BAYESIAN STATISTICS AND UNCERTAINTY QUANTIFICATION

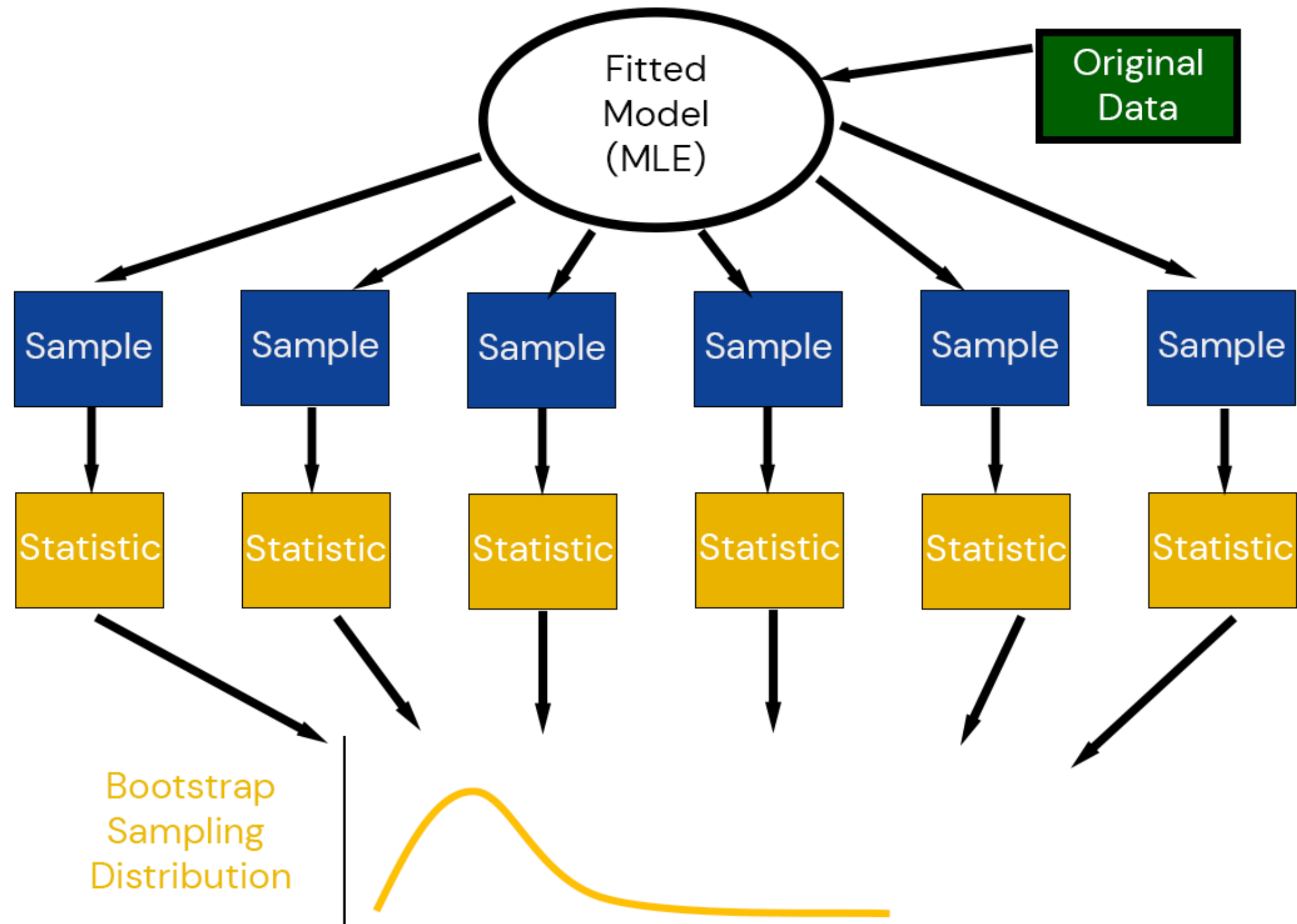BEE 6940 LECTURE 6

FEBRUARY 29, 2023

TABLE OF CONTENTS

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*):

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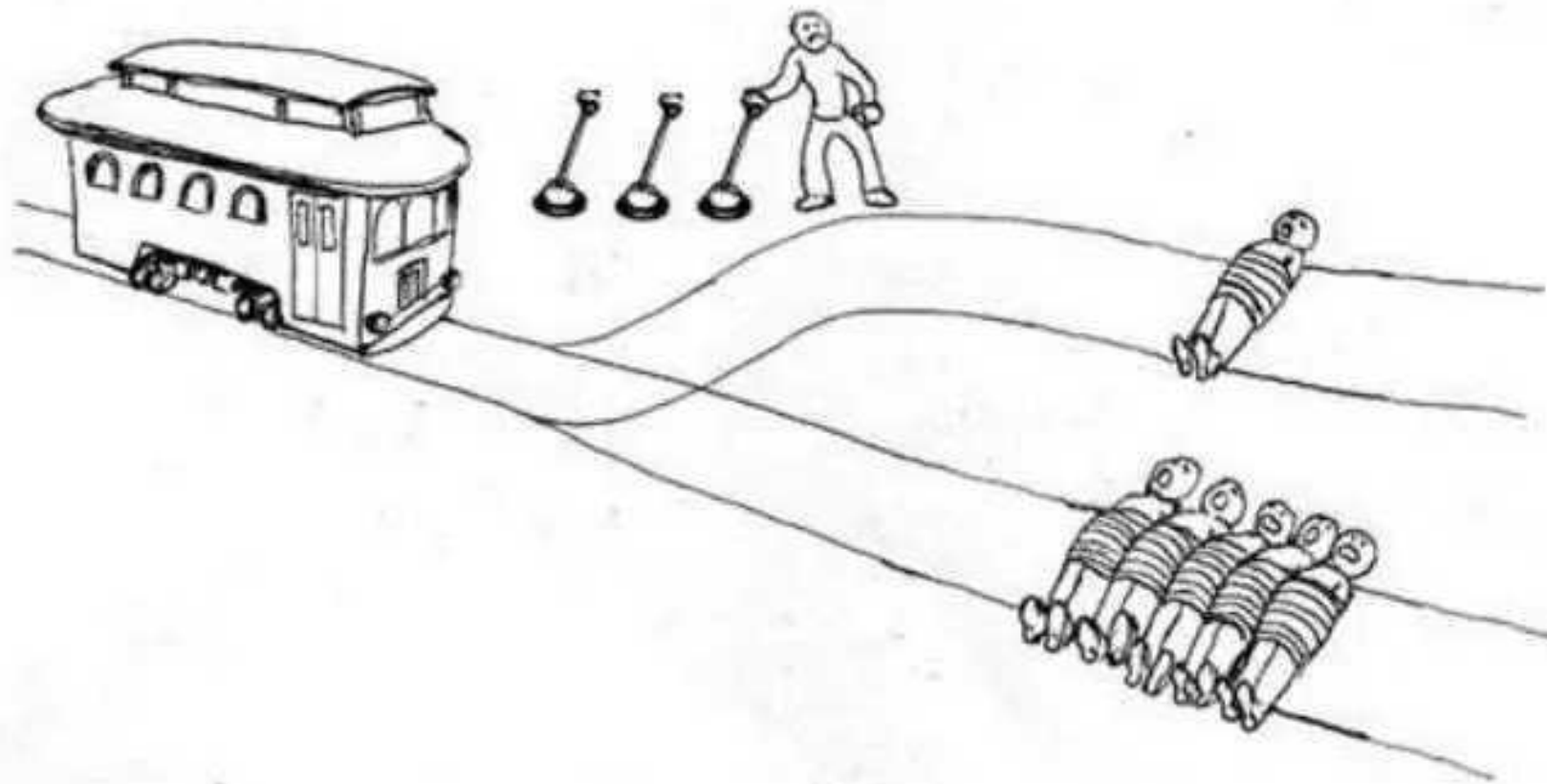
Bayes' Rule

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

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

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?

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*):

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

- $p(\theta|y)$ are *posterior* probabilities
- $p(\theta)$ are *prior* probabilities;
- $p(y|\theta)$ is the *likelihood*;
- $p(y) = \int p(y|\theta)p(\theta)d\theta$ is a normalizing constant.

BAYES' RULE

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..."*

BAYES' RULE

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..."*

The normalizing constant (called the **marginal likelihood**) matters for some things (model selection), but mostly vanishes as a constant.

Bayesian Model Components

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

1. Prior distributions over the parameters, $p(\theta)$

    ○ Can be independent or joint

2. Probability model for the data given the parameters (the *likelihood*), $p(y|\theta)$

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

SIMPLE EXAMPLE OF BAYESIAN UPDATING

Problem: Estimate the mean of normally-distributed data with known variance

$$y_i \sim \text{Normal}(\theta, 2).$$

This model defines the likelihood:

$$p(\mathbf{y}|\theta) = \prod_i \frac{1}{\sqrt{4\pi}} e^{-(y_i - \theta)^2/4}$$
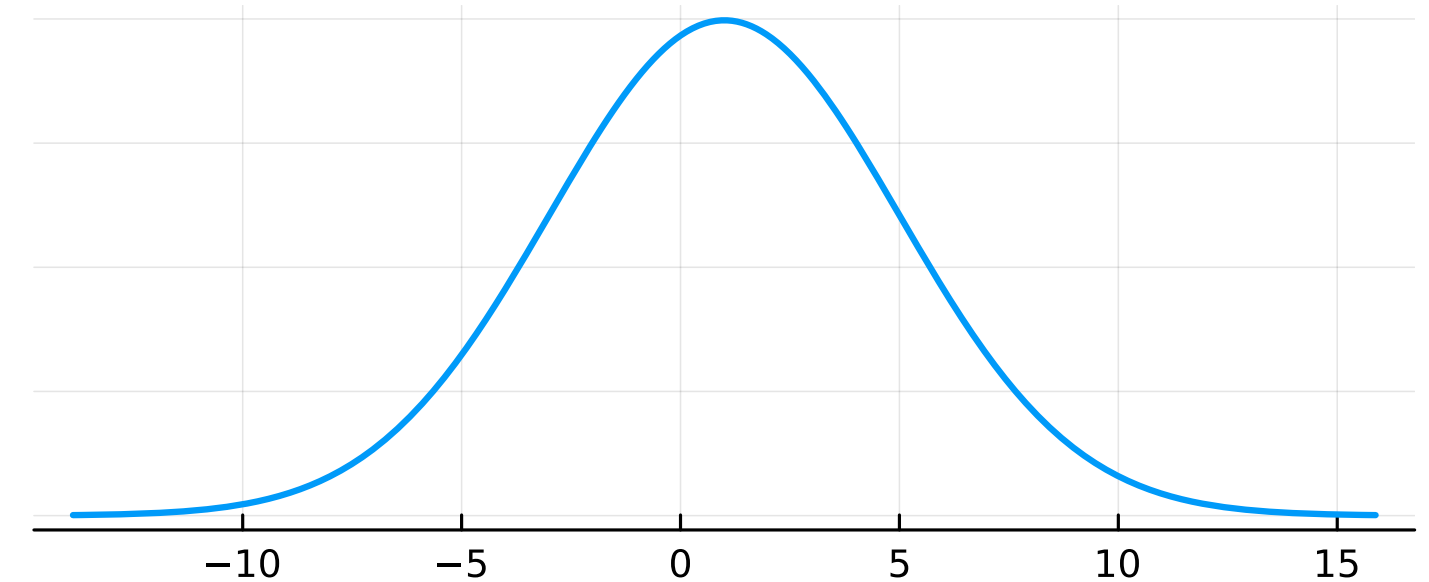
SIMPLE EXAMPLE OF BAYESIAN UPDATING

We also need a prior. Suppose that we suspect the mean is near 0, and is more likely to be positive, but we can't rule out any particular values.

What might we choose?

Simple Example of Bayesian Updating

We also need a prior. Suppose that we suspect the mean is near 0, and is more likely to be positive, but we can't rule out any particular values.

What might we choose?

One example:

$$\theta \sim \text{Normal}(1, 4)$$

SIMPLE EXAMPLE OF BAYESIAN UPDATING

Then the prior density is:

$$p(\theta) = \frac{1}{\sqrt{8\pi}} e^{(\theta-1)^2/8}$$

SIMPLE EXAMPLE OF BAYESIAN UPDATING

Then the prior density is:

$$p(\theta) = \frac{1}{\sqrt{8\pi}} e^{(\theta-1)^2/8}$$

Using Bayes' Rule (ignoring constants):

$$p(\theta|\mathbf{y}) \propto e^{-(\theta-1)^2/8} \prod_{i=1}^{n} e^{-(y_i-\theta)^2/4}$$

SIMPLE EXAMPLE OF BAYESIAN UPDATING

After doing some algebra and completing the square:

$$\theta | \mathbf{y} \sim \text{Normal} \left(\frac{1 + 2 \sum_{i=1}^{n} y_i}{2n + 1}, \frac{4}{2n + 1} \right)$$

Simple Example of Bayesian Updating

After doing some algebra and completing the square:

$$\theta | \mathbf{y} \sim \text{Normal} \left(\frac{1 + 2 \sum_{i=1}^{n} y_i}{2n + 1}, \frac{4}{2n + 1} \right)$$

- Bayesian updating acts as an *information filter*

  - 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)

Simple Example of Bayesian Updating
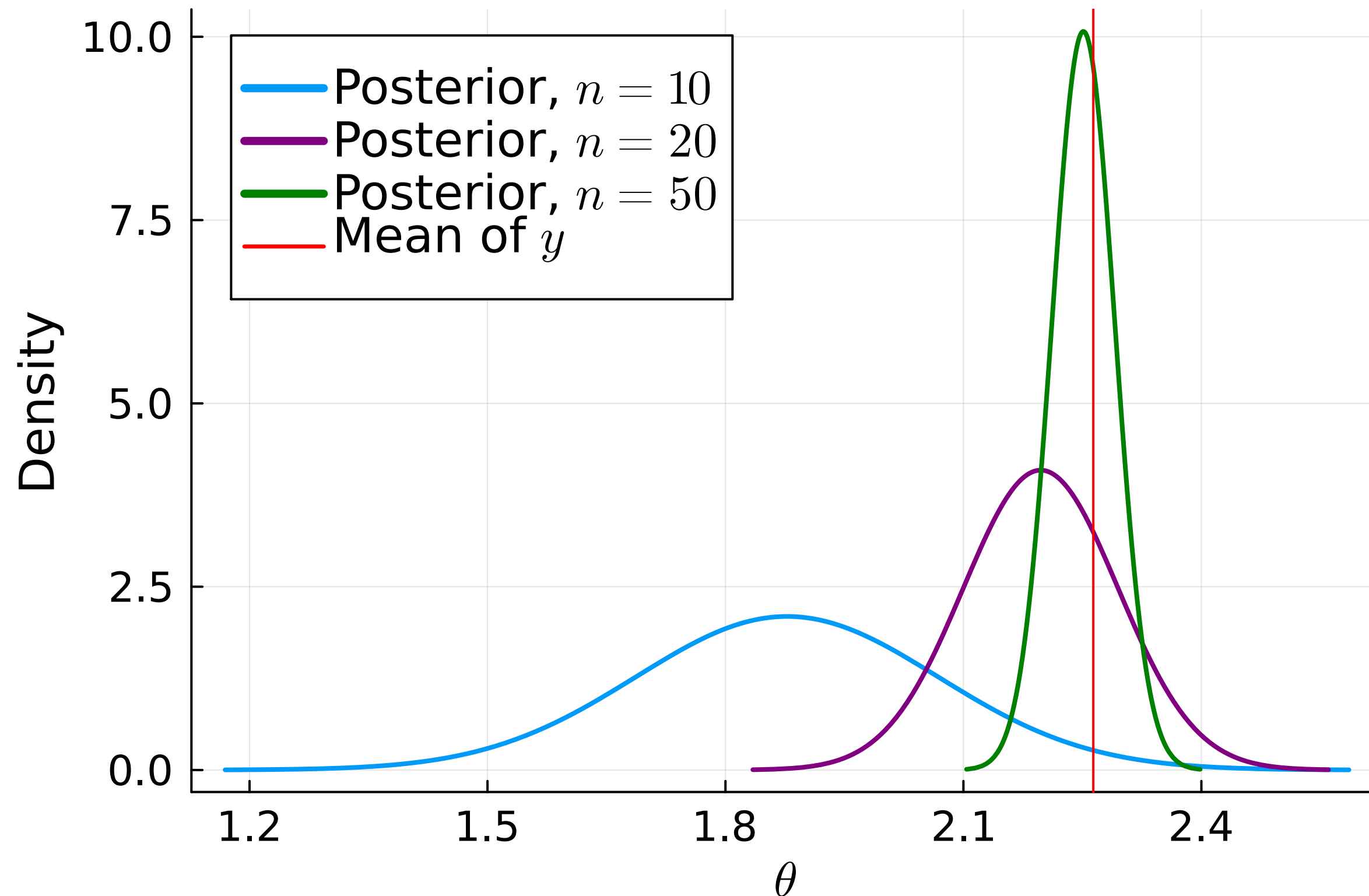
REPRESENTING UNCERTAINTY

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.

MORE COMPLEX MODELS

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

- Models with heteroskedastic residual structures:

$$y(t) = f(\theta, t) + R(\phi, t)$$
$$R(\phi, t) = \zeta(\phi, t) + \varepsilon t$$

MORE COMPLEX MODELS

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

- Hierarchical models:

$$y_j | \theta_j, \phi \sim P(y_j | \theta_j, \phi)$$
$$\theta_j | \phi \sim P(\theta_j | \phi)$$
$$\phi \sim P(\phi)$$

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.
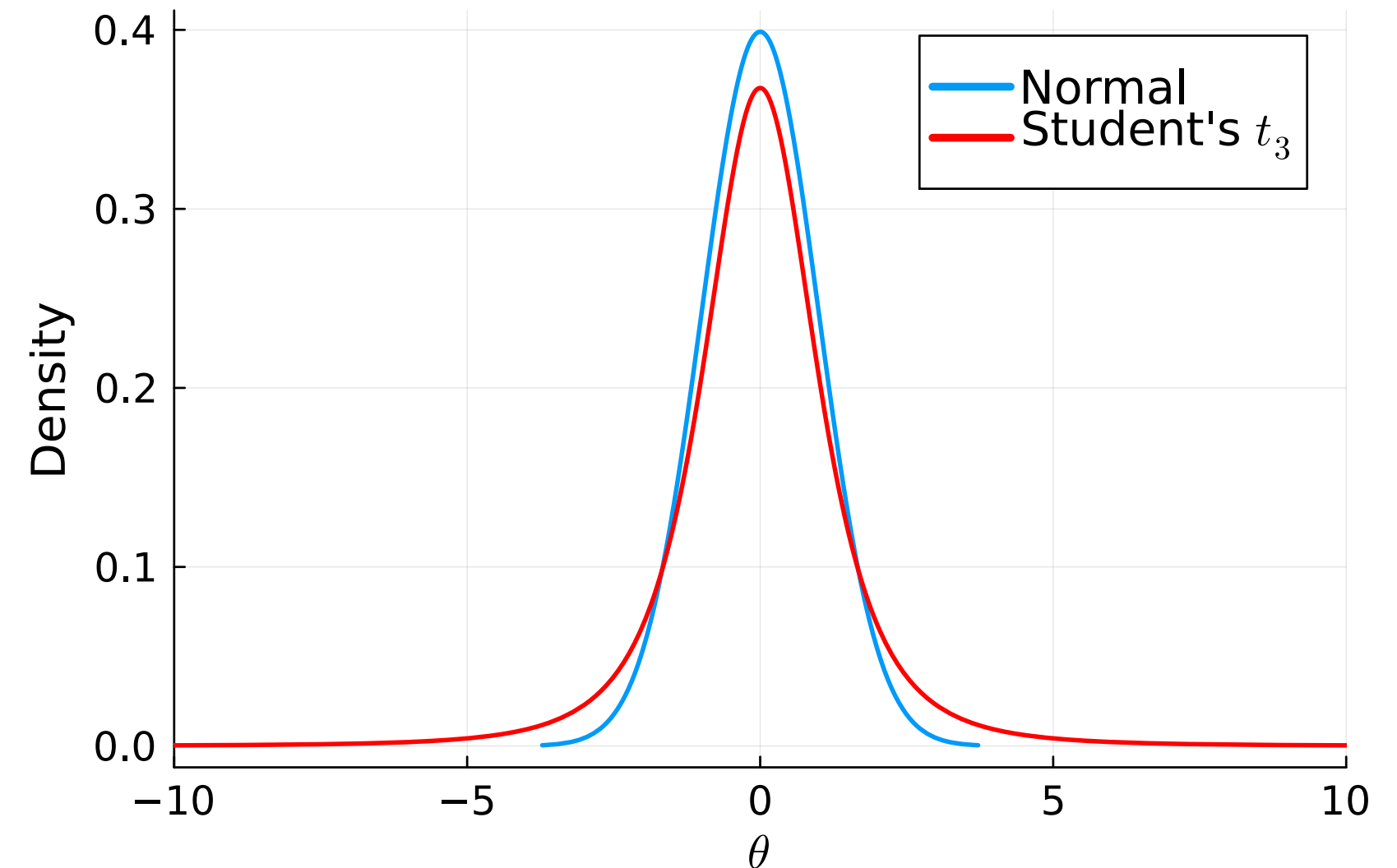
REJECTION SAMPLING ALGORITHM

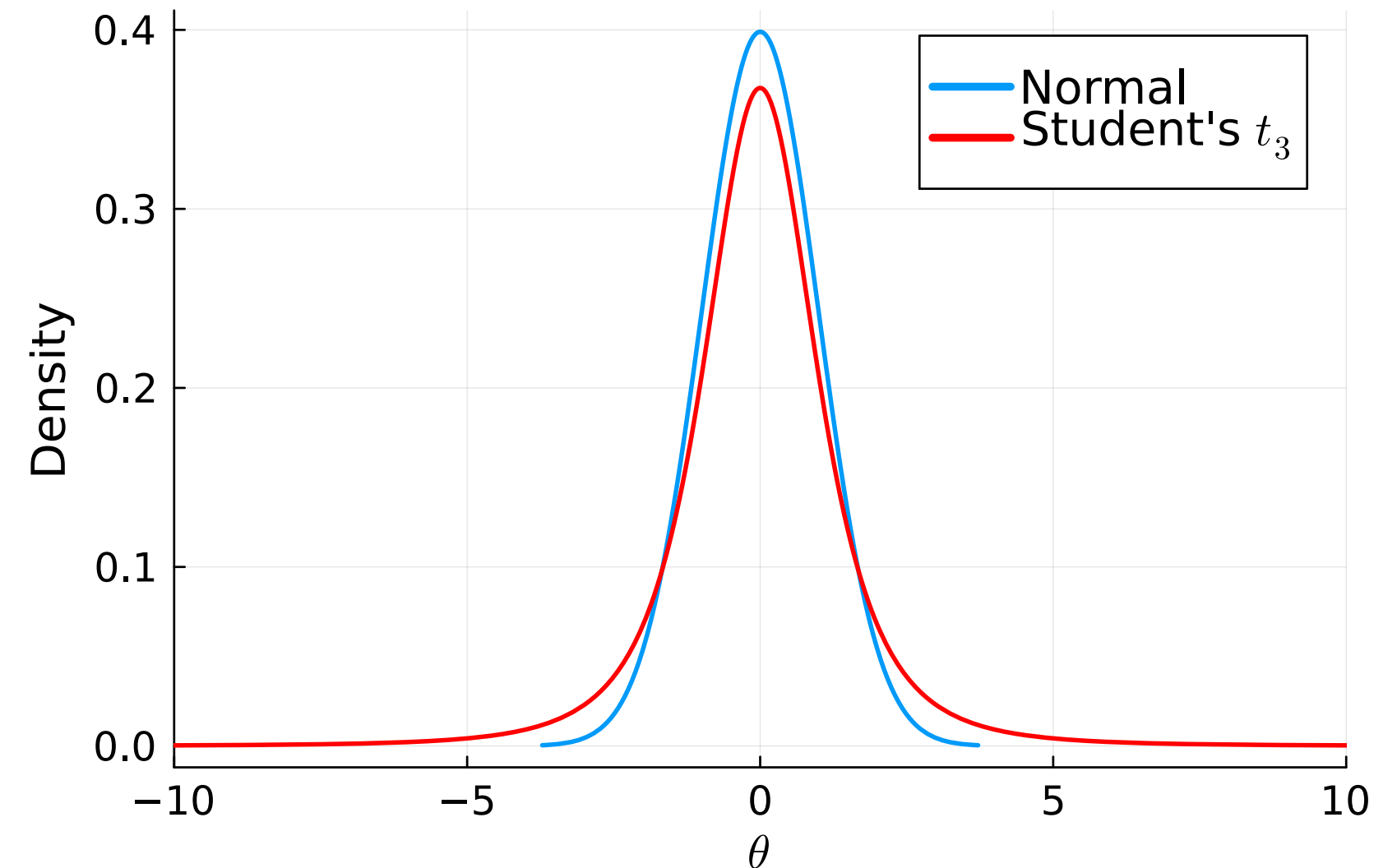Suppose $p(\theta|\mathbf{y}) \leq Mg(\theta)$ for some $1 < M < \infty$.

1. Simulate $u \sim \mathbf{Unif}(0,1)$.

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).
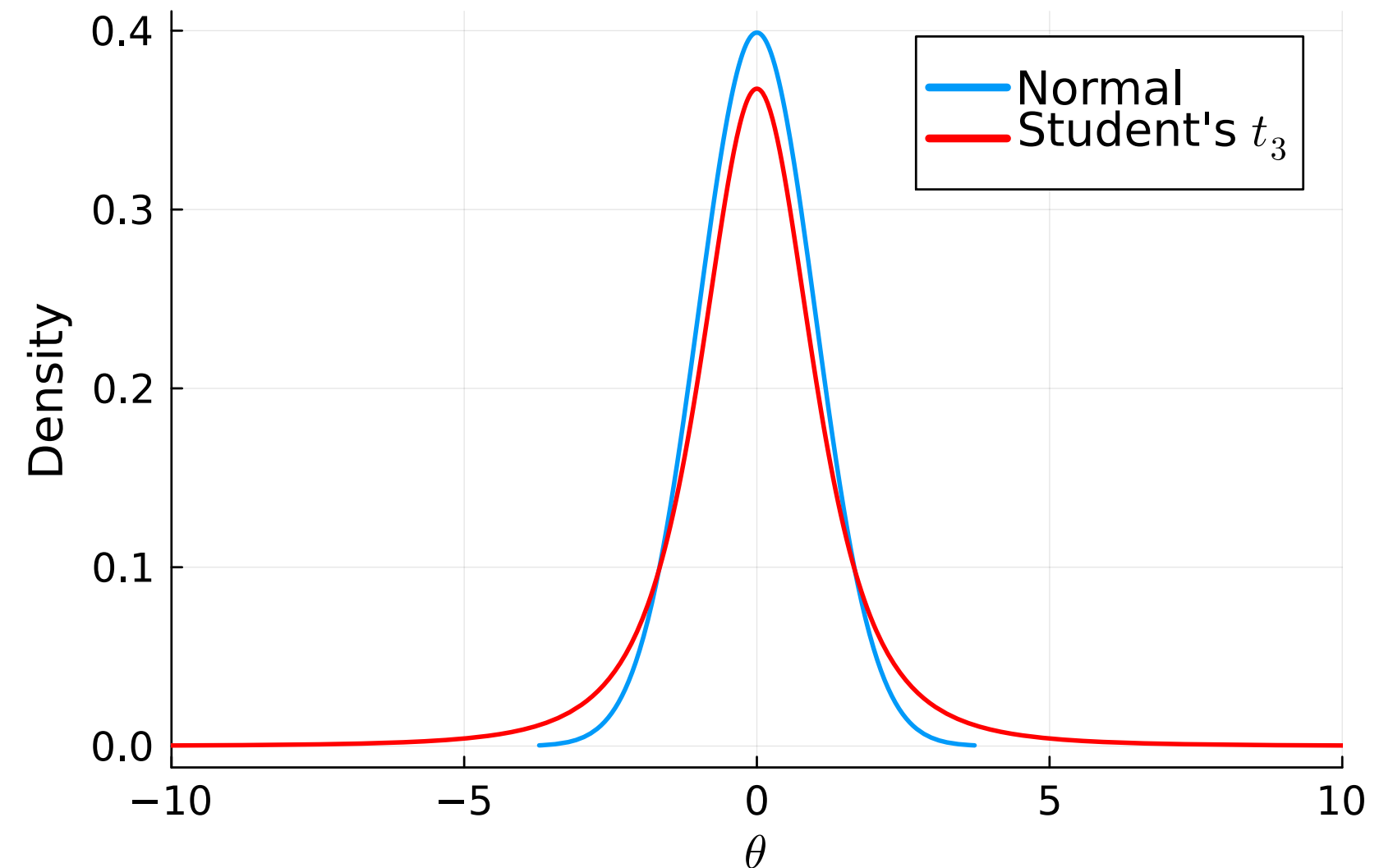
REJECTION SAMPLING INTUITION

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