# Mathematical Statistics II

## The Bayesian Approach to Parameter Estimation

Jesse Wheeler

## Contents

## 1 Introduction

### Bayesian Estimation

- Much of this work is based on Rice (2007, Section 8.6).
- We have already discussed the philosophy of Bayesian statistics.
- We start with a prior belief about parameter values, and update these beliefs using observed data.
- The resulting *distribution* is called the *posterior*, and it represents our updated belief after observing data.
- This is very natural idea that is closely related to the idea of likelihood: likelihood quantifies some degree of belief about a parameter value.

## 2 Review

### Some Review

- Before we begin, we will first do a bit of review.

- In the context of Bayesian inference, we treat unknown parameter vectors as random variables, which I will denote $\Theta$.

- Thus, our probability model can be expressed as $f(x|\Theta = \theta)$, which we often shorten to $f(x|\theta)$.

**Bayes' Theorem**
Let $X$ be the random vector representing observed data, and $\Theta$ the random parameter vector, and $x^*$ the observed data. Bayes Theorem states:

$$\begin{aligned}\pi_{\Theta|X}(\theta|x^*) &= \frac{f_{X|\Theta}(x^*|\theta)\pi_\Theta(\theta)}{f_X(x^*)} \\ &= \frac{f_{X|\theta}(x^*|\theta)\pi_\Theta(\theta)}{\int f_{X|\Theta}(x^*|\tau)\pi_\Theta(\tau)\,d\tau}\end{aligned}$$

- There are a few things to note in the equation above. First, the likelihood function $L(\theta) = f(x^*|\theta)$ makes its presence on the right hand side of the equation.

- Next, the denominator is not a function of $\theta$. As a result, it is just a normalizing constant to ensure that the posterior is a proper probability distribution.

- With this in mind, we often say that the posterior distribution $\pi_{\Theta|X}(\theta|x^*)$ is a product of the likelihood $L(\theta)$ and the prior $\pi_\Theta(\theta)$.

- There is a large number of notations that are often used. For instance, the symbol $f$ is often used instead of $\pi$ as a function. The most common is perhaps:

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)\,d\theta}.$$

- The above notation does a bit of "function overload", but it is often clear from context and the symbols used as input what is meant.

- As before, $f$ is taken to be either a pmf or pdf, depending on the problem.

*Flipping 10 coins*
Our friend hands us a coin from another country, and we want to estimate $\theta = p$, the probability that the coin lands heads. Suppose we flip a coin 10 times, and see $n$ heads. Find a Bayesian estimate for $\theta$.

- The probability model describing the data is Binomial($N = 10, p = \theta$), which has mass function:

$$f_{X|\Theta}(n|\theta) = \binom{N}{n}\theta^n(1-\theta)^{N-n}.$$

- After picking the model for the data $f(x|\theta)$, the next step is to define our prior belief about the coin, characterized by $\pi_\Theta(\theta)$.

- A natural prior might be: "I know nothing about the coin, all probabilities are possible". Then, our prior would be uniform:

$$\pi_\Theta(\theta) = 1(0 \leq \theta \leq 1)$$

- Thus, we previously found the posterior to be:

$$\begin{aligned}\pi(\theta|n) &= \frac{\binom{N}{n}\theta^n(1-\theta)^{N-n}}{\int_0^1 \binom{N}{n}\theta^n(1-\theta)^{N-n}\,d\theta}1[0 \leq \theta \leq 1] \\ &= \frac{\Gamma(N+2)}{\Gamma(n+1)\Gamma(N-n+1)}\theta^n(1-\theta)^{N-n}1[0 \leq \theta \leq 1]\end{aligned}$$

- We then demonstrated that this corresponds to a beta distribution. Thus:

$$\Theta|X = n \sim \text{Binom}(\text{Beta}(n+1, N-n+1))$$

- Now our belief about $\Theta$ has been updated using the data $X = n$. This belief is represented not be a single point, but an entire distribution.
- There are multiple ways to get a single point estimation.
- One idea is the mean of the posterior, $E[\Theta|X = n]$. In this case, the mean of the Beta distribution is known, and we find:

$$E[\Theta|X = n] = \frac{n+1}{N+2}.$$

  As mentioned, this is like a regularized version of the MLE ($n/N$), as the estimate is "pulled" toward the center $\theta = 0.5$.

- Another common approach is similar to what we did with the MLE: if the posterior represents our updated belief as a distribution, why don't we let our point estimate be the *maximum* of that belief? In this setting, the maximum of a probability density is the *mode*. The mode of the Beta($\alpha, \beta$) distribution is given by:

$$\frac{\alpha - 1}{\alpha + \beta - 2}.$$

- For our specific posterior, this implies that the mode is:

$$\hat{\theta} = \frac{(n+1) - 1}{(n+1) + (N-n+1) - 2} = \frac{n}{M}.$$

- Note that the mode in this case matches the MLE!
- The mode of the posterior distribution is called the Maximum A Posteriori (MAP) estimate. This is a common choice for a point estimate, in particular by Frequentists who use Bayesian methods to solve a given problem. There are some advantages and disadvantages of this approach, one being that it is not a properly weighted version of our belief; it also lacks some of the proprieties and guarantees that Bayesian statisticians like. The MAP can also be difficult (or impossible) to compute in many situations, whereas the posterior mean and median can readily be approximated using samples from a distribution.
- Another possible point estimate is the median of the posterior distribution. There's not a closed-form expression for the median of a Beta-distribution, but it can be calculated via software.
- Even in the simple problem above, we see two of the primary challenges with Bayesian parameter estimation:

- How do we choose the prior distribution $\pi(\theta)$? A generally safe and accepted approach is a uniform prior. However, this formally only exists if $\theta$ is bounded, which is not always the case. Also, it represents a prior belief: given a new coin, do we really think all values of $p$ are equally likely, or maybe values close to $p = 0.5$ are more likely than extreme values $p = 0, 1$? Since the prior represents our beliefs about $\theta$, is a uniform prior actually appropriate? If it isn't appropriate, how exactly should we specify the prior?
- Even in this very simple model and prior, the denominator $f(x)$ was difficult to compute. What about more complex models and priors? A large amount of Bayesian computation and theory is dedicated to solving this problem.

**Proposition: the MAP and MLE**
Let $\theta$ be a parameter of interest, and $x^*$ the observed data. If our prior distribution is proportional to 1, i.e., $\pi(\theta) \propto 1$ (which is effectively a uniform prior on a bounded interval), then

$$\hat{\theta}_{\text{MAP}} = \hat{\theta}_{\text{MLE}}.$$

- A proof sketch is given in class, but is left as an exercise in these notes.

- This is true for the Coin-tossing example; look back at the likelihood function and posterior, and use R to plot them both.

# 3 Examples

### Bayesian point-estimate examples

*Poisson model*
Suppose we have observations $n$ observations, which we wish to model as IID Poisson($\lambda$). Find a Bayesian estimate of $\Lambda = \lambda$ given the observed data $x^*$.

- First we find the density of $X|\Lambda = \lambda$, which begins with finding the density of a single observation $X_i|\Lambda = \lambda$:

$$f_{X_i|\Lambda}(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x \in \{0, 1, 2, \ldots\}.$$

- Under the IID assumption, the joint density of $X$ is:

$$\begin{aligned} f_{X|\Lambda}(x|\lambda) &= \prod_{i=1}^{n} f_{X_1|\Lambda}(x_i|\lambda) \\ &= \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \\ &= \frac{\lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda}}{\prod_{i=1}^{n} x_i!}. \end{aligned}$$

- Now we want to find the posterior of $\Lambda|X$, which is given by:

$$f_{\Lambda|X}(\lambda|x^*) = \frac{\lambda^{\sum_{i=1}^{n} x_i^*} e^{-n\lambda} f_\Lambda(\lambda)}{\int \lambda^{\sum_{i=1}^{n} x_i^*} e^{-n\lambda} f_\Lambda(\lambda)\, d\lambda}.$$

  Above, $f_\Lambda(\lambda)$ is the prior distribution of $\Lambda$, and the product $\prod_{i=1}^{N} x_i!$ canceled out as it was in both the numerator and denominator.

- Now there are two remaining steps, the parts that are often challenging in Bayesian estimations: (1) choosing the prior, (2) computing the integral in the denominator.

- We will consider two different approaches for picking the prior. The first is the traditional / orthodox Bayesian, who takes very seriously the philosophy that the prior distribution captures their prior opinion.

- In this orthodox approach, the prior density $f_\Lambda(\lambda)$ should be specified *before* ever seeing the data (the whole point is this is our prior belief before observing data).

- This itself is not an easy task; even in this scenario, we may pick a prior based both on belief, and convenience.

- That is, suppose that we believe the prior mean $E[\Lambda] = \mu = 15$, with variance $\text{Var}(\Lambda) = \sigma^2 = 25$. There's a lot of distributions out there that have these features, but we will pick the Gamma distribution because it will be mathematically convenient.

- Since the Gamma$(\alpha, \beta)$ has mean $15 = \alpha/\beta$ and variance $25 = \alpha/\beta^2$, we can solve and get our prior density for $\Lambda$ as:
$$\Lambda \sim \text{Gamma}(\alpha = 9, \beta = 3/5).$$

- Note that the choice of the mean and variance can (and should) be aided by plotting the function, and typically the Gamma distribution has parameter values $\alpha$ and $\lambda$, but we're already using $\lambda$ for something else here.

- Thus, the prior density is:
$$f_\Lambda(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}, \quad \lambda > 0.$$

- After canceling constants (anything not involving $\lambda$) and combining like-terms, we get:
$$f_{\Lambda|X}(\lambda|x^*) = \frac{\lambda^{\sum_{i=1}^n x_i^* + \alpha - 1} e^{-(n+\beta)\lambda}}{\int_0^\infty \lambda^{\sum_{i=1}^n x_i^* + \alpha - 1} e^{-(n+\beta)\lambda}\, d\lambda}$$

- Now we encounter a common trick (and the reason we picked a Gamma prior). Note that the denominator is *only* a function of $x$, and not $\lambda$. Thus,
$$\begin{aligned} f_{\Lambda|X}(\lambda|x^*) &\propto \lambda^{\sum_{i=1}^n x_i^* + \alpha - 1} e^{-(n+\beta)\lambda} \\ &= \lambda^{\text{something}} e^{-\text{something}\lambda}. \end{aligned}$$

  Here, $\lambda$ is the variable of interest (not a constant). Thus, we want to compare this statement to other *kernels* that look like:
$$f(x) \propto x^a e^{-b\lambda}.$$

- This kernel matches that of the standard Gamma$(\gamma, \zeta)$ distribution:
$$f(x) \propto x^{\gamma-1} e^{-\zeta x}$$

  (again, swapping variables $\gamma$, $\zeta$ for $\alpha$ and $\lambda$ for obvious reasons).

- We can immediately conclude that the posterior MUST be a Gamma distribution (since it will integrate to one). What is left is to pick the corresponding parameter values. To do this, we must have:
$$\gamma - 1 = \sum_{i=1}^n x_i^* + \alpha - 1 \implies \gamma = \sum_{i=1}^n x_i^* + \alpha,$$

and

$$-\zeta = -(n+\beta) \implies \zeta = n+\beta.$$

Thus, the posterior distribution is:

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

- Then all we need to do is plug is our specific data values $x_i^*$, and the specific values of our prior $\alpha = 9$, $\beta = 3/5$.
- From this, we can get various point estimates: posterior mean, MAP, or posterior median. We can also talk about posterior variance if we want.
- **This trick of avoiding calculating the integral in the denominator is extremely common, and it will appear again. Make sure this makes sense to you**.

### Real-data example: Poisson Distribution

- Now let's look at a real-data example. These data are the 23 observations from the asbestos-filter problem.

```
x <- c(
  31, 29, 19, 18, 31, 28, 34, 27, 34, 30, 16, 18,
  26, 27, 27, 18, 24, 22, 28, 24, 21, 17, 24
)
x
```

```
 [1] 31 29 19 18 31 28 34 27 34 30 16 18 26 27 27 18 24 22
[19] 28 24 21 17 24
```

*Comparing Estimates*
Using the data above, compare estimates using the MoM, MLE, and the Bayesian approach with a Gamma prior. Also, discuss the corresponding errors related to these estimates.

## 4 Conjugate Priors

### Conjugate priors

- The first approach to the Poisson($\lambda$) example was the traditional (subjective) Bayesian, who takes seriously the choice of prior, and chose a Gamma density to aid computations.
- This approach was aided by the choice of a Gamma prior, which helped the calculation.
- This type of prior is known as a *conjugate prior*.

**Definition: Conjugate priors**
Suppose the prior distribution belongs to a family of distributions, $G$, and the data come from a family of distributions $H$.
$G$ is said to be conjugate to $H$ if the posterior is in the family $G$.

- Example: If the data-model is Poisson($\lambda$), then the family $H$ is the family of Poisson distributions. The Gamma family ($G$) of distributions is conjugate to the Poisson family, because if Gamma is selected as the prior distribution, then the posterior distribution (under data model $H$) is still Gamma ($G$), with updated parameters.
- Much of the Bayesian statistics of the 20th century relied on conjugate priors to help with integration, or were confined to models with very few parameters.
- Recent developments in computing, both hardware, software, and theory of Bayesian computing, has enabled fitting much more complex models using arbitrary priors.
- Still, it's worth discussing conjugate priors, and we will provide a few examples.

*Conjugate Normals*
Model $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. Treating $\sigma^2$ as fixed, consider the prior for $\mu \sim N(\mu_0, \sigma_0^2)$. Find the posterior of $\mu | X = x^*$.

*Proof.* The likelihood function is given by:

$$\begin{aligned} f(x^*|\mu, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x_i^* - \mu)^2} \\ &= \sigma^{-n}(2\pi)^{-n/2} \exp\Big\{ -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i^* - \mu)^2\Big\}, \end{aligned}$$

and the prior for $\mu$ is:

$$\pi(\mu) = \frac{1}{\sigma_0\sqrt{2\pi}} \exp\big\{ -\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\big\}.$$

Thus, the posterior is proportional to:

$$\begin{aligned} \pi(\mu|x^*) &\propto \sigma^{-n} \exp\Big\{ -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i^* - \mu)^2\Big\} \times \sigma_0^{-1} \exp\big\{ -\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\big\} \\ &\propto \exp\Big\{ -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i^* - \mu)^2\Big\} \times \exp\big\{ -\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\big\}. \end{aligned}$$

In the last step, we note that we are only interested in the parameter $\mu$; other parameters that arise can be treated as constants with respect to $\mu$.
We now apply a common trick for Normal distributions, which is combining the exponential terms and completing the square. Using the posterior expression above, we have:

$$\begin{aligned} \pi(\mu|x^*) &\propto \exp\Big\{ -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i^* - \mu)^2 - \frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\Big\} \\ &= \exp\Big\{ -\frac{1}{2\sigma^2}\big[\sum_{i=1}^{n}\left(x_i^*\right)^2 - 2n\mu\bar{x}^* + n\mu^2\big] - \frac{1}{2\sigma_0^2}\big[\mu^2 - 2\mu_0\mu + \mu_0^2\big]\Big\} \end{aligned}$$

Now recall that any factor not involving $\mu$ is a constant (with respect to $\mu$). Thus, we can further simplify be removing a term from each of the inner exponential arguments:

$$\pi(\mu|x^*) \propto \exp\Big\{ -\frac{1}{2\sigma^2}\big[ -2n\mu\bar{x}^* + n\mu^2\big] - \frac{1}{2\sigma_0^2}\big[\mu^2 - 2\mu_0\mu\big]\Big\}$$

Now we want to combine like terms in the exponential. In particular, we have terms for $\mu^2$ and $\mu$:

$$\begin{aligned}-\frac{1}{2\sigma^2}n\mu^2-\frac{1}{2\sigma_0^2}\mu^2&=\Big(-\frac{1}{2\sigma_0^2}-\frac{n}{2\sigma^2}\Big)\mu^2\\&=-\frac{1}{2}\Big(\frac{n\sigma_0^2+\sigma^2}{\sigma^2\sigma_0^2}\Big)\mu^2\\&=-\frac{1}{2}a\mu^2\end{aligned}$$

Similarly, for the $\mu$ terms in the exponent,

$$\begin{aligned}\frac{2n\mu\bar{x}^*}{2\sigma^2}+\frac{2\mu_0\mu}{2\sigma_0^2}&=\frac{n\bar{x}^*}{\sigma^2}\mu+\frac{\mu_0}{\sigma_0^2}\mu\\&=\Big(\frac{n\bar{x}^*}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\Big)\mu\\&=\Big(\frac{n\bar{x}^*\sigma_0^2+\mu_0\sigma^2}{\sigma^2\sigma_0^2}\Big)\mu\\&=b\mu=-\frac{1}{2}(2b\mu)\end{aligned}$$

Making the substitutions above for $a$ and $b$, we now have the posterior in the form:

$$\pi(\mu|x^*)\propto\exp\Big\{-\frac{1}{2}(a\mu^2-2b\mu)\Big\}.$$

Now the final step is to complete the square. Within the exponential, we have

$$\begin{aligned}-\frac{1}{2}(a\mu^2-2b\mu)&=-\frac{a}{2}(\mu^2-\frac{2b}{a}\mu)\\&=-\frac{a}{2}\big(\mu^2-\frac{2b}{a}\mu+\frac{b^2}{a^2}\big)+\frac{b^2}{2a}\\&=-\frac{a}{2}(\mu-\frac{b}{a})^2+(\text{some constant}).\end{aligned}$$

Thus,

$$\pi(\mu|x^*)\propto\exp\Big\{-\frac{a}{2}\Big(\mu-\frac{b}{a}\Big)^2\Big\}$$

This is directly proportional to the pdf of a Normal distribution for $\mu$, with mean $b/a$, variance $1/a$. Solving for these:

$$\begin{aligned}E[\mu|X=x^*]&=b/a\\&=\Big(\frac{n\bar{x}^*\sigma_0^2+\mu_0\sigma^2}{\sigma^2\sigma_0^2}\Big)/\Big(\frac{n\sigma_0^2+\sigma^2}{\sigma^2\sigma_0^2}\Big)\\&=\frac{n\bar{x}^*\sigma_0^2+\mu_0\sigma^2}{n\sigma_0^2+\sigma^2},\end{aligned}$$

and

$$\begin{aligned}\text{Var}(\mu|X=x^*)&=1/a\\&=\frac{\sigma^2\sigma_0^2}{n\sigma_0^2+\sigma^2}\end{aligned}$$

Giving the final posterior distribution:

$$\mu|X=x^*\sim N\Big(\frac{n\bar{x}^*\sigma_0^2+\mu_0\sigma^2}{n\sigma_0^2+\sigma^2},\frac{\sigma^2\sigma_0^2}{n\sigma_0^2+\sigma^2}\Big)$$

□

*Beta-Binomial conjugate relation*
One example that we have actually seen already is the Beta-Binomial distributions.
The Beta$(\alpha, \beta)$ distribution is conjugate to Binomial$(n, p)$. In the coin flipping example, we selected a Beta$(1, 1)$ prior.

- This example will be a HW problem.

**Posteriors and Likelihood**

- In the Poisson-Gamma model, we saw that we get very similar estimates using MLE or Bayesian approaches, regardless of which prior we picked.
- We can argue why this will often be the case, especially for IID data.
- Previously, we saw:
$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$
- When $n$ gets large, the likelihood dominates in this equation. In the IID case:
$$\text{likelihood} = \prod_{i=1}^{n} f(x_i^* | \theta).$$
- In particular, each new data point scales the likelihood larger and larger, to the point where the prior has little impact on the posterior distribution.
- See the accompanying Lecture 4 R code for a visual demonstration of this using the Poisson distribution.

**Uniform priors**

- The choice of conjugate priors is useful if we want to actually use a prior and a conjugate is available.
- A common alternative choice is a uniform prior.
- This is saying: we don't have any prior knowledge or belief about a parameter.
- A uniform prior is not always possible (e.g., $\lambda > 0$ has no uniform prior), but we can approximate it.

*Poisson posterior, uniform prior*
Revisit the Poisson$(\lambda)$ model, while taking the alternative approach of using a uniform prior.

- The setup for the problem is the exact same, so note that the posterior distribution is:
$$f_{\Lambda|X}(\lambda|x^*) = \frac{\lambda^{\sum_{i=1}^{n} x_i^*} e^{-n\lambda} f_\Lambda(\lambda)}{\int \lambda^{\sum_{i=1}^{n} x_i^*} e^{-n\lambda} f_\Lambda(\lambda)\, d\lambda}.$$
- Now suppose we really don't have a good guess for the parameter $\lambda$, or we want to more utilitarian / noncommittal approach. Now what?
- The default is a uniform probability, but the possible values of $\lambda$ is the interval $(0, \infty)$; we can't have a uniform prior on this interval (it doesn't exist)
- Instead, we will feign confidence that $\lambda$ must be smaller than some fixed number based on the problem at hand. For instance, maybe $0 < \lambda \leq 100$ is reasonable.

- Then, the prior would be:
$$f_\Lambda(\lambda) = \frac{1}{100}1(0 < \lambda \leq 100),$$
and the posterior would be:
$$\begin{aligned} f_{\Lambda|X}(\lambda|x^*) &= \frac{\frac{1}{100}\lambda^{\sum_{i=1}^n x_i^*} e^{-n\lambda}}{\frac{1}{100}\int_0^{100} \lambda^{\sum_{i=1}^n x_i^*} e^{-n\lambda}\, d\lambda} 1(0 < \lambda \leq 100) \\ &= \frac{\lambda^{\sum_{i=1}^n x_i^*} e^{-n\lambda}}{\int_0^{100} \lambda^{\sum_{i=1}^n x_i^*} e^{-n\lambda}\, d\lambda} 1(0 < \lambda \leq 100). \end{aligned}$$

- In this case, we can't just do the denominator-integration trick! It looks very similar, because the denominator is still a constant, and the kernel in the numerator looks very similar to a Gamma kernel, **but** we have a new bound $0 < \lambda \leq 100$ that makes it distinct from the Gamma distribution, since the Gamma distribution has support on $(0, \infty)$.

- Unfortunately, there is not an easy way to compute the integral either (partly because of the bound). Thus, the integral needs to be computed numerically.

- For now, we will just use the `integrate` function in R, which is really good at integrating univariate functions.

- The posterior mean can similarly be found using a numeric integration technique, and posterior mode can be found using numeric optimization strategies covered in the last set of lecture notes.

# 5 Introduction to Numeric Integration

### Numeric Integration

- As we saw in the previous examples, one of the primary challenges of Bayesian estimation is the integration in the denominator of the posterior.

- Bayesian statistics has really exploded since the late 20th century, largely thanks to improved computational tools that help with the numeric integration.

- For this set of lectures, we only briefly introduce this topic. Depending on time and interest, we can explore this topic more later in the semester.

- For univariate functions, there are numerous approaches to well-approximate an integral.

- Traditional approaches are very simply, and are often based on Riemann-sum approximations.

- In R, one reliable function is the `integrate` function.

- Consider the integral $f(x) = x^2$,
$$\int_0^3 x^2\, dx = 9$$

```
x_sq <- function(x) x^2
integrate(x_sq, lower = 0, upper = 3)

9 with absolute error < 1e-13
```

- Let $m$ be the number of points used to evaluate the integral:

$$\int_A f(x)\,dx.$$

- If $x$ is univariate, then the approach above is the take $m$ points and do a numeric approximation.
- The standard Riemann approximation can be shown to converge to the true value at rate $O(1/m)$ for univariate functions, and can even be improved to faster rates (Liu and Liu, 2001, Chatper 1).
- However, these approaches scale very poorly as the dimensions of $x$ and the integration area $A$ increase.
- For instance, suppose that $A$ is a 10-dimensional area (not even that large). Then, in order to achieve the $O(1/m)$ promised rate, you need to evaluate $O(m^{10})$ different points!
- The primary alternative approach is known as *Monte Carlo* approximation.
- Let $f(x;\theta)$ denote a pdf of some random variable, $X$. Then, if $I$ is the integral

$$I = E[g(x)] = \int_A g(x)f(x;\theta)\,dx,$$

then the law of large numbers states that

$$\hat{I}_m = \frac{1}{m}\sum_{i=1}^m g(X_i) \overset{a.s.}{\to} I,$$

where $X_i$ is sampled from the distribution with density $f(x;\theta)$.

- Because we have an average of samples, the CLT gives us a way to approximate the error:

$$\sqrt{m}(I_m - I) \overset{d}{\to} N(0,\sigma^2),$$

where $\sigma^2 = \mathrm{Var}\big(g(X)\big)$.

- Thus, the error rate of the Monte-Carlo method is $O(m^{-1/2})$, regardless of the dimension of $A$.
- The most common integral of this type is by letting $f$ be uniform over the area $A$, in which case $f(x) = \frac{1}{|A|}$, and

$$I = \int_A g(x)\,dx, \quad I_m = \frac{|A|}{m}\sum_{i=1}^m g(X_i),\ X_i \sim \mathrm{Uniform}(A).$$

- If $x$ is univariate, this approach is worse than standard deterministic approaches $O(1/m)$), but it has better performance in higher dimensional settings.
- Example: $f(x) = x^2$ on the region $A = [0, 3]$

```
set.seed(12345)
m <- 10000
X <- runif(n = m, 0, 3)
3 * mean(x_sq(X))

[1] 8.992983
```

- Using the CLT, we can get a standard error of this estimate. $\text{Var}(3X^2) = 64.8$, and therefore:

$$SE \approx \sqrt{\frac{64.8}{m}}.$$

```
sqrt(64.8/m)

[1] 0.08049845
```

```
# Numeric Approximation:
sd(3 * X^2) / sqrt(m)

[1] 0.07999898
```

- Theoretically, the Monte-Carlo approximation converges at a rate $O(m^{-1/2})$. There are two primary problems:
    1. The numerator in finite-sample approximations of the variance $\sigma^2/m$ might be very large.
    2. Drawing uniform samples from $A$ might be hard.
- The solution to these two problems is more advanced Monte-Carlo designs. We'll introduce *importance sampling*.
- Idea: not intervals of $x$ contribute equally the function $f(x)$ and it's integral.
- For instance, if $f(x) = x^2$, then values of $x$ close to zero mean that the function $f \approx 0$. However, values of $x$ near 3 have larger influence on the integral evaluation.
- Because of this, we don't need lots of samples from $x$ near zero, and we should focus more on samples near 3.
- We can do this mathematically:

$$\int g(x)f(x)dx = E_{X\sim f(x)}[g(X_i)] \approx \frac{1}{m}\sum_{i=1}^{m} g(X_i),$$

which is the same as

$$\begin{aligned}\int \frac{g(x)f(x)}{\pi(x)}\pi(x)\,dx &= E_{X\sim\pi(x)}\left[\frac{g(X)f(X)}{\pi(X)}\right] \\ &\approx \frac{1}{m}\sum_{i=1}^{m}\frac{g(X_i)f(X_i)}{\pi(X_i)}.\end{aligned}$$

- The above approximation looks more complicated, but it has several advantages.
- The ratio $f(X_i)/\pi(X_i) = w_i$ is called the *importance weight*.
- Now an important part of this is picking an appropriate sampling distribution $\pi(x)$.
- There is no "correct" way to do this, other than we want to have more samples that are concentrated in more "important" regions.
- We'll look at a couple concrete examples to make this more clear.

### Importance Sampling, Example 1

- Consider approximating the integral:

$$\int_0^3 x^2 \, dx$$

.

- We did the standard Monte-Carlo approach, using Uniform(0, 3) random variables:

```r
set.seed(12345)
m <- 10000
X <- runif(n = m, 0, 3)
3 * mean(x_sq(X))

[1] 8.992983
```

- We numerically approximated the standard error of this estimate to be:

```r
# Numeric Approximation:
sd(3 * X^2) / sqrt(m)

[1] 0.07999898
```

- Now let's try importance sampling. We don't want many low-values of $X$, and values should be between 0-3. Let's let $B_i \sim \text{Beta}(\alpha, \beta)$, and then $X_i = 3B_i$.
- After some checking, a good distribution might be:

$$X_i \sim 3 \times \text{Beta}(2, 1)$$

- A quick change-of-variables application gives:

$$\begin{aligned}
\pi(x) &= \frac{1}{3}\frac{(x/3)^{\alpha-1}(1-x/3)^{\beta-1}}{B(\alpha,\beta)}, \quad x \in [0,3] \\
&= \frac{1}{3}\frac{(x/3)}{B(2,1)}, \quad x \in [0,3] \\
&= \frac{2}{9}x, \quad x \in [0,3]
\end{aligned}$$

- Now that we have a sampling distribution $\pi(x)$, there are two different ways to think about the problem. The first is just a direct integral:

$$\int g(x)\,dx = \int \frac{g(x)}{\pi(x)}\pi(x)\,dx = E_{X\sim\pi(x)}\Big[\frac{g(X)}{\pi(X)}\Big],$$

or via the "target" distribution $f(x)$ approach:

$$\begin{aligned}
\int \frac{g(x)f(x)}{\pi(x)}\pi(x)\,dx &= E_{X\sim\pi(x)}\Big[\frac{g(X)f(X)}{\pi(X)}\Big] \\
&\approx \frac{1}{m}\sum_{i=1}^{m}\frac{g(X_i)f(X_i)}{\pi(X_i)}.
\end{aligned}$$

- Both are correct, and sometimes the second is a useful way to think about a problem.

- In this case, the first approach is obvious, and the second approach we would pick the target distribution $f$ to correspond to a uniform$(0, 3)$ distribution.
- Thus, approximating the integral:

```
m <- 10000
X <- 3 * rbeta(n = m, 2, 1)
pi_x <- function(x) 2*x/9
mean(x_sq(X)/pi_x(X))

[1] 8.993616
```

- We can once again get an estimate of the standard error using the CLT:

$$SE \approx \frac{\sigma^2_{w_i}}{\sqrt{m}},$$

  where $w_i$ is the importance weight.

```
sd( (X^2) / (2 * X / 9) ) / sqrt(m)

[1] 0.03181511
```

  This error was about half of the default Monte-Carlo method.

- This example is a bit trivial, because the integral is not hard to compute.
- Also, our choice of sampling distribution meant we could get a closed-form solution for sampling weights:

$$\begin{aligned} I_m &= \frac{1}{m}\sum_{i=1}^{m}\frac{X^2}{\frac{2X}{9}} \\ &= \frac{1}{m}\sum_{i=1}^{m}\frac{9X}{2} \\ &= \frac{9}{2}\bar{X}. \end{aligned}$$

- We could actually simplify this even further by picking a Beta(3, 1) sampling distribution:

$$\begin{aligned} \pi(x) &= \frac{1}{3}\frac{(x/3)^{\alpha-1}(1-x/3)^{\beta-1}}{B(\alpha,\beta)}, \quad x \in [0,3] \\ &= \frac{1}{3}\frac{(x/3)^2}{B(3,1)}, \quad x \in [0,3] \\ &= \frac{1}{9}x^2, \quad x \in [0,3] \end{aligned}$$

- Thus, if we picked this distribution, we could actually get the exact answer!

$$\begin{aligned} I_m &= \frac{1}{m}\sum_{i=1}^{m}\frac{X^2}{\frac{X^2}{9}} \\ &= \frac{9}{m}\sum_{i=1}^{m}1 \\ &= 9. \end{aligned}$$

- This leads us to the identity that if $\pi(x) \propto g(x)$, such that $g(x) = c\pi(x)$, then the integral is:

$$\int g(x)\, dx = \int g(x)/\pi(x)\pi(x)\, dx = c \int \pi(x)\, dx = c$$

- This effectively never happens in real worl-scenarios, because if it did there wouldn't be a reason to do Monte-Carlo.
- However, it does help us decide a useful principle: try to pick a $\pi(x)$ such that $\pi(x) \propto g(x)$ as much as possible.

### Importance Sampling, Example II

*Importance Sampling: extreme events*
If $Z \sim N(0, 1)$, approximate the probability:

$$P(Z > 3) = I = \int_3^\infty \frac{1}{2\sqrt{\pi}} e^{-x^2/2}\, dx.$$

- Once again, we have an exact method to calculate the integral, so we can see how good our approximation is:

```
# Actual value
1-pnorm(3)

[1] 0.001349898
```

- We can use indicator functions to help write this in the standard Monte-Carlo form:

$$h(x) = I(x > 3),$$

then

$$I = \int_3^\infty \frac{1}{2\sqrt{\pi}} e^{-x^2/2}\, dx = \int_{-\infty}^\infty h(x) \frac{1}{2\sqrt{\pi}} e^{-x^2/2}\, dx.$$

- Thus, the standard Monte-Carlo approach may be:

$$\hat{I}_m = \frac{1}{m} \sum_{i=1}^m h(X_i), \quad X_i \sim N(0, 1).$$

```
m <- 10000
X1 <- rnorm(n = 10000)
h <- function(x) ifelse(x > 3, 1, 0)

mean(h(X1))

[1] 0.0013
```

- Because $h(X_i)$ is binary, the standard error is:

$$SE \approx \sqrt{\frac{\hat{I}_m(1 - \hat{I}_m)}{m}}$$

```
sqrt(mean(h(X1)) * (1 - mean(h(X1))) / m)

[1] 0.0003603207
```

- The problem with this approach, however, is that if the $X_i \sim N(0,1)$, then almost none of the samples will be larger than 3 (very rare event), so the estimate is high-variance!
- That is, we will end up with $h(X_i) = 0$ for *nearly all* samples of $X_i$.
- We want to focus our samples in the "important" regions to reduce variance.
- Thus, consider instead sampling from $\pi(x) \sim N(4,1)$, so we have more weight in the important part. Then:

$$\hat{I}_m = \frac{1}{m}\sum_{i=1}^{m} h(X_i)\frac{f(X_i)}{\pi(X_i)}, \quad X_i \sim N(4,1).$$

```
X2 <- rnorm(m, mean = 4)
weights2 <- dnorm(X2) / dnorm(X2, mean = 4)
mean(h(X2) * weights2)

[1] 0.001326697
```

- This time, the standard error is given numerically by:

$$SE \approx \frac{\mathrm{sd}(h(X_i)w_i)}{\sqrt{m}}$$

```
sd(h(X2)*weights2) / sqrt(m)

[1] 3.047903e-05
```

- There is an issue that many of the sampled $X_i$ values are less than 3, so we might want to try a completely different sampling approach.
- Consider the function $g(x)$, when $x > 3$, then $g(3) \propto e^{-9}$.
- Thus, we might consider sampling $X_i$ such that:

$$X_i \sim 3 + \mathrm{Exp}(-9)$$

- We again need to do a variable transformation, but this one is easy. If $Y \sim \mathrm{Exp}(\lambda)$, then for $X = Y + 3$,

$$\pi(x) = f_y(x-3).$$

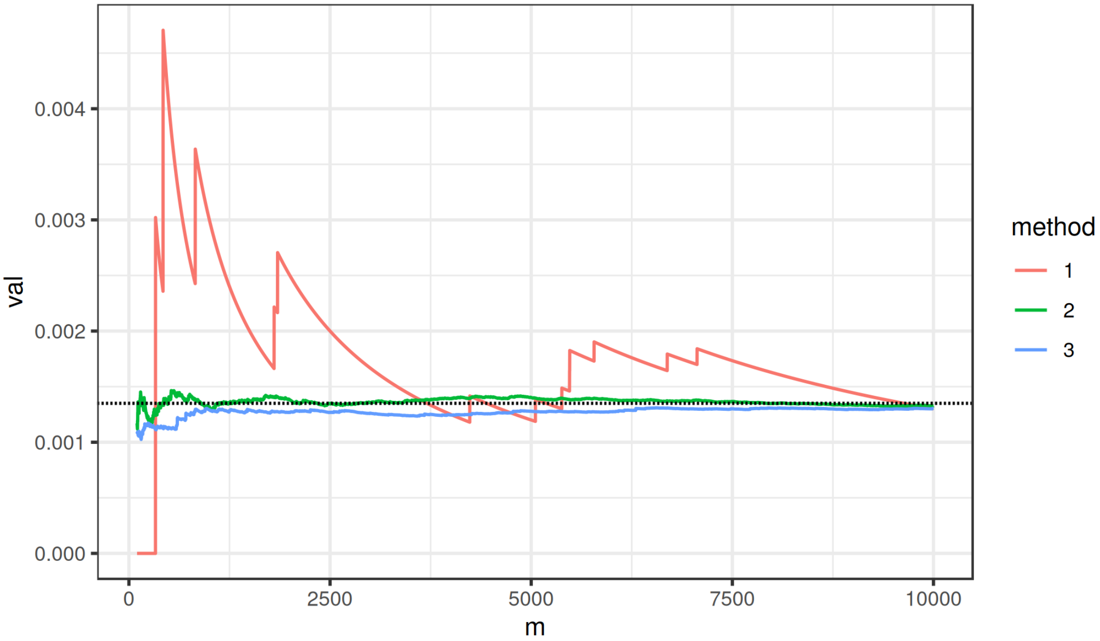
```
X3 <- 3 + rexp(m, 9)
weights3 <- dnorm(X3) / dexp(X3-3, 9)
mean(h(X3) * weights3)

[1] 0.001303543
```

- Finally, let's compute these 3 different estimate for every value of $m \in \{2, 3, \ldots, 10000\}$.

```
library(ggplot2)
est1 <- cumsum(h(X1)) / 1:m
est2 <- cumsum(h(X2) * weights2) / 1:m
est3 <- cumsum(h(X3) * weights3) / 1:m
all_vals <- data.frame(
  val = c(est1, est2, est3),
  method = rep(1:3, each = m) |> factor(),
  m = rep(1:m, 3)
)

ggplot(all_vals |> dplyr::filter(m > 100), aes(x = m, y = val, col = method)) +
  geom_line() +
  geom_hline(yintercept = 1-pnorm(3), linetype = 'dashed') +
  theme_bw()
```



### Rejection Sampling

- We can now see some recurring themes in Bayesian computing (and more generally, approximating integrals).
- For Monte-Carlo integral calculations, we first need to find an appropriate distribution to get samples $X_i \sim \pi(x)$, then apply importance sampling.
- Often, there's not an obvious choice for $\pi$, or we can't directly sample from $\pi(x)$.
- In many cases, we can do *rejection sampling* to get samples from $\pi(x)$, as discussed last semester.

### Variance reduction techniques

- We won't go into a lot of details here, but there are some approaches highlighted in Liu and Liu (Section 2.3 of 2001) that can help with Monte-Carlo evaluations.

- One of the main problems is the variance $\sigma^2$ that arises in the numerator. These approaches are intended to reduce the variance, while not adding any bias.

**Stratified Sampling**
Suppose we wish to calculate the integral

$$\int_A f(x)\,dx.$$

Using Monte-Carlo sampling, the variance is $\sigma^2/n$, where $\sigma^2$ is the variance of the function $f$ over the domain $A$.
Instead, consider breaking the domain $A$ into distinct areas: $A_1$, $A_2$, …, $A_k$, such that the variance of $f$ over these areas is roughly constant, then get Monte-Carlo samples from these regions independently.

- For each sub-region $A_i$, we get $m_i$ samples, and estimate the integral:

  $$\mu_i = \int_{A_i} f(x)\,dx \approx \frac{1}{m_i}\sum_{j=1}^{m_i} f(X_{i,j}),$$

  where $X_{i,j}$ is uniformly sampled on $A_i$.
- Then, if the variance on each region is roughly constant, we can get an improved estimate:

  $$\int_A f(x)\,dx \approx \sum_{i=1}^{k} \hat{\mu_i},$$

  where

  $$\hat{\mu_i} = \sum_{j=1}^{m_i} f(X_{i,j}).$$
- If done properly, the areas are independent, so the total overall variance is just a sum of variances, which should be smaller than the default variance approach:

  $$\mathrm{Var}(\hat{\mu}) = \sum_{i=1}^{k} \frac{\sigma_i^2}{m_i} \le \frac{\sigma^2}{m}$$
- However, if the variance over the individual regions is *not* constant, this can actually result in worse variance.

**Control Variates Method**
In this method, one uses a control variate $C$ which is correlated with the sample $X$, to produce a better estimate.
Suppose we want to estimate $\mu = E[X]$, and $\mu_C = E[C]$ is known. Then, we can produce Monte Carlo samples of the form:

$$X_i^* = X_i - b(C - \mu_C).$$

Then, $E[X_i^*] = E[X] = \mu$, but the variance is:

$$\mathrm{Var}(X_i^*) = \mathrm{Var}(X) - 2b\mathrm{Cov}(X, C) + b^2\mathrm{Var}(C).$$

You can show that there is always a choice of $b$ such that $\mathrm{Var}(X_i^*) < \mathrm{Var}(X_i)$.

- If the covariance and variance terms are easy, we can select an optimal value of $b$ as:

$$b = \frac{\text{Cov}(X, C)}{\text{Var}(C)},$$

  in which case

$$\text{Var}(X_i^*) = (1 - \rho_{XC}^2)\text{Var}(X) < \text{Var}(X_i).$$

- As a concrete example, consider approximating $\mu = E[e^X]$ if $X$ is a standard normal distribution.
- In this case, the true value can be given by the MGF, in particular $\mu = m(1) = e^{-1/2}$
- A standard Monte-Carlo evaluation would be to sample $X_i \sim N(0, 1)$, and compute:

$$\hat{\mu} = \frac{1}{m}\sum_{i=1}^{m} e^{X_i}.$$

- However, consider introducing a covariate $C_i$ that is correlated with $X_i$ to reduce the variance.
- One choice is $C_i = X_i$, in which case $\mu_C = 0$, and we can let:

$$b = \frac{\text{Cov}(e^{X_i}, X_i)}{\text{Var}(X_i)} = \text{Cov}(e^{X_i}).$$

- Since $b$ might be hard to get analytically, we'll just approximate using the covariance function numerically: $\hat{b} =$ `cov(exp(X), x)`.
- Then, setting:

$$X_i^* = e^{X_i} + \hat{b}(X_i - 0),$$

  then $E[X_i^*] = E[e^{X_i^*}] = \mu$, but the variance has been reduced.

- You can try this in R by simulating $e^{X_i}$, calculating the variance, then calculating $b$ and the transformed variable $X_i^* = e^{X_i} - bX_i$, and calculating the variance. They have the same expectation, but the later estimate is lower variance.

**Antithetic Variates Method**

Suppose we want to compute an integral $\int_0^1 f(x)dx$. We can sample $U_i \sim U(0, 1)$, and do Monte-Carlo as usual with $f(U_i)$. However, if $f$ is monotonic, then $U_i' = 1 - U_i$ is *antithetic* to $U_i$, and

$$\text{Cov}(U_i, 1 - U_i) = -Var(U_i).$$

Thus, for every sample $U_i \sim U(0, 1)$ drawn, calculate instead:

$$\int_0^1 f(x)dx \approx \frac{1}{m}\sum_{i=1}^{m}(f(u_i) + f(1 - u_i))/2$$

- More generally, we don't necessarily need to use Monte-Carlo samples from the uniform distribution.
- Recall from the last semester, we proved that if $U_i \sim U(0, 1)$, then we can simulate $X$ from any distribution function $F$ via:

$$X_i = F^{-1}(U_i).$$

  Similarly, if $U_i$ is uniform, then $1 - U_i$ is also uniform, so

$$X_i' = F^{-1}(1 - U_1)$$

also comes from the same distribution. However, we fix a specific $U_i = u$, then $X_i' \neq X_i$, and we can show that $\mathrm{Cov}(X_i, X_i') < 0$. Thus for any monotonic function $f$, if we want to do Monte Carlo integration using $X \sim F$ with $F$ known, then a more efficient approach is $U_i \sim U(0,1)$, $X_i = F^{-1}(U_i)$, $X_i' = F^{-1}(1-U_i)$, and then:

$$\int f(x)\,dx \approx \frac{1}{m}\sum_{i=1}^{m}\frac{f(X_i)+f(X_i')}{2}.$$

- As a more concrete example, note that if $U_i \sim U(0,1)$ is uniform, then $X_i = -\log(U_i)$ is exponential1. Thus, a better approximation of $\mu = \int g(x)f(x)\,dx$, where $f(x) = e^{-x}$ and $g$ is any monotone function, is:

$$\hat{\mu} = \frac{1}{m}\sum_{i=1}^{m}\frac{g\big(-\log(U_i)\big)+g\big(-\log(1-U_i)\big)}{2}$$

# 6 Choice of Priors

**Jeffrey's priors**

- A common safe-choice for a prior distribution is a uniform prior.
- However, there is some (not universal) belief that there is no such thing as a uniform prior.
- Consider the Binomial$(N,p)$ data examples. A "uniform" prior for $p$ might look like:

$$\pi(p) = 1(0 \leq p \leq 1).$$

- This Binomial model arises quite frequently in biomedical sciences. In this setting, we often are interested in the *log-odds*:

$$\theta = \log\Big(\frac{p}{1-p}\Big).$$

- If we pick a uniform prior for $p$, what does the prior for $\theta$ look like?
- The transformation is $g(p) = \log\big(p/(1-p)\big)$.
- For $0 \leq p \leq 1$, this transformation is monotonic. Thus, solving for $\theta$, we get the inverse function that matches the logistic-regression model:

$$g^{-1}(\theta) = \frac{e^\theta}{1+e^\theta},$$

  which, taking the derivative with respect to $\theta$ gives:

$$\frac{d}{d\theta}g^{-1}(\theta) = \frac{e^\theta}{(1+e^\theta)^2}.$$

- Thus, the change of variables theorem states:

$$\begin{aligned}\pi_\Theta(\theta) &= \pi_P\big(g^{-1}(\theta)\big)\big|\frac{d}{d\theta}g^{-1}(\theta)\big| \\ &= 1(0 \leq g^{-1}(\theta) \leq 1)\frac{e^\theta}{(1+e^\theta)^2} \\ &= \frac{e^\theta}{(1+e^\theta)^2}.\end{aligned}$$

- Thus, picking a uniform prior for $p$ gives a non-uniform prior for the log-odds, $\theta = \log(p/1-p)$.
- From a purely Bayesian perspective, this is interpreted by saying that: "although we have no prior knowledge about $p$ (i.e., it could be anything between 0–1), we *do* have some knowledge about the log-odds."
- This calls to question the existence of a uniform prior; while it may be uniform for one parameterization, it is *not* uniform for an alternative parameterization. In this sense, the choice of parameterization (which is independent of the model and data) has a direct impact on final conclusions.
- Note that this breaks the principle of *invariance* that we previously discussed. The estimate for $\hat{\theta} = \log(p/1-p)$ is *not* the same as the estimate $\log(\hat{p}/(1-\hat{p}))$.
- Still, there is a very valid argument why this shouldn't even matter! Consider a binomial model, with parameter $p$. What about the transformation $p^{100}$? While we don't know anything about $p$, we *do* have some valid prior belief about $p^{100}$ that isn't uniform. That is, I would believe that $p^{100}$ would be close to one, so a uniform prior for this transformation probably doesn't make much sense anyway.
- Still we might desire to pick a prior such that for any parameterization $\theta$ or $\varphi$:

$$\pi_{\Phi|X}(\varphi) = \pi_{\Theta|X}(\theta).$$

- We won't go into specific details, but there is such a prior distribution, known as the *Jeffreys prior.*

# 7 Hierarchical Bayes

### Hierarchical Bayes

- The idea behind Hierarchical Bayes is simple: our model $f$ depends on parameters $\theta$.
- We can get a prior for $\theta$, $\pi(\theta)$.
- The prior itself depends on parameters, say $\pi(\theta; \theta_1)$.
- How do we choose $\theta_1$? Sometimes we might know $\theta_1$, but sometimes not.
- In a pure Bayesian paradigm, if we don't know the value of $\theta_1$, then it is also a random variable $\Theta_1$, and we should put a prior on this as well!
- In some way, this allows us to be less-committal about the parameters in the prior model, and instead allow the data to inform our choice of priors (to some degree).
- Philosophically, this situation naturally arises if we want to pick a conjugate prior for $\Theta$, but are not committal about the *hyperparameters* $\Theta_1$ that define the distribution of $\Theta$.
- We could continue doing this many times if we wanted!
- The prior for $\Theta_1$ might depend on parameters $\theta_2$, which we model as a random variable $\Theta_2, \ldots$.
- This leads to a model for $(X, \Theta, \Theta_1, \ldots, \Theta_N)$.
- However, there is a conditional structure to this model:

$$\Theta_N \longrightarrow \Theta_{N-1} \longrightarrow \ldots \longrightarrow \Theta \longrightarrow X.$$

- Thus, $X$ depends only on $\Theta$, and $\Theta_n$ only on $\Theta_{n+1}$:
$$X|\Theta = \theta \sim f(x|\theta),\ \Theta|\Theta_1 = \theta_1 \sim \pi_1(\theta|\theta_1)\ \ldots\ \Theta_N \sim \pi_N(\theta_n).$$

- Using rules of marginal probability and conditional probability, then
$$\begin{aligned}\pi(\theta) &= \int \pi(\theta, \theta_1, \ldots, \theta_N)\, d\theta_{1:N} \\ &= \int \pi(\theta|\theta_{1:N})\pi(\theta_{1:N})\, d\theta_{1:N} \\ &= \int \pi(\theta|\theta_1)\pi(\theta_1|\theta_{2:N})\pi(\theta_{2:N})\, d\theta_{1:N} \\ &= \vdots \\ &= \int \pi(\theta|\theta_1)\pi(\theta_1|\theta_2)\ldots, \pi(\theta_{N-1}|\theta_N)\pi(\theta_N)\, d\theta_{1:N}\end{aligned}$$

- Thus, the hierarchical model is functionally equivalent to the standard Bayesian model:
$$X|\Theta = \theta \sim f(x|\theta) \quad \Theta \sim \pi(\theta),$$
where $\pi(\theta)$ is given by the integral above.

- Why would we want to do this?
  1. Sometimes the data / problem give rise to a natural hierarchical structure, and this idea will be useful. Here, we might actually be interested in the hyperparameters $\theta_1, \ldots, \theta_N$.
  2. We can now be less committal about our priors, while still using desirable structures.
  3. It can sometimes aid computations.

*Trivial case: hierarchical Normal-Normal*
Suppose that the data $X_i$ are iid $N(\theta, 1)$. Set a prior for $\theta$ as $\Theta|\Theta_1 = \theta_1 \sim N(\theta_1, 1)$, and $\Theta_1 \sim N(0, 1)$.

- This model seems rather odd, unless there is a good reason to get the posterior $(\Theta, \Theta_1)|X = x^*$.
- Otherwise, we can show that the *hyperparameter* $\Theta_1$ can be eliminated from the model.
- In particular, consider the marginal distribution of $\Theta$:
$$\pi_\Theta(\theta) = \int \pi_{\Theta|\Theta_1}(\theta|\theta_1)\pi_{\Theta_1})(\theta_1)\, d\theta_1,$$
- Some algebra (i.e., completing the square in the exponential terms), shows that this is equivalent to: $\Theta \sim N(0, 2)$.
- Thus, if we're not interested in the hyper-parameters themselves, then what we are interested in is:
$$\Theta|X = x^*,$$
which can just be calculated by setting the prior $\Theta \sim N(0, 2)$, and following the standard Bayesian approach.

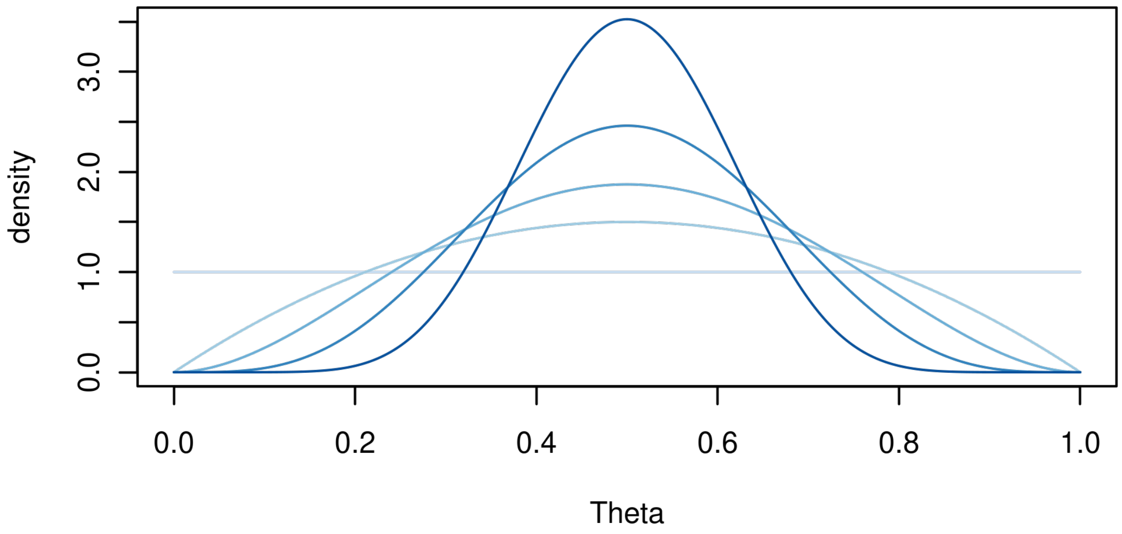
*More realistic example: Coin-toss experiment*
Suppose your friend gives you a coin from another country, and you want to estimate $\theta = p$, the probability of heads. Thus, in $N$ tosses, the natural model for $X$, the number of heads, $X \sim \text{Bin}(N, \theta)$. You're believe that the proportion is close to 1/2, but not quite sure. A nice prior would be the Beta$(\alpha, \beta)$-distribution, since it is conjugate for the binomial family.
If $\Theta \sim \text{Beta}(\alpha, \beta)$, then $E[\Theta] = \frac{\alpha}{\alpha+\beta}$. Thus, if I want a prior centered at 1/2, I can pick: $\theta_1 = \alpha = \beta$, and $E[\Theta] = \theta_1/2\theta_1 = 1/2$. We can now give a prior for $\Theta_1$.

**Hierarchical Bayes (continued)**

- For now, we will restrict our choices of $\Theta_1$ to be integers.

```
Theta <- seq(1e-8, 1-1e-8, length.out = 1000)
B1 <- dbeta(Theta, 1, 1)
B2 <- dbeta(Theta, 2, 2)
B3 <- dbeta(Theta, 3, 3)
B5 <- dbeta(Theta, 5, 5)
B10 <- dbeta(Theta, 10, 10)

plot(x = Theta, y = B1, type = ’l’, ylim = c(0, 3.5), col = "#c6dbef", xlab = "Theta", ylab = "density")
lines(x = Theta, y = B2, type = ’l’, col = ’#9ecae1’)
lines(x = Theta, y = B3, type = ’l’, col = ’#6baed6’)
lines(x = Theta, y = B5, type = ’l’, col = ’#3182bd’)
lines(x = Theta, y = B10, type = ’l’, col = ’#08519c’)
```



- As $\theta_1$ grows, the variance of $\Theta$ shrinks at a rate $O(\theta_1)$.
- Thus, to be noncommittal about our prior on $\Theta$, we will set a *hyperprior* on $\Theta_1$ that has more weight on smaller values of $k$:

$$\pi_{\Theta_1}(k) = \frac{1}{2\log(2)k(2k-1)}, \quad k \in \{1, 2, \dots\}$$

- This hyper-prior was selected somewhat out of convenience (The Catalan Numbers), which will allow us to get the marginal prior of $\Theta$:

$$\pi(\theta) = \sum_{k=1}^{\infty} \pi_{\Theta|\Theta_1}(\theta|k)\pi_{\Theta_1}(k) = \frac{1-|1-2\theta|}{4\log(2)\theta(1-\theta)}, \quad 0 < \theta < 1$$

*Proof.*

The proof relies primarily on the Binomial Theorem. Some algebraic manipulation has already been done for us in the derivation of the Catalan numbers (Wikipedia contributors, 2026).

$$\begin{aligned}\pi(\theta) &= \sum_{k=1}^{\infty} \pi_{\Theta|\Theta_1}(\theta|k)\pi_{\Theta_1}(k) \\ &= \sum_{k=1}^{\infty} \frac{\theta^{k-1}(1-\theta)^{k-1}}{B(k,k)} \frac{1}{2\log(2)k(2k-1)}\end{aligned}$$

Since we picked $\theta_1 = k$ to be an integer (replacing with $k$ because the notation will be easier),

$$B(k,k) = \frac{\Gamma(k)\Gamma(k)}{\Gamma(2k)} = \frac{(k-1)!(k-1)!}{(2k-1)!}.$$

thus,

$$\begin{aligned}\pi(\theta) &= \sum_{k=1}^{\infty} \frac{\theta^{k-1}(1-\theta)^{k-1}(2k-1)!}{(k-1)!(k-1)!2\log(2)k(2k-1)} \\ &= \frac{1}{2\log 2}\sum_{k=1}^{\infty} \binom{2k-1}{k-1} \frac{\theta^{k-1}(1-\theta)^{k-1}}{(2k-1)} \\ &= \frac{1}{2\log 2}\sum_{j=0}^{\infty} \binom{2j+1}{j} \frac{\big(\theta(1-\theta)\big)^j}{2j+1},\end{aligned}$$

where the second inequality uses the identity:

$$\binom{2k-1}{k-1} = \frac{(2k-1)!}{(k-1)!(k-1)!k},$$

and then we made the substitution $j = k-1$ to get the third equality.

Now if we make the substitution $\theta(1-\theta) = x$, such that $0 \le x \le 1/4$, then we get:

$$\pi(\theta) = \frac{1}{2\log 2}\sum_{j=0}^{\infty} \frac{1}{j+1}\binom{2j}{j} x^j = \frac{1}{2\log 2} c(x),$$

where $c(x)$ is the generating unction for the Catalan numbers:

$$c(x) = \frac{1-\sqrt{1-4x}}{2x}.$$

Replacing $x = (\theta)(1-\theta)$, and noting that $1 - 4(\theta)(1-\theta) = (1-2\theta)^2$, we can then simplify to:

$$\pi(\theta) = \frac{1 - |1-2\theta|}{4\log(2)\theta(1-\theta)}, \quad 0 \le \theta \le 1.$$

□

- In this case, we can get a closed-form expression for $\pi(\theta)$, but as you can tell, it can often get very difficult to do this mathematically.

- Thus, while the hierarchical structure is equivalent to just setting $\pi(\theta)$ as our prior (and not worrying about hierarchical model), this additional structure can aid in computations.
- If we are looking to estimate, for instance, the posterior mean:

  $$E_{\Theta|X}[\Theta],$$

  Then the law of total expectation gives:

  $$E_{\Theta|X}[\Theta|X] = E_{\Theta_1|X}\left[E_{\Theta|\Theta_1,X}[\Theta|\Theta_1, X]\right].$$

- Thus, the calculation of the posterior mean of $\Theta|X$ can be done without needing explicit form of the posterior $\Theta|X$, which can simplify the problem.
- Our particular choice of likelihood and prior makes it easy to calculate the marginal-likelihood of $\Theta_1 = k$:

  $$\begin{aligned}\pi_{X|\Theta_1}(x|k) &= \int_0^1 f(x|\theta, k)\pi_{\Theta|\Theta_1}(\theta; k)\, d\theta \\ &= \binom{N}{x}\frac{B(x+k, N-x+k)}{B(k,k)}.\end{aligned}$$

- Also, The Beta distribution was picked because it is conjugate, so the posterior mean $\Theta|\Theta_1 = k, X$ is readily available:

  $$E_{\Theta|\Theta_1=k,X} = \mu_k = \frac{x+k}{N+2k}.$$

- Now we need to take the expectation of this, with respect to the marginal posterior (un-normalized weights) $\pi_{\Theta_1|x}(k|x)$:

  $$\begin{aligned}\pi_{\Theta_1|x}(k|x) &\propto w_k \\ &= \pi_{X|\Theta_1}(x|k)\pi_{\Theta_1}(k) \\ &= \frac{B(x+k, N-x+k)}{2\log(2)B(k,k)k(2k-1)}.\end{aligned}$$

- Then, the normalized weights are:

  $$\bar{w}_k = \frac{w_k}{\sum_j w_j} = p(k|x),$$

  and the posterior mean is:

  $$E[\Theta|x] = \sum_{k=1}^{\infty} \bar{w}_k \mu_k.$$

- For this particular example, the sum can be calculated exactly. However, we can also approximate this using software by taking the first $K$ partial sums. Check out the provided HB-code R code.

**Baseball statistics**

- One place that Hierarchical Bayes is particularly useful is for grouped data:
- Let $i = 1, 2, \ldots, k$ denote $k$ distinct groups, and $j = 1, 2, \ldots, n_i$ be the number of observations, per each group.

- Then, the data look like:
$$Y_{1,1},\dots,Y_{1,n_1},Y_{2,1},\dots,Y_{2,n_2},\dots Y_{k,n_k}$$

- We then have a natural Hierarchical structure to our model, where within each group $i$, the observations $1,\dots,n_i$ are iid:
$$Y_{i,j}\overset{\text{iid}}{\sim} f(y;\theta_i)$$

- Then, we create a prior for the group-parameters:
$$\Theta_i\overset{\text{iid}}{\sim}\pi(\theta).$$

- This approach is extremely useful as a form of regularlization

*Baseball: Hitting Percentages*
Consider two baseball players, A and B. Player A has a career batting average of 0.353, whereas Player B as an average 0.400. Which do you prefer on your team?

- From the percentages alone, Player B is a natural choice.
- What if we were told, however, that Player A had 1000 at bats, with success rate 353/1000, and Player B only has 10 at bats (successful 4/10 times)?
- From this example, there are two things to consider:
  1. The best batting average of all time is: .3662 (Ty Cobb). Roughly only 200 players have hit over .3000 for their entire career, the average being somewhere closer to 0.2500. This information is very useful, and maybe should be accounted for.
  2. If there are many observations within a group, we should highly value these data, but if there are not many observations, we should "penalize" towards the prior distribution.
- Continuing this example, we effectively have two binomial likelihoods. If $X_A$ and $X_B$ are the number of successful at-bats for players $A$ and $B$, respectively, then
$$X_A\sim\text{Binom}(1000,p_A)X_B\sim\text{Binom}(10,p_B)$$

- For the hyper-prior of $P_i$, we will use a Beta distribution, since it's a convenient conjugate prior:
$$P_i\sim Beta(\alpha,\beta).$$

- Now we need to pick the values of $\alpha,\beta$. We want a Beta with mean around 0.25. Since the expected value is $\alpha/(\alpha+\beta)$, setting this equal to 1/4 gives the relation: $3\alpha=\beta$.
- Using a simple normal approximation, we want roughly $0.25+2\sigma=0.4$, so that the upper-end of this prior gives little-to-no weight for batting averages greater than 0.4. Thus, we want $\sigma\approx 15/200$, which we can approximate with 2/25. Using our relation $3\alpha=\beta$, this gives:
$$\left(\frac{2}{25}\right)^2=\frac{\alpha\beta}{(\alpha+\beta)^{(}\alpha+\beta+1)}=\frac{3\alpha^2}{16\alpha^2(4\alpha+1)},$$
which implies $\alpha\approx 7$.
- Thus, the hyper-prior we will use is Beta(7, 21), shown in Figure 1

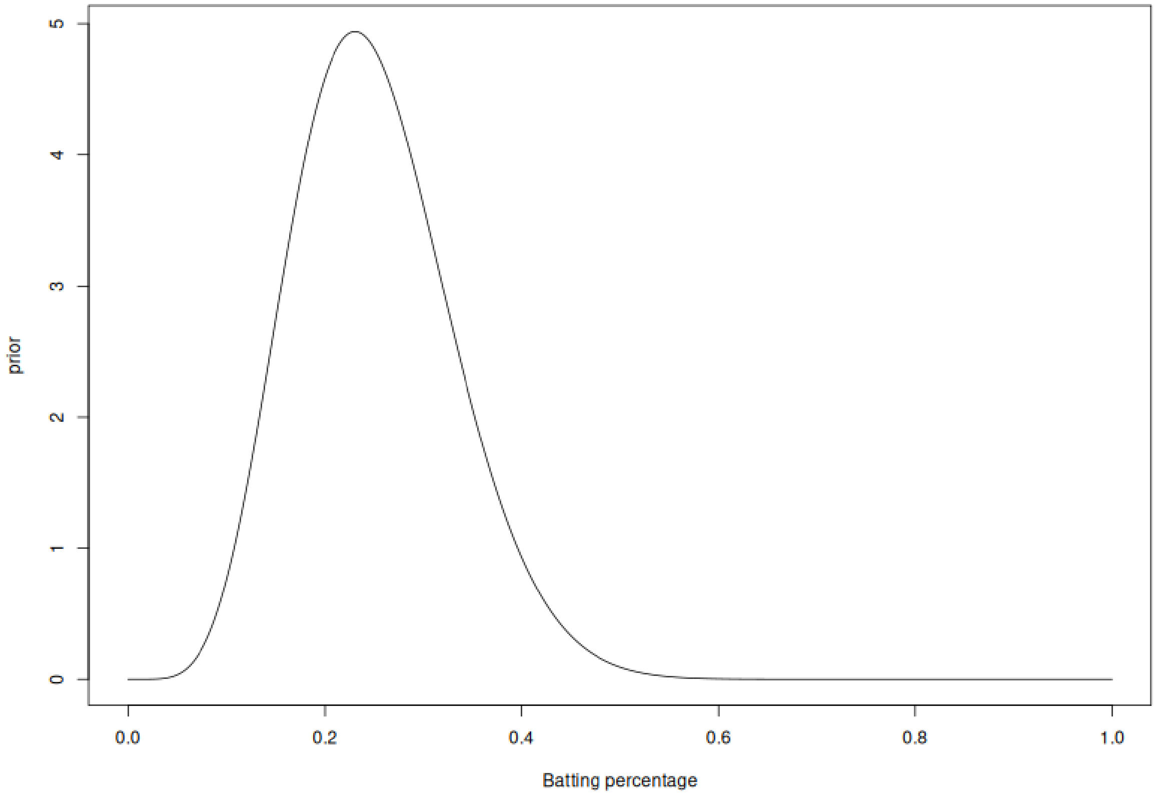


Figure 1: Prior distribution, Beta$(7, 21)$ for the batting percentages.

- Recall that for the Beta-Binomial conjugate distributions, the posterior for the probability of success is given by:
$$P_i|X_i = x_i \sim \text{Beta}(\alpha + x_i, \beta + n_i - x_i).$$
- For the current example, of the two-baseball players, we get posterior distributions:
$$P_A|X_A = x_A \sim \text{Beta}(7 + 353, 21 + 647)$$
$$P_B|X_B = x_B \sim \text{Beta}(7 + 4, 21 + 6).$$

  The posterior means for each of these distributions are 0.350 and 0.289, respectively. The respective posteriors are plotted in Figure 2.

# 8 Empirical Bayes

### Empirical Bayes

- The idea because Empirical Bayes is simple: Use the data to estimate the prior distribution for the parameters.
- This is a bit controversial, because we're "double dipping". Often makes sense when used in conjunction with a hierarchical structure.

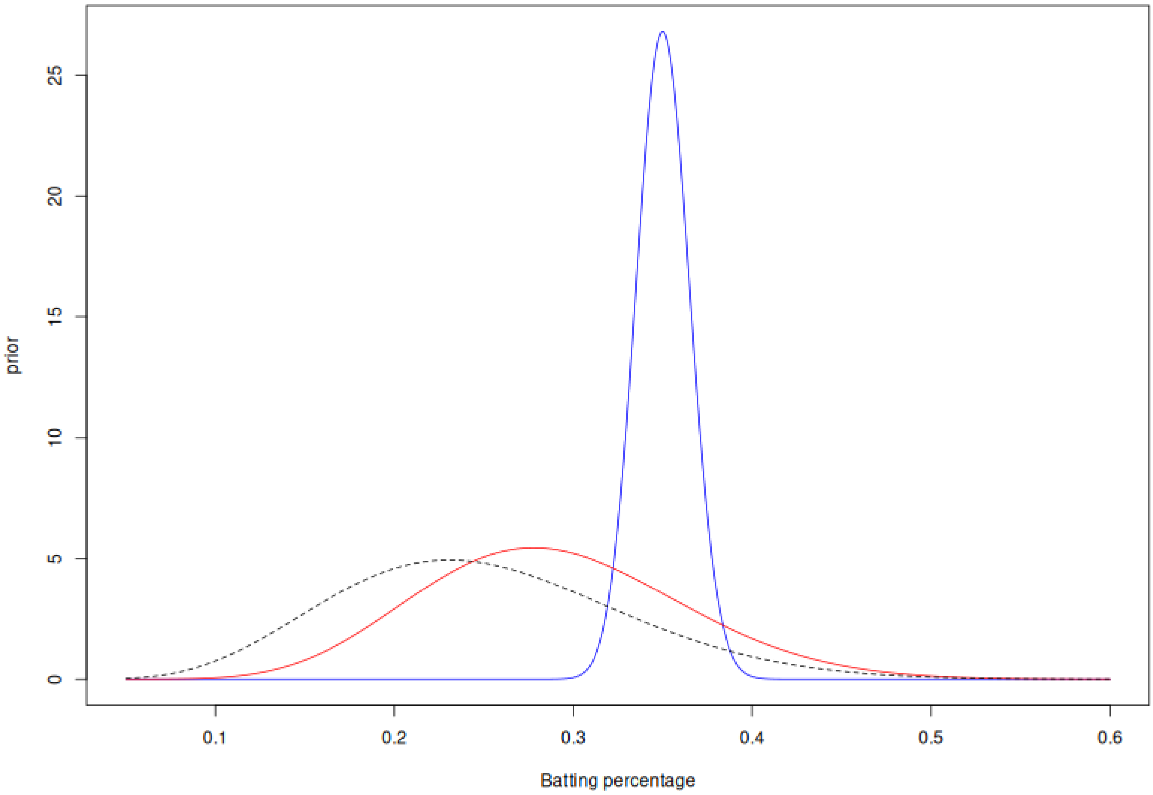


Figure 2: Posterior distribution estimates for the estimated batting percentage of the two Baseball players. Player A is in blue, and Player B is in Red. For refefence, the prior distribution is given as a dashed black line.

*Baseball Players*
Let's revisit the Baseball-batting percentage problem. Now, we want to estimate the posterior via maximum likelihood of IID Beta distribution, using dataset of historic baseball-player data.

# 9 Uncertainty quantification

**Uncertainty in Bayes estimates**

- Uncertainty quantification is very natural in Bayesian models.
- We won't focus on too many examples, but will introduce the basic concepts.

*Approach 1: Posterior variance*
Since our estimates correspond to an entire distribution of $\Theta|X = x^*$, one natural idea is to use the posterior variance to report uncertainty:

$$\mathrm{Var}(\Theta|X = x^*).$$

This approach is often convenient, since we can approximate this variance even if we don't have a closed-form solution for the posterior, and instead just have samples from the distribution. However, variance is not a great measure of uncertainty for asymmetric distributions, which is often the case with our posterior distribution. It also doesn't make as much sense to use if we want our point estimate to correspond to the MAP instead of the posterior mean.

*Approach 2: quantiles (credible intervals)*
Another common approach is the get desired percentiles / quantiles of the posterior. For instance, picking $\alpha = 0.05$, then we might want to select an interval of $I_\alpha = (\theta_{\alpha/2}, \theta_{1-\alpha/2})$ to represent our confidence, where $\alpha_x$ corresponds to the $x$th percentile of the posterior. In this case,

$$\begin{aligned} P(\Theta \in I_\alpha | X = x^*) &= P(\theta_{\alpha/2} \leq \Theta \leq \theta_{1-\alpha/2} | X = x^*) \\ &= F_{\Theta|X}(\theta_{1-\alpha/2}) - F_{\Theta|X}(\theta_{\alpha/2}) \\ &= \alpha \end{aligned}$$

Note that this interval above closely corresponds to the frequentist "Confidence Interval". In Bayesian analysis, we often call this a "credible interval", and it often has very good frequentist properties. Note a few key differences:

- In this interval, we are conditioning on the data $X = x^*$. Thus, the interval is fixed, and $\Theta$ is random. In the frequentist confidence interval, this is reversed! The parameter is fixed, and the interval is random. The Bayesian interpretation is more natural in this case, as it corresponds to the probability that $\Theta$ is in the given interval.
- Note that really the probability should be written as:

  $$P(\Theta \in I_\alpha | X = x^*, M),$$

  where $M$ is the selected model, which depends of course on the prior. Thus, a different choice in prior will lead to a different interval; even though these intervals are more natural to interpret, one should be careful to not forget that this probability statement formally is a statement about *belief* about a parameter, not long-term frequencies. Thus, if our prior belief $M$ is wildly different than the real-situation, this probability may not correspond to the number of times this experiment will cover the truth.

*Approach 3: high-density region (or high-posterior density)*
Also called credible intervals, but constructed in a different way, is the high-density-region (HDR) approach. Here, we take as our set that measures uncertainty the set:

$$R_\alpha(\Theta) = \{\Theta : \pi_{\Theta|X}(\theta|x^*) \geq 1 - \alpha\}.$$

This is similar to the previously proposed credible interval, and if the posterior is unimodal and symmetric, they give the same answer. This is like a top-down approach: Starting from the MAP, you draw a horizontal line down until $(1-\alpha)\%$ of the posterior is above the line. This makes the biggest difference if the posterior is multi-modal, in which case $R_\alpha(\Theta)$ might not even be an interval, but rather the union of intervals.

# 10 Conclusion

### Concluding Rermarks

- Bayesian statistics in more than just a philosophy of what probability is, but also a useful way to solve hard problems.
- At the earliest stages of the statistics discipline, most approaches were fundamentally Bayesian, until Fisher in the early 1900s

- The approach lost some popularity until the late 1900s, do to the numeric difficulties that often arise.

- Advances in hardware, software, and methodology caused an explosion of Bayesian statistics research, which continues today.

- Some of the algorithms that led to this success include:

- Markov-chain Monte Carlo (MCMC) algorithms. Importance sampling is hard in high-dimensions, and often fails in time-series / spatial statistics. Further, a good importance distribution is not always available. MCMC is a class of algorithms that largely solve this issue, by using *dependent* samples; the next sample in a "Chain" is a permutation of previous samples, guided by the likelihood function.

- Sequential Monte Carlo (SMC) algorithms. In many time-series examples, even MCMC fails because of the unique dependence of the model. SMC addresses this by doing Monte-Carlo one step (observation) at a time. The most famous example of this is the *particle filter*.

- Despite some heated arguments on the topic, a well-done analysis with sufficient data usually result in the same practical solutions, whether a Bayes or Frequentist approach is taken.

- There's somewhat of a general consensus that Bayesian methods are particularly useful when the sample size is small; here, the data alone might not have enough information to tell us about the model, and we benefit from using a prior belief.

- Frequentist methods still dominate in the areas of hypothesis testing, mean comparison, etc.; this is where statistics is used most often by non-statisticians, so it remains the dominant approach in science.

- However, there are increasing complaints about the accepted approach (by statisticians and non-statisticians alike), and this has given rise to more Bayesian applications.

# Acknowledgments

- Compiled on February 13, 2026 using R version 4.5.2.

- We acknowledge students and instructors for previous versions of this course / slides.

# References

Liu JS, Liu JS (2001). *Monte Carlo strategies in scientific computing*, volume 10. Springer. 8, 12

Rice JA (2007). *Mathematical statistics and data analysis*, volume 371. 3 edition. Thomson/Brooks/Cole Belmont, CA. 1

Wikipedia contributors (2026). "Catalan Number#Proof of the formula: Wikipedia, The Free Encyclopedia." [Online; accessed 30-January-2026], URL `https://en.wikipedia.org/wiki/Catalan_number#Proof_of_the_formula`. 15