Introduction to Bayesian Statistics I

Dawie van Lill

ECO872: Advanced Time Series Econometrics

July 16, 2020

Introduction: Which textbook?

Primary textbooks used to compile the slides and class notes.

- **Gelman, et al. (AG)** (2014) Bayesian Data Analysis (NB)
- **McElroy (RM)** (2019) Statistical Rethinking: A Bayesian Course with R
- **Kruschke (JK)** (2015) Doing Bayesian Data Analysis
- **Koop (GK)** (2003) Bayesian Econometrics (NB)
- **Geweke (JG)** (2005) Contemporary Bayesian Econometrics
- **Marin and Robert (MR)** (2014) Bayesian Essentials with R
- **Albert (JA)** (2009) Bayesian Computation with R
- **Gamerman and Lopes (GL)** (2006) Markov Chain Monte Carlo
- **Kotze (KK)** (2019) Time Series Analysis Notes (NB)
- **Commandeur and Koopman (CK)** (2007) An Introduction to State Space Time Series Analysis
- **Hamilton (JH)** (1994) Time Series Analysis

Chapters and articles will be **posted on SUNLearn**!

Introduction: What are we going to cover?

- My assumption $\rightarrow$ you know almost **nothing** about Bayesian econometrics
- We will consider the following topics
  - Introduction to Bayesian statistics (2 Sessions)
  - Bayesian computation (3 Sessions)
  - Bayesian regression models (1 Session)
  - State space modelling + Kalman filter (2 Sessions)
  - Bayesian VARs (2 Sessions)
  - TVP-VAR with stochastic volatility (2 Sessions)
- Coding will be done mostly in R, but it is possible that MATLAB will also be used.
- Face-to-face lectures will take place on Friday, with Tuesday reserved for online sessions. We will mostly do computer related work in the Tuesday sessions.

Important readings

- These notes draw heavily from **AG, RM, JK, GK**.
- We follow the conventions in **AG**. However, the book is quite technical, so if you want something more approachable try **RM** or **JK**.
- For background reading look at Ch 2, 4, 5 from **JK**.
- **RM** also has some excellent lecture videos online
- I will also post some general Bayesian readings on SUNLearn

- So, let's get started with Chapter 1 + 2 from **AG**!

Broad outline of the lecture for today

- Bayesian paradigm
- Binomial model (repeated experiment with binary outcome)
- The posterior as a compromise between data and prior information
- Posterior summaries
- Informative prior distributions (skip exponential families and sufficient statistics)
- Gaussian model with known variance

Bayesian data analysis

- Count all the ways data can happen according to assumptions
- Assumptions with more ways that are consistent with data are more plausible

- Consider an example: The garden of forking data (**RM**)

Bayesian paradigm

- Consider an economic model that describes an AR(1) process,

$$y_t = \mu + \alpha y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}\left[0, \sigma^2\right]$$

- where μ, α and σ^2 are parameters in a vector θ
- There may be many possible values for θ from population Θ
- Bayesians will test initial assertions regarding θ using data on y_t and y_{t-1} to investigate probability of assertions
- Provides probability distribution over possible values for $\theta \in \Theta$

Bayesian paradigm

- Unobserved variables (such as those mentioned in the previous slide) are usually called **parameters** and can be inferred from other variables
- θ represents the unobservable parameter of interest, where y is the observed data
- In the context of economics these parameters could be the coefficients in a regression model (or components of AR(1) model as discussed before)

- Bayesian conclusions about the parameter θ is made in terms of **probability statements**
- Statements are conditional on the observed values of y and can be written $p(\theta|y)$ $\leftarrow$ given the data, what do we know about θ?

Bayesian paradigm

- To make probability statements about θ given y, we begin with a model providing a joint probability distribution for θ and y
- Joint probability density can be written as product of two densities: the prior $p(\theta)$ and sampling distribution $p(y|\theta)$

$$p(\theta, y) = p(\theta)p(y|\theta)$$

- However, using the properties of conditional probability we can also write the joint probability density as

$$p(\theta, y) = p(y)p(\theta|y)$$

- Setting these equations equal and rearranging provides us with Bayes' theorem / rule

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

Bayesian paradigm

For our model we start with a numerical formulation of joint beliefs about y and θ expressed in terms of probability distributions

- For each $\theta \in \Theta$ the prior distribution $p(\theta)$ describes belief about true population characteristics
- For each $\theta \in \Theta$ and $y \in \mathcal{Y}$, our sampling model $p(y|\theta)$ describes belief that y would be the outcome of the study if we knew θ to be true.

Once data is obtained, the last step is to update beliefs about θ

- For each $\theta \in \Theta$ our posterior distribution $p(\theta|y)$ describes our belief that θ is the true value having observed the dataset.

Bayesian paradigm

- We can safely ignore $p(y)$ in Bayes' rule since it does not involve the parameter of interest (θ), which means we can write

$$p(\theta|y) \propto p(\theta)p(y|\theta)$$

- The **posterior density** $p(\theta|y)$ summarises all we know about θ after seeing the data
- The **prior density** $p(\theta)$ does not depend on the data (what you know about θ prior to seeing data)
- The **likelihood function** $p(y|\theta)$ is the data generating process (density of the data conditional on the parameters in the model)
- Finally, $p(y \mid \theta)p(\theta) = p(y, \theta)$ is the **econometric model** (joint probability distribution of data and parameters)

Model vs. likelihood (notational sloppiness)

- Bayes' rule $p(\theta|y) \propto p(\theta)p(y|\theta)$
- Sampling model: $p_Y(Y|\Theta = \theta) = p(y|\theta)$ as a function of y given fixed θ describes the aleatoric uncertainty
- Likelihood: $p_\Theta(Y = y|\Theta) = p(y|\theta) = L(\theta|y)$ as a function of θ given fixed y provides information about epistemic uncertainty, but is **not a probability distribution**
- Bayes' rule combines the **likelihood** with **prior** uncertainty $p(\theta)$ and transforms them to updated **posterior** uncertainty

Prediction in a single parameter model

- Prediction is based on the predictive density $p(y^*|y)$
- Since a marginal density can be obtained from a join density through integration, we have:

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

- The term inside the integral can be written as:

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

- Prediction then involves the posterior and $p(y^* \mid y, \theta)$.
- Sometimes we will use the notation $\tilde{y}$ to indicate y^*
- Important to note that y^* is also an unobservable component

Binomial: known θ

- There are two events in a trial
- Probability of event 1 in the trial is θ
- Probability of event 2 in the trial is $1 - \theta$
- Probability of several events in independent trials is $\theta\theta(1-\theta)\theta(1-\theta)(1-\theta)\ldots$
- If there are n trials and we don't care about the order of the events, then the probability that event 1 happens y times is

$$\begin{aligned} p(y|\theta, n) &= \frac{n!}{y!(n-y)!}\theta^y(1-\theta)^{n-y} \\ &= \binom{n}{y}\theta^y(1-\theta)^{n-y} \end{aligned}$$

- Let's also do this example in R, to get some practice

Binomial: known θ

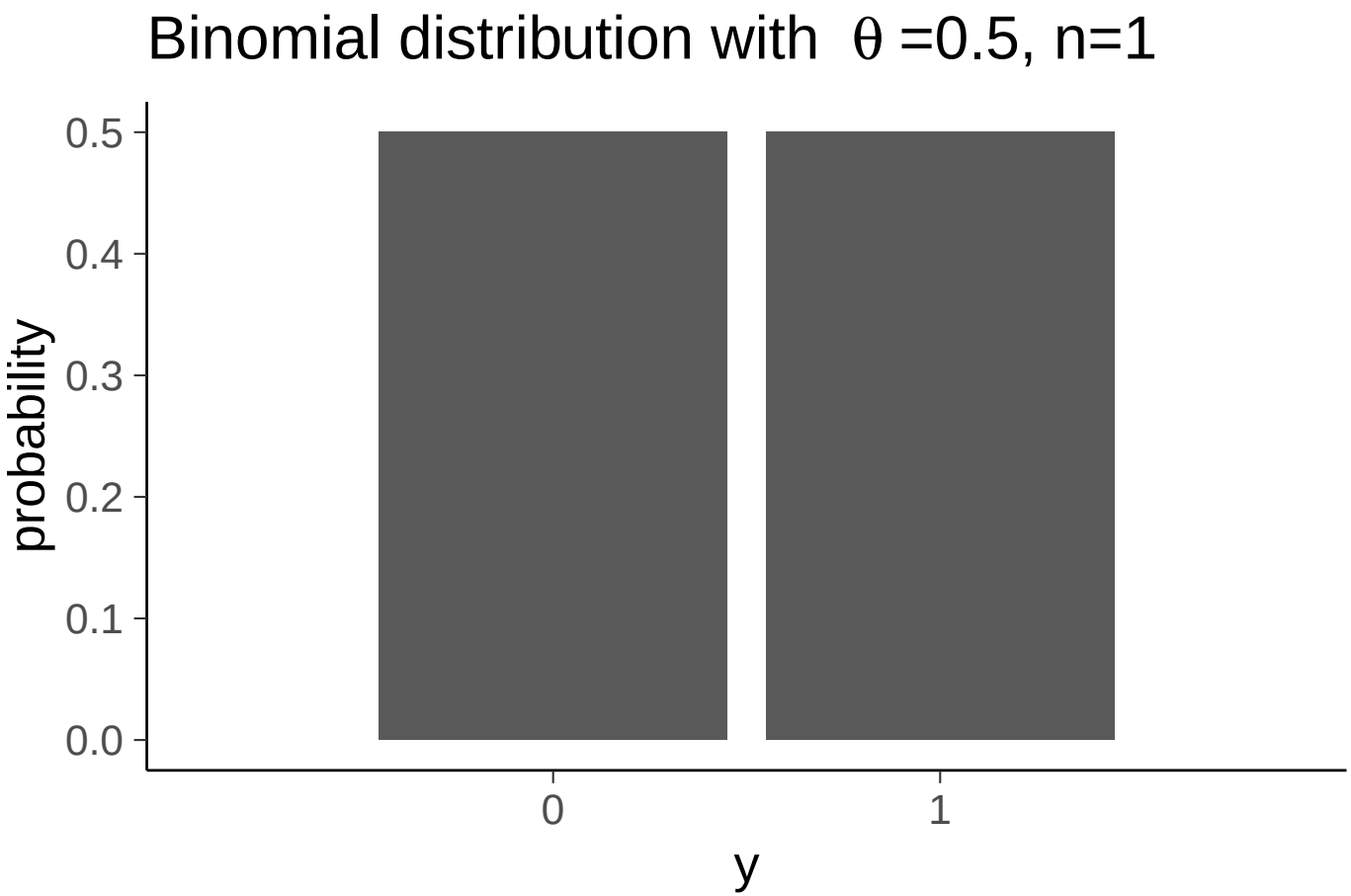
- Observation model (function of y, discrete)

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

Binomial: known θ

- Observation model (function of y, discrete)

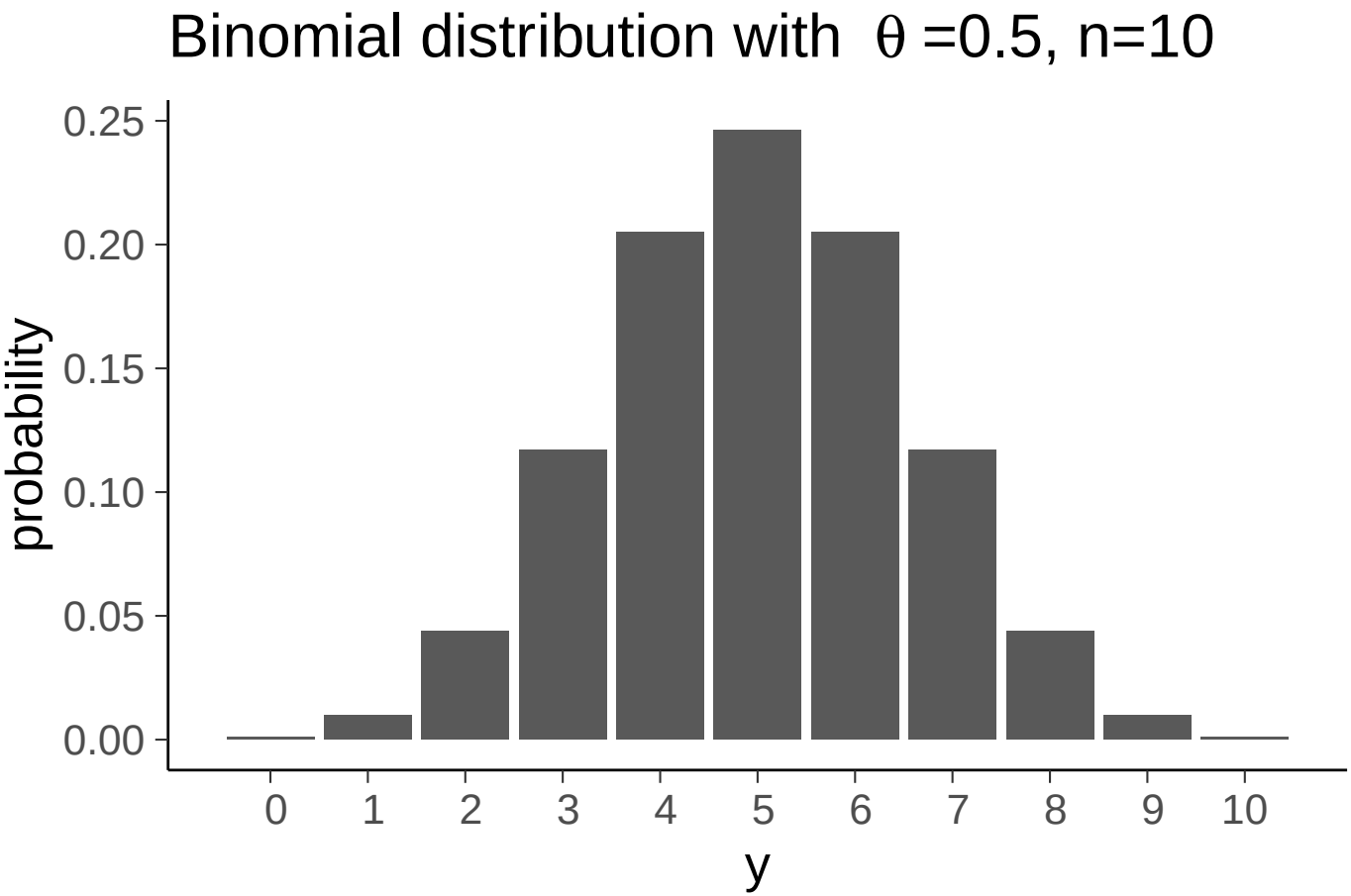
$$p(y|\theta, n) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$



Binomial: known θ

- Observation model (function of y, discrete)

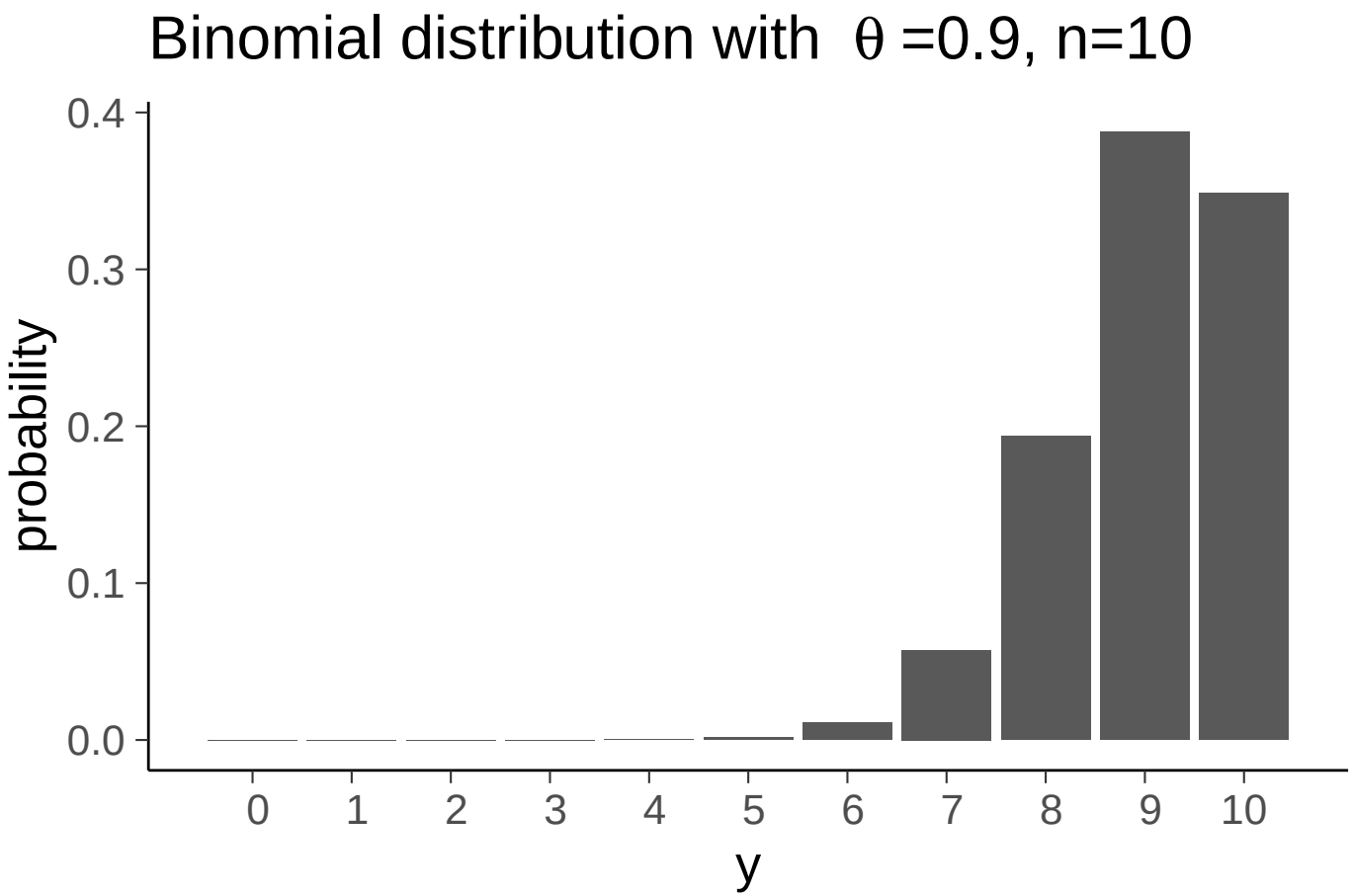
$$p(y|\theta, n) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$



Binomial: known θ

- Observation model (function of y, discrete)

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



Estimating bias in a coin

- We observe the number of heads that result from flipping a coin and we estimate its underlying probability of coming up heads.
- Want to create a descriptive model with meaningful parameters
- The outcome of a flip will be given by y, with $y = 1$ indicating heads and $y = 0$ tails.
- We need underlying probability of heads as value of parameter θ
- Formally, this can be written as $p(y = 1|\theta) = \theta \leftarrow$ the probability that the outcome is heads, given a parameter value of θ, is the value θ.
- We also need the probability of tails, which is the complement of probability of heads $\rightarrow p(y = 0|\theta) = 1 - \theta$

Estimating bias in a coin

- Combine the equations for the probability of heads and tails

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

- This probability distribution is called the Bernoulli distribution
- This is a distribution over two discrete values of y for a fixed value of $\theta \leftarrow$ construction of our **sampling model**
- The sum of the probabilities is 1 (which must be the case for a probability distribution)

$$\sum_y p(y|\theta) = p(y=1|\theta) + p(y=0|\theta) = \theta + (1-\theta) = 1$$

- If we consider y fixed and the value of θ as variable, then our equation is a **likelihood function** of θ
- This likelihood function is not a probability distribution, suppose that $y = 1$ then $\int_0^1 \theta^y(1-\theta)^{1-y}\mathrm{d}\theta = \int_0^1 \theta^y \mathrm{d}\theta = 1/2$

Estimating bias in a coin

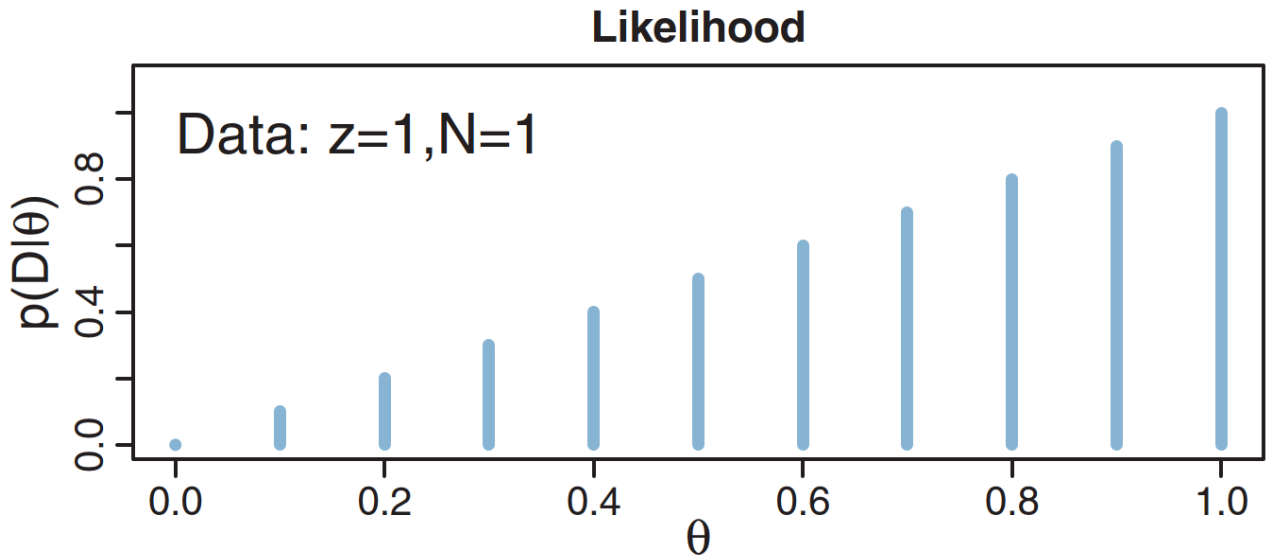
- Let us take a look at what happens for multiple flips
- Outcome of ith flip is given by y_i and set of outcomes is $\{y_i\}$
- Formula for the probability of the set of outcomes is given by

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

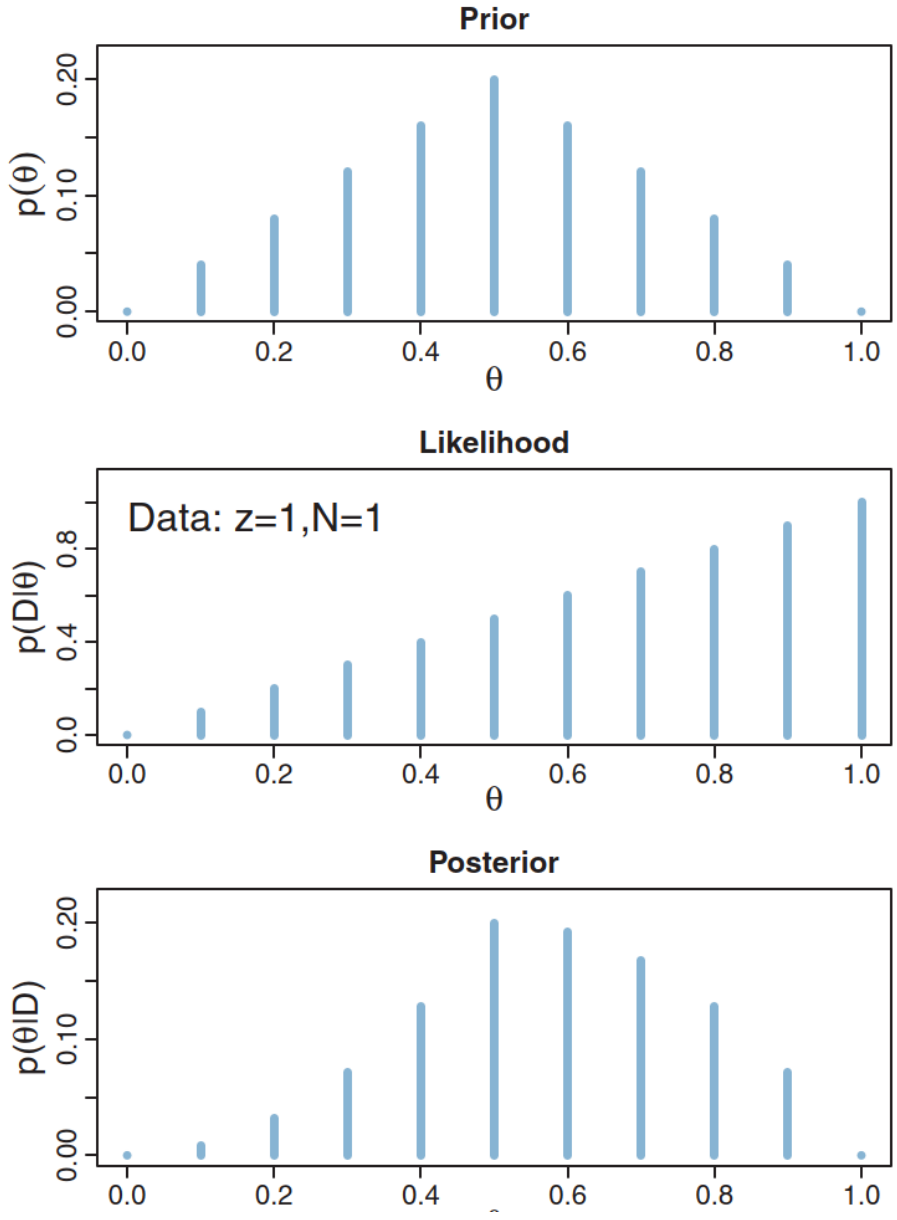
- Next we establish the prior, which will be an arbitrary choice here
- One assumption could be that the factory producing the coins tends to produce mostly fair coins (normal distribution on prior)
- Indicate number of heads by z and number of flips by N
- Suppose that we flip the coin only once and observe heads, then the data D consists of $y = 1$ or $z = 1$ and $N = 1$

Estimating bias in a coin

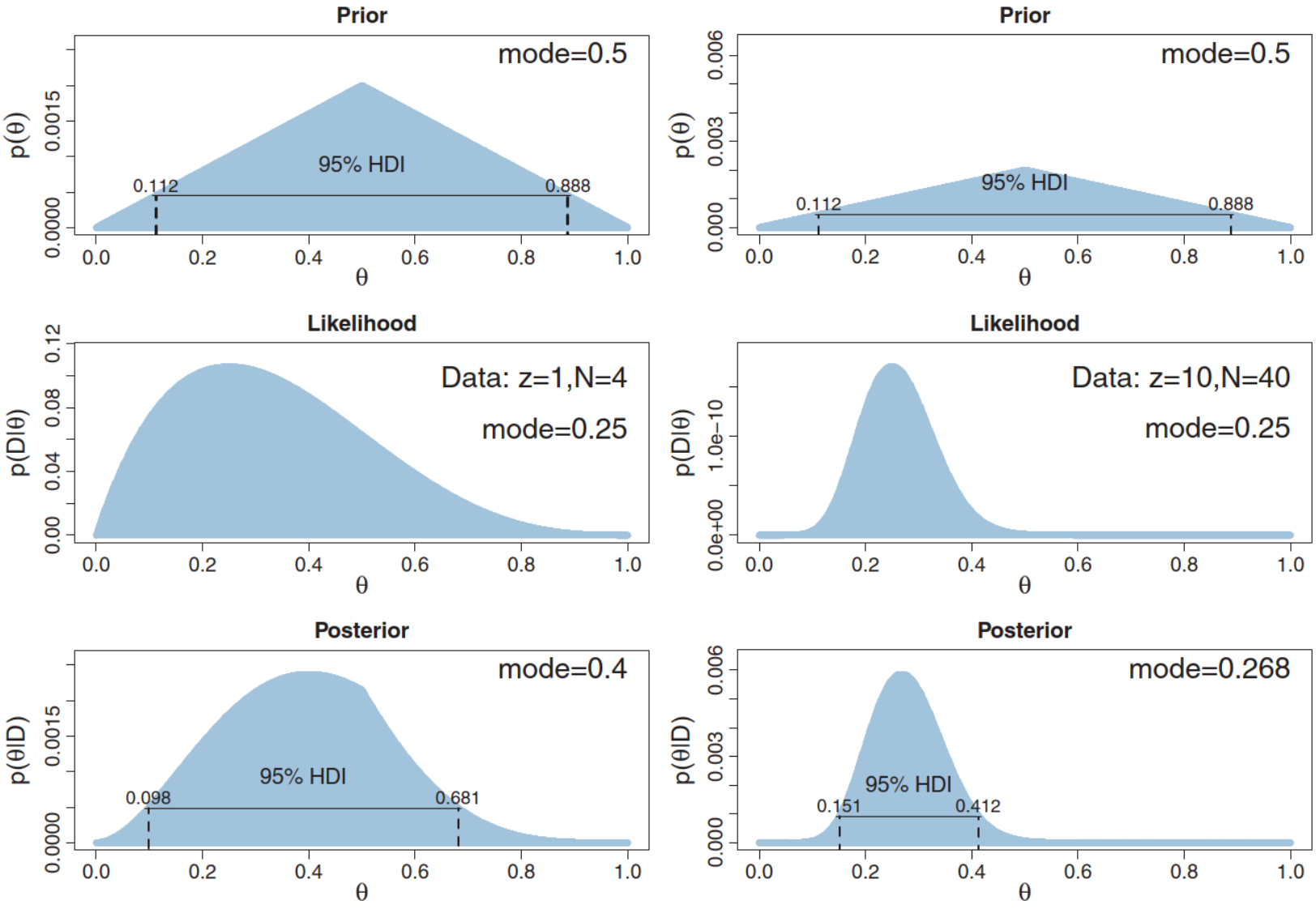
- We see that our likelihood function becomes $p(D|\theta) = \theta$
- This means, for example, that $p(D|\theta = 0.9) = 0.9$ and $p(D|\theta = 0.2) = 0.2$ which is reflected in the graph
- In other words, the bar at $\theta = 0.9$ has height $p(D|\theta) = 0.9$
- Likelihood that $\theta = 0.2$ given the that heads was flipped is 0.2



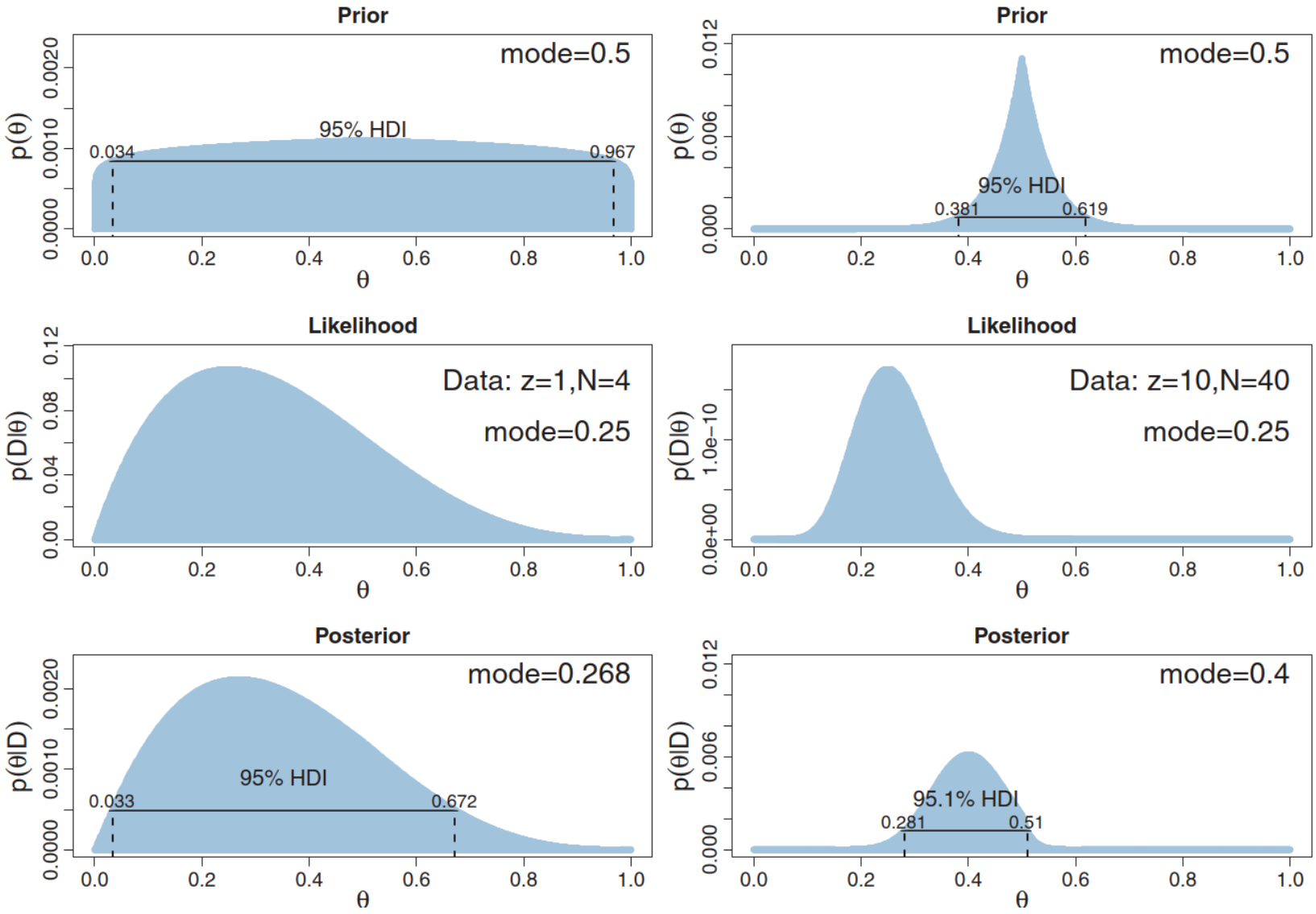
Estimating bias in a coin



Influence of sample size on posterior



Influence of prior on posterior



Quick digression on the Beta distribution

- Suppose that for our coin flipping model we want to derive the posterior credibilities of parameter values.
- We would need a mathematical description of the prior probability for each value of the parameter θ on interval $[0, 1]$.
- Any relevant probability density function would work, but there are two desiderata for mathematical tractability.
  1. Product of $p(y|\theta)$ and $p(\theta)$ results in same form as $p(\theta)$.
  2. $\int p(y|\theta)p(\theta)\mathrm{d}\theta$ should be solvable analytically
- When the forms of $p(y|\theta)$ and $p(\theta)$ combine so that the posterior has the same form as the prior distribution then $p(\theta)$ is called conjugate prior for $p(y|\theta)$
- Prior is conjugate with respect to particular likelihood function
- We are looking for a functional form for a prior density over θ that is conjugate to the **Bernoulli likelihood function**

Quick digression on the Beta distribution

- If the prior is of the form, $\theta^a(1-\theta)^b$ then when you multiply with Bernoulli likelihood you will get

$$\theta^{y+a}(1-\theta)^{(1-y+b)}$$

- A probability density of this form is called the Beta distribution
- Beta distribution has two parameters, called a and b

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

- In this case $B(a,b)$ is a normalising constant, to make sure area under Beta density integrates to 1.
- Beta function is given by $\int_0^1 \theta^{a-1}(1-\theta)^{(b-1)}\mathrm{d}\theta$

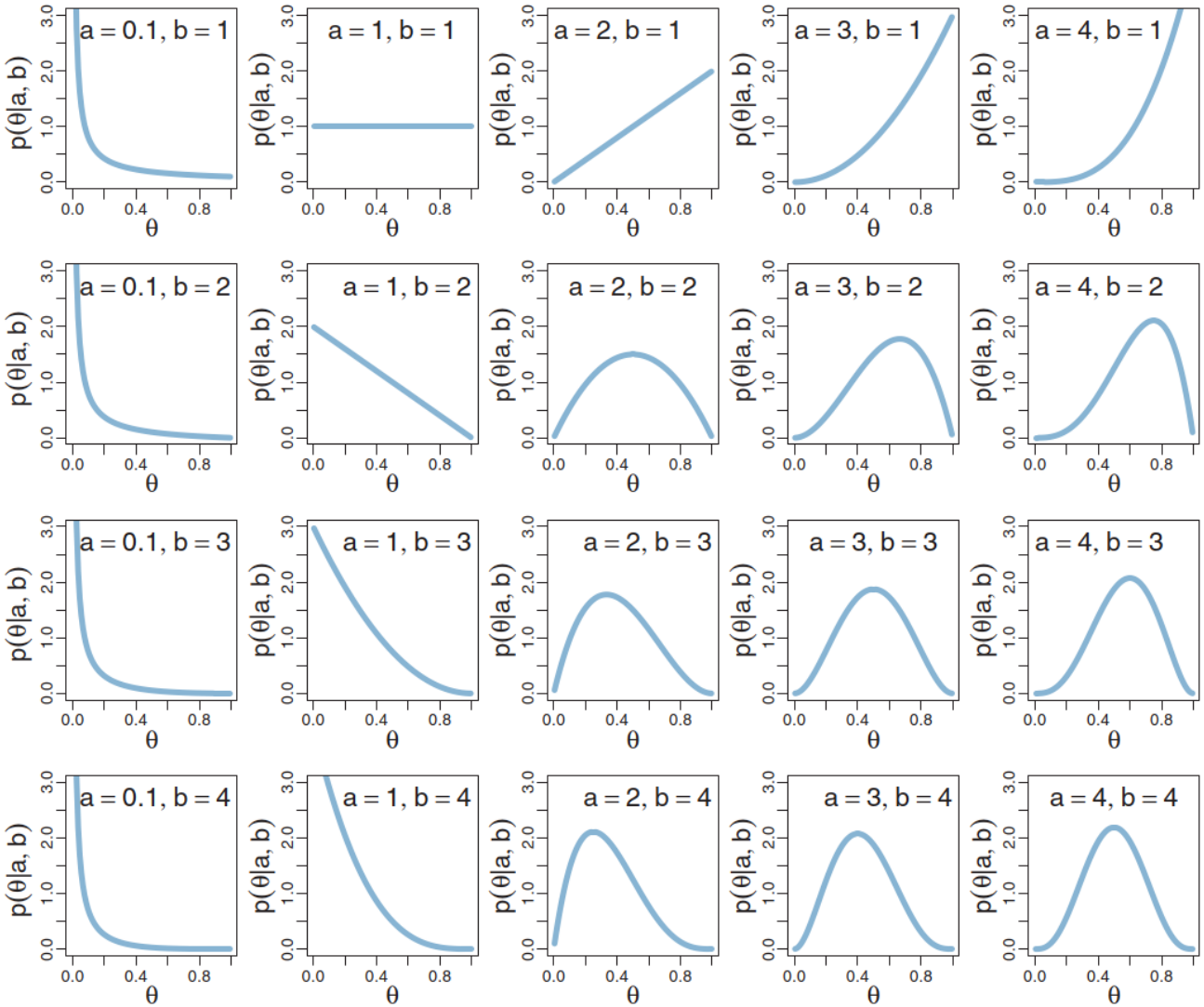
Quick digression on the Beta distribution

- Another way in which we can express the Beta function,

$$B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b) \quad \text{where} \quad \Gamma(a) = \int_0^\infty t^{(a-1)}\exp(-t)$$

- The variables a and b are called the shape parameters of the Beta distribution (they determine the shape)
- Refer to the next slide for some examples of the distribution for different values of a and b over θ

Quick digression on the Beta distribution



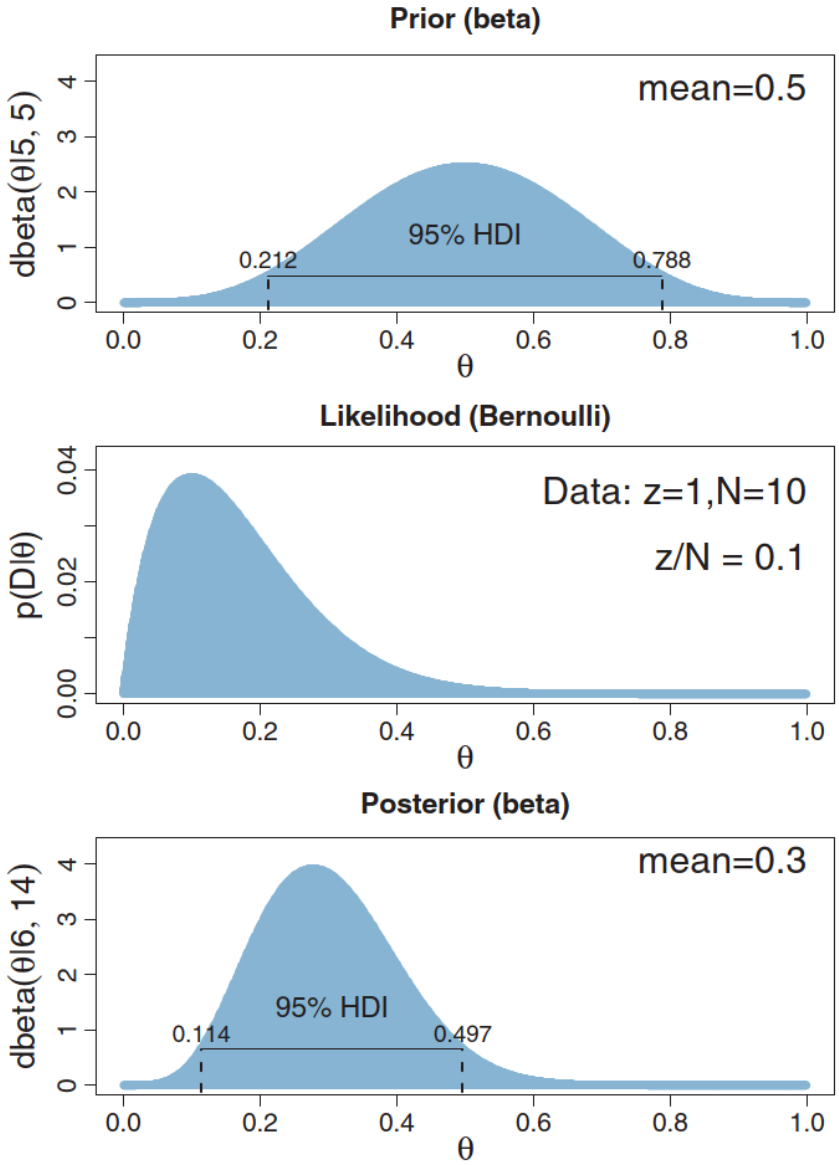
Quick digression on the Beta distribution

- Suppose we have set of data with N flips and z heads, then

$$\begin{aligned}
p(\theta|z,N) =& p(z,N|\theta)p(\theta)/p(z,N) \\
=& \theta^z(1-\theta)^{(N-z)}\frac{\theta^{a-1}(1-\theta)^{(b-1)}}{B(a,b)}/p(z,N) \\
=& \theta^z(1-\theta)^{(N-z)}\theta^{a-1}(1-\theta)^{(b-1)}/[B(a,b)p(z,N)] \\
=& \theta^{((z+a)-1)}(1-\theta)^{((N-z+b)-1)}/[B(a,b)p(z,N)] \\
=& \theta^{((z+a)-1)}(1-\theta)^{((N-z+b)-1)}/B(z+a,N-z+b)
\end{aligned}$$

- Last step was made by considering what the normalising factor should be for the numerator of the Beta distribution.
- From this we see that if prior is Beta$(\theta|a,b)$ then the posterior will be Beta$(\theta|z+a,N-z+b)$

Quick digression on the Beta distribution



Quick digression on the Beta distribution

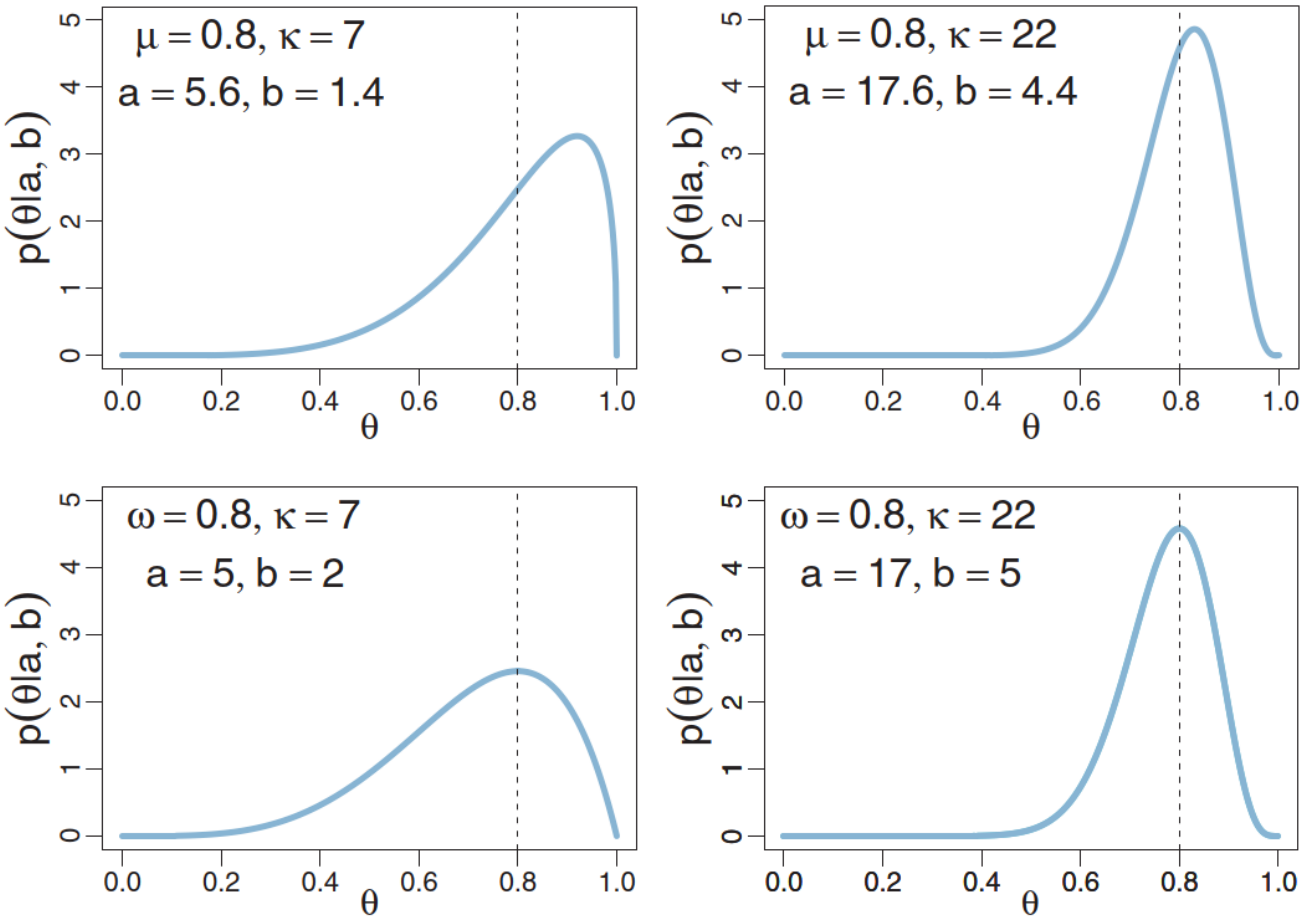
- Mean of Beta distribution is given by $\mu = a/(a+b)$ and mode is $\omega = (a-1)/(a+b-2)$ for $a, b > 1$
- When $a = b$ the mean and mode are 0.5, so if $a > b$, the mean and mode are greater than 0.5
- Spread of Beta distribution is related to concentration $\kappa = a + b$
- As κ gets larger, the Beta distribution gets narrower
- Solving for a and b in terms of mean, mode and concentration,

$$a = \mu\kappa, \quad b = (1-\mu)\kappa,$$
$$a = \omega(\kappa - 2) + 1, \quad b = (1-\omega)(\kappa - 2)$$

Quick digression on the Beta distribution

- Value we choose for prior κ can be thought of as the number of new flips of the coin that we would need to make us teeter between new data and prior belief about μ.
- Suppose that a coin is fair, so $\mu = 0.5$, but I'm not entirely sure.
- I have only seen $\kappa = 8$ previous flips, so $a = \mu\kappa = 4$ and $b = (1 - \mu)\kappa = 4$, which is a beta distribution which peaks at $\theta = 0.5$ (see the graph two slides back)
- The mode can be more intuitive measure of central tendency for skewed distributions than the mean.
- Some examples in the next slide that show the difference between modes and means for different parameterisations of the Beta distribution.

Quick digression on the Beta distribution



- See the slides from **RM** for another illustration of this example.

Binomial: unknown θ

- Now for a more general treatment of the binomial distribution
- Posterior with Bayes rule (function of θ, continuous)

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

Binomial: unknown θ

- Now for a more general treatment of the binomial distribution
- Posterior with Bayes rule (function of θ, continuous)

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

where $p(y|n,M) = \int p(y|\theta,n,M)p(\theta|n,M)d\theta$

- Start with uniform prior

$$p(\theta|n,M) = p(\theta|M) = 1, \text{ with } 0 \leq \theta \leq 1$$

- Then

$$p(\theta|y,n,M) = \frac{p(y|\theta,n,M)}{p(y|n,M)} = \frac{\binom{n}{y}\theta^y(1-\theta)^{n-y}}{\int_0^1 \binom{n}{y}\theta^y(1-\theta)^{n-y}d\theta}$$

$$= \frac{1}{Z}\theta^y(1-\theta)^{n-y}$$

Binomial: unknown θ

- Normalization term Z (constant given y)

$$Z = p(y|n, M) = \int_0^1 \theta^y (1-\theta)^{n-y} d\theta = \frac{\Gamma(y+1)\Gamma(n-y+1)}{\Gamma(n+2)}$$

- Normalisation term has Beta function form
  - When integrated over $(0, 1)$ the result can presented with Gamma functions
  - With integers $\Gamma(n) = (n-1)!$

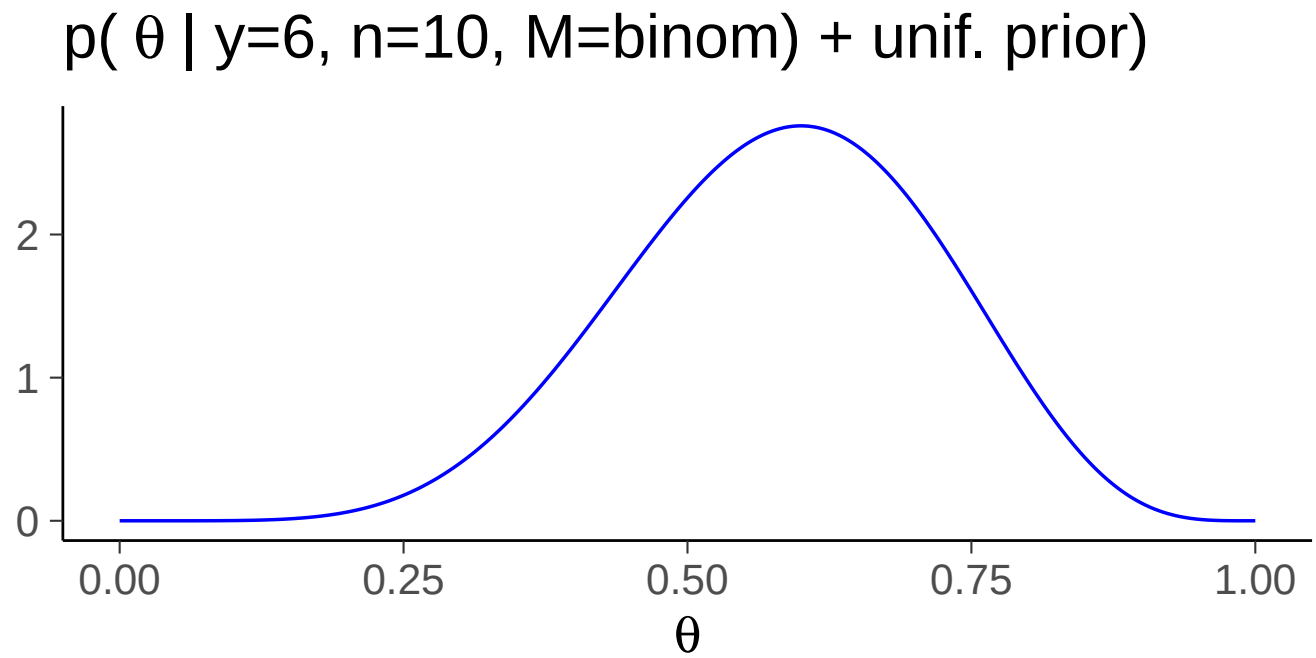
Binomial: unknown θ

- Posterior is

$$p(\theta|y,n,M) = \frac{\Gamma(n+2)}{\Gamma(y+1)\Gamma(n-y+1)}\theta^y(1-\theta)^{n-y},$$

which is called Beta distribution

$$\theta|y,n \sim \text{Beta}(y+1, n-y+1)$$

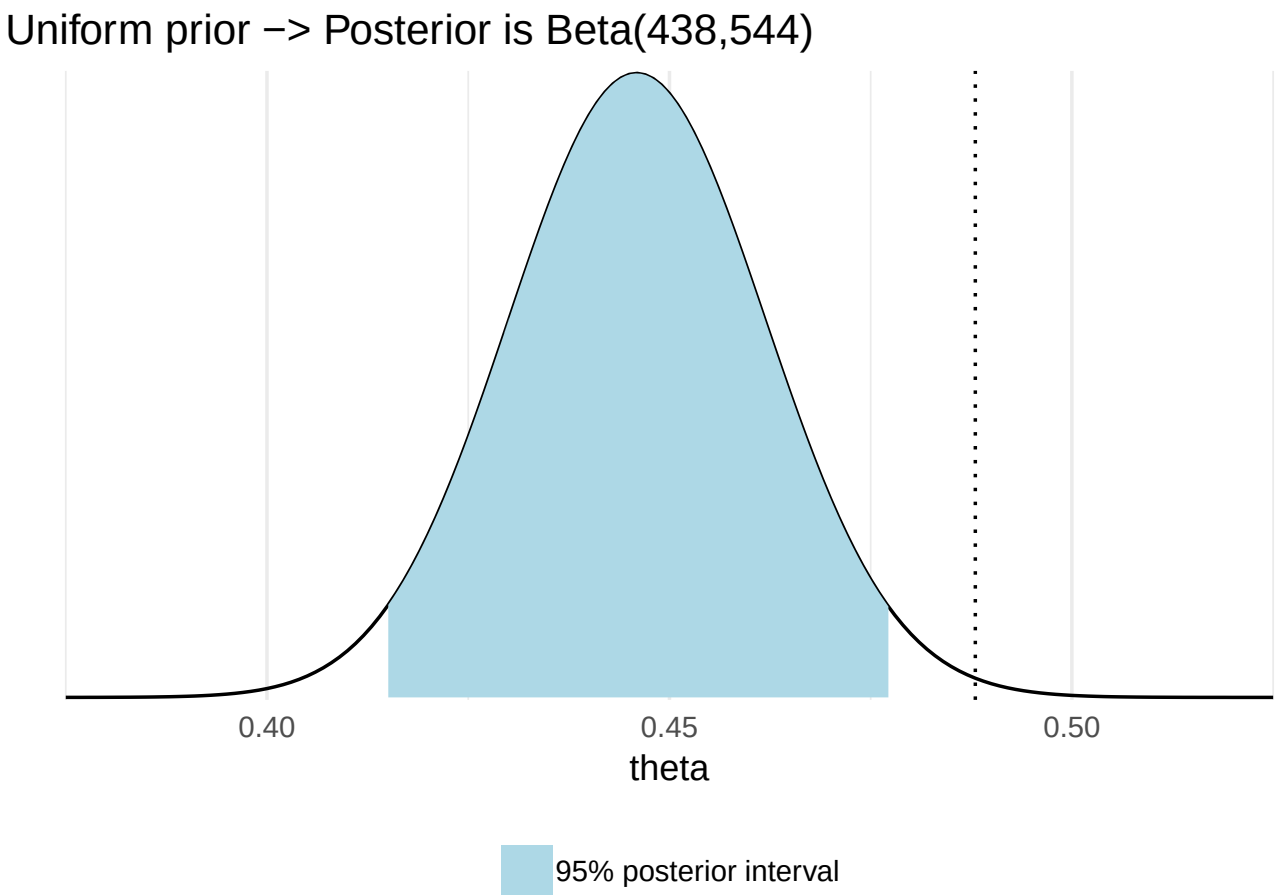


Placenta previa

- Probability of a girl birth given placenta previa (p. 37)
  - 437 girls and 543 boys have been observed
  - is the ratio 0.445 different from the population average 0.485?

Placenta previa

- Probability of a girl birth given placenta previa (p. 37)
  - 437 girls and 543 boys have been observed
  - is the ratio 0.445 different from the population average 0.485?



Priors

- Below is a short list of priors discussed in **AG**

- Conjugate prior (p. 35)
- Noninformative prior (p. 51)
- Proper and improper prior (p. 52)
- Weakly informative prior (p. 55)
- Informative prior (p. 55)
- Prior sensitivity (p. 38)

Conjugate prior

- Prior and posterior have the same form
  - Only for exponential family distributions (plus for some irregular cases)
- Used to be important for computational reasons (1990s), and still sometimes used for special models to allow partial analytic marginalization
  - With Hamiltonian Monte Carlo (as an example) being used in Stan there is no computational benefit
- We won't place much emphasis on conjugate priors, since we will look at state-of-the-art computational methods
- Simple example of Beta prior for Binomial model on next slide (in order to give the general idea)

Beta prior for Binomial model

- Prior

$$\text{Beta}(\theta|\alpha,\beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

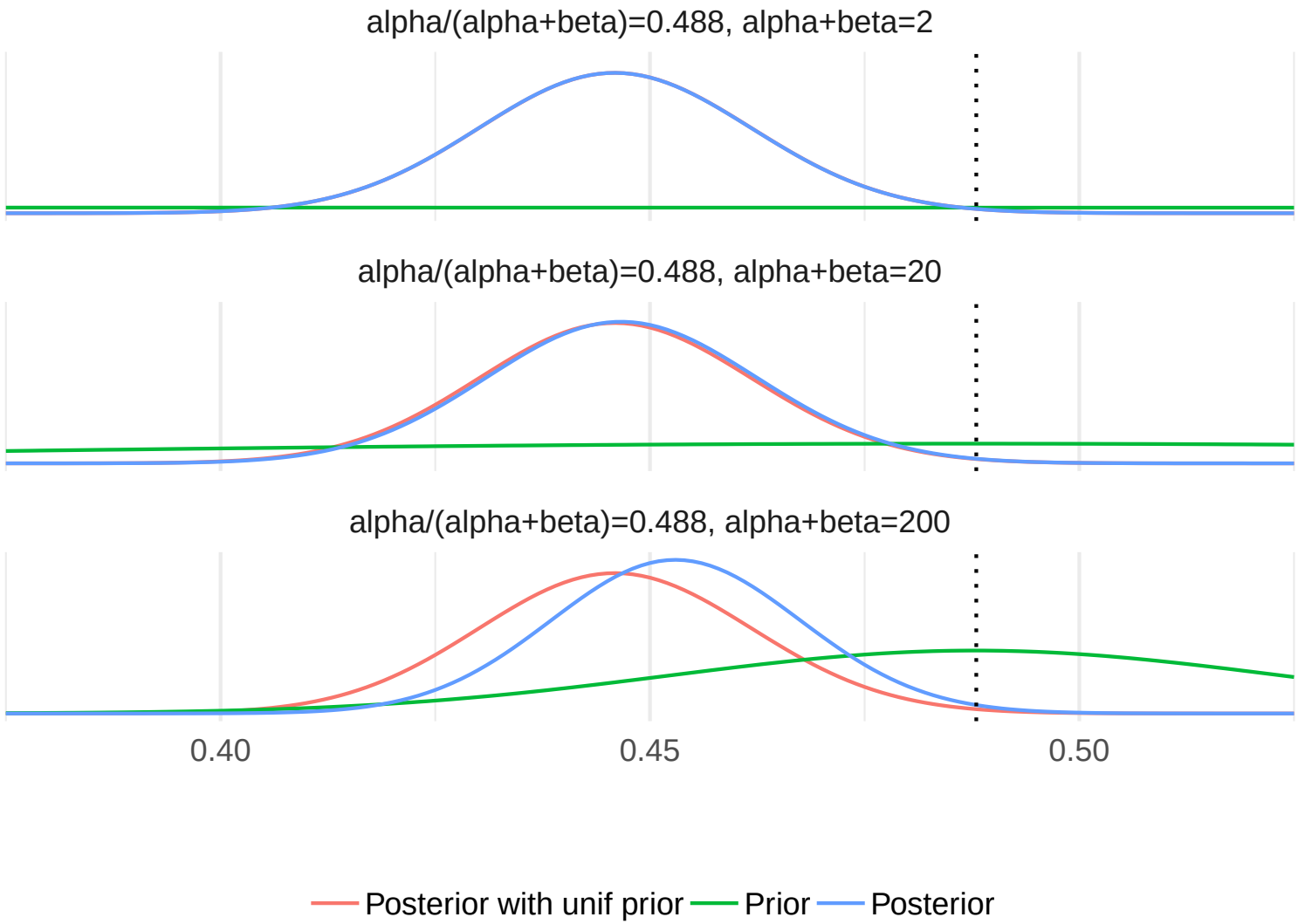
- Posterior

$$\begin{aligned} p(\theta|y,n,M) &\propto \theta^{y}(1-\theta)^{n-y}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1} \\ &= \text{Beta}(\theta|\alpha+y,\beta+n-y) \end{aligned}$$

- $(\alpha-1)$ and $(\beta-1)$ can considered to be number of prior observations
- Uniform prior when $\alpha = 1$ and $\beta = 1$

Placenta previa

- Beta prior centered on population average 0.485



Beta prior for Binomial model

- Posterior

$$p(\theta|y, n, M) = \text{Beta}(\theta|\alpha + y, \beta + n - y)$$

- Posterior mean

$$\text{E}[\theta|y] = \frac{\alpha + y}{\alpha + \beta + n}$$

  - combination prior and likelihood information
  - when $n \to \infty$, $\text{E}[\theta|y] \to y/n$

Beta prior for Binomial model

- Posterior

$$p(\theta|y, n, M) = \text{Beta}(\theta|\alpha + y, \beta + n - y)$$

- Posterior mean

$$\text{E}[\theta|y] = \frac{\alpha + y}{\alpha + \beta + n}$$

  - combination prior and likelihood information
  - when $n \to \infty$, $\text{E}[\theta|y] \to y/n$

- Posterior variance

$$\text{Var}[\theta|y] = \frac{\text{E}[\theta|y](1 - \text{E}[\theta|y])}{\alpha + \beta + n + 1}$$

  - decreases when n increases
  - when $n \to \infty$, $\text{Var}[\theta|y] \to 0$

Noninformative prior, proper and improper prior

- Vague, flat, diffuse of noninformative
  - try to “to let the data speak for themselves”
  - flat is not non-informative
  - flat can be stupid
  - making prior flat somewhere can make it non-flat somewhere else
- Proper prior has $\int p(\theta) = 1$
- Improper prior density doesn’t have a finite integral
  - the posterior can still sometimes be proper

Weakly informative priors

- Weakly informative priors produce computationally better behaving posteriors
  - quite often there's at least some knowledge about the scale
  - useful also if there's more information from previous observations, but not certain how well that information is applicable in a new case uncertainty

Weakly informative priors

- Weakly informative priors produce computationally better behaving posteriors
  - quite often there's at least some knowledge about the scale
  - useful also if there's more information from previous observations, but not certain how well that information is applicable in a new case uncertainty
- Construction
  - Start with some version of a noninformative prior distribution and then add enough information so that inferences are constrained to be reasonable.
  - Start with a strong, highly informative prior and broaden it to account for uncertainty in one's prior beliefs and in the applicability of any historically based prior distribution to new data.
- Stan team prior choice recommendations https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations

Example of informative prior

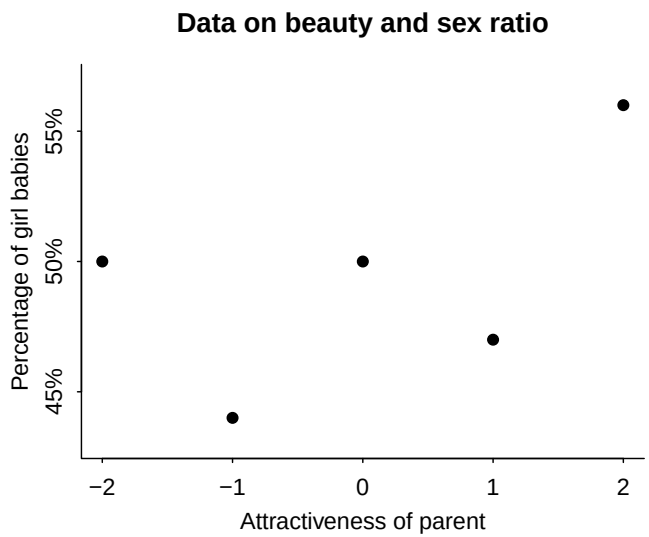
- The percentage of girl births is remarkably stable at about 48.8% girls, rarely varying by more than 0.5% from this rate

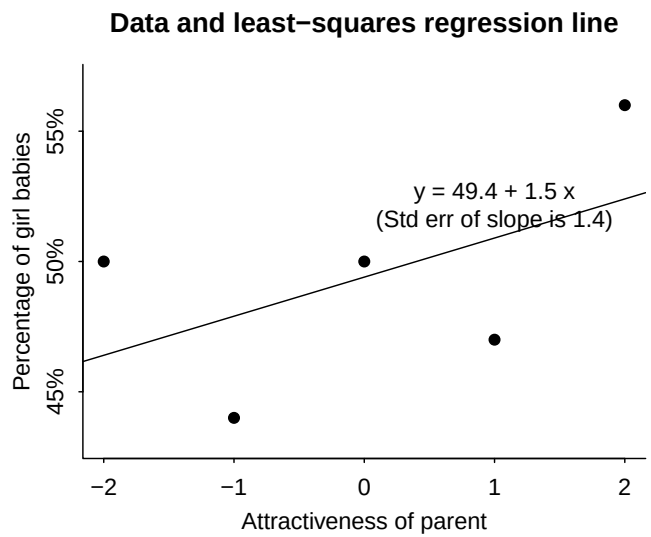
Example of informative prior

- The percentage of girl births is remarkably stable at about 48.8% girls, rarely varying by more than 0.5% from this rate
- There was a study on the percentage of girl births among parents in attractiveness categories 1–5 (assessed by interviewers in a face-to-face survey)

Example of informative prior

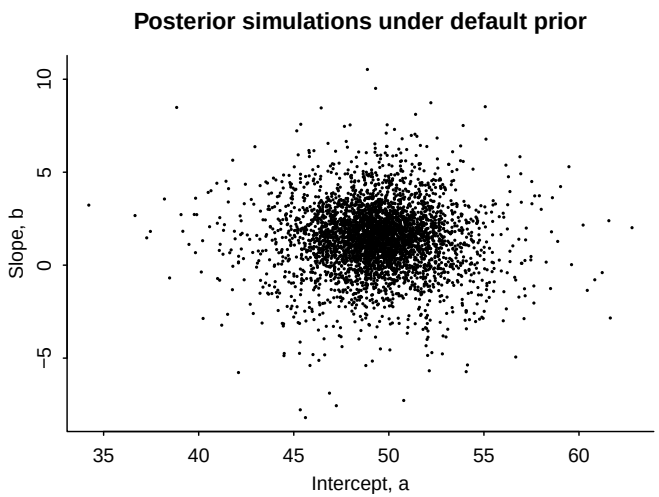
- The percentage of girl births is remarkably stable at about 48.8% girls, rarely varying by more than 0.5% from this rate
- There was a study on the percentage of girl births among parents in attractiveness categories 1–5 (assessed by interviewers in a face-to-face survey)

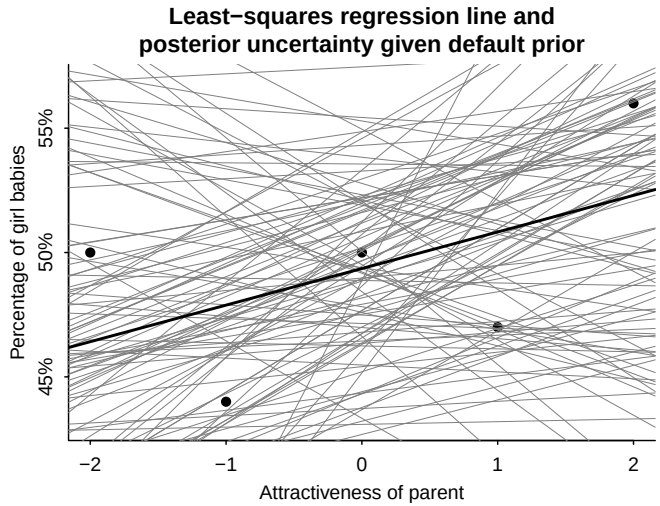


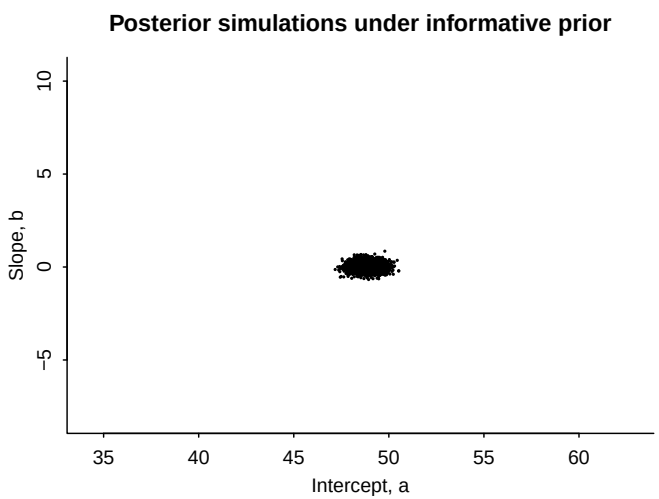


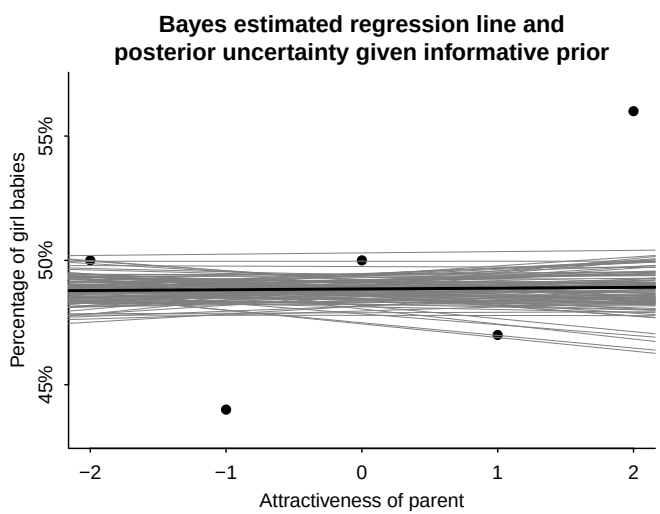
Example of informative prior

- The percentage of girl births is remarkably stable at about 48.8% girls, rarely varying by more than 0.5% from this rate

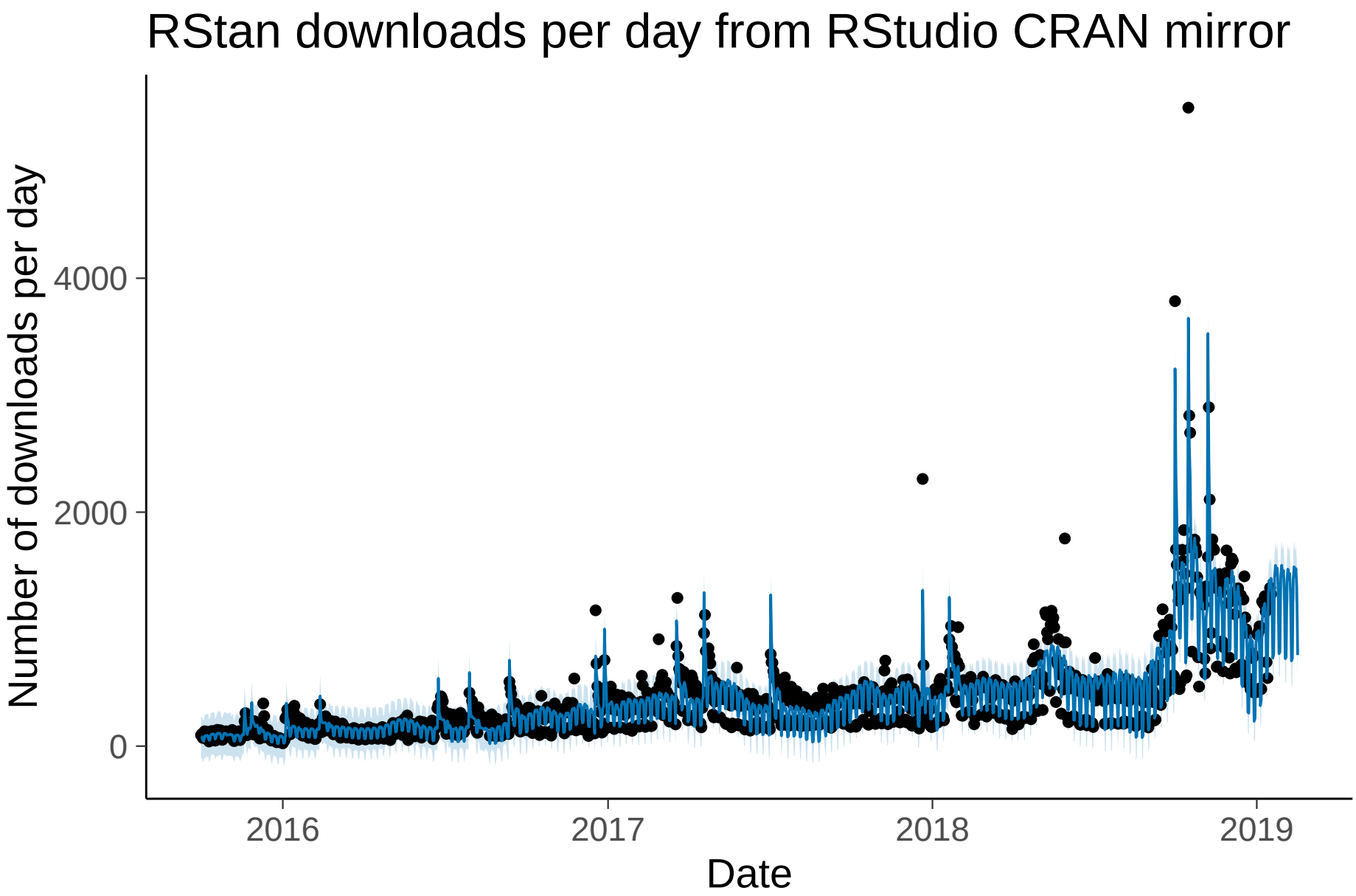




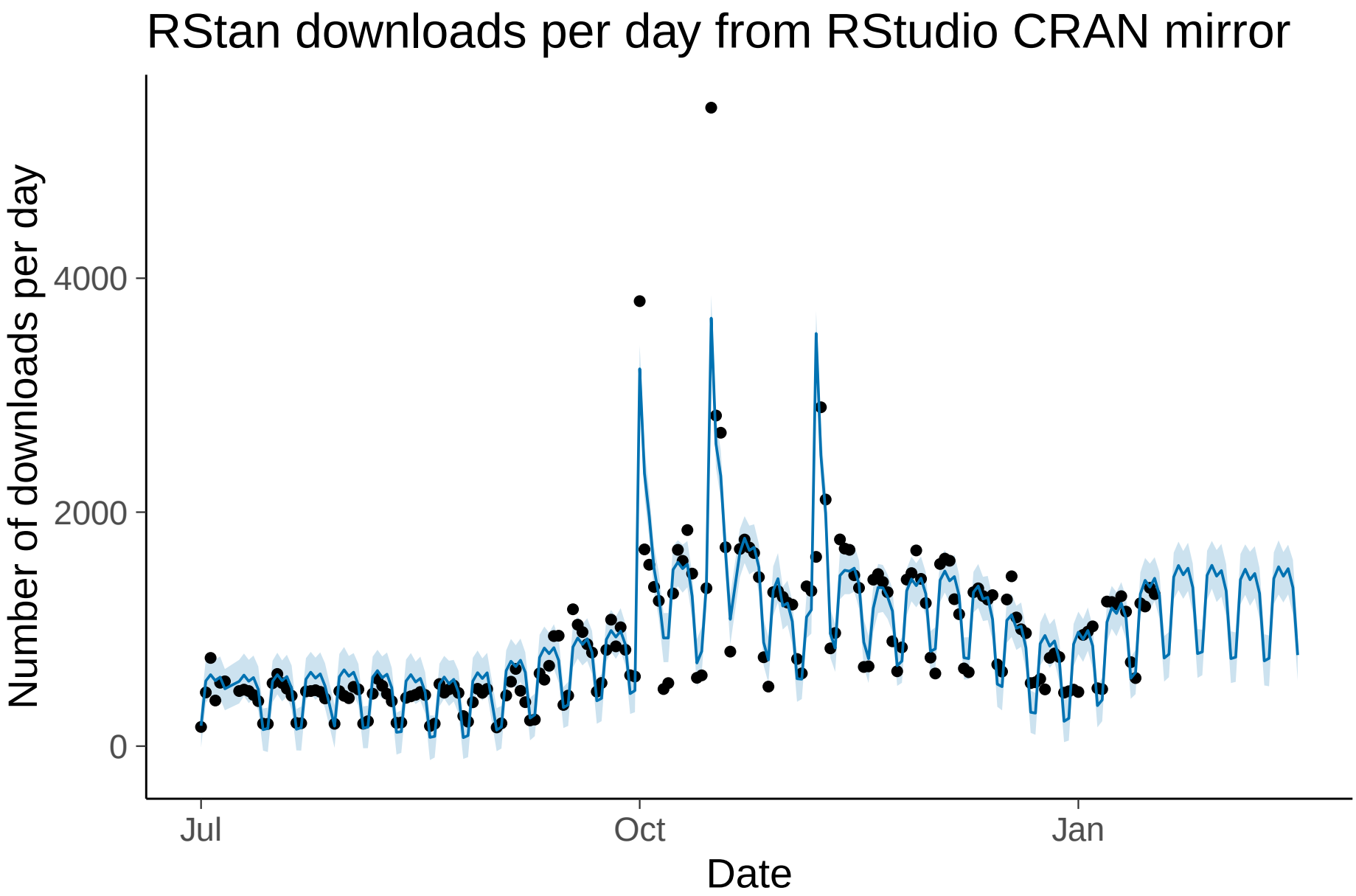




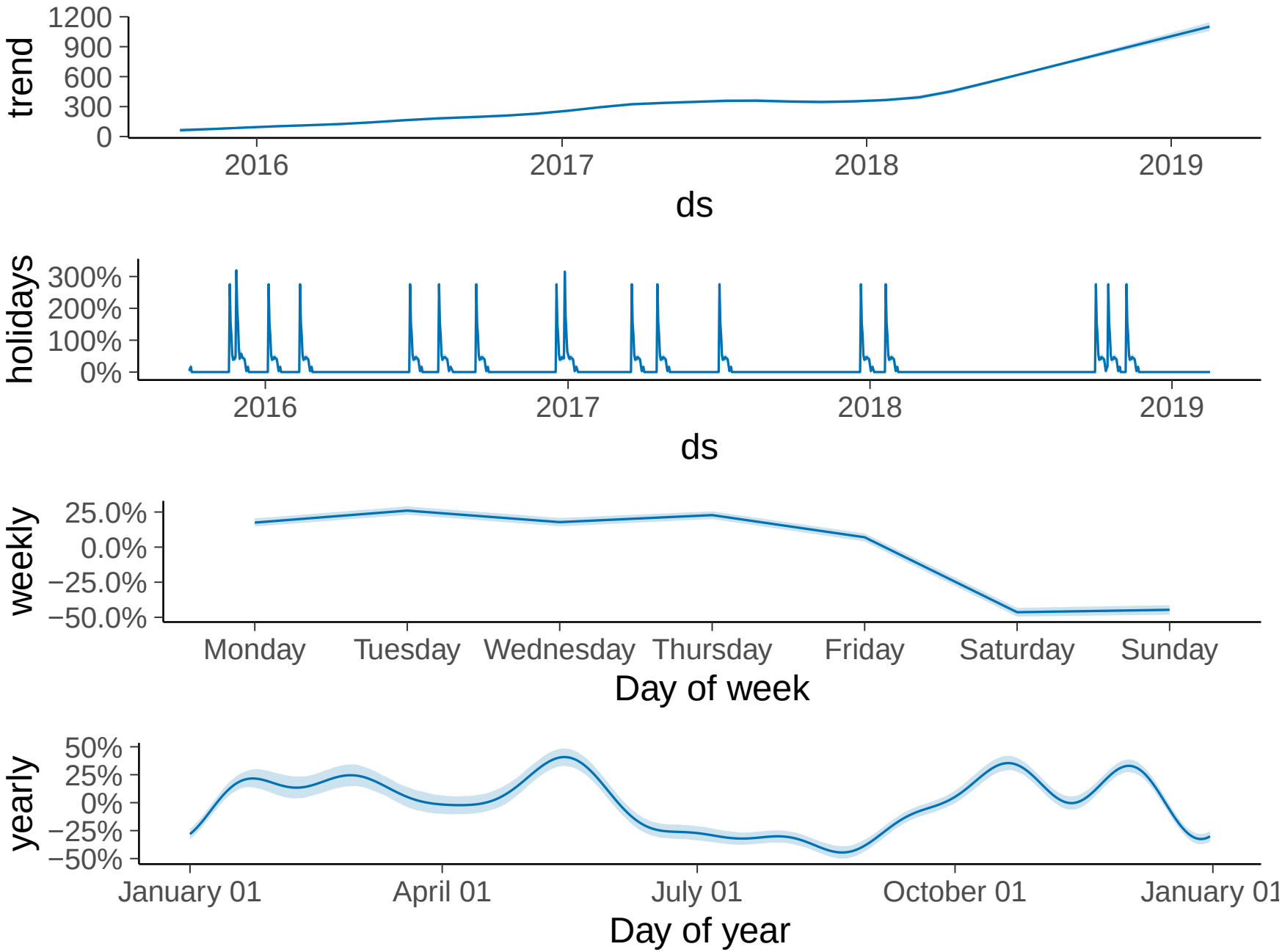
Structural information in predicting future



Structural information in predicting future

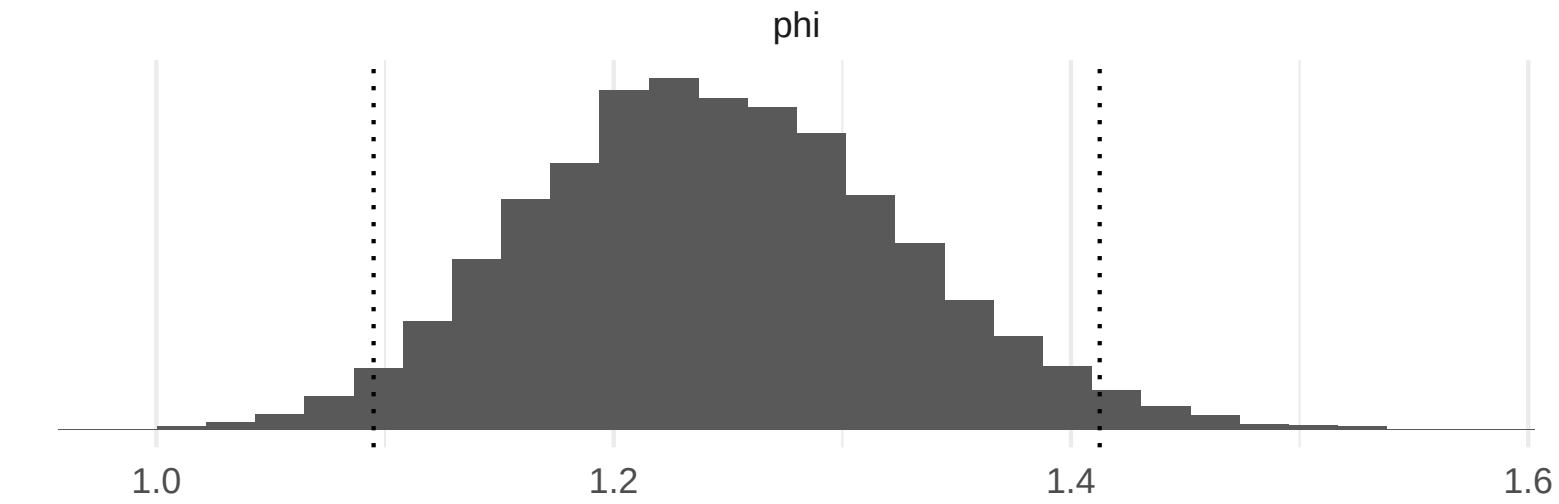


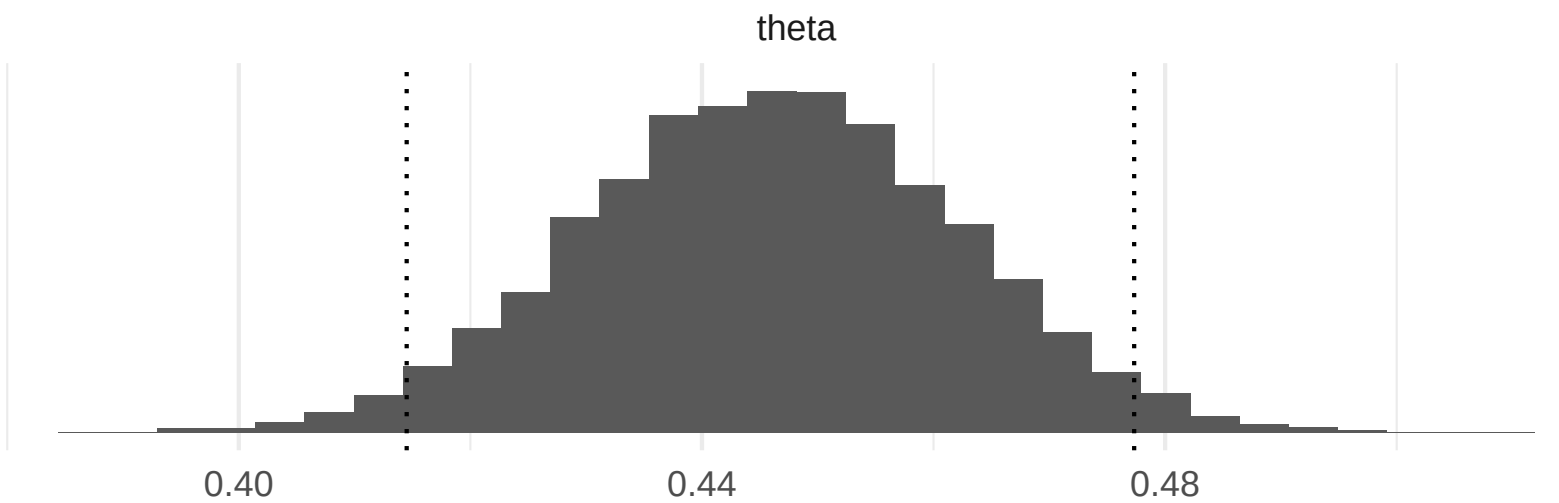
Structural information – Prophet by Facebook



Posterior visualization and inference

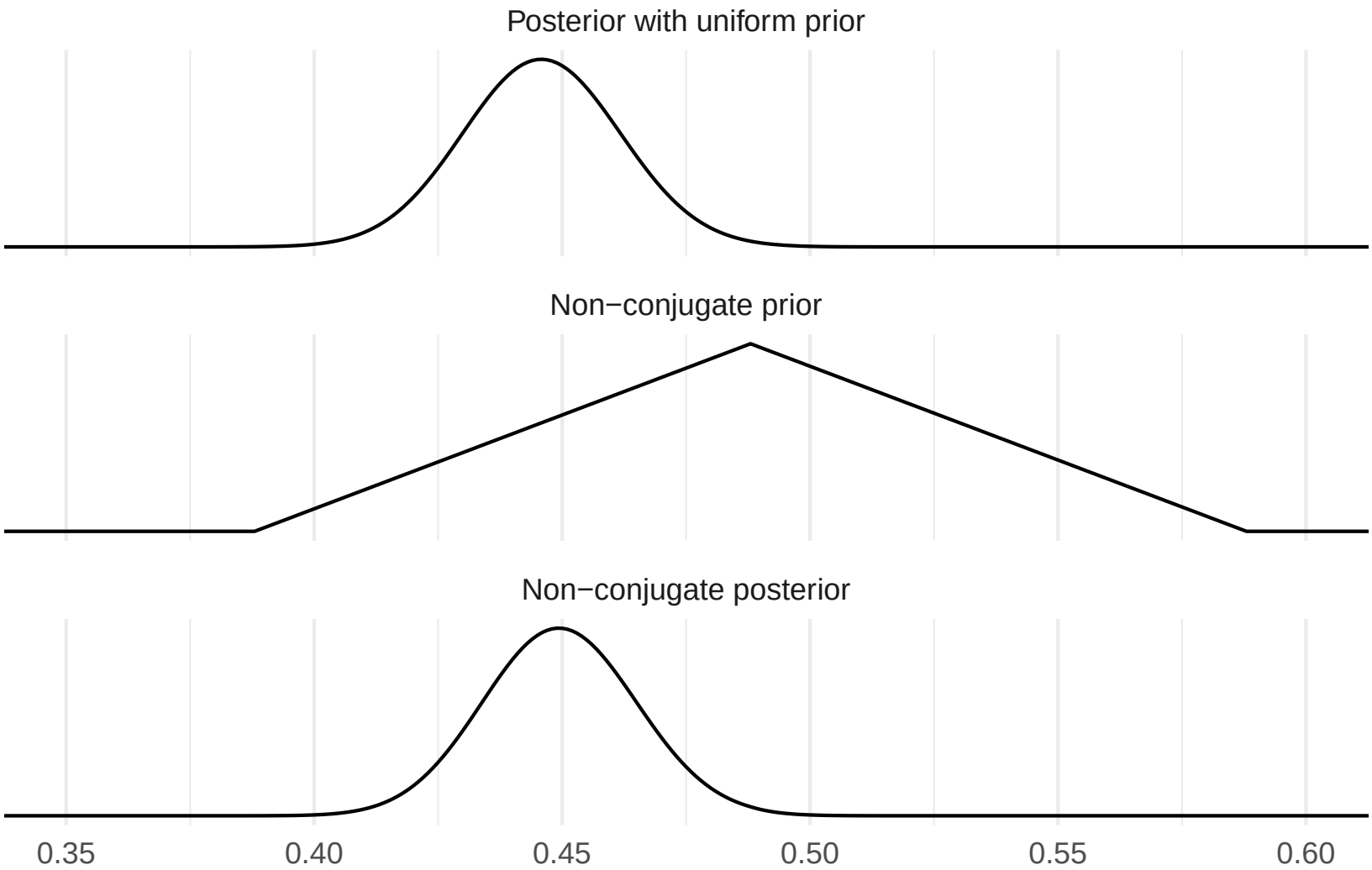
- Simulate samples from Beta(438,544), and draw a histogram of θ with quantiles.





Posterior visualization and inference

- Compute posterior distribution in a grid.



Central limit theorem

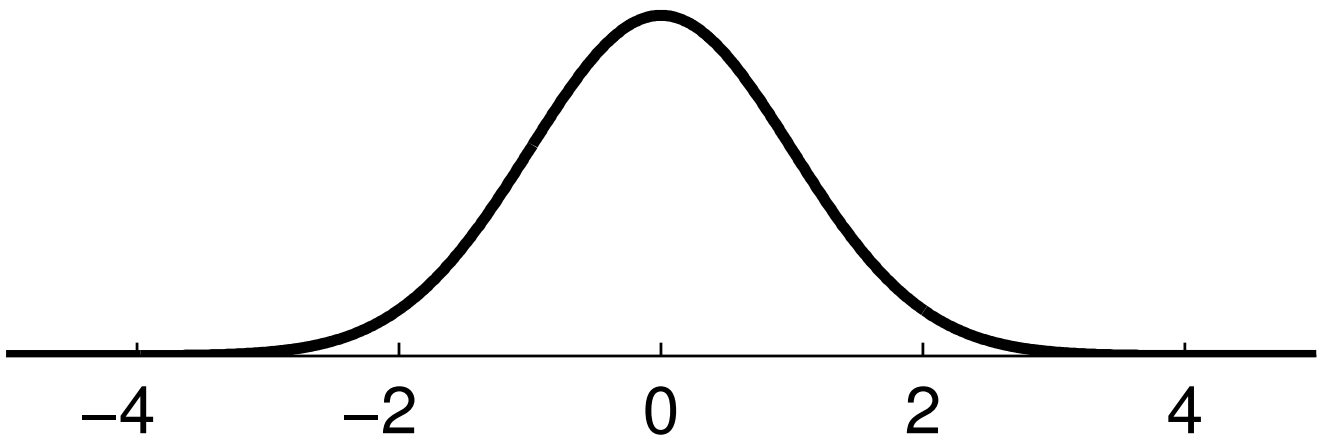
- The central limit theorem helps justify our usage of the normal distribution
- Important authors responsible for development of this theorem include De Moivre, Laplace, Gauss, Chebysev, Liapounov and Markov
- Given certain conditions the sum (and mean) of independent random variables approach a Gaussian distribution as $n \to \infty$ even if original variables are not normally distributed
- There are some problems
  - It does not hold for all distributions, e.g., Cauchy
  - This may require very large values of n. See the case of Binomial, when θ close to 0 or 1
  - Does not hold if one the variables has much larger scale

Normal / Gaussian

- Observations y real valued
- Mean θ and variance σ^2 (first assume σ^2 known)

$$p(y|\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(y-\theta)^2\right)$$

$$y \sim \mathrm{N}(\theta, \sigma^2)$$



Normal distribution - conjugate prior for θ

- Assume σ^2 known

Likelihood $$p(y|\theta) \propto \exp\left(-\frac{1}{2\sigma^2}(y-\theta)^2\right)$$

Prior $$p(\theta) \propto \exp\left(-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2\right)$$

Normal distribution - conjugate prior for θ

- Assume σ^2 known

Likelihood $$p(y|\theta) \propto \exp\left(-\frac{1}{2\sigma^2}(y-\theta)^2\right)$$

Prior $$p(\theta) \propto \exp\left(-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2\right)$$

$$\exp(a)\exp(b) = \exp(a+b)$$

Normal distribution - conjugate prior for θ

- Assume σ^2 known

Likelihood $$p(y|\theta) \propto \exp\left(-\frac{1}{2\sigma^2}(y-\theta)^2\right)$$

Prior $$p(\theta) \propto \exp\left(-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2\right)$$

$$\exp(a)\exp(b) = \exp(a+b)$$

Posterior $$p(\theta|y) \propto \exp\left(-\frac{1}{2}\left[\frac{(y-\theta)^2}{\sigma^2} + \frac{(\theta-\mu_0)^2}{\tau_0^2}\right]\right)$$

Normal distribution - conjugate prior for θ

- Posterior (see ex 2.14a)

$$p(\theta|y) \propto \exp\left(-\frac{1}{2}\left[\frac{(y-\theta)^2}{\sigma^2} + \frac{(\theta-\mu_0)^2}{\tau_0^2}\right]\right)$$
$$\propto \exp\left(-\frac{1}{2\tau_1^2}(\theta-\mu_1)^2\right)$$

$$\theta|y \sim \mathrm{N}(\mu_1, \tau_1^2), \quad \text{where} \quad \mu_1 = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{1}{\sigma^2}y}{\frac{1}{\tau_0^2} + \frac{1}{\sigma^2}} \quad \text{and} \quad \frac{1}{\tau_1^2} = \frac{1}{\tau_0^2} + \frac{1}{\sigma^2}$$

Normal distribution - conjugate prior for θ

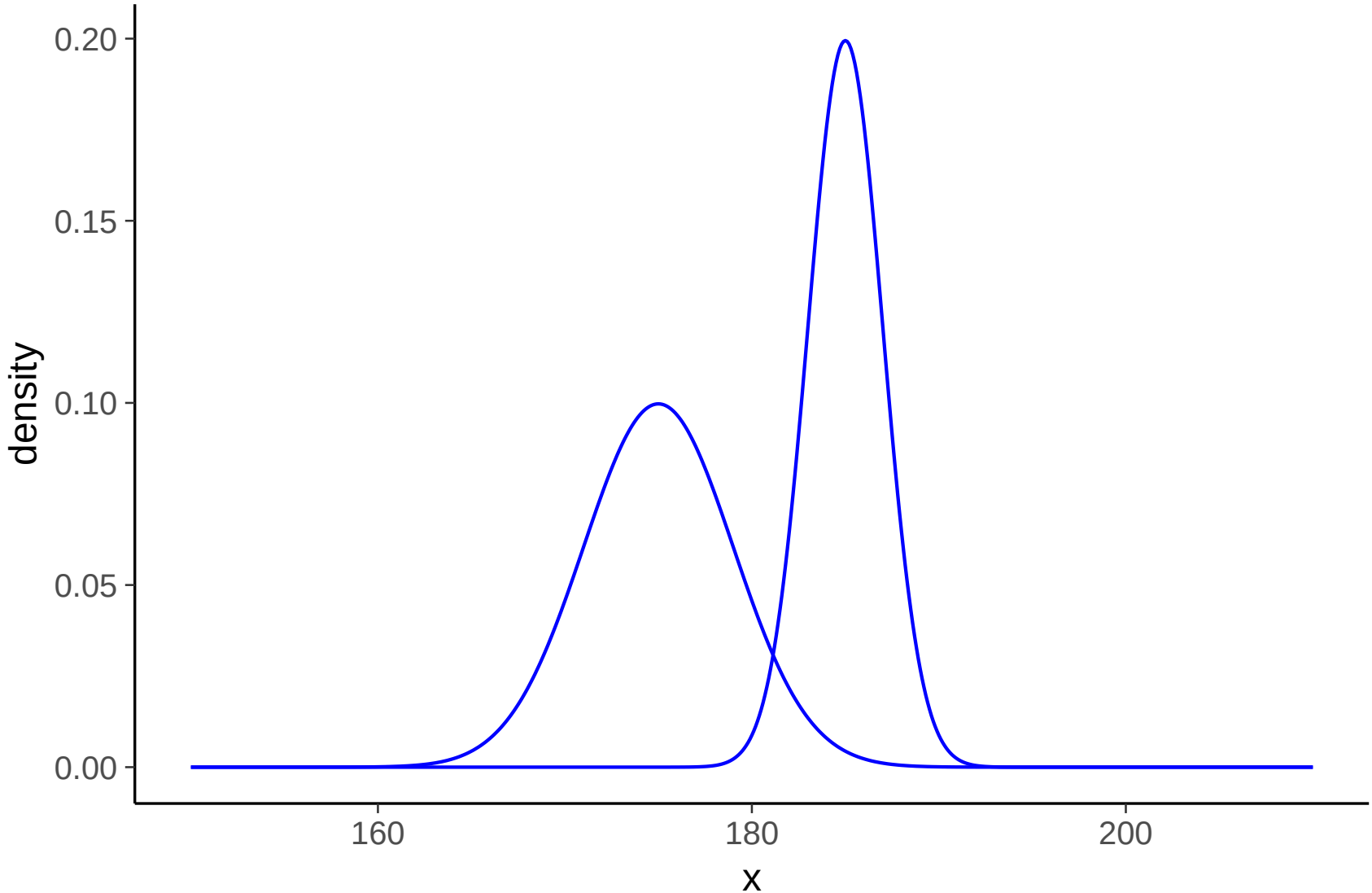
- Posterior (see ex 2.14a)

$$p(\theta|y) \propto \exp\left(-\frac{1}{2}\left[\frac{(y-\theta)^2}{\sigma^2} + \frac{(\theta-\mu_0)^2}{\tau_0^2}\right]\right)$$
$$\propto \exp\left(-\frac{1}{2\tau_1^2}(\theta-\mu_1)^2\right)$$

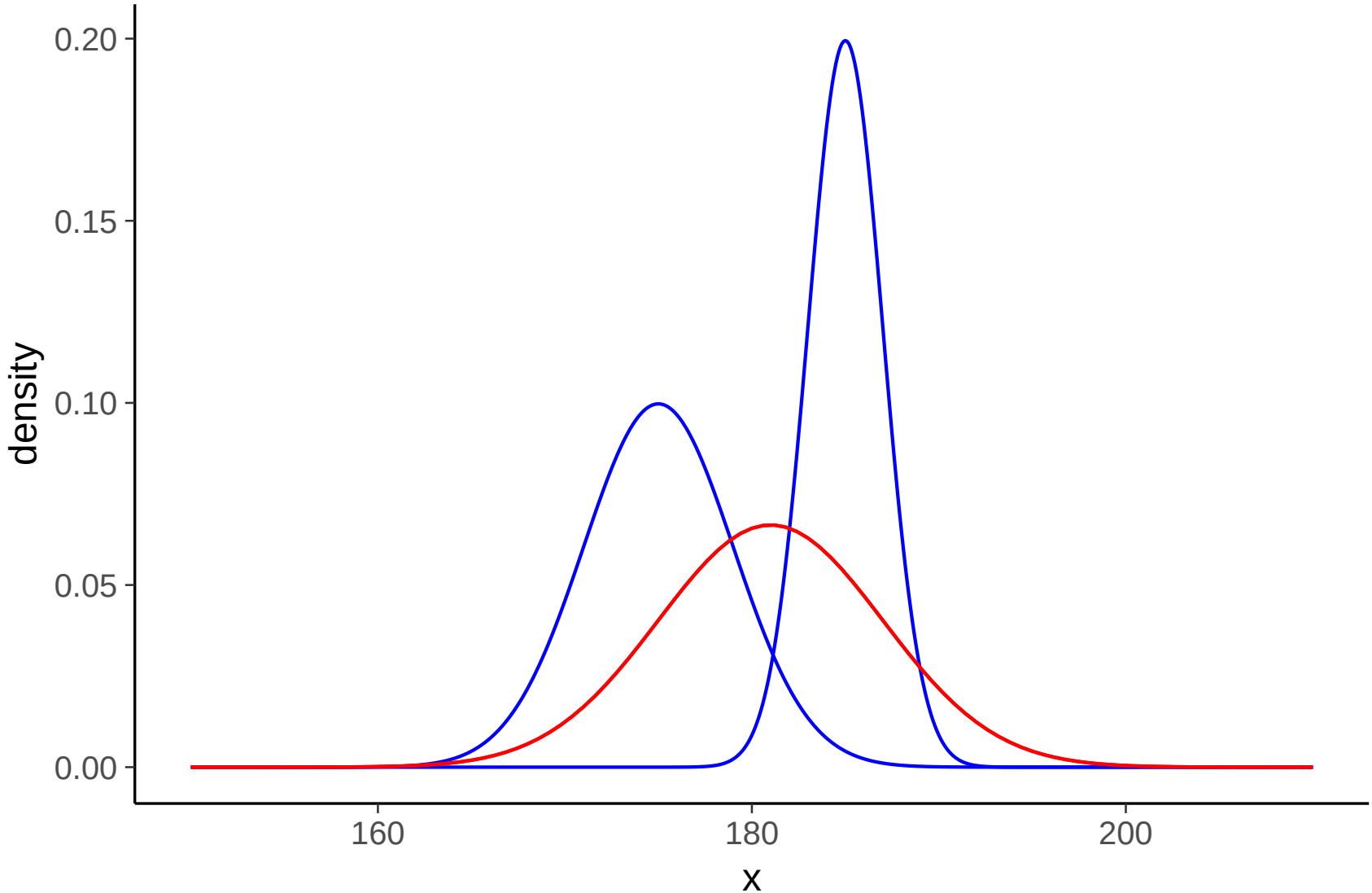
$$\theta|y \sim \mathrm{N}(\mu_1, \tau_1^2), \quad \text{where} \quad \mu_1 = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{1}{\sigma^2}y}{\frac{1}{\tau_0^2} + \frac{1}{\sigma^2}} \quad \text{and} \quad \frac{1}{\tau_1^2} = \frac{1}{\tau_0^2} + \frac{1}{\sigma^2}$$

- 1/variance = precision
- Posterior precision = prior precision + data precision
- Posterior mean is precision weighted mean

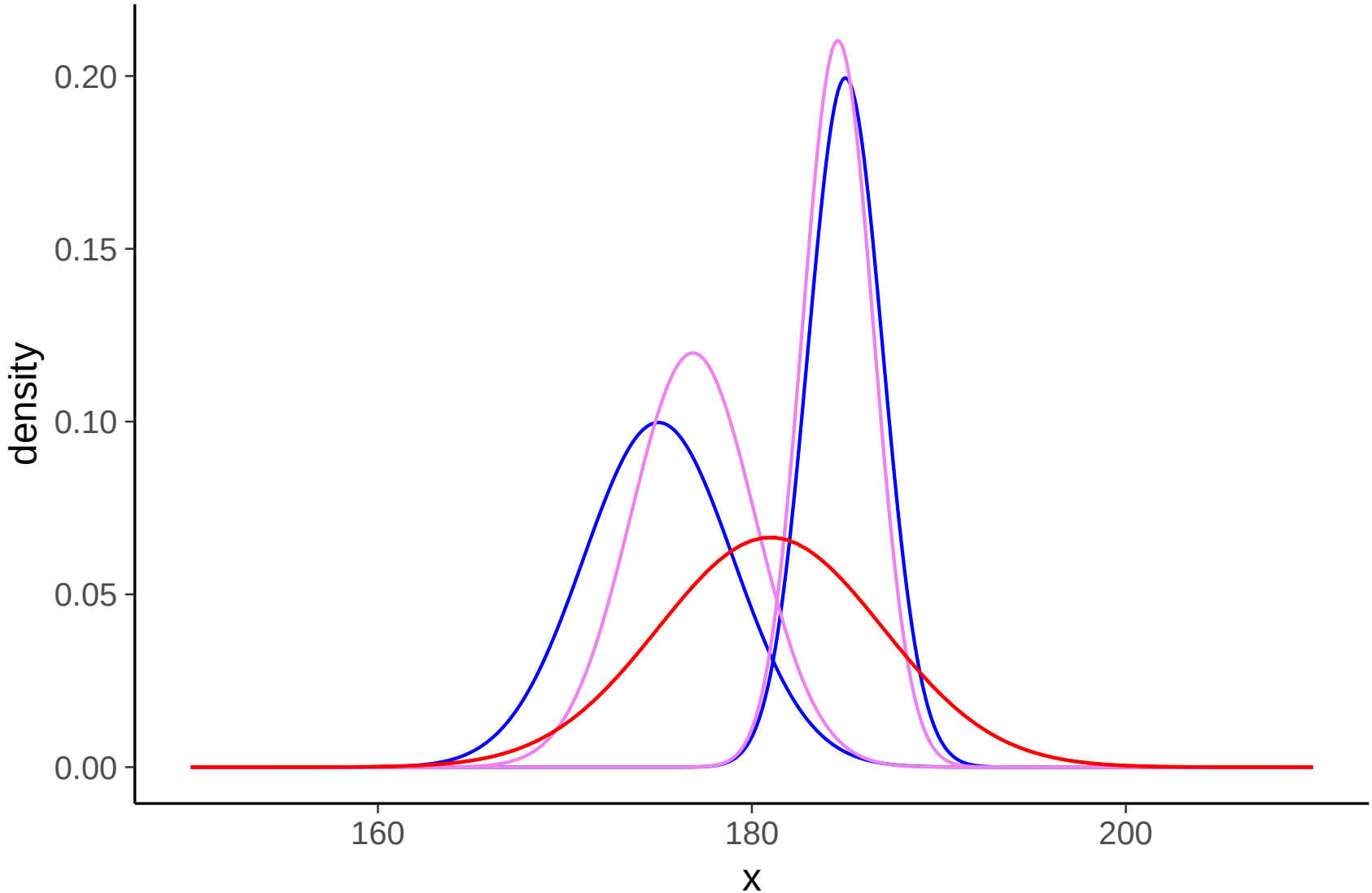
Normal distribution - example



Normal distribution - example



Normal distribution - example



Normal distribution - conjugate prior for θ

- Several observations $y = (y_1, \ldots, y_n)$

$$p(\theta|y) = \mathrm{N}(\theta|\mu_n, \tau_n^2)$$

$$\text{where} \quad \mu_n = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} \quad \text{and} \quad \frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}$$

- If $\tau_0^2 = \sigma^2$, prior corresponds to one virtual observation with value μ_0

Normal distribution - conjugate prior for θ

- Several observations $y = (y_1, \ldots, y_n)$

$$p(\theta|y) = \mathrm{N}(\theta|\mu_n, \tau_n^2)$$

$$\text{where} \quad \mu_n = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} \quad \text{and} \quad \frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}$$

- If $\tau_0^2 = \sigma^2$, prior corresponds to one virtual observation with value μ_0
- If $\tau_0 \to \infty$ when n fixed
  or if $n \to \infty$ when τ_0 fixed

$$p(\theta|y) \approx \mathrm{N}(\theta|\bar{y}, \sigma^2/n)$$