BIOS 6342: Contemporary Statistical Inference

Andrew J. Spieker, Ph.D.

Associate Professor of Biostatistics
Vanderbilt University

Set 5: Bayesian statistics

Version: 03/16/2026

Table of Contents

The Bayesian framework:

- The purpose of this set of slides is to orient you toward the mathematical and computational considerations of Bayesian statistics.
- As a reminder, I am not taking a stance on the "Bayesian vs. frequentist" debate in this course, nor will I ask you to take a stance.
  - In my view, the debate is overblown in some ways.
  - Bayesian methods have pragmatic advantages in certain circumstances that are orthogonal to any philosophical considerations.
  - Frequentist methods are, in some (but not all) circumstances, procedurally easier.
  - More commentary on the debate scattered throughout these slides.
- To orient you to these ideas, I will focus mostly on one-parameter models. Many of the ideas generalize to higher dimensions.

Bayesian thinking

The Bayesian framework:

- Frequentist framework: a probability is a number between zero and one that characterizes long-term frequency under replications.
- Bayesian framework: a probability can serve as a statement regarding extent of knowledge/belief.
  - 0 and 1: Certainty ("no" and "yes," respectively); densities appropriate for characterizing continuous quantities.
  - Indeed, can be based on prior data (but need not be).
- Therefore, all unknown quantities contained in a probability model can be treated as random variables in a Bayesian framework, **including the values of fixed, unknown parameters**.
  - The axioms necessary to justify this are mild and defensible.
  - The term "random variable" is very often a misnomer and can make Bayesian thinking confusing.
- What does the statement "there is a 70% chance of rain tomorrow" mean under the frequentist and under the Bayesian paradigm?

The Bayesian framework: The prior and posterior distribution

- Let θ denote the parameter of interest.
- Before the current data are observed, knowledge/belief/uncertainty regarding θ is expressed via a *prior*, $\pi(\theta)$.
- After we observe data, $\mathbf{x}$, inference can be made on the basis of the *posterior*, $\pi(\theta|\mathbf{X} = \mathbf{x})$, which is expressed using Bayes' theorem:

$$\pi(\theta|\mathbf{X} = \mathbf{x}) = \frac{p_{\mathbf{X}}(\mathbf{x};\theta)\pi(\theta)}{p_{\mathbf{X}}(\mathbf{x})} = \frac{p_{\mathbf{X}}(\mathbf{x};\theta)\pi(\theta)}{\int_{\Theta} p_{\mathbf{X}}(\mathbf{x};\theta)d\Pi_{\Theta}(\theta)}.$$

- $p_{\mathbf{X}}(\mathbf{x};\theta)$ is the *likelihood* (when considered as a function of θ).

The Bayesian framework: Notes

- A common misconception is that, to a Bayesian, a prior on θ must reflect actual randomness in θ (e.g., different treatment effects across different hospital sites).
- A Bayesian is completely permitted to think of θ as a fixed, unknown quantity in the population; $\pi(\theta)$ as the representation of his or her extent of belief about that fixed, unknown quantity before observing study data; and $\pi(\theta|\mathbf{X} = \mathbf{x})$ as the representation of his or her extent of belief about that fixed, unknown quantity after observing study data.

The Bayesian framework: The posterior distribution

- Ignoring the normalizing constant in the denominator:

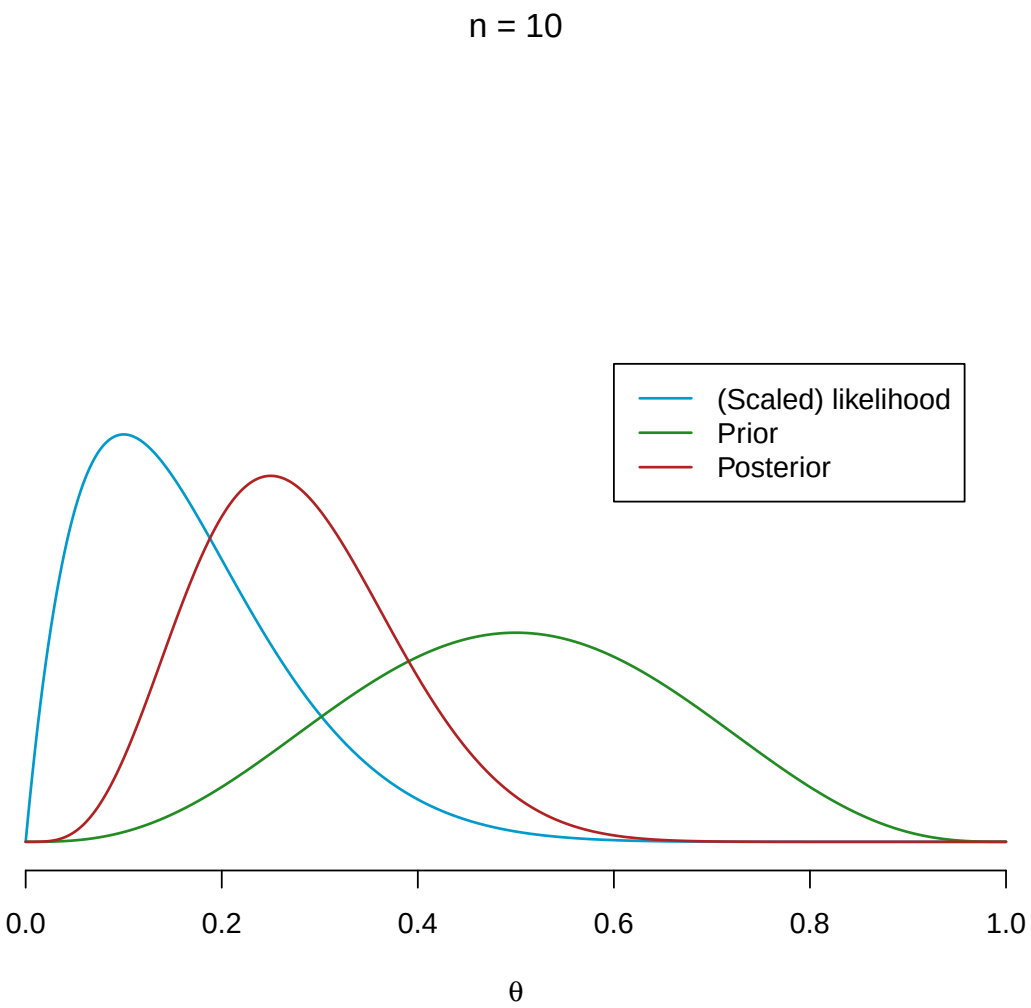
$$\pi(\theta|\mathbf{X} = \mathbf{x}) \propto_\theta p_{\mathbf{X}}(\mathbf{x}; \theta)\pi(\theta).$$

- "Posterior" $\propto$ "Likelihood" $\times$ "Prior."
- Conceptually appealing: probabilistically combines information on the parameters arising from the data and from prior beliefs or knowledge.
  - Simple scalar multiplication with no integration nor need to incorporate intentions (contrast with messy sampling distributions that must incorporate intention to analyze data).
- In the long run (asymptotics in n), likelihood should dominate prior.
  - Caveat: $\pi(\theta) = 0 \Rightarrow \pi(\theta|\mathbf{X} = \mathbf{x}) = 0$. We need to be careful.

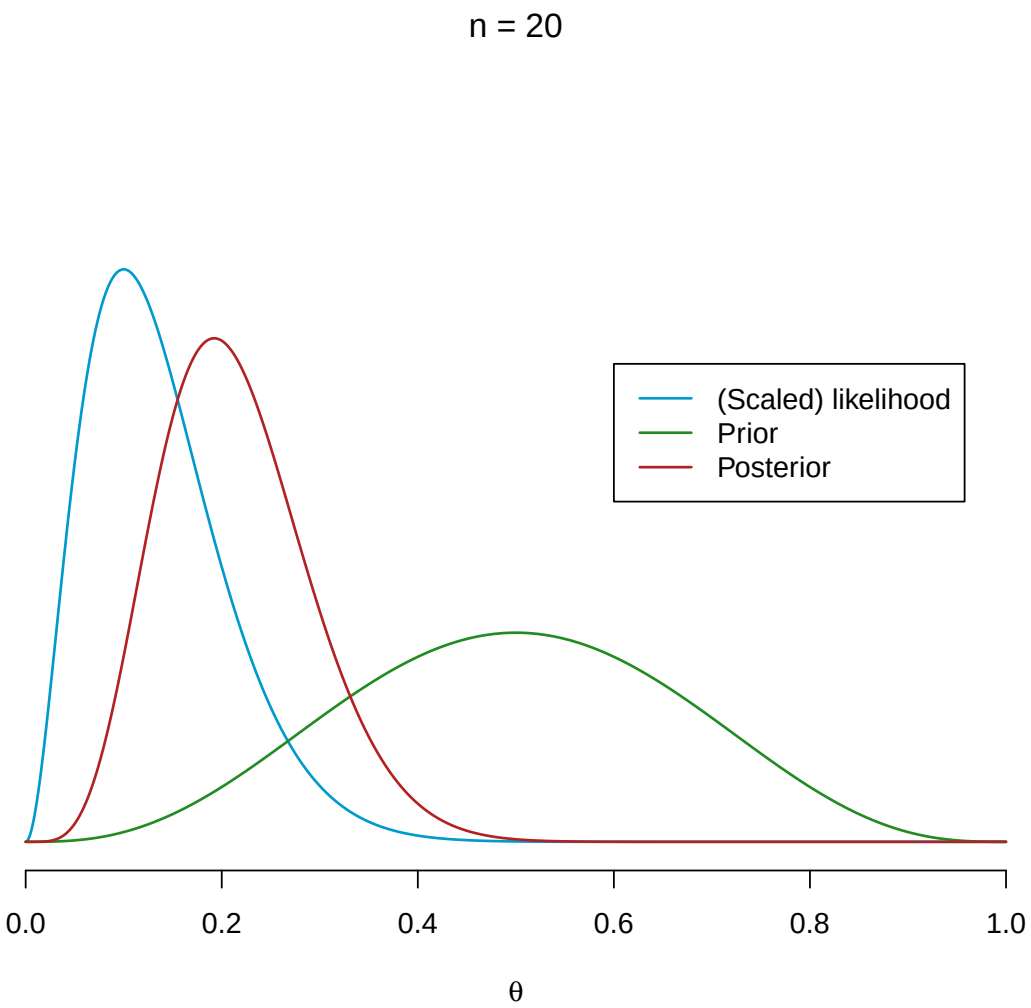
Likelihood overwhelms the prior: In large samples

- Suppose $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\theta)$.
- I am interested in Bayesian estimation of θ. Consider a prior on θ with mean, median, mode at 0.5; and with moderate variability.
- Suppose the MLE of θ is given by $\widehat{\theta}_n = 0.10$. Can you draw a reasonable representation of the likelihood, prior, and posterior from this information alone?
  - Answer: No, because knowing the value of n is critically important information needed to get the picture right.
- On the following three slides, I show the same prior, but update the sample size. Don't worry (yet) about the calculations—just focus on the pictures.
- Anticipating behavior heuristically: the total likelihood is a product of n individual likelihoods (i.i.d. observations); the prior only gets one chance to contribute regardless of sample size.

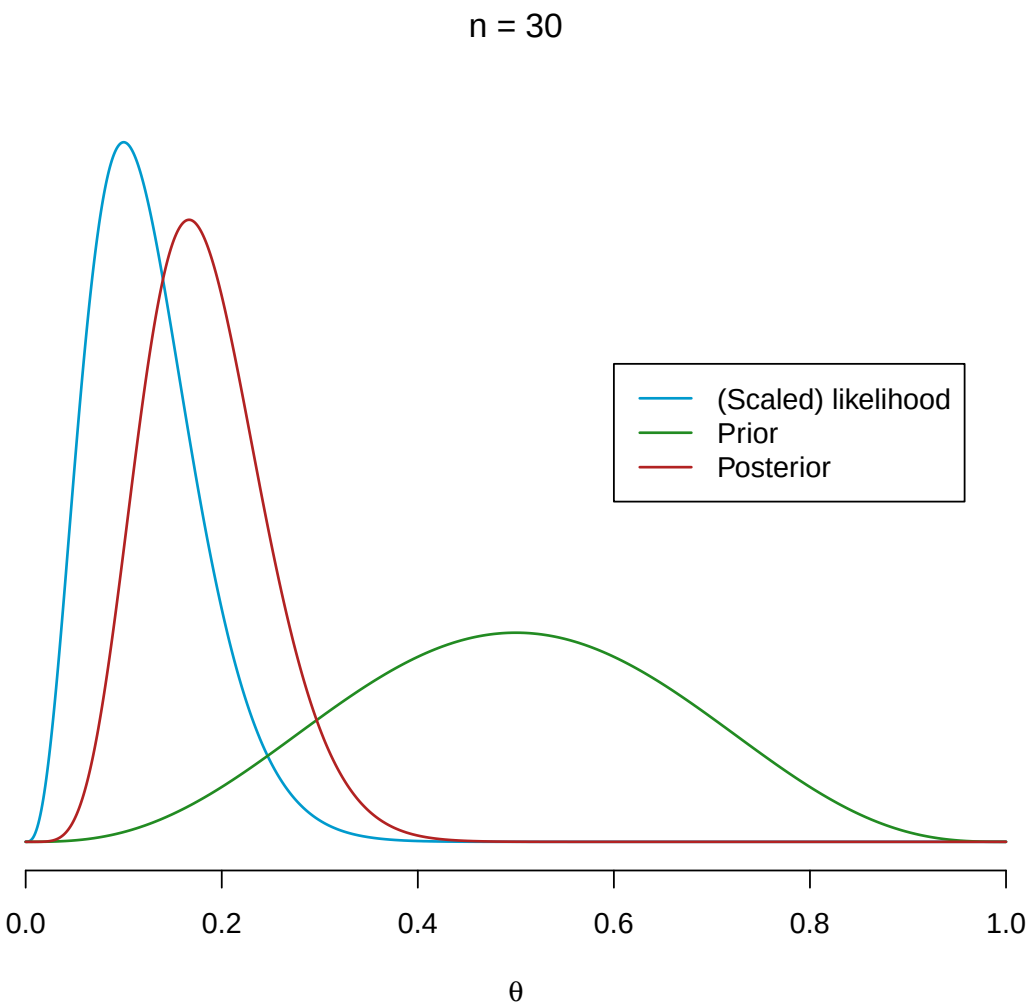
Likelihood overwhelms the prior: In large samples



Likelihood overwhelms the prior: In large samples



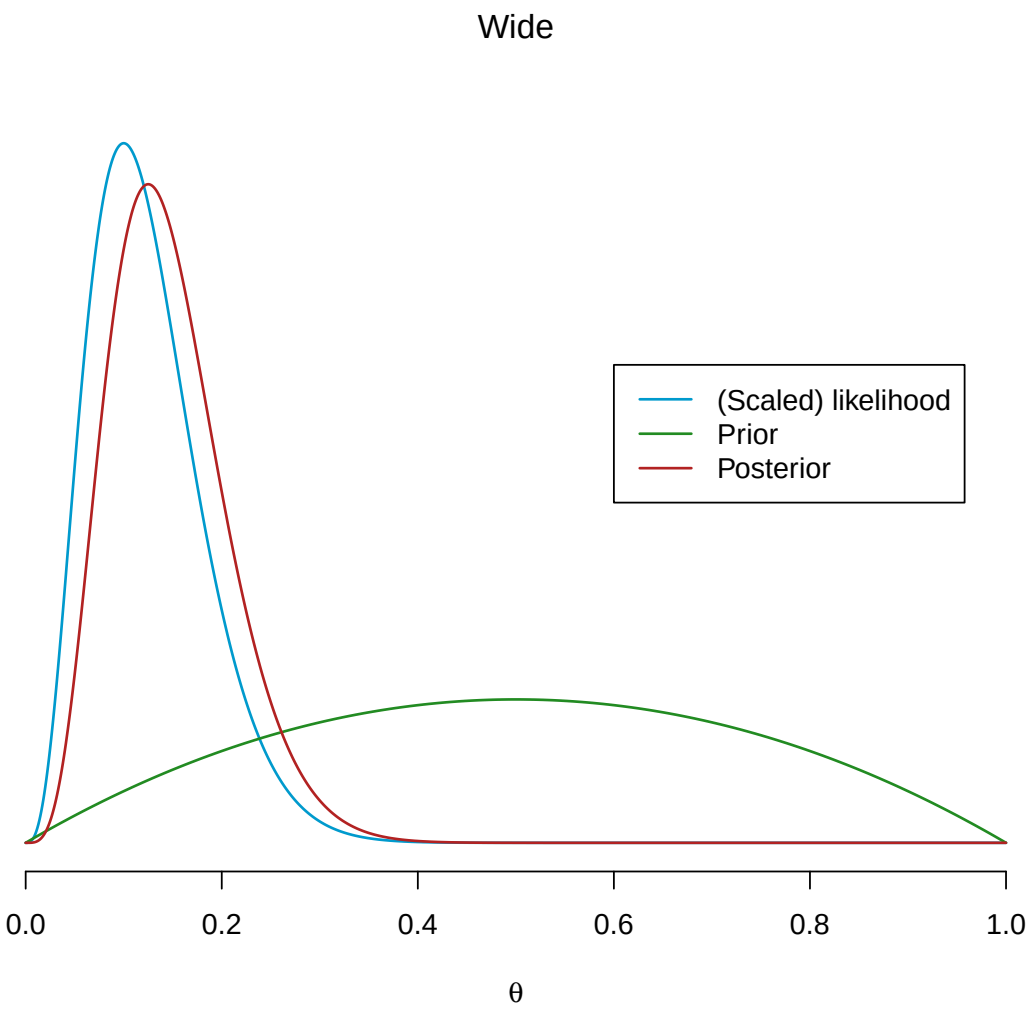
Likelihood overwhelms the prior: In large samples



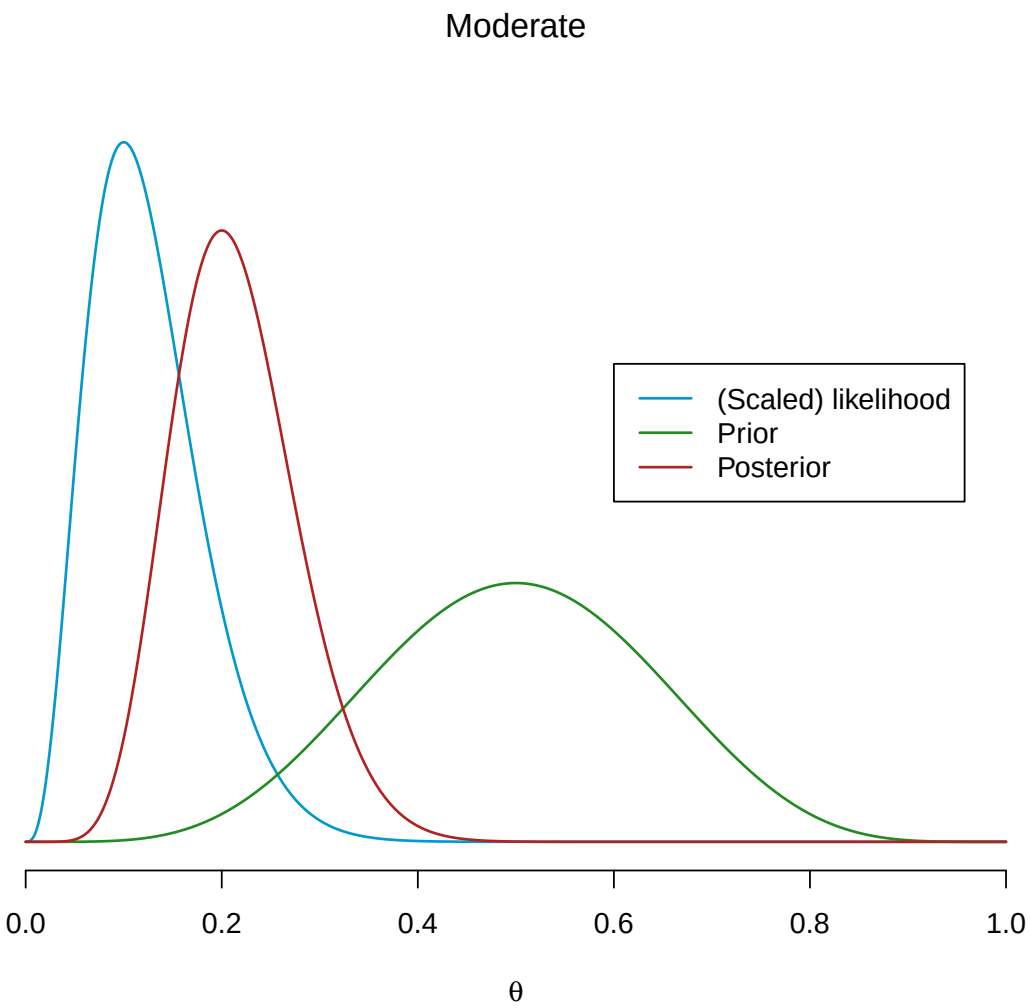
Prior information influences posterior: In fixed samples

- On the other hand, suppose I fix the sample size at $n = 30$. Let's take a look at what happens when I force the prior to "narrow in" on its center of 0.5.
- On the following three slides, I fix the sample size, but update the prior. Don't worry (yet) about the calculations—just focus on the pictures.
- Anticipating behavior heuristically: a prior can be manipulated to exert influence on the posterior in a fixed sample.

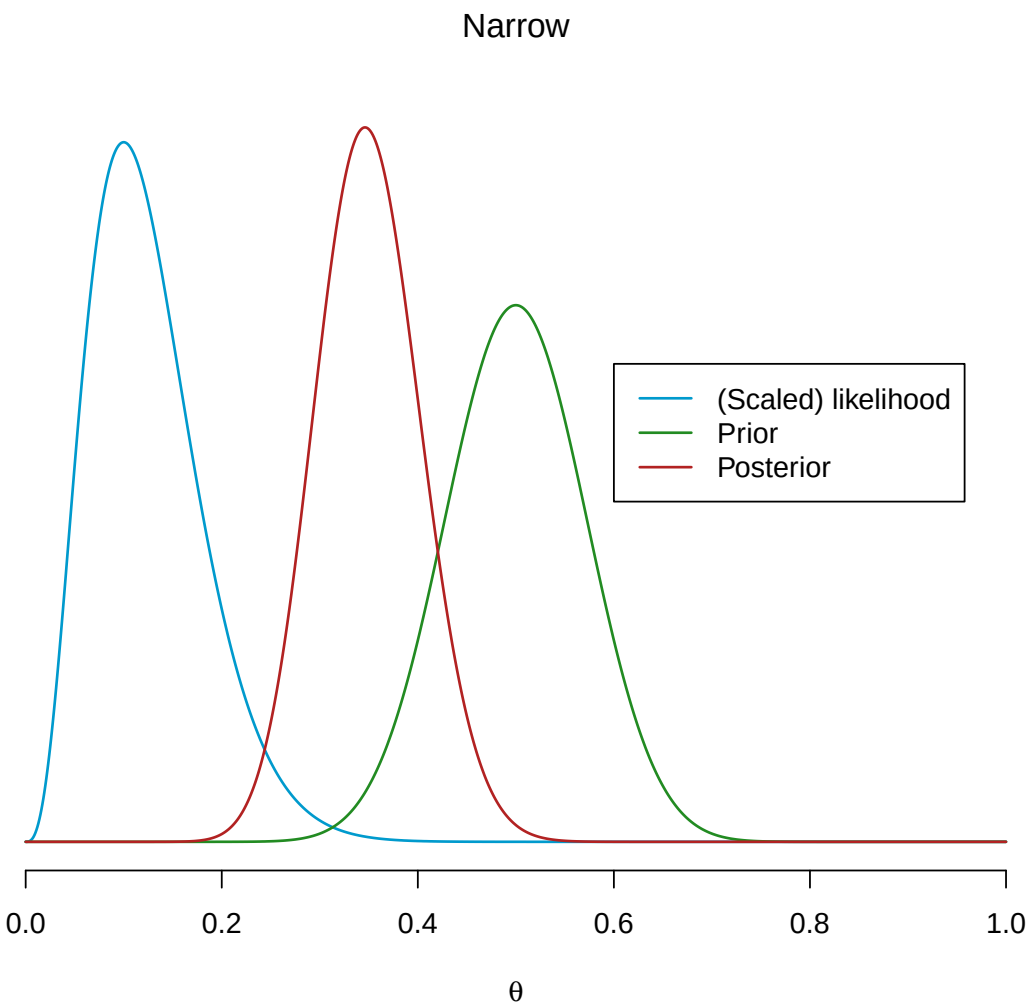
Prior information influences posterior: In fixed samples



Prior information influences posterior: In fixed samples



Prior information influences posterior: In fixed samples



Prior/likelihood/sample size: A trade-off

- Understand the important trade-off being described in the previous slides (and why each side of the trade-off can make one comfortable or uncomfortable with Bayesian methods).
  - To inject a little reality to those camping out near either extreme, note first that we never have an infinite sample size. Note second that a frequentist has just as many opportunities to "cheat" as a Bayesian does.
  - Again, I'm not taking sides—I'm just characterizing the trade-offs.
- Make your prior explicit to avoid unnecessary problems.

The Bayesian framework: Notes

- We previously developed theory for point estimation and inferences on the basis of likelihood theory (along with other methods that aren't quite as relevant for this set of slides).
- Our goal now is to develop a bit more theory for estimation and inferences as a Bayesian on the basis of the *posterior*, which incorporates information from the likelihood and the prior.

The Bayesian framework: Notes

- Bayesian methods are deceptively simple to describe.
- Must overcome the following barriers:
  - Difficulty in specification of a prior distribution.
  - Difficulty in evaluating integrals required for inference.
    - ⋆ Particularly if wanting an analytic solution.
- In these notes, we will discuss:
  - Both analytic *and* computational techniques.
  - Matters surrounding both implementation *and* interpretation (commentary scattered throughout these slides).
- The *next* set of slides will get a little bit more "in-the-weeds" on decision-theoretic matters.

Prior specification: Two flavors

- Weakly informative/baseline/diffuse priors.
  - Wish for prior to have minimal influence on results (so that the likelihood is driving the analysis).
  - Note that the degree of "informativeness" is not always scale-invariant. More on this later.
- Substantive/informative priors.
  - Wish for the prior to have a greater role in informing belief after collecting data.

Posterior:

- There are many ways to obtain the posterior distribution implied by the likelihood and prior.
  - Mathematically convenient choices (e.g., conjugate priors).
  - Computationally (approximations, simulation-based).
- We will soon discuss different ways of obtaining the posterior. For now, assume we've obtained the posterior and are prepared to use it for analysis.
- How do we use/interpret it?

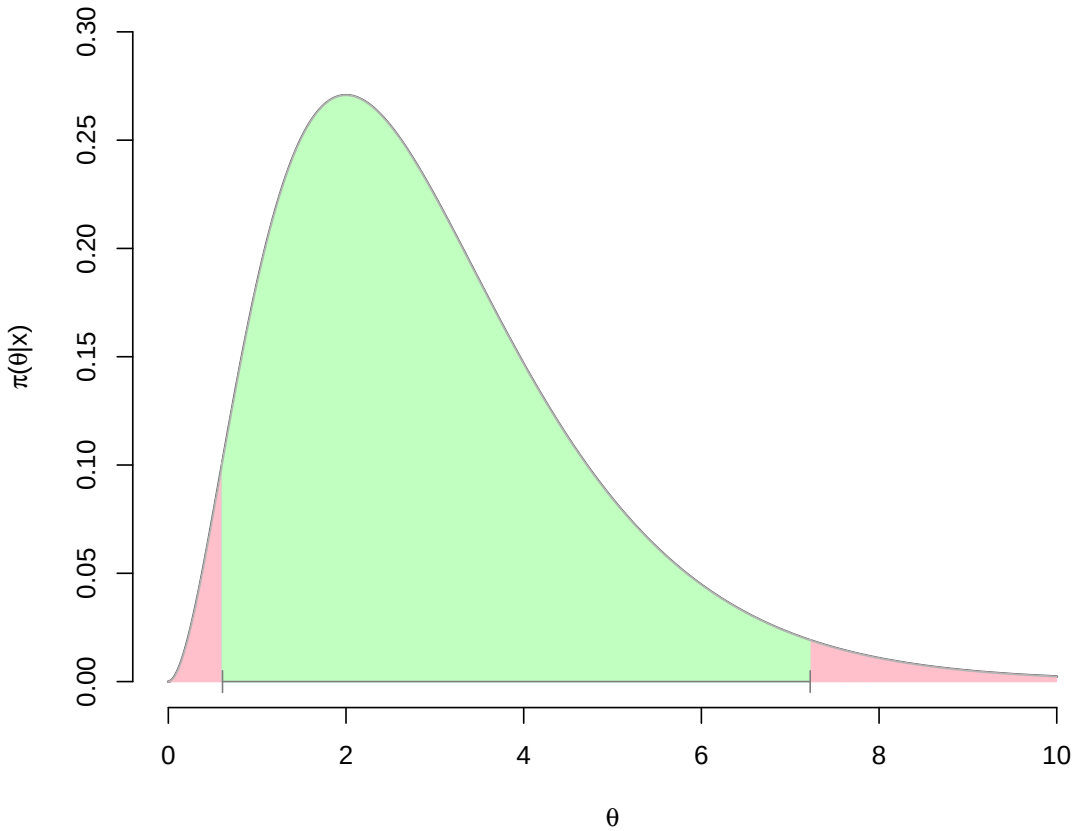
Point estimation:

- Can derive point estimator, $\widehat{\theta}_n$, for θ.
- Posterior mean: $\widehat{\theta}_n = \mathrm{E}[\theta|\mathbf{X}]$ (the mean of the posterior).
  - Characterized as the "Bayes estimator" under squared error loss (and under the given prior). More on this in the next unit.
  - This will be our default approach in this set of slides unless otherwise specified.
- Posterior median: Bayes estimator under *absolute* loss.
- Posterior *mode* most closely resembles the maximum likelihood estimator in spirit, although not necessarily widely implemented.

Interval estimation:

- Intervals formed on the basis of the posterior distribution are referred to as *credible intervals*.
- They characterize the range of θ that are "most believable" on the basis of the data (and the prior belief regarding θ).
- As a Bayesian, we say there is a 95% probability that θ lies between the endpoints of a 95% credible interval (I prefer "probability" in place of "chance," as we're conceptualizing probability in the subjective sense, and *not* in the long-term frequency sense).
- A confidence interval does not afford us the same interpretation because the number "95%" emerges from conceptualizing the coverage of random endpoints around the fixed, unknown parameter.

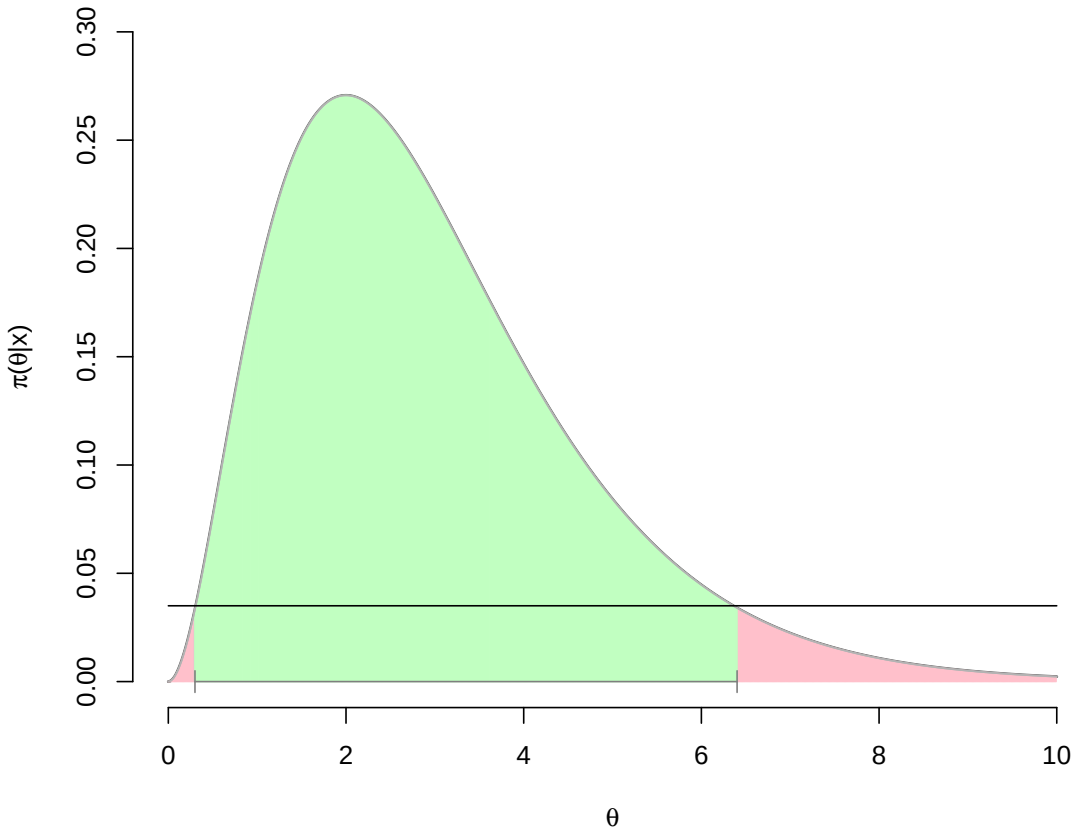
Interval estimation: Posterior quantiles



Interval estimation: Problems with the quantile method

- When the posterior is right-skewed, the right-hand side of the quantile-based interval has a lower density than regions that were ruled out on the right-hand side.
- Further, you may also be able to see how things could go wrong if the posterior is "severely bimodal."
- Highest posterior density (HPD) method more frequently used.
  - Convince yourself that this approach could (but may not) result in a union of two disjoint intervals if applying it to a bimodal posterior.
  - Consider $(a, b) \cup (c, d)$ as a credible region, where $b < c$.
  - There is no reason to restrict our characterization of a "range of plausible estimates" to be path-connected.
- HPD method provides smallest-sized $(1 - \alpha)$ range.

Interval estimation: Highest posterior density



Interval estimation: Highest posterior density (R code)

```
HPDint <- function(x, prob = 0.95) {
  x      <- sort(x)
  nsamp <- length(x)
  gap    <- max(1, min(nsamp - 1, round(nsamp * prob)))
  init   <- 1:(nsamp - gap)
  inds   <- which.min(x[init + gap] - x[init])
  c(Lower =  x[inds], Upper = x[inds + gap])
}
```

Table of Contents

Conjugate priors

Prior specification: Conjugate priors

- It is sometimes possible to choose a family of priors so that the posterior is in the same family as the prior. Such priors are referred to as *conjugate priors*.
  - Broadly, this is convenience-based, rather than evidence-based.
  - You can choose which specific probability function to use for the prior within that family—that choice can be evidence-based.
- The parameters of a conjugate prior control the informativeness of the prior (on the scale of θ).

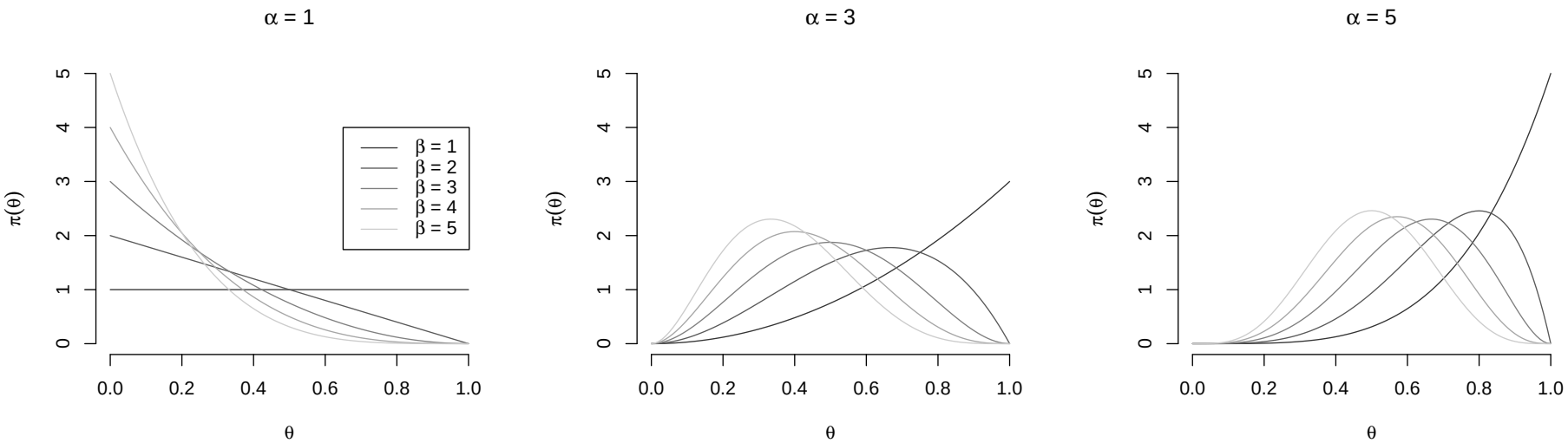
Example 5.1: Bernoulli and Beta

- Suppose $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\theta)$, with θ unknown. You seek to estimate θ as a Bayesian.
- Mathematically convenient to choose $\theta \sim \text{Beta}(\alpha, \beta)$:

$$\pi(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}.$$

- I omitted the notation for support (i.e., the term $I_{(0,1)}(\theta)$), taking it as implicit—but it *must* be in the back of your mind.
- α and β are referred to as the prior hyperparameters.
- How do α and β control the shape and position of the prior?

Example 5.1: Bernoulli and Beta (various beta density plots)



Pay attention to symmetries.

Example 5.1: Bernoulli and Beta

- Note: $\pi(\theta|\mathbf{X}=\mathbf{x}) \propto_\theta p_{\mathbf{X}}(\mathbf{x};\theta)\pi(\theta)$. How does this help us?

$$\begin{aligned}\pi(\theta|\mathbf{X}=\mathbf{x}) &\propto_\theta \left(\prod_{i=1}^{n}\theta^{x_i}(1-\theta)^{1-x_i}\right)\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \theta^{\sum_{i=1}^{n}x_i+\alpha-1}(1-\theta)^{n-\sum_{i=1}^{n}x_i+\beta-1}.\end{aligned}$$

- Now, note that the kernel matches that of the prior to define the parameters of the posterior:

$$\theta|\mathbf{X}=\mathbf{x} \sim \text{Beta}\left(\alpha^* = \alpha + \sum_{i=1}^{n}x_i, \beta^* = \beta + n - \sum_{i=1}^{n}x_i\right).$$

- α^* and β^* are referred to as the posterior hyperparameters.

Example 5.1: Bernoulli and Beta

- Posterior mean (Bayes estimator under squared-error loss):

$$\begin{aligned}\widehat{\theta}_n = \mathrm{E}[\theta|\mathbf{X}] &= \frac{\alpha^*}{\alpha^* + \beta^*} = \frac{\alpha + \sum_{i=1}^n X_i}{\alpha + \beta + n} \\ &= \frac{\alpha}{\alpha + \beta + n} + \frac{\sum_{i=1}^n X_i}{\alpha + \beta + n} \\ &= \frac{\alpha + \beta}{\alpha + \beta + n} \cdot \frac{\alpha}{\alpha + \beta} + \frac{n}{\alpha + \beta + n} \cdot \frac{\sum_{i=1}^n X_i}{n}.\end{aligned}$$

- Recognize this as a convex combination (weighted average) of the prior mean and $\widehat{\theta}_n$, with relative weights controlled by α and β.
- How does n affect the weights?
- Under what conditions is $\widehat{\theta}_n$ unbiased for θ?
- Note the extremely convenient interpretation of $\alpha + \beta$ as if a prior number of trials and α as if the prior number of successes.

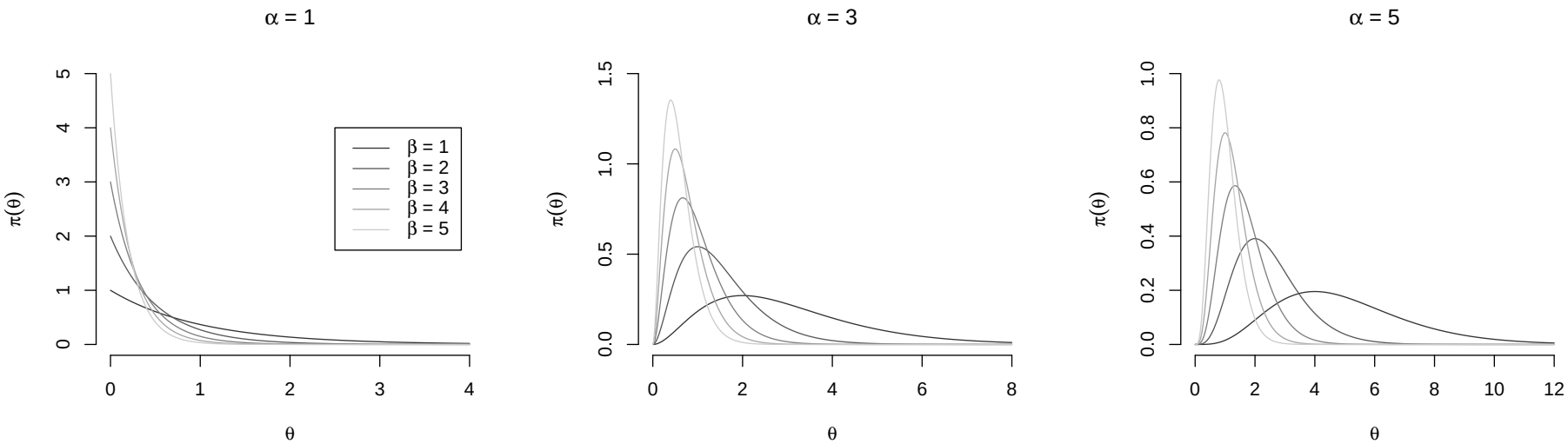
Example 5.2: Poisson and Gamma

- Suppose $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\theta)$, with θ unknown. You seek to estimate θ as a Bayesian.
- Mathematically convenient to choose $\theta \sim \text{Gamma}(\alpha, \beta)$:

$$\pi(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} \exp(-\beta\theta).$$

- I omitted the notation for support (i.e., the term $I_{(0,\infty)}(\theta)$), taking it as implicit—but it *must* be in the back of your mind.
- α and β are referred to as the prior hyperparameters.
- How do α and β control the shape and position of the prior?

Example 5.2: Poisson and Gamma (various Gamma density plots)



Left-skewed priors are notably ruled out by the Gamma family.

Example 5.2: Poisson and Gamma

- Note: $\pi(\theta|\mathbf{X} = \mathbf{x}) \propto_\theta p_{\mathbf{X}}(\mathbf{x}|\theta)\pi(\theta)$. How does this help us?

$$\pi(\theta|\mathbf{X} = \mathbf{x}) \quad \propto_\theta \quad \left(\prod_{i=1}^{n} \theta^{x_i} \exp(-\theta)\right) \theta^{\alpha-1} \exp(-\beta\theta).$$

- Now, note that the kernel matches that of the prior to define the parameters of the posterior:

$$\theta|\mathbf{X} = \mathbf{x} \sim \text{Gamma}\left(\alpha^* = \alpha + \sum_{i=1}^{n} x_i, \beta^* = \beta + n\right).$$

- α^* and β^* are referred to as the posterior hyperparameters.

Example 5.2: Poisson and Gamma

- Posterior mean (Bayes estimator under squared-error loss):

$$\begin{aligned}\widehat{\theta}_n = \mathrm{E}[\theta|\mathbf{X}] &= \frac{\alpha^*}{\beta^*} = \frac{\alpha + \sum_{i=1}^n X_i}{\beta + n} = \frac{\alpha}{\beta + n} + \frac{\sum_{i=1}^n X_i}{\beta + n} \\ &= \frac{\beta}{\beta + n} \cdot \frac{\alpha}{\beta} + \frac{n}{\beta + n} \cdot \frac{\sum_{i=1}^n X_i}{n}.\end{aligned}$$

- Recognize this as a convex combination (weighted average) of the prior mean and $\widehat{\theta}_n$, with relative weights controlled by α and β.
- How does n affect the weights?
- Under what conditions is $\widehat{\theta}_n$ unbiased for θ?
- Note the extremely convenient interpretation of β as if a prior sample size and α as if the sum of prior count observations.

Bayesian sufficiency:

- You may have noticed that the sufficient statistic shows up in these examples. That's no coincidence.
- We say that $T(\mathbf{X})$ is *Bayesian sufficient* if the posterior $\pi(\theta|\mathbf{X} = \mathbf{x})$ necessarily depends upon $\mathbf{x}$ only through the value of $T(\mathbf{x})$.
- Sufficiency implies Bayesian sufficiency (immediate from factorization theorem).
- If you know the distribution of the sufficient statistic, you can use it directly to obtain the posterior:

$$\pi(\theta|T = t) \propto_\theta p_T(t;\theta)\pi(\theta).$$

Conjugate priors

Prior mean and MLE:

- You may have noticed that in both cases, the posterior mean could be expressed as a weighted average of the prior mean and the MLE.
- This will *always* happen when you have a conjugate prior for a likelihood in the exponential family.
- When expressed in this way, you can typically think of it as "data augmentation."

Conjugate priors

Other conjugate priors: Examples (not exhaustive)

- Normal likelihood with unknown mean μ/known variance σ^2.
  - Conjugate prior: Normal on μ.
- Normal likelihood with known mean μ/unknown variance σ^2.
  - Conjugate prior: Inverse-gamma on σ^2.
- Uniform$(0, \theta)$ likelihood.
  - Conjugate prior: Pareto.
- Weibull likelihood with known shape β/unknown scale θ.
  - Conjugate prior: Inverse gamma.
- The Bernoulli-Beta, Poisson-Gamma, and Normal-Normal ones are worth committing to memory.

A two-parameter model for your enjoyment:

- Suppose $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, with both μ and σ^2 unknown.
- We now need a *joint* prior on $\boldsymbol{\theta} = (\mu, \sigma^2)$.
- We typically parameterize hierarchically based on precision, $\tau = 1/\sigma^2$:

$$\begin{aligned} X_i|\boldsymbol{\theta} &\sim \mathcal{N}(\mu, 1/\tau), \\ \mu|\tau &\sim \mathcal{N}(\delta, (\lambda\tau)^{-1}), \\ \tau &\sim \text{Gamma}(\alpha, \beta). \end{aligned}$$

- This is a conjugate prior for this likelihood; the posterior for $\boldsymbol{\theta}|\mathbf{X} = \mathbf{x}$ is also (hierarchically) in the Normal/Inverse-gamma family, with posterior hyperparameters:
  - $\delta^* = (\lambda\delta + n\overline{x}_n)/(\lambda + n)$.
  - $\lambda^* = \lambda + n$.
  - $\alpha^* = \alpha + n/2$.
  - $\beta^* = \beta + (1/2)s_n^2 + [n\lambda/(\lambda + n)]((\overline{x}_n - \delta)^2)/2)$.

A two-parameter model for your enjoyment:

- If one seeks to carry out inference regarding μ, knowing the posterior $\mu|\tau, \mathbf{X} = \mathbf{x}$ isn't the most helpful.
- Instead, one wants the *marginal* posterior, $\mu|\mathbf{X} = \mathbf{x}$.
- It turns out—with some math—that this will be a location-scale t-distribution:

$$\mu|\mathbf{X} = \mathbf{x} \quad \sim \quad t_{2\alpha^*}\left(\delta^*, \sqrt{\frac{\beta^*}{\lambda^*\alpha^*}}\right),$$

  which is another way of saying that $(\mu - \delta^*)/\sqrt{\beta^*/(\lambda^*\alpha^*)} \sim t_{2\alpha^*}$.
- Truly, the computational approaches (to be addressed very shortly) are sometimes a little easier than the messy algebra that would lead to these conclusions.

Conjugate priors

Motivating other approaches:

- There is no general reason for the posterior to take a mathematically "nice" form.
- If it doesn't, we often rely on computational tools, which can include:
- Integral approximation (we will not discuss these):
  - Laplace approximation.
  - Gauss-Hermite quadrature.
  - Variational methods.
- Markov Chain Monte Carlo (MCMC) and other algorithms:
  - Rejection sampling.
  - Metropolis-Hastings algorithm.
  - Gibbs sampler.
  - Hamiltonian methods.
- We are in the parametric Bayesian framework, but we are now moving to Bayesian computation that addresses the intractability of the marginal distribution (e.g., when there is no conjugacy).

Table of Contents

Grid approximation

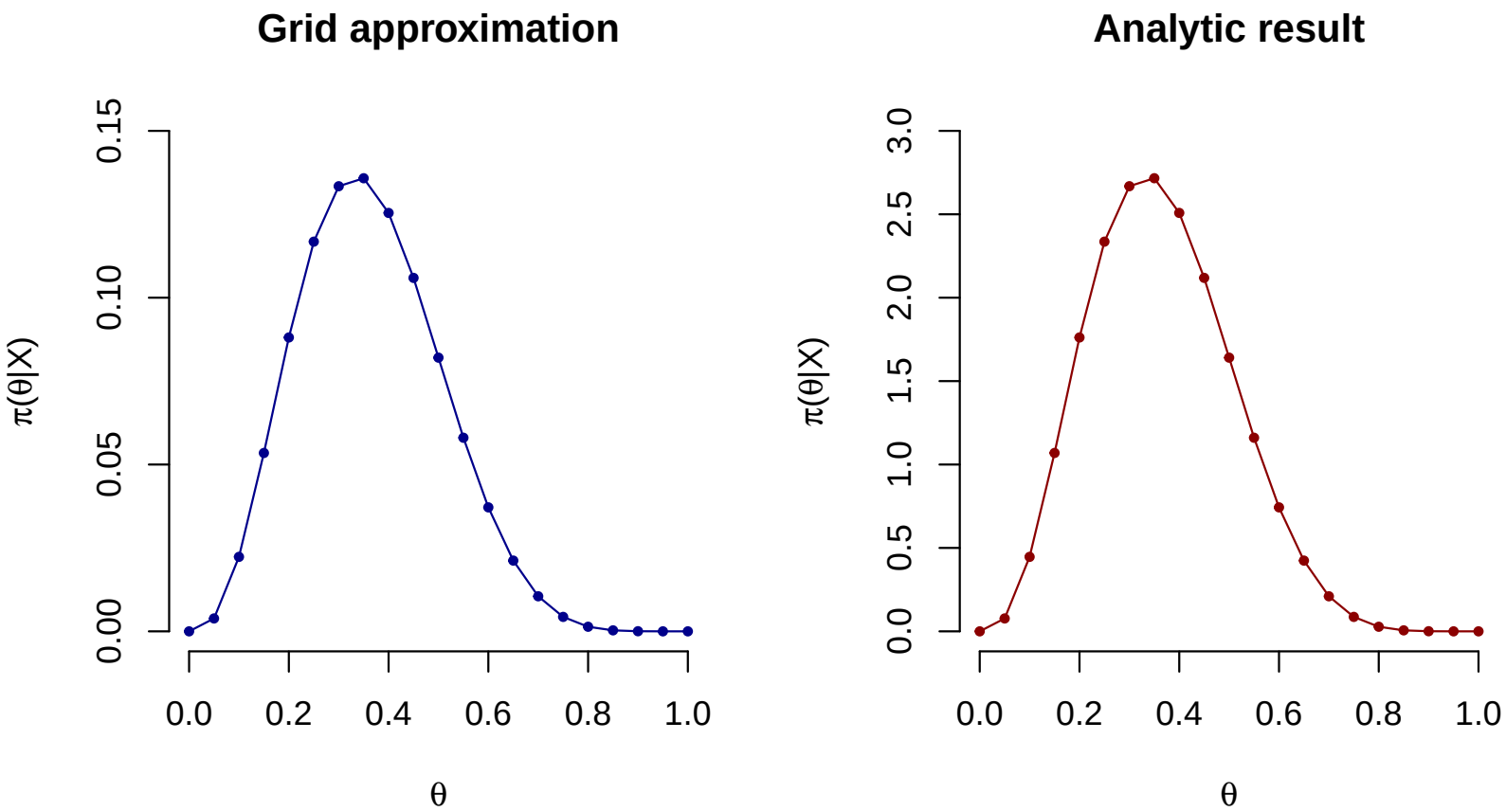
Posterior: Computational trick

- Suppose, I lack an analytically "nice" form for the posterior. We can use a grid approximation to the posterior as follows:
  1. Define a grid of discrete values for $\theta_1, \ldots, \theta_J$.
  2. Compute the likelihood at each value of θ_j: $p_{\mathbf{X}}(\mathbf{x}; \theta_j)$.
  3. Compute the prior at each value of θ_j: $\pi(\theta_j)$
  4. Compute the unstandardized posterior: $\pi^*(\theta_j|\mathbf{X} = \mathbf{x}) = p_{\mathbf{X}}(\mathbf{x}; \theta_j)\pi(\theta_j)$.
  5. Standardize: $\pi(\theta_j|\mathbf{X} = \mathbf{x}) = \pi^*(\theta_j|\mathbf{X} = \mathbf{x}) / \sum_{k=1}^{J} \pi^*(\theta_k|\mathbf{X} = \mathbf{x})$.
- This does not work well for high-dimensional settings.
- However—as an example—suppose that I'd forgotten that the Bernoulli-Beta example is conjugate and I wanted to use a grid approximation.

Example 5.3: Grid approximation for Bernoulli/Beta

```
## Define grid
theta.grid <- seq(0, 1, 0.05)

## Suppose theta ~ Beta(1, 1) == Uniform(0, 1) is my prior
prior <- rep(1, length(theta.grid))

## Suppose there were three successes in nine trials
likelihood <- dbinom(x = 3, size = 9, prob = theta.grid)

## Compute unstandardized posterior
unstd.posterior <- likelihood * prior

## Standardize the posterior
posterior <- unstd.posterior/sum(unstd.posterior)
```

Example 5.3: Grid approximation for Bernoulli/Beta



Why don't the y-axis scales match?

Grid approximation

Example 5.4: Obligatory coin-toss example (Setup)

- You are provided a coin; you are told it is a biased coin that handily favors one direction, but you're not told which direction.
  - We had a similar setup in the previous set of slides.
- Likelihood: $X \sim \text{Bernoulli}(\theta)$.
- You want to answer this problem as a Bayesian by placing a bimodal prior on θ and updating with incoming binary data.

Example 5.4: Obligatory coin-toss example (Setting up the prior)

- Consider a *mixture* distribution to specify the prior for θ:

$$\theta \sim \text{“Beta}(\alpha = 2, \beta = 9) \times 0.5 + \text{Beta}(\alpha = 9, \beta = 2) \times 0.5.\text{”}$$

- Put more formally, the prior has a density given by:

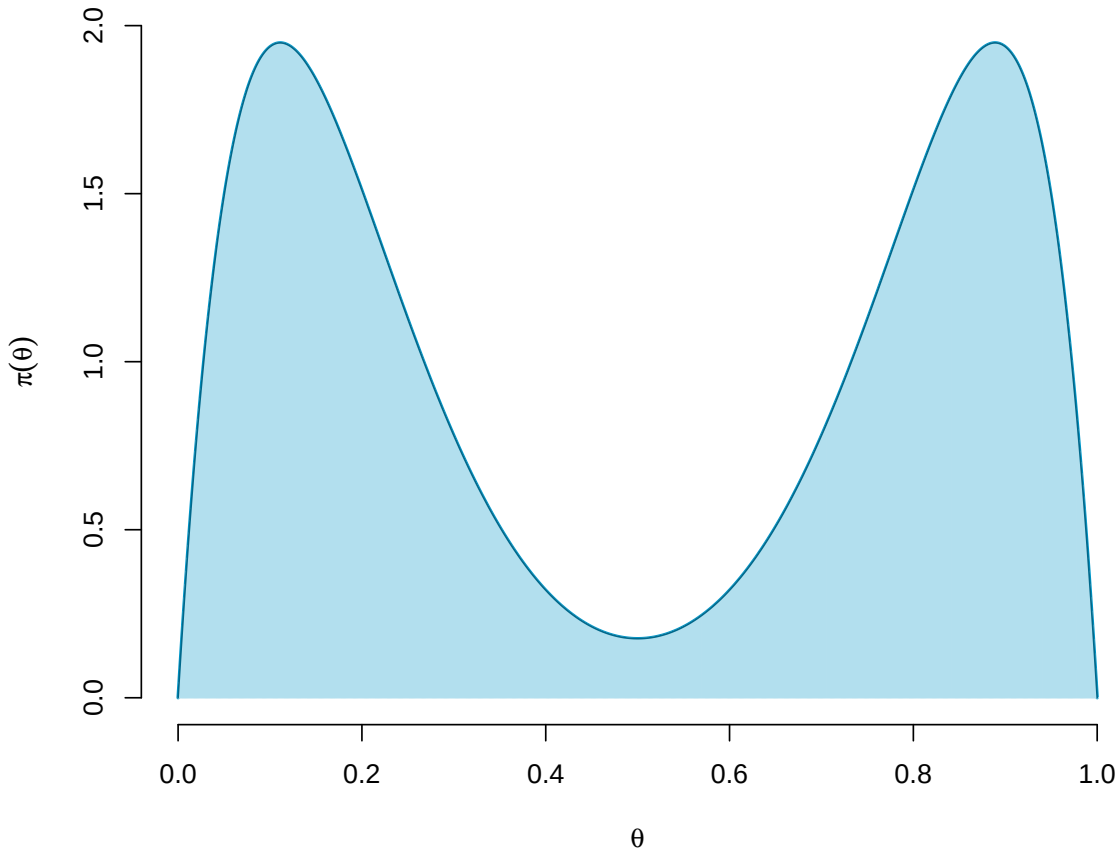
$$\pi(\theta) = \frac{1}{2} f_{\text{Beta}}(\theta; \alpha = 2, \beta = 9) + \frac{1}{2} f_{\text{Beta}}(\theta; \alpha = 9, \beta = 2),$$

  where $f_{\text{Beta}}(\theta; \alpha, \beta)$ is the Beta(α, β) density on $\theta \in (0, 1)$.

- One way to think of a mixture distribution is hierarchically:

$$\begin{aligned} Z &\sim \text{Bernoulli}(p = 0.5) \\ \theta_1 &\sim \text{Beta}(\alpha = 2, \beta = 9) \\ \theta_2 &\sim \text{Beta}(\alpha = 9, \beta = 2) \\ \theta &= \theta_1 Z + \theta_2 (1 - Z). \end{aligned}$$

Example 5.4: Obligatory coin-toss example (Setting up the prior)

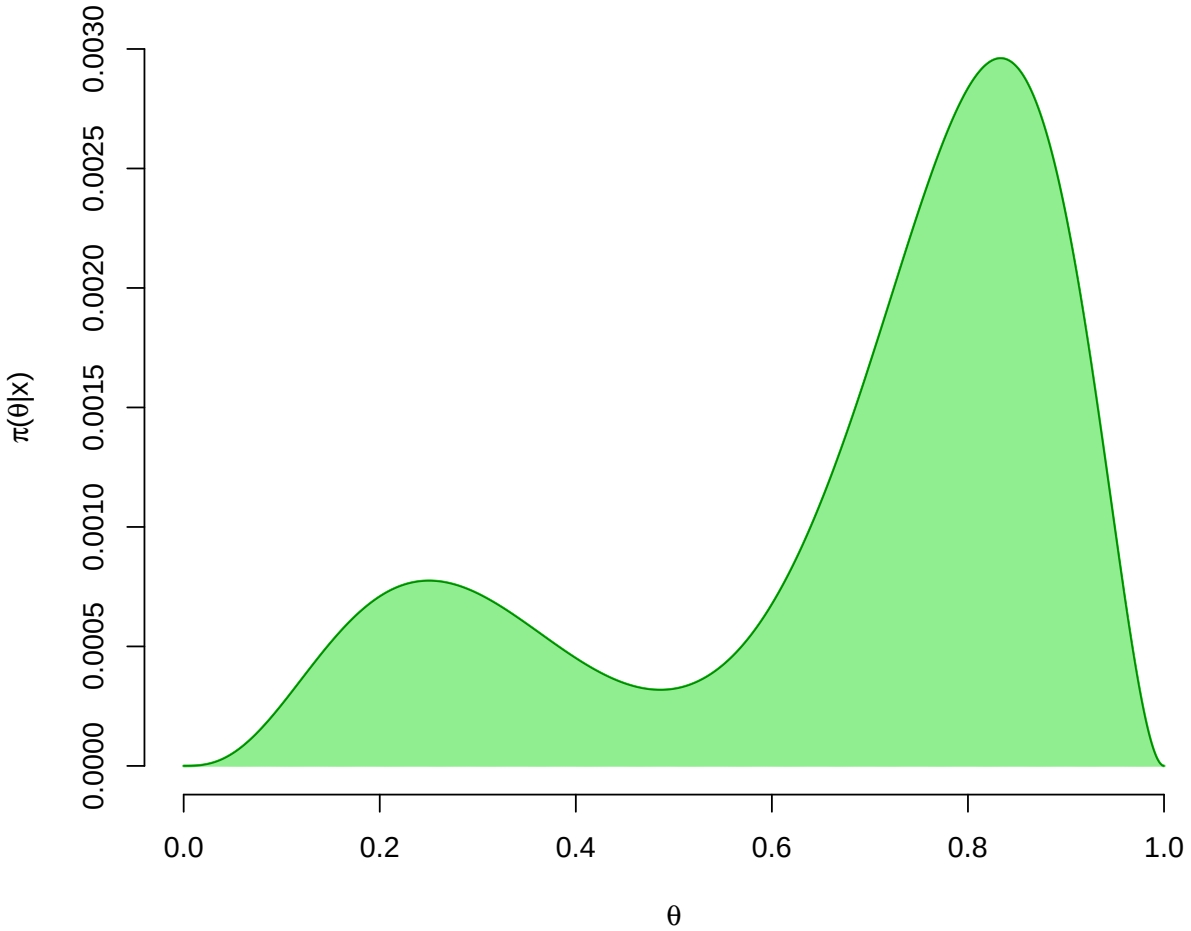


Grid approximation

Example 5.4: Obligatory coin-toss example (In come the data!)

- As an exercise, imagine what the posterior might look like under the following hypothetical data sets:
  1. $\{H\}$
  2. $\{T\}$
  3. $\{H, T\}$
  4. $\{T, H\}$
  5. $\{H, H, T\}$
  6. $\{H, H, T, H\}$
  7. $\{H, T, T, T, H, T, H, T, H, H\}$
- Conjugate priors won't save us. But some of our computational methods might! Let's choose hypothetical data set #5 for now. I believe this model is simple enough for the grid approximation to be just fine (try it as an exercise).

Example 5.4: Obligatory coin-toss example (Data: $\{H, H, T\}$)



Grid approximation

Example 5.4: Obligatory coin-toss example (Continued data collection)

- Consider tossing the coin and recording your response until you are able to arrive at a decision. In particular, consider the following decision rule:
  - Decide bias toward tails if $\pi(\theta < 0.50 | \mathbf{X} = \mathbf{x}) > 0.95$.
  - Decide bias toward heads if $\pi(\theta > 0.50 | \mathbf{X} = \mathbf{x}) > 0.95$.
- One of the challenges is that if $\theta \approx 0.5$, this decision rule may suffer. To address this, you could impose another rule:
  - Decide approximate equivalence if $\pi(|\theta - 0.5| < 0.05 | \mathbf{X} = \mathbf{x}) > 0.95$.
- If you have a budget constraint, you could impose another rule:
  - Declare futility (in the sense of inability to decide among the above) if $n = n_{\text{max}}$ if none of the above criteria are met by some fixed sample size n_{max}.

Grid approximation

Data collection and interim monitoring:

- In Bayesian world, there is no fundamental difference between:
  - Updating with $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$ altogether,
  - Updating with $\mathbf{x}_1$ and then $\mathbf{x}_2$, sequentially.
  - Updating with $\mathbf{x}_2$ and then $\mathbf{x}_1$, sequentially.
- Every time you update your prior with data, that posterior serves as the prior for data "yet-to-be-collected."
- This is the basis for sequential monitoring in Bayesian world.
- Simply put, it is **much, much** harder to implement these sorts of stopping boundaries as a frequentist because type 1 error control is very challenging (group-sequential trials).
  - Pocock boundary.
  - O'Brien-Fleming boundary.
  - Lan-DeMets boundary.
- A Bayesian is not targeting a specific type 1 error rate. This can cause friction between the frequentist and the Bayesian.

Table of Contents

Rejection sampling

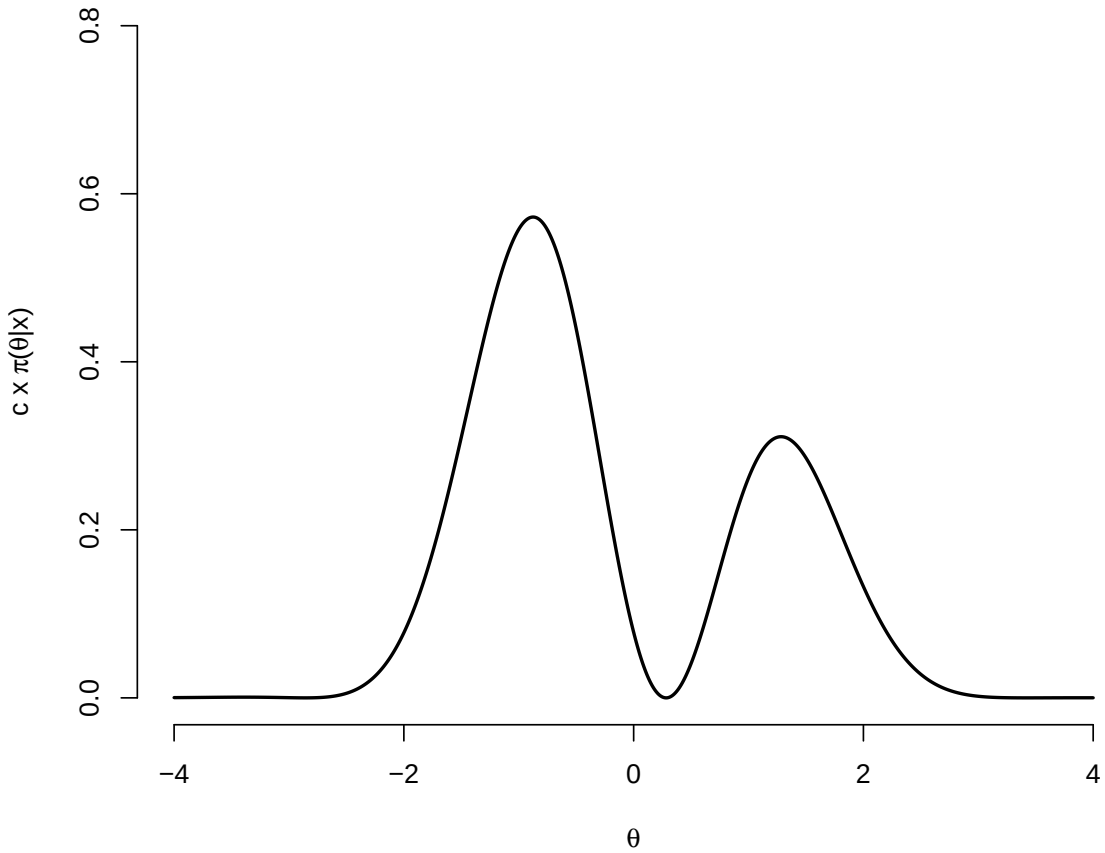
The problem:

- You're likely familiar with an algorithm to generate draws from a known distribution (namely, the inverse-CDF method).
  - This relies on knowing the CDF of your target distribution and being able to invert it. It therefore has limited utility in practice.
- The posterior is proportional to the likelihood times the prior, but we may simply not be able to figure out the normalizing constant (which depends upon $\mathbf{x}$ in practice):

$$\pi(\theta|\mathbf{X}=\mathbf{x}) = \frac{1}{c} \times h(\theta); \quad c \text{ is such that } \int_{\Theta} \pi(\theta|\mathbf{X}=\mathbf{x}) d\theta = 1.$$

- Example: Suppose we believe $h(\theta) = \exp(-\theta^2/2) \times \sin^2(6+\theta)$ but do not know the value of the normalizing constant, c. Note: $\Theta = \mathbb{R}$.
  - In this example, $c \approx 0.901$, but make believe we don't know that...
- Recall: From the vantage point of the posterior, $\mathbf{x}$ is a constant.

Example 5.5: $h(\theta) = \exp(-\theta^2/2) \times \sin^2(6 + \theta)$



Example 5.5: The solution

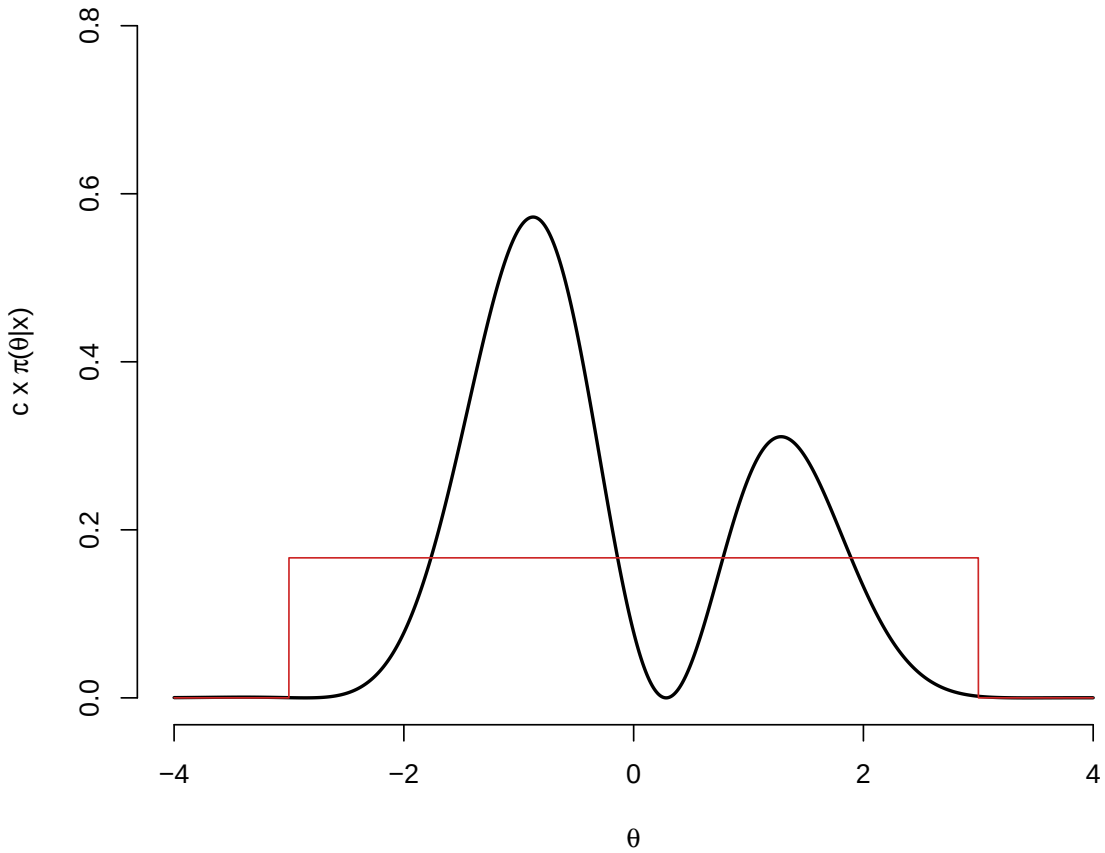
- The rejection sampling algorithm works as follows:
  1. Propose a density, g, that is "easy to draw from."
  2. Scale g by a constant, M, so that $M \times g(\theta) > h(\theta)$ for "all θ."
     - ⋆ Not always possible for all θ over the real line, but this should be possible over a finite range of θ (choose carefully).
  3. Sample random draws from g, $d_1, \ldots, d_R$, accepting probabilistically:

$$\text{P(Accept } d_r) \quad = \quad \frac{h(d_r)}{M \times g(d_r)}.$$

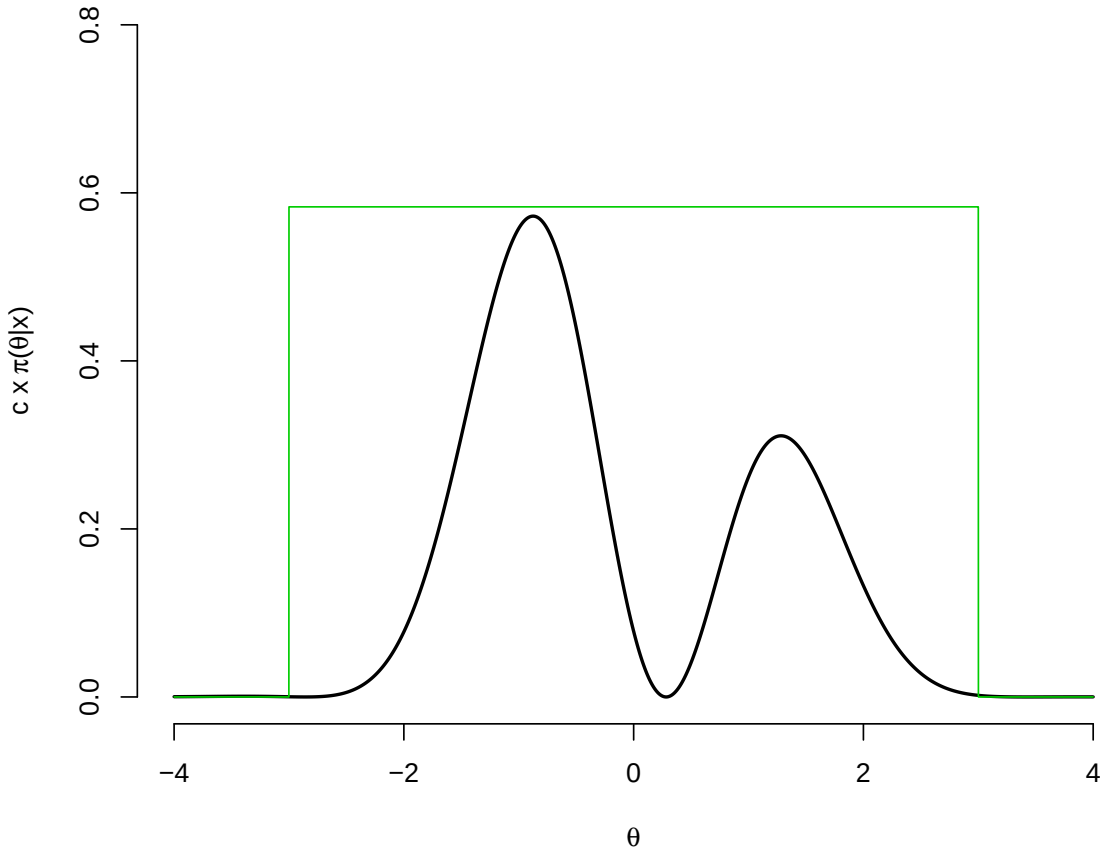
  For example, accept d_r if $\text{P(Accept } d_r) > U_r$; $U_r \sim \text{Uniform}(0, 1)$.

- *Accepted* proposals are draws from target posterior, $\pi(\theta|\mathbf{X} = \mathbf{x})$.
- The reason for scaling g is to guarantee that the acceptance probabilities are smaller than one for any draw we could reasonably sample.

Example 5.5: Try proposing g as Uniform$(-3, 3)$



Example 5.5: Scale up by $M = 3.5$ so that $M \times g(x) > h(x)$

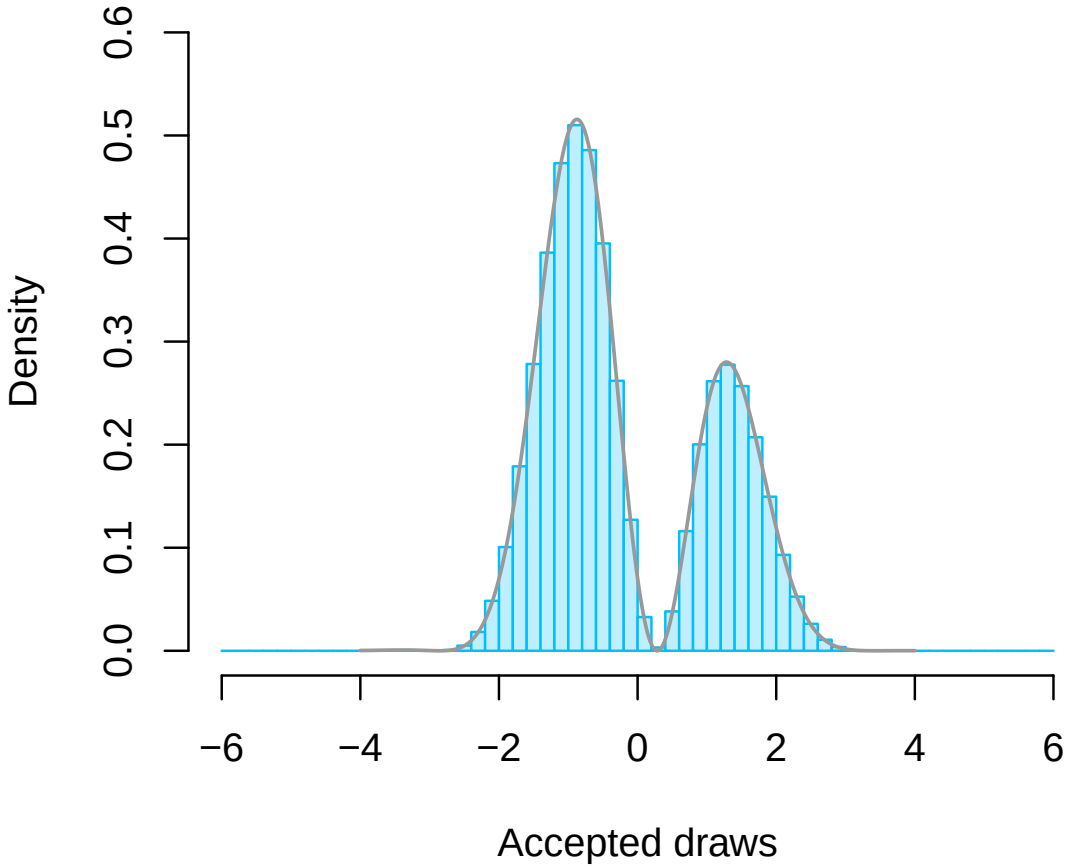


Clear limitation: No candidates outside $(-3, 3)$ *permitted.*

Example 5.5: Algorithm under uniform proposal

```
## Function proportional to "posterior"
h <- function(theta) {
  out <- (exp(-(theta^2)/2))*((sin(6 + theta))^2)
  return(out)
}

## Set seed for reproducibility
set.seed(6342)

## Large number of draws
R <- 5000000
candidates <- runif(R, -3, 3)

## Scale (this is a visual exercise, often)
M <- 3.5

## Algorithm
num <- h(candidates)
denom <- M * dunif(candidates, -3, 3)
prob <- num/denom
accept <- as.numeric(runif(R, 0, 1) < prob)
sample <- candidates[accept == 1]

## What proportion of draws were accepted?
> mean(accept)
[1] 0.3170628
```

Example 5.5: Result under uniform proposal



The true density (unknown in practice) is shown in gray for reference.

Aside: Why it works

- Let A denote the indicator of accepting a draw, D. The density of accepted draws can be derived as follows:

$$p_D(d|A=1) = \frac{\mathrm{P}(A=1|D=d)p_D(d)}{\mathrm{P}(A=1)} = \frac{\frac{h(d)}{M\times g(d)} \times g(d)}{\mathrm{P}(A=1)}.$$

- Now,

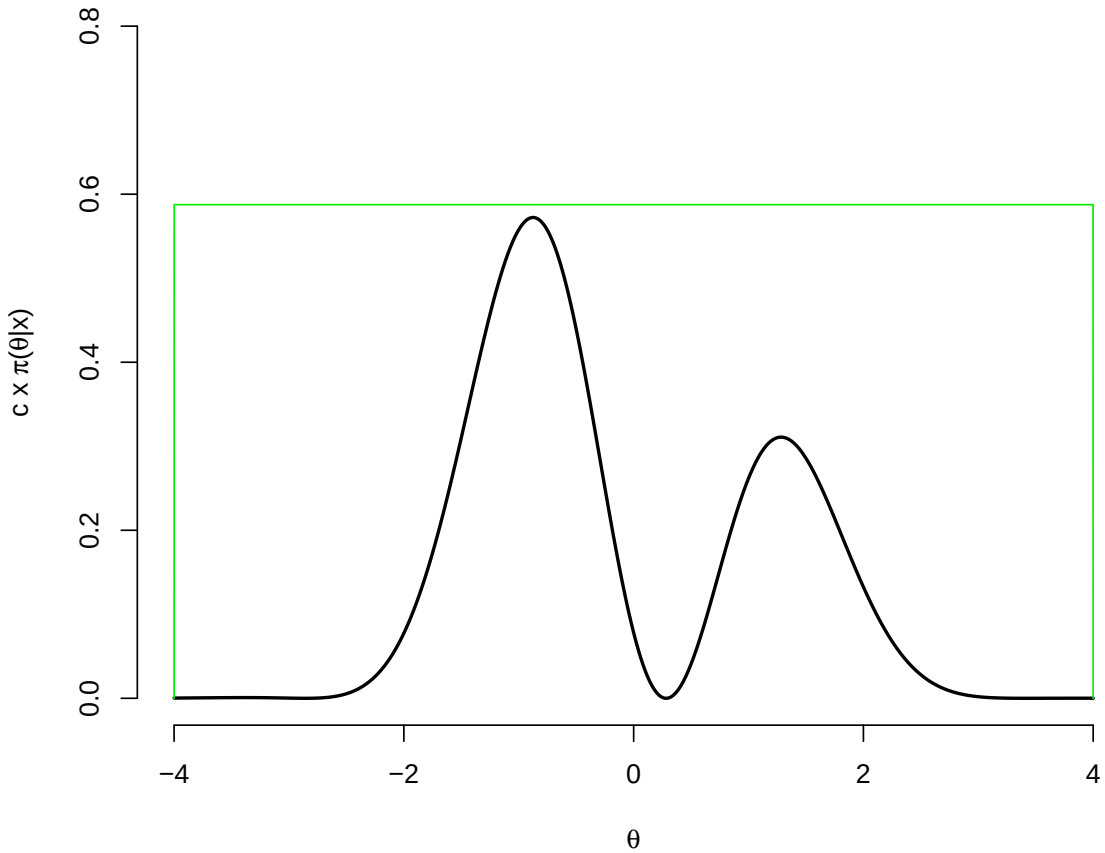
$$\mathrm{P}(A=1) = \int_{\mathbb{R}} g(t) \cdot \frac{h(t)}{M \times g(t)} dt = \frac{1}{M} \int_{\mathbb{R}} c \times \pi(t|\mathbf{X}=\mathbf{x}) dt = \frac{c}{M}.$$

- Therefore,

$$p_D(d|A=1) = \frac{h(d)/M}{c/M} = \frac{h(d)}{c} = \pi(d|\mathbf{X}=\mathbf{x}).$$

- This sheds light on some of the challenges we can run into (and some of the ways we might be able to improve).

Example 5.5: Consider g as a Uniform$(-4, 4)$, with $M = 4.7$

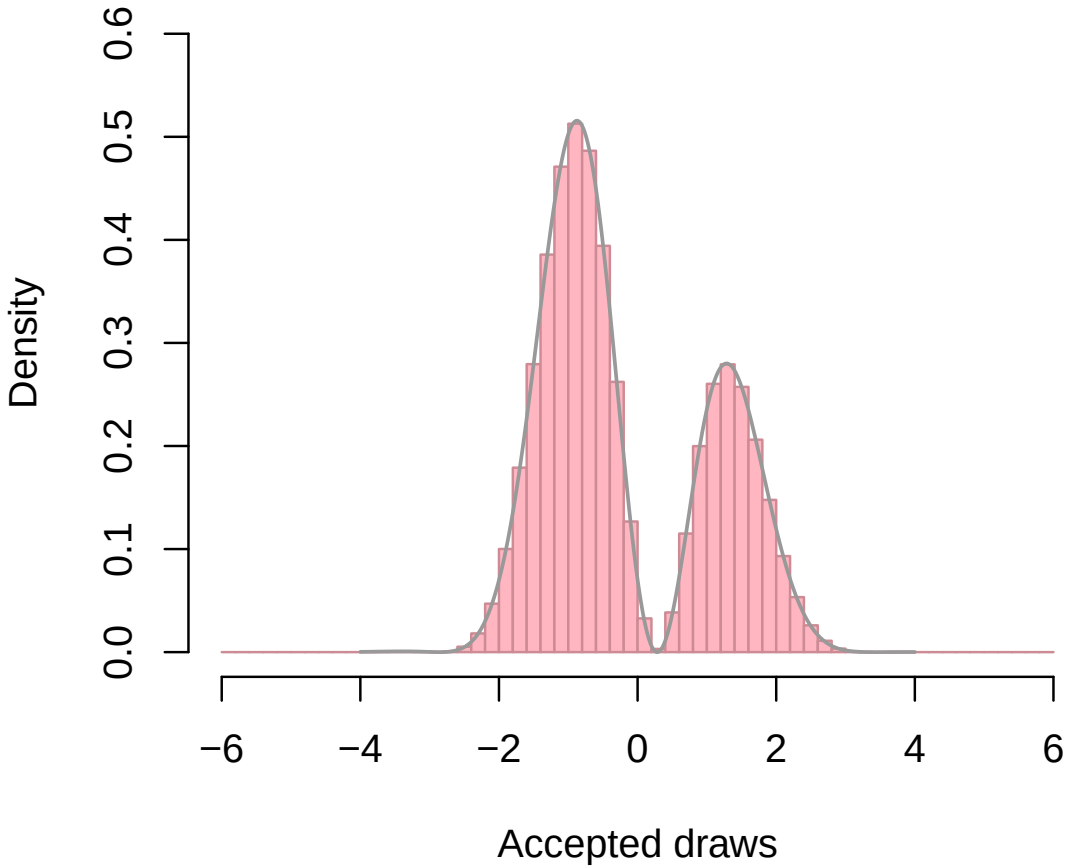


What do you think will happen?

Example 5.5: Algorithm under uniform proposal (larger window)

```
## Set seed for reproducibility
set.seed(6342)

## Updated draws
candidates <- runif(R, -4, 4)

## Scale accordingly
M <- 4.7

## Algorithm
num <- h(candidates)
denom <- M * dunif(candidates, -4, 4)
prob <- num/denom
accept <- as.numeric(runif(R, 0, 1) < prob)
sample <- candidates[accept == 1]

## What proportion of draws were accepted?
> mean(accept)
[1] 0.2362536
```

Example 5.5: Uniform proposal (larger window)



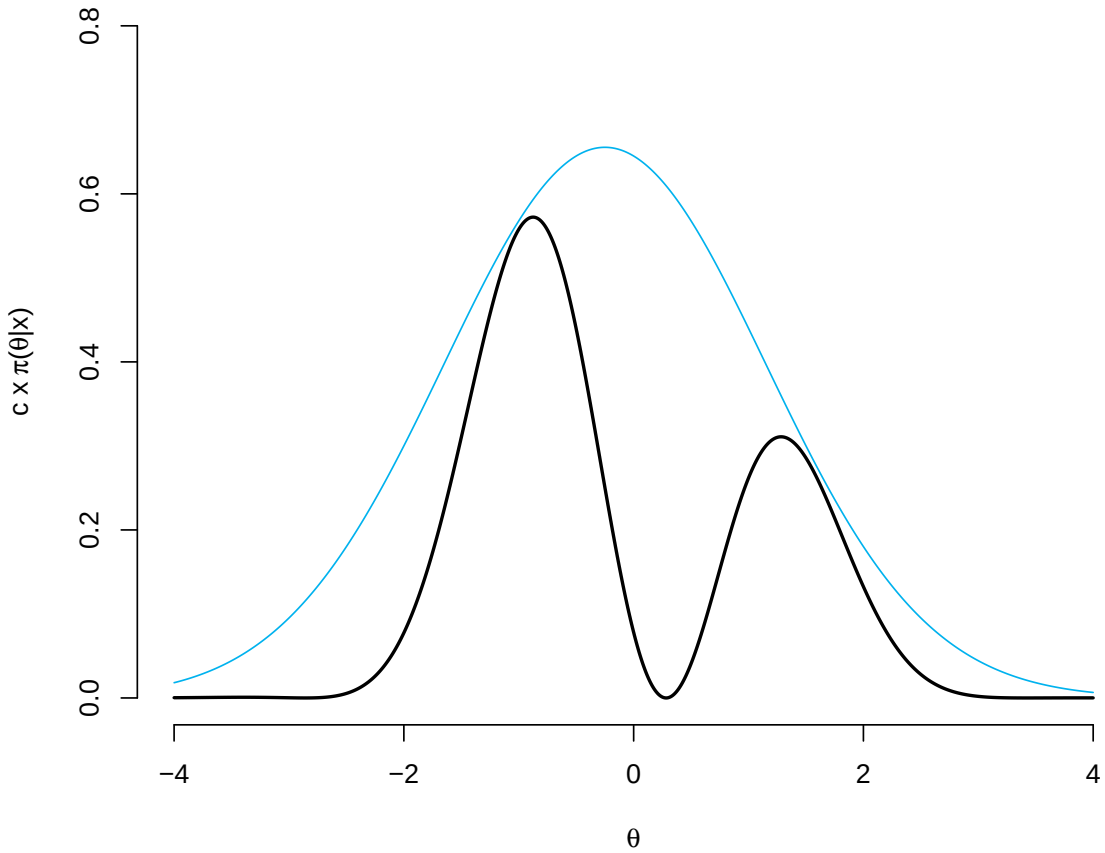
Was it worth it? Probably not.

Example 5.5: Key ideas

- The proportion of accepted draws went down because expanding our range forced us to consider more draws that, according to h, were not very likely.
- Intuitively, it's better (from a computational efficiency standpoint) to be able to propose a g with a similar enough shape to h so that the algorithm is not forced to propose too many draws that are poor candidates.
- You should also be able to see the advantage of keeping M as low as possible.

Rejection sampling

Example 5.5: g corresponds to $\mathcal{N}(\mu = -0.25, \sigma^2 = 1.4^2)$; $M = 2.3$

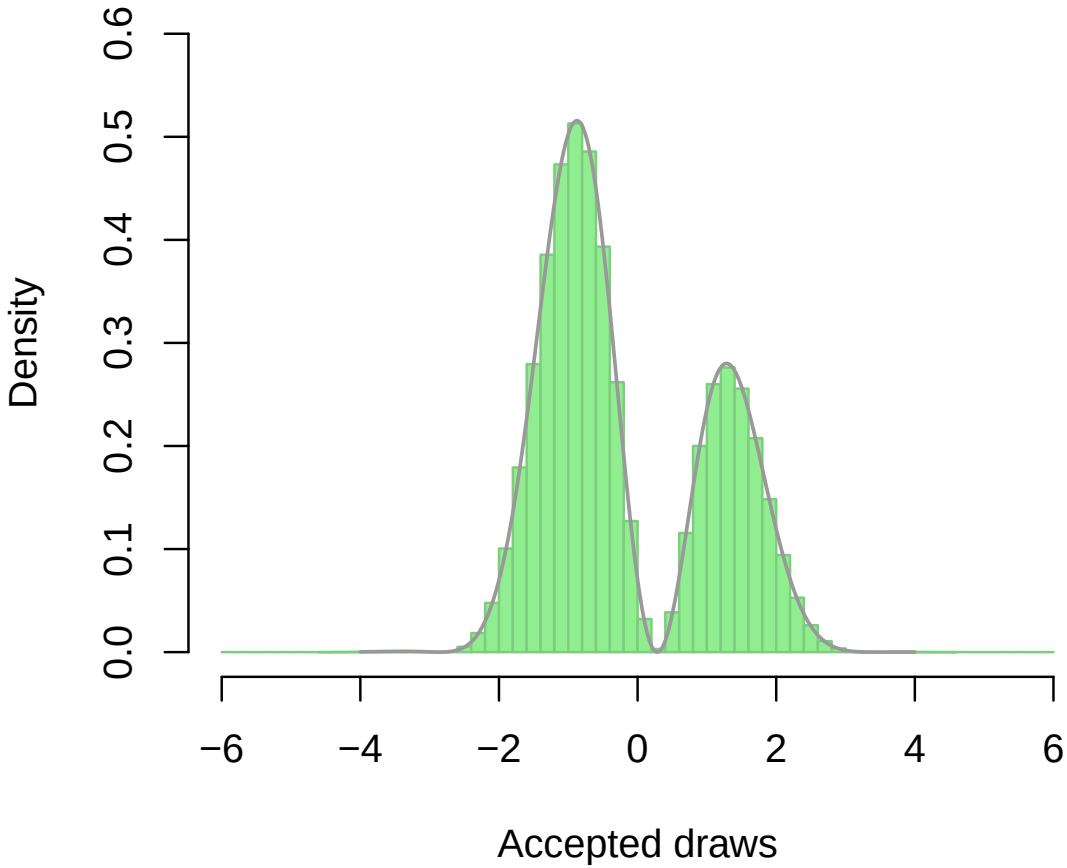


Limitation: I cannot guarantee $M \times g(\theta) > h(\theta)$ *outside* $[-4, 4]$.

Example 5.5: Algorithm under normal proposal

```
## Set seed for reproducibility
set.seed(6342)

## Updated draws
candidates <- rnorm(R, -0.25, 1.4)

## Scale accordingly
M <- 2.3

## Algorithm
num <- h(candidates)
denom <- M * dnorm(candidates, -0.25, 1.4)
prob <- num/denom
accept <- as.numeric(runif(R, 0, 1) < prob)
sample <- candidates[accept == 1]

## What proportion of draws were accepted?
> mean(accept)
[1] 0.482505
```

Example 5.5: Normal proposal



Was it worth it? Probably!

Rejection sampling

Other points:

- Rejection sampling can be inefficient, especially when the ratio of the target distribution to the proposal is close to zero for much of the support.
- Rejection sampling requires a known bounding function, which wasn't so hard to find **in one dimension**, but you can imagine this would be much more challenging for multi-dimensional parameters.
- The proposals are all independent. However, an accepted proposal could be used as an opportunity to explore "nearby" values, which is a motivation for the sampler we'll pivot to next.

Table of Contents

Metropolis-Hastings

Idea:

- Suppose you have specified a likelihood and a prior for θ.
- You want to generate random draws from the implied posterior.
- The Metropolis-Hastings algorithm produces random draws from the posterior.
- Let's base our discussion on an example in which we could use our understanding of conjugate priors to "check" our work:
  - $X_1, \ldots, X_{10} \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\theta)$; $\theta > 0$ unknown, to be estimated.
  - Prior: $\theta \sim \text{Exponential}(\lambda = 0.5) \equiv \text{Gamma}(\alpha = 1, \lambda = 0.5)$.
  - Observed sufficient statistic: $t = \sum_{i=1}^{10} x_i = 30$.
- Now, $\theta|\mathbf{X} = \mathbf{x} \sim \text{Gamma}(\alpha^* = 31, \beta^* = 10.5)$, which can be determined from our knowledge of conjugate priors that, but let's see if the algorithm could learn this for us.

Example 5.6: Metropolis-Hastings algorithm (Big picture)

- General idea of the Metropolis-Hastings algorithm:
  1. Initialize a posterior draw as $\theta^{(0)}$ (it doesn't have to be *great*, but it shouldn't be *awful*—consider drawing from the prior).
  2. Starting from a posterior draw, $\theta^{(j)}$, take a random (but structured, in a way we will discuss) step away to a "nearby" value of θ, call it $\theta_{\text{prop}}^{(j+1)}$.
  3. With stochasticity (we will discuss), evaluate $\theta_{\text{prop}}^{(j+1)}$ as a candidate to include as a valid posterior draw.
  4. If algorithm says "yes" in Step 3, that new draw can be added to your collection of posterior draws **and is your new basis to go back to Step 2 and repeat the process** (that is $\theta^{(j+1)} = \theta_{\text{prop}}^{(j+1)}$). If "no," go back to Step 2 and try again (i.e., from $\theta^{(j)}$, the last accepted draw).
  5. Iterate Steps 2-4 until there are enough samples from the apparent stationary distribution.
- Under some assumptions, the stationary distribution is $\pi(\theta|\mathbf{X} = \mathbf{x})$.

Example 5.6: Elaboration on Step 2 in context of example

- Reminder of example:
  - $X_1, \ldots, X_{10} \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\theta)$; $\theta > 0$ unknown, to be estimated.
  - Prior: $\theta \sim \text{Exponential}(\lambda = 0.5)$.
  - Observed sufficient statistic: $t = \sum_{i=1}^{10} x_i = 30$.
- Starting from $\theta^{(j)}$, consider taking a random draw, $\theta_{\text{prop}}^{(j+1)}$ from the following proposal distribution:

$$\theta_{\text{prop}}^{(j+1)} \sim \text{Gamma}(\theta^{(j)}/\omega, 1/\omega).$$

- The distribution for the draw has mean at the previous accepted proposal; $\theta^{(j)}$; ω controls the width of the search window (variance).
  - $\text{E}[\theta_{\text{prop}}^{(j+1)}] = \theta^{(j)}$ and $\text{Var}[\theta_{\text{prop}}^{(j+1)}] = \omega\theta^{(j)}$.
- Why are my proposals of the same family as my prior?
  - The support is the same; this is convenient!

Example 5.6: Elaboration on Step 2 in context of example

- The value of ω controls how close to $\theta^{(j)}$ you'd like your subsequent proposal to be, inviting the following trade-off:
  - Larger window: search for new draws in unexplored areas. Too large, and the algorithm may fail to converge to a stationary distribution in the number of iterations you've specified.
    - ⋆ In our example, a higher value of ω casts a wider net as compared to smaller values, though the algorithm does not *require* this.
  - Narrower window: refine search closer to the areas where good candidates have been identified. Too narrow, and the algorithm may not explore the support of the target density in a reasonable amount of time.
- The value of ω does not update across iterations.
- You may already be developing some feeling for "what can go wrong."

Example 5.6: Elaboration on Step 3 in context of example

- Once we have our candidate, $\theta_{\text{prop}}^{(j+1)}$, we have to decide whether to accept it or to say "thanks anyway, but I'm going to try again." We compare it to our previously accepted step, $\theta^{(j)}$, in the following way:

$$Q = \frac{p_{\mathbf{X}}(\mathbf{x}; \theta_{\text{prop}}^{(j+1)})}{p_{\mathbf{X}}(\mathbf{x}; \theta^{(j)})} \cdot \frac{\pi(\theta_{\text{prop}}^{(j+1)})}{\pi(\theta^{(j)})} \cdot \frac{g_\omega(\theta^{(j)} | \theta_{\text{prop}}^{(j+1)})}{g_\omega(\theta_{\text{prop}}^{(j+1)} | \theta^{(j)})},$$

  where $g_\omega(\theta_{\text{prop}}^{(j+1)} | \theta^{(j)})$ evaluates the proposal density of $\theta_{\text{prop}}^{(j+1)}$ assuming parameters $\theta^{(j)}$ (and vice versa).
- Recognize the first two ratios as comparing the *posterior* evaluated at $\theta_{\text{prop}}^{(j+1)}$ (numerator) and $\theta^{(j)}$ (denominator).
- Let $U \sim \text{Uniform}(0, 1)$ denote an iteration-specific random draw. If $Q > U$, we accept the candidate, and $\theta^{(j+1)} = \theta_{\text{prop}}^{(j+1)}$. Otherwise, go back to Step 2 and try generating another proposal.
  - This is the same as saying "accept proposal with probability $\min(Q, 1)$."

Example 5.6: That's it!

- That's the Metropolis-Hastings algorithm!
- It is not in any way "obvious" that this algorithm would have the property of converging to the posterior as a stationary distribution (the technical details are not in the scope of this course).
- However, we can gain some intuition for the value of Q.
  - First note that—all else being equal—higher values of Q are more likely to be accepted as compared to lower values. Q is higher when...
    - ★ $\theta_{\text{prop}}^{(j+1)}$ is more consistent with the likelihood as compared to $\theta^{(j)}$.
    - ★ $\theta_{\text{prop}}^{(j+1)}$ is more consistent with the prior as compared to $\theta^{(j)}$.
    - ★ $\theta_{\text{prop}}^{(j+1)}$ is *less* consistent with the proposal as compared to $\theta^{(j)}$.
- Intuitively, this procedure encourages the acceptance of draws that are more consistent with your model and prior belief, but can also be designed to encourage exploration of under-explored areas.
- If the proposal distribution is symmetric, the last fraction is one (Metropolis-Hastings $\rightarrow$ Metropolis).

Example 5.6: Reminder of setup

- $X_1, \ldots, X_{10} \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\theta)$; $\theta > 0$ unknown, to be estimated.
- Prior: $\theta \sim \text{Exponential}(\lambda = 0.5)$.
- Observed sufficient statistic: $t = \sum_{i=1}^{10} x_i = 30$.
- Let's code it up and you'll see it's not too bad!

Example 5.6: Coding the prior (easy peasy)!

```
## Prior
pr <- function(theta, lambda) {
  d.theta <- dexp(theta, rate = lambda)
  out <- log(d.theta)
  return(out)
}
```

Please note that use of the log-density has better computational stability.

Example 5.6: Setting the proposal

- Now, starting from $\theta^{(j)}$, we want to have a suitable window in which to propose our next draw:

$$\theta_{\text{prop}}^{(j+1)} \quad \sim \quad \text{Gamma}(\theta^{(j)}/\omega, 1/\omega).$$

- Consider $\omega = 0.1$.
  - Modifying window will affect the properties of the algorithm.
- The most essential part is to center the draws to have mean $\theta^{(j)}$.
- That ω is directly proportional to the variance is nice, but not always straightforward to achieve (nor is it essential). However, you must understand how the windows "control" the variance.

Example 5.6: Coding the proposal (also easy peasy)!

```
## Proposal
proposal <- function(theta, omega) {
  prop <- rgamma(1, shape = theta/omega, rate = 1/omega)
  return(prop)
}
```

Example 5.6: Coding the log-likelihood (again, easy peasy)!

```
## Log-likelihood
loglik <- function(theta, t = 30, n = 10) {
  out <- log(dpois(t, lambda = n*theta))
  return(out)
}
```

I am coding this using the **sufficient statistic**, $T \sim \text{Poisson}(n\theta)$, because I have not given you the actual values of the observations (nor do they matter). Otherwise, I might have used the following code:

```
## Log-likelihood
loglik <- function(theta, x) {
  out <- sum(log(dpois(x, lambda = theta)))
  return(out)
}
```

Example 5.6: Coding the proposal ratio (believe it or not, easy peasy)!

```
## Proposal ratio
prop.ratio <- function(theta.j, theta.proposed, omega) {
  d.1 <- dgamma(theta.j, shape = theta.proposed/omega, rate = 1/omega)
  d.2 <- dgamma(theta.proposed, shape = theta.j/omega, rate = 1/omega)
  out <- log(d.1) - log(d.2)
  return(out)
}
```

Example 5.6: Setting up MCMC ($\sum_k (\text{easy peasy})_k = \text{easy peasy}^*$)!

```
## Set seed for reproducibility
set.seed(6342)

## Track number accepted/rejected
accepted <- 0
rejected <- 0

## Prior hyperparameter
lambda <- 0.5

## Initialize based on a random draw from prior
theta.j <- rexp(1, rate = lambda)

## Number of MCMC iterations
M = 500000

## Search window
omega <- 0.1

## Store accepted thetas
accepted.thetas <- matrix(0, nrow = M, ncol = 1)
```

* Forgive the truly abysmal logic. I just want you to believe that this isn't so bad!

Example 5.6: Running MCMC

```
## Run MCMC
for (j in 1:M) {
  theta.prop <- proposal(theta.j, omega = omega)
  ll.th.prop <- loglik(theta.prop)
  ll.th.j <- loglik(theta.j)
  prior.prop <- pr(theta.prop, lambda = lambda)
  prior.theta.j <- pr(theta.j, lambda = lambda)
  p.ratio <- prop.ratio(theta.j, theta.prop, omega = omega)
  logQ <- ll.th.prop - ll.th.j + prior.prop - prior.theta.j + p.ratio
  U <- runif(1)
  if (exp(logQ) > U) {
    accepted <- accepted + 1
    accepted.thetas[accepted, 1] <- theta.prop
    theta.j <- theta.prop
  }
  if (exp(logQ) <= U) {rejected <- rejected + 1}
  if (round(j/25000) == (j/25000)) {print(paste(j, "iterations!"))}
}
```

Example 5.6: Algorithm results

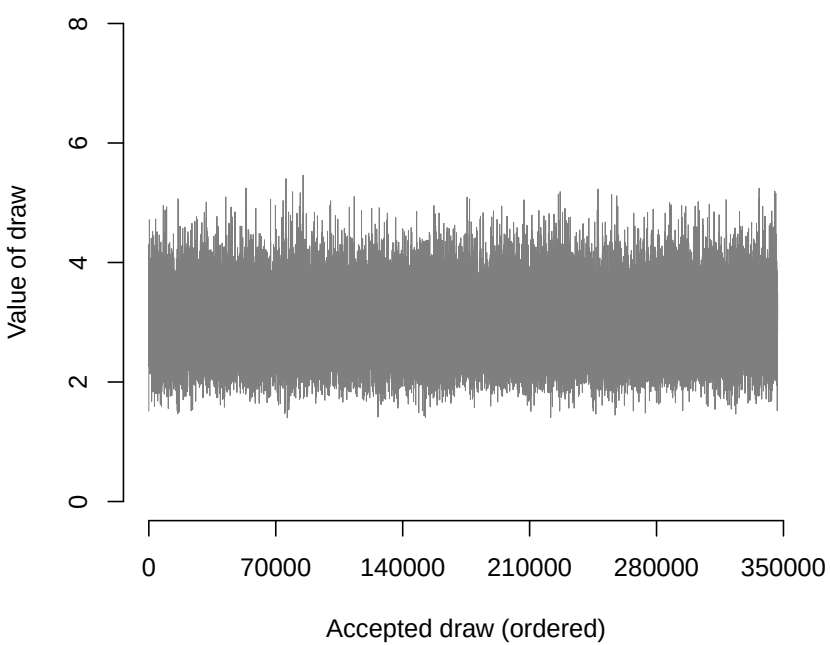
```
## Number of accepted draws
> accepted
[1] 346773

## Number of rejected draws
> rejected
[1] 153227

## Don't forget that you have more storage than necessary
accepted.thetas <- accepted.thetas[1:accepted,1]
```

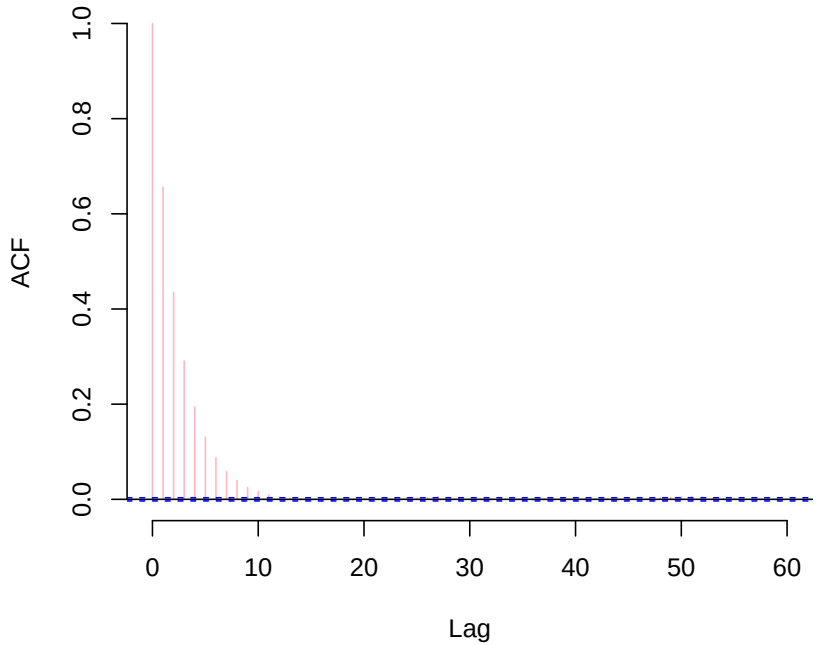
Metropolis-Hastings

Diagnostics:

- I know we're excited to see the results!
- With computational techniques, we have to do our due diligence.
- Two popular methods:
  - Trace plots: Effectively a line graph of the accepted draws, in the order which they were accepted.
    - ★ Goal: "Fuzzy caterpillar."
  - Autocorrelation plots (`acf()`): Correlation between adjacent pairs, pairs separated by one, two, three, etc.
    - ★ Goal: Fairly rapid descent toward zero.
- Another rule of thumb for low-dimensional problems is to aim for 30% to 60% acceptance. By this metric, we overshot a touch. Again, though, it's a rule of thumb, and not a hard rule.

Example 5.6: Trace and autocorrelation plots

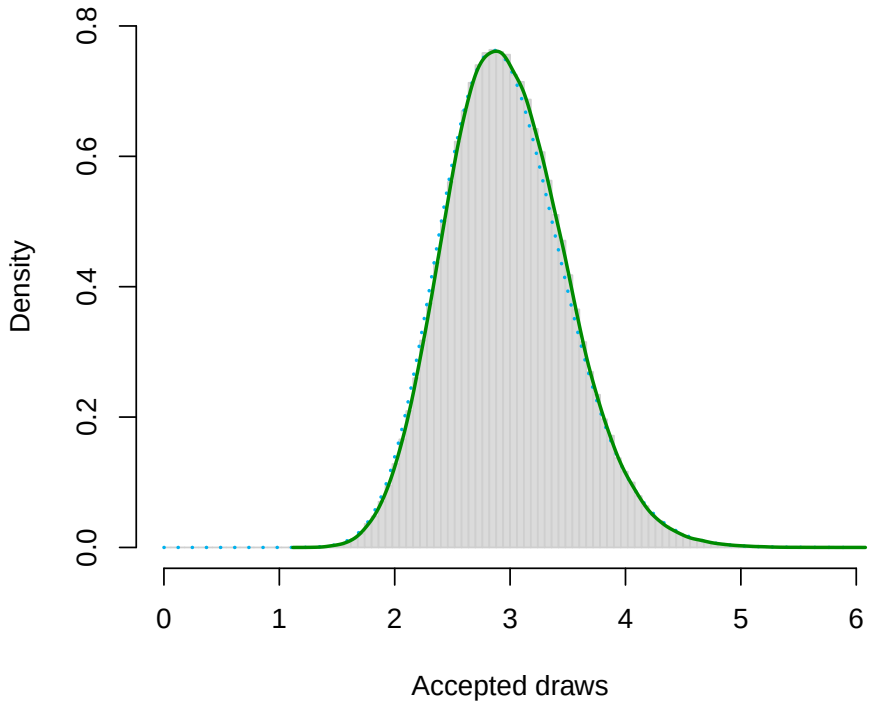




Example 5.6: Burn-in periods

- Particularly in more complicated models, it may take many iterations (hundreds? thousands?) for the algorithm to stabilize toward draws from the posterior.
- For these reasons, the first B draws are sometimes discarded; we call this subjectively-defined warm-up the *burn-in* period.
- With a simple problem such as this one, there is no apparent need to do this (so we won't).

Example 5.6: Histogram of accepted draws (posterior)



Kernel density smoother in solid green; true density in dashed blue.

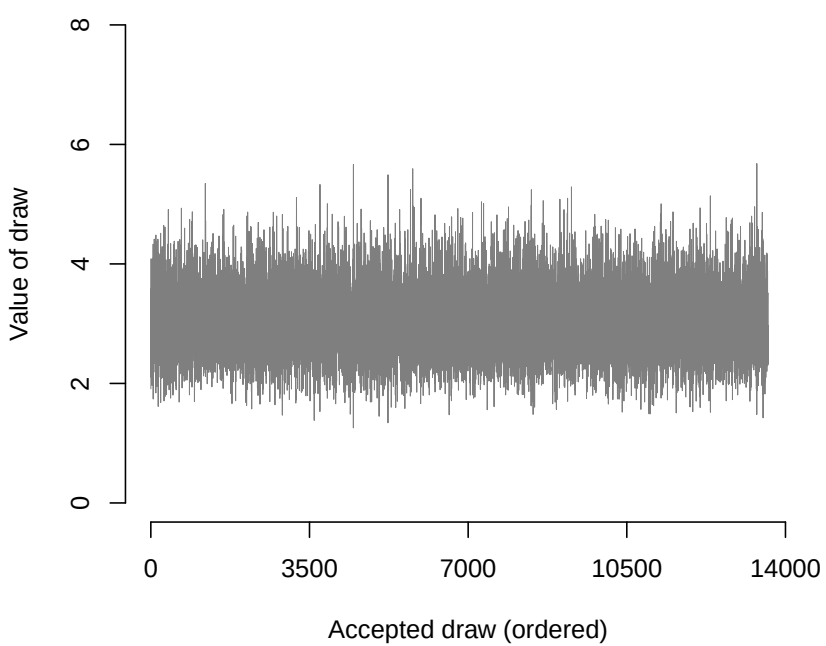
Example 5.6: Results

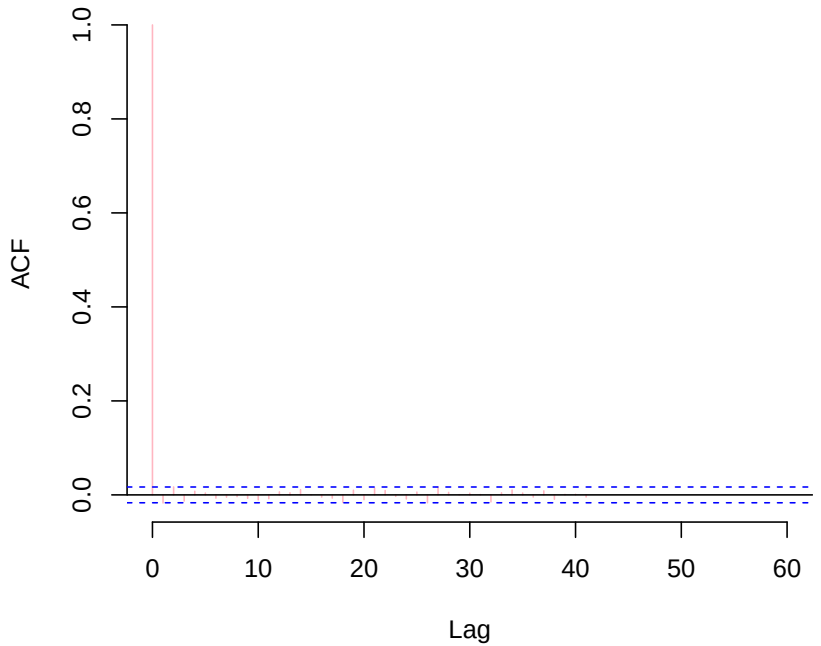
```
## Estimated
> mean(accepted.thetas)
[1] 2.972203

> sd(accepted.thetas)
[1] 0.5196276

## Compare to truth (easy to determine: conjugate prior)
> 31/10.5
[1] 2.9524

> sqrt(31/10.5^2)
[1] 0.53026
```

Example 5.6: Altering ω

- If ω is too *large*, then we are too liberal in exploring the parameter space and are prone to rejecting too many draws.
- If ω is too *small*, then we are lingering around the area too long and have potentially too much redundancy.
- The following two slides illustrate what happens when ω is too large or too small.
- There are ways to measure the "effective number of iterations." This also affects the Monte Carlo standard error. This is a topic for another course—just be aware that these are considerations that need to be addressed in practice.

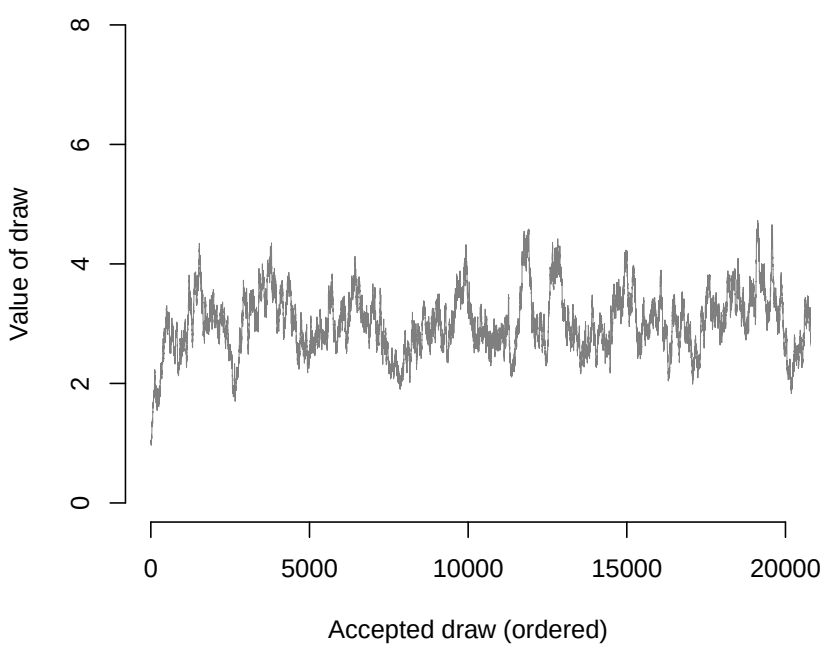
Example 5.6: Choosing $\omega = 50$

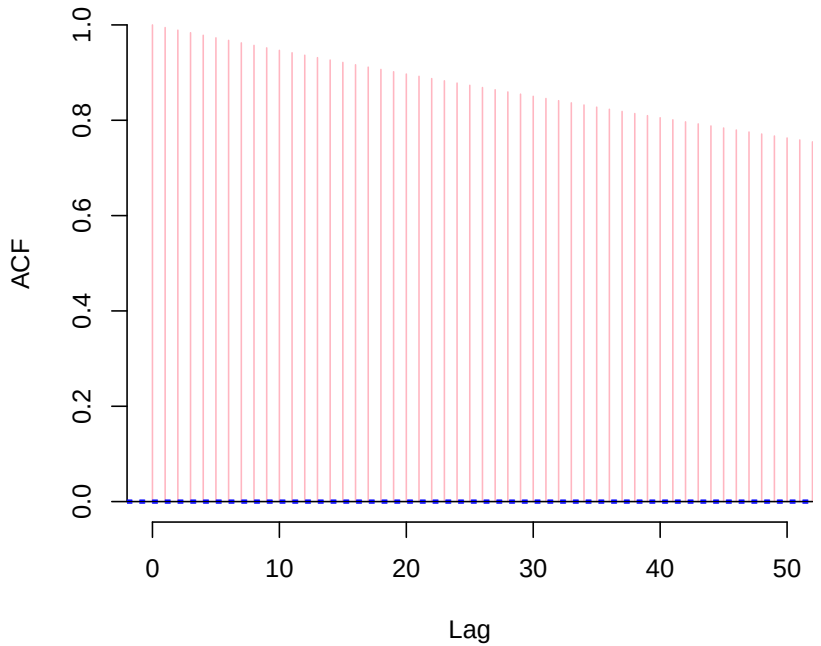




Way too few draws are accepted!

Example 5.6: Choosing $\omega = 0.001$





Only first 20,000 draws shown to illustrate the high autocorrelation.

Example 5.6: Multiple chains

- In practice, we initialize the Metropolis-Hasting algorithm under multiple draws and assess whether the stationary distributions appear to overlap.
  - This is referred to as "good mixing."
- In this course, we will not do examples that are sufficiently complicated to warrant this. However, I want you to be aware that this better reflects the *practice* of Bayesian analysis.

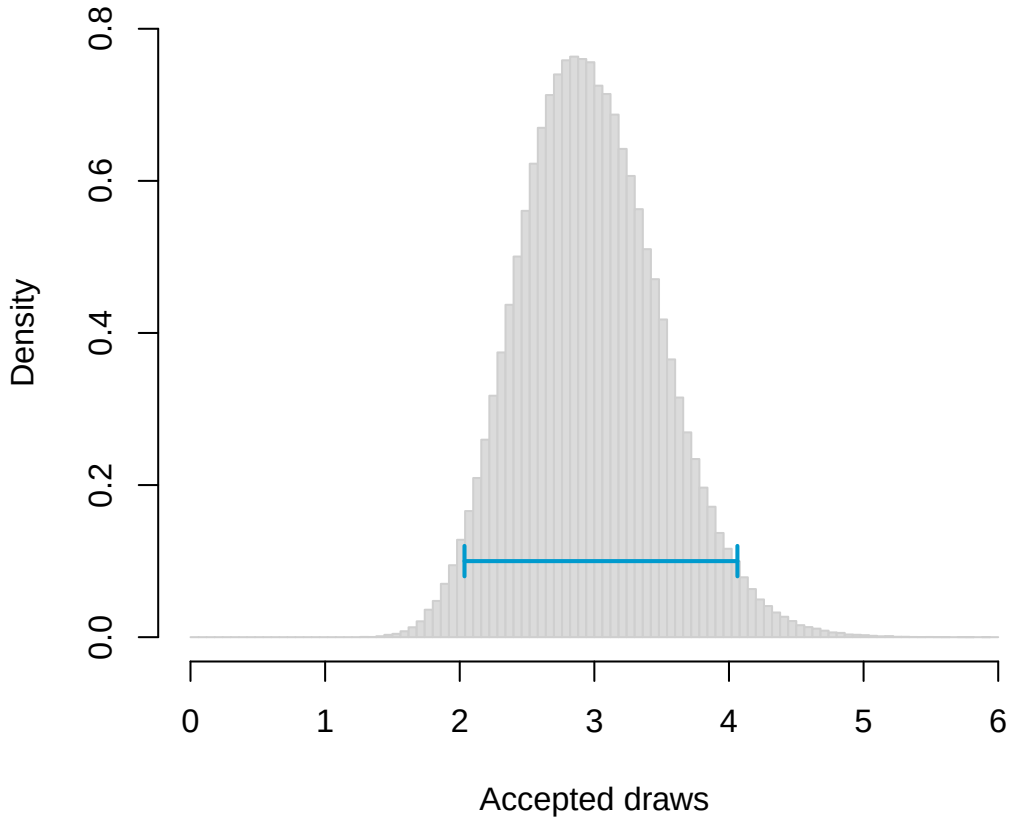
Example 5.6: Interval estimation

- So, with all that out of the way, let's stick with our original choice of a window ($\omega = 0.1$).
- Let's form a quantile-based and HPD-based interval.
- Because the posterior distribution is reasonably approximated by a normal distribution, we expect these intervals to be similar in this example.

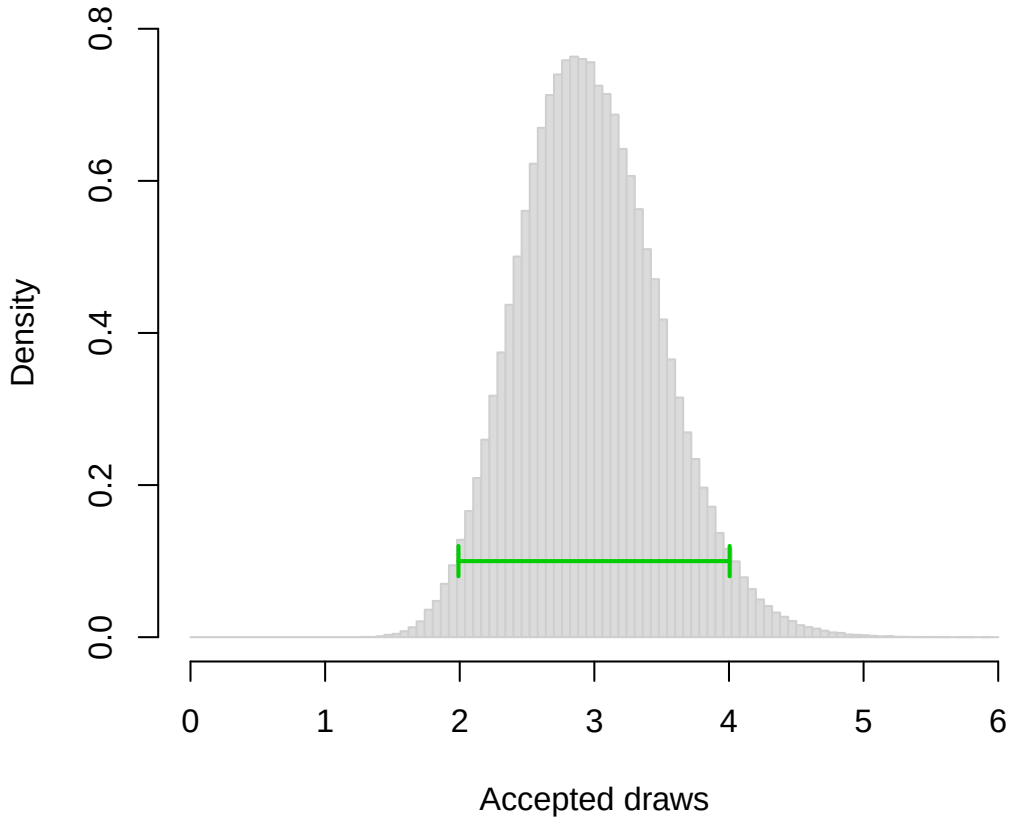
Example 5.6: Results

```
## Quantile-based credible interval
as.numeric(quantile(accepted.thetas, c(0.025, 0.975)))
[1] 2.035303 4.062816

## HPD-based credible ineterval
as.numeric(HPDint(accepted.thetas, 0.95))
[1] 1.990439 4.004968
```

Example 5.6: Quantile-based 95% credible interval



Example 5.6: HPD-based 95% credible interval



Example 5.6: Interpretation

- Let's take the HPD-based interval.
- We may say, as a Bayesian, that there is a 95% probability that θ lies in the interval $[1.99, 4.00]$.
- Reminder: The word "probability" is not being used the same way here as it is in the frequentist sense. That is, we are not thinking of this as a long-run frequency with respect to some sampling mechanism.

Example 5.6: Other uses

- We can use the draws to learn about other quantities related to the posterior:
  - Example: $\pi(\theta > 2|\mathbf{X} = \mathbf{x})$.
  - In R: `mean(accepted.thetas > 2)`.
- We can also learn about the posterior for transformations of the parameter by transforming the draws:
  - Example: If $\phi = \exp(\theta)$, what is $\pi(\phi|\mathbf{X} = \mathbf{x})$?
  - In R: `hist(exp(accepted.thetas))`.
- In some ways, having the draws makes Bayesian inference easier!
  - Less math...

Table of Contents

Predictive distributions

Prior predictive distribution:

- When you've specified a prior and a likelihood, you can aggregate that information to tell you the values of X you believe to be plausible before ever having collected any data.
- Prior predictive distribution:

$$p_{\tilde{X}}(x) = \int_{\Theta} dP_{X,\Theta}(x, \theta) = \int_{\Theta} p_X(x; \theta) d\Pi_{\Theta}(\theta).$$

- Real world: prior reflects knowledge about X more than about θ. We rely on information from the prior predictive distribution to tell us that our prior is sensible.
  - Example: If prior predictive distribution produces SBP values of -40 mm Hg with high frequency, our prior might be off.
- Regarding the argument against Bayesian "I want to be completely agnostic about θ," consider this slide a *rebuttal*. You usually do have *some* information even if only through the nature of the data.

Example 5.7: Bernoulli and Beta

- Suppose $X \sim \text{Bernoulli}(\theta)$, and we utilize the prior $\theta \sim \text{Beta}(\alpha, \beta)$.
- The prior predictive distribution is straightforward to determine:

$$\begin{aligned} p_{\widetilde{X}}(x) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_{\theta=0}^{\theta=1} \theta^x (1-\theta)^{1-x} \theta^{\alpha-1} (1-\theta)^{\beta-1} d\theta \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_{\theta=0}^{\theta=1} \theta^{x+\alpha-1} (1-\theta)^{\beta-x} d\theta \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \frac{\Gamma(x+\alpha)\Gamma(\beta+1-x)}{\Gamma(\alpha+\beta+1)} \\ &= \frac{\alpha^x \beta^{1-x}}{\alpha+\beta} = \left(\frac{\alpha}{\alpha+\beta}\right)^x \left(\frac{\beta}{\alpha+\beta}\right)^{1-x}. \end{aligned}$$

- Prior predictive: $\widetilde{X} \sim \text{Bernoulli}(\alpha/(\alpha+\beta))$. We often use the "$\widetilde{\text{tilde}}$" notation to mark a predictive distribution.

Example 5.8: Poisson and Gamma

- Suppose $X \sim \text{Poisson}(\theta)$, and we utilize the prior $\theta \sim \text{Gamma}(\alpha, \beta)$.
- It happens that $\widetilde{X} \sim \text{NegativeBinomial}(\alpha, \beta/(1+\beta))$.
  - It's just some algebra/calculus, but it's not intellectually stimulating.
- Note that the negative binomial distribution does not require the "k" parameter to be an integer (though indeed the interpretation is nicer when it is).

Predictive distributions

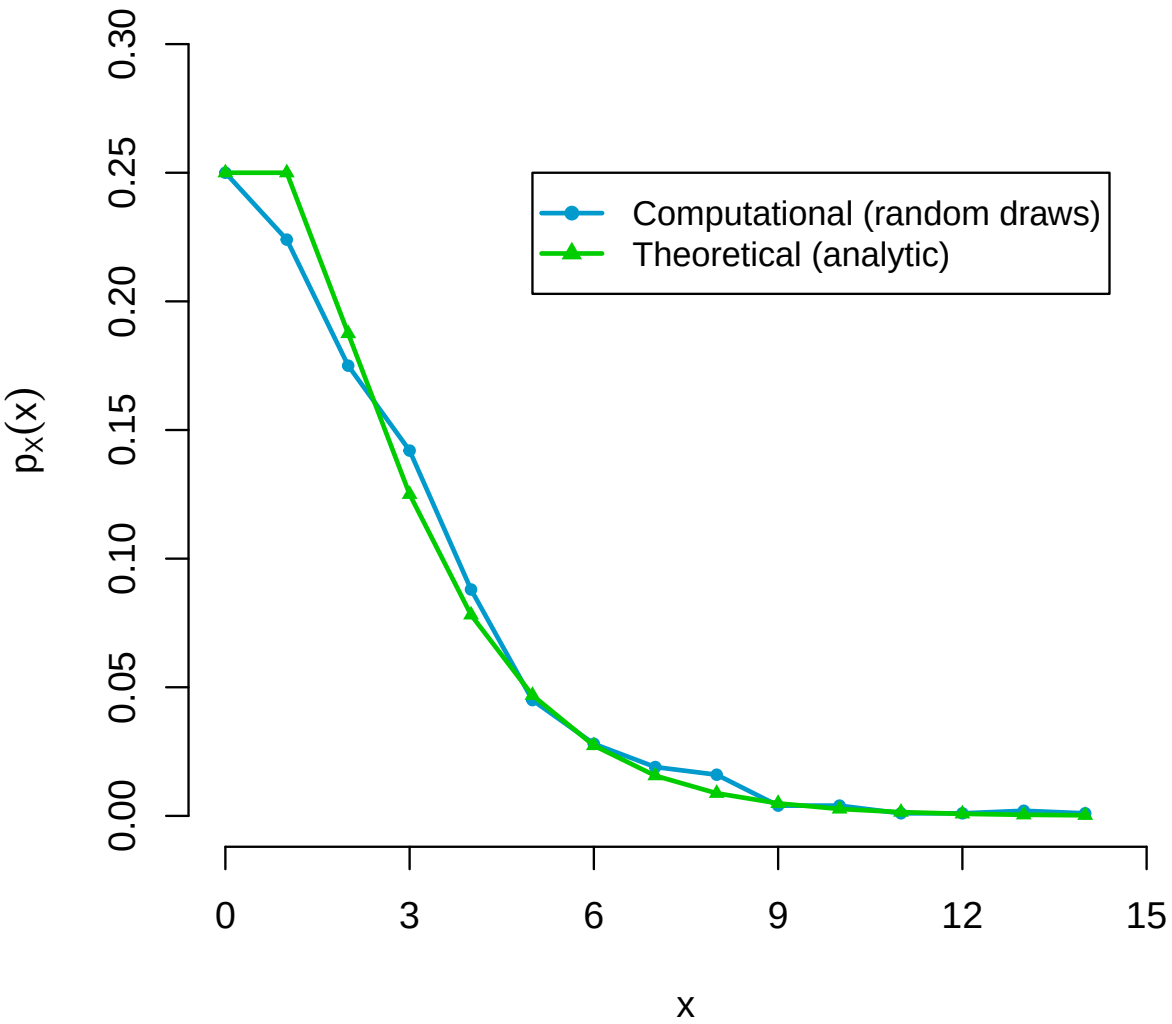
Prior predictive distribution:

- Evaluating the integral analytically can sometimes be tough.
- However, generating draws from the prior predictive distribution *computationally* is easier than it might look. For some large enough number of samples, M:
  1. Generate a random draw from the prior, θ_m.
  2. Generate a draw, $\widetilde{x}_m$, from the likelihood $p_X(x; \theta_m)$.
- The draws $\widetilde{x}_1, \ldots, \widetilde{x}_M$ constitute draws from the prior predictive distribution.

Example 5.9: Prior predictive (computational)

```r
## Set seed for reproducibility
set.seed(6342)

## Set values of parameters
alpha <- 2
beta <- 1

## Random draw approach
M <- 1000
theta.draws <- rgamma(M, shape = alpha, rate = beta)
x.d <- rpois(M, theta.draws)

## Analytic approach
x.a <- dnbinom(seq(0, 11, 1), size = alpha, prob = beta/(1 + beta))
```

Predictive distributions



Approximation even better with larger number of draws.

Predictive distributions

Posterior predictive distribution:

- As you might have guessed, there is also a *posterior* predictive distribution:

$$p_{\tilde{X}|\mathbf{X}}(x|\mathbf{x}) = \int_{\Theta} p_X(x;\theta) d\Pi_{\Theta|\mathbf{X}}(\theta, \mathbf{x}).$$

- In practice, the posterior predictive distribution is used after collecting data to ascertain that it resembles the observed data.
  - Not looking for an exact match.
- A posterior predictive distribution that wildly differs from the observed data suggests a poor fit.

Predictive distributions

Posterior predictive distribution: Bernoulli/Beta and Poisson/Gamma

- When using a conjugate prior, recall that the prior and posterior are in the same family. In such settings, the prior predictive and the posterior predictive distributions share a conjugate relationship.
- When we've "already done the math" for the prior predictive distributions, we can jump immediately to the analogous conclusions for the following posterior predictive distributions.
- Example: Bernoulli(θ) likelihood/Beta(α, β) prior:
  - $\widetilde{X}|\mathbf{X} = \mathbf{x} \sim \text{Bernoulli}\big((\alpha + \sum_{i=1}^{n} x_i)/(\alpha + \beta + n)\big)$.
- Example: Poisson(θ) likelihood/Gamma(α, β) prior:
  - $\widetilde{X}|\mathbf{X} = \mathbf{x} \sim \text{NegativeBinomial}\big(\alpha + \sum_{i=1}^{n} x_i, (\beta + n)/(\beta + n + 1)\big)$.
- Note how the posterior hyperparameters make their appearance in these examples analogously to the way the prior hyperparameters do in the prior predictive distribution. This is characteristic of "conjugate prior" math.

Posterior predictive distribution:

- For settings in which the integral is not tractable, generating draws is again easier than it might appear. For some large enough number of samples, M:
  1. Generate a random draw from the posterior, $\widetilde{\theta}_m$.
     - ⋆ This can come from knowledge of the form of the posterior, or it can even come from random selection of a posterior draw if using a computational approach such as Metropolis-Hastings!
  2. Generate a draw, $\widetilde{x}_m$, from the likelihood $p_X(x; \widetilde{\theta}_m)$.
- The draws $\widetilde{x}_1, \ldots, \widetilde{x}_M$ constitute draws from the posterior predictive distribution.

Table of Contents

Scaling and improper priors

Prior specification: The scaling problem

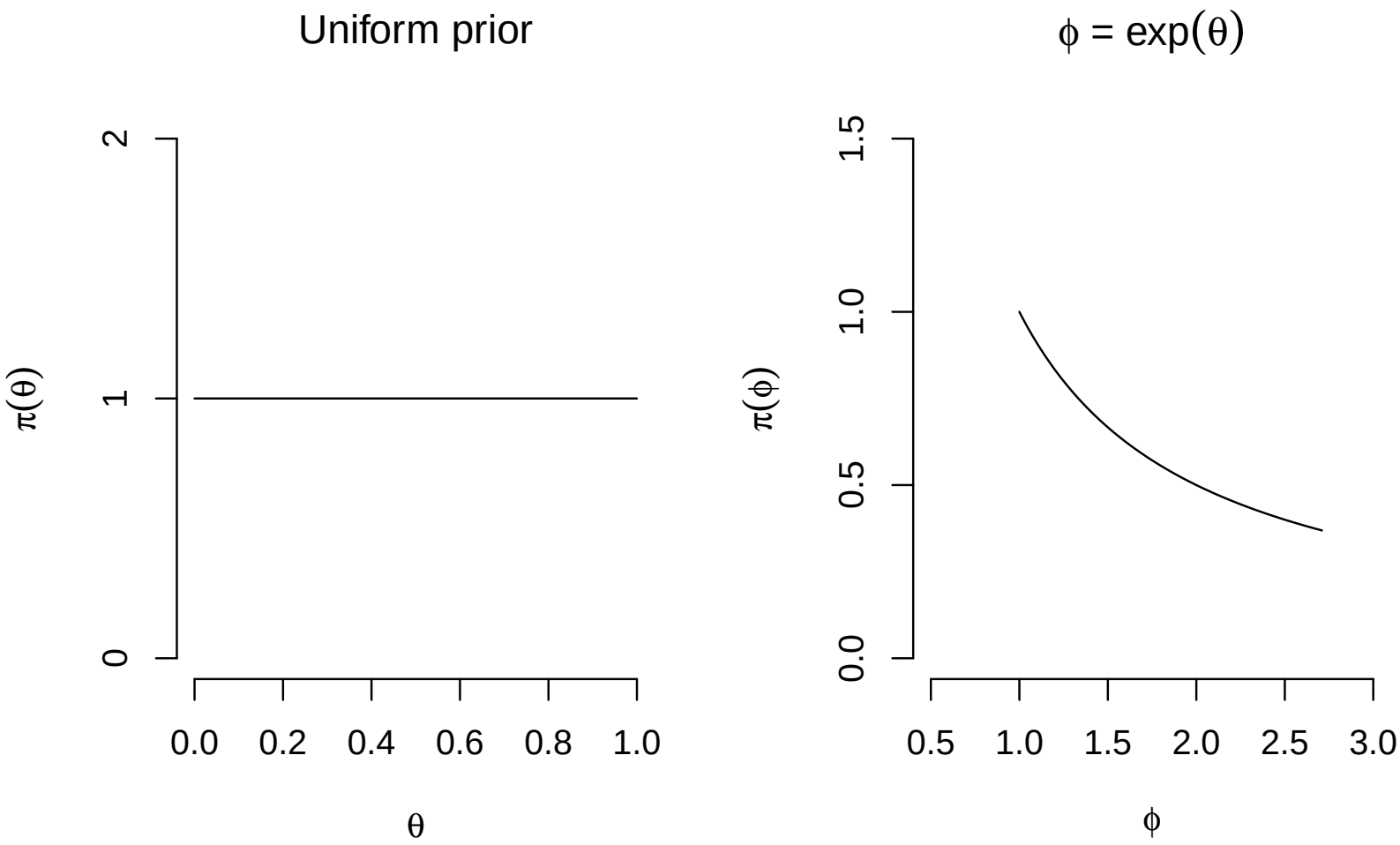
- Suppose I posit a prior, $\pi_\theta(\theta)$, for θ.
- Suppose $\phi = g(\theta)$ is the parameter of interest (assume $g(\cdot)$ is strictly monotone and differentiable for simplicity). Then,

$$\pi_\phi(\phi) \quad = \quad \pi_\theta(g^{-1}(\phi)) \left| \frac{d}{d\phi} g^{-1}(\phi) \right|.$$

- The implication here is that the parameterization of the model matters as it pertains to inference on ϕ.
- Example: Suppose $\Theta = [0, 1]$; consider a uniform prior on θ, and that $\phi = \exp(\theta)$. Then, $g^{-1}(\phi) = \log(\phi)$, and

$$\pi_\phi(\phi) \quad = \quad \frac{d}{d\phi} g^{-1}(\phi) = \frac{1}{\phi} I_{[1,e]}(\phi).$$

Prior specification: Scale-dependence



Scaling and improper priors

Prior specification: The scaling problem

- Key point: "flatness" on one scale does not translate to "flatness" on another. This is part of the reason why we say there is really no such thing as a totally non-informative prior.
- Even still, a uniform prior on θ can't even reasonably be deemed non-informative on the scale of θ. It implies, for instance, equal prior belief in the statements $\theta \in (0, 0.5)$ and $\theta \in (0.5, 1)$.

Scaling and improper priors

Prior specification: Flat priors

- Let's continue, though, with the idea of a prior $\pi(\theta) \propto 1$, so that the posterior depends only upon the likelihood.
- Suppose now that Θ is not bounded. The prior does not actually constitute a density because it does not integrate to one.
  - We call it an *improper* prior.
- In some cases, an improper prior may actually nevertheless produce a *proper* posterior.
  - Example: Under the likelihood $X \sim \mathcal{N}(\theta, \sigma^2 = 1)$, the improper prior $\pi(\theta) \propto 1$ produces a *proper* posterior $\theta|X = x \sim \mathcal{N}(x, \sigma^2 = 1)$.
- We don't always get so lucky.
- The other side: a proper prior does not alone guarantee a proper posterior (sad trombone)...

Scaling and improper priors

The Jeffreys Prior:

- If one seeks to overcome scale-dependence, one can use the *Jeffreys* prior: $\pi(\theta) \propto_\theta \mathcal{I}^{1/2}(\theta)$. This choice of a prior turns out to be scale-invariant, meaning that the density/mass assigned to a region will be the same regardless of how you parameterize the model.
- As an example, the Bernoulli(θ) distribution can be parameterized by the odds, $\phi = \theta/(1-\theta)$. A uniform prior on θ is not the same as a uniform prior on ϕ, but the Jeffreys prior on one reparameterizes to the Jeffreys prior on the other.
  - That is, $\pi(\phi) \propto_\phi \mathcal{I}^{1/2}(\phi)$.
- Careful: Posterior resulting from Jeffreys prior may not be proper.

Example 5.10: Jeffreys prior for Poisson

- Suppose $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\theta)$.
- It is straightforward to show that $\mathcal{I}(\theta) = \theta^{-1}$.
- Therefore, the Jeffreys prior is given by $\pi(\theta) \propto_\theta \theta^{-1/2}$. This is not a proper prior.
- The posterior can be determined readily (let $t = \sum_{i=1}^n x_i$):

$$\pi(\theta|T = t) \propto_\theta \theta^t \exp(-n\theta)\theta^{-1/2} = \theta^{t-1/2} \exp(-n\theta).$$

- We recognize this as the kernel of a Gamma density. In particular, we conclude that $\theta|\mathbf{X} = \mathbf{x} \sim \text{Gamma}(\alpha^* = t + 1/2, \beta^* = n)$.

Reading

Corresponding sections of the textbook:

- C&B Section 7.2.3
  - Bayes estimators and conjugate priors.
- C&B is quite sparse on the Bayesian material; I have included far more detail in these notes.
- Further reading: Statistical Rethinking (McElreath)
- Further reading: Bayesian Data Analysis (Gelman *et al.*)

Summary: So far

- Sufficiency, ancillarity, and completeness.
- Point estimation.
- Large sample theory.
- Hypothesis testing and confidence intervals.
- Bayesian methods.

Summary: Coming up

- Foundations of decision theory.
- Model misspecification.
- Nonparametric methods.

Table of Contents

Enrichment: Some additional considerations

Marginal posteriors:

- If $\boldsymbol{\theta}$ is our K-dimensional multivariate parameter, but only θ_k is of interest, one can summarize the posterior distribution for θ_k only:

$$\pi(\theta_k|\mathbf{X}=\mathbf{x}) \quad = \quad \int_{\boldsymbol{\theta}_{-k}} \pi(\boldsymbol{\theta}|\mathbf{X}=\mathbf{x}) d\boldsymbol{\theta}_{-k},$$

  where $\boldsymbol{\theta}_{-k} = (\theta_1, \theta_2, \ldots, \theta_{k-1}, \theta_{k+1}, \theta_{k+2}, \ldots, \theta_K)$.
- For instance, a regression model has parameters $(\boldsymbol{\beta}, \sigma^2)$, but I may only be interested in β_1.
- The integral looks complicated, but this is unusually easy when sampling from the posterior (marginalization of most kinds of parameters amounts to simply ignoring the parameters over which we're marginalizing).

Enrichment: Some additional considerations

Prior specification: Substantive priors

- One possible procedure to form a more substantive prior is to base it upon previously collected data.
- If one believes that the data-generation mechanism for both sets of data to be identical (or at least sufficiently similar), then it would be logical to base the posterior on the combined data.
- Example: Seamless Phase-II/Phase-III randomized controlled trials.

Enrichment: Some additional considerations

Bayesian thinking:

- Quantifying uncertainty for any estimate derived as a function of the parameters is as simple as plugging in the posterior samples (we illustrated this).
- Because a Bayesian's concern is about $\boldsymbol{\theta}$ given (data), he or she does not have to navigate the landmines that a frequentist does when there are nuances in the sampling scheme.
  - Group-sequential designs.
  - Negative binomial or binomial sampling.
- A Bayesian can answer questions that a frequentist would have nearly no hope of knowing how to approach.
  - For instance: "what is the probability that treatment A improves at least three of five outcomes as compared to treatment B?"

Enrichment: Some additional considerations

Barriers to widespread implementation: At least by my judgment

- Reluctance to allow subjective belief to be a factor in an analysis.
  - Not a great argument. A frequentist incorporates subjectivity into an analysis as well but it's often difficult if not impossible to account for that subjectivity in the model.
- Computational complexity.
  - This is *less* of a barrier as our computing power increases, but prior specification is very much a hands-on approach and modern Bayesian implementation requires some level of computational sophistication, particularly for non-"cookie cutter" problems.
- Old dog/new tricks.
  - Frequentist methods have been taught as the default for decades and people are comfortable with them.
  - Or so they say. But I have to correct phrases such as "there was no association" more often than I would care to.

Enrichment: Some additional considerations

Bayesian implementation:

- Generally speaking, we do not hard-code our own samplers.
- Bread-and-butter Bayesian modeling: `rstanarm`/`brms`.
  - `https://cran.r-project.org/web/packages/rstanarm/index.html`
- CmdStan implements a more state-of-the-art approach based on Hamiltonian MCMC, which is quite a bit more complex than the Metropolis-Hastings algorithm.
- The tuning and diagnostic methods are specific to Hamiltonian MCMC; the skills we learned for Metropolis-Hastings are helpful but do not generalize perfectly.
- My understanding is that the Hamiltonian methods take "longer to run," but have less autocorrelation such that you achieve a greater effective sample size.

Enrichment: Some additional considerations

Nonparametric Bayesian:

- Our discussion has focused on parametric Bayesian methods.
- There are nonparametric Bayesian methods, in which model "size" is allowed to grow with sample size.
- Indeed, there are semi-parametric Bayesian models, in which only one part of the model "size" grows with n (e.g., proportional odds models).
- There isn't adequate time to cover these materials well in this course (nor am I particularly familiar with them), but I do want you to be aware that they exist.

Table of Contents

Enrichment: The Gibbs sampler

The algorithm:

- Another popular algorithm of choice is the Gibbs sampler.
- Suppose priors are specified in a way such that the conditional posterior, $\pi(\theta_k|\boldsymbol{\theta}_{-k}, \mathbf{X} = \mathbf{x})$, can be derived for each k (may not always be the case). The algorithm is as follows.
- Initialize some $\boldsymbol{\theta}^{(0)}$, and set $j = 0$. Then, repeat:
  - Generate $\theta_1^{(j+1)} \sim \pi(\theta_1|\theta_2^{(j)}, \ldots, \theta_K^{(j)}, \mathbf{X} = \mathbf{x})$.
  - Generate $\theta_2^{(j+1)} \sim \pi(\theta_2|\theta_1^{(j+1)}, \theta_3^{(j)}, \ldots, \theta_K^{(j)}, \mathbf{X} = \mathbf{x})$.
  - $\vdots$
  - Generate $\theta_{K-1}^{(j+1)} \sim \pi(\theta_{K-1}|\theta_1^{(j+1)}, \ldots, \theta_{K-2}^{(j+1)}, \theta_K^{(j)}, \mathbf{X} = \mathbf{x})$.
  - Generate $\theta_K^{(j+1)} \sim \pi(\theta_K|\theta_1^{(j+1)}, \ldots, \theta_{K-1}^{(j+1)}, \mathbf{X} = \mathbf{x})$.
  - Set $j = j + 1$.
- Under some assumptions, the stationary distribution is $\pi(\boldsymbol{\theta}|\mathbf{X} = \mathbf{x})$.

Enrichment: The Gibbs sampler

Considerations:

- Sometimes the conditional posteriors can be derived explicitly for some of the parameters but not the others. It is valid to mix and match algorithms (i.e., using Gibbs sampling for some parameters, but Metropolis-Hastings to generate a draw from posteriors for which you do not have an analytic form).
  - This is known as "adding a Metropolis step."
- If some of the parameters are too tightly correlated, it may be challenging to converge to a stationary distribution. In such cases, we can derive the conditional posterior for a collection of parameters given the others.
  - This is a strategy known as *blocking*.
- Were we to have an extra couple of weeks in the semester, I might be more inclined to include examples. :)

Table of Contents

Enrichment: The Jeffreys-Lindley paradox

Motivating example: Setup

- In a certain town, 50,310 males and 49,690 females are born over a five-year period.
- Let $\theta \in (0, 1)$ denote the population proportion of male births.
- We seek to test $H_0 : \theta = 0.5$ against $H_1 : \theta \neq 0.5$.
- The Jeffreys-Lindley paradox is a counterintuitive situation in which the frequentist and Bayesian paradigms produce (wildly) different conclusions about the same data, despite a prior that favors H_0 and H_1 equally.

Enrichment: The Jeffreys-Lindley paradox

Motivating example: Approaching as a frequentist

- Let $X \sim \text{Binomial}(n = 100{,}000,\ \theta)$ denote the number of male births.
- Central limit theorem $\Rightarrow X \dot{\sim} \mathcal{N}(\mu = 50{,}000, \sigma^2 = 25{,}000)$ under H_0.
- $\text{P}(X \geq 50{,}400 | X \leq 49{,}690) \approx 0.0499$.
- We reject $H_0 : \theta = 0.5$, concluding the proportion of male births is different from 0.5 (this is a *significance test*).

Enrichment: The Jeffreys-Lindley paradox

Motivating example: Approaching as a Bayesian

- If there is no reason to favor one hypothesis over the other, one could assign prior probabilites of $\pi(H_0) = \pi(H_1) = 0.5$, with a uniform prior on θ for $\theta \in (0.0, 0.5) \cup (0.5, 1.0)$. This is sometimes called a *spike and slab* prior.
- By Bayes' theorem:

$$\mathrm{P}(H_0|X = x) = \frac{\mathrm{P}(X = x|H_0)\pi(H_0)}{\mathrm{P}(X = x|H_0)\pi(H_0) + \mathrm{P}(X = x|H_1)\pi(H_1)}.$$

- $\mathrm{P}(X = 50{,}310|H_0) = \binom{100{,}000}{50{,}310}(0.5)^{50{,}310}(0.5)^{49{,}690} \approx 0.000369171.$
- Further (averaging over alternative hypotheses),

$$\mathrm{P}(X = 50{,}310|H_1) = \int_0^1 \binom{100{,}000}{50{,}310}\theta^{50{,}310}(1-\theta)^{49{,}690}d\theta \approx 0.000009999.$$

- From this, we conclude that $\mathrm{P}(H_0|X = 50{,}310) \approx 0.9736$.

Enrichment: The Jeffreys-Lindley paradox

Motivating example: Noting and reconciling the difference

- These results (seemingly) conflict because the frequentist approach seems to strongly support the conclusion of uneven birth probabilities, while the Bayesian approach seems to strongly support the null.
- However, there really is no paradox here, *per se*. The Bayesian approach aggregates a wide range of values of θ that were wildly inconsistent with the data and averaged over alternative hypotheses. The support for H_0 was *relative* to $\theta \in \Theta_1$—"most" of which were even *less* plausible.
  - Furthermore, assigning *half* of the prior weight to a single point $\theta_0 \in \Theta \equiv (0, 1)$ severely tips the scales in favor of H_0.
- Bayesian hypothesis testing is very controversial. Part of the advantage of Bayesian thinking is being able to make probabilistic statements about θ directly. Turning it into an indirect problem is regarded by many (including myself) as unnecessary.
- Under modern sensibilities, I think basing decisions on credible intervals and posterior probabilities makes far more sense.

Table of Contents

Enrichment: The likelihood principle

Motivating example: Bernoulli trials

- Suppose two successes and eight failures occur in a set of Bernoulli experiments. Is this information alone enough to conduct inference on the success probability?
- A frequentist might say, "no, not without more context regarding the sampling scheme?"
  - Did two successes occur in a fixed set of ten independent trials?
  - Were trials conducted until the first two successes occurred?
- A Bayesian, however, might say, "sure, so long as I have a prior on the unknown probability."

Enrichment: The likelihood principle

Proportional likelihoods:

- Binomial likelihood ($n = 10$ trials and $x = 2$ successes):

$$p_X(2;\theta) = \binom{10}{2}\theta^2(1-\theta)^8.$$

- Negative binomial likelihood ($r = 2$ successes and $x = 8$ failures):

$$p_X(8;\theta) = \binom{9}{8}\theta^2(1-\theta)^8.$$

- Under a fixed prior, the posteriors will be the same regardless of the likelihood.
- A frequentist would balk at this (but one adhering to strict application of the likelihood principle would not care).
- There are many illustrations of this problem. You can look this up if it interests you!