Bayesian Statistics and Uncertainty Quantification

BEE 6940 Lecture 6

February 29, 2023

Table of Contents

1. Review of the Bootstrap
2. Introduction to Bayesian Statistics
3. Preview of Bayesian Computation
4. Key Takeaways
5. Upcoming Schedule

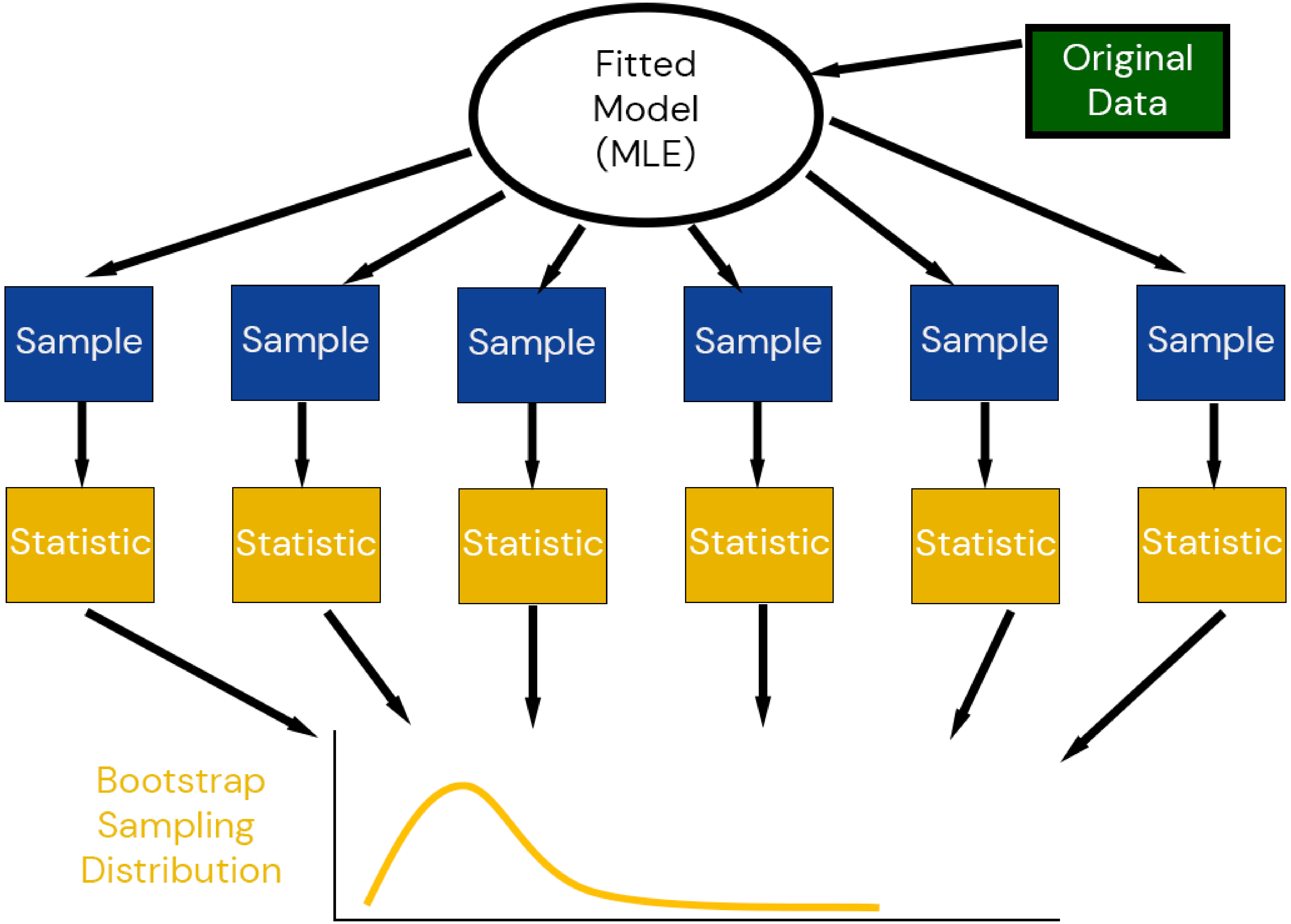
Review of the Bootstrap

Last Class: Frequentist Statistics and Sampling Distributions

Frequentist UQ: Capture the *sampling distribution* of relevant parameters.

This takes into account uncertainty in the data sample and reflects the impact of inference.

Review: The Parametric Bootstrap



Review: The Parametric Bootstrap

Key assumptions:

- Data is sufficiently representative of the population;
- Model structure reasonably captures data-generating process and/or residuals

Sources of Bootstrap Error

The bootstrap has three potential sources of error:

1. **Sampling error**: error from using finitely many replications
2. **Statistical error**: error in the bootstrap sampling distribution approximation
3. **Specification error** (parametric): Error in the data-generating model

But What If We Don't Care About Frequency?

For what might be called the "lab science paradigm," frequency properties are central to make inferences about relevant scientific laws.

But for risk analysis, do we care about them?

But What If We Don't Care About Frequency?

For what might be called the "lab science paradigm," frequency properties are central to make inferences about relevant scientific laws.

But for risk analysis, do we care about them?

Perhaps not! A more relevant perspective: how much should we believe in a given level of a future or present risk? This is the perspective of *Bayesian statistics*.

Introduction to Bayesian Statistics

The Bayesian Perspective

From the Bayesian perspective, probability is interpreted as the degree of belief in an outcome or proposition.

The Bayesian Perspective

From the Bayesian perspective, probability is interpreted as the degree of belief in an outcome or proposition.

There are two different two types of random quantities:

- Observable quantities, or data (also random for frequentists);
- Unobservable quantities, or parameters.

We can also speak of probabilities on model structures, rather than framing model selection as hypothesis-testing.

Conditional Probability Notation

Then it makes sense to discuss the *probability* of

- model parameters θ
- unobserved data $\tilde{\mathbf{y}}$

conditional on the observations $\mathbf{y}$, which we can denote:

$$p(\theta|\mathbf{y}) \text{ or } p(\tilde{\mathbf{y}}|\mathbf{y})$$

Conditioning on Observations

This fundamental conditioning on observations $\mathbf{y}$ is a distinguishing feature of Bayesian inference.

Compare: frequentist approaches are based on re-estimated over the distribution of possible $\mathbf{y}$ conditional on the "true" parameter value.

Bayesian Updating

Bayesian probabilities are conditional on observations.

This means that as make new observations, we can *update* them.

We do this using **Bayes' Rule**.

Bayes' Rule

Original version (Bayes (1763), *An Essay towards solving a Problem in the Doctrine of Chances*):

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \quad \text{if} \quad P(B) \neq 0.$$

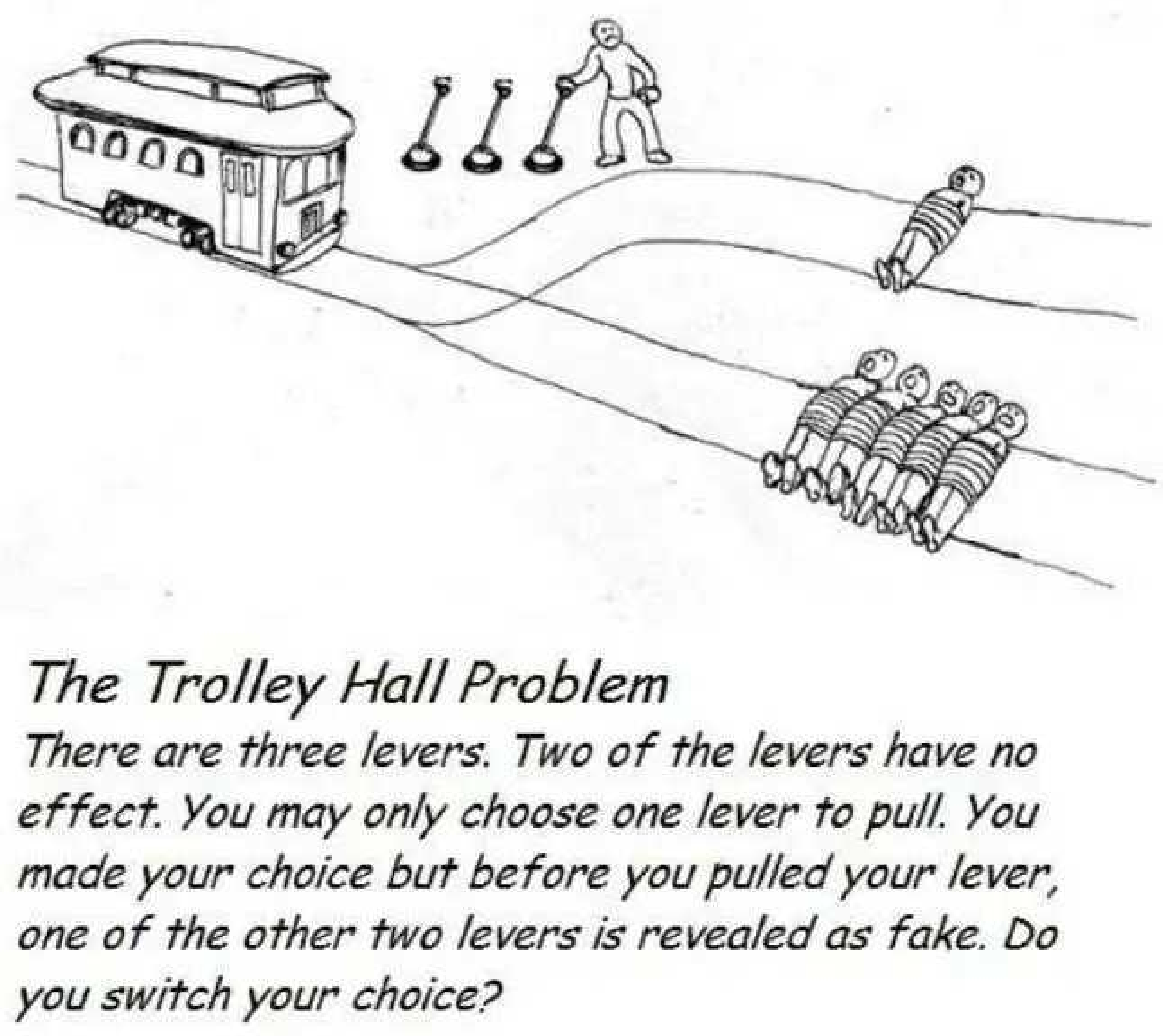
Bayes' Rule

Original version (Bayes (1763), *An Essay towards solving a Problem in the Doctrine of Chances*):

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \quad \text{if} \quad P(B) \neq 0.$$

- Standard but important result in conditional probability
- As previously seen (hopefully remembered!) in Intro Stats
    - Monty Hall Problem

Bayes' Rule



The Trolley Hall Problem

There are three levers. Two of the levers have no effect. You may only choose one lever to pull. You made your choice but before you pulled your lever, one of the other two levers is revealed as fake. Do you switch your choice?

Source: Unsure! Got it from Facebook, also seen on r/PhilosophyMemes

Bayes' Rule

Original version (Bayes (1763), *An Essay towards solving a Problem in the Doctrine of Chances*):

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \quad \text{if} \quad P(B) \neq 0.$$

- Bayes used this to estimate the distribution of a probability p of a binomial outcome (think success/failure).
- Richard Price (actual writer of quite a bit of Bayes (1763); see Stigler (2018)) "rebutted" Hume by "demonstrating" we ought to believe the sun will continue to rise.

Bayes' Rule

"Modern" version (Laplace (1774), *Mémoire sur la probabilité des causes par les événements*):

$$p(\theta|y) = \frac{p(y|\theta)}{p(y)} p(\theta)$$

Bayes' Rule

"Modern" version (Laplace (1774), *Mémoire sur la probabilité des causes par les événements*):

$$\underbrace{p(\theta|y)}_{\text{posterior}} = \frac{\overbrace{p(y|\theta)}^{\text{likelihood}}}{\underbrace{p(y)}_{\text{normalization}}} \overbrace{p(\theta)}^{\text{prior}}$$

On The Normalizing Constant

The normalizing constant (also called the **marginal likelihood**) is the integral

$$p(y) = \int_\Theta p(y|\theta)p(\theta)d\theta.$$

Since this *generally* doesn't depend on θ, it can often be ignored, as the relative probabilities don't change.

On The Normalizing Constant

The normalizing constant (also called the **marginal likelihood**) is the integral

$$p(y) = \int_\Theta p(y|\theta)p(\theta)d\theta.$$

Since this *generally* doesn't depend on θ, it can often be ignored, as the relative probabilities don't change.

One big exception: model selection (will discuss later...)

Bayes' Rule (Ignoring Normalizing Constants)

The version of Bayes' rule which matters the most for 95% (approximate) of Bayesian statistics:

$$p(\theta|y) \propto p(y|\theta) \times p(\theta)$$

> *"The posterior is the prior times the likelihood..."*

Bayesian Model Components

This means that a Bayesian model specification requires two key components:

1. Probability model for the data given the parameters (the *likelihood*), $p(y|\theta)$t
2. Prior distributions over the parameters, $p(\theta)$
   - Can be independent or joint

Bayesian Model Components

This means that a Bayesian model specification requires two key components:

1. Probability model for the data given the parameters (the *likelihood*), $p(y|\theta)$t
2. Prior distributions over the parameters, $p(\theta)$
   - Can be independent or joint

Bayesian updating: Using the likelihood to "update" the prior probabilities of the parameters.

A Coin Flipping Example

We would like to understand if a coin-flipping game is fair.
We've observed the following sequence of flips:

$$\text{H, H, H, T, H, H, H, H, H}$$

A Coin Flipping Example

We would like to understand if a coin-flipping game is fair. We've observed the following sequence of flips:

$$H, H, H, T, H, H, H, H, H$$

8/9 are heads, which might seem suspicious, but randomness can result in outliers like this.

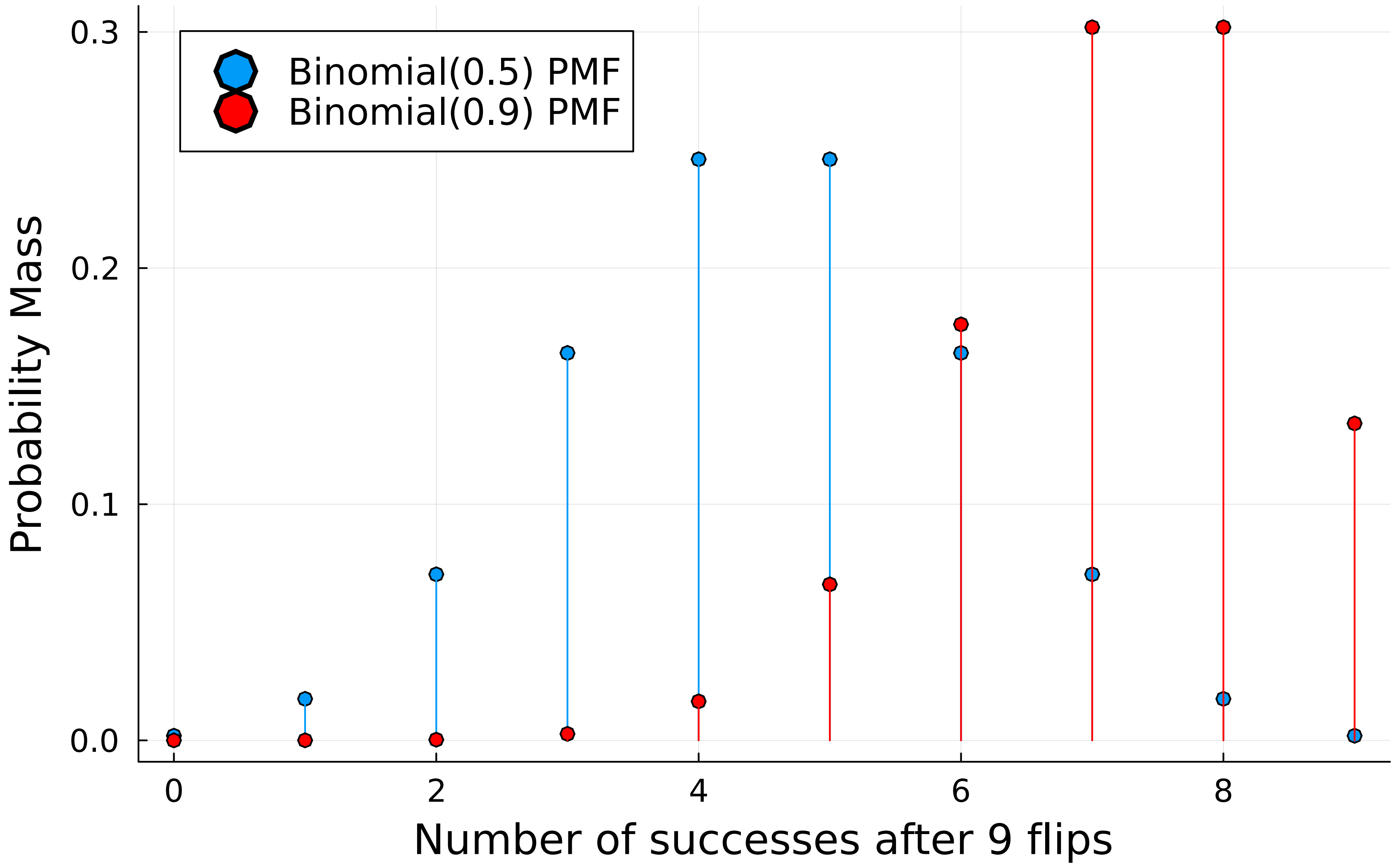
Coin Flipping Likelihood

The data-generating process here is straightforward: we can represent a coin flip with a heads-probability of θ as a sample from a Binomial distribution,

$$y \sim \mathrm{Binomial}(\theta).$$

Coin Flipping Probability Mass

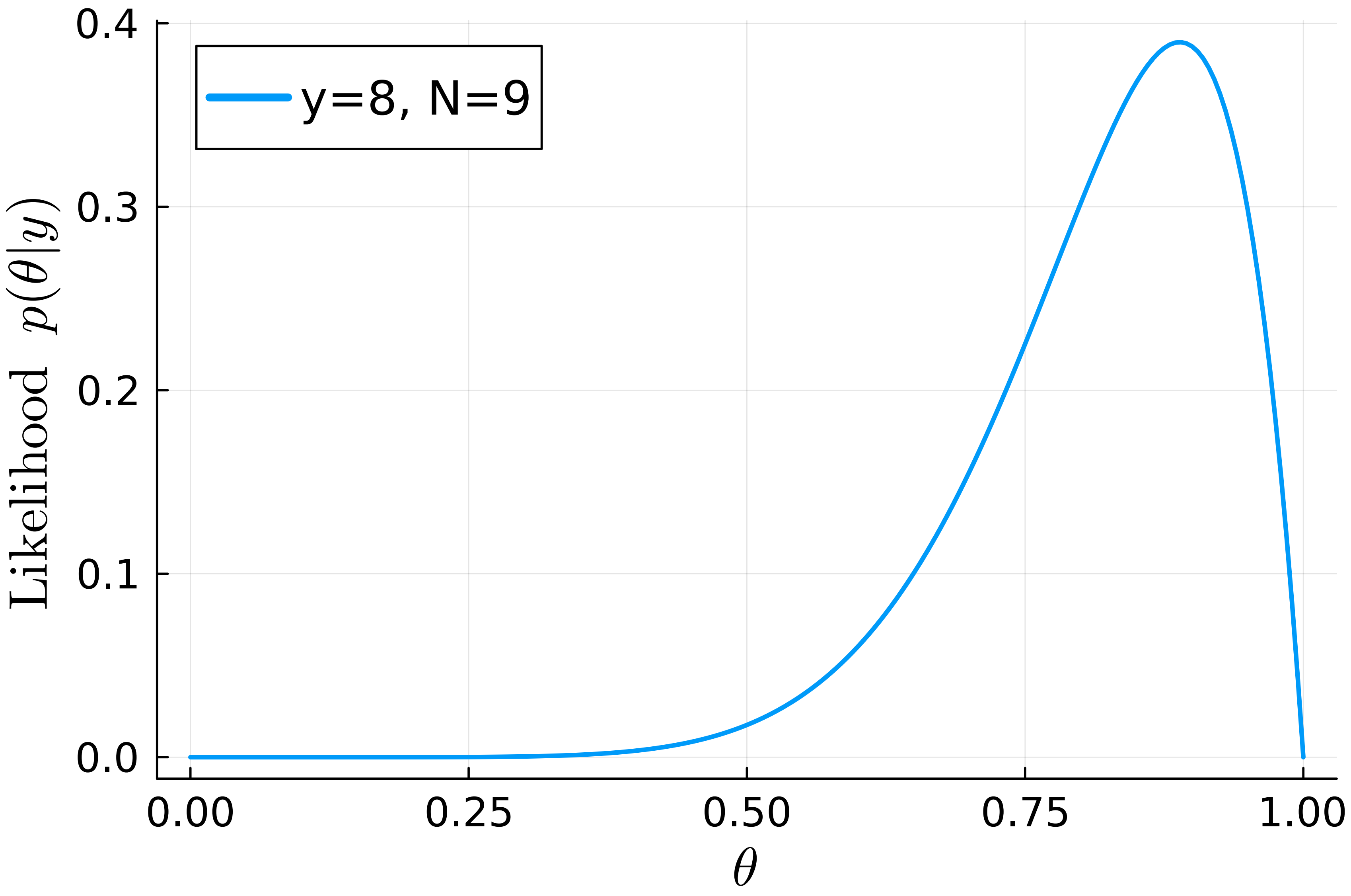
Let's compare what the **probability mass functions** of these distributions look like for $\theta = 0.5$ and $\theta = 0.9$.



Likelihood Function

The PMF told us what the probability of a given dataset was given a fixed parameter θ.

But we can view this same function from a different perspective: given the number of successes, how *likely* is a given parameter?



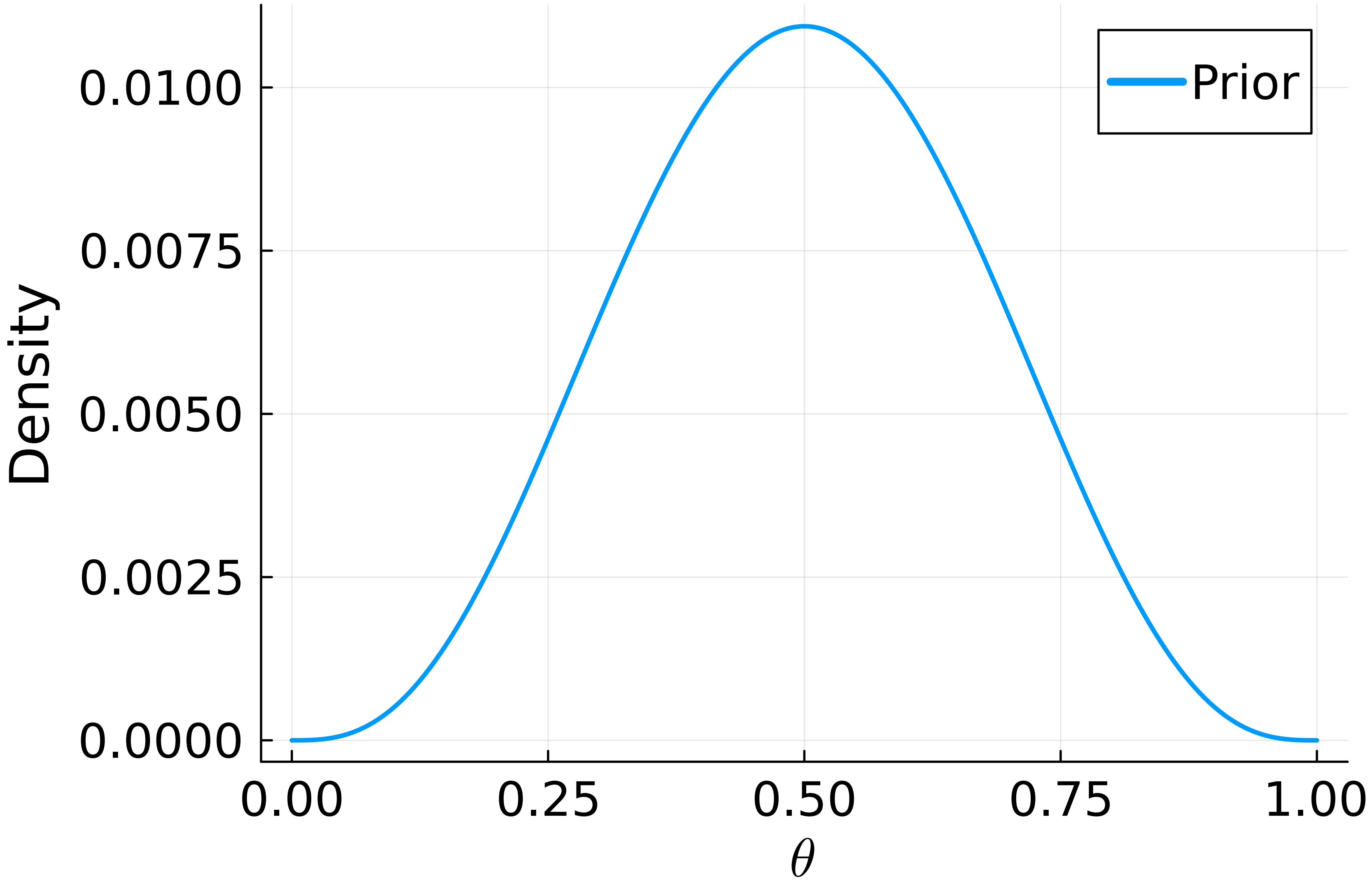
Prior Distribution

For frequentist approaches, we could stop there and maximize the likelihood, and we'd get a maximum likelihood estimate of $\theta \approx 0.88$.

But suppose that we spoke to a friend who knows something about coins, and she tells us that it is extremely difficult to make a passable weighted coin which comes up heads more than 75% of the time. Since we have a relatively small amount of data, this seems like valuable information to include, and we can do this through our prior.

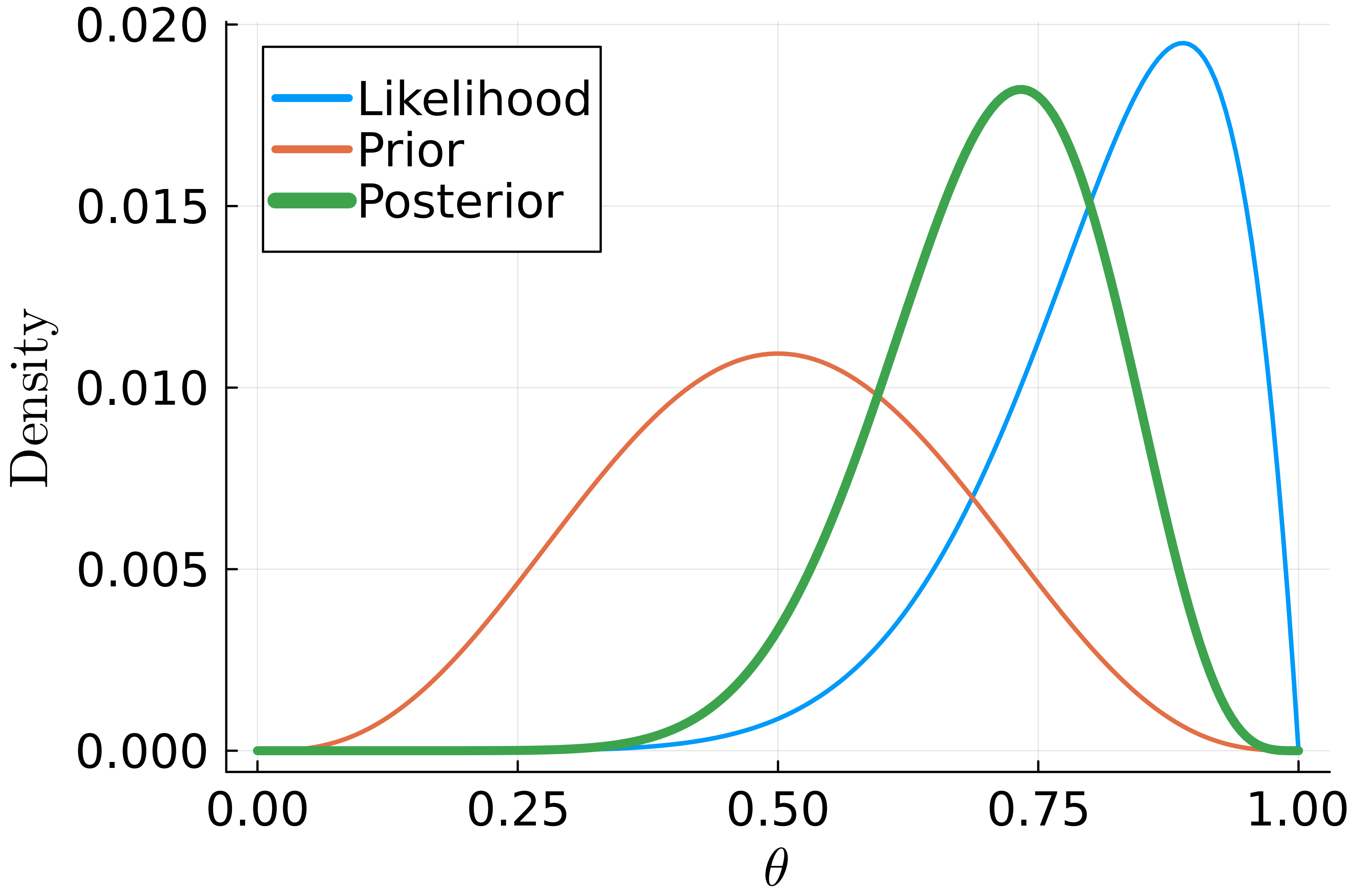
Prior Distribution

Since θ is bounded between 0 and 1, we'll settle on a Beta distribution for our prior, specifically $\mathrm{Beta}(4, 4)$, which covers a reasonable spread of possibilities while maintaining symmetry.



Posterior Distribution

Combining using Bayes' rule gives us a **maximum *a posteriori* (MAP)** estimate of $\theta \approx 0.74$.



Bayesian Updating As An Information Filter

- The posterior is a "compromise" between the prior and the data.
- The posterior mean is a weighted combination of the data and the prior mean
- The weights depend on the prior and the likelihood variances
- More data makes the posterior more confident (lower variance)

Representing Uncertainty

As with frequentist approaches, can reflect posterior inferences through a point estimate (mean, median, or some other *Bayes estimator*).

But more often, we want to capture the degree of uncertainty associated with a particular value.

Credible Intervals

Bayesian **credible intervals** are straightforward to interpret: θ is in I with probability α.

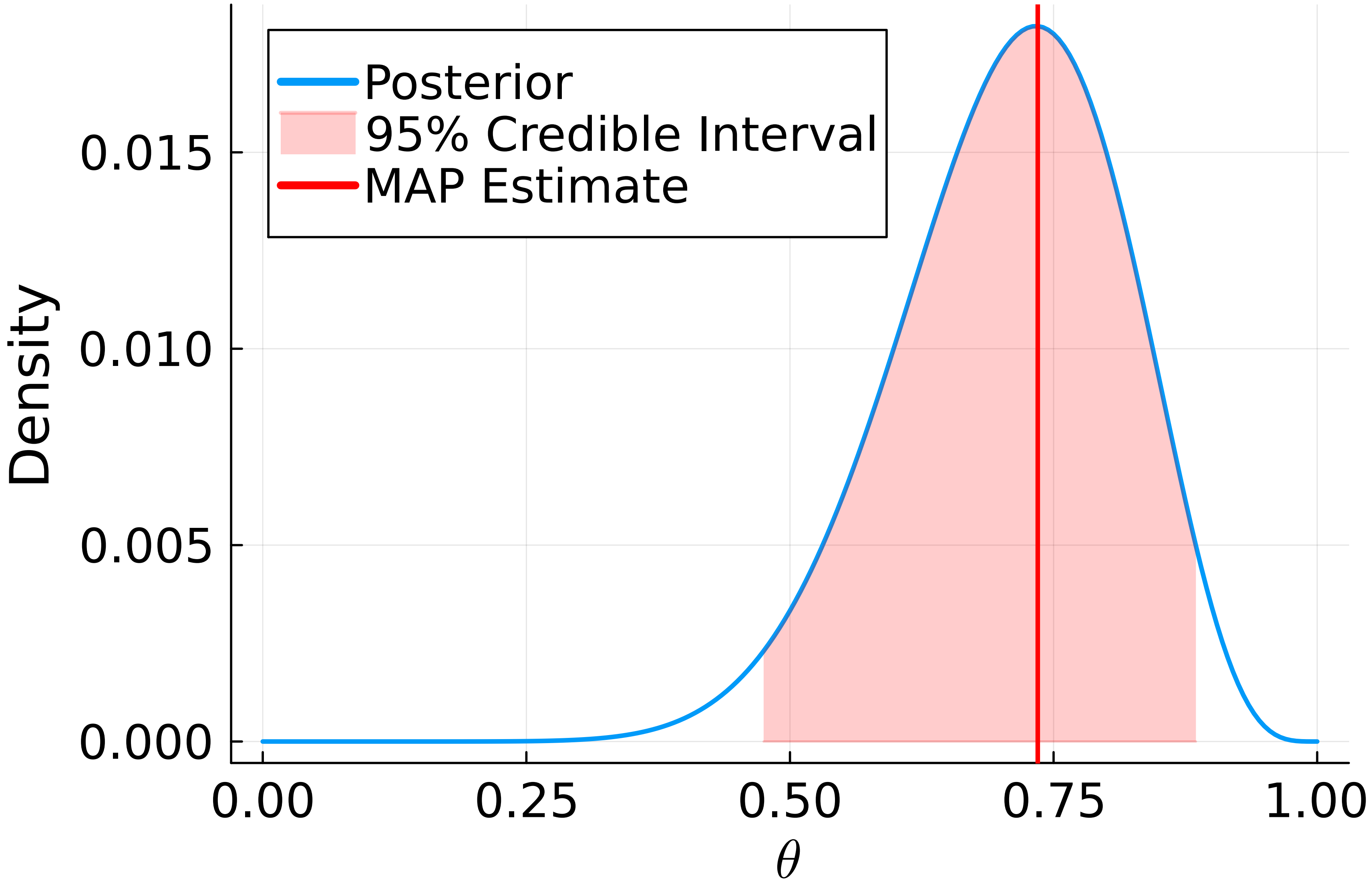
In other words, choose I such that

$$p(\theta \in I | \mathbf{y}) = \alpha.$$

This is not usually a unique choice, but the "equal-tailed interval" between the $(1-\alpha)/2$ and $(1+\alpha)/2$ quantiles is a common choice.

Credible Intervals: Coin Flipping Example

In the case of our coin flipping example, how much uncertainty is there? Let's say we want to capture the 95% credible interval, which is $(0.48, 0.89)$.



Credible Intervals: Coin Flipping Example

But that was for a simple example where it was easy to compute the posterior for a large number of values.

The easiest way to do this in general is through Monte Carlo: draw a lot of samples from the posterior and compute the empirical quantiles. We'll discuss later today/next week.

More Complex Models

One advantage of the Bayesian framework is it can be extended to more complex problems:

- Models with heteroskedastic residual structures:

$$\begin{aligned} y(t) &= f(\theta, t) + R(\phi, t) \\ R(\phi, t) &= \zeta(\phi, t) + \varepsilon t \end{aligned}$$

More Complex Models

One advantage of the Bayesian framework is it can be extended to more complex problems:

- Hierarchical models:

$$
\begin{aligned}
y_j | \theta_j, \phi &\sim P(y_j | \theta_j, \phi) \\
\theta_j | \phi &\sim P(\theta_j | \phi) \\
\phi &\sim P(\phi)
\end{aligned}
$$

Generative Modeling

Bayesian models can also be used to generate new data $\tilde{y}$ through the *posterior predictive distribution*:

$$p(\tilde{y}|\mathbf{y}) = \int_{\Theta} p(\tilde{y}|\theta)p(\theta|\mathbf{y})d\theta$$

This allows us to test the model through simulation (*e.g.* hindcasting) and generate probabilistic predictions.

Role of the Prior

There are two views on how to select prior distributions:

1. Priors as capturing by constraints/prior knowledge;

Role of the Prior

There are two views on how to select prior distributions:

1. Priors as capturing by constraints/prior knowledge;
2. Priors as part of the data-generating process.

On Uniform Priors

What about uniform priors?

- Unbounded uniform priors: often chosen to reflect ignorance, but nonsensical for data-generation;
- Bounded uniform priors: sudden transition from positive to zero probability is rarely justified.

Considerations When Selecting Priors

- **Informativeness**: How much information does the prior encode?
- **Structure**: Does the prior encode modeling features (*e.g.* symmetry)?
- **Regularization**: Does the prior yield more "stable" inferences (*e.g.* penalizing extreme parameter values)?

Also: what values are assigned zero prior probability? These values are ruled out from the posterior.

Bayesian Computation: A Preview

Goals of Bayesian Computation

1. Sampling from the *posterior* distribution

$$p(\theta|\mathbf{y})$$

2. Sampling from the *posterior predictive* distribution

$$p(\tilde{y}|\mathbf{y})$$

by generating data.

Bayesian Computation and Monte Carlo

In other words, Bayesian computation involves Monte Carlo simulation from the posterior (predictive) distribution.

These samples can then be analyzed to identify estimators, credible intervals, etc.

Sampling from the Posterior

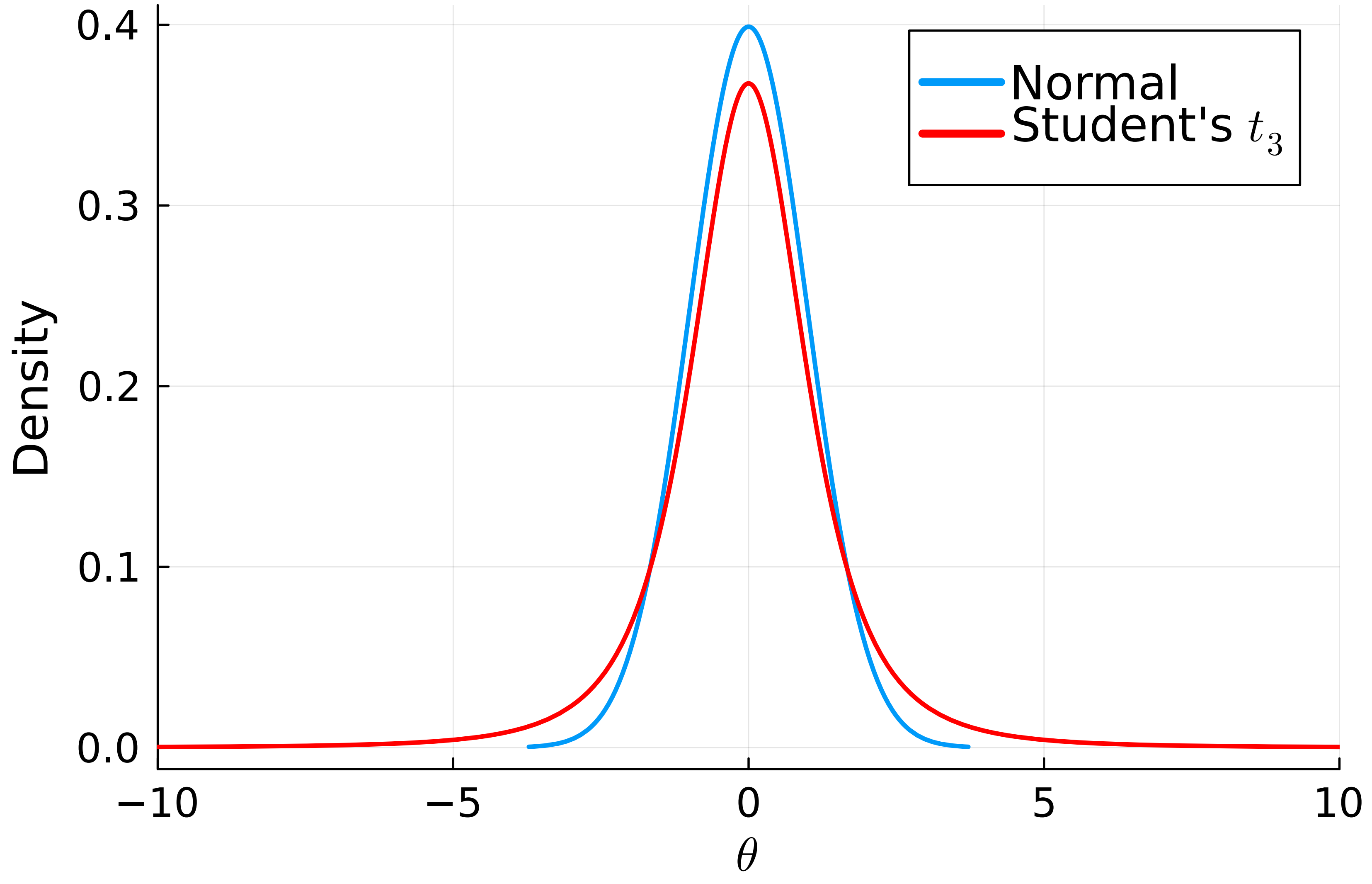
Trivial for extremely simple problems: low-dimensional and with "conjugate" priors (which make the posterior a closed-form distribution).

What to do when problems are more complex and/or we don't want to rely on priors for computational convenience?

A First Algorithm: Rejection Sampling

Idea:

1. Generate proposed samples from another distribution $g(\theta)$ which covers the target $p(\theta|\mathbf{y})$;
2. Accept those proposals based on the ratio of the two distributions.



Rejection Sampling Algorithm

Suppose $p(\theta|\mathbf{y}) \leq Mg(\theta)$ for some $1 < M < \infty$.

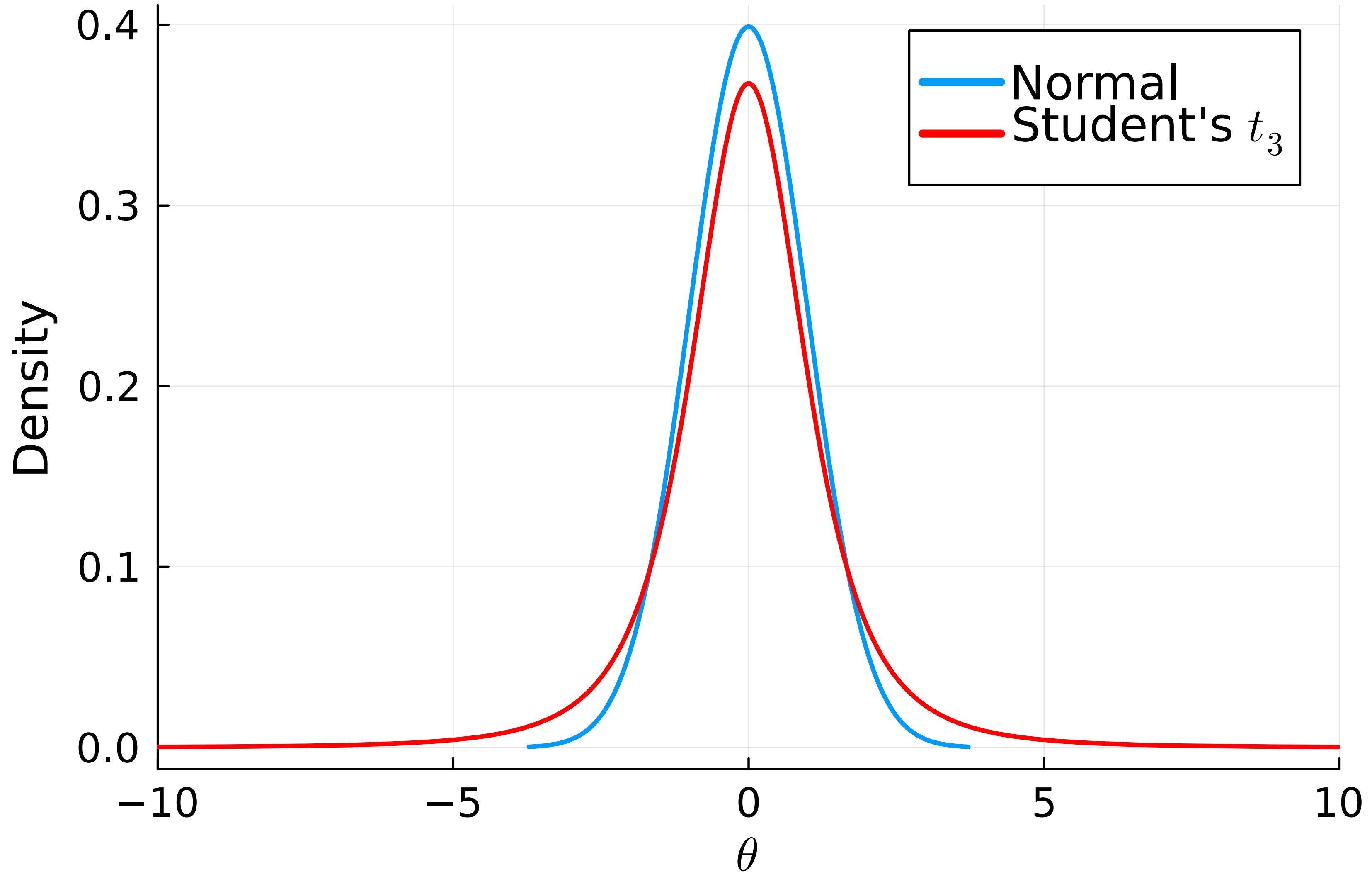
1. Simulate $u \sim \mathrm{Unif}(0, 1)$.
2. Simulate a proposal $\hat{\theta} \sim g(\theta)$.
3. If

$$u < \frac{p(\hat{\theta}|\mathbf{y})}{Mg(\hat{\theta})},$$

accept $\hat{\theta}$. Otherwise reject.

Rejection Sampling Intuition

We want to keep more samples from the areas where $g(\theta) < p(\theta|\mathbf{y})$ and reject where g is heavier (in this case, the tails).



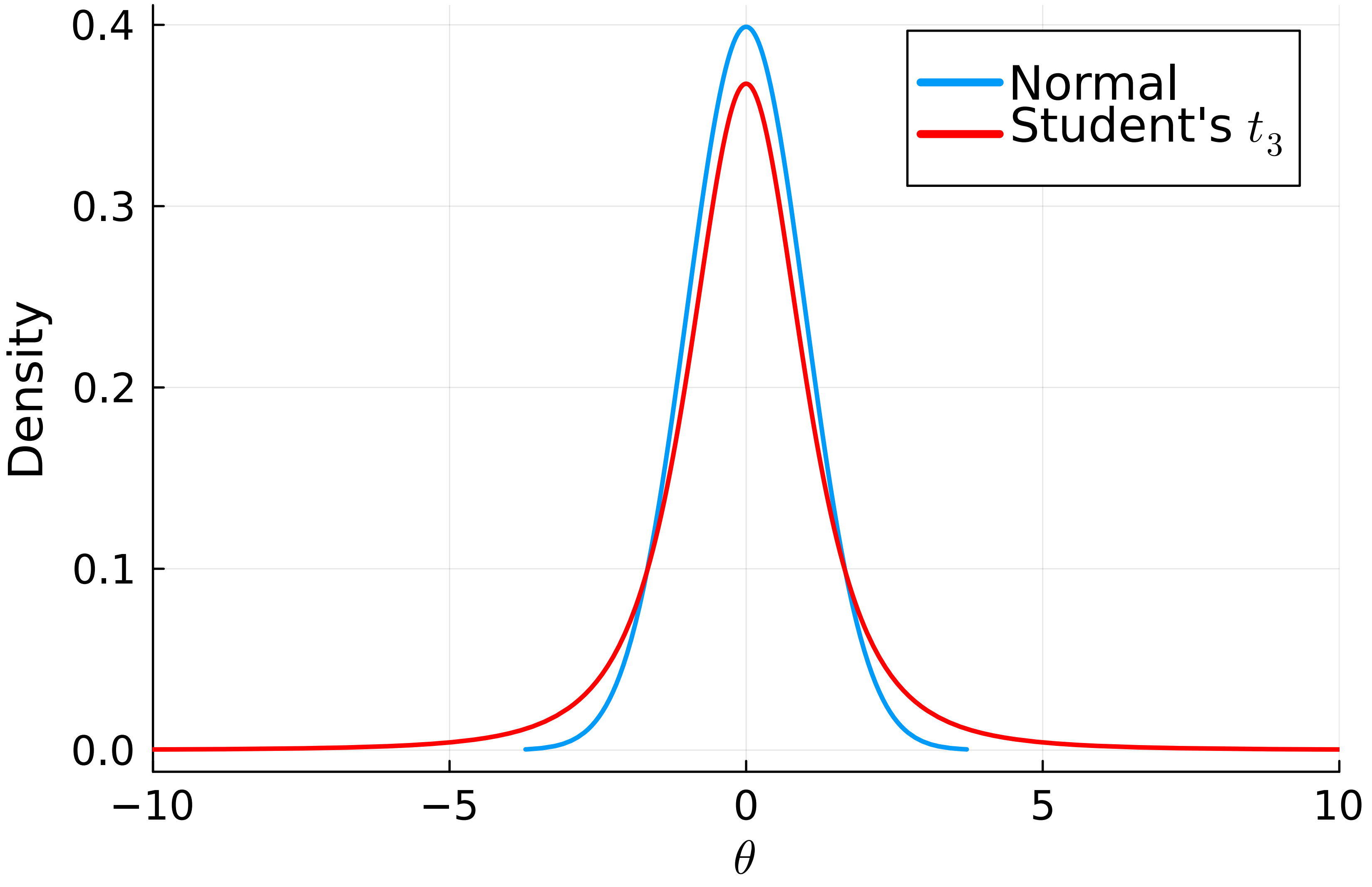
Rejection Sampling Intuition

In the bulk: $Mg(\theta)$ is closer to $p(\theta|\mathbf{y})$, so the acceptance ratio is higher.

In the tails:

$$Mg(\theta) \gg p(\theta|\mathbf{y}),$$

so the acceptance ratio is much lower.



Rejection Sampling Considerations

1. Probability of accepting a sample is $1/M$, so the "tighter" the proposal distribution coverage the more efficient the sampler.
2. Need to be able to compute M.

Finding a good proposal and computing M may not be easy for complex posteriors!

How can we do better?

Key Takeaways

Key Takeaways: Bayesian Statistics

- Probability as degree of belief
- Emphasis on explicit conditioning on the data
- Bayes' Rule as the fundamental theorem of conditional probability
- Bayesian updating as an information filter
- Prior selection important: lots to consider!
- Rejection sampling as a first Monte Carlo algorithm for sampling from "arbitrary" distributions.

Upcoming Schedule

Upcoming Schedule

Next Monday: Markov chains and Markov chain Monte Carlo