Module 1: Introduction to Bayesian Statistics

Rebecca C. Steorts

Agenda

- Motivations
- Traditional inference
- Bayesian inference
- Bernoulli, Beta
- Posterior of Bernoulli-Beta
- Conjugacy
- 2012 Election (Obama vs Romney)
- Marginal likelihood
- Posterior Prediction
- Additional problems at the end of lecture (derivation + applied)

What should you learn?

- You should learn the main principles of Bayesian inference/prediction and how to apply these to real data analysis.
- You will continue with this in lab/homework to make sure that you understand these key principles.

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.
- Please review unbiased estimation and finding an MLE.
- Please also review other background material such as likelihoods, sufficient statistics, basic probability concepts, etc. Most of this material can be reviewed in Chapters 1-3 in Hoff.

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

How can we derive $p(\theta \mid x)$?

Derivation 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 $1 - \theta$.

Consider $X \sim \text{Bernoulli}(\theta)\mathbb{1}(0 < \theta < 1)$

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?
- What is the connection with the Bernoulli and the Binomial distribution?

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 $x_1, \ldots, x_n$

Bernoulli and Binomial Connection

$X_1, \ldots, X_n \overset{iid}{\sim}$ Bernoulli(θ).[1]

Suppose $Y = \sum_i X_{i=1}^n$. Then $Y \sim \mathit{Binomial}(n, \theta)$.[2]

Remark: A binomial random variable with parameter $n = 1$ is equivalent to a Bernoulli random variable, i.e. there is only one trial.

[1] This represents n coin flips with success probability θ.

[2] This represents n Bernoulli trials with success probability θ.

Likelihood

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

Beta distribution

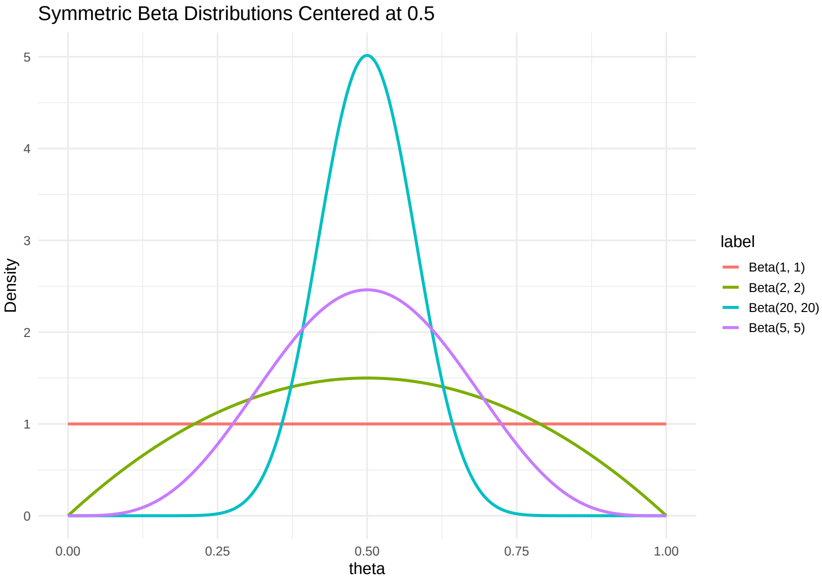
Given $a, b > 0$, we write $\theta \sim \text{Beta}(a, b)$ to mean that θ has pdf

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

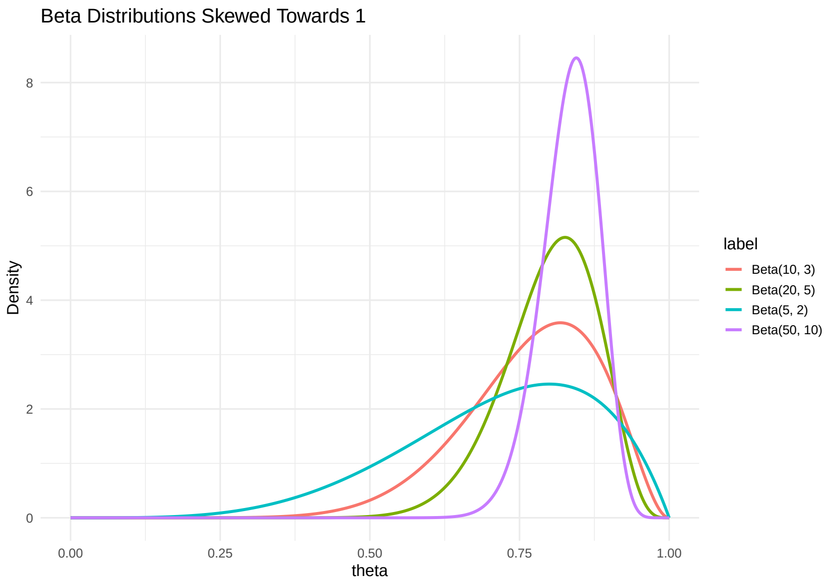
i.e., $p(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$ on the interval from 0 to 1.

- Here,
$$B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$
.
- Parameters a, b control the shape of the distribution.
- This distribution models random behavior of percentages/proportions.

Beta distribution



Beta distribution



Posterior of Bernoulli-Beta

Let's derive the posterior of $\theta \mid x_{1:n}$

Conjugacy

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

Conjugacy

A class P of prior distributions for θ is called **conjugate** for the likelihood $p(x \mid \theta)$ if

$$p(\theta) \in P \implies p(\theta \mid x) \in P.$$

Tip: In practice, we check to see if the posterior has an updated form of the prior.

Conjugacy

Benefits

- We do minimal or often no math. In fact, https://en.wikipedia.org/wiki/Conjugate_prior provides many conjugate families.
- We have an exact posterior distribution. No approximations are needed.
- Computation is fast and simple!

Downside

- Sometimes an unrealistic assumption, however, might provide guidance to us.

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. Likelihood
- The national approval rating (Zogby poll) of President Obama in mid-September 2010 was 50%. We'll assume that in PA his approval rating is also 50%. Prior
- Based on this prior information, we'll use a Beta prior for θ and we'll choose a and b.

Obama Example

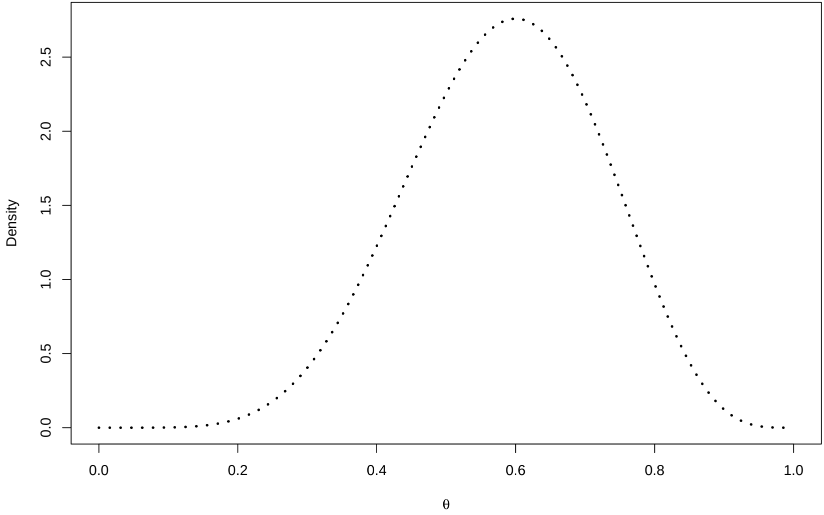
```r
n <- 10
# Fixing values of a,b. Chosen skewed Beta.
#a = 21/8
#b = 0.04
a <- 2
b <- 2
th <- seq(0, 1, length = 500)
x <- 6
# we set the likelihood, prior, and posteriors with
# THETA as the sequence that we plot on the x-axis.
# Beta(c,d) refers to shape parameter
like <- dbeta(th, x + 1 , n - x + 1)
prior <- dbeta(th, a, b)
print(a / (a + b))
```

```
## [1] 0.5
```

```r
post <- dbeta(th, x + a, n - x + b)
```

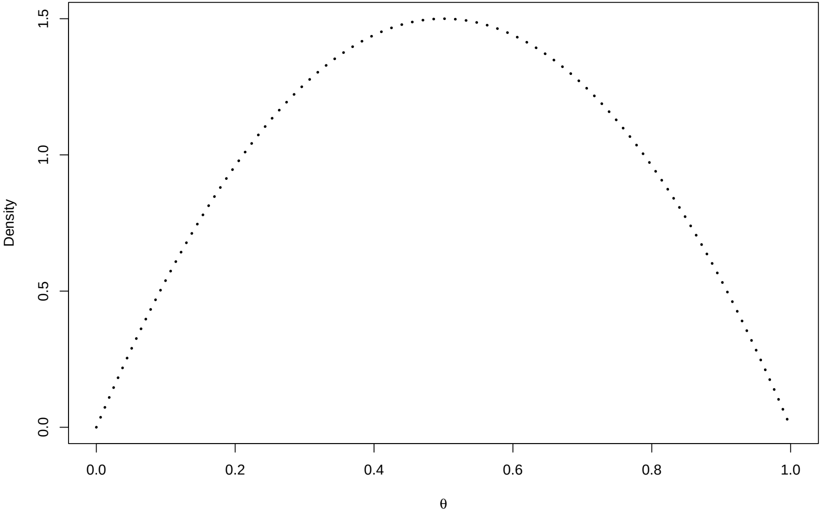
Likelihood

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



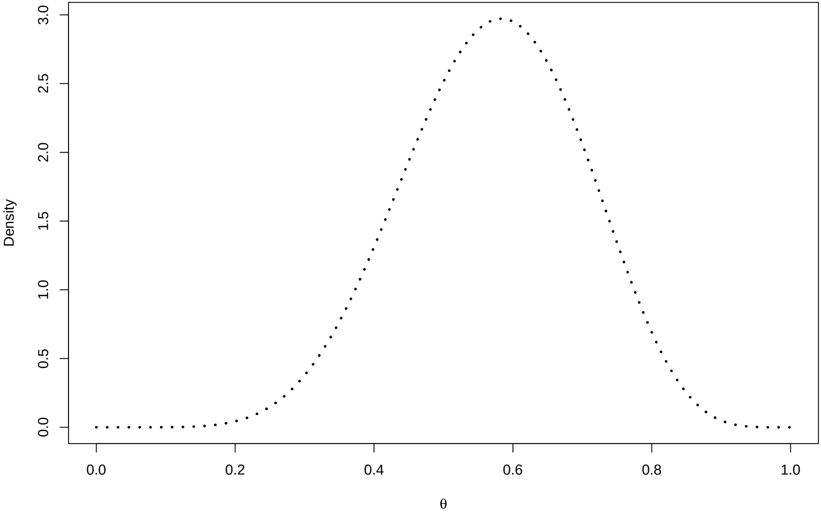
Prior

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

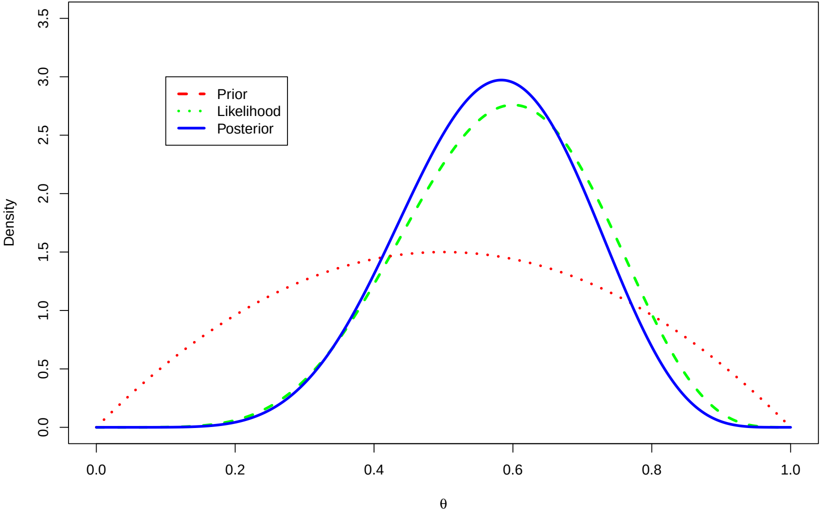


Posterior

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



Likelihood, Prior, and Posterior



Back to the Prior

- We choose the prior here two different ways. What do you observe?
- In the supplemental material (end of lecture), find an example where we have more information and can set a, b from in a more subjective and principled manner.

Cast of characters

- Observed data: x
- This often involves many data points, e.g., $x = x_{1:n} = (x_1, \dots, x_n)$.

likelihood	$p(x_{1:n}\|\theta)$
prior	$p(\theta)$
posterior	$p(\theta\|x_{1:n})$
marginal likelihood	$p(x_{1:n})$
posterior predictive	$p(x_{n+1}\|x_{1:n})$

Marginal likelihood

The **marginal likelihood** is defined as

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

Example: Back to the Bernoulli-Beta

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

and

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

What is the marginal likelihood for the Bernoulli-Beta?

Marginal Likelihood: Bernoulli-Beta

Then the marginal likelihood is

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

by the integral definition of the Beta function.

Posterior predictive distribution

- At times, we may wish to find the conditional distribution of x_{n+1} given $x_{1:(n+1)}$.
- Assumption 1: Assume that $x_{1:(n+1)}$ are independent given θ

$$\begin{aligned}
p(x_{n+1}|x_{1:n}) &= \int p(x_{n+1}, \theta | x_{1:n})\, d\theta \\
&\int \frac{p(x_{n+1}, \theta, x_{1:n})}{p(x_{1:n})}\, d\theta \quad \text{(Conditional probability)} \\
&\int \frac{p(x_{n+1}|\theta, x_{1:n}) p(\theta|x_{1:n}) p(x_{1:n})}{p(x_{1:n})}\, d\theta \quad \text{(Product rule)} \\
&= \int p(x_{n+1}|\theta, x_{1:n}) p(\theta|x_{1:n})\, d\theta \\
&= \int p(x_{n+1}|\theta) p(\theta|x_{1:n})\, d\theta \quad \text{By Assumption 1.}
\end{aligned}$$

Posterior predictive distribution: Bernoulli-Beta

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

and

$$\theta \sim \textit{Beta}(a, b).$$

The posterior distribution can be shown to be $p(\theta|x_{1:n}) = \text{Beta}(\theta|a_n, b_n)$, where $a_n = a + \sum x_i$ and $b_n = b + n - \sum x_i$.

Posterior predictive distribution: Bernoulli-Beta

The posterior predictive can be derived to be

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

Similarly,

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

Posterior predictive distribution (continued)

This implies that

$$p(x_{n+1}|x_{1:n}) = \begin{cases} \frac{a_n}{a_n+b_n} = \frac{(a+\sum_i x_i)}{a+b+n} & \text{if } x_{n+1} = 1 \\ \frac{b_n}{a_n+b_n} = \frac{b+\sum_i (1-x_i)}{a+b+n} & \text{if } x_{n+1} = 0 \end{cases}$$

More formally,

$$\begin{aligned} p(x_{n+1}|x_{1:n}) &= \frac{a_n^{x_{n+1}} b_n^{1-x_{n+1}}}{a_n + b_n} \mathbb{1}(x_{n+1} \in \{0,1\}) \\ &= \frac{(a + \sum_i x_i)^{x_{n+1}} (b + \sum_i (1 - x_i))^{1-x_{n+1}}}{(a+b+n)} \mathbb{1}(x_{n+1} \in \{0,1\}) \end{aligned}$$

Either solution above is correct. (See page 40 of Hoff for a similar derivation of this result).

Posterior predictive distribution

Observe that the posterior predictive distribution:

1. Does not depend on unknown parameters.
2. The predictive distribution depends on the observed data.

Overall 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
- Marginal distribution
- Posterior predictive

Background Knowledge

- Familiar with Discrete and Continuous Distributions
- Can calculate expectations and variances
- Change of variables
- Mean squared error
- Sufficiency
- Confident calculating the likelihood and log-likelihood
- Confident in working with partial derivatives
- Familiar maximizing or minimizing functions (and proving they are global max/min)

Detailed Summary for Exam

- Bayes Theorem
- Likelihood
- Prior
- Posterior derivation
- Marginal likelihood
- Posterior predictive distribution
- Conjugacy
- Proportionality
- Understanding when models are appropriate for data given to you (Ex: Approval ratings for Obama)
- What is an informative prior
- What is a non-informative prior
- Proper posterior
- How do you incorporate a pilot study into your posterior analysis (Ex: See sleep study)

Supplemental Material

Below you will find supplemental material, such as exercises to help you for the exam with solutions provided.

Exercise 1

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?

Solution

Since the data is independent given θ, the likelihood factors and we get

$$\begin{aligned} p(x_{1:n}|\theta) &= \prod_{i=1}^{n} p(x_i|\theta) \\ &= \prod_{i=1}^{n} e^{-\theta}\theta^{x_i}/x_i! \\ &\underset{\theta}{\propto} e^{-n\theta}\theta^{\sum x_i}. \end{aligned}$$

Solution 1

Thus, using Bayes' theorem,

$$\begin{aligned} p(\theta|x_{1:n}) &\propto p(x_{1:n}|\theta)p(\theta) \\ &\propto e^{-n\theta}\theta^{\sum x_i}\theta^{a-1}e^{-b\theta}\mathbb{1}(\theta > 0) \\ &\propto e^{-(b+n)\theta}\theta^{a+\sum x_i - 1}\mathbb{1}(\theta > 0) \\ &\propto \text{Gamma}\left(\theta \mid a + \textstyle\sum x_i,\ b + n\right). \end{aligned}$$

Therefore, since the posterior density must integrate to 1, we have

$$p(\theta|x_{1:n}) = \text{Gamma}\left(\theta \mid a + \textstyle\sum x_i,\ b + n\right).$$

Exercise 2

Suppose that $Y = \sum_i X_i$, where $X_i \mid \theta \overset{iid}{\sim} \text{Bernoulli}(\theta)$ for $i = 1, \ldots, n$.

a. What is the distribution of Y.
b. What is a conjugate prior? (Provide the distribution and parameters).
c. What is the posterior update for $\theta \mid Y$ assuming the conjugate prior in part b.
d. Write the posterior mean $E[\theta \mid Y]$ as a weighted average of the prior mean and the sample mean, where you specify the weights.

Solution

a. $Y \sim \text{Binomial}(n, \theta)$.
b. A conjugate prior is $\theta \sim \text{Beta}(a, b)$ for $a, b > 0$ and known.
c. The posterior update is $\theta \mid Y \sim \text{Beta}(a + y, n + b - y)$.

Solution

Recall the prior mean is $a/(a+b)$ and the sample mean is y/n.

d. The posterior mean is

$$
\begin{aligned}
E[\theta \mid Y] &= \frac{a+y}{a+b+n} && (1)\\
&= \frac{a}{a+b+n} + \frac{y}{a+b+n} && (2)\\
&= \frac{a}{a+b+n} \times \frac{a+b}{a+b} + \frac{y}{a+b+n} \times \frac{n}{n} && (3)\\
&= \frac{a+b}{a+b+n} \times \frac{a}{a+b} + \frac{n}{a+b+n} \times \frac{y}{n} && (4)
\end{aligned}
$$

Above the prior mean and sample mean is in blue and the respective weights are multipled by either prior mean or sample mean.

Solution

The weights are proportional to $a+b$ for the prior mean and n for the sample mean.

This leads to an interpretation of a and b as "prior data":

- $a \approx$ "prior number of 1's."
- $b \approx$ "prior number of 0's."
- $a+b \approx$ "prior sample size"

Remark: If $n >> a+b$ then we would inform θ according to the data. However, if $n << a+b$, we would inform θ according to our prior sample or historical data. (This is explained more in depth on page 39 of Hoff).

Module 1 Derivations

Class notes from Module 1 can be found below:

https://github.com/resteorts/modern-bayes/blob/master/lecturesModernBayes20/lecture-1/notes-module1.pdf

https://github.com/resteorts/modern-bayes/blob/master/lecturesModernBayes20/lecture-1/notes-module1.pdf

Additional Applied Example

Below, there is an additional applied example that you may find useful regarding this material.

How Much Do You Sleep Example

We are interested in a population of American college students and the proportion of the population that sleep at least eight hours a night, which we denote by θ.

How Much Do You Sleep Example

- *The Gamecock*, at the USC printed an internet article "College Students Don't Get Enough Sleep" (2004).
  - Most students spend six hours sleeping each night.
- 2003: University of Notre Dame's paper, *Fresh Writing*.
  - The article reported took random sample of 100 students:
  - "approximately 70% reported to receiving only five to six hours of sleep on the weekdays,
  - 28% receiving seven to eight,
  - and only 2% receiving the healthy nine hours for teenagers."

How Much Do You Sleep

- Have a random sample of 27 students is taken from UF.
- 11 students record that they sleep at least eight hours each night.
- Based on this information, we are interested in estimating θ.

How Much Do You Sleep

- From USC and UND, believe it's probably true that most college students get less than eight hours of sleep.
- Want our prior to assign most of the probability to values of $\theta < 0.5$.
- From the information given, we decide that our best guess for θ is 0.3, although we think it is very possible that θ could be any value in $[0, 0.5]$.

Our Model

Our model can be summarized by the Binomial-Beta distribution

$$X|\theta \sim \text{Binomial}(n, \theta) \tag{5}$$
$$\theta \sim \text{Beta}(a, b) \tag{6}$$

You can show that the posterior of

$$\theta \mid X \sim \text{Beta}(x + a, n - x + b)$$

Choice of a,b for Beta Prior

- Given this information, we believe that the median of θ is 0.3 and the 90th percentile is 0.5.
- Knowing this allows us to estimate the unknown values of a and b.
- How do we actually calculate a and b?

Choice of a,b for Beta Prior

We would need to solve the following equations:

$$\int_0^{0.3} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1}\, d\theta = 0.5$$
$$\int_0^{0.5} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1}\, d\theta = 0.9$$

In non-calculus language, this means the 0.5 quantile (50th percentile)$_=$0.3. The 0.9 quantile (90th percentile) $=$ 0.5.

The equations are written as percentiles above!

- We can easily solve this numerically in `R` using a numerical solver `BBsolve` using the `BB` package. .
- The documentation for this package is not great, so beware.

How Much Do You Sleep

```r
## load the BB package
library(BB)

## using percentiles
myfn <- function(shape) {
    test <- pbeta(q = c(0.3, 0.5), shape1 = shape[1],
     shape2 = shape[2]) - c(0.5, 0.9)
   return(test) }
BBsolve(c(1, 1), myfn)
```

```
##   Successful convergence.
```

```
## $par
## [1] 3.263743 7.185121
##
## $residual
## [1] 5.905161e-08
##
## $fn.reduction
```

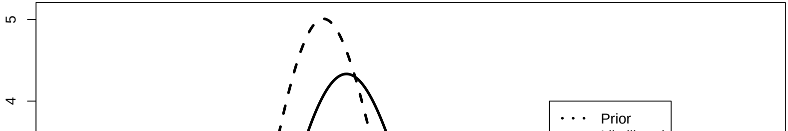
How Much Do You Sleep

Using our calculations from the Beta-Binomial our model is

$$
\begin{aligned}
X \mid \theta &\sim \text{Binomial}(27, \theta) \\
\theta &\sim \text{Beta}(3.3, 7.2) \\
\theta \mid x &\sim \text{Beta}(x + 3.3, 27 - x + 7.2) \\
\theta \mid 11 &\sim \text{Beta}(14.3, 23.2)
\end{aligned}
$$

How Much Do You Sleep

```
th <- seq(0,1,length=500)
a <- estimated$par[1]
b <- estimated$par[2]
n <- 27
x <- 11
prior <- dbeta(th, a, b)
like <- dbeta(th, x + 1, n - x + 1)
post <- dbeta(th, x + a, n - x + b)
plot(th, post, type = "l", ylab = "Density", lty = 2, lwd =
xlab = expression(theta))
lines(th, like, lty = 1, lwd = 3)
lines(th, prior, lty = 3, lwd = 3)
legend(0.7, 4, c("Prior", "Likelihood", "Posterior"),
lty = c(3,1,2), lwd = c(3,3,3))
```



How Much Do You Sleep