Lecture 1: Bayes and Beyond

Jeff Rouder

June, 2024

All Stats Begins With Probability

- Probability in Two Parts:
    - technical part for parceling probability across things
    - the soul or meaning

The Technical Part

- X, sample space, set of all outcomes
- A, B, subset of X, event
- $P(A)$ is probability of A, how much of X does it measure.
- Kolmogorov Axioms:
  - $P(A) \geq 0$
  - $P(X) = 1$
  - if $A \cap B = \emptyset$ then $P(A \cup B) = P(A) + P(B)$
- Kolmogorov Axioms describe a system of *relative weights*.
- Everyone agrees on probability as a relative weight system that obeys the Kolmogorov Axioms

Some Questions About Meaning?

Flip a coin:

- Is the probability of a head the property of a coin much as the coin has properties of weight, composition, and circumference? Alternatively, is probability a property of the observer? Both?
- Does the probability of a head change depending on the situation?
- Is probability subjective or objective?
- Can we talk about the probability of one-off events, say that the Ukraine war ends in 2025?

Usual, Classical, Frequentist

- Probability of heads is the proportion of heads in the long-run limit of many flips.
    - belongs to coin
    - objective fact
    - long-run limit is an abstraction, nonetheless useful

Bayesian

- Probability is the observers belief about the plausibility of events.
- Goal is to update rationally in light of data using Kolmogov's Axioms and the Law of Conditional Probability
- Probability of a head:
    - belongs to observer
    - subjective opinion
    - mutable
    - no need for long-run abstraction, works always
- Probability as wagers
    - p=.25 (1-to-3 odds)
    - I might wager a dollar if I win more than three.

Bayesian Updating

- Before The Bad Apple Catastrophe
  - A: 20%
  - B: 30%
  - C: 50%
- After The Bad-Apple Catastrophe
  - A: ?
  - B: ??
  - C: 0%

Refresher

- $Pr(A \cap B)$: joint probability.
  - the foundation
  - it all starts here
- $Pr(A)$: Marginal Probability
- $Pr(A|B)$: Conditional Probability

Refresher

Joint:

```
##      A=0 A=1
## B=0 0.1 0.2
## B=1 0.3 0.4
```

Marginals For A:

```
## A=0 A=1
## 0.4 0.6
```

Marginals For B:

```
## B=0 B=1
## 0.3 0.7
```

Refresher

Joint:

```
##      A=0 A=1
## B=0 0.1 0.2
## B=1 0.3 0.4
```

$Pr(A|B)$:

```
##       A=0  A=1
## B=0 0.33 0.67
## B=1 0.43 0.57
```

Your Turn

What is $Pr(B|A)$?

- You need a 2-by-2 matrix of responses.

Ans

$Pr(B|A)$:

```
##      A=0  A=1
## B=0 0.25 0.33
## B=1 0.75 0.67
```

Puzzler

If I provided $P(A|B)$ and $P(B|A)$, could you compute $P(A \cap B)$?

- Yes / No ?
- Try the above example with 2-by-2. You have $Pr(A|B)$ and $Pr(B|A)$. Can you compute $Pr(A \cap B)$?

Law of Conditional Probability

Classic:

$$P(A|B) = \frac{P(B|A)Pr(A)}{P(B)}$$

Updating Version:

$$P(A|B) = \left[\frac{P(B|A)}{P(B)}\right] P(A)$$

Ratio Version:

$$\frac{P(A|B)}{P(A)} = \frac{P(B|A)}{P(B)}$$

Two Bayesian Orientations

- Posterior orientation:
  - What you learn in most courses
  - Closest to conventional statistics
  - Benefit of Bayesian machinery without a strong Bayesian commitment
  - Me: intellectually dangerous
  - Most of today
- Updating Orientation
  - Deep stuff
  - Controversial
  - Hard
  - Higher plane of Bayesian consciousness
  - I'll touch on it

Posterior Orientation

$$P(\theta|Y) = \frac{P(Y|\theta)P(\theta)}{P(Y)}$$

- Posterior: $P(\theta|Y)$
- Prior: $P(\theta)$
- Likelihood: $P(Y|\theta)$
- P(Y)? : marginal probability of data
    - uniquely Bayesian
    - doesn't depend on θ
    - constant

Posterior Orientation

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

Multiple Parameters

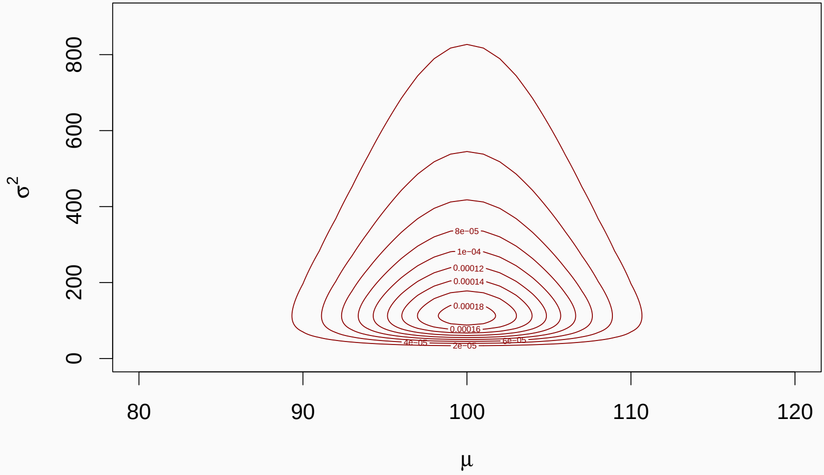
Kids On Smarties Take An IQ Test?

Data: 100,104,105,124

Kids On Smarties Take An IQ Test?

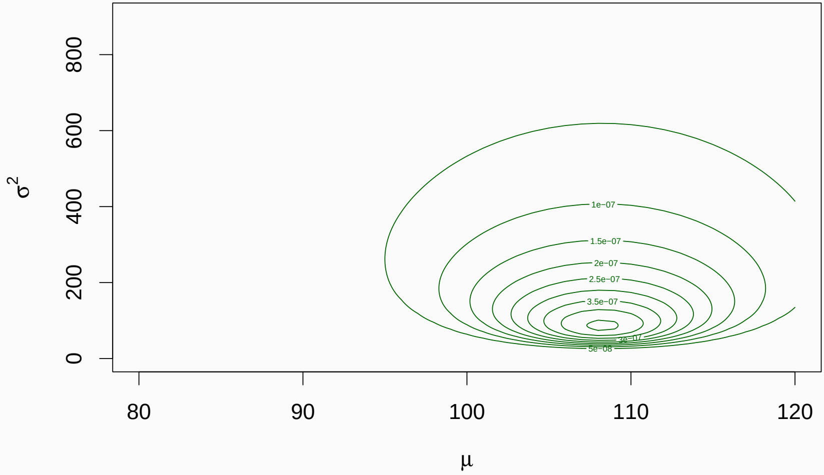
$$
\begin{aligned}
Y_i|\mu,\sigma^2 &\sim \text{Normal}(\mu,\sigma^2)\\
\mu &\sim \text{Norma}(100,5^2)\\
\sigma^2 &\sim \text{InvGamma}(1,15^2)
\end{aligned}
$$

My Prior



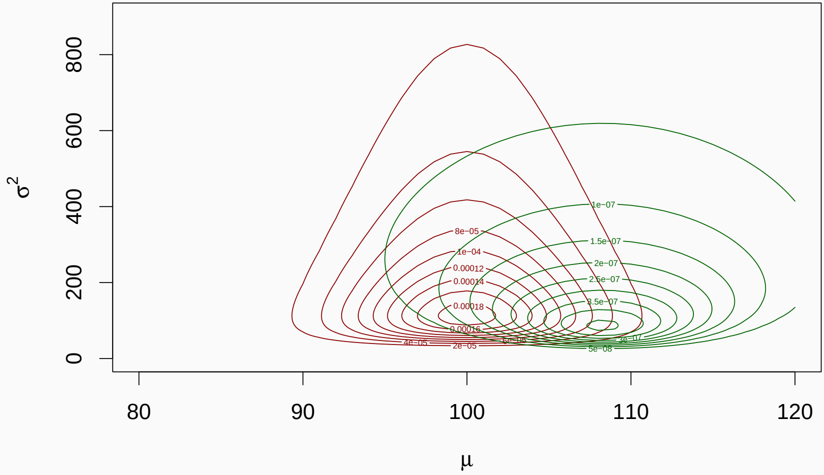
800
600
400
200
0
σ^2
80
90
100
110
120
μ
8e−05
1e−04
0.00012
0.00014
0.00018
0.00016
4e−05
2e−05
6e−05

Likelihood

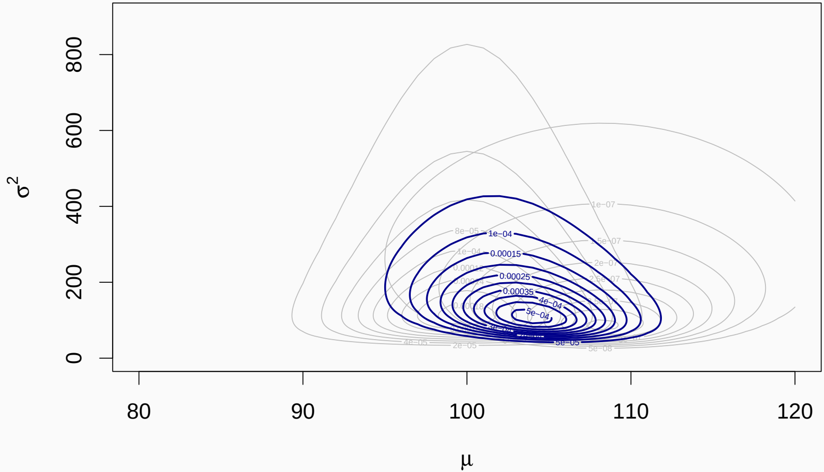


1e−07
1.5e−07
2e−07
2.5e−07
3.5e−07
3e−07
5e−08
0
200
400
600
800
80
90
100
110
120
σ^2
μ

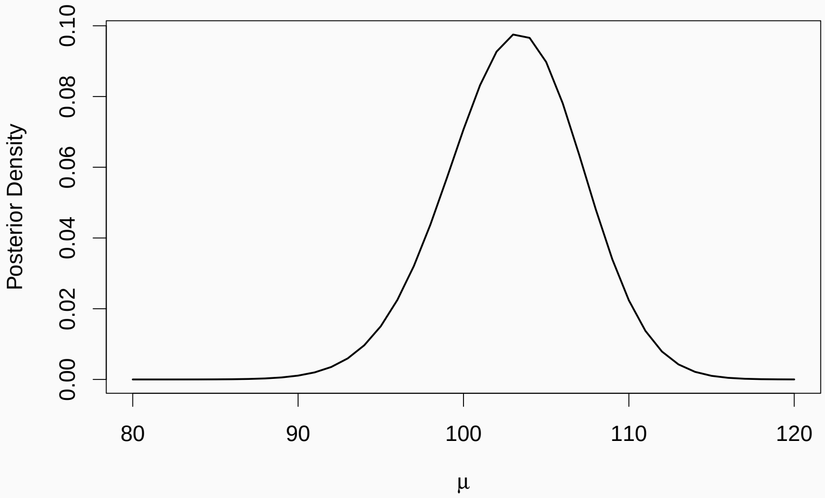
Posteiror: Multiply These, Point By Pont



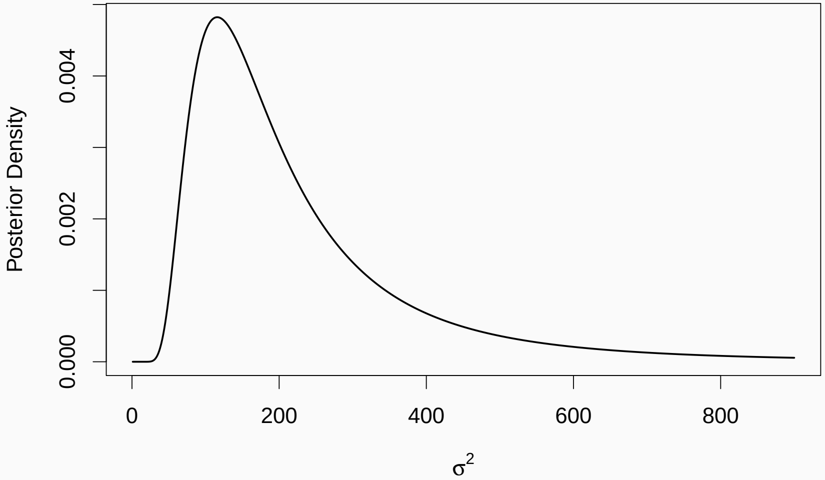
Posteiror: Multiply These, Point By Pont



Marginal Posterior for μ



Marginal Posterior for σ^2



Your Turn

Code up the normal in stan or jags:

- do you get the same joint posterior?
- do you get the same marginal posteriors?

Many People

- Example: 100 people each read 10 words. Q: How fast does each person read?
- Example data at `dat1.RDS`
- Your Turn: Write a model to describe this case. You can typeseet in Rmarkdown/Latex or just write it on paper.

Many People, Fixed Effects

$$
\begin{aligned}
Y_{ij}|\theta_i, \delta^2 &\sim \mathrm{N}(\theta_i, \tau^2) \\
\theta_i &\sim \mathrm{N}(a, b) \\
\tau^2 &\sim \mathrm{InvGamma}\left(\frac{1}{2}, \frac{r^2}{2}\right)
\end{aligned}
$$

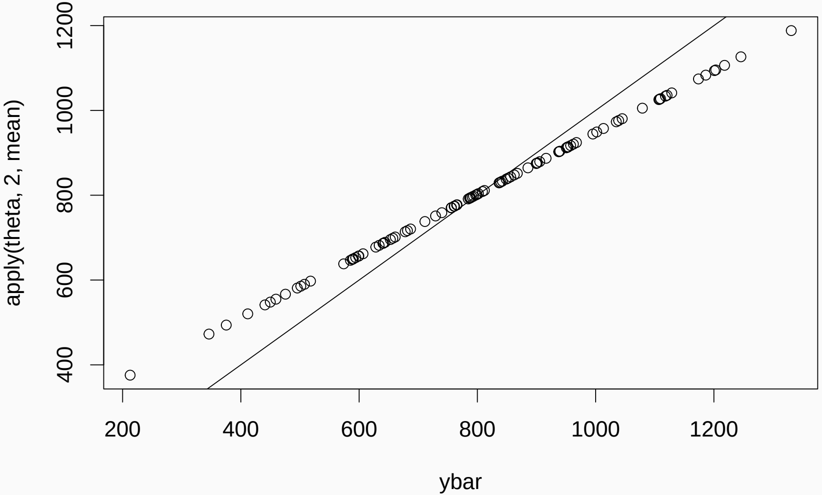
- τ^2 describes variance within a person across trials (trial noise)
- Settings: $a = 600$, $b = 400^2$, $r^2 = 200^2$
- Go code this up in JAGS/stan
- Compare to sample mean.

Going Hierarchical

$$
\begin{aligned}
Y_{ij}|\theta_i, \delta^2 &\sim \mathrm{N}(\theta_i, \tau^2) \\
\theta_i|\mu, \sigma^2 &\sim \mathrm{N}(\mu, \sigma^2) \\
\mu &\sim \mathrm{N}(a, b) \\
\delta^2 &\sim \mathrm{InvGamma}\left(\frac{1}{2}, \frac{r_w^2}{2}\right) \\
\sigma^2 &\sim \mathrm{InvGamma}\left(\frac{1}{2}, \frac{r_b^2}{2}\right)
\end{aligned}
$$

- Settings: $a = 600$, $b = 400^2$, $r_w^2 = r_b^2 = 200^2$
- Go code this up in JAGS/stan
- Compare to sample mean.

Bayes Estimates



Up To Now...

Bayes' Rule:

$$\pi(\theta|Y) = \frac{p(Y|\theta)\pi(\theta)}{}$$

Critique

Problems:

- No sense of what $p(Y)$ means
- Stress posterior rather than the process of updating.
- Strive to minimize influence of priors.
- Separation of estimation and model comparison.

Ratio Form of Bayes Rule

Solution, restate Bayes Rule as follows:

$$\frac{\pi(\theta|Y)}{\pi(\theta)} = \frac{p(Y|\theta)}{p(Y)}.$$

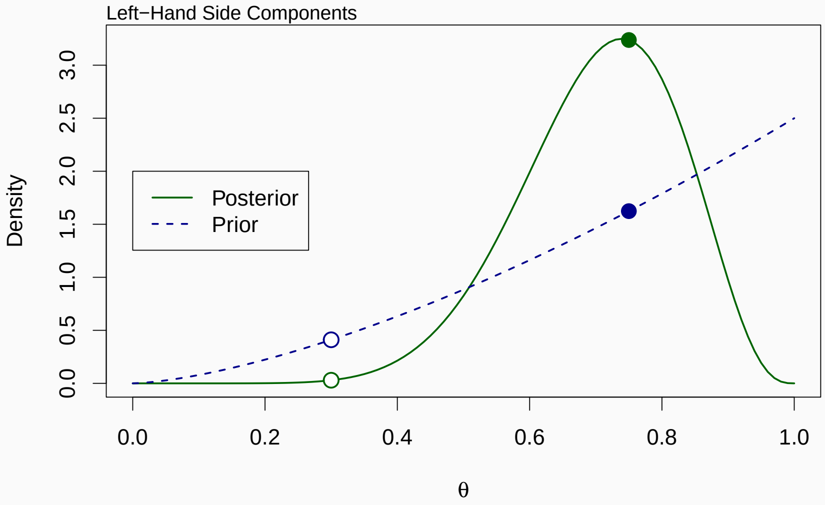
And study what this equation means!

For Our Example

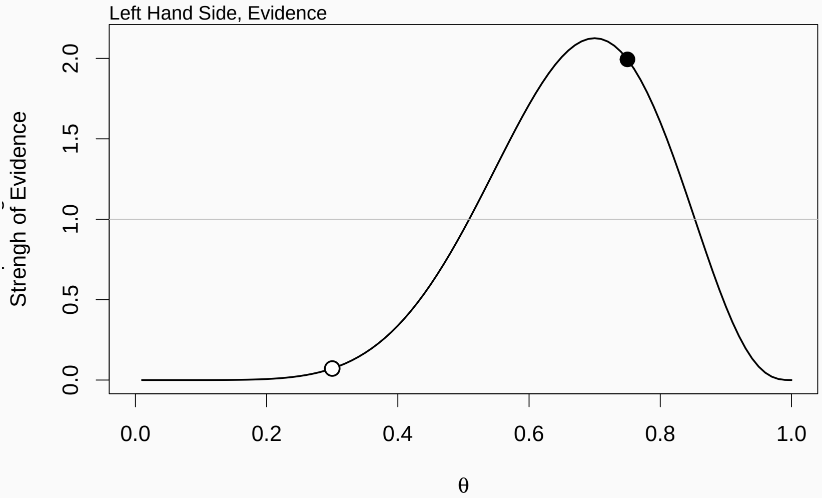
- 7 successes in 10 trials

Left-Hand Side

- Bayes Rule:
$$\frac{\pi(\theta|Y)}{\pi(\theta)} = \frac{p(Y|\theta)}{p(Y)}$$
- Left-Hand Side:
$$\frac{\pi(\theta|Y)}{\pi(\theta)}$$



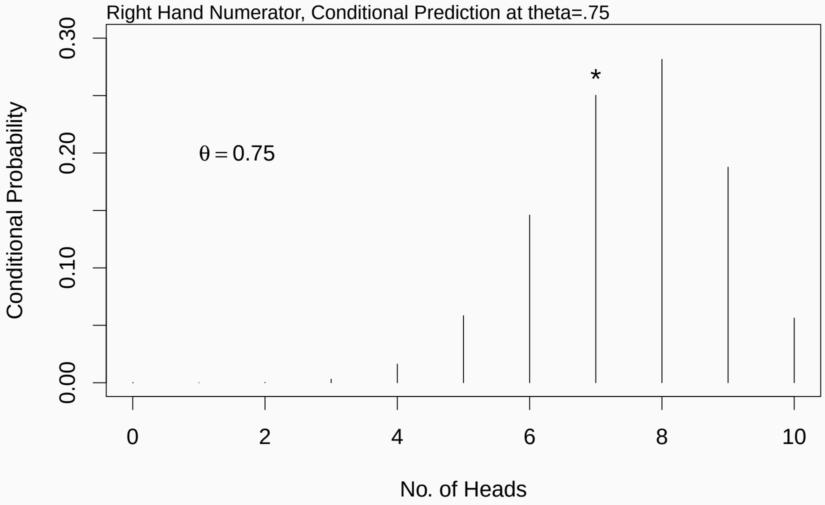
Left−Hand Side Components
Posterior
Prior
Density
θ
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0.0
0.2
0.4
0.6
0.8
1.0



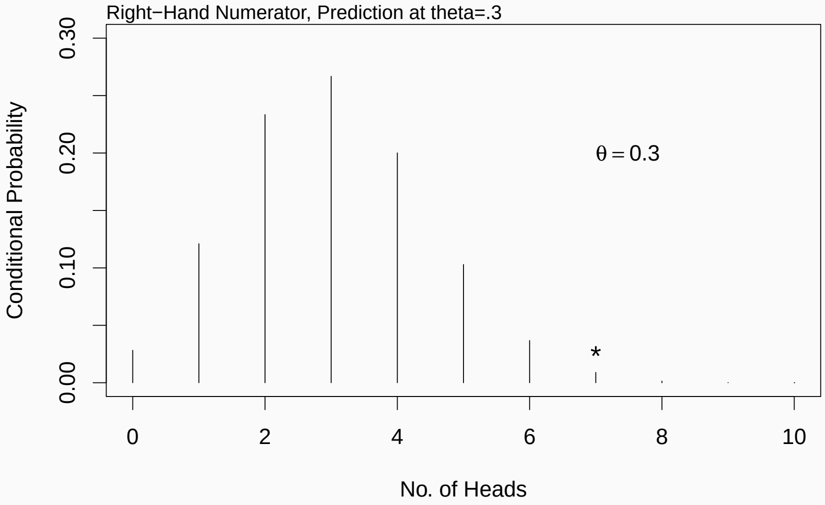
Left Hand Side, Evidence
Strengh of Evidence
0.0
0.5
1.0
1.5
2.0
0.0
0.2
0.4
0.6
0.8
1.0
θ

Right-Hand Side

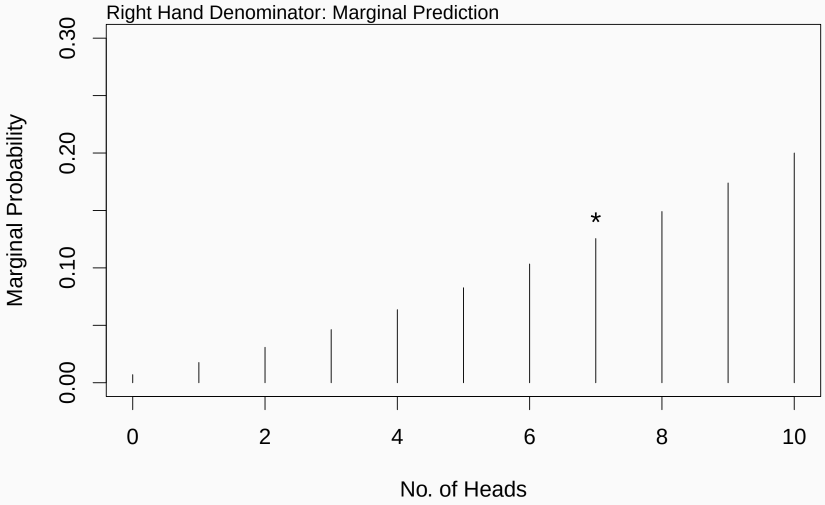
- Bayes Rule:
$$\frac{\pi(\theta|Y)}{\pi(\theta)} = \frac{p(Y|\theta)}{p(Y)}$$
- Right-Hand Side:
$$\frac{p(Y|\theta)}{p(Y)}$$
- $p(Y|\theta)$: Conditional Predictive Accuracy
- $p(Y)$: Marginal Predictive Accuracy

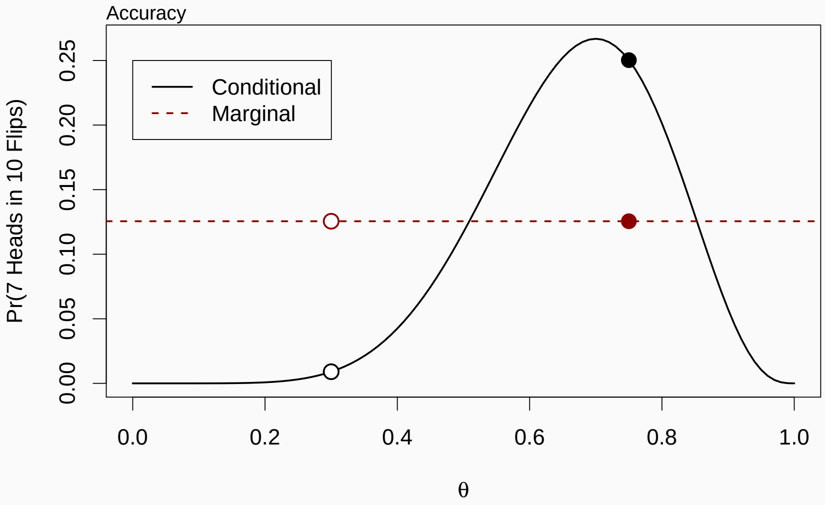


Right Hand Numerator, Conditional Prediction at theta=.75
θ = 0.75
*
Conditional Probability
0.00
0.10
0.20
0.30
0
2
4
6
8
10
No. of Heads

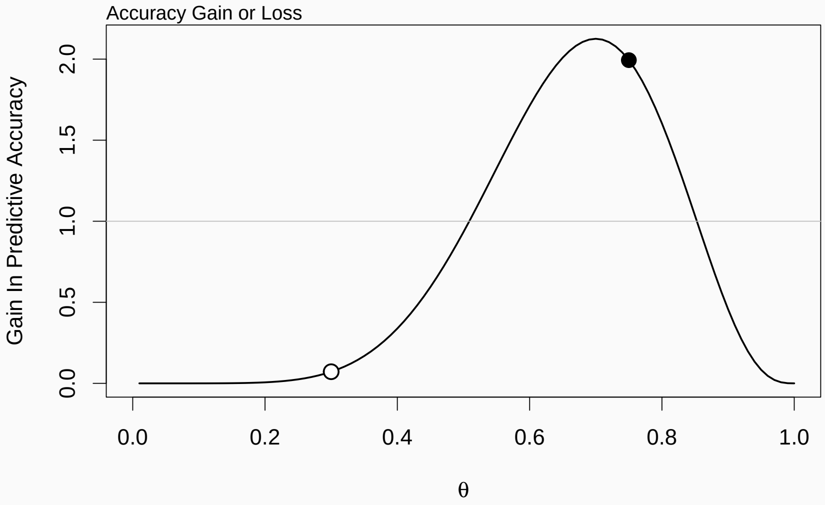


Right−Hand Numerator, Prediction at theta=.3
θ = 0.3
*
Conditional Probability
0.00
0.10
0.20
0.30
No. of Heads
0
2
4
6
8
10


Right Hand Denominator: Marginal Prediction
Marginal Probability
0.00
0.10
0.20
0.30
*
0
2
4
6
8
10
No. of Heads



Accuracy
Conditional
Marginal
Pr(7 Heads in 10 Flips)
0.00
0.05
0.10
0.15
0.20
0.25
0.0
0.2
0.4
0.6
0.8
1.0
θ


Accuracy Gain or Loss
Gain In Predictive Accuracy
0.0
0.5
1.0
1.5
2.0
0.0
0.2
0.4
0.6
0.8
1.0
θ

Evidence for a point is the degree to which it improves the prediction for the observed data.

- "Evidence = Prediction"
- way catchier than "Posterior is Proportional To The Likelihood Times The Prior"

Your Turn: Which is the better parameter value?

- A coin has probability either $\theta = 1/3$ or $\theta = 2/3$.
- We observe Y out of N successes.
- Q: What is the relative evidence of $\theta = 1/3$ vs. $\theta = 2/3$

Need a Hint?

Compute

$$\frac{\frac{\pi(\theta_1|Y)}{\pi(\theta_1)}}{\frac{\pi(\theta_2|Y)}{\pi(\theta_2)}}$$

- where $\theta_1 = 1/3$ and $\theta_2 = 2/3$.
- derive generally and then put in values.

Fair Coin?

We have observed 7 of 10 successes. What is the evidence for a fair coin?

General Model, Model M_a

$$\begin{aligned} Y \mid \theta &\sim \text{Binomial}(\theta, N), \\ \theta &\sim \text{Uniform}(0, 1). \end{aligned}$$

Fair-Coin Model, Model M_b:

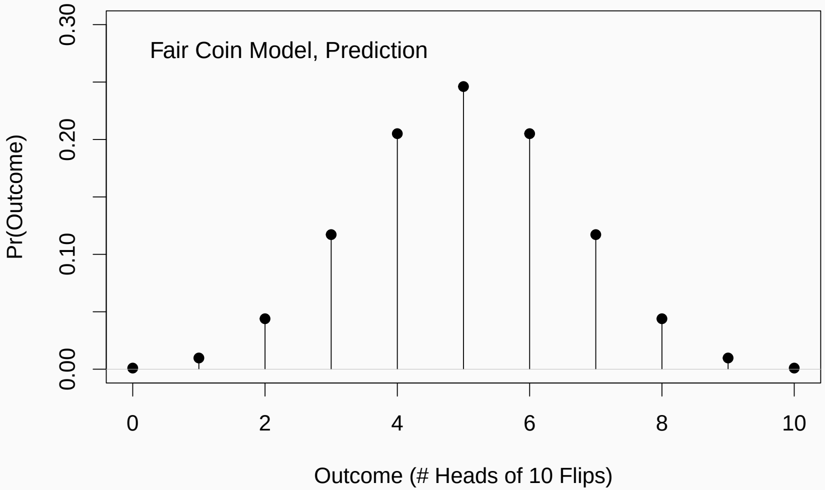
$$Y \sim \text{Binomial}(.5, N).$$

Bayes Rule:

$$\frac{\frac{\pi(M_a|Y)}{\pi(M_a)}}{\frac{\pi(M_b|Y)}{\pi(M_b)}} = \frac{\frac{p(Y|M_a)}{p(Y)}}{\frac{p(Y|M_b)}{p(Y)}} = \frac{p(Y|M_a)}{p(Y|M_b)} = B_{ab}.$$

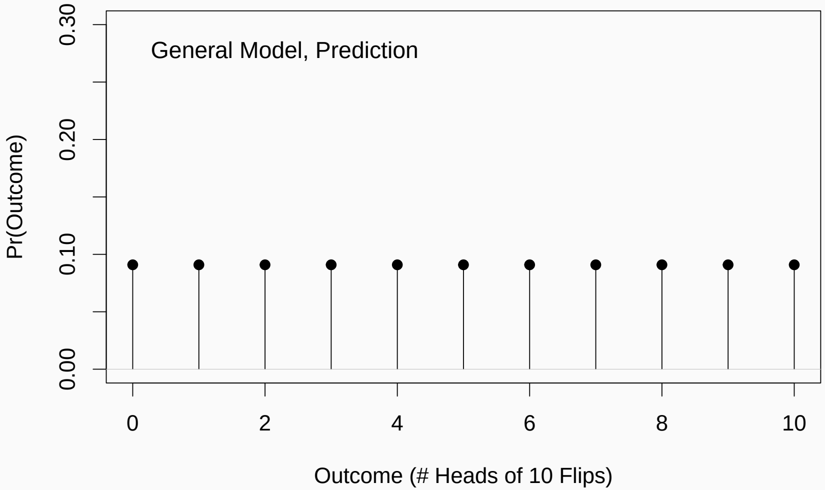
Fair Coin Prediction

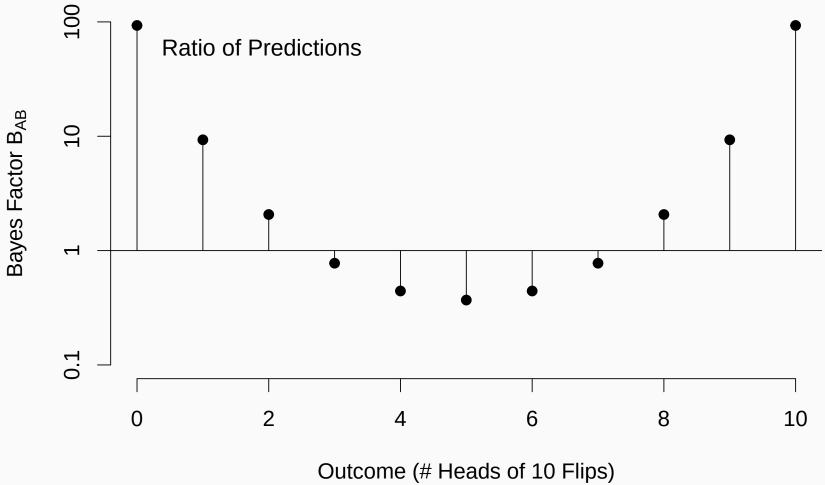
$$Pr(Y = y) = \binom{N}{y} .5^y, 5^{(N-y)}$$


Fair Coin Model, Prediction
Pr(Outcome)
0.00
0.10
0.20
0.30
0
2
4
6
8
10
Outcome (# Heads of 10 Flips)

General Model Prediction

$$Pr(Y = y) = \int_0^1 \binom{N}{y} \theta^y, (1-\theta)^{(N-y)}\, d\theta$$


General Model, Prediction
Pr(Outcome)
0.00
0.10
0.20
0.30
0
2
4
6
8
10
Outcome (# Heads of 10 Flips)



Ratio of Predictions
Bayes Factor B_AB
100
10
1
0.1
0
2
4
6
8
10
Outcome (# Heads of 10 Flips)

Spike and Slab Priors

I have a coin in my hand. It has a true probability of heads of θ. I am going to flip it 1000 times. Guess how many heads of 1000. Let's call your guess G, let's flip it for Y heads. Your error is $e = G - Y$ and I am going to pay you $\max(0, 10000 - e^2)$, so guess good. Now, to make this fair I am going to show flip the coin 1000 times twice. The first time of 1000 flips is to help you formulate your guess; the second time of 1000 flips is for the money.

- if there is 743 of 1000 initially, what is your guess?
- if there are 510 of 1000 initially, what is your guess?

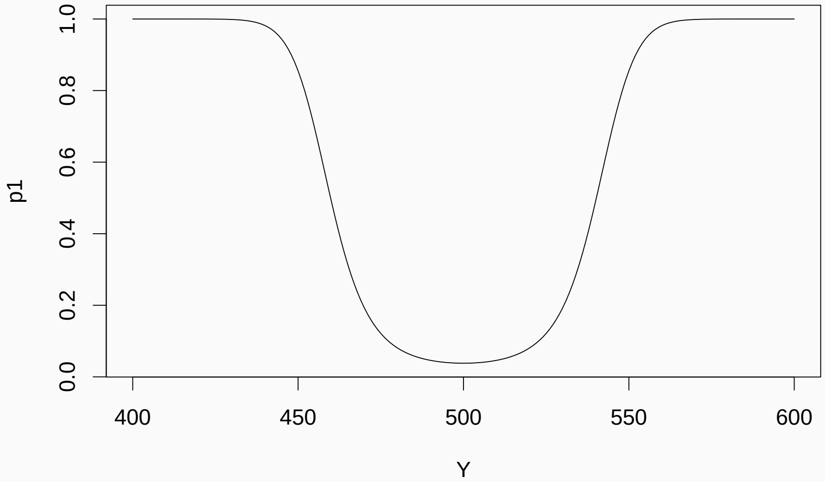
Spike and Slab Priors

$$
\begin{aligned}
Y_0 &\sim \text{Binomial}(\theta, 1000) \\
(\theta \mid \eta = 0) &= .5 \\
(\theta \mid \eta = 1) &\sim \text{Unif}(0, 1) \\
\eta &\sim \text{Bernoulli}(.5)
\end{aligned}
$$

Spike and Slab, Probabilities

$$
\begin{aligned}
BF_{01} = \frac{\Pr(\eta = 0|Y)}{\Pr(\eta = 1|Y)} &= \frac{\Pr(Y|\eta = 0)}{\Pr(Y|\eta = 1)} \times \frac{Pr(\eta = 0)}{Pr(\eta = 1)} \\
&= \frac{\Pr(Y|\eta = 0)}{\Pr(Y|\eta = 1)} \\
&= 1001 \times \binom{N}{Y} \times .5^N
\end{aligned}
$$

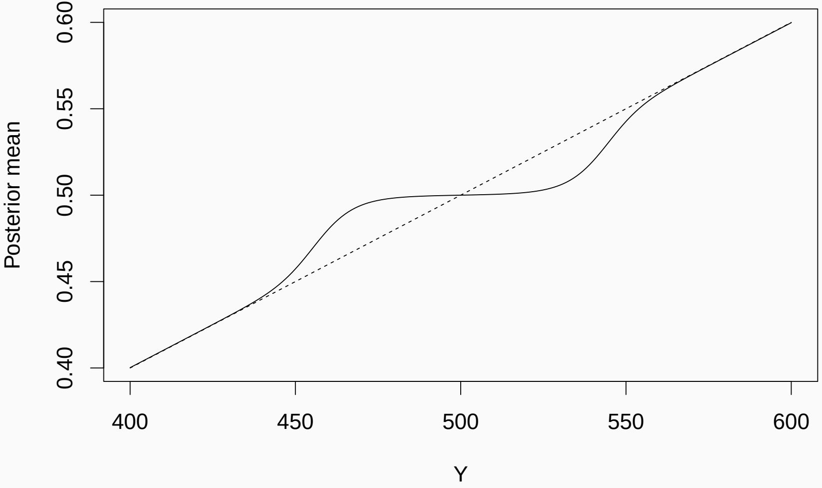
- $\Pr(\eta = 0|Y) = BF_{01}/(BF_{01} + 1)$
- $\Pr(\eta = 1|Y) = 1 - \Pr(\eta = 0|Y)$



1.0
0.8
0.6
0.4
0.2
0.0
p1
400
450
500
550
600
Y

Estimates, Posterior Mean

$$
\begin{aligned}
E(\theta \mid Y) = & E(\theta \mid Y, \eta = 0) \times \Pr(\eta = 0 \mid Y) + \\
& E(\theta \mid Y, \eta = 1) \times \Pr(\eta = 1 \mid Y). \\
E(\theta \mid Y) = & .5 \times \Pr(\eta = 0 \mid Y) + \\
& \frac{Y+1}{1002} \times \Pr(\eta = 1 \mid Y)
\end{aligned}
$$



Posterior mean
0.40
0.45
0.50
0.55
0.60
400
450
500
550
600
Y

JAGS spike and slab code

```
spikeSlab.mod<-"
model{
eta ~ dbern(.5)
theta ~ dunif(ifelse(eta==0,.499,0),
              ifelse(eta==0,.501,1))
Y ~ dbinom(theta,1000)}
"
```

Spike and Slab in Modern Regression

Let Y_n be an observation, $i = 1, \ldots, N$. Here is a linear model for R covariates:

$$Y_n = \mu + X_{1n}\alpha_1 + \ldots + X_{Rn}\alpha_R + \epsilon_n$$

Suppose R is large, potentially $R \gg N$. Yet, we know, several α's are at or near zero.

$$\begin{aligned}
(\alpha_r \mid \eta_r = 0) &= 0 \\
(\alpha_r \mid \eta_r = 1) &\sim \text{Normal}(0, b) \\
\eta_r &\sim \text{Bernoulli}(\pi) \\
\pi &\sim \text{Beta}(c, d)
\end{aligned}$$

- Potential to zero out many, many coefficients.
- VSS (variable selection search)

Adaptive Priors

Modern shrinkage is *adaptive* and, in the Bayesian case, reflects mixture priors where the mixture is over different models (or different priors, it's all the same).

- Spike and slab is a dichotomous mixture
- Lasso is a continuous mixture