Module 1: Introduction to Bayesian Statistics

Rebecca C. Steorts

Agenda

- Motivations
- Traditional inference
- Bayesian inference
- Bernoulli, Beta
- Posterior of Beta-Bernoulli
- Conjugacy
- Example with 2012 election data
- Marginal likelihood
- Posterior Prediction

Traditional inference

You are given **data** X and there is an **unknown parameter** you wish to estimate θ

How would you estimate θ?

- Find an unbiased estimator of θ.
- Find the maximum likelihood estimate (MLE) of θ by looking at the likelihood of the data.
- If you cannot remember the definition of an unbiased estimator or the MLE, review these before our next class.

Bayesian inference

Bayesian methods trace its origin to the 18th century and English Reverend Thomas Bayes, who along with Pierre-Simon Laplace discovered what we now call **Bayes' Theorem**

- $p(x \mid \theta)$ likelihood
- $p(\theta)$ prior
- $p(\theta \mid x)$ posterior
- $p(x)$ marginal distribution

Derive the posterior distribution of $p(\theta \mid x)$.

Bernoulli distribution

The Bernoulli distribution is very common due to binary outcomes.

- Consider flipping a coin (heads or tails).
- We can represent this a binary random variable where the probability of heads is θ and the probability of tails is $1 - \theta$.

The write the random variable as $X \sim \text{Bernoulli}(\theta)\mathbb{1}(0 < \theta < 1)$

It follows that the likelihood is

$$p(x \mid \theta) = \theta^x (1 - \theta)^{(1-x)} \mathbb{1}(0 < \theta < 1).$$

- Exercise: what is the mean and the variance of X?

Bernoulli distribution

- Suppose that $X_1, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}(\theta)$. Then for $x_1, \ldots, x_n \in \{0, 1\}$ what is the likelihood?

Notation

- $\propto$: means "proportional to"
- $x_{1:n}$ denotes $x_1, \ldots, x_n$

Likelihood

$$
\begin{aligned}
p(x_{1:n}|\theta) &= \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n \mid \theta) \\
&= \prod_{i=1}^{n} \mathbb{P}(X_i = x_i \mid \theta) \\
&= \prod_{i=1}^{n} p(x_i|\theta) \\
&= \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} \\
&= \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}.
\end{aligned}
$$

Beta distribution

Given $a, b > 0$, we write $\theta \sim \text{Beta}(a, b)$ to mean that θ has pdf

$$p(\theta) = \text{Beta}(\theta|a, b) = \frac{1}{B(a, b)} \theta^{a-1}(1-\theta)^{b-1}\mathbb{1}(0 < \theta < 1),$$

i.e., $p(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$ on the interval from 0 to 1.

- Here,
$$B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$
.
- The mean is $E(\theta) = \int \theta\, p(\theta) d\theta = a/(a+b)$.

Posterior of Bernoulli-Beta

Lets derive the posterior of $\theta \mid x_{1:n}$

$$
\begin{aligned}
& p(\theta|x_{1:n}) \\
& \propto p(x_{1:n}|\theta)p(\theta) \\
& = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}\frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}I(0<\theta<1) \\
& \propto \theta^{a+\sum x_i-1}(1-\theta)^{b+n-\sum x_i-1}I(0<\theta<1) \\
& \propto Beta\big(\theta \mid a+\sum x_i,\ b+n-\sum x_i\big).
\end{aligned}
$$

Conjugacy

What do you notice about the prior and the posterior from the Bernoulli-Beta example that we just considered?

Conjugacy

If the posterior distribution comes from the same family of distributions as the prior, we say that the prior and posterior are conjugate distributions.

More formally, the prior is called a conjugate family for the likelihood funtion.

Example

- The Gaussian family is conjugate to itself with respect to a Gaussian likelihood.
- That is, if the likelihood function is Gaussian, choosing a Gaussian prior over the mean will ensure that the posterior distribution is also Gaussian.
- We will return to conjugacy later.
- First, we will look at a simple example to illustrate the idea of the bernoulli-beta.

Approval ratings of Obama

What is the proportion of people that approve of President Obama in PA?

- We take a random sample of 10 people in PA and find that 6 approve of President Obama.
- The national approval rating (Zogby poll) of President Obama in mid-September 2015 was 45%. We'll assume that in PA his approval rating is approximately 50%.
- Based on this prior information, we'll use a Beta prior for θ and we'll choose a and b.

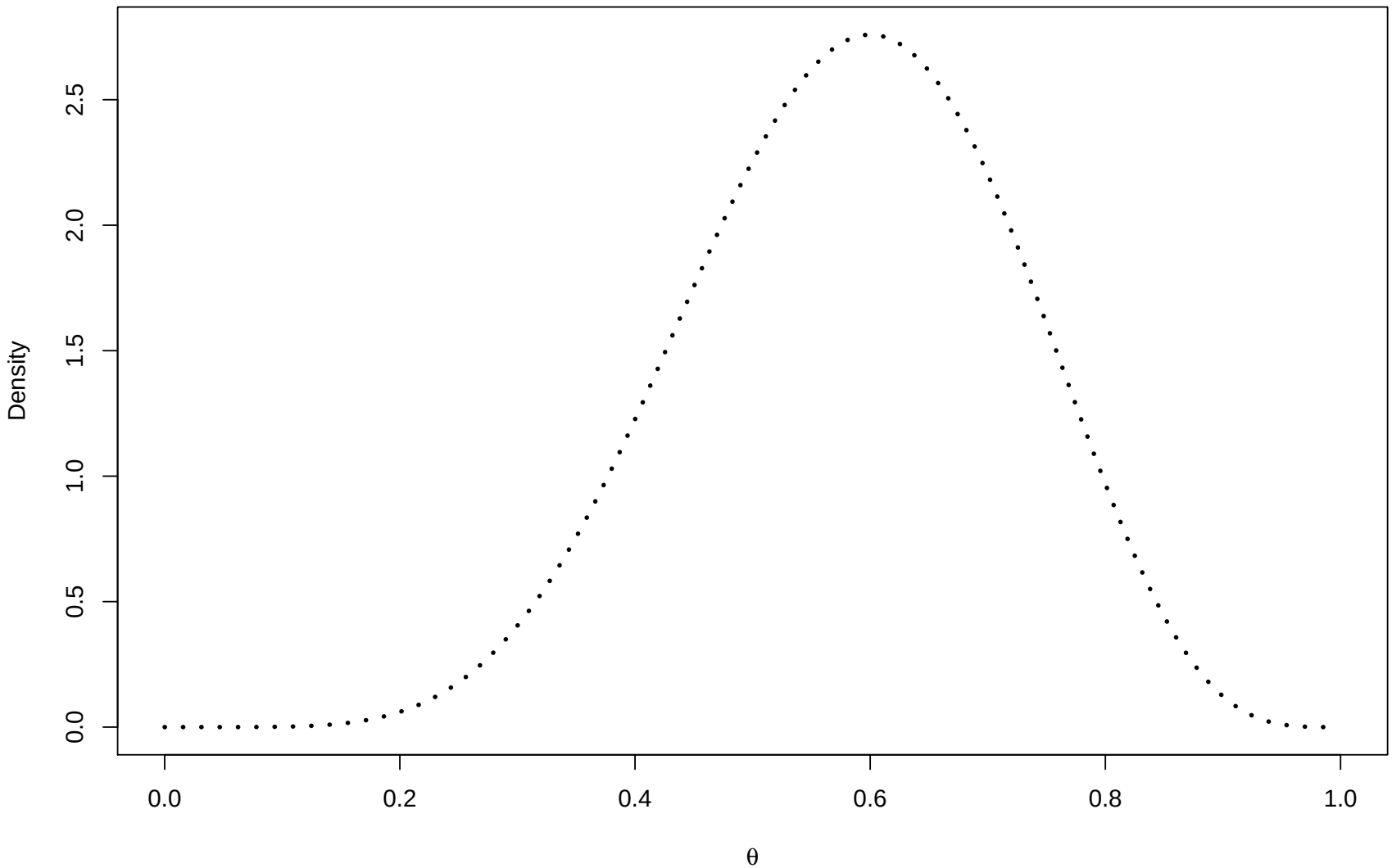
Obama Example

```
n = 10
# Fixing values of a,b.
a = 21/8
b = 0.04
th = seq(0,1, length=500)
x = 6

# we set the likelihood, prior, and posteriors with
# THETA as the sequence that we plot on the x-axis.
# Beta(c,d) refers to shape parameter
like = dbeta(th, x+1, n-x+1)
prior = dbeta(th, a, b)
post = dbeta(th, x+a, n-x+b)
```

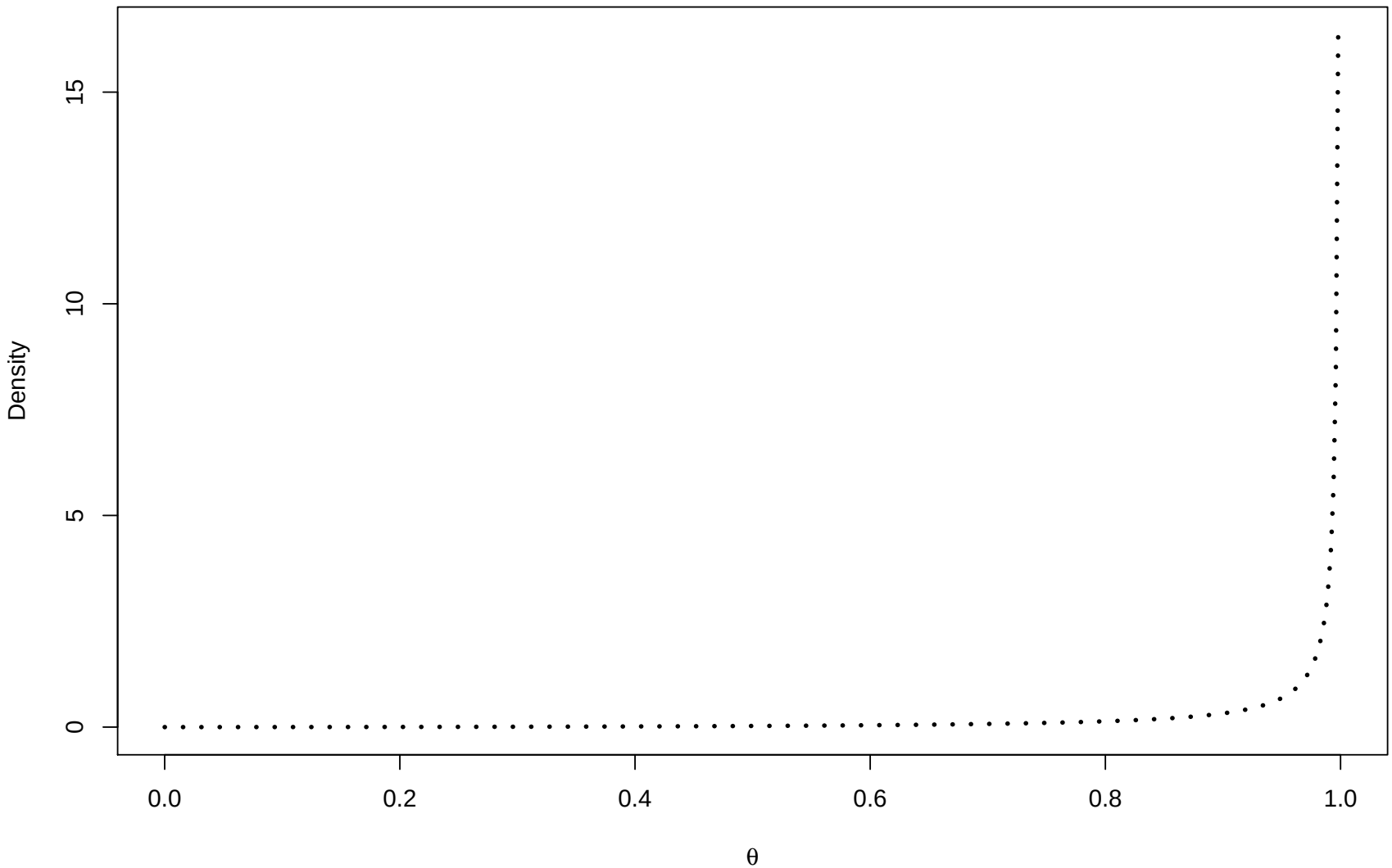
Likelihood

```
plot(th, like, type='l', ylab = "Density",
     lty = 3, lwd = 3, xlab = expression(theta))
```



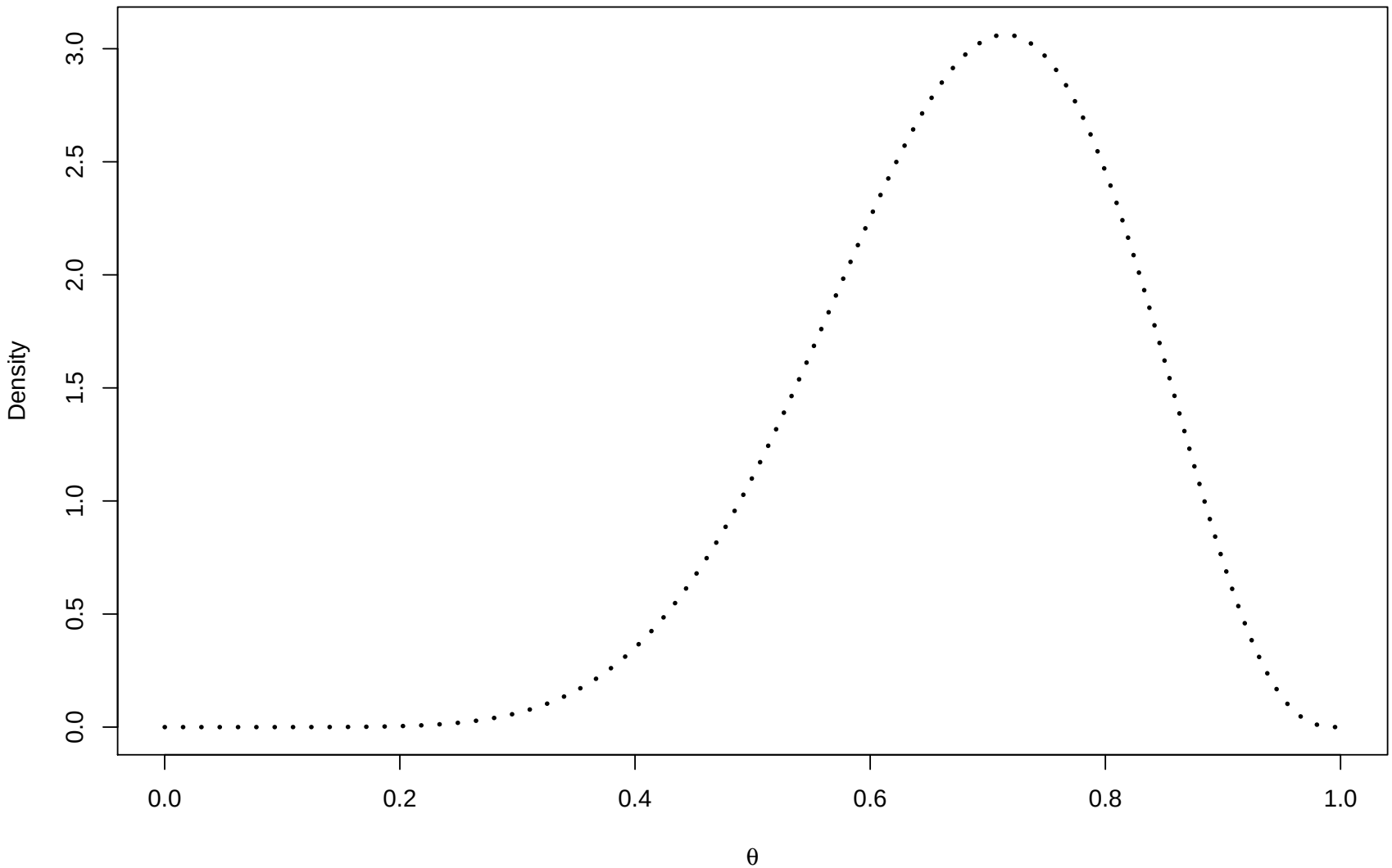
Prior

```
plot(th, prior, type='l', ylab = "Density",
     lty = 3, lwd = 3, xlab = expression(theta))
```

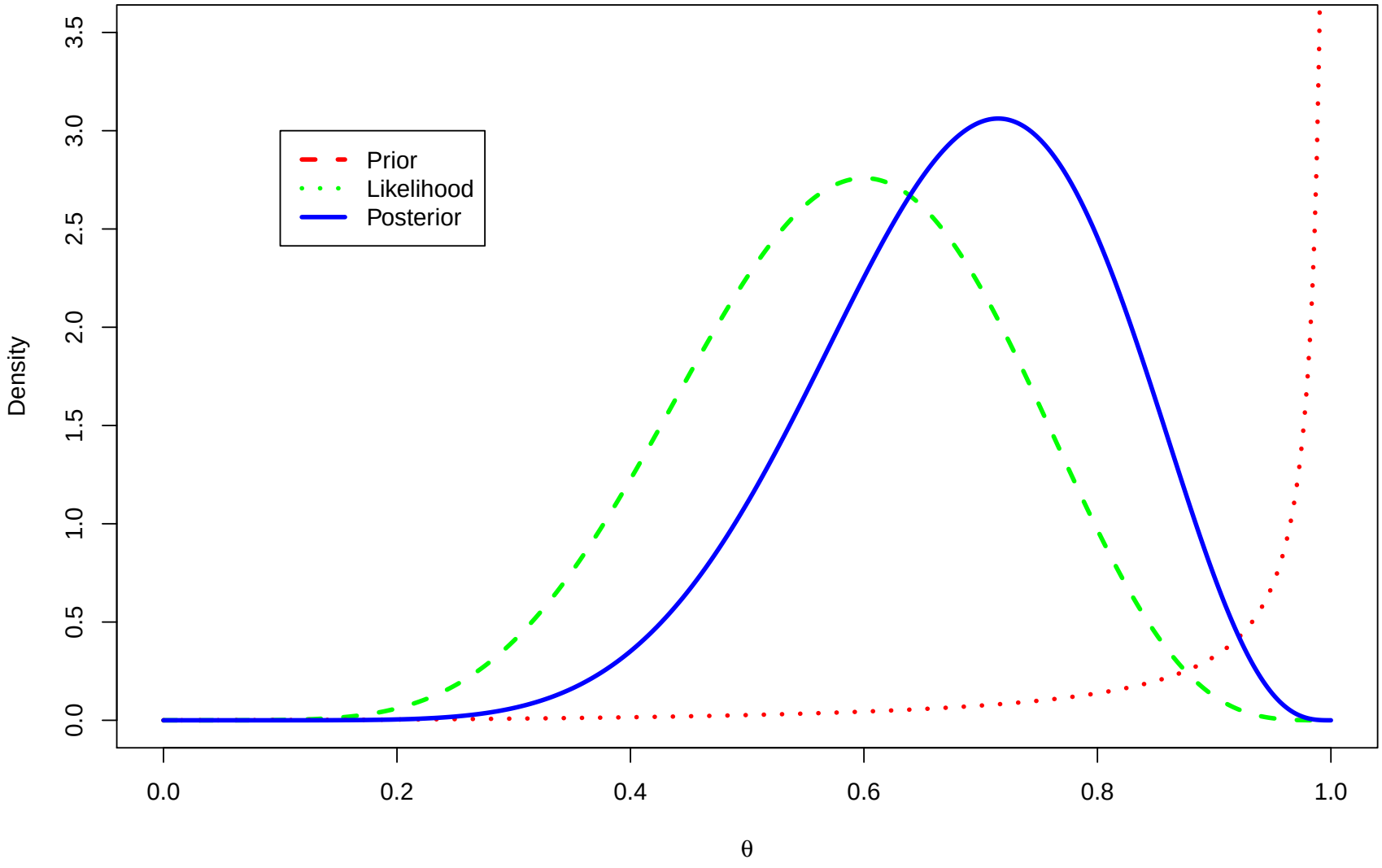


Posterior

```
plot(th, post, type='l', ylab = "Density",
     lty = 3, lwd = 3, xlab = expression(theta))
```



Likelihood, Prior, and Posterior



Cast of characters

- Observed data: x
- Note this could consist of many data points, e.g., $x = x_{1:n} = (x_1, \ldots, x_n)$.

likelihood	$p(x\|\theta)$
prior	$p(\theta)$
posterior	$p(\theta\|x)$
marginal likelihood	$p(x)$
posterior predictive	$p(x_{n+1}\|x_{1:n})$
loss function	$\ell(s, a)$
posterior expected loss	$\rho(a, x)$
risk / frequentist risk	$R(\theta, \delta)$
integrated risk	$r(\delta)$

Marginal likelihood

The **marginal likelihood** is

$$p(x) = \int p(x|\theta)p(\theta)\, d\theta$$

- What is the marginal likelihood for the Bernoulli-Beta?

Posterior predictive distribution

- We may wish to predict a new data point x_{n+1}
- We assume that $x_{1:(n+1)}$ are independent given θ

$$
\begin{aligned}
p(x_{n+1}|x_{1:n}) &= \int p(x_{n+1}, \theta|x_{1:n})\, d\theta \\
&= \int p(x_{n+1}|\theta, x_{1:n}) p(\theta|x_{1:n})\, d\theta \\
&= \int p(x_{n+1}|\theta) p(\theta|x_{1:n})\, d\theta.
\end{aligned}
$$

Example: Back to the Beta-Bernoulli

Suppose

$$\theta \sim Beta(a, b)$$

and

$$X_1, \ldots, X_n \mid \theta \overset{iid}{\sim} Bernoulli(\theta)$$

Then the marginal likelihood is

$$\begin{aligned}
&p(x_{1:n}) \\
&= \int p(x_{1:n}|\theta)p(\theta)\, d\theta \\
&= \int_0^1 \theta^{\sum x_i}(1-\theta)^{n-\sum x_i} \frac{1}{B(a,b)} \theta^{a-1}(1-\theta)^{b-1} d\theta \\
&= \frac{B\left(a + \sum x_i,\ b + n - \sum x_i\right)}{B(a,b)},
\end{aligned}$$

by the integral definition of the Beta function.

Posterior predictive distribution

Let $a_n = a + \sum x_i$ and $b_n = b + n - \sum x_i$.

Recall that the posterior distribution is $p(\theta|x_{1:n}) = \text{Beta}(\theta|a_n, b_n)$.

Let's derive the posterior preditive distribution.

Posterior predictive distribution

The posterior predictive can be derived to be

$$\begin{aligned}\mathbb{P}(X_{n+1} = 1 \mid x_{1:n}) &= \int \mathbb{P}(X_{n+1} = 1 \mid \theta) p(\theta | x_{1:n}) d\theta \\ &= \int \theta \ \text{Beta}(\theta | a_n, b_n) = \frac{a_n}{a_n + b_n},\end{aligned}$$

Similarly,

$$\mathbb{P}(X_{n+1} = 0 \mid x_{1:n}) = 1 - \mathbb{P}(X_{n+1} = 1 \mid x_{1:n}) = \frac{b_n}{a_n + b_n}$$

hence, the posterior predictive p.m.f. is

$$p(x_{n+1} | x_{1:n}) = \frac{a_n^{x_{n+1}} b_n^{1-x_{n+1}}}{a_n + b_n} \mathbb{1}(x_{n+1} \in \{0, 1\}).$$

Summary

- We covered the "cast of characters" needed to work with Bayesian models
- These include the likelihood, prior, posterior, marginal likelihood, and posterior predictive distribution
- We derived Bayes' Theorem
- Bernoulli-Beta
- Conjugacy

Exercise

We write $X \sim \text{Poisson}(\theta)$ if X has the Poisson distribution with rate $\theta > 0$, that is, its p.m.f. is

$$p(x|\theta) = \text{Poisson}(x|\theta) = e^{-\theta}\theta^x/x!$$

for $x \in \{0, 1, 2, \dots\}$ (and is 0 otherwise). Suppose $X_1, \dots, X_n \overset{iid}{\sim} \text{Poisson}(\theta)$ given θ, and your prior is

$$p(\theta) = \text{Gamma}(\theta|a, b) = \frac{b^a}{\Gamma(a)}\theta^{a-1}e^{-b\theta}\mathbb{1}(\theta > 0).$$

What is the posterior distribution on θ?