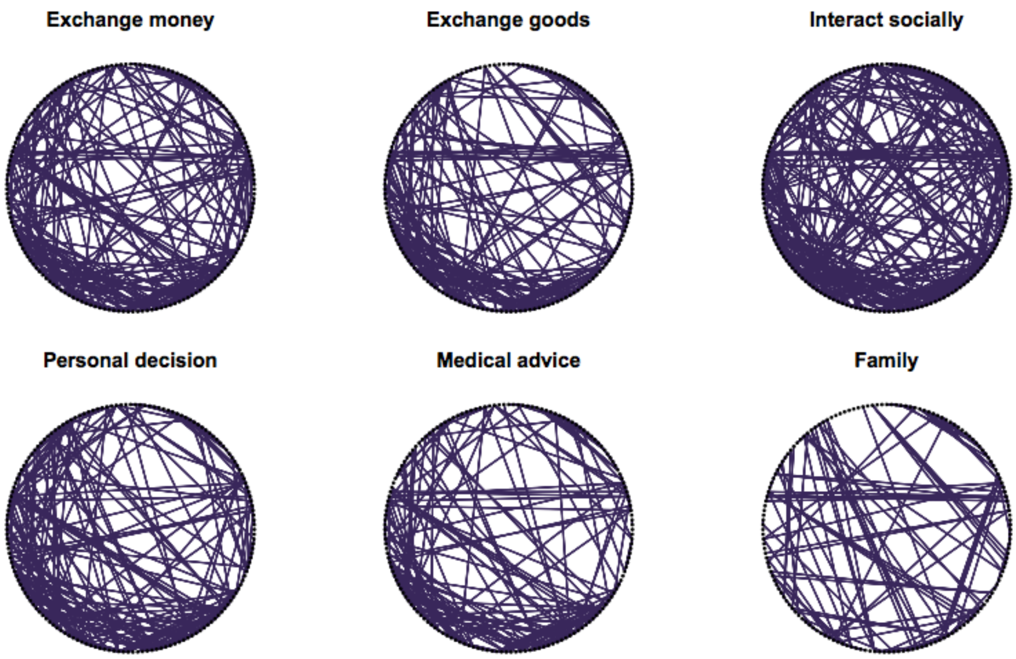
Module 1: Introduction to Bayesian Statistics, Part I

Rebecca C. Steorts

Agenda

- Motivations
- Traditional inference
- Bayesian inference
- Bernoulli, Beta
- Connection to the Binomial distribution
- Posterior of Beta-Bernoulli
- Example with 2012 election data
- Marginal likelihood
- Posterior Prediction

Social networks



Precision Medicine

Editor: **Daniel B. Neill, H.J. Heinz III College,** Carnegie Mellon University, neill@cs.cmu.edu

A $3 Trillion Challenge to Computational Scientists: Transforming Healthcare Delivery

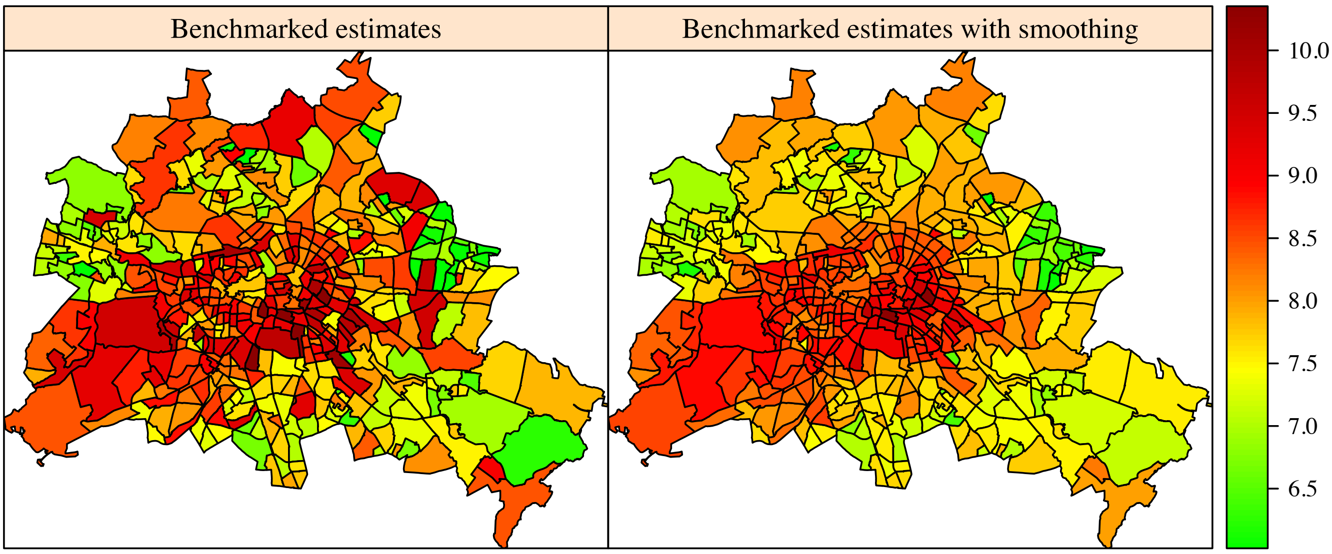
Suchi Saria, *Johns Hopkins University*

Healthcare spending in the US is nearing $3 trillion per year, but in spite of this expenditure, the US is outpaced by most developed countries in terms of health and quality of life outcomes—for example, it ranks 36th internationally in life expectancy.[1] The share of health spending in its gross domestic product has increased sharply, from 5 percent of GDP in 1960 to more than 17 percent today,[2] a rate of increase that's widely believed to be unsustainable.[3]

Policy and regulatory reform have important roles to play in addressing these challenges. Yet one of the largest underexplored avenues is the better use of information derived from the vast amount of health data now being collected in digital format.[4] I believe that one of the most significant open fron- paper records that weren't amenable to retrospective, automated analyses. The Health Information Technology for Economic and Clinical Health (HITECH) Act, a program that was part of the American Recovery and Reinvestment Act of 2009, incentivized the adoption of Electronic Health Records (EHRs) to encourage the shift from paper to digital records. That program has made more than $15.5 billion available to hospitals and healthcare professionals conditioned on their meeting certain EHR benchmarks for so-called "meaningful use." It's one of the largest investments in healthcare infrastructure ever made by the federal government.

A survey by the American Hospital Association showed that adoption of EHRs has doubled from 2009 to 2011. Today, much of an individual's health

Estimating Rent Prices in Small Domains



Traditional inference

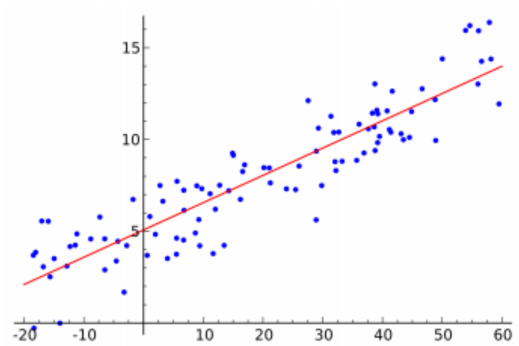
You are given **data** X and there is an **unknown parameter** you wish to estimate θ
How would you estimate θ?

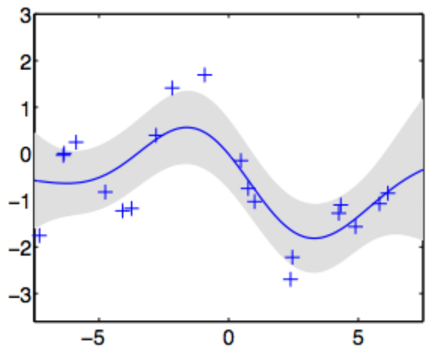
- Find an unbiased estimator of θ.
- Find the maximum likelihood estimate (MLE) of θ by looking at the likelihood of the data.
- If you cannot remember the definition of an unbiased estimator or the MLE, review these before our next class.

Bayesian Motivation

Parameters

$$P(X|\theta) \quad = \quad \text{Probability}[\text{data}|\text{pattern}]$$





Inference idea

data = underlying pattern + independent noise

[credit: Peter Orbanz, Columbia University]

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

$$p(\theta|x) = \frac{p(\theta, x)}{p(x)} = \frac{p(x|\theta)p(\theta)}{p(x)} \propto p(x|\theta)p(\theta)$$

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 $X_1, \dots, X_n \overset{iid}{\sim} \text{Bernoulli}(\theta)$. Then for $x_1, \dots, x_n \in \{0, 1\}$ 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 $\theta \mid x_{1:n}$

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

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.

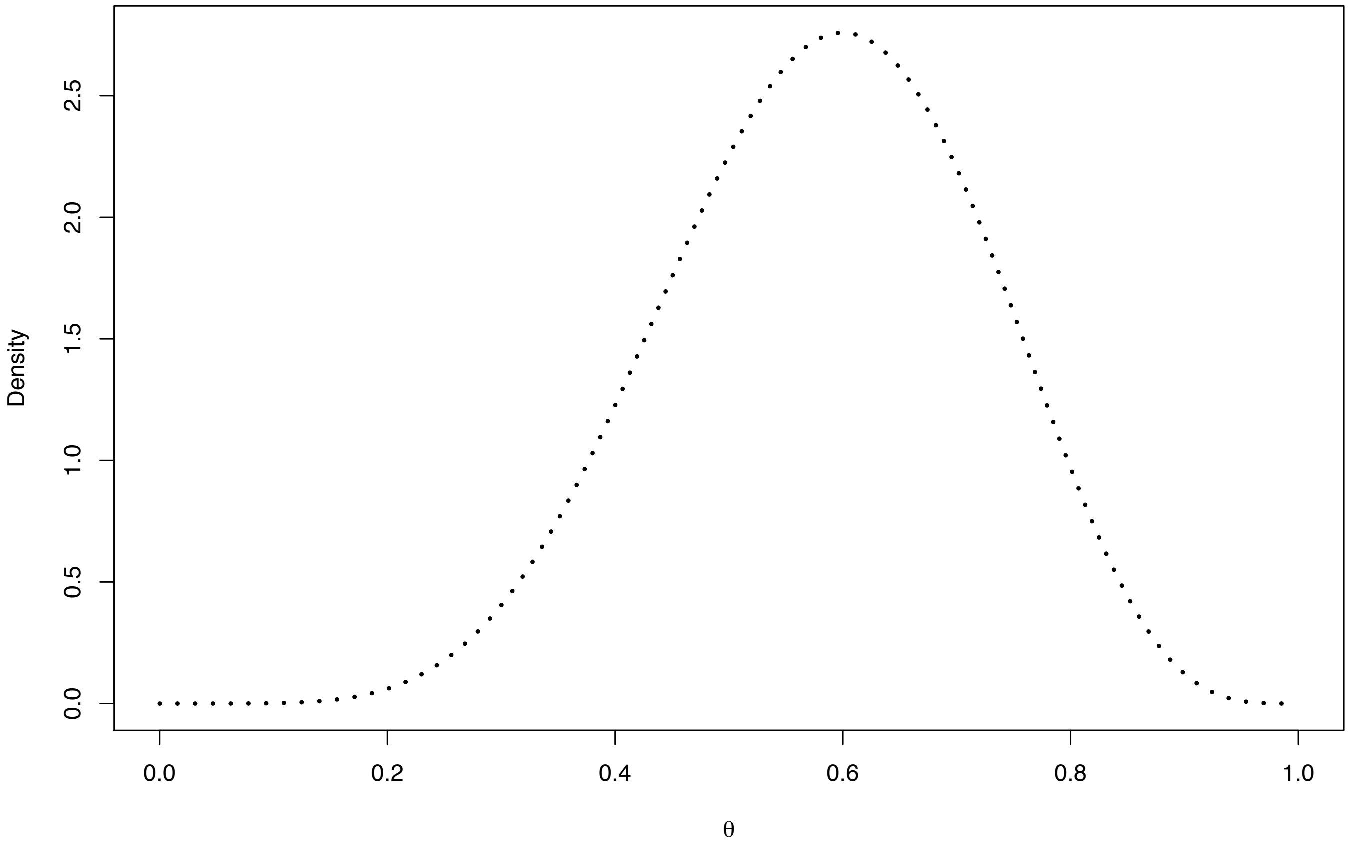
Obama Example

```
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)
```

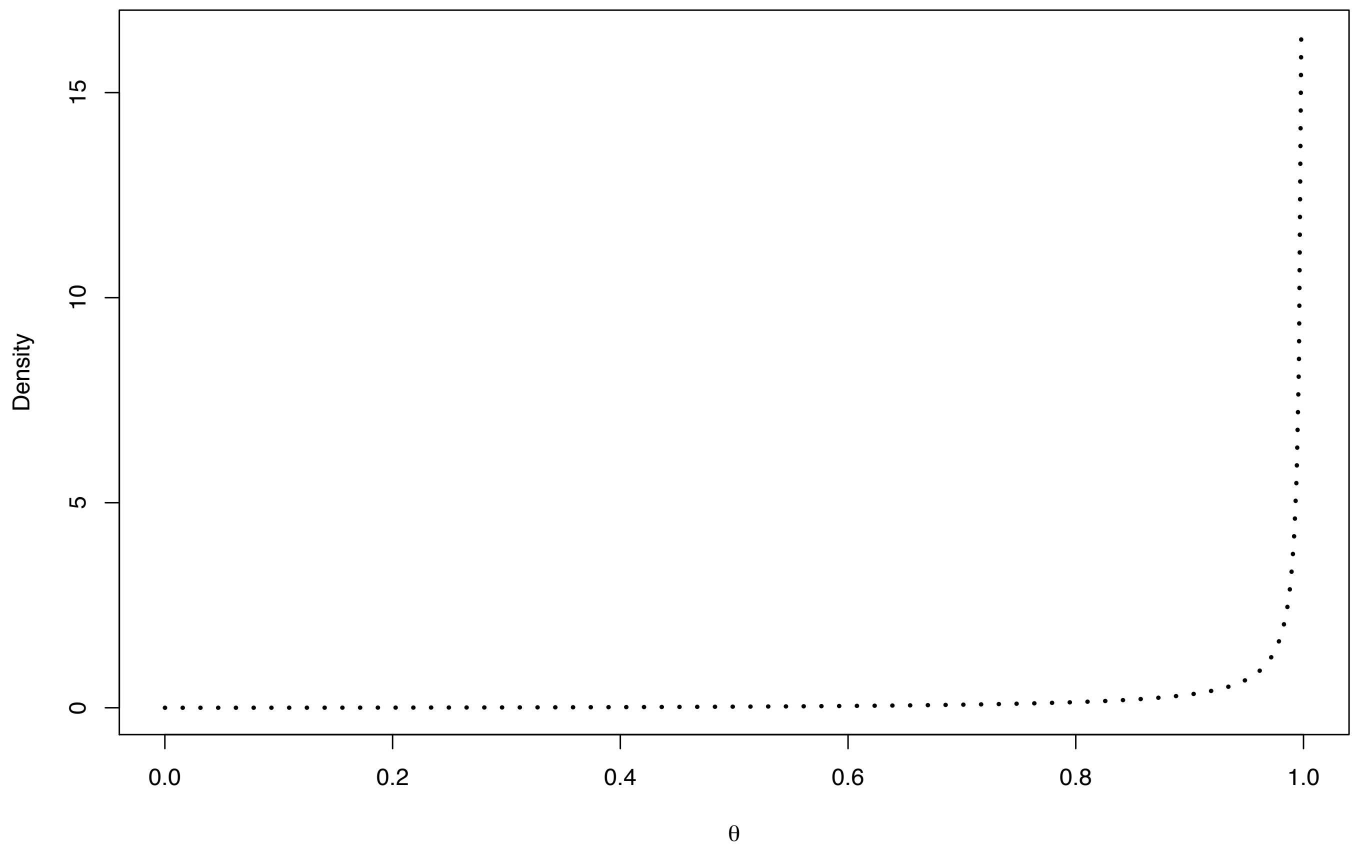
Likelihood

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



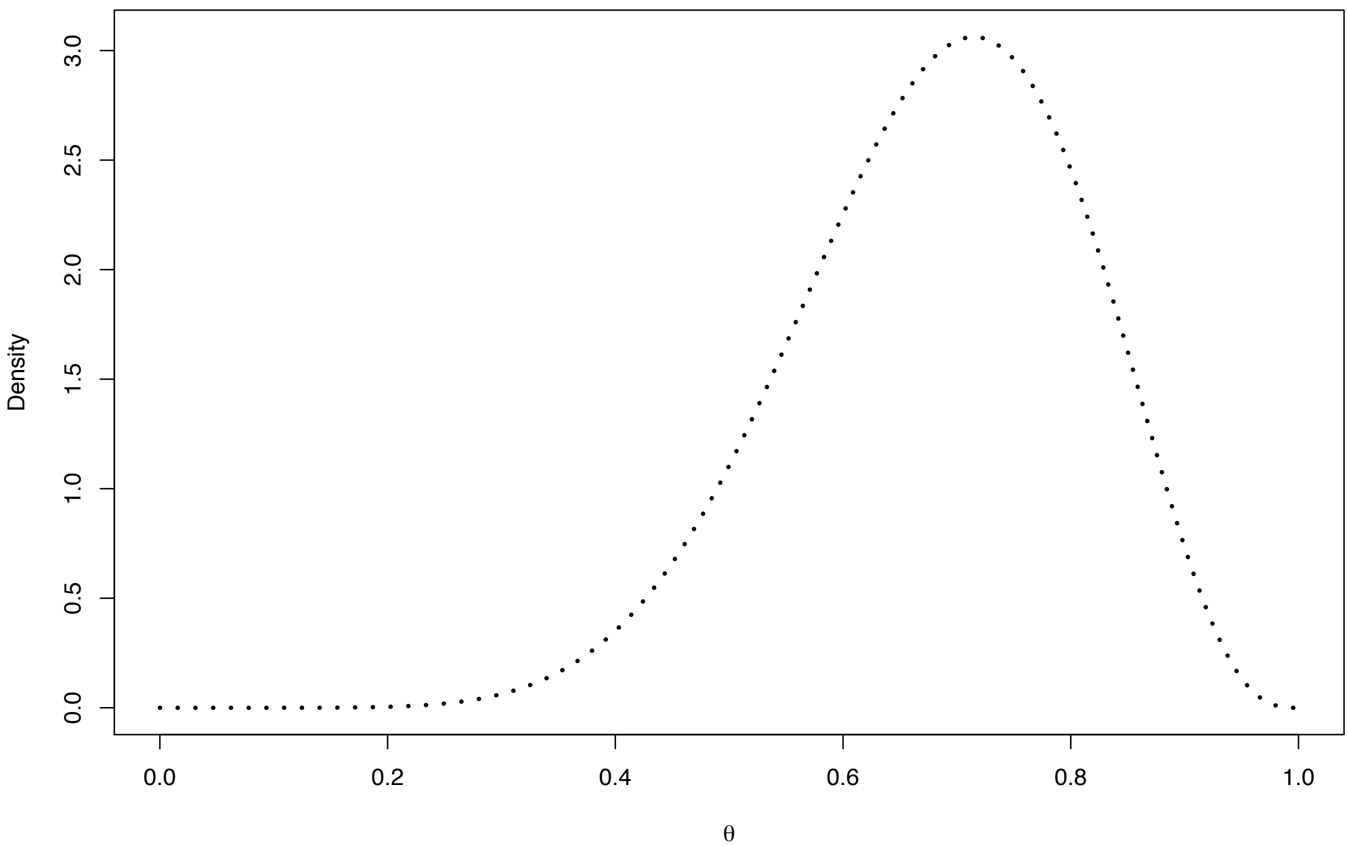
Prior

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

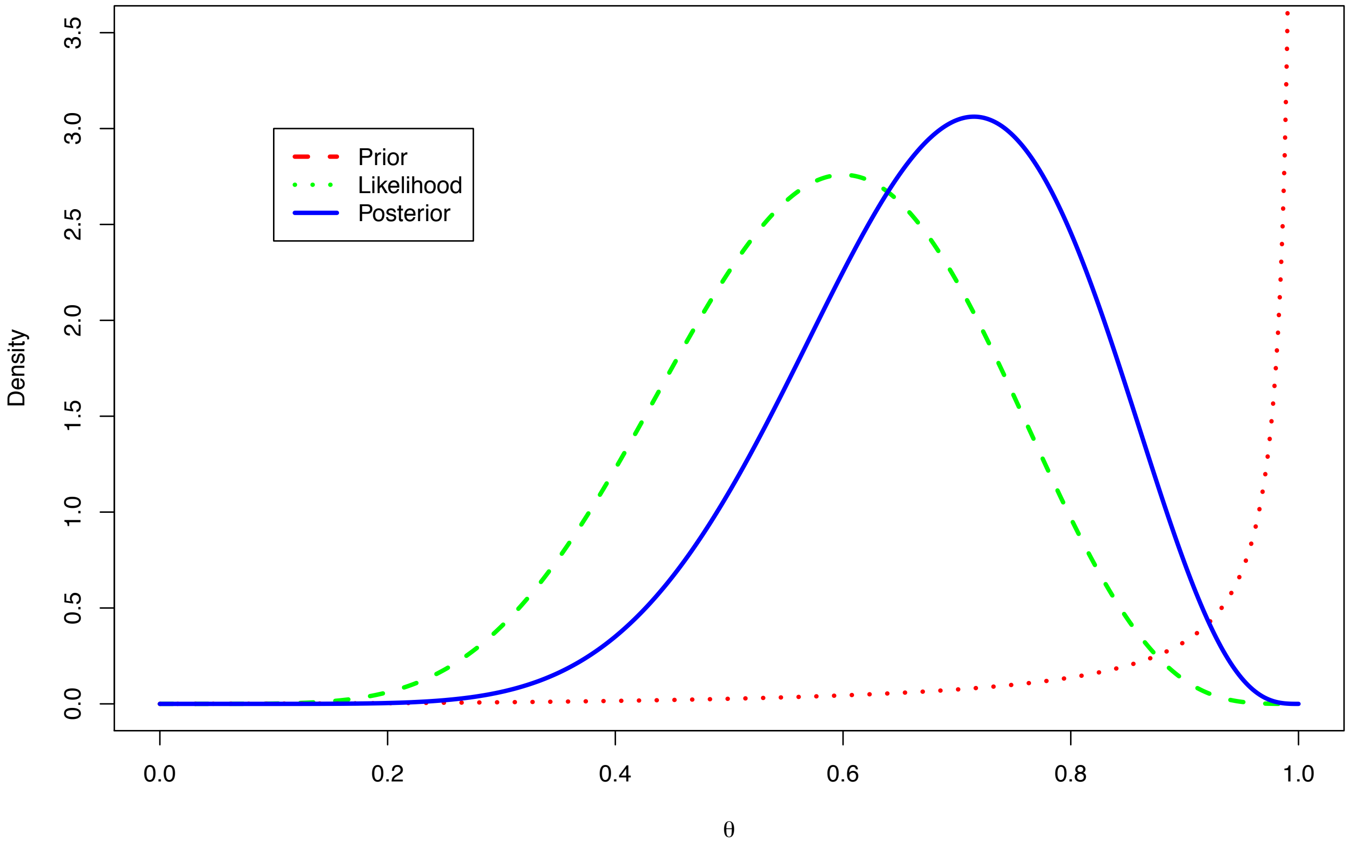


Posterior

```
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.

Example continued

Let $a_n = a + \sum x_i$ and $b_n = b + n - \sum x_i$.
It follows that the posterior distribution is
$p(\theta|x_{1:n}) = \text{Beta}(\theta|a_n, b_n)$.
The posterior predictive can be derived to be

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

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

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