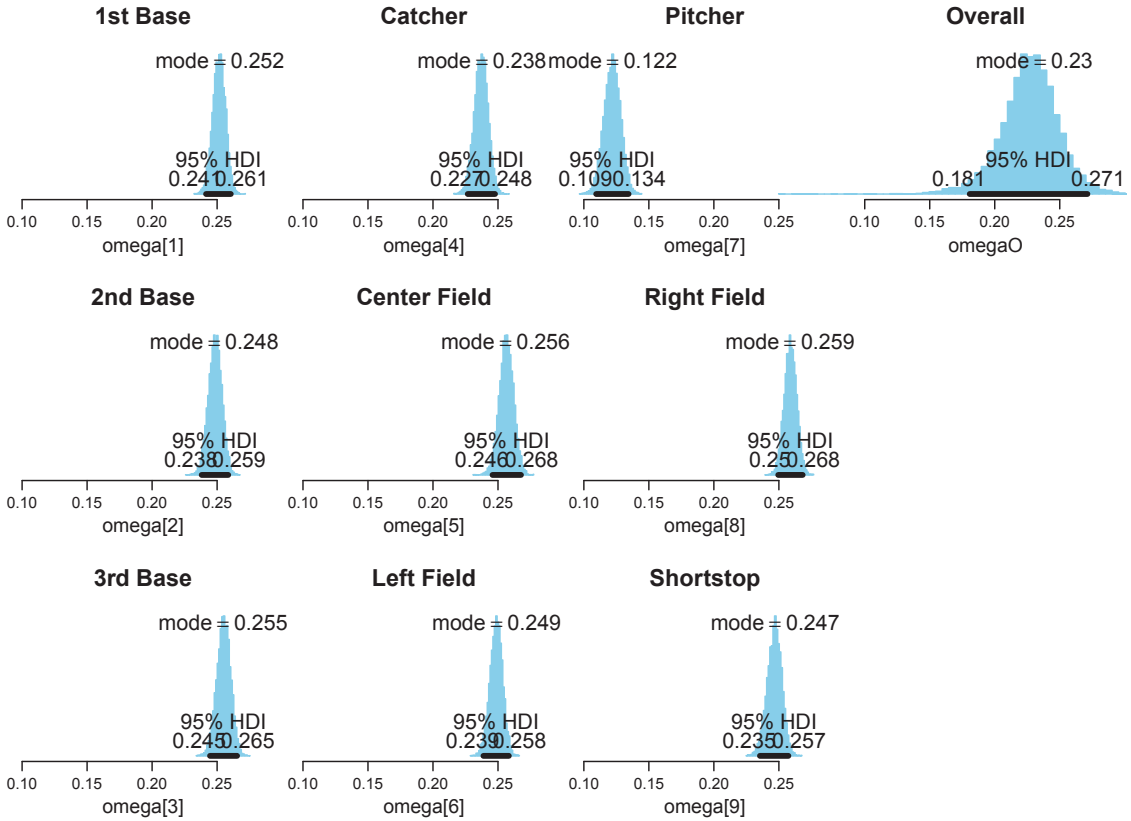


Figure 9.17: Marginal posterior distributions for baseball batting data. Notice that the estimate of the overall mode `omega0` is less certain (wider HDI) than the estimate of position modes `omega[c]`. One reason for the different certainties is that there are dozens or hundreds of individuals contributing to each position, but only nine positions contributing to the overall level. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.