STA261 - Module 6

Bayesian Statistics

Rob Zimmerman

University of Toronto

August 9-11, 2022

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

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

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

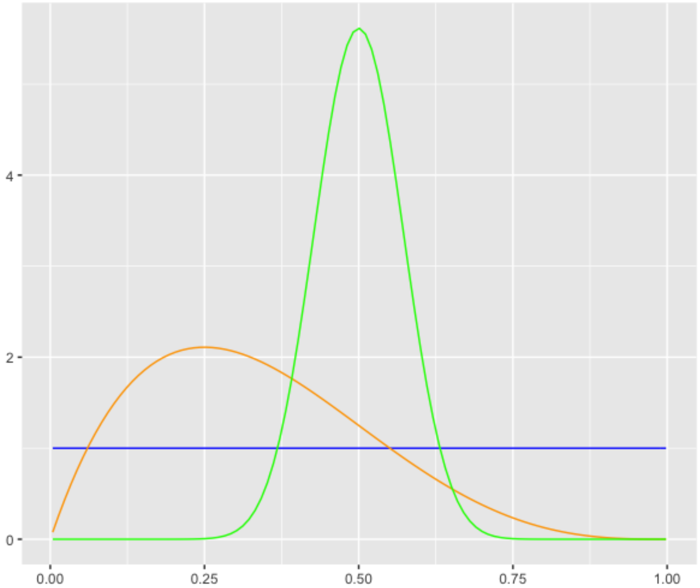
- If $X = \mathbb{1}_{\text{heads}}$, then we want to make inferences about the random variable p, where $X \mid p \sim \text{Bernoulli}\,(p)$

- 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

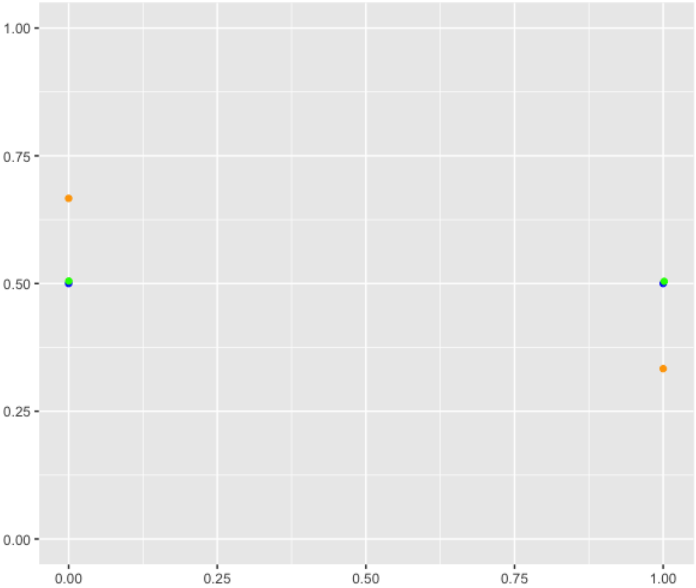
$$\int_{\Theta} \mathbb{P}_p(X = 1) \cdot \pi(p)\, \mathrm{d}p = \int_0^1 p \cdot \pi(p)\, \mathrm{d}p$$

- 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

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

Prior Predictive Distributions for the Coin Example

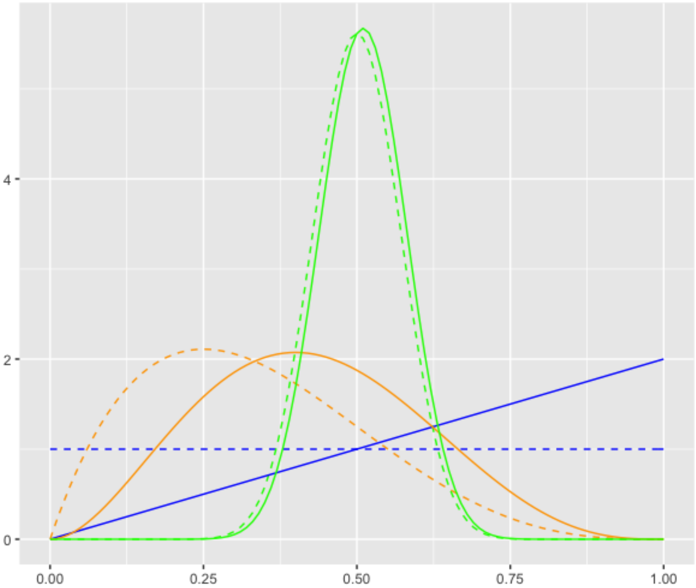


The Posterior Distribution - A Motivation

- 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

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

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 $(\mathbf{X}, \theta)$

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

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

- 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

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

Poll Time!

More on the Posterior

- The posterior $\pi(\theta \mid \mathbf{x})$ is a function of θ, and the data $\mathbf{x}$ is *observed*
- So we could write $\pi(\theta \mid \mathbf{x}) \propto f_\theta(\mathbf{x}) \cdot \pi(\theta)$
- 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 $f_\theta(x) \cdot \pi(\theta) \propto \theta^{x+1}(1-\theta)^{2-x}$ for $\theta \in (0,1)$. What is $\pi(\theta \mid x)$?

More on the Posterior

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

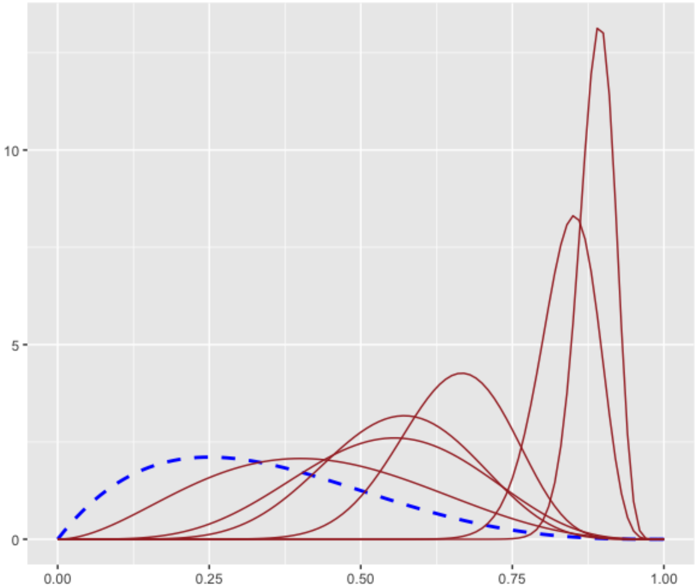
- Consider three different priors:
  - π_1 is highly concentrated at $\theta_1 \in \Theta$
  - π_2 is highly concentrated at $\theta_2 \in \Theta$
  - π_3 is Unif (Θ)

- 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?
  - $\pi_1(\cdot \mid \mathbf{x})$
  - $\pi_2(\cdot \mid \mathbf{x})$
  - $\pi_3(\cdot \mid \mathbf{x})$

- 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



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 $\pi(p \mid \mathbf{x})$.

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 $\pi(\lambda \mid \mathbf{x})$.

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 $\pi(\theta \mid \mathbf{x}) = \pi(\theta \mid T(\mathbf{x}))$.

Computing Posteriors: Examples

- 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 $\pi(p \mid \sum_{i=1}^{n} x_i)$.

Hyperparameters

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

- 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)

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

- 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...

Poll Time!

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
- Some priors are simply not sensible given the parametric family for the data
- Example 6.5:
- 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:

- Example 6.7:

Conjugate Priors

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

Conjugate Priors

- 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:

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

  If we choose an exponential family prior of the form

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

  where ν and $\eta_1, \ldots, \eta_k$ are hyperparameters, then $\pi(\theta)$ is a conjugate prior for f_θ.

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:

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!

Elicitation: Examples

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

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 $\mathbb{R}$
- And yet, if we take $\pi(\mu) = 1$,

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**.

- Example 6.11:

- 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., $\int_\Theta \pi(\theta) \, \mathrm{d}\theta$ is divergent), then it is called an **improper prior**.

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

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

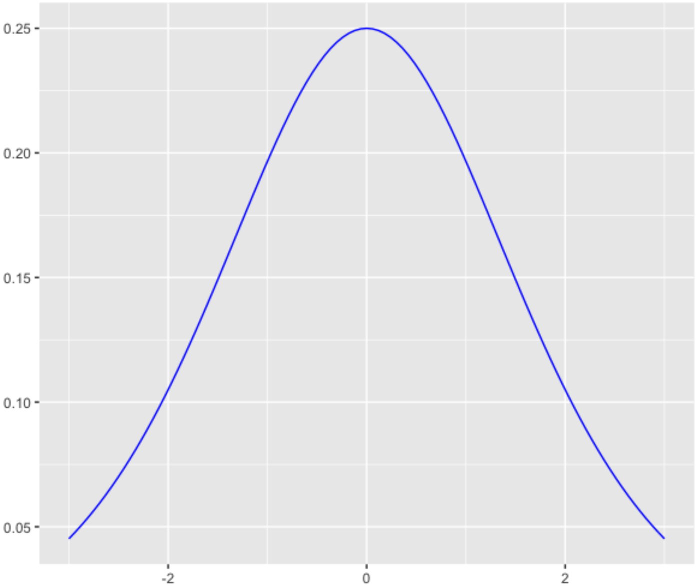
Problems With Uninformative Priors

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

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 $\tau = \log\left(\frac{p}{1-p}\right)$?

Oh No



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 $\pi_\theta^J(\theta) \propto \sqrt{I_1(\theta)}$.

- 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

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

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

Proof.

Jeffreys' Prior: Examples

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

Jeffreys' Prior: Examples

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

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

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?
- 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:
$$\hat{\theta}_{\mathsf{MAP}}(\mathbf{X}) = \operatorname*{argmax}_{\theta \in \Theta} \pi(\theta \mid \mathbf{X}).$$

- 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:

$$\hat{\theta}_{\mathsf{B}}(\cdot) = \underset{T(\cdot)}{\operatorname{argmin}} \, \mathbb{E}\left[\mathsf{MSE}_{\theta}\left(T(\mathbf{X})\right)\right]$$

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.

Bayesian Point Estimation: Examples

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

Poll Time!

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

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 $\theta^2 \mid \mathbf{x}$.

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

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

  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

- 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?

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

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

- 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

$$BF_{H_0} = \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})}.$$

Bayes Factors: Examples

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

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 $(\mathbf{1} - \boldsymbol{\alpha})$**-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 $\mathbf{(1-\alpha)}$**-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 $\mathbf{(1-\alpha)}$**-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 95% HPD intervals and ETIs for μ look like? What happens as $\tau^2 \to \infty$?

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?

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 $\tau(\theta)$.

Proof.

- Example 6.22:

Poll Time!

The Bernstein-von Mises Theorem

- Bayesian and frequentist inferences 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 on θ, 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

- 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

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