STA261 - Module 6

Bayesian Statistics

Rob Zimmerman

University of Toronto

August 6-8, 2024

(Long Spiel)

What is "p"?

Coffee cup lid flips: # H = 19
T = 4. Interesting!

The Bayesian Model

- So θ is now treated as a *random variable* with its own distribution expressing our beliefs

- The Bayesian framework for inference contains the statistical model $\{f_\theta : \theta \in \Theta\}$ and adds a **prior probability measure** $\Pi : \Theta \to [0,1]$ describing our beliefs about θ *before* we observe the data (like "$\mathbb{P}$", but the prior version)

- We usually refer to the prior by its pdf/pmf, which we denote generically as $\pi(\cdot)$

For example: $\pi(p) = \mathbb{1}_{p \in (0,1)} \iff \pi(p)$ is a Unif(0,1) prior on p

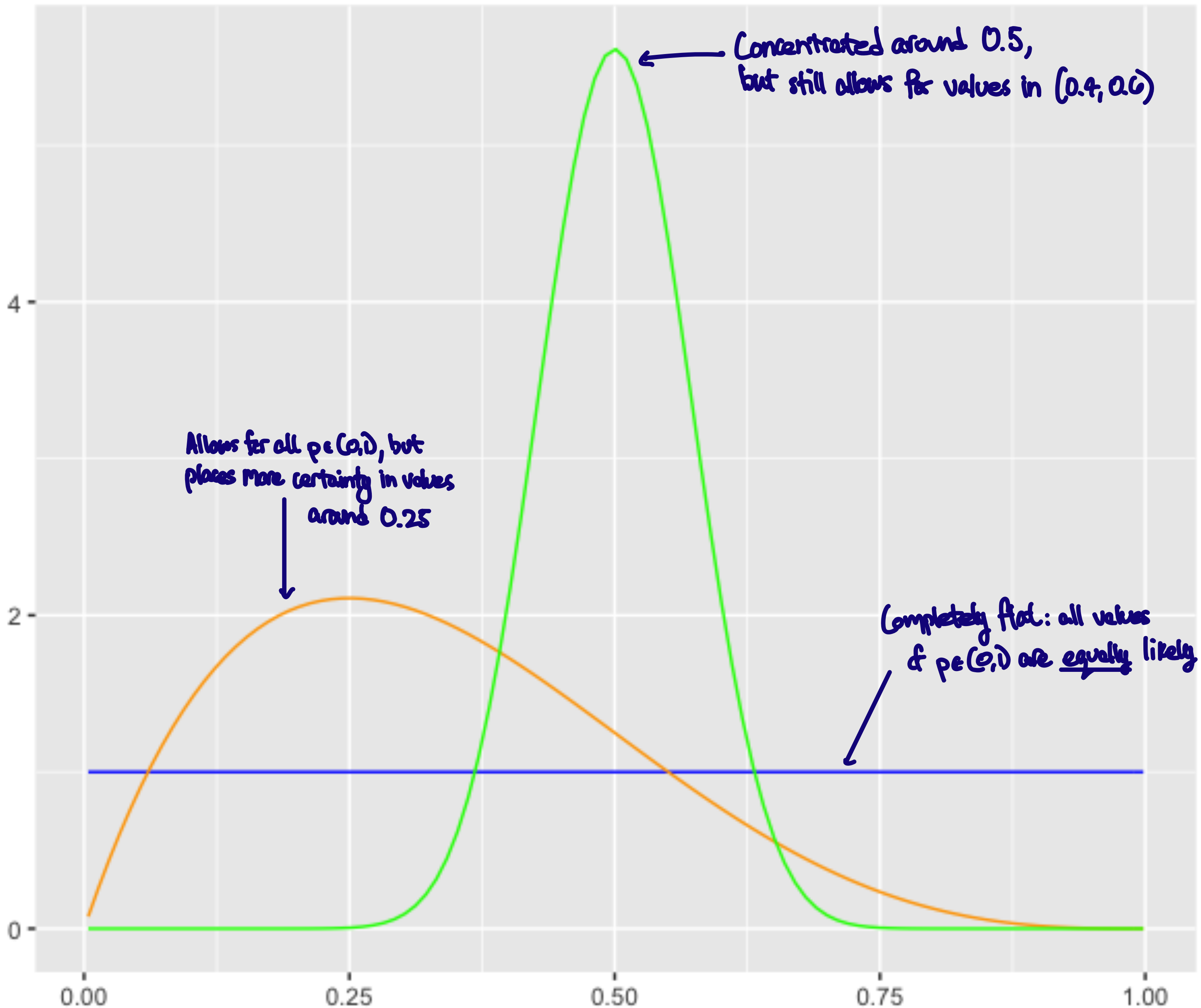
$\pi(\theta) = 3e^{-3\theta}, \ \theta > 0 \iff \pi(\theta)$ is an Exp(3) prior on θ

$\pi(\lambda) = \frac{1}{\sqrt{2\pi}} e^{-\lambda^2/2}, \ \lambda \in \mathbb{R} \iff \pi(\lambda)$ is a $N(0,1)$ prior on λ

A Simple Example of a Prior

- Suppose we're shown a coin, and we are told to infer whether it's biased or not just from looking at it (i.e., before flipping it)
- If $X = \mathbb{1}_{\text{heads}}$, then we want to make inferences about the random variable p, where $X \mid p \sim \text{Bernoulli}\,(p)$
- What should our prior on $\Theta = [0, 1]$ look like?
- It depends on what we know (or don't know) about the coin
- Here are three of many possible choices

Prior Distributions for the Coin Example



The Prior Predictive Distribution

- What if we were asked to predict the likelihood of the coin coming up heads at this point?

- It's reasonable to take a weighted average of all possible Bernoulli (p) distributions, each one weighted by our prior confidence $\pi(p)$, which is

$$\int_{\Theta} \mathbb{P}_p(X = 1) \cdot \pi(p)\, \mathrm{d}p = \int_0^1 p \cdot \pi(p)\, \mathrm{d}p$$

- There's a name for this

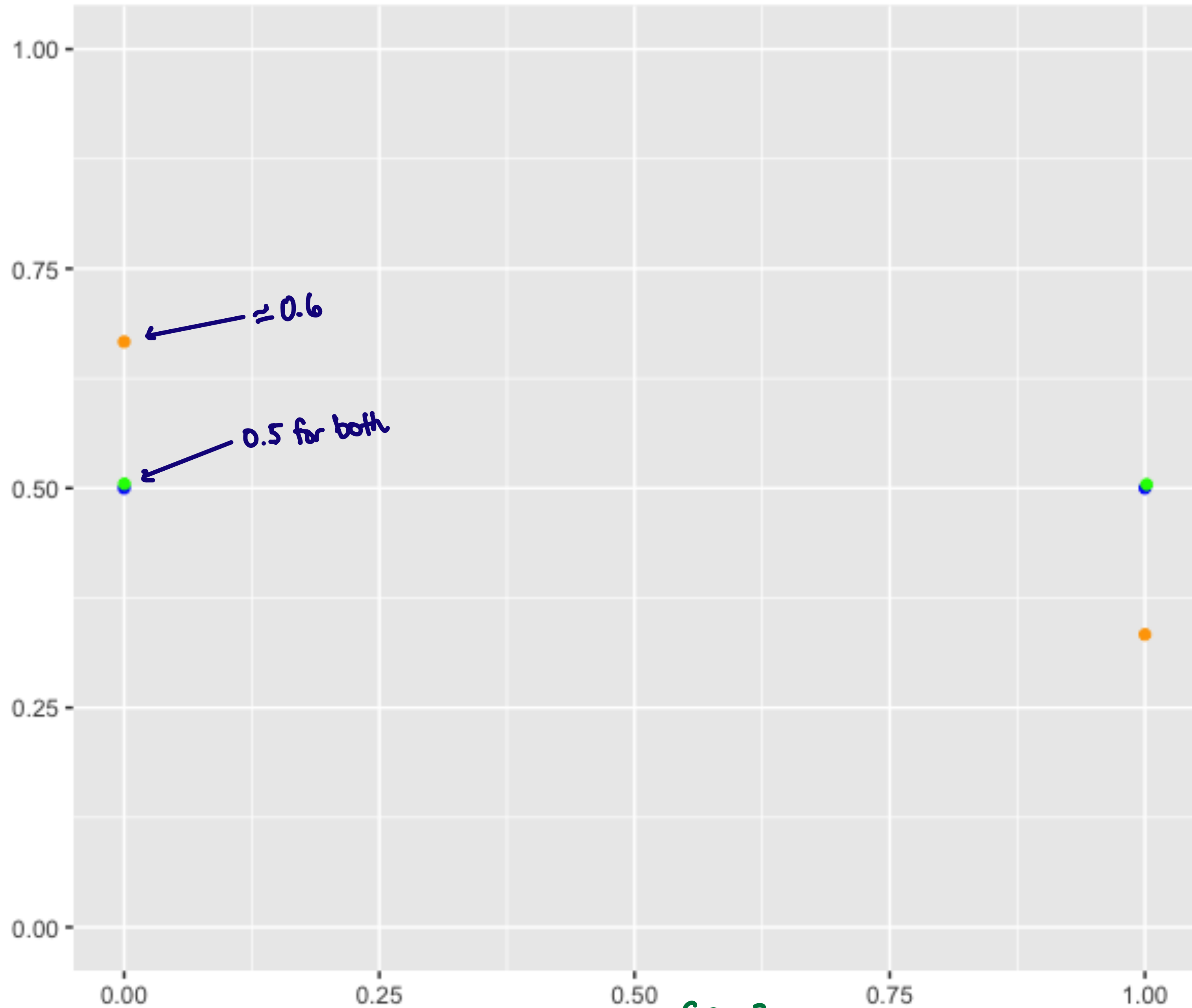
- Definition 6.1: Given a pdf f_θ and a prior distribution π on θ, the **prior predictive distribution** of the data $\mathbf{x}$ is given by the pdf

$$f(\mathbf{x}) = \int_{\Theta} f_\theta(\mathbf{x}) \cdot \pi(\theta)\, \mathrm{d}\theta.$$

$\longleftarrow$ If $X \sim \text{Bernoulli}(\theta)$, this is $\int_0^1 \theta^x (1-\theta)^{1-x} \cdot \pi(\theta)\, d\theta$

Prior Predictive Distributions for the Coin Example

$f(x) = \int_0^1 p^x (1-p)^{1-x} \cdot \pi(p)\, dp$



≈ 0.6

0.5 for both

$x \in \{0,1\}$

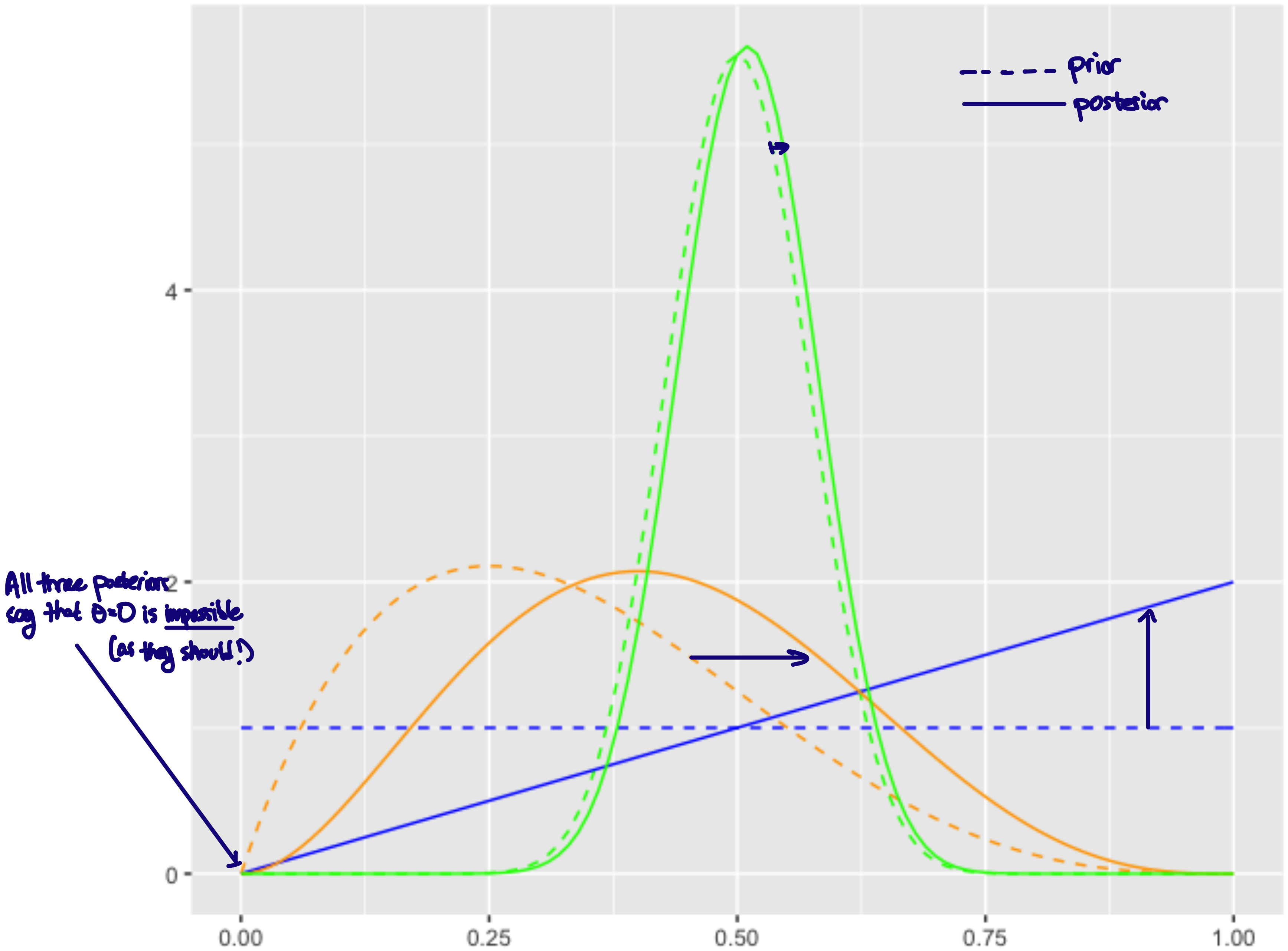
The Posterior Distribution - A Motivation

- Now, suppose we actually flip the coin once and observe $X = 1$
- If we were asked what the likelihood of some $p' \in [0, 1]$ is now, we could take our prior probability $\pi(p')$ and weigh it down by the likelihood of observing $X = 1$ if the "true" parameter really were p'
- That is, it's reasonable to answer with $\mathbb{P}_{p'}(X = 1) \cdot \pi(p')$, since data in support of p' will make this relatively high, while data in support of some p'' far away from p' will make it relatively low
- To put everything on the same scale, may as well normalize those quantities over all possible $p \in [0, 1]$ and answer instead with

$$\frac{\mathbb{P}_{p'}(X = 1) \cdot \pi(p')}{\int_0^1 \mathbb{P}_p(X = 1) \cdot \pi(p) \,\mathrm{d}p} = \frac{p' \cdot \pi(p')}{\int_0^1 p \cdot \pi(p) \,\mathrm{d}p}$$

EXERCISE:
show this is a valid pdf on $\Theta = [0,1]$
(as a function of p')

Posterior Distributions for the Coin Example $(X = 1)$



The Posterior Distribution - A Derivation

- In general, $f_\theta(\mathbf{x}) \cdot \pi(\theta)$ is the joint pdf of $(\mathbf{X}, \theta)$ $f_\theta(\vec{x}) = f_{\vec{X}|\theta}(\vec{x}|\theta)$

- From Bayes' rule, the conditional pdf of $\theta \mid \mathbf{X}$ is given by

$$\frac{f_\theta(\mathbf{x}) \cdot \pi(\theta)}{f(\mathbf{x})}$$

$\leftarrow$ prior predictive distribution!

$f(\vec{x}) = \int_\Theta f_\theta(\vec{x}) \cdot \pi(\theta)\, d\theta$

- There's also a name for this

- Definition 6.2: The **posterior distribution of $\boldsymbol{\theta}$** is the conditional distribution of $\theta \mid (\mathbf{X} = \mathbf{x})$, given by the pdf

$$\pi(\theta \mid \mathbf{x}) = \frac{f_\theta(\mathbf{x}) \cdot \pi(\theta)}{\int_\Theta f_\theta(\mathbf{x}) \cdot \pi(\theta)\, \mathrm{d}\theta}.$$

$\pi(\theta)$ is the prior distribution

$\pi(\theta|\vec{x})$ is the posterior

Poll Time!

On Quercus: Module 6 - Poll 1

More on the Posterior

"proportional to"

$f(x) \propto g(x)$ if there exists some $c \neq 0$ $\underline{\text{free of } x}$ s.t. $f(x) = c \cdot g(x)$

- The posterior $\pi(\theta \mid \mathbf{x})$ is a function of θ, and the data $\mathbf{x}$ is *observed*
- So we could write $\pi(\theta \mid \mathbf{x}) \propto f_\theta(\mathbf{x}) \cdot \pi(\theta)$ because $\pi(\theta|\vec{x}) = \dfrac{f_\theta(\vec{x}) \cdot \pi(\theta)}{\underbrace{\int f_\theta(\vec{x}) \cdot \pi(\theta)\, d\theta}_{\text{constant w.r.t. } \theta}}$
- Thus, $[\int_\Theta f_\theta(\mathbf{x}) \cdot \pi(\theta)\, \mathrm{d}\theta]^{-1}$ plays the role of normalizing constant for the unnormalized pdf $f_\theta(\mathbf{x}) \cdot \pi(\theta)$
- If the functional form of $f_\theta(\mathbf{x}) \cdot \pi(\theta)$ looks familiar, then we'll know what $(\int_\Theta f_\theta(\mathbf{x}) \cdot \pi(\theta)\, \mathrm{d}\theta)^{-1}$ must be, and we can get $\pi(\theta \mid \mathbf{x})$ for free
- Example 6.1: Suppose we calculate $f_\theta(x) \cdot \pi(\theta) \propto \overbrace{\theta^{x+1}(1-\theta)^{2-x}}^{\text{"kernel" of a distribution}}$ for $\theta \in (0,1)$. What is $\pi(\theta \mid x)$?

(i.e. the part without the normalizing constant)

It's a Beta! What are its parameters? If $Z \sim \text{Beta}(\alpha, \beta)$, then $f_Z(z) \propto z^{\alpha-1}(1-z)^{\beta-1}$.

So $\theta|\vec{x} \sim \text{Beta}(x+2, 3-x)$. Therefore, $\pi(\theta|\vec{x}) = \dfrac{\Gamma(5)}{\Gamma(x+2)\cdot\Gamma(3-x)} \cdot \theta^{x+1}(1-\theta)^{2-x}$

Integration exercise: check that $\int_0^1 \theta^{x+1}(1-\theta)^{2-x}\, d\theta = \dfrac{\Gamma(x+2)\cdot\Gamma(3-x)}{\Gamma(5)}$

More on the Posterior

- The observed data dictates how much the posterior distribution differs from the prior

- Consider three different priors:
  - π_1 is highly concentrated at $\theta_1 \in \Theta$
  - π_2 is highly concentrated at $\theta_2 \in \Theta$
  - π_3 is $\text{Unif}(\Theta)$

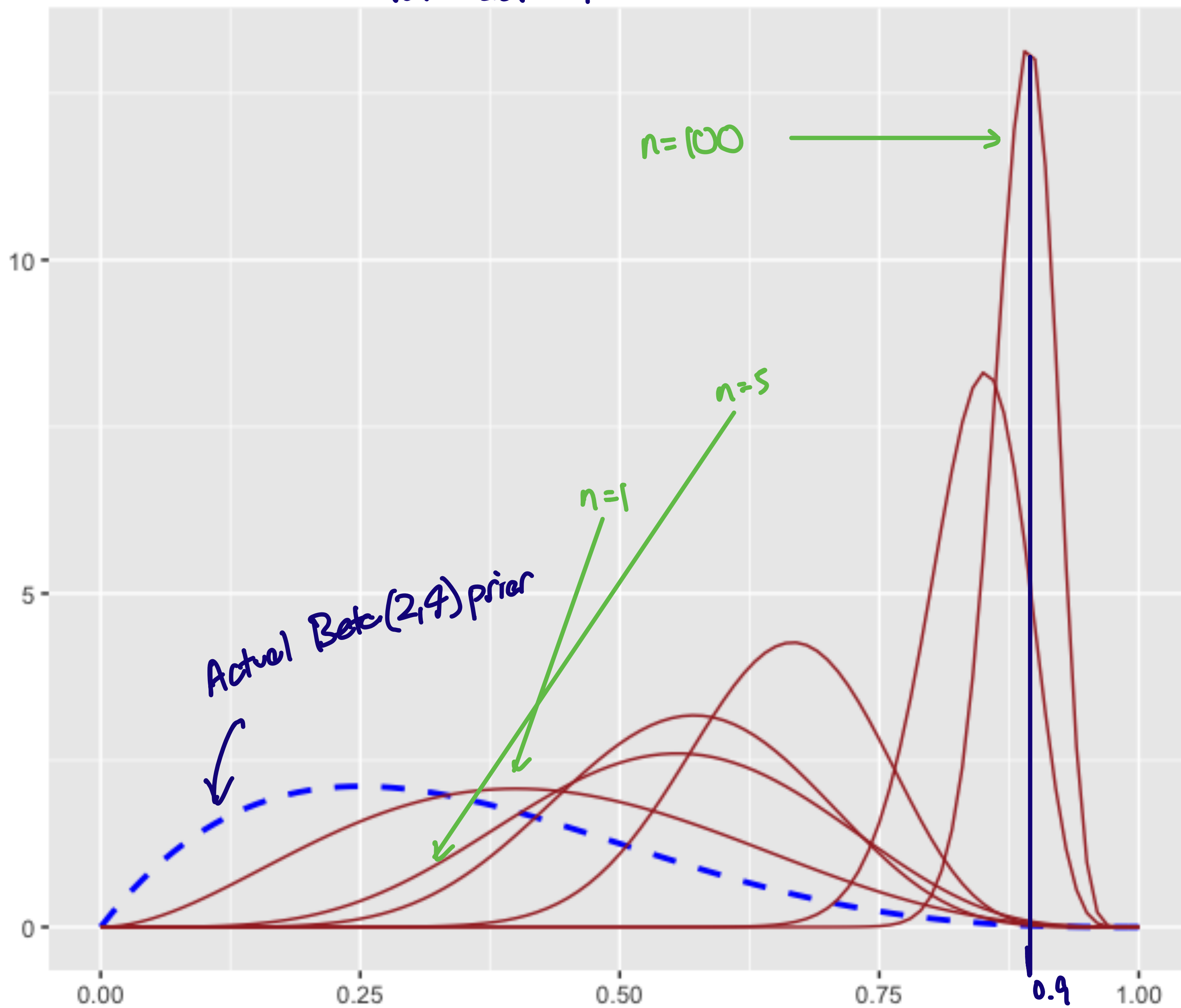
- Now we observe $\mathbf{x}$; suppose the likelihood $L(\theta \mid \mathbf{x}) = f_\theta(\mathbf{x})$ "supports" θ_2 in the frequentist sense

- What do the posteriors look like?
  - $\pi_1(\cdot \mid \mathbf{x})$ will be less concentrated at θ_1
  - $\pi_2(\cdot \mid \mathbf{x})$ will be even <u>more</u> concentrated at θ_2
  - $\pi_3(\cdot \mid \mathbf{x})$ will be (somewhat) concentrated at θ_2

- Even if the prior is strong, the likelihood will eventually "overpower" it as the sample size n grows

When the Prior and the Data Disagree

Actual data: $X \sim Bin(n, 0.9)$



Computing Posteriors: Examples

- Example 6.2: Suppose that $\pi(p) = \text{Beta}\,(\alpha, \beta)$ and $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}\,(p)$. Find the posterior $\pi(p \mid \mathbf{x})$.

$$\pi(p|\vec{x}) \propto \pi(p)\cdot f_p(\vec{x}) = \pi(p)\cdot L(p|\vec{x})$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\cdot\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}\left(\prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}\right)$$

$$\propto p^{\sum x_i+\alpha-1}(1-p)^{n-\sum x_i+\beta-1}$$

This is an unnormalized $\text{Beta}(\alpha', \beta')$ pdf, where $\alpha' = \sum x_i + \alpha$ and $\beta' = n - \sum x_i + \beta$

$$\Rightarrow \pi(p|\vec{x}) = \frac{\Gamma(\sum x_i + \alpha + n - \sum x_i + \beta)}{\Gamma(\sum x_i + \alpha)\cdot\Gamma(n - \sum x_i + \beta)} \cdot p^{\sum x_i+\alpha-1}(1-p)^{n-\sum x_i+\beta-1}$$

$$\Rightarrow p|\vec{x} \sim \text{Beta}\left(\sum x_i + \alpha, n - \sum x_i + \beta\right).$$

Same parametric family as $\pi(p)$, but with the original parameters "updated" in light of $\vec{X} = \vec{x}$

Computing Posteriors: Examples

- Example 6.3: Suppose that $\pi(\lambda) = \text{Gamma}(\alpha, \beta)$ and $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \text{Poisson}(\lambda)$. Find the posterior $\pi(\lambda \mid \mathbf{x})$.

$$\pi(\lambda|\vec{x}) \propto \pi(\lambda) \cdot L(\lambda|\vec{x})$$

$$\propto \underbrace{\lambda^{\alpha-1} e^{-\beta\lambda}}_{\text{Unnormalized Gamma}(\alpha,\beta)\text{ pdf}} \cdot \left(\prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \right)$$

$$\propto \lambda^{\sum x_i + \alpha - 1} e^{-(n+\beta)\lambda}$$

$$\Longrightarrow \lambda|\vec{x} \sim \text{Gamma}\left(\sum x_i + \alpha,\ n+\beta\right)$$

The Return of Sufficiency

- What if instead of observing $\mathbf{x}$, we only have access to a sufficient statistic $T(\mathbf{x})$?

- Sufficiency kind of carries over to the Bayesian setting, in the following sense

- Theorem 6.1: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} f_\theta$ and let $\pi(\theta)$ be a prior on θ. If $T(\mathbf{X})$ is a sufficient statistic for θ (in the frequentist sense), then $\underbrace{\pi(\theta \mid \mathbf{x})}_{\text{Posterior given } \pi(\theta) \text{ and } \vec{X}=\vec{x}} = \underbrace{\pi(\theta \mid T(\mathbf{x}))}_{\text{Posterior given } \pi(\theta) \text{ and } T=t}$.

$t = T(\vec{x})$

Proof: EXERCISE!

Computing Posteriors: Examples

$T(\vec{X}) = \sum_{i=1}^{n} X_i$ is sufficient for p

- Example 6.4: Suppose that $\pi(p) = \text{Beta}(\alpha, \beta)$ and $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}(p)$. Find the posterior $\pi(p \mid \sum_{i=1}^{n} x_i)$.

$T(\vec{X}) \sim \text{Bin}(n, p)$

Let $t = \sum x_i$.

$$\begin{aligned}
\pi(p \mid t) &\propto \pi(p) \cdot f_p(t) \\
&\propto p^{\alpha-1}(1-p)^{\beta-1} \cdot \left(\binom{n}{t} \cdot p^t (1-p)^{n-t} \right) \\
&\propto p^{t+\alpha-1}(1-p)^{n-t+\beta-1} \\
&= p^{\sum x_i + \alpha - 1}(1-p)^{n - \sum x_i + \beta - 1}
\end{aligned}$$

$\Rightarrow p \mid \sum x_i \sim \text{Beta}(\sum x_i + \alpha, n - \sum x_i + \beta)$. Same posterior as before!

Hyperparameters

- In the previous example, the prior $\pi(\theta) = \cancel{\text{Gamma}}\ \text{Beta}(\alpha, \beta)$ had its own set of parameters: α and β

  a generic parameter (like "θ" used to be), which could be a vector; e.g., $\lambda = (\alpha, \beta)$

- Definition 6.3: The parameters λ of a prior distribution $\pi_\lambda(\cdot)$ in a parametric family $\{\pi_\lambda : \lambda \in \Lambda\}$ are called **hyperparameters**.

- Sometimes the hyperparameter λ is a given constant (either known from prior experience or chosen based on the situation) , e.g., $\lambda = (\alpha, \beta)$

- Other times, we go meta and assign a prior distribution to λ itself (called a **hyperprior**, possibly with its own **hyperhyperparameters**)

  an actual word!

- Models of this sort are called **hierarchical Bayesian models**

- We could keep going and assign a hyperhyperprior to the hyperhyperparameters, and a hyperhyperhyperprior to the hyperhyperhyperparameters, and... ... but we've gotta stop somewhere!

Poll Time!

On Quercus: Module 6 - Poll 2

Choosing Priors

- How do we choose an appropriate prior (both for the parameter associated with the data, as well as any hyperparameters)?

- There's no single answer to this question

- One of a Bayesian statistician's key roles is arguing with other statisticians about prior selection Almost every paper that applies Bayesian statistics will justify their choices & priors... It's important!

- Some priors are simply not sensible given the parametric family for the data

- Example 6.5: $X_1, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}(p)$ $\pi(p) = \text{Unif}(-1, 0)$ makes no sense! $\pi(p) = N(-10, 20)$ makes no sense!

  $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. $\pi(\sigma^2) = \text{Unif}(\{5, 8\})$ probably not that sensible...

- We'll discuss several commonly used methods of prior selection, but these certainly aren't the only ones (nor are they mutually exclusive)

Objectivity Versus Subjectivity

- One can very roughly classify Bayesians into two groups: *objective Bayesians* and *subjective Bayesians*
- Subjective Bayesians prefer to integrate personal beliefs about the world – or lack thereof – into their inferences, and they would choose priors that reflect their beliefs (to the extent possible)
- Of course, these would influence the posterior, so two subjective Bayesians might come up with different posteriors (even if they both agree on a model for the data itself); these reflect their differing opinions
- Objective Bayesians prefer to let the data speak for itself, and they would choose priors that do not reflect any personal biases
- To an objective Bayesian, there should be a fixed procedure for choosing a prior, and therefore everyone should agree on the same posterior

Conjugate Priors

- In the previous examples, the posterior distribution was in the same parametric family as the prior (albeit with "updated" parameters)

- This doesn't always happen – most of the time, the posterior will be an unfamiliar distribution – but when it does happen, there's a special name for it

- Definition 6.4: A family of priors $\{\pi_\lambda : \lambda \in \Lambda\}$ for the parameter θ of the model $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ is called **conjugate for $\mathcal{F}$** if, for all data $\mathbf{x} \in \mathcal{X}^n$ and all $\lambda \in \Lambda$, the posterior $\pi(\cdot \mid \mathbf{x}) \in \{\pi_\lambda : \lambda \in \Lambda\}$

- Example 6.6: Beta(α, β) is conjugate for Bernoulli(p) (and Bin(n,p)) (and others)

- Example 6.7: Gamma(α, β) is conjugate for Poisson(λ)

Conjugate Priors

- Example 6.8: Suppose that $\pi(\mu) = \mathcal{N}(\theta, \tau^2)$ and $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ where σ^2 is known. Find the posterior $\pi(\mu \mid \mathbf{x})$.

We know that $T(\vec{x}) = \bar{X}_n \sim N(\mu, \sigma^2/n)$ is sufficient for μ.

Let $t = \bar{x}_n$. Then by Theorem 6.1,

$$\pi(\mu \mid \vec{x}) = \pi(\mu \mid t)$$

$$\propto \pi(\mu) \cdot f_\mu(t)$$

$$\propto \exp\left(\frac{-(\mu-\theta)^2}{2\tau^2}\right) \cdot \exp\left(\frac{-(t-\mu)^2}{2\sigma^2/n}\right)$$

$$= \exp\left(\frac{-\mu^2 + 2\mu\theta - \theta^2}{2\tau^2} + \frac{-t^2 + 2\mu t - \mu^2}{2\sigma^2/n}\right)$$

$$\propto \exp\left(\frac{-\mu^2 + 2\mu\theta}{2\tau^2} + \frac{2\mu t - \mu^2}{2\sigma^2/n}\right)$$

looks like $\exp\left(\frac{-(\mu-a)^2}{b^2}\right)$ for some a, b...

What's inside the exponential function?

$$\frac{2\mu\theta - \mu^2}{2\tau^2} + \frac{2\mu t - \mu^2}{2\sigma^2/n}$$

$$= \mu^2\left[\frac{-1}{2\tau^2} - \frac{n}{2\sigma^2}\right] + \mu\left[\frac{\theta}{\tau^2} + \frac{nt}{\sigma^2}\right]$$

$$= \left[\frac{-1}{2\tau^2} - \frac{n}{2\sigma^2}\right]\left(\mu^2 - 2\mu\frac{\left[\frac{\theta}{\tau^2} + \frac{nt}{\sigma^2}\right]}{\left[\frac{1}{\tau^2} + \frac{n}{\sigma^2}\right]}\right)$$

$$= \left[\frac{-1}{2\tau^2} - \frac{n}{2\sigma^2}\right]\left(\mu^2 - 2\mu\frac{\left[\frac{\theta}{\tau^2} + \frac{nt}{\sigma^2}\right]}{\left[\frac{1}{\tau^2} + \frac{n}{\sigma^2}\right]} + \frac{\left[\frac{\theta}{\tau^2} + \frac{nt}{\sigma^2}\right]^2}{\left[\frac{1}{\tau^2} + \frac{n}{\sigma^2}\right]^2}\right) + c, \text{ where } c \text{ is free is } \mu$$

$$= -\left[\frac{1}{2\tau^2} + \frac{n}{2\sigma^2}\right]\left(\mu - \frac{\left[\frac{\theta}{\tau^2} + \frac{nt}{\sigma^2}\right]}{\left[\frac{1}{\tau^2} + \frac{n}{\sigma^2}\right]}\right)^2 + c$$

$$= \frac{-\left(\mu - \frac{\left[\frac{\theta}{\tau^2} + \frac{nt}{\sigma^2}\right]}{\left[\frac{1}{\tau^2} + \frac{n}{\sigma^2}\right]}\right)^2}{2\left[\frac{1}{\tau^2} + \frac{n}{\sigma^2}\right]^{-1}} + c$$

$$\pi(\mu \mid \vec{x}) \propto \exp\left(\frac{-\left(\mu - \frac{\left[\frac{\theta}{\tau^2} + \frac{nt}{\sigma^2}\right]}{\left[\frac{1}{\tau^2} + \frac{n}{\sigma^2}\right]}\right)^2}{2\left[\frac{1}{\tau^2} + \frac{n}{\sigma^2}\right]^{-1}}\right)$$

$$\Rightarrow \pi(\mu \mid \vec{x}) = N\left(\frac{\frac{\theta}{\tau^2} + \frac{n\bar{x}}{\sigma^2}}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}}, \frac{1}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}}\right)$$

Conjugate Priors

What happens when τ^2 is really close to 0?
Or when n is very large?

- In those examples, it was no coincidence that both prior and likelihood were in exponential families

- Theorem 6.2: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} f_\theta$ where f_θ is in an exponential family:

$$f_\theta(x) = h(x) \cdot g(\theta) \cdot \exp\left(\sum_{j=1}^{k} \eta_j(\theta) \cdot T_j(x)\right).$$

If we choose an exponential family prior of the form

$$\pi(\theta) \propto g(\theta)^\nu \cdot \exp\left(\sum_{j=1}^{k} \eta_j(\theta) \cdot \xi_j\right)$$

where ν and $\xi_1, \ldots, \xi_k$ are hyperparameters, then $\pi(\theta)$ is a conjugate prior for f_θ.

Proof: EXERCISE! Identify the "updated" parameters, too!

Why Conjugate Priors?

- Conjugacy is very mathematically convenient
- But is a conjugate family actually *relevant* to whatever the statistical situation is?
- It's widely acknowledged that most conjugate families are rich enough to express a wide spectrum of prior beliefs
- Example 6.9: The $N(\theta, \tau^2)$ prior for μ in the $N(\mu, \sigma^2)$ model: if we're encoding "symmetric" and "unimodal" prior knowledge about μ, then this prior accommodates a lot

The $\text{Beta}(\alpha, \beta)$ prior for p in the $\text{Bernoulli}(p)$ model: can handle uniform prior beliefs, any "mode" in $(0,1)$, etc...

Elicitation

- Even if we do have a particular parametric family $\{\pi_\lambda : \lambda \in \Lambda\}$ selected for our prior, how do we actually set the hyperparameters?

- Ideally, we'll have some experts in the field (possibly ourselves) available to give us their thoughts on what they believe is plausible, based on their own past experiences

- We can't expect them to just tell us raw numbers for λ, but with enough information, we can try and work out the best match

- Translating those thoughts into a choice of hyperprior is called **prior elicitation**

Poll Time!

On Quercus: Module 6 - Poll 3

Elicitation: Examples

- Example 6.10: Suppose we're sampling from an $\mathcal{N}\left(\mu, \sigma^2\right)$ distribution with μ unknown and σ^2 known, and we restrict attention to the family $\{\mathcal{N}\left(\mu_0, \tau^2\right) : \mu_0 \in \mathbb{R}, \tau^2 > 0\}$. If an expert tells us they're 50% certain that μ lies between 2 and 3, how can we elicit our prior?

Gotta choose $\mu_0 = 2.5$. What about τ^2?

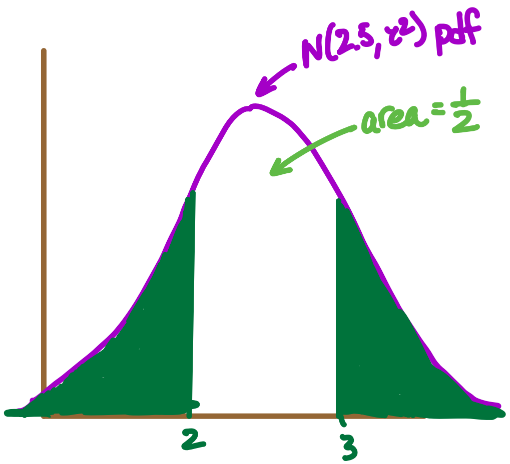
Suppose $\mu \sim N(2.5, \tau^2)$. We know that $\frac{1}{2} = \mathbb{P}(\mu \in (2,3))$.

Then $1 - \frac{1}{2} = \frac{1}{2} = \mathbb{P}(\mu \le 2) + \mathbb{P}(\mu \ge 3)$

$$= \mathbb{P}\left(\frac{\mu - 2.5}{\tau} \le \frac{2-2.5}{\tau}\right) + \mathbb{P}\left(\frac{\mu-2.5}{\tau} \ge \frac{3-2.5}{\tau}\right)$$

$$= \mathbb{P}(Z \le {-0.5}/\tau) + \mathbb{P}(Z \ge 0.5/\tau) \text{ where } Z \sim N(0,1)$$

$$= 2 \cdot \Phi\left(\frac{-0.5}{\tau}\right)$$



$$\Rightarrow \tau = \frac{-0.5}{\Phi^{-1}(0.25)}$$

So we should choose $\pi(\mu) = N\left(2.5, \left(\frac{-0.5}{\Phi^{-1}(0.25)}\right)^2\right)$

Expressing Ignorance

- What if the experts are keeping quiet and we have nothing to work with?
- Or maybe we're objective Bayesians and "expert advice" is irrelevant to us
- How do we choose a prior that expresses *complete* ignorance about θ?
- In the coin example, choosing $\pi(p) = \text{Unif}\,(0, 1)$ would work
- What about a completely objective prior on μ in the $\mathcal{N}\left(\mu, \sigma^2\right)$ model? There's no uniform distribution on $\mathbb{R}$

  $\int_{-\infty}^{\infty} c \, d\mu$ does not exist for any $c \neq 0$ (:)
- And yet, if we take $\pi(\mu) = 1$, (or even $\pi(\mu) \propto 1$)

$$\pi(\mu \mid \vec{x}) \propto 1 \cdot \exp\left(-\frac{(\bar{x}-\mu)^2}{2\sigma^2/n}\right) = \exp\left(\frac{-(\bar{x}-\mu)^2}{2\sigma^2/n}\right)$$

$$\Rightarrow \mu \mid \vec{x} \sim N(\bar{x}, \sigma^2/n)$$

This is a completely legitimate posterior!
It's clearly letting the data do all the talking...

Uninformative Priors

- Definition 6.5: A function $\pi(\theta)$ used in place of a true prior distribution that does not relect any prior beliefs about θ is called an **uninformative** (or **noninformative** or **default** or **reference**) **prior**.

  $\pi(\mu) \propto 1$ in the $N(\mu, \sigma^2)$ model, $\mu \in \mathbb{R}$, σ^2 known

- Example 6.11: $\pi(\theta) \propto 1$ in the $\text{Unif}(0, \theta)$ model, $\theta > 0$

  $\pi(p) = 1$ in the Bernoulli(p) model, $p \in (0,1)$

- We have a special name for choices like $\pi(\mu) = 1$ above

- Definition 6.6: If an uninformative prior $\pi(\theta)$ is not a true distribution (i.e., $\int_\Theta \pi(\theta)\, d\theta$ is divergent), then it is called an **improper prior**.

- Improper priors are controversial, and they're difficult to interpret probabilistically

  $\pi(\theta)$ is improper iff $c \cdot \pi(\theta)$ is improper, for any $c > 0$

- Moreover, if chosen haphazardly they can lead to improper posteriors (which are truly meaningless)

Problems With Uninformative Priors

- Example 6.12: Suppose that $X \sim \text{Bernoulli}(p)$. What is the posterior $\pi(p \mid x)$ based on the **Haldane prior** $\pi(p) = \frac{1}{p(1-p)}$?

This is improper! $\int_0^1 \pi(p)\,dp = \infty$.

$p \in (0,1) \Rightarrow (1-p) \in (0,1) \Rightarrow \frac{1}{1-p} \geq 1$

$\Rightarrow \int_0^1 \frac{1}{p(1-p)}\,dp \geq \int_0^1 \frac{1}{p}\,dp = \log(1) - \log(0) = \infty$

$$\pi(p \mid \vec{x}) \propto \frac{1}{p(1-p)} \cdot p^x (1-p)^{1-x}$$

$$= p^{x-1}(1-p)^{-x}.$$

Is this a pdf? $\int_0^1 \pi(p \mid \vec{x})\,dp = \int_0^1 p^{x-1}(1-p)^{-x}\,dp$

$= \pi \cdot \csc(\pi x)$ Calculus exercise!

$= \frac{\pi}{\sin(\pi x)}$

$= \pm\infty$ when $x \in \mathbb{Z}$ ... which is the case here!

Not a pdf! This is an "improper posterior" — useless!

This can never happen when we choose a proper prior...

Problems With Uninformative Priors

- Example 6.13: Suppose that $X \sim \text{Bernoulli}\,(p)$ and we choose $\pi(p) = \text{Unif}\,(0,1)$. What prior does this correspond to for the log-odds $\tau = \log\left(\frac{p}{1-p}\right)$?

inverses

$p(\tau) = \frac{1}{1+e^{-\tau}}$ "expit function" maps $\mathbb{R}$ to $(0,1)$

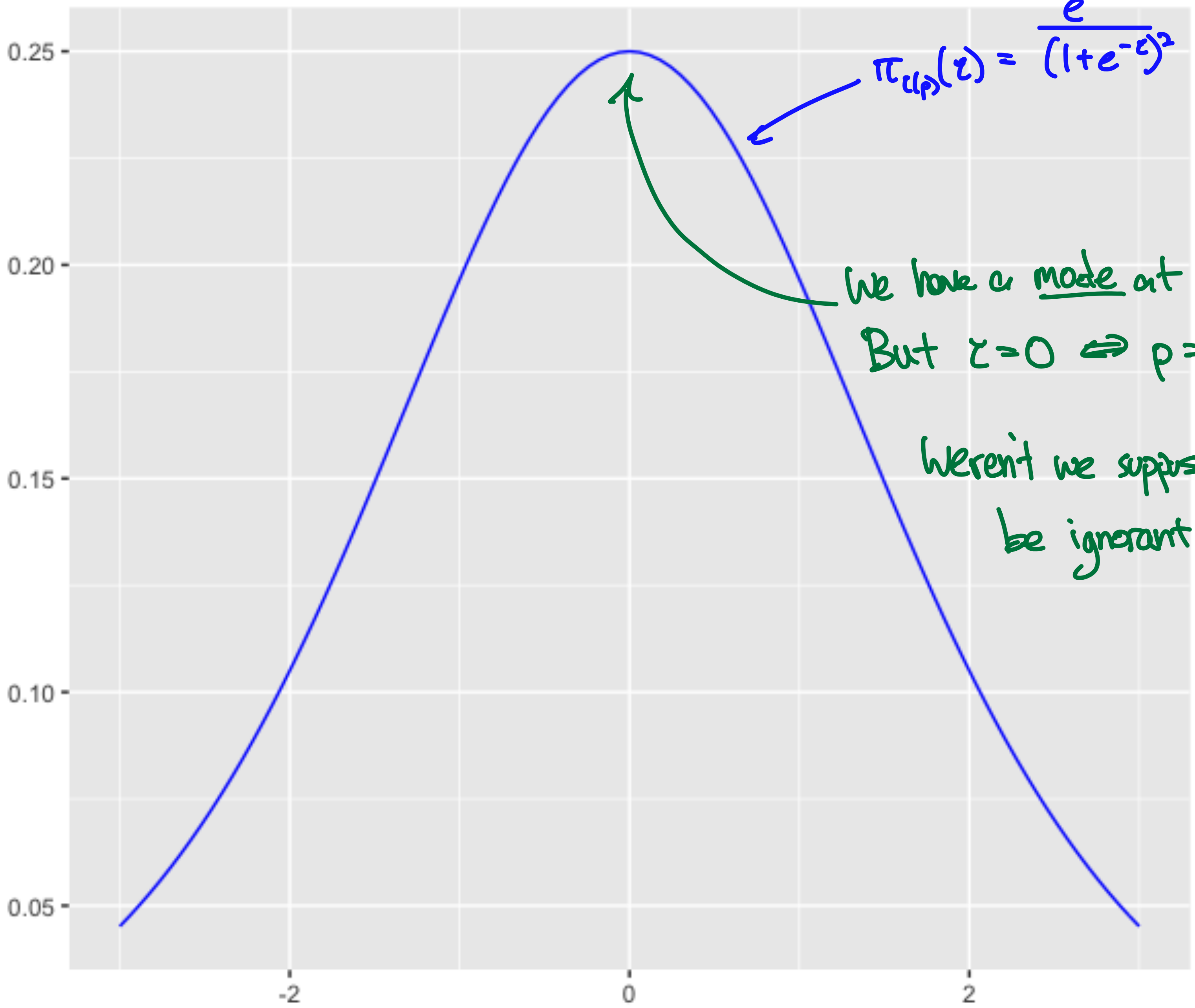
$\tau(p) = \log\left(\frac{p}{1-p}\right)$ "logit function" maps $(0,1)$ to $\mathbb{R}$.

$$\pi_{\tau(p)}(\tau) = \pi_p(p(\tau)) \cdot \left|\frac{d}{d\tau} p(\tau)\right|$$

$$= 1 \cdot \left|\frac{e^{-\tau}}{(1+e^{-\tau})^2}\right|$$

$$= \frac{e^{-\tau}}{(1+e^{-\tau})^2}$$

Oh No



$\pi_{\tau(p)}(\tau) = \dfrac{e^{-\tau}}{(1+e^{-\tau})^2}$

We have a mode at $\tau = 0$...

But $\tau = 0 \iff p = 0.5$.

Weren't we supposed to be ignorant?!

Ignorance From All Perspectives

- The previous example shows that ignorance about θ does not necessarily translate to the same ignorance about $\tau(\theta)$

- In other words, if π_θ is a prior for the model parameterized by θ and π_τ is a prior for the model parameterized by $\tau = \tau(\theta)$,

$$\pi_\tau(t) \neq \pi_\theta(\tau^{-1}(t)) \cdot \left| \frac{\mathrm{d}}{\mathrm{d}t} \tau^{-1}(t) \right|$$

  in general

- What if we insisted on "equivalent" ignorance for all monotone re-parametrizations of θ?

- It turns out there's a way to make this happen using the Fisher information

Jeffreys' Prior

- Definition 6.7: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} f_\theta$ where θ is univariate. **Jeffreys' prior** for θ is given by $\pi^J_\theta(\theta) \propto \sqrt{I_1(\theta)}$.

- Notice that this prior *depends only the model* – there's no room for any subjectivity beyond the choice of model

- Jeffreys felt that invariance under monotone transformations is a suitably uninformative property for a prior

- Theorem 6.3: Under the regularity conditions of the Cramér-Rao Lower Bound, Jeffreys' prior is invariant under monotone transformations, in the sense that

$$\pi^J_\tau(t) = \pi^J_\theta(\tau^{-1}(t)) \left| \frac{\mathrm{d}}{\mathrm{d}t} \tau^{-1}(t) \right|$$

  if $\tau : \Theta \to \mathbb{R}$ is monotone and differentiable.

Proof. Let $f_\theta(\vec{x})$ be the original pdf, and let $g_\tau(\vec{x})$ be the pdf under the $\tau(\theta)$ transformation.
Let $I_\theta(\theta)$ and $I_\tau(\tau)$ be the Fisher information under the two parameterizations.

Then... $I_\theta(\theta) = \mathbb{E}_\theta\left[\left(\frac{d}{d\theta}\log(f_\theta(\vec{X}))\right)^2\right]$ by definition

$= \mathbb{E}_\theta\left[\left(\frac{d}{d\theta}\log(g_{\tau(\theta)}(\vec{X}))\right)^2\right]$ because $f_\theta(x) = g_\tau(x)$; reparameterization doesn't change the likelihood

$= \mathbb{E}_\theta\left[\left(\frac{d\tau}{d\theta}\cdot\frac{d}{d\tau}\log(g_{\tau(\theta)}(\vec{X}))\right)^2\right]$ by the chain rule

$= \left(\frac{d\tau}{d\theta}\right)^2 \cdot \mathbb{E}_\tau\left[\left(\frac{d}{d\tau}\log(g_\tau(\vec{X}))\right)^2\right]$

$= \left(\frac{d\tau}{d\theta}\right)^2 \cdot I_\tau(\tau)$

Thus... $\pi_\tau^J(\tau(\theta)) \propto \sqrt{I_\tau(\tau)}$ by definition of Jeffreys' prior

$= \sqrt{I_\theta(\theta)} \cdot \left|\frac{d\tau}{d\theta}\right|^{-1}$

$= \sqrt{I_\theta(\theta)} \cdot \left|\frac{d\theta}{d\tau}\right|.$

The result follows upon letting $t = \tau(\theta)$. $\square$

Jeffreys' Prior: Examples

- Example 6.14: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim}$ Bernoulli (p). Determine Jeffreys' prior for this model, and determine the posterior $\pi(p \mid \mathbf{x})$ based on it.

We know from old stuff that $I_1(p) = \frac{1}{p(1-p)}$, so that $\pi^J(p) \propto \sqrt{\frac{1}{p(1-p)}} = p^{-\frac{1}{2}}(1-p)^{-\frac{1}{2}}$.

It's a Beta$(\frac{1}{2}, \frac{1}{2})$ distribution!

Our posterior is
$$\pi(p \mid \vec{x}) \propto \pi^J(p) \cdot f_p(\vec{x})$$
$$\propto p^{-\frac{1}{2}}(1-p)^{-\frac{1}{2}} p^{\sum x_i}(1-p)^{n-\sum x_i}$$
$$= p^{\sum x_i - \frac{1}{2}}(1-p)^{n-\sum x_i - \frac{1}{2}}$$

$$\Rightarrow p \mid \vec{x} \sim \text{Beta}\left(\sum x_i + \tfrac{1}{2},\ n - \sum x_i + \tfrac{1}{2}\right)$$

What if $\tau(p) = \arcsin(\sqrt{p})$? $\Rightarrow p(\tau) = \sin^2(\tau)$

$$\Rightarrow \pi^J_\tau(\tau) \propto \pi^J_p(p(\tau)) \cdot \left|\frac{d}{d\tau} p(\tau)\right| \text{ by Theorem 6.3}$$
$$= \sin^2(\tau)^{-\frac{1}{2}}(1-\sin^2(\tau))^{-\frac{1}{2}}\left|2 \cdot \sin(\tau) \cdot \cos(\tau)\right|$$
$$= 2 \Rightarrow \pi^J_\tau(\tau) \propto 2 \cdot \mathbb{1}_{\tau \in (0, \pi/2)} \Rightarrow \pi^J(\tau) = \text{Unif}(0, \pi/2)$$

$p \in (0,1)$
$\Leftrightarrow \sqrt{p} \in (0,1)$
$\Leftrightarrow \arcsin(\sqrt{p}) \in (0, \pi/2)$
$\Leftrightarrow \tau \in (0, \pi/2)$

Jeffreys' Prior: Examples

- Example 6.15: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ with σ^2 known. Determine Jeffreys' prior for this model, and determine the posterior $\pi(\mu \mid \mathbf{x})$ based on it.

From many examples past, $I_1(\mu) = 1/\sigma^2$. So $\pi^J(\mu) \propto \sqrt{1/\sigma^2} \propto 1$.

That's improper, because $\int_{-\infty}^{\infty} \pi^J(\mu)\,d\mu$ DNE!

Our posterior is $\pi(\mu \mid \vec{x}) \propto \pi^J(\mu) \cdot f_\mu(\vec{x})$

$$\propto 1 \cdot \exp\left(\frac{-(\bar{x}-\mu)^2}{2\sigma^2/n}\right)$$

$$\Rightarrow \mu \mid \vec{x} \sim N(\bar{x}, \sigma^2/n).$$

Inferences Based On the Posterior

- If we're satisfied with a choice of prior and we've computed (or estimated) the posterior, what do we actually do with this distribution?
- The inferential techniques of Modules 2-4 (point estimation, hypothesis testing, and confidence intervals) can't be directly applied here, since $\theta \mid \mathbf{x}$ is not a fixed constant
- Our goal is to find Bayesian analogues of these techniques

There are LOTS of Bayesian analogues of frequentist concepts,

but (almost) none are fully agreed upon by all Bayesians...

Bayesian Point Estimation

- If $\mathbf{X} \sim f_\theta$, how do we "estimate" either θ itself or some quantity $\tau = \tau(\theta)$ in the Bayesian context?

- We have a posterior distribution $\pi(\theta \mid \mathbf{x})$ to work with

- What quantities can we extract from it that can meaningfully take the place of our frequentist estimates?

  eg, the mean, the median, some quantile...

- If we use some characteristic $\hat{\theta}$ of $\pi(\theta \mid \mathbf{x})$, then it must be a function of the data $\mathbf{x}$ and we can write $\hat{\theta} = \hat{\theta}(\mathbf{x})$

- That makes $\hat{\theta}(\mathbf{X})$ a genuine point estimator, which we can compare to our favourite frequentist estimators like the MLE

- To keep the notation simple, we'll work with θ itself, but everything carries over to $\tau(\theta)$

MAP Estimators

- One reasonable approach is to choose the value that the posterior says is most probable – that is, the mode of the posterior

- Definition 6.8: Given a posterior distribution $\pi(\theta \mid \mathbf{x})$, a **maximum *a posteriori* (MAP) estimator** of θ is given by the conditional mode of the posterior:

$$\hat{\theta}_{\mathsf{MAP}}(\mathbf{X}) = \underset{\theta \in \Theta}{\operatorname{argmax}}\, \pi(\theta \mid \mathbf{X}).$$

(assuming the posterior is unimodal)

- If we want the MAP estimator of $\tau = \tau(\theta)$, we'll need to maximize $\pi(\tau \mid \mathbf{x})$

- But that's the same as maximizing $f(\mathbf{x}) \cdot \pi(\tau \mid \mathbf{x}) = \pi(\tau) \cdot f_\tau(\mathbf{x})$, so we don't need to bother with the normalizing constant $f(\mathbf{x})$, which is usually a nasty integral

Posterior Means

- We might prefer to take a weighted average of all $\theta' \in \Theta$, each weighed down by how probable the posterior says it is – that is, the expectation of the posterior

- Definition 6.9: Given a posterior distribution $\pi(\theta \mid \mathbf{x})$, the **posterior mean estimator** – if it exists – is given by the conditional expectation of the posterior:

$$\hat{\theta}_{\mathsf{B}}(\mathbf{X}) = \mathbb{E}\left[\theta \mid \mathbf{X}\right] = \int_{\Theta} \theta \cdot \pi(\theta \mid \mathbf{x})\, \mathrm{d}\theta.$$

- The posterior mean estimator is nice because it minimizes the *expected MSE* under the posterior:

$$\hat{\theta}_{\mathsf{B}}(\cdot) = \underset{T(\cdot)}{\text{argmin}}\, \overbrace{\mathbb{E}\left[\mathrm{MSE}_{\theta}\left(T(\mathbf{X})\right)\right]}^{\int_{\Theta} \mathrm{MSE}_\theta(T(\vec{x}))\cdot\pi(\theta \mid \vec{x})\, d\theta}$$

taken with respect to $\pi(\theta \mid \vec{x})$

minimum over all functions $T(\cdot)$ which give us estimators $T(\vec{x})$

Bayesian Point Estimation: Examples

- Example 6.16: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim}$ Bernoulli (p), and suppose we place a Beta (α, β) prior on p. Find the MAP estimator and the posterior mean estimator for p, and describe how they compare to the MLE.

From Example 6.2, $\pi(p|\vec{x}) = \text{Beta}(\alpha + \sum x_i, \beta + n - \sum x_i)$.

MAP: gotta maximize a Beta pdf $f_{\alpha,\beta}(\theta)$ in θ. That's the same as maximizing $\log(f_{\alpha,\beta}(\theta))$.

$$\frac{d}{d\theta}\log(f_{\alpha,\beta}(\theta)) = \frac{d}{d\theta}\left((\alpha-1)\cdot\log(\theta) + (\beta-1)\cdot\log(1-\theta)\right) = \frac{\alpha-1}{\theta} - \frac{\beta-1}{1-\theta} \overset{\text{set}}{=} 0 \Rightarrow \hat{\theta} = \frac{\alpha-1}{\alpha+\beta-2}$$

Provided that $\alpha,\beta>1$

check!

So $$\hat{p}_{MAP}(\vec{X}) = \frac{\alpha + \sum X_i - 1}{\alpha + \sum X_i + \beta + n - \sum X_i - 2} = \frac{\sum X_i + \alpha - 1}{\alpha+\beta+n-2}.$$

Posterior mean: the mean of a Beta(α,β) is $\frac{\alpha}{\alpha+\beta}$. So $$\hat{p}_B(\vec{X}) = \frac{\sum X_i + \alpha}{\alpha + \sum X_i + \beta + n - \sum X_i} = \frac{\sum X_i + \alpha}{\alpha+\beta+n}$$

MLE: $\hat{p}_{MLE}(\vec{X}) = \overline{X}_n = \frac{\sum X_i}{n}$.

All three are pretty similar... but the posterior mean and MAP estimators reflect prior information (ie, choices of α and β) in different ways. But when n is large, the differences become negligible!

EXERCISE: what (if anything) happens as $n \to \infty$? What if we "chose" α and β to make $\hat{p}_{MAP} = \hat{p}_{MLE}$? Or $\hat{p}_B = \hat{p}_{MLE}$? Would that be a legitimate prior?

Bayesian Point Estimation: Examples

- Example 6.17: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ with σ^2 known, and suppose we place a $\mathcal{N}(\theta, \tau^2)$ prior on μ. Find the MAP estimator and the posterior mean estimator for μ, and describe how they compare to the MLE.

From Example 6.8, $\mu|\vec{x} \sim N\left(\dfrac{\frac{\theta}{\tau^2} + \frac{n\bar{x}}{\sigma^2}}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}}, \dfrac{1}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}}\right)$

MAP estimator: $\hat{\mu}_{MAP}(\vec{X}) = \dfrac{\frac{\theta}{\tau^2} + \frac{n\bar{X}_n}{\sigma^2}}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}} = \hat{\mu}_B(\vec{X})$: Posterior mean estimator

MLE: $\hat{\mu}_{MLE}(\vec{X}) = \bar{X}_n$

For the normal distribution, the mean equals the mode (equals the median)

As n gets large, $\hat{\mu}_{MAP}(\vec{X}) = \hat{\mu}_B(\vec{X}) \simeq \hat{\mu}_{MLE}(\vec{X})$

EXERCISE: what do they converge to as $n \to \infty$?

Poll Time!

On Quercus: Module 6 - Poll 4

Bayesian Hypothesis Testing

- What about Bayesian hypothesis testing?
- We might think to test every hypothesis by simply computing probability under $\pi(\theta \mid \mathbf{x})$, we'd quickly run into problems
- For example, if the posterior is continuous, then we'd reject every simple hypothesis $H : \theta = \theta_0$
- We might try to get around this by computing a **Bayesian p-value** $\Pi(\{\theta : \pi(\theta \mid \mathbf{x}) \leq \pi(\theta_0 \mid \mathbf{x})\} \mid \mathbf{x})$, but there can be problems with that as well

Capital π Π

$\Pi(\cdot \mid \vec{x})$ is the posterior probability measure. Just like $\mathbb{P}(\cdot)$, but Bayesian!

Interpretation: we have evidence against $H_0: \theta = \theta_0$ if θ_0 is in a region of low posterior probability (i.e., regions where $\pi(\cdot \mid \vec{x})$ is small)

Bayesian p-Values Aren't Great

- Example 6.18: Suppose $\pi(\theta \mid \mathbf{x}) = \text{Beta}\,(2,1)$. Compute Bayesian p-values for $H_0 : \theta = \frac{3}{4}$ under the posterior of $\theta \mid \mathbf{x}$ and the posterior of $\theta^2 \mid \mathbf{x}$.

$\pi(\theta|\vec{x}) = 2\theta$ for $\theta \in (0,1)$. Now, $\pi(\theta|\vec{x}) \leq \pi(3/4|\vec{x})$
$\Rightarrow 2\theta \leq 2 \cdot 3/4$
$\Rightarrow \theta \leq 3/4$

So our Bayesian p-value is $\Pi\left((0,3/4) \mid \vec{x} \right) = \int_0^{3/4} \pi(\theta|\vec{x})\,d\theta = \int_0^{3/4} 2\theta\,d\theta = 9/16$.

What about under $\pi(\theta^2|\vec{x})$? Then we're testing $H_0: \theta^2 = (3/4)^2 = 9/16$.

We can get $\theta^2|\vec{x} \sim \text{Beta}(1,1) = \text{Unif}(0,1)$, so $\pi(\theta^2|\vec{x}) = 1 \;\; \forall \theta^2 \in (0,1)$ $(\forall \theta \in (0,1))$
check!

But $1 \leq 1 \iff \pi(\theta^2|\vec{x}) \leq \pi(9/16|\vec{x})$. That's always true! So $\Pi(\{\theta : 1 \leq 1\} | \vec{x}) = 1$, regardless of $\vec{x}$. So there can never be any evidence against H_0!

Not so great...

Tweaking the Prior

- These issues happen when the prior $\pi(\theta)$ assigns zero probability to H_0, and can be avoided by tweaking the prior in such a way to fix this

- This isn't unreasonable; if we have reason to test $H : \theta \in A$, then we suspect it *could* be true, which would be contradicted if $\Pi(\theta \in A) = 0$

- If we start with a continuous prior π_2, we can create a new one using

$$\pi(\theta) = \alpha \cdot \pi_1(\theta) + (1 - \alpha) \cdot \pi_2(\theta),$$

General form of a "finite mixture distribution", whose pdf/pmf is of the form $f(x) = \sum_{j=1}^{k} \alpha_j \cdot f_j(x)$ where each $\alpha_j > 0$, $\sum_{j=1}^{k} \alpha_j = 1$, and each f_j is a pdf/pmf.

EXERCISE: show $f(x)$ is a valid pdf/pmf.

where π_1 is degenerate at θ_0 and $\alpha \in (0, 1)$

- This gives

$$\Pi(\{\theta_0\} \mid \mathbf{x}) = \frac{\alpha f_1(\mathbf{x})}{\alpha f_1(\mathbf{x}) + (1 - \alpha) f_2(\mathbf{x})},$$

where $f_i(\mathbf{x})$ is the prior predictive distribution under the prior π_i

Bayes Factors

In a general probability space $(\mathcal{F}, \Omega, \mathbb{P})$, the odds of an event $A \subseteq \Omega$ is/are defined as $\frac{\mathbb{P}(A)}{1-\mathbb{P}(A)}$

- There's a popular approach to Bayesian hypothesis testing involves the odds

- Definition 6.10: Let $\pi(\theta)$ be a prior, let $\mathbf{X} \sim f_\theta(\mathbf{x})$, and let $\pi(\theta \mid \mathbf{x})$ be the posterior for the model. Suppose that $H_0 : \theta \in \Theta_0$ and $H_A : \theta \in \Theta_0^c$ are two competing hypotheses about plausible values of θ.

  The **prior odds** in favour of H_0 is the ratio $\dfrac{\Pi(\Theta_0)}{\Pi(\Theta_0^c)} = \dfrac{\Pi(\Theta_0)}{1 - \Pi(\Theta_0)}$.

  The **posterior odds** in favour of H_0 is the ratio $\dfrac{\Pi(\Theta_0 \mid \mathbf{x})}{\Pi(\Theta_0^c \mid \mathbf{x})} = \dfrac{\Pi(\Theta_0 \mid \mathbf{x})}{1 - \Pi(\Theta_0 \mid \mathbf{x})}$.

  Provided that $\Pi(\Theta_0) > 0$, the **Bayes factor** in favour of H_0 is given by the ratio of the posterior odds to the prior odds:

$$BF_{H_0} = \frac{\Pi(\Theta_0 \mid \mathbf{x})}{1 - \Pi(\Theta_0 \mid \mathbf{x})} \Bigg/ \frac{\Pi(\Theta_0)}{1 - \Pi(\Theta_0)}.$$

Bayes Factors

- What's the point of Bayes factors? i.e., $r = \frac{\Pi(\Theta_0)}{1-\Pi(\Theta_0)} = \frac{\Pi(\Theta_0)}{\Pi(\Theta_0^c)}$

- For one, if we let r be the prior odds, then

$$\Pi(\Theta_0 \mid \mathbf{x}) = \frac{r \cdot BF_{H_0}}{1 + r \cdot BF_{H_0}}$$ EXERCISE: show!

- So a small/large Bayes factor means a small/large posterior probability of H_0

- Moreover, Bayes factors have a surprising connection to likelihood ratios

- Theorem 6.4: If we want to test $H_0 : \theta \in \Theta_0$ and we choose a prior mixture $\pi(\theta) = \alpha \cdot \pi_1(\theta) + (1-\alpha) \cdot \pi_2(\theta)$ such that $\Pi_1(\Theta_0) = \Pi_2(\Theta_0^c) = 1$, then

$$BF_{H_0} = \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})}.$$ free of θ (!)

Here f_i is the prior predictive distribution under π_i — i.e., $f_i(\vec{x}) = \int \pi_i(\theta) \cdot f_\theta(\vec{x})\, d\theta$.

Bayes Factors: Examples

- Example 6.19: Suppose that $X_1, X_2, \ldots, X_n \stackrel{iid}{\sim}$ Bernoulli (θ) and we place a Unif $(0,1)$ prior on θ. Compute the Bayes factor in favour of $H_0 : \theta = \theta_0$.

Let π_1 be degenerate at θ_0, so $\Pi_1(\{\theta_0\}) = 1$.

Let $\pi_2 = \text{Unif}(0,1)$, so $\Pi_2(\{\theta_0\}) = 0 \iff \Pi_2((0,\theta_0) \cup (\theta_0, 1)) = 1$

By Theorem 6.4, $BF_{H_0} = \dfrac{f_1(\vec{x})}{f_2(\vec{x})}$.

Prior predictive under π_1: Π_1 is degenerate at θ_0, so $f_1(\vec{x}) = \theta_0^{\sum x_i}(1-\theta_0)^{n-\sum x_i}$ (*)

Prior predictive under π_2: $f_2(\vec{x}) = \displaystyle\int_0^1 1 \cdot \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}\, d\theta = \frac{\Gamma(\sum x_i + 1)\cdot\Gamma(n - \sum x_i + 1)}{\Gamma(n+2)}$

$$\text{So } BF_{H_0} = \frac{\theta_0^{\sum x_i}(1-\theta_0)^{n-\sum x_i}}{\Gamma(\sum x_i+1)\cdot\Gamma(n-\sum x_i+1)/\Gamma(n+2)}.$$

(*) FYI: the "pdf/pmf" of a degenerate r.v. θ_0 is a "Dirac delta function" $\delta_{\theta_0}(\cdot)$ (not actually a function) which satisfies $\int_{-\infty}^{\infty} \delta_{\theta_0}(\theta) = 1$ and (informally) satisfies $\int \delta_{\theta_0}(\theta)\cdot g(\theta)\, d\theta = g(\theta_0)$ for any function $g(\cdot)$ $\Rightarrow f_1(\vec{x}) = \int \delta_{\theta_0}(\theta)\cdot\theta^{\sum x_i}(1-\theta)^{n-\sum x_i}\, d\theta$

Credible Intervals

- Assuming that $\Theta \subseteq \mathbb{R}$, what's a reasonable Bayesian analogue of confidence intervals?

- Now, it's perfectly reasonable to ask what the probability is that $l \leq \theta \leq u$ for $l, u \in \Theta$

- Definition 6.11: Let $\pi(\theta \mid \mathbf{x})$ be a posterior distribution on Θ. A $\boldsymbol{(1-\alpha)}$**-credible interval** for θ is an interval $[L(\mathbf{x}), U(\mathbf{x})] \subseteq \Theta$ such that

$$\Pi(L(\mathbf{x}) \leq \theta \leq U(\mathbf{x}) \mid \mathbf{x}) = \int_{L(\mathbf{x})}^{U(\mathbf{x})} \pi(\theta \mid \mathbf{x})\, \mathrm{d}\theta \geq 1 - \alpha.$$

- As with confidence intervals, there are usually plenty of credible intervals available for a given posterior, so we look for some desirable properties

Two Types of Credible Intervals

- Definition 6.12: If $\pi(\theta \mid \mathbf{x})$ is unimodal, the $(1-\alpha)$-credible interval $[L(\mathbf{x}), U(\mathbf{x})]$ such that the length $U(\mathbf{x}) - L(\mathbf{x})$ is minimized is called the **$(1-\alpha)$-highest posterior density (HPD) interval** for θ

- An HPD interval really does capture the most likely values in Θ, since any region outside of it will be assigned a lower posterior probability

- Definition 6.13: The $(1-\alpha)$-credible interval $[L(\mathbf{x}), U(\mathbf{x})]$ which satisfies

$$\Pi((-\infty, L(\mathbf{x})] \mid \mathbf{x}) = \Pi([U(\mathbf{x}), \infty) \mid \mathbf{x}) = \alpha/2$$

  is called the **$(1-\alpha)$-equal tailed interval (ETI)** for θ

- An ETI exists for any continuous posterior, unimodal or otherwise

- One can show that if $\pi(\theta \mid \mathbf{x})$ is symmetric, unimodal, and continuous, then the HPD interval and the ETI will be equal

Credible Intervals: Examples

- Example 6.20: Suppose that $X_1, X_2, \dots, X_n \overset{iid}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$ where σ^2 is known, and we place a $\mathcal{N}\left(\theta, \tau^2\right)$ prior on μ. What do $(1-\alpha)$-HPD intervals and ETIs for μ look like? What happens as $\tau^2 \to \infty$?

The posterior $\pi(\mu|\vec{x})$ is normal, which is continuous, unimodal, and symmetric. So the HPD and ETI intervals will be the same! From Example 6.8, $\mu|\vec{x} \sim N\left(\dfrac{\frac{\theta}{\tau^2}+\frac{n\bar{x}}{\sigma^2}}{\frac{1}{\tau^2}+\frac{n}{\sigma^2}}, \dfrac{1}{\frac{1}{\tau^2}+\frac{n}{\sigma^2}}\right)$.

We need
$$1-\alpha = \Pi\left(z_{1-\alpha/2} < \left(\mu - \frac{\frac{\theta}{\tau^2}+\frac{n\bar{x}}{\sigma^2}}{\frac{1}{\tau^2}+\frac{n}{\sigma^2}}\right) \Big/ \left(\frac{1}{\tau^2}+\frac{n}{\sigma^2}\right)^{-\frac{1}{2}} < z_{\alpha/2} \;\Big|\; \vec{x}\right)$$
$$= \Pi\left(\frac{\frac{\theta}{\tau^2}+\frac{n\bar{x}}{\sigma^2}}{\frac{1}{\tau^2}+\frac{n}{\sigma^2}} + z_{1-\alpha/2}\cdot\left(\frac{1}{\tau^2}+\frac{n}{\sigma^2}\right)^{-\frac{1}{2}} < \mu < \frac{\frac{\theta}{\tau^2}+\frac{n\bar{x}}{\sigma^2}}{\frac{1}{\tau^2}+\frac{n}{\sigma^2}} + z_{\alpha/2}\left(\frac{1}{\tau^2}+\frac{n}{\sigma^2}\right)^{-\frac{1}{2}} \;\Big|\; \vec{x}\right)$$

So our $(1-\alpha)$-credible intervals are both
$$\left[\frac{\frac{\theta}{\tau^2}+\frac{n\bar{x}}{\sigma^2}}{\frac{1}{\tau^2}+\frac{n}{\sigma^2}} + z_{1-\alpha/2}\left(\frac{1}{\tau^2}+\frac{n}{\sigma^2}\right)^{-\frac{1}{2}},\ \frac{\frac{\theta}{\tau^2}+\frac{n\bar{x}}{\sigma^2}}{\frac{1}{\tau^2}+\frac{n}{\sigma^2}} + z_{\alpha/2}\left(\frac{1}{\tau^2}+\frac{n}{\sigma^2}\right)^{-\frac{1}{2}}\right]$$

What happens as $\tau^2 \to \infty$? (i.e., as the prior becomes improper?)

Then that $(1-\alpha)$-credible interval becomes $\left(\bar{x} + z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}},\ \bar{x} + z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}\right)$... it's our z-interval!

Credible Intervals: Examples

- Example 6.21: Suppose that $X_1, X_2, \ldots, X_n \overset{iid}{\sim}$ Poisson (λ) and we place a Gamma (α, β) prior on λ. What do 95% HPD intervals and ETIs for λ look like?

From Example 6.3, $\lambda | \vec{x} \sim \text{Gamma}(\alpha + \sum x_i, \beta + n)$.

Let $G(\cdot | \vec{x})$ be the cdf of that thing.

ETI: need $\underbrace{\alpha/2 = \Pi\left((-\infty, L(\vec{x})] \,\middle|\, \vec{x} \right)}_{\Rightarrow \alpha/2 = G(L(\vec{x}) | \vec{x}) \;\Rightarrow\; L(\vec{x}) = G^{-1}(\alpha/2 | \vec{x})} = \underbrace{\Pi\left([U(\vec{x}), \infty) \,\middle|\, \vec{x} \right) = \alpha/2}_{\Rightarrow G(U(\vec{x}) | \vec{x}) = 1 - \alpha/2 \;\Rightarrow\; U(\vec{x}) = G^{-1}(1-\alpha/2 | \vec{x})}$

So our $(1-\alpha)$-ETI is $\left[G^{-1}(\alpha/2 | \vec{x}), G^{-1}(1-\alpha/2 | \vec{x}) \right]$.

HPD: impossible to do by hand! Gotta use a statistical software package (or maybe simulation) to estimate this.

ETIs are Invariant

- We've seen that posterior distributions can do unexpected things when we're interested in inferences of $\tau(\theta)$
- In general, a credible interval for θ may tell us nothing about a credible interval (or credible region) for $\tau(\theta)$
- But ETIs have a special property that bypasses this issue
- Theorem 6.5: ETIs are invariant under monotone transformations of θ, in the sense that if $(L(\mathbf{x}), U(\mathbf{x}))$ is a $(1-\alpha)$-ETI for θ and $\tau : \Theta \to \mathbb{R}$ is monotone increasing, then $(\tau(L(\mathbf{x})), \tau(U(\mathbf{x})))$ is a $(1-\alpha)$-ETI for $\tau(\theta)$. If τ is monotone decreasing, everything flips!

Proof. If $\Pi((-\infty, L(\vec{x})] \mid \vec{x}) = \Pi([U(\vec{x}), \infty) \mid \vec{x}) = \alpha/2$, then

$\Pi((-\infty, \tau(L(\vec{x}))] \mid \vec{x}) = \Pi([\tau(U(\vec{x})), \infty) \mid \vec{x}) = \alpha/2 \Rightarrow [\tau(L(\vec{x})), \tau(U(\vec{x}))]$ is a $(1-\alpha)$-ETI for $\tau(\theta)$. $\square$

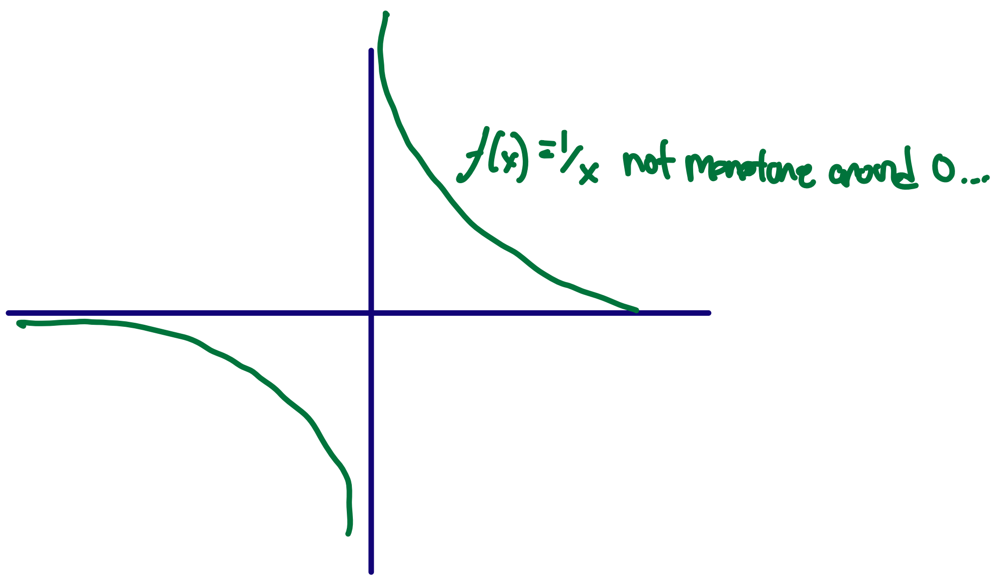
- Example 6.22:

For the $N(\mu, \sigma^2)$ model, a $(1-\alpha)$-ETI for μ^3 is given by...

Poll Time!

$$\left[\left(\frac{\frac{\theta}{\tau^2}+\frac{n\bar{x}}{\sigma^2}}{\frac{1}{\tau^2}+\frac{n}{\sigma^2}} + z_{1-\alpha/2}\left(\frac{1}{\tau^2}+\frac{n}{\sigma^2}\right)^{-\frac{1}{2}}\right)^3, \left(\frac{\frac{\theta}{\tau^2}+\frac{n\bar{x}}{\sigma^2}}{\frac{1}{\tau^2}+\frac{n}{\sigma^2}} + z_{\alpha/2}\left(\frac{1}{\tau^2}+\frac{n}{\sigma^2}\right)^{-\frac{1}{2}}\right)^3\right]$$

On Quercus: Module 6 - Poll 5



The Bernstein-von Mises Theorem

- Bayesian and frequentist inference unite in this monumental result

- Theorem 6.6 (**Bernstein-von Mises**): Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} f_{\theta_0}$, let $\pi(\theta)$ be a prior distribution placing positive mass around θ_0, and let $\theta_n \sim \pi(\theta \mid \mathbf{x}_n)$. Under suitable regularity conditions,

$$\sqrt{n}\left(\theta_n - \hat{\theta}_{\mathsf{MLE}}(\mathbf{x}_n)\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{I_1(\theta_0)}\right).$$

- This statement is a *vast* simplification of the actual Bernstein-von Mises theorem, but it preserves the essence

FYI: the actual mode of convergence is "convergence in total variation", which implies convergence in probability (and hence in distribution)

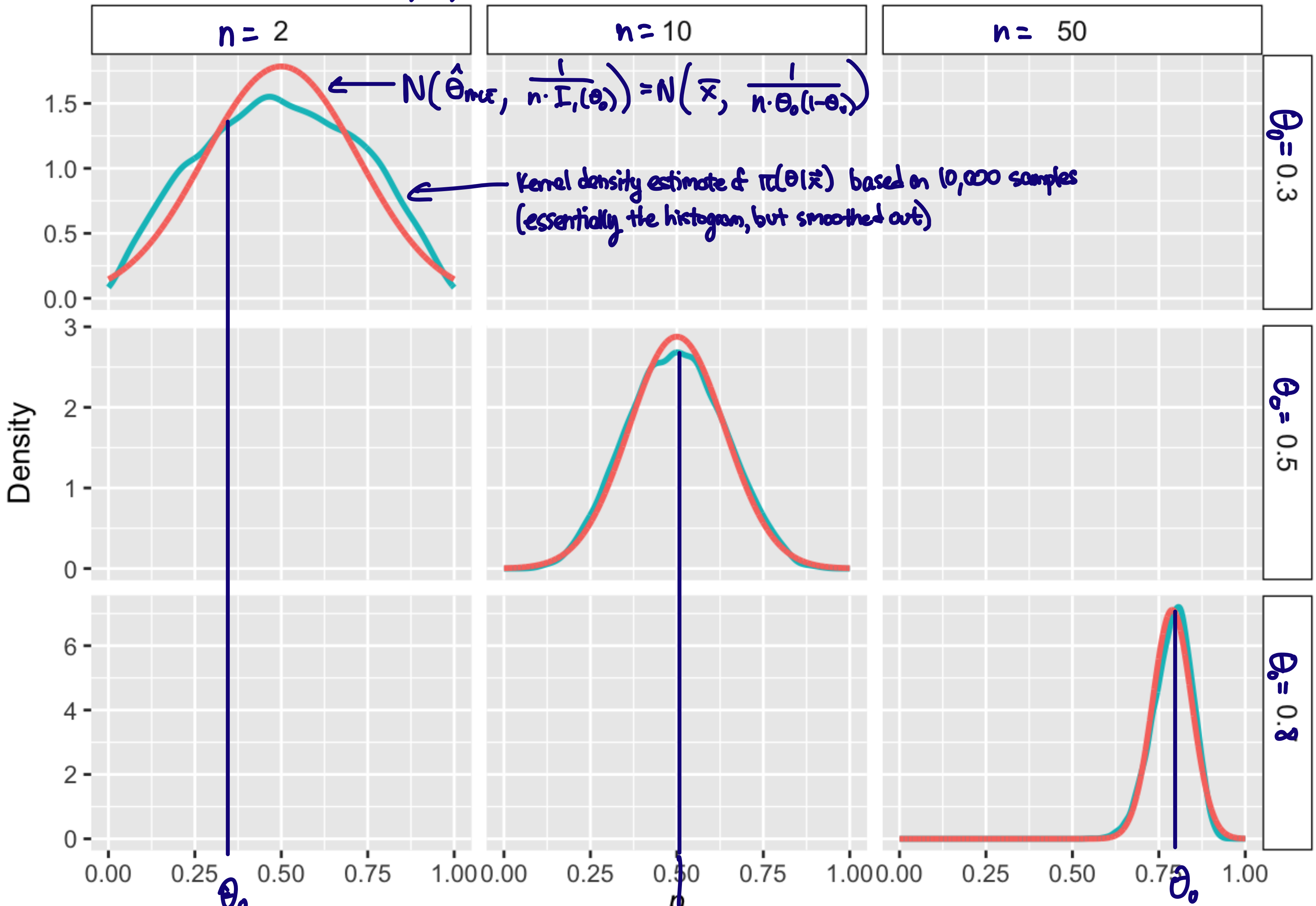
- The takeaway is that as the sample size of our data n gets larger, the choice of $\pi(\theta)$ matters less and the likelihood dominates

The posterior tends to center around the MLE... but the MLE tends to approach θ_0

- Roughly speaking, the posterior $\pi(\theta \mid \mathbf{x}_n)$ converges to a degenerate distribution on θ_0, for *any* well-behaved prior (!)

The Bernstein-von Mises Theorem: It's True

$X_1, \ldots, X_n \sim \text{Bernoulli}(\theta)$, $\pi(\theta) = \text{Beta}(1,1) = \text{Unif}(0,1)$



The End ?! :(