Intro to Bayesian Methods

Rebecca C. Steorts
Predictive Modeling and Data Mining: STA 521

October 2015

Topics

- Why Bayes'
- Bayes' Theorem
- Hierarchical models
- Conjugacy
- Examples
- Lab: applied examples

- Why should we learn about Bayesian concepts?
- Natural if thinking about unknown parameters as random.
- They naturally give a full distribution when we perform an update.
- We automatically get uncertainty quantification.
- Drawbacks: They can be slow and inconsistent.

Suppose we have some noisy data. How can we recover the underlying structure of the data?

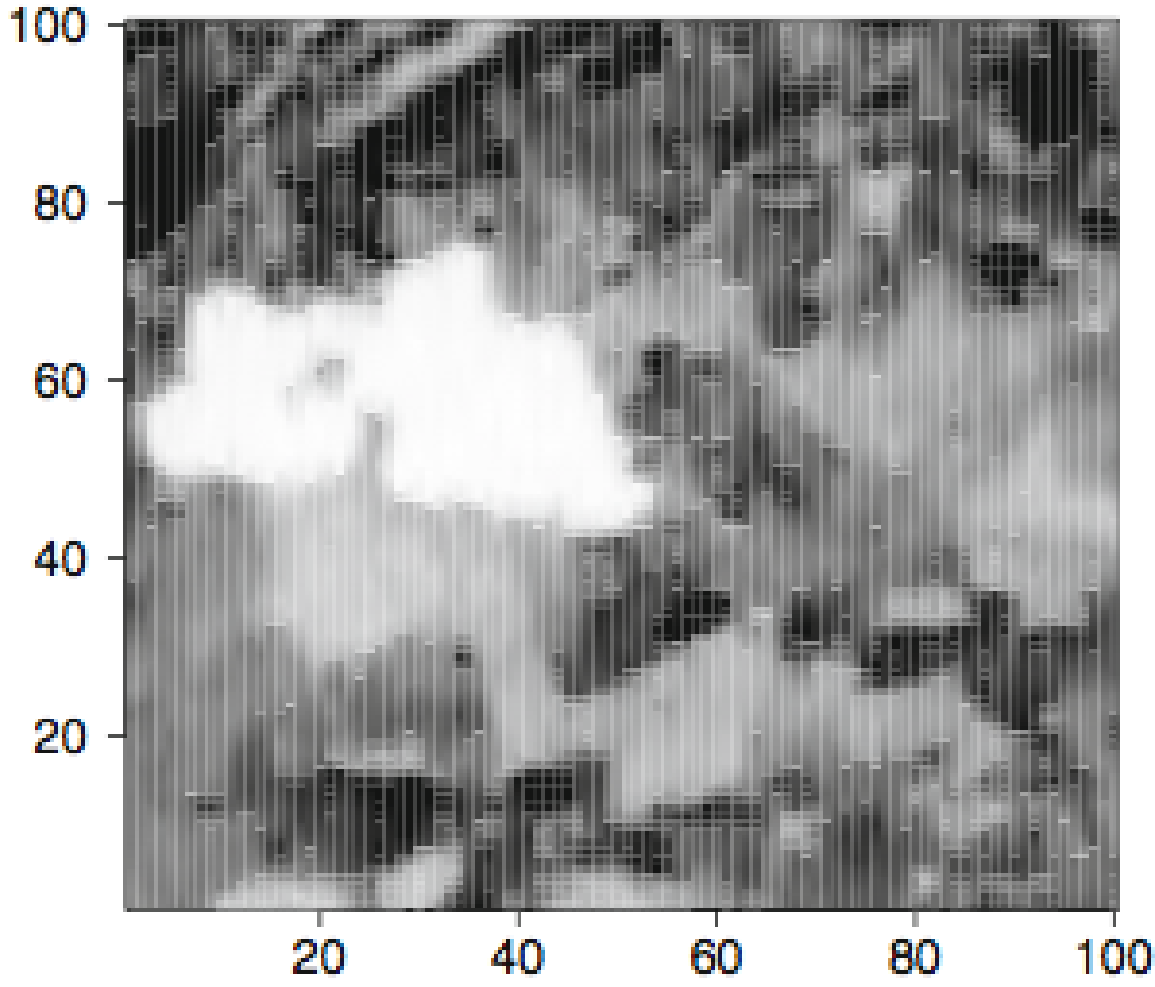


Figure 1: Satellite image of the lake of Menteith, Scotland

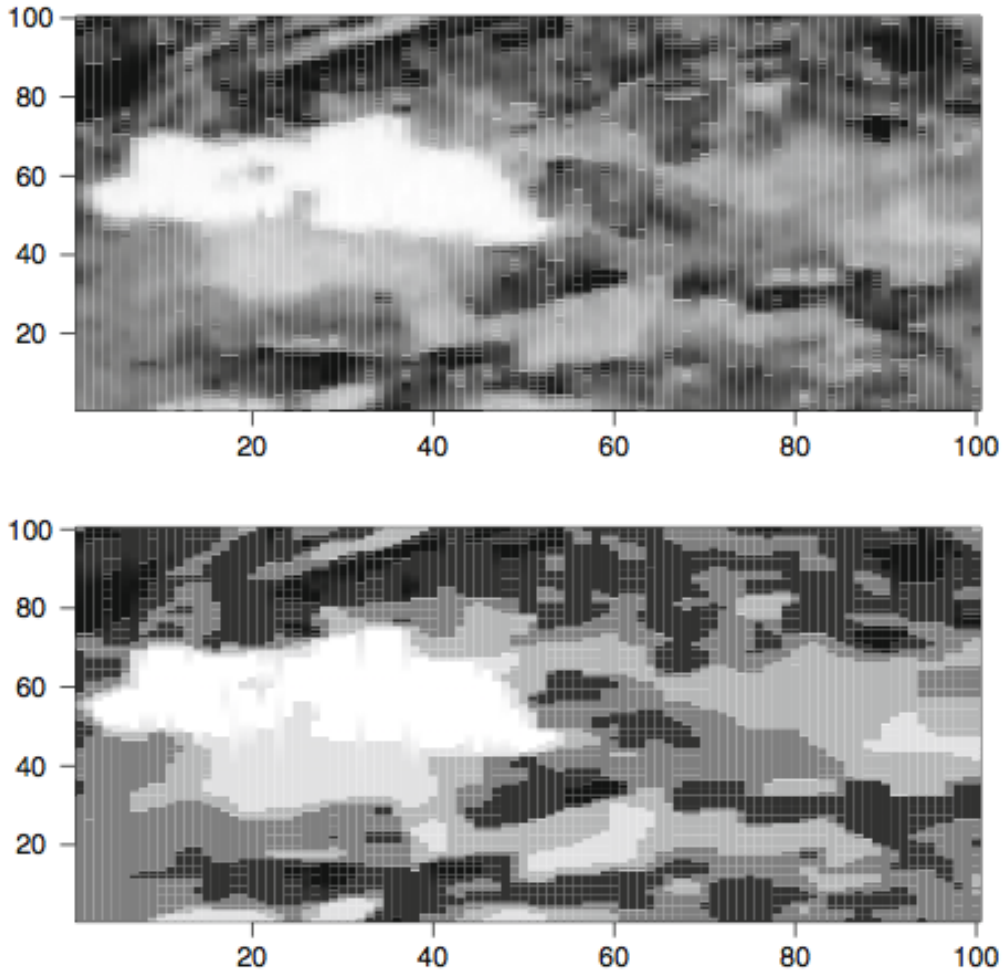


Figure 2: Dataset Menteith: (top) the observed image and (bottom) Segmented image

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

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

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \propto p(x|\theta)p(\theta). \tag{1}$$

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

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \propto p(x|\theta)p(\theta). \tag{1}$$

We can decompose Bayes' Theorem into three principal terms:

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

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \propto p(x|\theta)p(\theta). \tag{1}$$

We can decompose Bayes' Theorem into three principal terms:

$p(\theta|x)$ posterior

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

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \propto p(x|\theta)p(\theta). \tag{1}$$

We can decompose Bayes' Theorem into three principal terms:

$p(\theta\|x)$	posterior
$p(x\|\theta)$	likelihood

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

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \propto p(x|\theta)p(\theta). \tag{1}$$

We can decompose Bayes' Theorem into three principal terms:

$p(\theta\|x)$	posterior
$p(x\|\theta)$	likelihood
$p(\theta)$	prior

Polling Example 2012

Let's apply this to a real example! We're interested in the proportion of people that approve of President Obama in PA.

Polling Example 2012

Let's apply this to a real example! We're interested in 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.

Polling Example 2012

Let's apply this to a real example! We're interested in 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-December was 45%. We'll assume that in PA his approval rating is approximately 50%.

Polling Example 2012

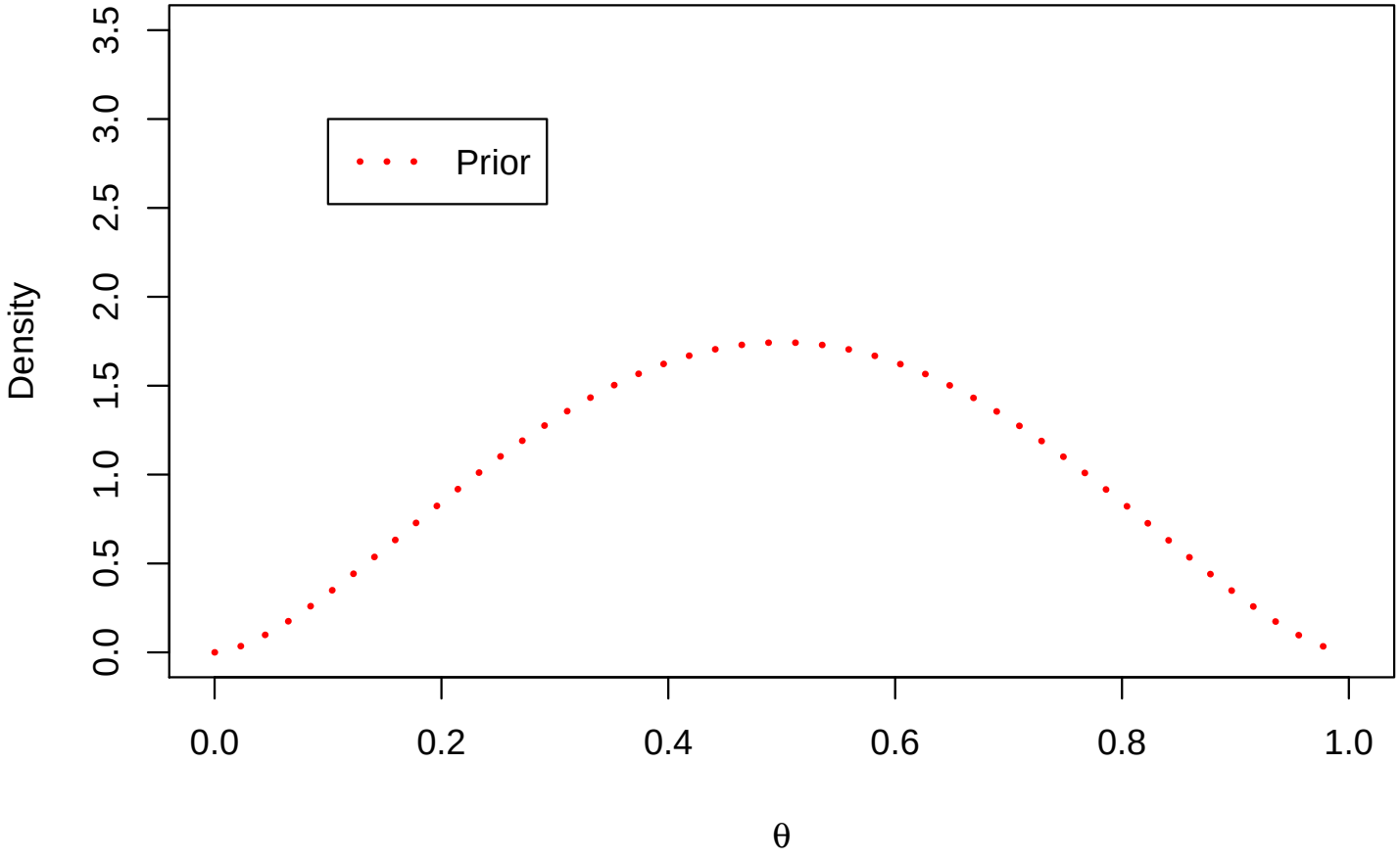
Let's apply this to a real example! We're interested in the proportion of people that approve of President Obama in PA.

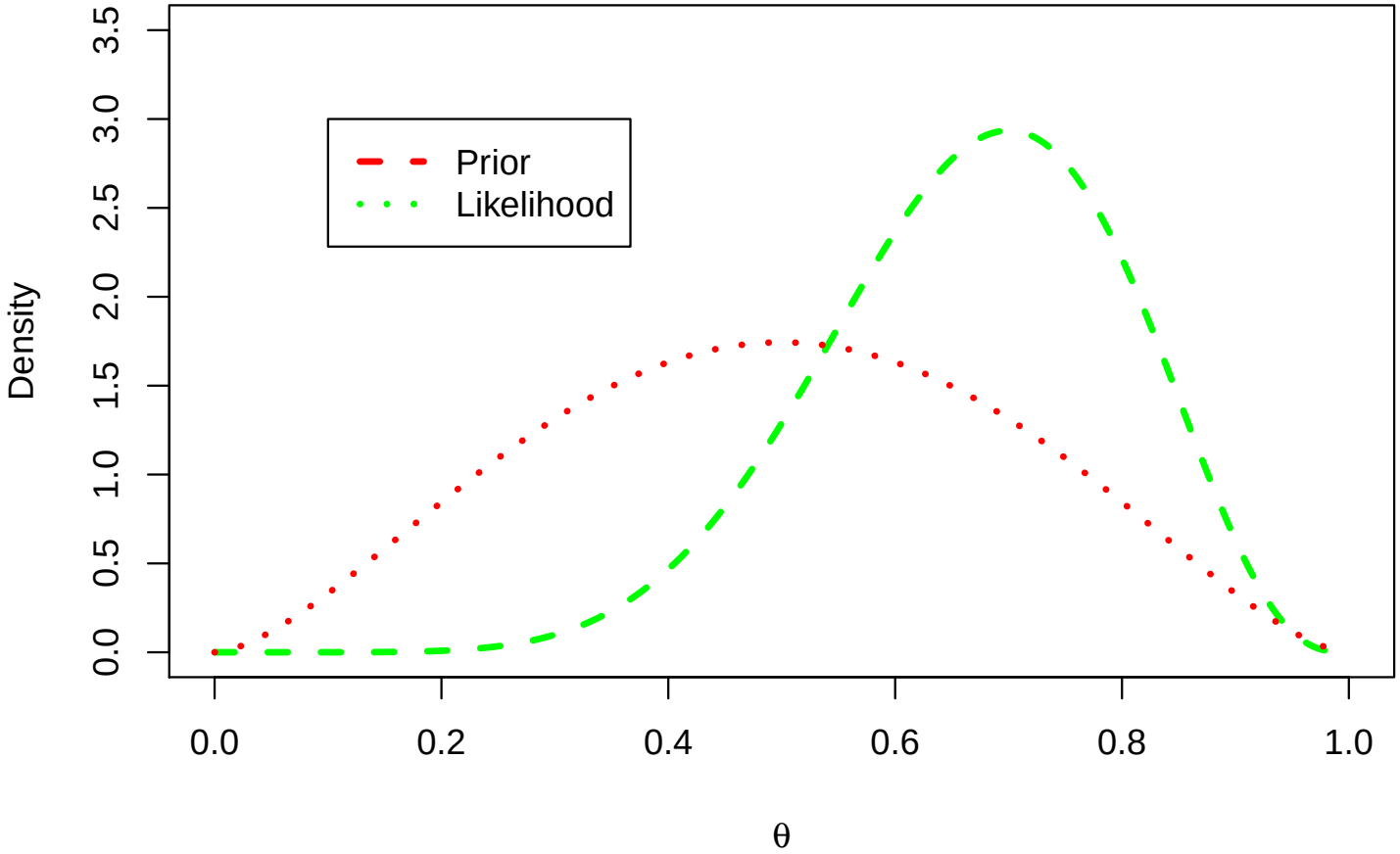
- We take a random sample of 10 people in PA and find that 6 approve of President Obama.
- The national approval rating (Zogby poll) of President Obama in mid-December was 45%. We'll assume that in PA his approval rating is approximately 50%.
- Based on this prior information, we'll use a Beta prior for θ and we'll choose a and b. (Won't get into this here).

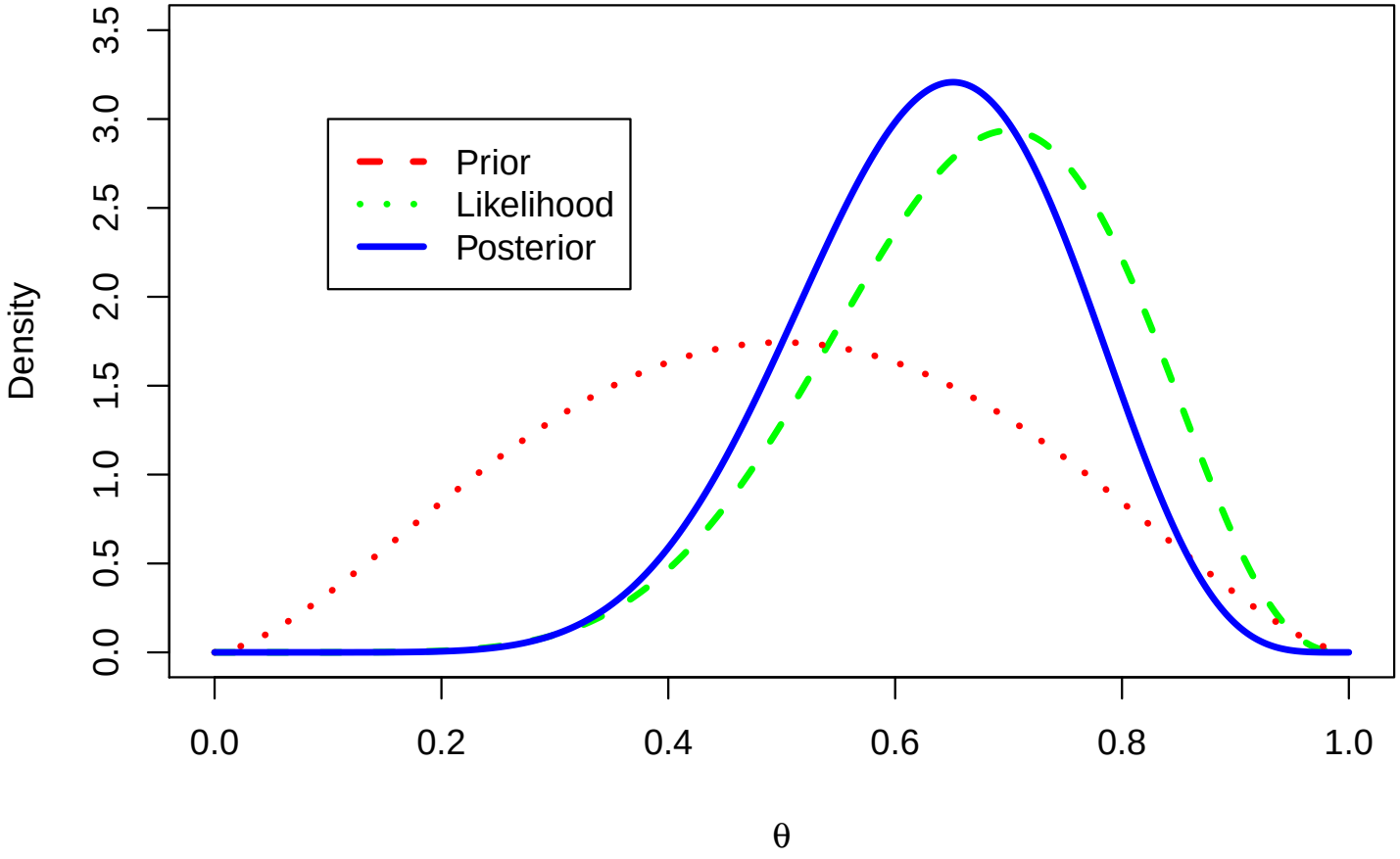
Polling Example 2012

Let's apply this to a real example! We're interested in 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-December 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. (Won't get into this here).
- We can plot the prior and likelihood distributions in `R` and then see how the two mix to form the posterior distribution.


Prior
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0.0
0.2
0.4
0.6
0.8
1.0
θ


Prior
Likelihood
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0.0
0.2
0.4
0.6
0.8
1.0
θ


Prior
Likelihood
Posterior
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0.0
0.2
0.4
0.6
0.8
1.0
θ

The basic philosophical difference between the frequentist and Bayesian paradigms is that

- Bayesians treat an unknown parameter θ as *random*.

The basic philosophical difference between the frequentist and Bayesian paradigms is that

- Bayesians treat an unknown parameter θ as *random*.
- Frequentists treat θ as unknown but *fixed*.

Stopping Rule

Let θ be the probability of a particular coin landing on heads, and suppose we want to test the hypotheses

Stopping Rule

Let θ be the probability of a particular coin landing on heads, and suppose we want to test the hypotheses

$$H_0 : \theta = 1/2, \qquad H_1 : \theta > 1/2$$

at a significance level of $\alpha = 0.05$. Suppose we observe the following sequence of flips:

heads, heads, heads, heads, heads, **tails** (5 heads, 1 tails)

Stopping Rule

Let θ be the probability of a particular coin landing on heads, and suppose we want to test the hypotheses

$$H_0 : \theta = 1/2, \qquad H_1 : \theta > 1/2$$

at a significance level of $\alpha = 0.05$. Suppose we observe the following sequence of flips:

heads, heads, heads, heads, heads, **tails** (5 heads, 1 tails)

- To perform a frequentist hypothesis test, we must define a random variable to describe the data.

Stopping Rule

Let θ be the probability of a particular coin landing on heads, and suppose we want to test the hypotheses

$$H_0 : \theta = 1/2, \qquad H_1 : \theta > 1/2$$

at a significance level of $\alpha = 0.05$. Suppose we observe the following sequence of flips:

heads, heads, heads, heads, heads, **tails** (5 heads, 1 tails)

- To perform a frequentist hypothesis test, we must define a random variable to describe the data.
- The proper way to do this depends on exactly which of the following two experiments was actually performed:

- Suppose the experiment is "**Flip six times and record the results.**"

- Suppose the experiment is "**Flip six times and record the results.**"
  - X counts the number of heads, and $X \sim \text{Binomial}(6, \theta)$.
  - The observed data was $x = 5$, and the p-value of our hypothesis test is

- Suppose the experiment is "**Flip six times and record the results.**"
  - X counts the number of heads, and $X \sim \text{Binomial}(6, \theta)$.
  - The observed data was $x = 5$, and the p-value of our hypothesis test is

$$\begin{aligned} \text{p-value} &= P_{\theta=1/2}(X \geq 5) \\ &= P_{\theta=1/2}(X = 5) + P_{\theta=1/2}(X = 6) \end{aligned}$$

- Suppose the experiment is "**Flip six times and record the results.**"
  - X counts the number of heads, and $X \sim \text{Binomial}(6, \theta)$.
  - The observed data was $x = 5$, and the p-value of our hypothesis test is

$$\begin{aligned} \text{p-value} &= P_{\theta=1/2}(X \geq 5) \\ &= P_{\theta=1/2}(X = 5) + P_{\theta=1/2}(X = 6) \\ &= \frac{6}{64} + \frac{1}{64} = \frac{7}{64} = 0.109375 > 0.05. \end{aligned}$$

- Suppose the experiment is "**Flip six times and record the results.**"
  - X counts the number of heads, and $X \sim \text{Binomial}(6, \theta)$.
  - The observed data was $x = 5$, and the p-value of our hypothesis test is

$$\begin{aligned} \text{p-value} &= P_{\theta=1/2}(X \geq 5) \\ &= P_{\theta=1/2}(X = 5) + P_{\theta=1/2}(X = 6) \\ &= \frac{6}{64} + \frac{1}{64} = \frac{7}{64} = 0.109375 > 0.05. \end{aligned}$$

So we fail to reject H_0 at $\alpha = 0.05$.

- Suppose now the experiment is "**Flip until we get tails.**"

- Suppose now the experiment is "**Flip until we get tails.**"
  - X counts the number of the flip on which the first tails occurs, and $X \sim$ Geometric$(1-\theta)$.
  - The observed data was $x = 6$, and the p-value of our hypothesis test is

$$\text{p-value} \quad = \quad P_{\theta=1/2}(X \geq 6)$$

- Suppose now the experiment is "**Flip until we get tails.**"
  - X counts the number of the flip on which the first tails occurs, and $X \sim$ Geometric$(1-\theta)$.
  - The observed data was $x = 6$, and the p-value of our hypothesis test is

$$\begin{aligned} \text{p-value} &= P_{\theta=1/2}(X \geq 6) \\ &= 1 - P_{\theta=1/2}(X < 6) \end{aligned}$$

- Suppose now the experiment is "**Flip until we get tails.**"
  - X counts the number of the flip on which the first tails occurs, and $X \sim$ Geometric$(1-\theta)$.
  - The observed data was $x = 6$, and the p-value of our hypothesis test is

$$\begin{aligned} \text{p-value} &= P_{\theta=1/2}(X \geq 6) \\ &= 1 - P_{\theta=1/2}(X < 6) \\ &= 1 - \sum_{x=1}^{5} P_{\theta=1/2}(X = x) \end{aligned}$$

- Suppose now the experiment is "**Flip until we get tails.**"
  - X counts the number of the flip on which the first tails occurs, and $X \sim$ Geometric$(1-\theta)$.
  - The observed data was $x = 6$, and the p-value of our hypothesis test is

$$\begin{aligned}
\text{p-value} &= P_{\theta=1/2}(X \geq 6) \\
&= 1 - P_{\theta=1/2}(X < 6) \\
&= 1 - \sum_{x=1}^{5} P_{\theta=1/2}(X = x) \\
&= 1 - \left(\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \frac{1}{32}\right) = \frac{1}{32} = 0.03125 < 0.05.
\end{aligned}$$

So we reject H_0 at $\alpha = 0.05$.

- The conclusions differ, which seems strikes *some people* as absurd.

- The conclusions differ, which seems strikes *some people* as absurd.
- P-values aren't close—one is 3.5 times as large as the other.

- The conclusions differ, which seems strikes *some people* as absurd.
- P-values aren't close—one is 3.5 times as large as the other.
- The result our hypothesis test depends on whether we would have stopped flipping if we had gotten a tails sooner.

- The conclusions differ, which seems strikes *some people* as absurd.
- P-values aren't close—one is 3.5 times as large as the other.
- The result our hypothesis test depends on whether we would have stopped flipping if we had gotten a tails sooner.
- The tests are dependent on what we call the *stopping rule*.

- The likelihood for the actual value of x that was observed is the same for both experiments (up to a constant):

$$p(x|\theta) \propto \theta^5(1-\theta).$$

- The likelihood for the actual value of x that was observed is the same for both experiments (up to a constant):

$$p(x|\theta) \propto \theta^5(1-\theta).$$

- A Bayesian approach would take the data into account only through this likelihood.

- The likelihood for the actual value of x that was observed is the same for both experiments (up to a constant):

$$p(x|\theta) \propto \theta^5(1-\theta).$$

- A Bayesian approach would take the data into account only through this likelihood.
- This would provide the same answers regardless of which experiment was being performed.

The Bayesian analysis is independent of the stopping rule since it only depends on the likelihood (show this at home!).

Hierarchical Bayesian Models

In a hierarchical Bayesian model, rather than specifying the prior distribution as a single function, we specify it as a hierarchy.

Hierarchical Bayesian Models

$$
\begin{aligned}
X|\theta &\sim f(x|\theta) \\
\Theta|\gamma &\sim \pi(\theta|\gamma) \\
\Gamma &\sim \phi(\gamma),
\end{aligned}
$$

where we assume that $\phi(\gamma)$ is known and not dependent on any other unknown *hyperparameters*.

Conjugate Distributions

Let F be the class of sampling distributions $p(y|\theta)$.

Conjugate Distributions

Let F be the class of sampling distributions $p(y|\theta)$.

- Then let P denote the class of prior distributions on θ.

Conjugate Distributions

Let F be the class of sampling distributions $p(y|\theta)$.

- Then let P denote the class of prior distributions on θ.
- Then P is said to be conjugate to F if for every $p(\theta) \in P$ and $p(y|\theta) \in F$, $p(y|\theta) \in P$.

Simple definition: A family of priors such that, upon being multiplied by the likelihood, yields a posterior in the same family.

Beta-Binomial

If $X|\theta$ is distributed as binomial(n, θ), then a conjugate prior is the beta family of distributions, where we can show that the posterior is

Beta-Binomial

If $X|\theta$ is distributed as binomial(n, θ), then a conjugate prior is the beta family of distributions, where we can show that the posterior is

$$\pi(\theta|x) \quad \propto \quad p(x|\theta)p(\theta)$$

Beta-Binomial

If $X|\theta$ is distributed as binomial(n, θ), then a conjugate prior is the beta family of distributions, where we can show that the posterior is

$$\begin{aligned} \pi(\theta|x) &\propto p(x|\theta)p(\theta) \\ &\propto \binom{n}{x}\theta^x(1-\theta)^{n-x}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} \end{aligned}$$

Beta-Binomial

If $X|\theta$ is distributed as binomial(n, θ), then a conjugate prior is the beta family of distributions, where we can show that the posterior is

$$\begin{aligned}
\pi(\theta|x) &\propto p(x|\theta)p(\theta) \\
&\propto \binom{n}{x}\theta^x(1-\theta)^{n-x}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} \\
&\propto \theta^x(1-\theta)^{n-x}\theta^{a-1}(1-\theta)^{b-1}
\end{aligned}$$

Beta-Binomial

If $X|\theta$ is distributed as binomial(n, θ), then a conjugate prior is the beta family of distributions, where we can show that the posterior is

$$\begin{aligned}
\pi(\theta|x) &\propto p(x|\theta)p(\theta) \\
&\propto \binom{n}{x}\theta^x(1-\theta)^{n-x}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} \\
&\propto \theta^x(1-\theta)^{n-x}\theta^{a-1}(1-\theta)^{b-1} \\
&\propto \theta^{x+a-1}(1-\theta)^{n-x+b-1} \implies
\end{aligned}$$

Beta-Binomial

If $X|\theta$ is distributed as binomial(n, θ), then a conjugate prior is the beta family of distributions, where we can show that the posterior is

$$\begin{aligned}
\pi(\theta|x) &\propto p(x|\theta)p(\theta) \\
&\propto \binom{n}{x}\theta^x(1-\theta)^{n-x}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} \\
&\propto \theta^x(1-\theta)^{n-x}\theta^{a-1}(1-\theta)^{b-1} \\
&\propto \theta^{x+a-1}(1-\theta)^{n-x+b-1} \implies
\end{aligned}$$

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

How Much Do You Sleep

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.

Prior Data

- *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."

- 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.

- 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 $[0, 0.5]$.

- 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.
- After some calculations we find that $a = 3.3$ and $b = 7.2$. How did we actually calculate a and b?

- 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.
- After some calculations we find that $a = 3.3$ and $b = 7.2$. How did we actually calculate a and b?

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) $= 0.5$.
- We can easily solve this numerically in `R` using a numerical solver `BBsolve`.

Since you won't have used this command before, you'll need to install the package `BB` and load the library.

Here is the code in R to find a and b.

```
## install the BBsolve package
install.packages("BB", repos="http://cran.r-project.org")
library(BB)
fn = function(x){qbeta(c(0.5,0.9),x[1],x[2])-c(0.3,0.5)}
BBsolve(c(1,1),fn)

## alternative way
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)
```

Using our calculations from the Beta-Binomial our model is

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

```
th = seq(0,1,length=500)
a = 3.3
b = 7.2
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)
pdf("sleep.pdf",width=7,height=5)
plot(th,post,type="l",ylab="Density",lty=2,lwd=3,
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))
dev.off()
```

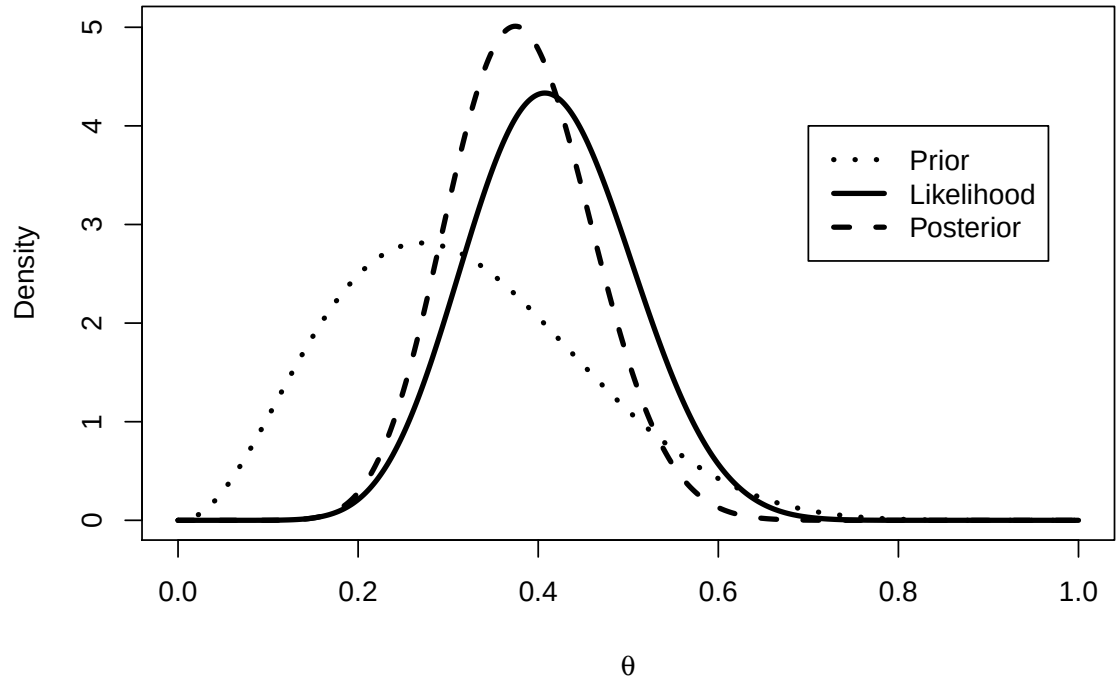


Figure 3: Likelihood $p(X|\theta)$, Prior $p(\theta)$, and Posterior Distribution $p(\theta|X)$