

# Mathematical Statistics II

## The Bayesian Approach to Parameter Estimation

---

Jesse Wheeler

# 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.

# 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  $f(x|\theta)$ .

## Some Review II

### 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:

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

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

## Some Review III

### 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  $\theta$ .

## Some Review IV

- 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

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

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

## Examples

---

# Bayesian point-estimate examples

## Poisson model

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

## 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
```

## Real-data example: Poisson Distribution II

### 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.

## Conjugate priors

- The first approach to the  $\text{Poisson}(\lambda)$  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.

# Conjugate priors II

## 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  $\text{Poisson}(\lambda)$ , 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.

## Conjugate priors III

- 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 priors IV

### Conjugate Normals

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

# 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:

$$\text{likelihood} = \prod_{i=1}^n f(x_i^* | \theta).$$

## Posteriors and Likelihood II

- 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  $\text{Poisson}(\lambda)$  model, while taking the alternative approach of using a uniform prior.

# 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.

## Numeric Integration II

- 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
```

# Numeric Integration III

- 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 to 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, Chapter 1).
- However, these approaches scale very poorly as the dimensions of  $x$  and the integration area  $A$  increase.



## Numeric Integration IV

- 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.

## Numeric Integration V

- 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

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

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

## Numeric Integration VI

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

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

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

- 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

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

## Numeric Integration VII

- If  $x$  is univariate, this approach is worse than standard deterministic approaches  $O(1/m)$ , but it has better performance in higher dimensional settings.

## Numeric Integration VIII

- 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:

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

## Numeric Integration IX

```
sqrt(64.8/m)
```

```
[1] 0.08049845
```

```
# Numeric Approximation:
```

```
sd(3 * X^2) / sqrt(m)
```

```
[1] 0.07999898
```

# Numeric Integration X

- 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**.

# Numeric Integration XI

- 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.



## Numeric Integration XII

- 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**.

## Numeric Integration XIII

- 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:

```
set.seed(12345)
m <- 10000
X <- runif(n = m, 0, 3)
3 * mean(x_sq(X))
[1] 8.992983
```

## Importance Sampling, Example 1 II

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

```
# Numeric Approximation:  
sd(3 * X^2) / sqrt(m)  
[1] 0.07999898
```

## Importance Sampling, Example 1 III

- 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:

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

## Importance Sampling, Example 1 IV

- 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:

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

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

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

## Importance Sampling, Example 1 V

- 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
```

## Importance Sampling, Example 1 VI

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

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

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.



## Importance Sampling, Example 1 VII

- 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:

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

## Importance Sampling, Example 1 VIII

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

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

## Importance Sampling, Example 1 IX

- 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:

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

## Importance Sampling, Example 1 X

- This effectively never happens in real world-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

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
```

## Importance Sampling, Example II II

- 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:

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

## Importance Sampling, Example II III

```
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
```

## Importance Sampling, Example II IV

- 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:

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



## Importance Sampling, Example II V

```
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:

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

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

[1] 3.047903e-05
```

## Importance Sampling, Example II VI

- 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:

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

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

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

## Importance Sampling, Example II VII

```
X3 <- 3 + rexp(m, 9)
weights3 <- dnorm(X3) / dexp(X3-3, 9)
mean(h(X3) * weights3)
[1] 0.001303543
```

## Importance Sampling, Example II VIII

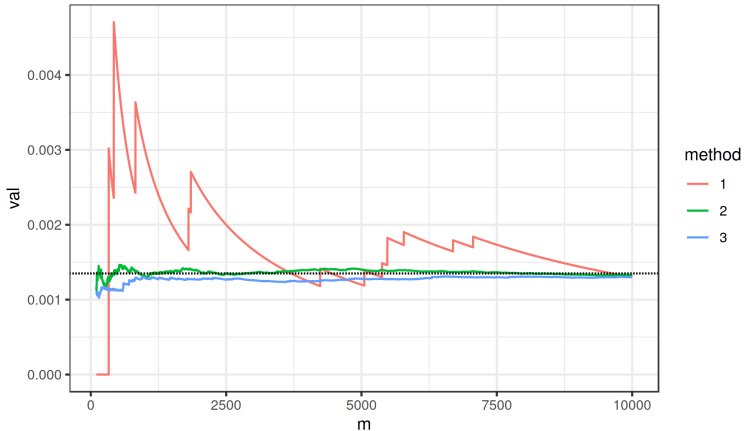
- Finally, let's compute these 3 different estimate for every value of  $m \in \{2, 3, \dots, 10000\}$ .

## Importance Sampling, Example II IX

```
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)) +
  geom_line() +
  geom_hline(yintercept = 1-pnorm(3), linetype = 'dashed') +
  theme_bw()
```

# Importance Sampling, Example II X



# 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.



## Variance reduction techniques II

### 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:  $A_1, A_2, \dots, A_k$ , such that the variance of  $f$  over these areas is roughly constant, then get Monte-Carlo samples from these regions independently.

## Variance reduction techniques III

### 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:

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

## Variance reduction techniques IV

### 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

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

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

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

## Choice of Priors

---

- TODO

# 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).

## Hierarchical Bayes II

- 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  $\Theta_2, \dots$
- This leads to a model for  $(X, \Theta, \Theta_1, \dots, \Theta_N)$ .
- However, there is a conditional structure to this model:

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

- Thus,  $X$  depends only on  $\Theta$ , and  $\Theta_n$  only on  $\Theta_{n+1}$ :

$$X|\Theta = \theta \sim f(x|\theta), \quad \Theta|\Theta_1 = \theta_1 \sim \pi_1(\theta|\theta_1) \quad \dots \quad \Theta_N \sim \pi_N(\theta_n).$$



## Hierarchical Bayes III

- Using rules of marginal probability and conditional probability, then

$$\begin{aligned}\pi(\theta) &= \int \pi(\theta, \theta_1, \dots, \theta_N) d\theta_{1:N} \\ &= \int \pi(\theta | \theta_{1:N}) \pi(\theta_{1:N}) d\theta_{1:N} \\ &= \int \pi(\theta | \theta_1) \pi(\theta_1 | \theta_{2:N}) \pi(\theta_{2:N}) d\theta_{1:N} \\ &= \vdots \\ &= \int \pi(\theta | \theta_1) \pi(\theta_1 | \theta_2) \dots \pi(\theta_{N-1} | \theta_N) \pi(\theta_N) d\theta_{1:N}\end{aligned}$$

# Hierarchical Bayes IV

- 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  $\theta_1, \dots, \theta_N$ .
  2. We can now be less committal about our priors, while still using desirable structures.
  3. It can sometimes aid computations.

# Hierarchical Bayes V

## 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  $\Theta_1 \sim N(0, 1)$ .

# Hierarchical Bayes VI

## 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  $\text{Beta}(\alpha, \beta)$ -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 = "#c6
```

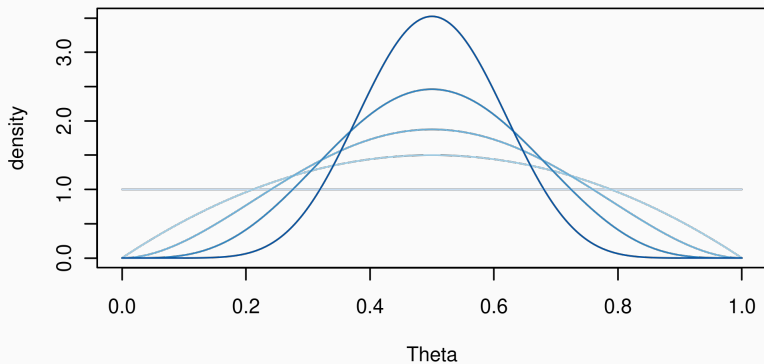
```
lines(x = Theta, y = B2, type = 'l', col = '#9ecae1')
```

```
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')
```

## Hierarchical Bayes (continued) II



## Hierarchical Bayes (continued) III

- 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  $\Theta$ :

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

## Hierarchical Bayes (continued) IV

- 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:

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



## Hierarchical Bayes (continued) V

- 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:

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

## Hierarchical Bayes (continued) VI

- 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:

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

## Hierarchical Bayes (continued) VII

- 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.

- TODO: HB example of on-base percentages for base-ball players. Move to EB?

# 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.

# Uncertainty quantification

---

# Uncertainty in Bayes estimates

- TODO



## References and Acknowledgements

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

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

- Compiled on February 3, 2026 using R version 4.5.2.
- Licensed under the [Creative Commons Attribution-NonCommercial](#) license. Please share and remix non-commercially, mentioning its origin.



## References and Acknowledgements II

- We acknowledge [students and instructors for previous versions of this course / slides](#).