

STAT 24400 Lecture 16
Section 8.6 The Bayesian Approach to Parameter
Estimation

Yibi Huang
Department of Statistics
University of Chicago

Example — Coin Tossing

Suppose there is a jar of hundreds of coins with various probabilities to land heads.

We randomly choose a coin from the jar, flip it n times and observe X heads.

Can you infer about the probability θ to land heads for the chosen coin?

Model:

$$X \mid \Theta \sim \text{Bin}(n, \Theta)$$

$$\Theta \sim \text{some distribution of coin probabilities}$$

Example — Coin Tossing (Discrete Prior)

Suppose the jar only contains 2 types of coins.

- ▶ 75% of the coins are fair with a prob. of $\Theta = 0.5$ to land heads;
- ▶ 25% of the coins are biased with a prob. of $\Theta = 0.8$ to land heads.

In other words, the distribution of Θ is

$$P(\Theta = 0.5) = 0.75, \quad P(\Theta = 0.8) = 0.25.$$

The joint PMF of X and Θ is

$$f(x, \theta) = f_{X|\Theta}(x | \theta) f_{\Theta}(\theta) = \begin{cases} \binom{n}{x} (0.5)^n \times 0.75 & \text{if } \theta = 0.5 \\ \binom{n}{x} (0.8)^x (0.2)^{n-x} \times 0.25 & \text{if } \theta = 0.8 \end{cases}$$

The marginal PMF of X is

$$f_X(x) = \sum_{\theta \in \{0.5, 0.8\}} f(x, \theta) = 0.75 \binom{n}{x} (0.5)^n + 0.25 \binom{n}{x} (0.8)^x (0.2)^{n-x}.$$

Given $X = x$, the conditional distribution of Θ would be

$$f_{\Theta|X}(\theta | x) = \frac{f(x, \theta)}{f_X(x)} = \begin{cases} \frac{0.75(0.5)^n}{0.75(0.5)^n + 0.25(0.8)^x(0.2)^{n-x}} & \text{if } \theta = 0.5 \\ \frac{0.25(0.8)^x(0.2)^{n-x}}{0.75(0.5)^n + 0.25(0.8)^x(0.2)^{n-x}} & \text{if } \theta = 0.8 \end{cases}$$

For $n = 10$ tosses,

$$P(\Theta = 0.5 | x) = f_{\Theta|X}(0.5 | x) = \begin{cases} 0.991 & \text{if } x = 4 \\ 0.965 & \text{if } x = 5 \\ 0.875 & \text{if } x = 6 \\ 0.636 & \text{if } x = 7 \\ 0.304 & \text{if } x = 8 \\ 0.0984 & \text{if } x = 9 \\ 0.0266 & \text{if } x = 10 \end{cases}$$

Example — Coin Tossing (Continuous Prior)

Suppose the coins in the jar have probabilities $\Theta \sim \text{Uniform}(0,1)$ to land heads with the PDF

$$f_{\Theta}(\theta) = 1, \quad \text{for } 0 \leq \theta \leq 1.$$

The joint distribution of X and Θ is

$$f(x, \theta) = f_{X|\Theta}(x | \theta)f_{\Theta}(\theta) = \binom{n}{x} \theta^x (1 - \theta)^{n-x} \cdot 1.$$

The marginal PMF of X is

$$\begin{aligned} f_X(x) &= \int_0^1 f(x, \theta) d\theta = \binom{n}{x} \overbrace{\int_0^1 \theta^x (1 - \theta)^{n-x} d\theta}^{=\text{Beta}(x+1, n-x+1)} \\ &=^* \binom{n}{x} \frac{\Gamma(x+1)\Gamma(n-x+1)}{\Gamma(n+2)} \\ &=^{**} \frac{n!}{x!(n-x)!} \frac{x!(n-x)!}{(n+1)!} = \frac{1}{n+1} \end{aligned}$$

where the step (*) is from the definition of the Beta function
Beta(u, v):

$$\text{Beta}(u, v) = \int_0^1 x^{u-1}(1-x)^{v-1}dx, \text{ and it's equal to } \frac{\Gamma(u)\Gamma(v)}{\Gamma(u+v)}.$$

and the step (**) comes from that $\Gamma(x+1) = x!$ if $x \geq 0$ is an integer.

The conditional PDF of Θ given $X = x$ would be

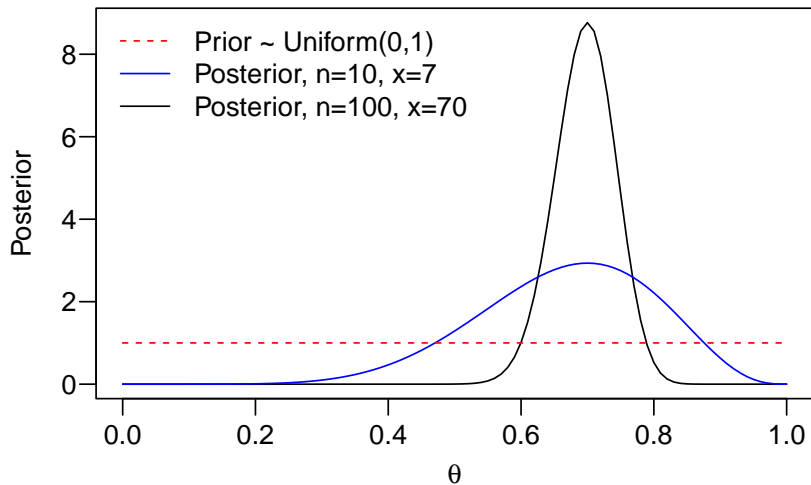
$$\begin{aligned} f_{\Theta|X}(\theta | x) &= \frac{f(x, \theta)}{f_X(x)} = (n+1) \frac{n!}{x!(n-x)!} \theta^x (1-\theta)^{n-x} \\ &= \frac{\Gamma(n+2)}{\Gamma(x+1)\Gamma(n-x+1)} \theta^x (1-\theta)^{n-x}, \quad 0 \leq \theta \leq 1. \end{aligned}$$

Thus

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

This is called the *posterior distribution* of Θ .

Prior v.s. Posterior Distribution



Bayesian Statistics

So far, we have been trying to infer about the unknown parameter(s) Θ of a known distribution $f(x | \Theta)$ from i.i.d. observations $X_1, X_2, \dots, X_n \sim f(x | \Theta)$.

- ▶ In *frequentist statistics*, the parameter(s) Θ are regarded as fixed number(s), not random.
- ▶ In *Bayesian statistics*, the **underlying parameter(s) Θ are treated as a random variable**, distributed according to a *prior distribution* $\Theta \sim g(\theta)$

The prior distribution may be interpreted as reflecting our subjective beliefs or our level of uncertainty about the parameter, or may reflect information gathered from past experience.

Bayesian Statistics — Posterior Distribution

Upon observing $X_1, X_2, \dots, X_n \sim f(x | \Theta)$, we calculate the **conditional distribution** of Θ given X_1, X_2, \dots, X_n , called the *posterior distribution*.

The posterior distribution is our updated belief on the possible value of Θ , after observing $X_1, X_2, \dots, X_n \sim f(x | \Theta)$.

Beta-Binomial Bayes Estimation

For Binomial observation $X \sim \text{Bin}(n, \Theta)$, a commonly used prior for Θ is the Beta(α, β) distribution with the PDF

$$f_{\Theta}(\theta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1 - \theta)^{\beta-1}, \quad \text{for } 0 \leq \theta \leq 1.$$

with mean and variance

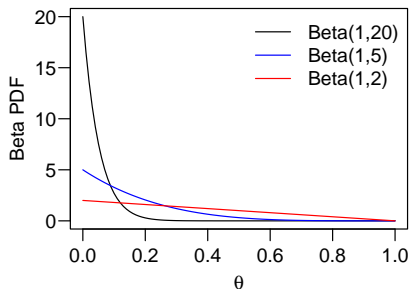
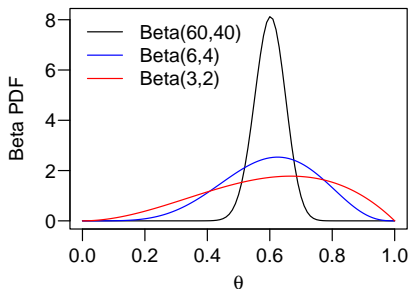
$$E[\Theta] = \frac{\alpha}{\alpha + \beta}, \quad \text{Var}(\Theta) = \frac{\alpha}{\alpha + \beta} \cdot \frac{\beta}{\alpha + \beta} \cdot \frac{1}{\alpha + \beta + 1}$$

The Beta family include a great variety of distributions on $[0, 1]$ that can reflect our belief on the possible range of Θ .

How to Choose a Beta Prior (1)

If you believe that $\Theta \approx \theta_0$, choose α and β that $\frac{\alpha}{\alpha + \beta} = \theta_0$.

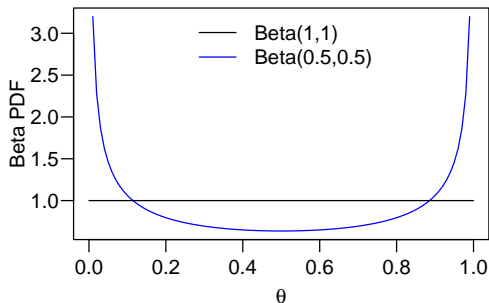
- ▶ choose large values of α and β with $\frac{\alpha}{\alpha + \beta} = \theta_0$ if you believe Θ is close to θ_0
- ▶ choose small values of α and β with $\frac{\alpha}{\alpha + \beta} = \theta_0$ if you are not sure whether Θ is close to θ_0



How to Choose a Beta Prior (2)

If you have no clue what Θ is, you can choose

- ▶ $\text{Beta}(\alpha = 1, \beta = 1) = \text{Uniform}[0,1]$, an uninformative prior
- ▶ $\text{Beta}(\alpha = 0.5, \beta = 0.5)$



Beta-Binomial Posterior (1)

If Θ has the prior $\text{Beta}(\alpha, \beta)$

$$f_{\Theta}(\theta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1 - \theta)^{\beta-1}, \quad \text{for } 0 \leq \theta \leq 1.$$

The joint distribution of X and Θ is

$$\begin{aligned} f(x, \theta) &= f_{X|\Theta}(x | \theta) f_{\Theta}(\theta) \\ &= \binom{n}{x} \theta^x (1 - \theta)^{n-x} \cdot \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1 - \theta)^{\beta-1} \\ &= h(x) \theta^{x+\alpha-1} (1 - \theta)^{n-x+\beta-1} \end{aligned}$$

where $h(x) = \binom{n}{x} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}$ only depends on x (and α, β) but not θ .

The marginal PMF of X is

$$\begin{aligned} p_X(x) &= \int_0^1 f(x, \theta) d\theta = h(x) \overbrace{\int_0^1 \theta^{x+\alpha-1} (1 - \theta)^{n-x+\beta-1} d\theta}^{=\text{Beta}(x+\alpha, n-x+\beta)} \\ &= h(x) \frac{\Gamma(x + \alpha)\Gamma(n - x + \beta)}{\Gamma(n + \alpha + \beta)} \end{aligned}$$

Beta-Binomial Posterior (2)

The conditional PDF of Θ given $X = x$ would be

$$f_{\Theta|X}(\theta | x) = \frac{f(x, \theta)}{f_X(x)} = \frac{\Gamma(n + \alpha + \beta)}{\Gamma(x + \alpha)\Gamma(n - x + \beta)} \theta^{x+\alpha-1} (1 - \theta)^{n-x+\beta-1}$$

Thus the *posterior distribution* of Θ given $X = x$ is

$$(\Theta | X = x) \sim \text{Beta}(x + \alpha, n - x + \beta).$$

Posterior Mean & Posterior Mode

The posterior gives a distribution of Θ .

What if we want a “point estimate”,

i.e. a single value that is a good estimate for θ ?

Two Common Options:

- ▶ Posterior mean:

$$\hat{\theta} = E(\theta \mid X_1, \dots, X_n) \leftarrow E(\cdot) \text{ with respect to posterior of } (\Theta \mid X_1, \dots, X_n)$$

- ▶ Posterior mode:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} f_{\Theta|X}(\theta \mid X_1, \dots, X_n)$$

Posterior Mean & Mode for Beta-Binomial Bayes

For the $\text{Beta}(\alpha, \beta)$ distribution,

$$\text{Mean} = \frac{\alpha}{\alpha + \beta}, \quad \text{Mode} = \frac{\alpha - 1}{\alpha + \beta - 2},$$

As the *posterior distribution* of Θ given X is

$$(\Theta | X) \sim \text{Beta}(X + \alpha, n - X + \beta).$$

$$\text{posterior mean} = \frac{X + \alpha}{n + \alpha + \beta}, \quad \text{posterior mode} = \frac{X + \alpha - 1}{n + \alpha + \beta - 2}.$$

- ▶ The posterior mean is like with MLE for Θ but adding α more heads and β more tails to the outcome.
- ▶ The posterior mode is like with MLE for Θ but adding $\alpha - 1$ more heads and $\beta - 1$ more tails to the outcome.
- ▶ For Uniform[0,1] prior ($\alpha = \beta = 1$),

$$\text{posterior mean} = \frac{X + 1}{n + 2}, \quad \text{posterior mode} = \frac{X}{n} = \text{MLE}.$$

Note the posterior mean is a weighted average of the MLE and the prior mean.

$$\begin{aligned}\text{posterior mean} &= \frac{X + \alpha}{n + \alpha + \beta} \\ &= \frac{n}{n + \alpha + \beta} \cdot \underbrace{\frac{X}{n}}_{=\text{MLE}} + \frac{\alpha + \beta}{n + \alpha + \beta} \cdot \underbrace{\frac{\alpha}{\alpha + \beta}}_{=\text{prior mean}}\end{aligned}$$

Likewise, the posterior mode is a weighted average of the MLE and the prior mode.

$$\begin{aligned}\text{posterior mode} &= \frac{X + \alpha - 1}{n + \alpha + \beta - 2} \\ &= \frac{n}{n + \alpha + \beta - 2} \cdot \underbrace{\frac{X}{n}}_{=\text{MLE}} + \frac{\alpha + \beta - 2}{n + \alpha + \beta - 2} \cdot \underbrace{\frac{\alpha - 1}{\alpha + \beta - 2}}_{=\text{prior mode}}\end{aligned}$$

For both of them, the greater the sample size n , the more weights go to the MLE.

Both are $\approx X/n = \text{MLE}$ for θ when n is large \Rightarrow prior has little effect on Bayesian estimate when the sample size n is large

Gamma-Exponential Bayes

Data: i.i.d. $X_1, \dots, X_n \mid \lambda \stackrel{\text{iid}}{\sim}$ Exponential(λ) with joint PDF

$$f_{X|\Lambda}(x_1, \dots, x_n \mid \Lambda = \lambda) = \lambda^n e^{-\lambda \sum_{i=1}^n x_i} = \lambda^n e^{-n\lambda \bar{X}}$$

Prior for Λ is $\Lambda \sim$ Gamma(a, b) with PDF

$$f_{\Lambda}(\lambda) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda}, \quad \lambda \geq 0$$

The joint distribution of X_1, \dots, X_n and Λ is

$$\begin{aligned} f(x_1, \dots, x_n, \lambda) &= f_{X|\Lambda}(x_1, \dots, x_n \mid \lambda) f_{\Lambda}(\lambda) \\ &= \lambda^n e^{-n\lambda \bar{X}} \cdot \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda} \\ &= h(a, b) \lambda^{n+a-1} e^{-(b+n\bar{X})\lambda} \end{aligned}$$

where $h(a, b) = \frac{b^a}{\Gamma(a)}$.

As the joint PDF is proportional to

$$f(x_1, \dots, x_n, \lambda) \propto \lambda^{n+a-1} e^{-(b+n\bar{X})\lambda},$$

and the marginal PDF of (X_1, \dots, X_n)

$$f_X(x_1, \dots, x_n) = \int f(x_1, \dots, x_n, \lambda) d\lambda$$

does not depend on λ , the posterior must be proportional to

$$f_{\lambda|X}(\lambda | x) = \frac{f(x, \lambda)}{f_X(x)} \propto \lambda^{n+a-1} e^{-(b+n\bar{X})\lambda},$$

\Rightarrow the posterior distribution is $\text{Gamma}(a + n, b + n\bar{X})$

Posterior Mean & Mode for Gamma-Exponential Bayes

For the Gamma(a, b) distribution

$$\text{Mean} = \frac{a}{b}, \quad \text{Mode} = \frac{b-1}{b},$$

As the *posterior distribution* of Λ given (X_1, \dots, X_n) is Gamma($a + n, b + n\bar{X}$)

$$\text{posterior mean} = \frac{a+n}{b+n\bar{X}}, \quad \text{posterior mode} = \frac{a+n-1}{b+n\bar{X}}.$$

Note both of them are $\approx 1/\bar{X} = \text{MLE}$ for Λ when n is large.
 \Rightarrow the prior has little influence on the point estimate when the sample size n is large.

Credible Intervals

A $(1 - \alpha)$ credible interval (L, U) (calculated as a function of X_1, \dots, X_n) contains $(1 - \alpha)$ posterior probability:

$$P(L \leq \Theta \leq U \mid X_1, \dots, X_n) = 1 - \alpha$$

There are various ways to construct a credible interval.

Two common options:

- ▶ Equal tailed interval
- ▶ High posterior density (HPD) interval

If the posterior distribution is symmetric & unimodal, the two options are equivalent

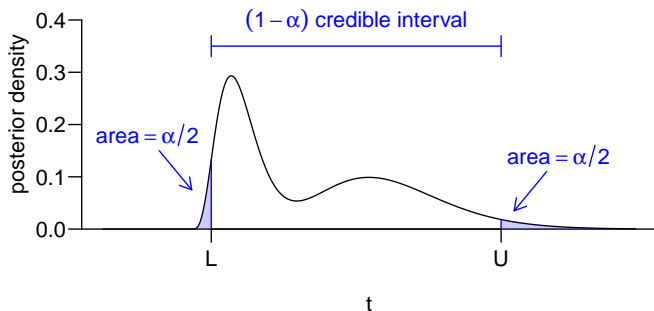
Equal-Tailed Credible Intervals

The $1 - \alpha$ equal-tailed credible interval (L, U) for Θ is

$$P(\Theta < L) = F_{\text{posterior}}(L) = \frac{\alpha}{2},$$

$$P(\Theta > U) = 1 - F_{\text{posterior}}(U) = \frac{\alpha}{2}.$$

where $F_{\text{posterior}}$ is the CDF for the posterior distribution of Θ .

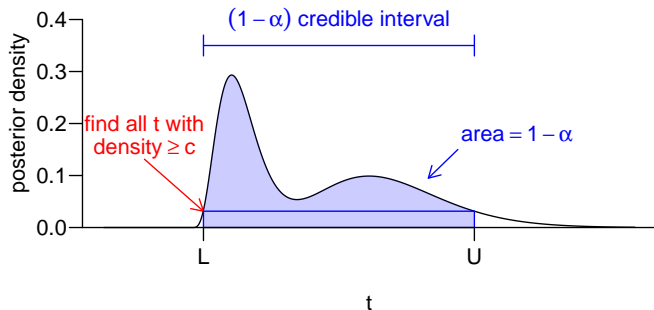


High Posterior Density (HPD) Interval

The our interval is given by

$$I = \{t : f_{\theta|X_1, \dots, X_n}(t | x_1, \dots, x_n) \geq c\}$$

where the density cutoff c is chosen so that $\text{prob.} = 1 - \alpha$



Note that an HPD interval I might not be a single interval!
(In the example above, if α is large, then I splits into two intervals)

Equal-Tailed Credible Interval for Gamma-Exponential

Model:

$$\begin{cases} \Lambda \sim \text{Gamma}(a, b) \\ X_1, \dots, X_n \mid \Lambda \stackrel{\text{iid}}{\sim} \text{Exponential}(\lambda) \end{cases}$$

Posterior:

$$\Lambda \mid X_1, \dots, X_n \sim \text{Gamma}(a + n, b + n\bar{X})$$

Equal-tailed credible interval: (L, U)

$$P(\Theta < L) = F_{\text{Gamma}(a+n, b+n\bar{X})}(L) = \frac{\alpha}{2},$$

$$P(\Theta > U) = 1 - F_{\text{Gamma}(a+n, b+n\bar{X})}(U) = \frac{\alpha}{2}.$$

where $F_{\text{Gamma}(a+n, b+n\bar{X})}$ is the CDF of the posterior.

A fact about Gamma distributions:

$$\text{Gamma}(a, b) \approx N\left(\frac{a}{b}, \frac{a}{b^2}\right) \quad \text{for large } a.$$

Thus,

$$\text{Gamma}(a + n, b + n\bar{X}) \approx N\left(\frac{a + n}{b + n\bar{X}}, \frac{a + n}{(b + n\bar{X})^2}\right)$$

Therefore, the $(1 - \alpha)$ credible interval is approximately equal to:

$$\approx \frac{a + n}{b + n\bar{X}} \pm z_{\alpha/2} \cdot \frac{\sqrt{a + n}}{b + n\bar{X}}$$

If n is large while a & b are constant. . .

$$\approx \frac{1}{\bar{X}} \pm z_{\alpha/2} \cdot \frac{1}{\sqrt{n \cdot \bar{X}}}$$

which is the confidence interval based on the asymp. normality of the MLE.