Lecture 2: Posteriors

Jeff Rouder

June, 2022

What Makes A Bayesian A Bayesian

- Everyone uses the Law of Conditional Probability wherever they will use Probability
- Freuqentists use probability on data (observations) but not on models or parameters
- Bayesians use probability on everything. Models, parameters, data, etc.
- Bayes rule is the Law of Conditional Probability applied to parameters and models and data.
- Bayesians use Bayes rule *always*

Motivating Problem

What is the probability that toast falls butter side down? Let's suppose we have observed 7 successes (butter-side down) in 10 trials. What does that tell us about buttered toast?

Conventional Specification

Let Y be a random variable that denotes the number of successes. We may model the random variable Y as a binomial. It is conventional to write:

$$Y \sim \text{Binomial}(\theta, n)$$

where θ is the parameter of interest and n is the number of trials.

The Complete Frequentist Specification

The above notation hides something though. In a frequentist framework, we should write:

$$Y(\theta) \sim \text{Binomial}(\theta, n)$$

to emphasize that θ is a fixed parameter and all models are defined as a function of (unknown) θ.

The Bayesian Specification

We should write:

$$Y|\theta \sim \text{Binomial}(\theta, n)$$

The data are stated conditional on parameters rather than as a function of parameters.

It is also clear this is incomplete. Someone has to tell us which θ.

The Bayesian Specification

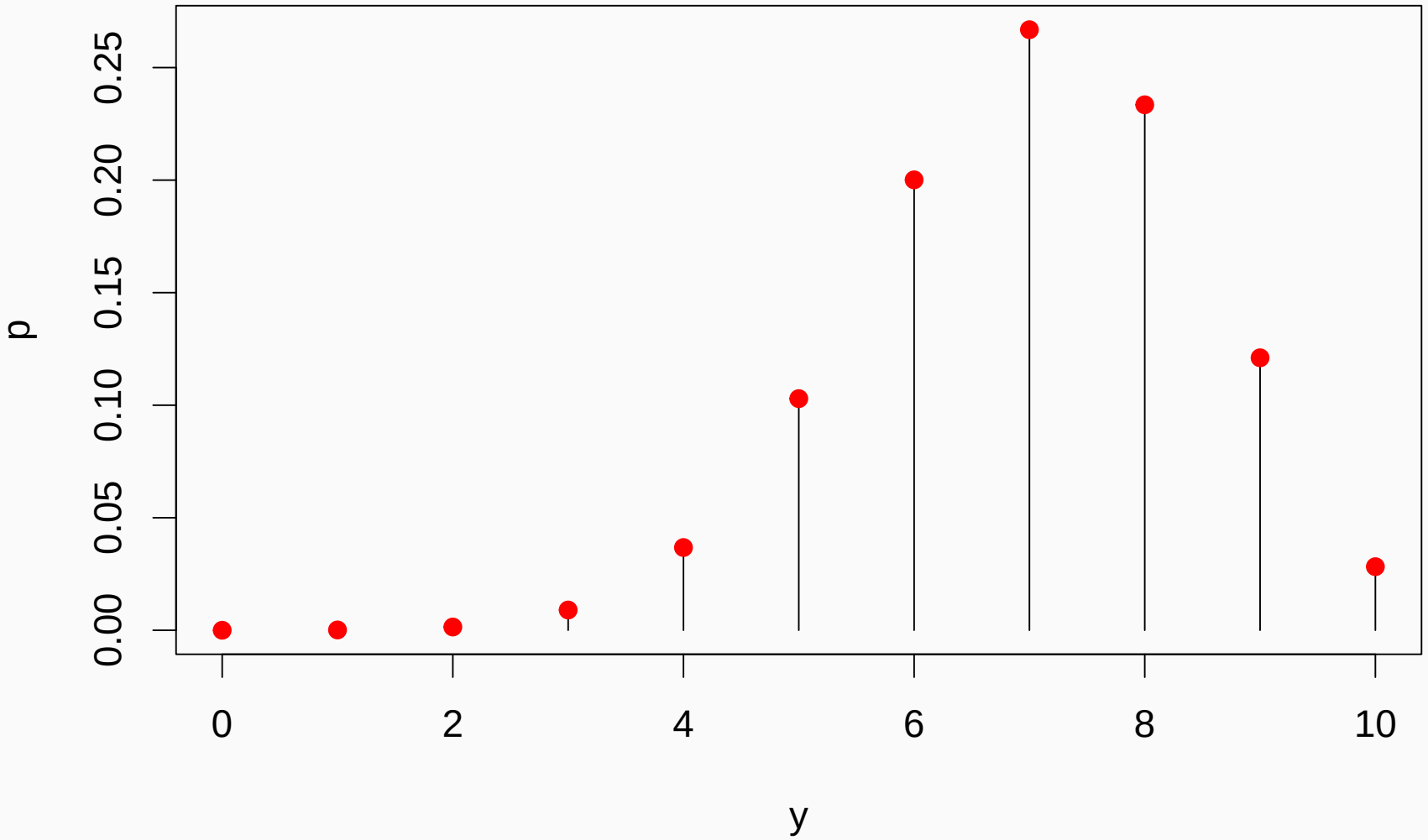
In Bayesian analysis, we must *complete* the specification.

$$Y|\theta \sim \text{Binomial}(\theta, n)$$

$$\theta \sim \text{Some Distribution}$$

Binomial Distribution on Data

Let's look at some predictions of the model for known θ. Here is $\theta = .7$:



Notation

The probability mass function, `dbinom()` in this case, in classical statistics is

$$f(y;\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}.$$

It Bayesian statistics, it is

$$Pr(Y = y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}.$$

Bayesian Analysis

Bayes' Rule for this case:

$$f(\theta|y) = \frac{Pr(Y = y|\theta)}{Pr(Y = y)} f(\theta)$$

1. $f(\theta)$, prior or marginal distribution of parameters. Our beliefs before seing the data. Flat, $\theta \sim \text{Uniform}(0, 1)$.
2. $Pr(Y = y|\theta)$, conditional probability of observed data (related to likelihood). Known from specification of model:

$$Pr(Y = y|\theta) = \binom{n}{y} \theta^y (1 - \theta)^{n-y}$$

3. $Pr(Y = y)$. Marginal probability of data. Uniquely Bayesian quantity without a frequentist analog.

$Pr(Y = y)$

Law of Total Probability

$$Pr(Y = y) = \int_0^1 Pr(Y = y|\theta) f(\theta)\, d\theta$$

R implementation

```
prior=function(theta)
  dunif(theta,0,1)
p.data.g.theta = function(theta,y,n)
  dbinom(y,n,theta)
integrand=function(theta,y,n)
  prior(theta)*p.data.g.theta(theta,y,n)
p.data=function(y,n)
  integrate(integrand,
            lower=0,
            upper=1,
            y=y,n=n)$value

print(p.data(7,10))
```

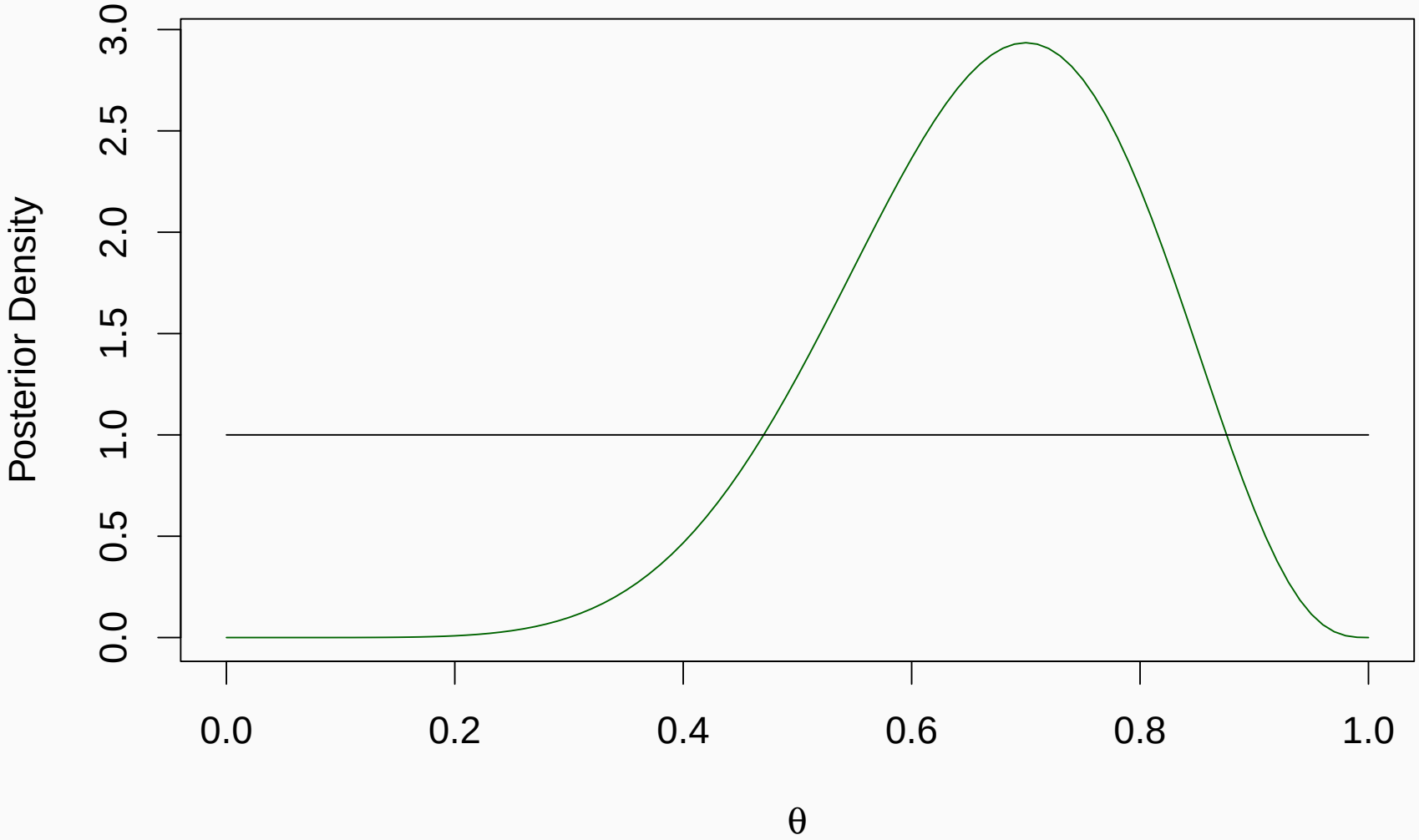
```
## [1] 0.09090909
```

Bayes Rule In R

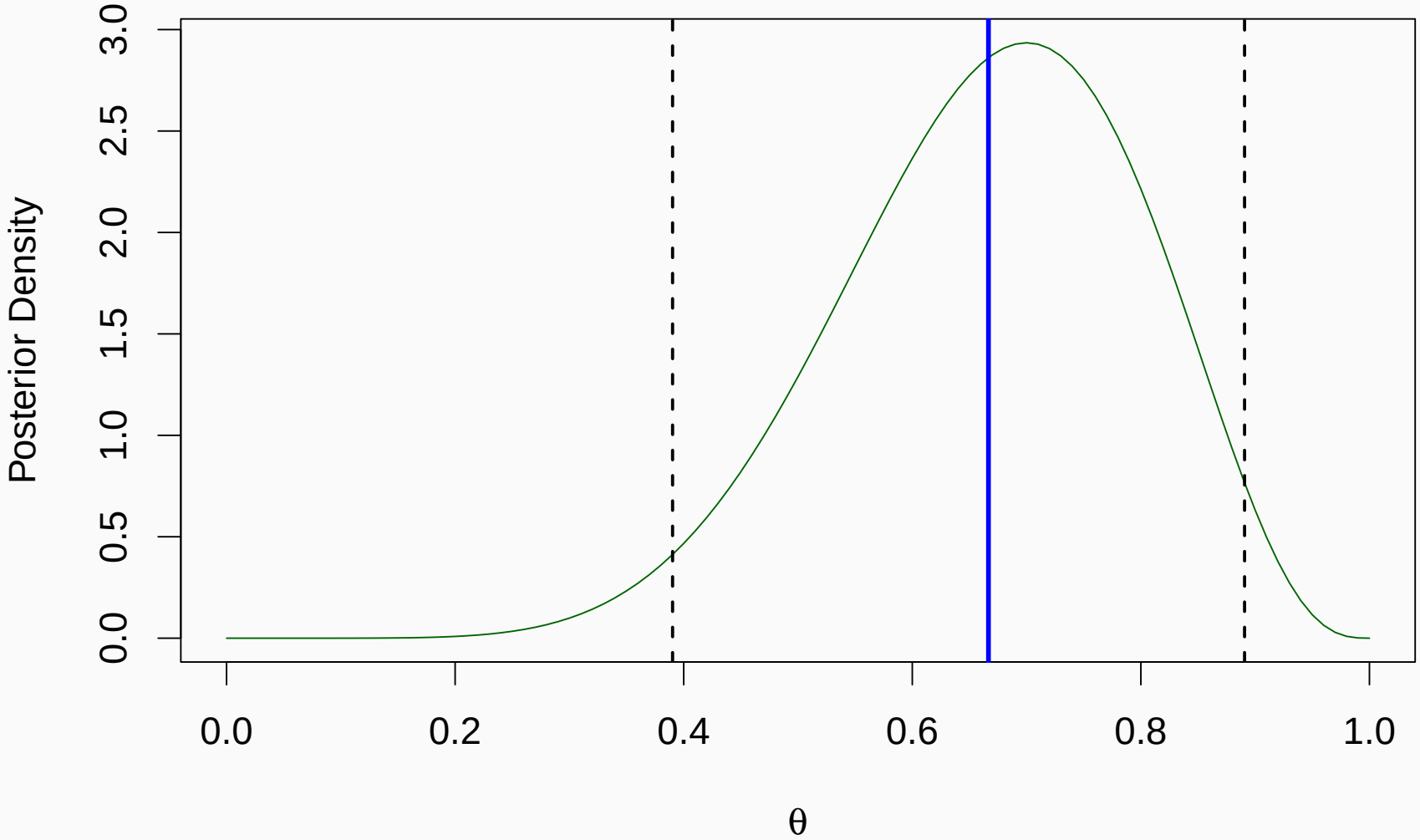
$$f(\theta|y) = \frac{Pr(Y = y|\theta)}{Pr(Y = y)} f(\theta)$$

```
posterior=function(theta,y,n)
  p.data.g.theta(theta,y,n) * prior(theta) / p.data(y,n)
```

Bayes Rule in R



Characterizing Posterior (not a fan)



Encoding Other Beliefs

Meet the beta distribution.

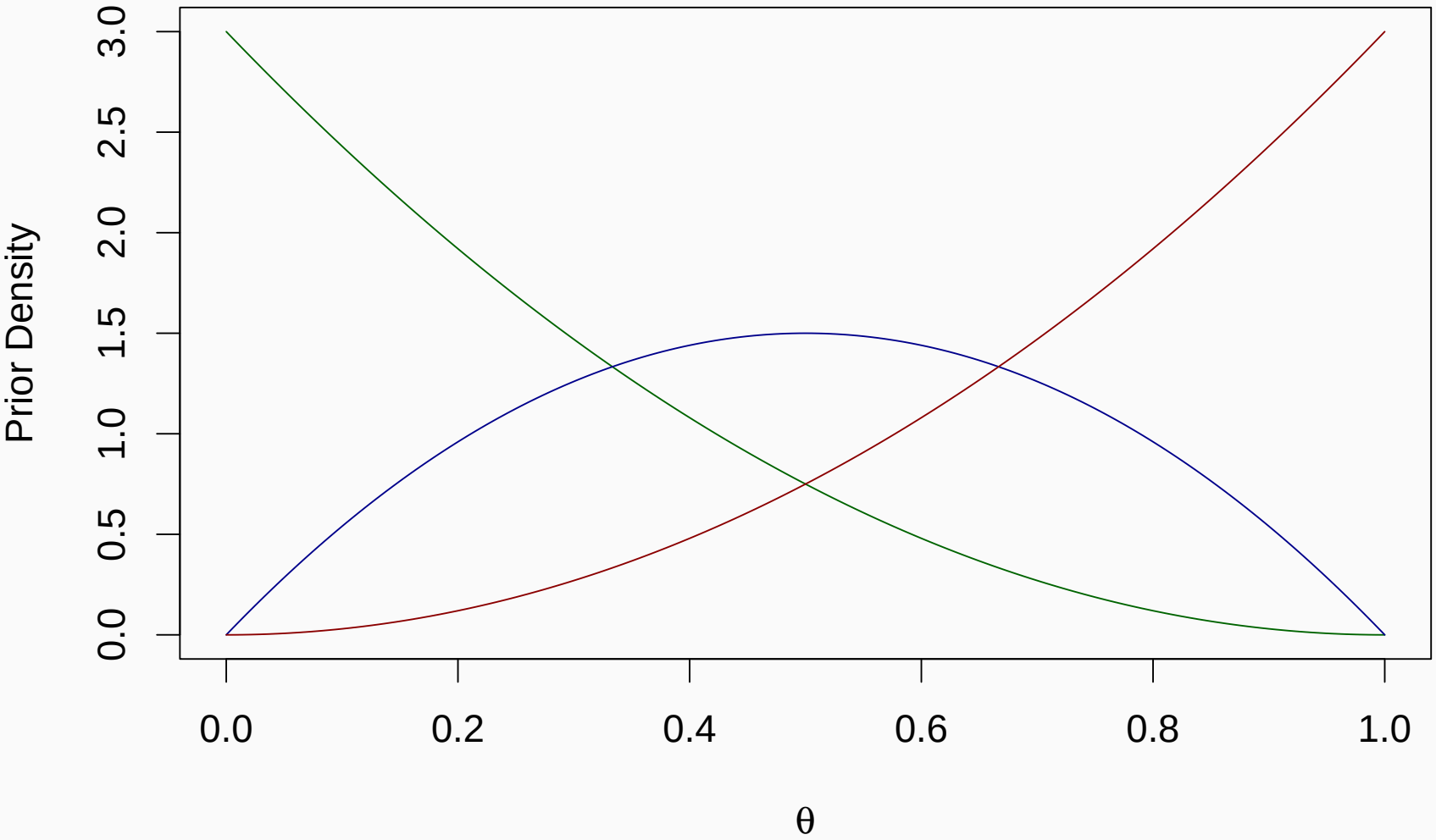
- Flexible Form That Lives on [0,1]
- Two parameters determine the shape
- beta(1,1) is uniform

Your Turn

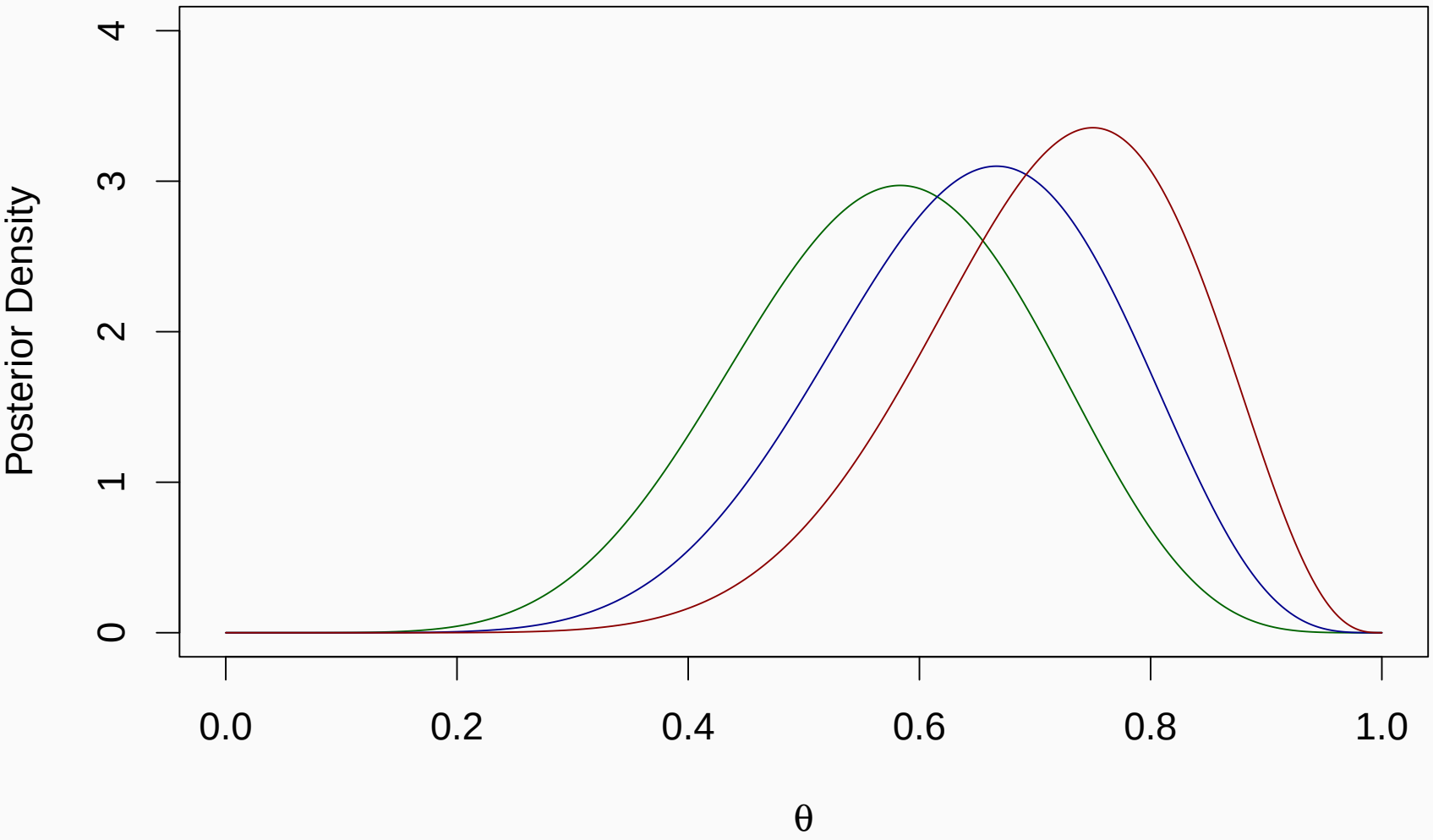
Play with the following code, what do parameters a and b do? Obey $a > 0$ and $b > 0$.

```
a=1
b=1
p=seq(0,1,.001)
plot(p,dbeta(p,a,b),typ='l')
```

Encoding Other Beliefs



Encoding Other Beliefs



Your Turn

Adapt my previous example with the uniform prior for the beta prior. Consider two radically different priors, say Beta(10,.5) and Beta(.5,10). What is the effect for 7 successes out of 10 flips? How about 70 out of 100? 700 out of 1000? For estimation, the effect of the prior seemingly fades as the sample size gets larget (often but not always true).

Two Interpretations: Posterior vs. Updating

You have naturally followed the posterior interpretation of Bayes rule. There is another, perhaps more fruitful interpretation that focuses on updating. The amount of updating rather than the posterior becomes the target of inquiry that informs how we learn about phenomena.

Both interpretations are important, and let's stick with the posterior interpretation for now. Under this interpretation, we focus on the posterior as our target of inquiry.

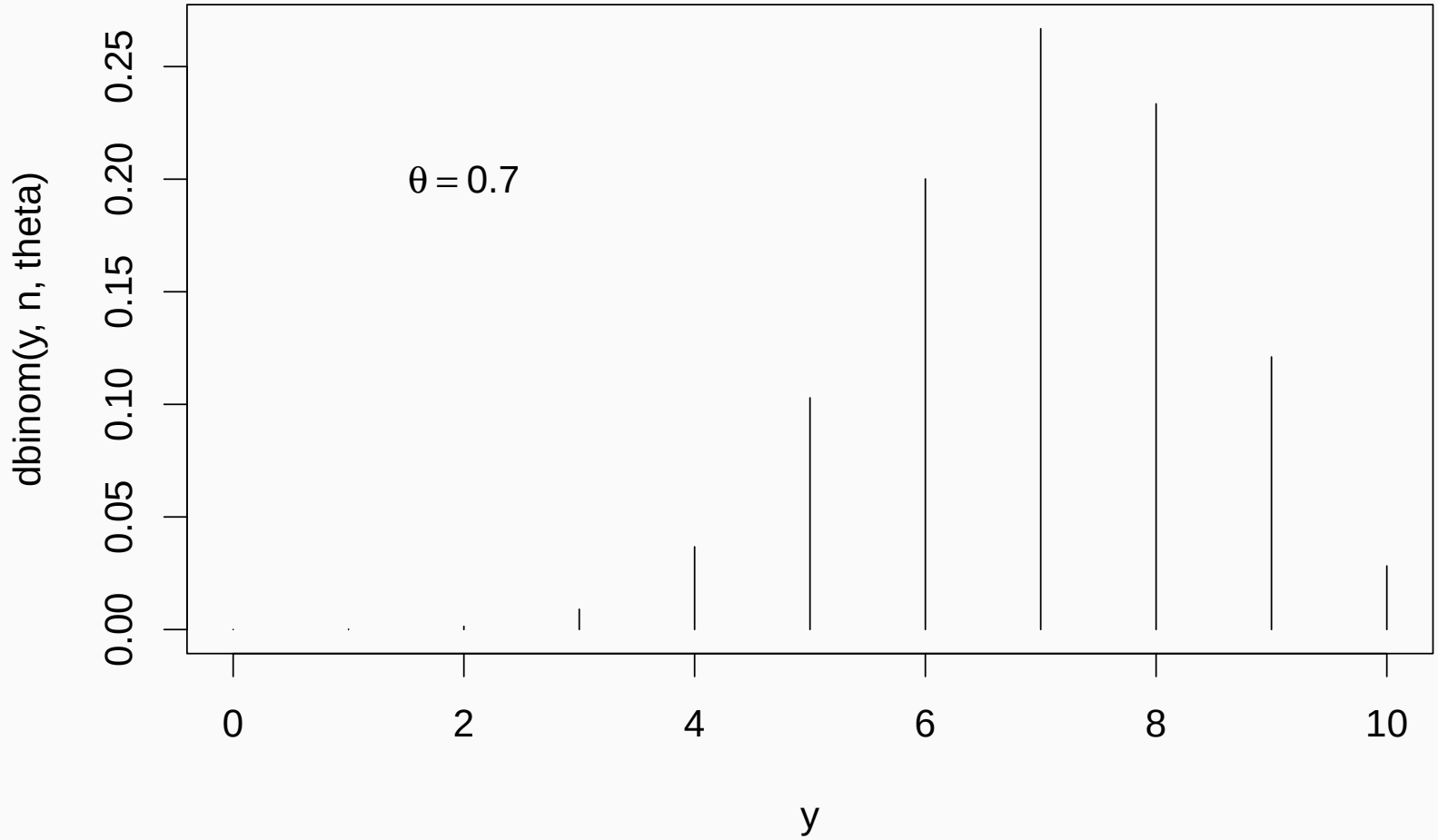
Likelihood, a means to posteriors

Here is Bayes rule again:

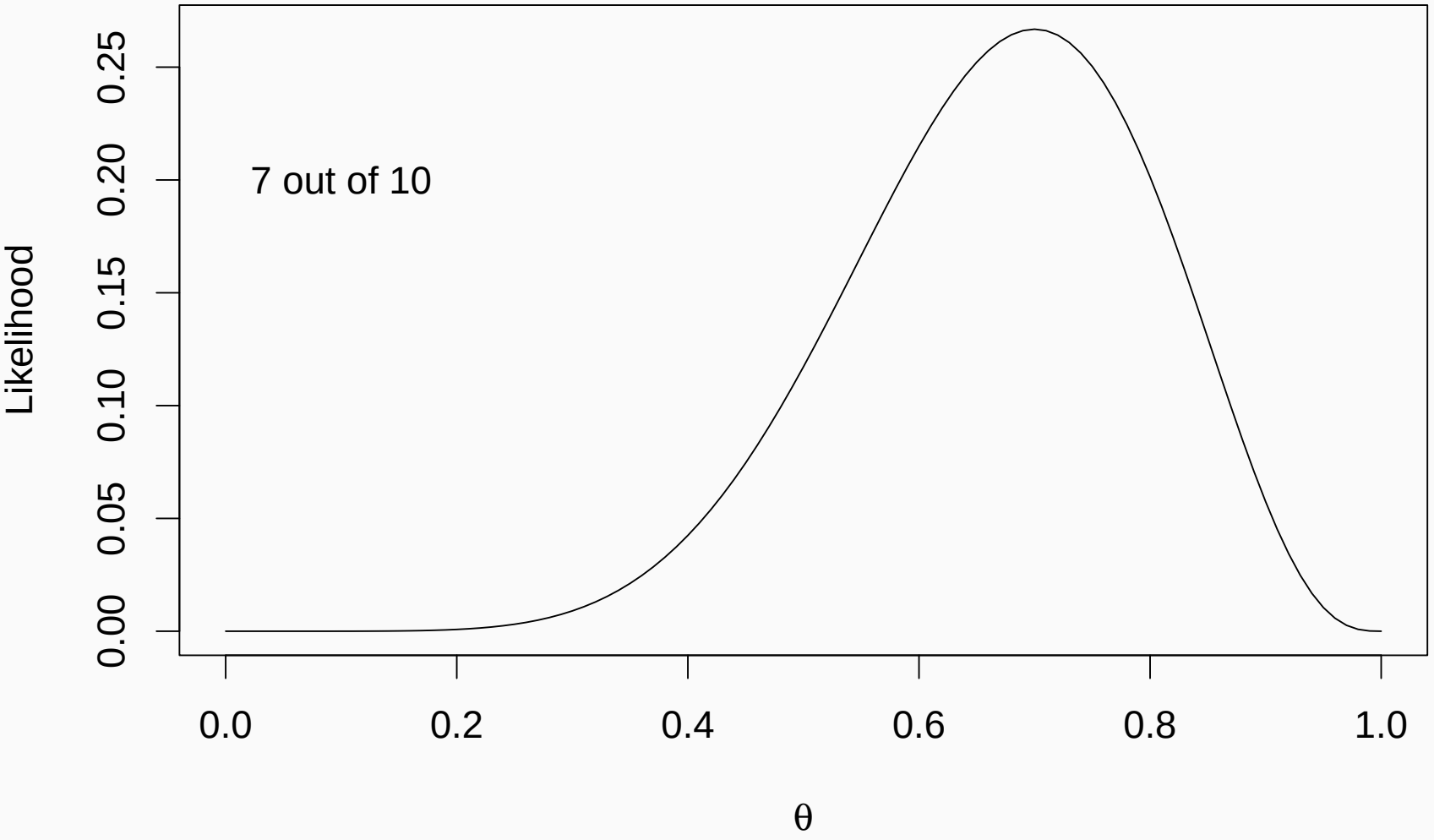
$$f(\theta|y) = \frac{Pr(Y = y|\theta)}{Pr(Y = y)} f(\theta)$$

We are interested in the posterior, $f(\theta|y)$, which is a function of θ for fixed y. So we can treat the RHS as a function of θ too.

1. Probility density, $Pr(Y = y|\theta)$, is treated as function of y



1. Probility density, $Pr(Y = y|\theta)$, may be treated as function of θ



Likelihood

For binomial and observed data,

$$L(\theta; y, N) = Pr(Y = y|\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$

Going Further

- $Pr(Y = y)$, the denominator, does not depend on θ. It is a constant.
- Constant of proprtionality, $\propto$, $y = Kx$ or $y \propto x$.

$$f(\theta|y) \propto L(\theta; y) \times f(\theta)$$

Computational Form of Bayes Rule

"Posterior is proportional to the likelihood times the prior"

Analysis of the Normal

The normal distribution is quite flexible, popular, and convenient. It is fairly ubiquitous and understanding how to analyze it is essential. The goal here is to discuss how to do so.

Let $Y_1, Y_2, \ldots, Y_N$ be a sequence of N random variables. The normal model is given by

$$Y_i \sim \text{Normal}(\mu, \sigma^2)$$

where the normal is a two-parameter symmetric distribution with parameters μ and σ^2 and density given by:

$$f(y) = f(y|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

Posterior of μ for Known σ^2

The goal here is to estimate μ conditional on the observed data $y_1, \ldots, y_N$.

1. We need prior beliefs on μ, let's use $\mu \sim \text{Normal}(a, b)$. Is b standard deviation or variance? These values of a and b are set before hand.
2. To update our beliefs we use, wait for it. . . . . , Bayes Rule:

$$f(\mu|\sigma^2, y_1, \ldots, y_N) \propto f(y_1, \ldots, y_N|\mu, \sigma^2) \times f(\mu|\sigma^2)$$

The term $f(\mu|\sigma^2, y_1, \ldots, y_N)$ is called the conditional posterior distribution of μ.

Posterior of μ for Known σ^2

3. The prior on μ holds for any σ^2, hence,

$$f(\mu|\sigma^2) = f(\mu) = \frac{1}{\sqrt{2\pi b}} \exp\left(-\frac{(\mu - a)^2}{2b}\right).$$

Posterior of μ for Known σ^2

4. The remaining part is $f(y_1, \ldots, y_N|\mu, \sigma^2)$. If there was one piece of datum, y_1, the density is

$$f(y_1|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_1-\mu)^2}{2\sigma^2}\right),$$

If there were two observations, y_1 and y_2, then

$$f(y_1, y_2|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_1-\mu)^2}{2\sigma^2}\right) \times \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_2-\mu}{2\sigma^2}\right.$$

Collecting terms yields:

$$f(y_1, y_2|\mu, \sigma^2) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(y_1-\mu)^2 + (y_2-\mu)^2}{2\sigma^2}\right).$$

So, continuing,

$$f(y_1, \ldots, y_N|\mu, \sigma^2) = \frac{1}{(2\pi)^{N/2}(\sigma^2)^{N/2}} \exp\left(-\frac{\sum_i (y_i-\mu)^2}{2\sigma^2}\right).$$

5. Putting it altogether:

$$f(\mu|\sigma^2, y_1, \ldots, y_N) \propto f(y_1, \ldots, y_N|\mu, \sigma^2) \times f(\mu|\sigma^2)$$

$$f(\mu|\sigma^2, y_1, \ldots, y_N) \propto \frac{1}{(2\pi)^{N/2}(\sigma^2)^{N/2}} \exp\left(-\frac{\sum_i (y_i - \mu)^2}{2\sigma^2}\right) \times \frac{1}{\sqrt{2\pi b}} \exp\left(-\frac{(\mu - a)^2}{2b}\right).$$

6. The above may be reduced (a small miracle occurs here, see Rouder and Lu, 2005) to

$$f(\mu|\sigma^2, y_1, \ldots, y_N) \propto \frac{1}{\sqrt{2\pi v}} \exp\left(-\frac{(y - cv)^2}{2v}\right),$$

where

$$v = \left(\frac{N}{\sigma^2} + \frac{1}{b}\right)^{-1}$$

and

$$c = \left(\frac{N\bar{y}}{\sigma^2} + \frac{a}{b}\right)$$

You may notice that this is the density of a normal with mean cv and variance v, e.g.,

$$\mu|\sigma^2; y_1, \ldots, y_N \sim \text{Normal}(cv, v).$$

Suppose $b = \infty$? Do the math please.

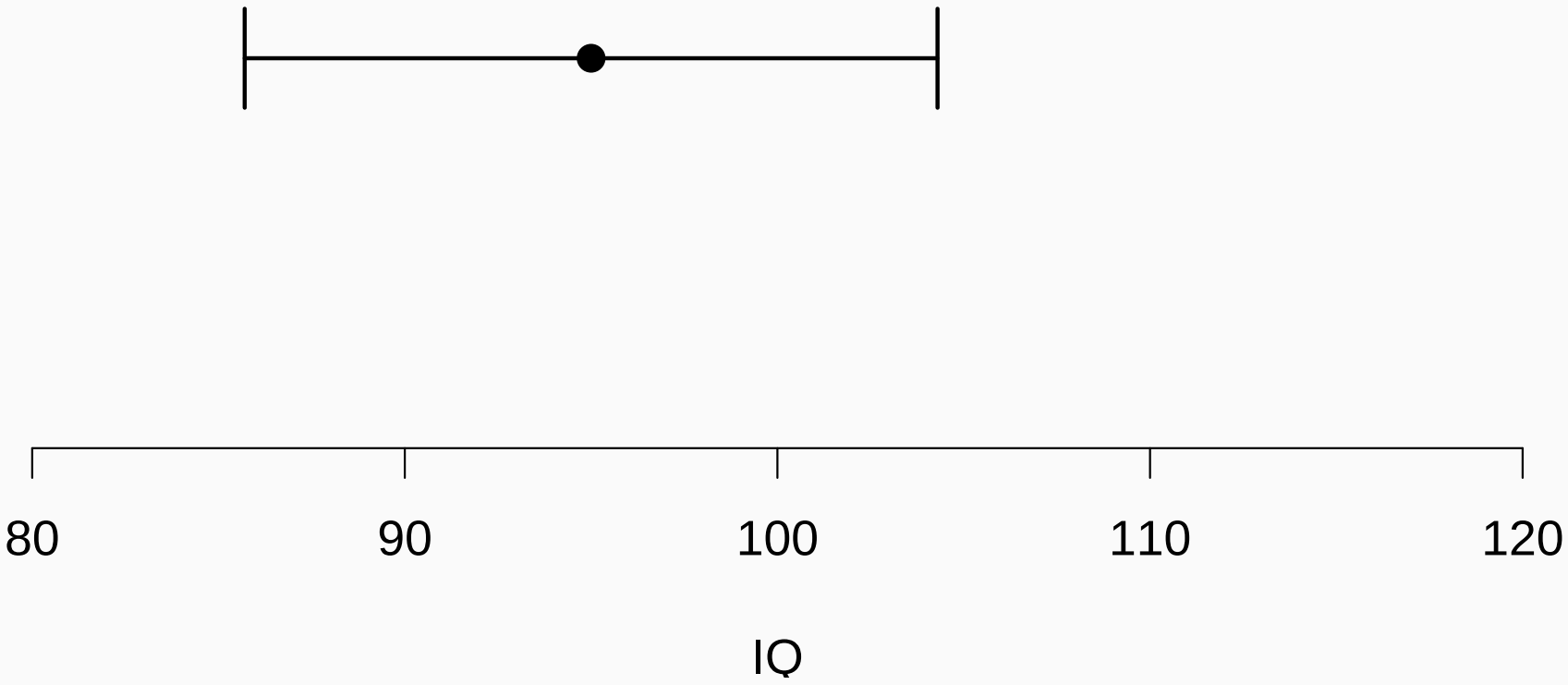
Example, Smarties:

Suppose we wished to know the effects of "Smarties," a brand of candy, on IQ. Certain children have been known to implore their parents for Smarties with the claim that it assuredly makes them smarter. Let's assume for argument's sake that we have fed Smarties to a randomly selected group of school children, and then measured their IQ, which we model as a normal. Let's assume the test we used has been normed so that $\sigma^2 = 15^2 = 225$.

Data

- $N = 10$ and
- $\bar{y} = 95$.

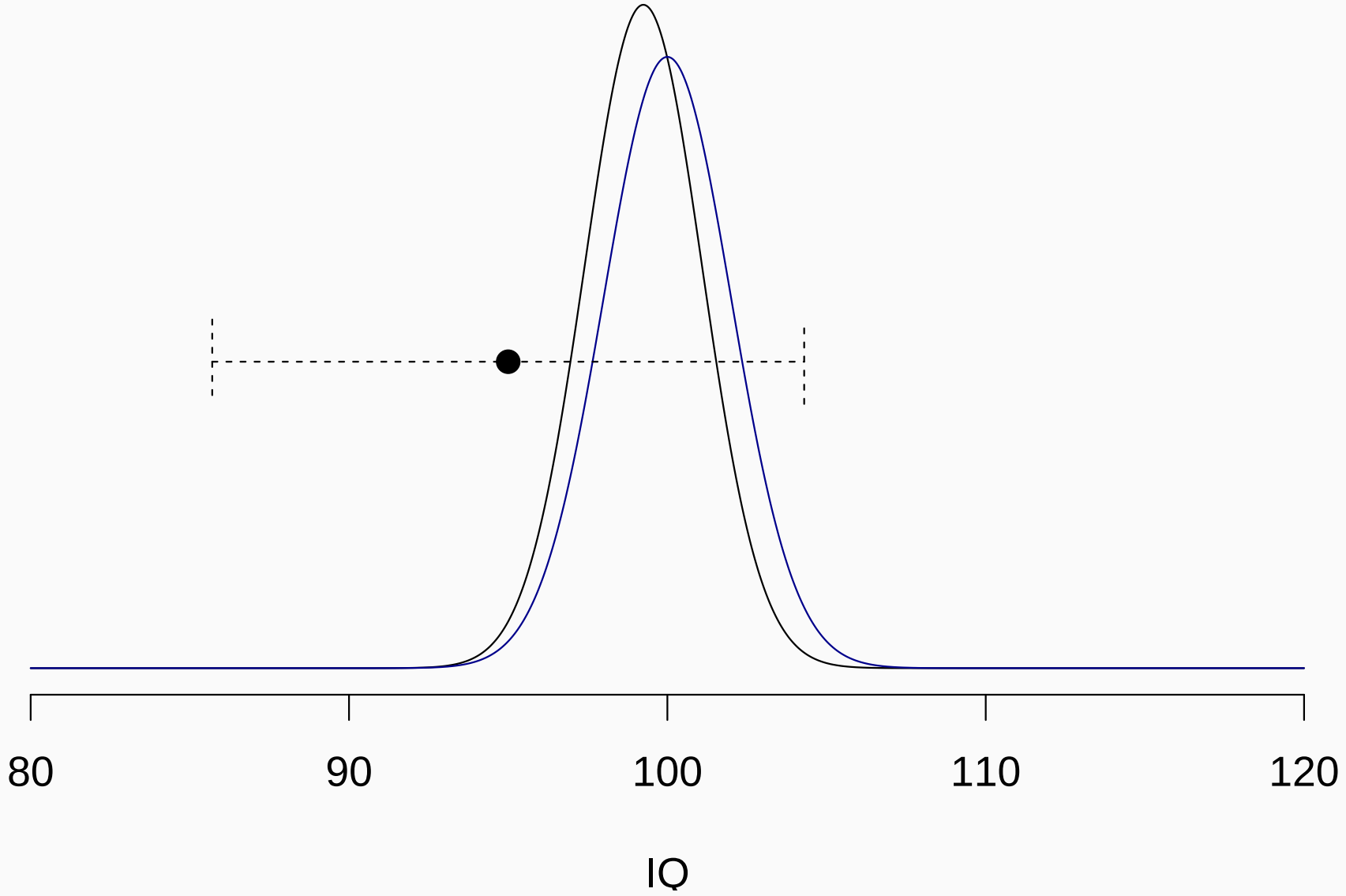
Frequentist Analysis, Mean + 95%CI



Bayesian Analysis, Prior + Posterior

I am pretty sure that even if there is an effect, it will be small, say on the order of 2 IQ points. So I am going to choose $a = 100$ and $b = 2^2$ as a prior. That means true IQ can reasonable vary between say 96 and 104 ($\pm$ 2 standard deviations).

Bayesian Analysis, Prior + Posterior



Suppose both μ and σ^2 are unknown

$$Y_i \sim \text{Normal}(\mu, \sigma^2)$$

$$\mu \sim \text{Normal}(a, b)$$

$$\sigma^2 \sim \text{Imverse Gamma}(q, s)$$

Your Turn

What is an inverse gamma distribution?

Use Wiki

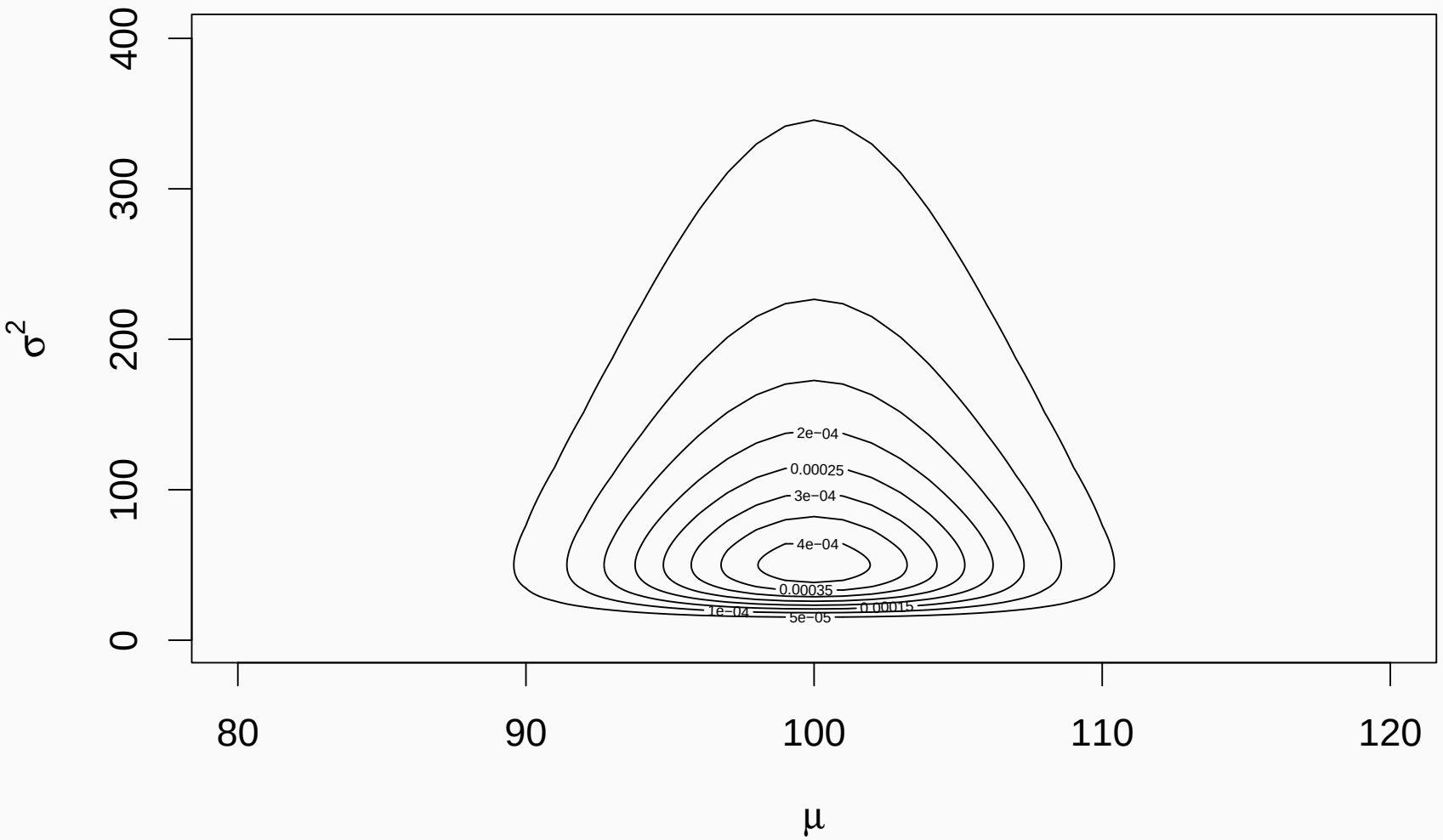
Suppose both μ and σ^2 are unknown

We can still use, wait for it, Bayes rule:

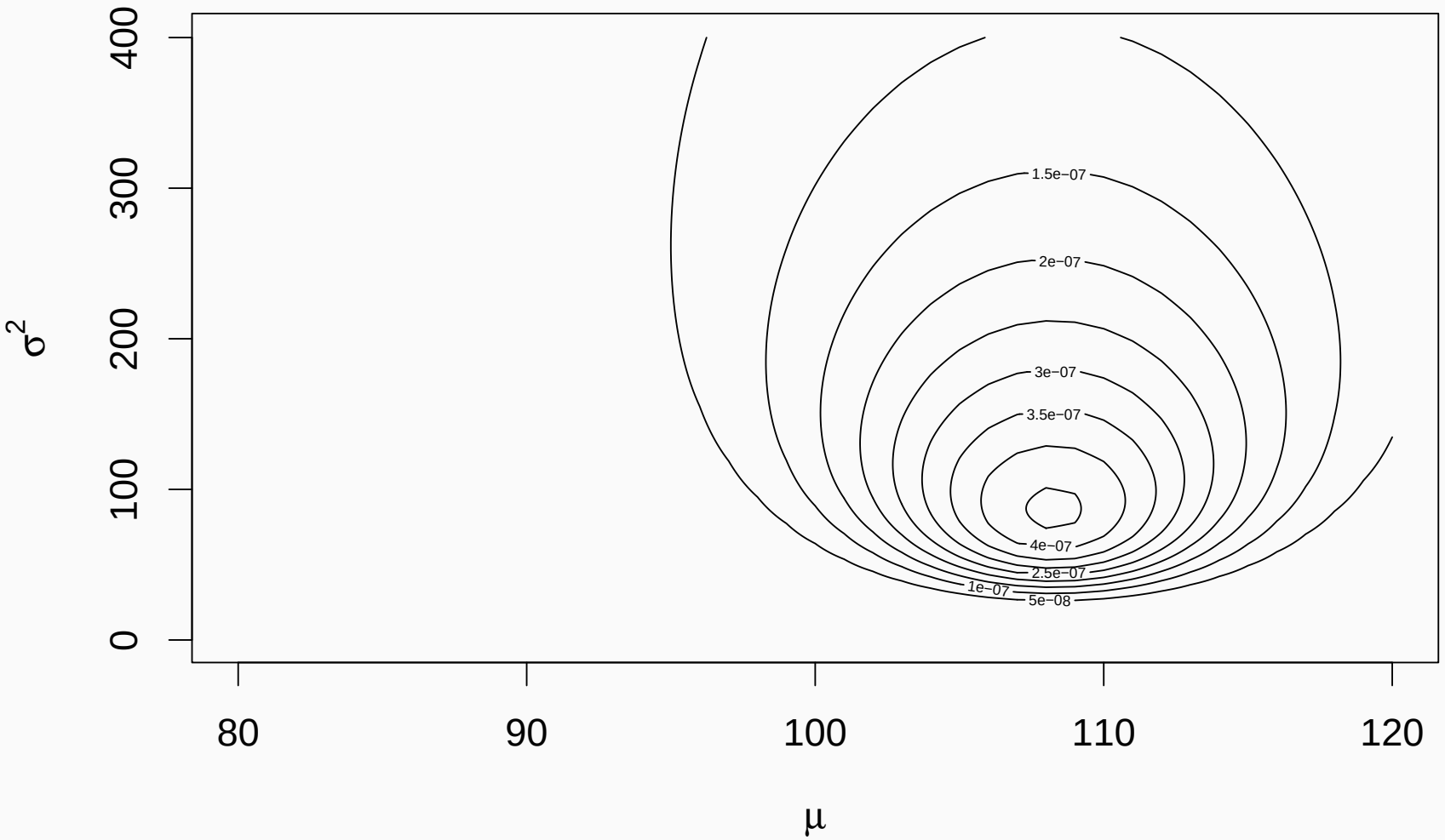
$$f(\mu, \sigma^2 | y_1, \ldots, y_N) \propto f(y_1, \ldots, y_N | \mu, \sigma^2) \times f(\mu, \sigma^2)$$

We are just going to do the rest in R rather than in math. So, lets do IQ with observations (105,100,124,104)

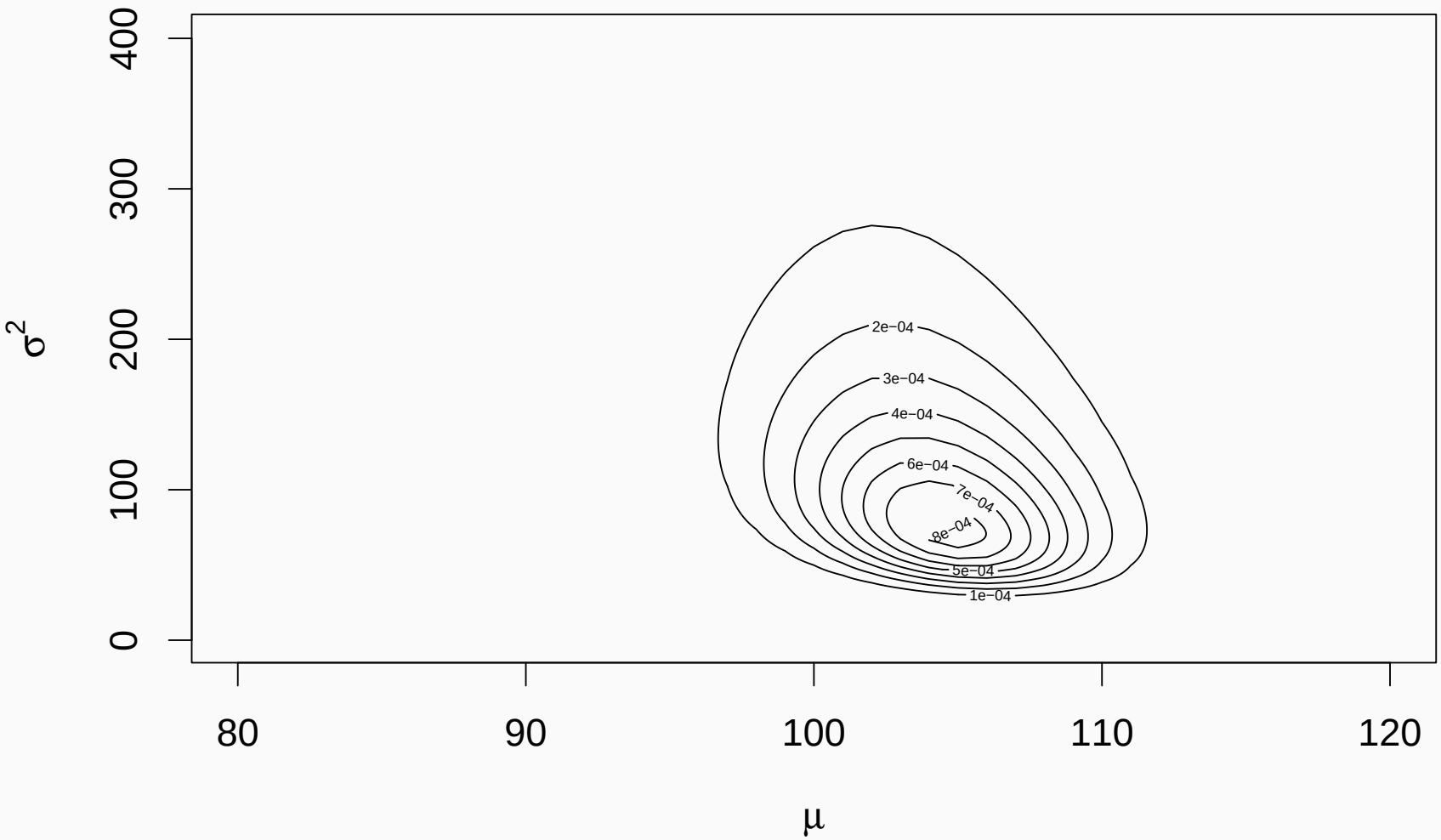
Joint Prior over μ and σ^2



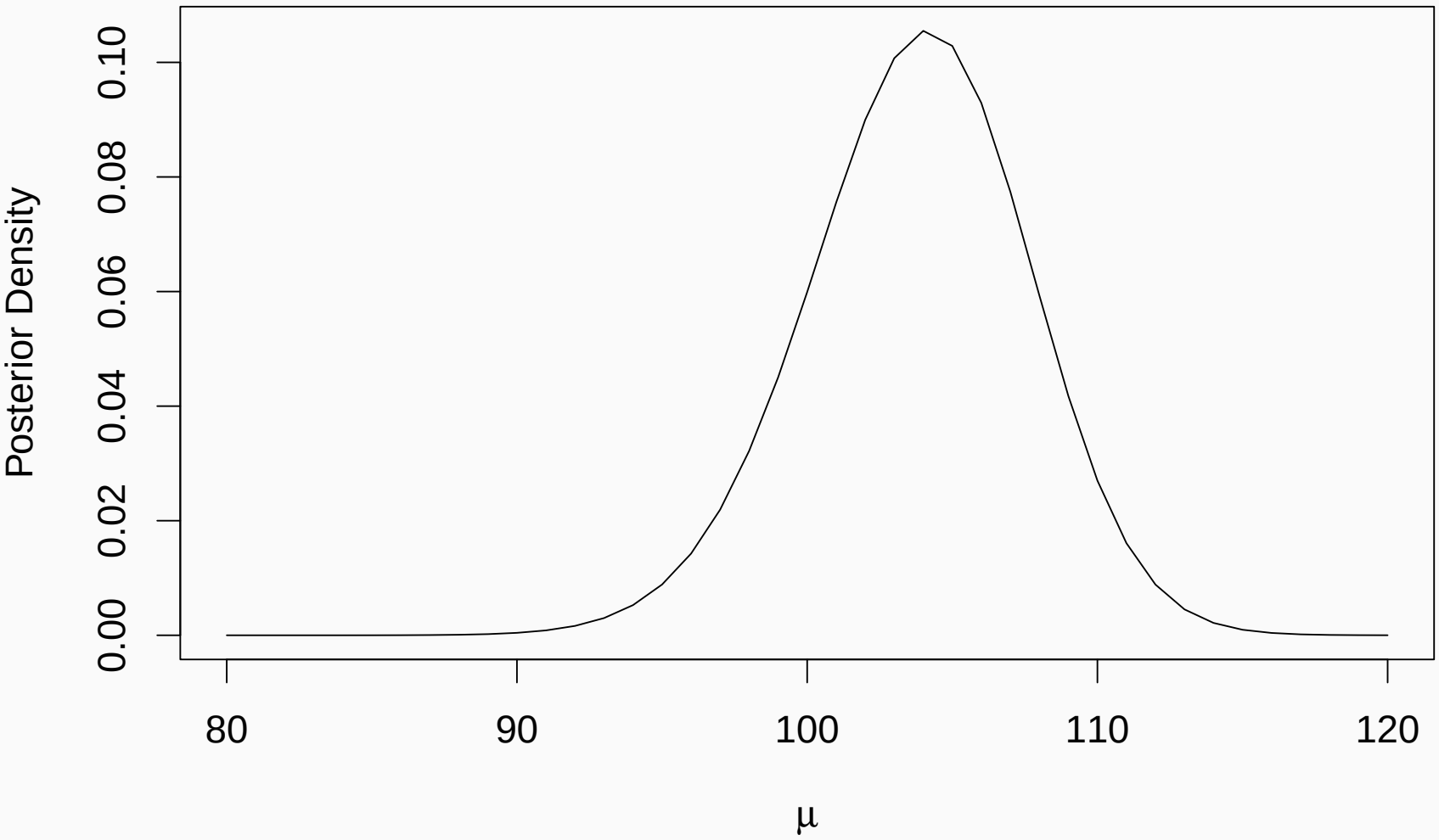
Likelihood Over μ and σ^2 for observations



Joint Posterior over μ and σ^2



Marginal Posterior for μ



Marginal Posterior for σ^2