

# Stat 710: Mathematical Statistics

## Lecture 1

Jun Shao

Department of Statistics  
University of Wisconsin  
Madison, WI 53706, USA

# Chapter 4: Estimation in Parametric Models

## Lecture 1: Prior, posterior, and Bayes formula

$X$  is from a population in a parametric family  $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ , where  $\Theta \subset \mathcal{R}^k$  for a fixed integer  $k \geq 1$

### Three topics

- Bayesian method
- Minimavity and admissibility
- Likelihood approach

### Bayes rules in §2.3.2

- Decision rules minimizing the average risk w.r.t. a given probability measure  $\Pi$  on  $\Theta$
- Optimal rules in the *Bayesian approach*, which is fundamentally different from the classical frequentist approach that we have been adopting

# Chapter 4: Estimation in Parametric Models

## Lecture 1: Prior, posterior, and Bayes formula

$X$  is from a population in a parametric family  $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ , where  $\Theta \subset \mathcal{R}^k$  for a fixed integer  $k \geq 1$

### Three topics

- Bayesian method
- Minimaxity and admissibility
- Likelihood approach

### Bayes rules in §2.3.2

- Decision rules minimizing the average risk w.r.t. a given probability measure  $\Pi$  on  $\Theta$
- Optimal rules in the *Bayesian approach*, which is fundamentally different from the classical frequentist approach that we have been adopting

# Chapter 4: Estimation in Parametric Models

## Lecture 1: Prior, posterior, and Bayes formula

$X$  is from a population in a parametric family  $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ , where  $\Theta \subset \mathcal{R}^k$  for a fixed integer  $k \geq 1$

### Three topics

- Bayesian method
- Minimavity and admissibility
- Likelihood approach

### Bayes rules in §2.3.2

- Decision rules minimizing the average risk w.r.t. a given probability measure  $\Pi$  on  $\Theta$
- Optimal rules in the *Bayesian approach*, which is fundamentally different from the classical frequentist approach that we have been adopting

## Bayesian approach

- $\theta$  is viewed as a realization of a random vector  $\vec{\theta} \in \Theta$  whose *prior* distribution is  $\Pi$
- Prior distribution: past experience, past data, or a statistician's belief (subjective)
- Sample  $X \in \mathcal{X}$ : from  $P_\theta = P_{X|\theta}$ , the conditional distribution of  $X$  given  $\vec{\theta} = \theta$
- Posterior distribution: updated prior distribution using the sample  $X = x$

### How to construct the posterior?

By Theorem 1.7, the joint distribution of  $X$  and  $\vec{\theta}$  is a probability measure on  $\mathcal{X} \times \Theta$  determined by

$$P(A \times B) = \int_B P_{X|\theta}(A) d\Pi(\theta), \quad A \in \mathcal{B}_{\mathcal{X}}, B \in \mathcal{B}_\Theta$$

The posterior distribution is the conditional distribution  $P_{\theta|x}$  whose existence is guaranteed by Theorem 1.7 a.s.  $x \in \mathcal{X}$

## Bayesian approach

- $\theta$  is viewed as a realization of a random vector  $\vec{\theta} \in \Theta$  whose *prior* distribution is  $\Pi$
- Prior distribution: past experience, past data, or a statistician's belief (subjective)
- Sample  $X \in \mathcal{X}$ : from  $P_\theta = P_{x|\theta}$ , the conditional distribution of  $X$  given  $\vec{\theta} = \theta$
- Posterior distribution: updated prior distribution using the sample  $X = x$

## How to construct the posterior?

By Theorem 1.7, the joint distribution of  $X$  and  $\vec{\theta}$  is a probability measure on  $\mathcal{X} \times \Theta$  determined by

$$P(A \times B) = \int_B P_{x|\theta}(A) d\Pi(\theta), \quad A \in \mathcal{B}_{\mathcal{X}}, B \in \mathcal{B}_\Theta$$

The posterior distribution is the conditional distribution  $P_{\theta|x}$  whose existence is guaranteed by Theorem 1.7 a.s.  $x \in \mathcal{X}$

When  $P_{x|\theta}$  has a p.d.f., Theorem 4.1 provides a formula for the p.d.f. of the posterior distribution

### Theorem 4.1 (Bayes formula)

Assume  $\mathcal{P} = \{P_{x|\theta} : \theta \in \Theta\}$  is dominated by a  $\sigma$ -finite measure  $\nu$  and  $f_\theta(x) = dP_{x|\theta}/d\nu$  is a Borel function on  $(\mathcal{X} \times \Theta, \sigma(\mathcal{B}_{\mathcal{X}} \times \mathcal{B}_{\Theta}))$ . Let  $\Pi$  be a prior distribution on  $\Theta$ . Suppose that  $m(x) = \int_{\Theta} f_\theta(x) d\Pi > 0$ .

(i) The posterior distribution  $P_{\theta|x} \ll \Pi$  and

$$dP_{\theta|x}/d\Pi = f_\theta(x)/m(x)$$

(ii) If  $\Pi \ll \lambda$  and  $d\Pi/d\lambda = \pi(\theta)$  for a  $\sigma$ -finite measure  $\lambda$ , then

$$dP_{\theta|x}/d\lambda = f_\theta(x)\pi(\theta)/m(x)$$

Proof:

Result (ii) follows from result (i) and Proposition 1.7(iii)

When  $P_{x|\theta}$  has a p.d.f., Theorem 4.1 provides a formula for the p.d.f. of the posterior distribution

### Theorem 4.1 (Bayes formula)

Assume  $\mathcal{P} = \{P_{x|\theta} : \theta \in \Theta\}$  is dominated by a  $\sigma$ -finite measure  $\nu$  and  $f_\theta(x) = dP_{x|\theta}/d\nu$  is a Borel function on  $(\mathcal{X} \times \Theta, \sigma(\mathcal{B}_{\mathcal{X}} \times \mathcal{B}_{\Theta}))$ . Let  $\Pi$  be a prior distribution on  $\Theta$ . Suppose that  $m(x) = \int_{\Theta} f_\theta(x) d\Pi > 0$ .

(i) The posterior distribution  $P_{\theta|x} \ll \Pi$  and

$$dP_{\theta|x}/d\Pi = f_\theta(x)/m(x)$$

(ii) If  $\Pi \ll \lambda$  and  $d\Pi/d\lambda = \pi(\theta)$  for a  $\sigma$ -finite measure  $\lambda$ , then

$$dP_{\theta|x}/d\lambda = f_\theta(x)\pi(\theta)/m(x)$$

Proof:

Result (ii) follows from result (i) and Proposition 1.7(iii)

When  $P_{x|\theta}$  has a p.d.f., Theorem 4.1 provides a formula for the p.d.f. of the posterior distribution

### Theorem 4.1 (Bayes formula)

Assume  $\mathcal{P} = \{P_{x|\theta} : \theta \in \Theta\}$  is dominated by a  $\sigma$ -finite measure  $\nu$  and  $f_\theta(x) = dP_{x|\theta}/d\nu$  is a Borel function on  $(\mathcal{X} \times \Theta, \sigma(\mathcal{B}_{\mathcal{X}} \times \mathcal{B}_\Theta))$ . Let  $\Pi$  be a prior distribution on  $\Theta$ . Suppose that  $m(x) = \int_\Theta f_\theta(x) d\Pi > 0$ .

(i) The posterior distribution  $P_{\theta|x} \ll \Pi$  and

$$dP_{\theta|x}/d\Pi = f_\theta(x)/m(x)$$

(ii) If  $\Pi \ll \lambda$  and  $d\Pi/d\lambda = \pi(\theta)$  for a  $\sigma$ -finite measure  $\lambda$ , then

$$dP_{\theta|x}/d\lambda = f_\theta(x)\pi(\theta)/m(x)$$

Proof:

Result (ii) follows from result (i) and Proposition 1.7(iii)

## Proof for (i)

$$\int_{\mathcal{X}} m(x) d\nu = \int_{\mathcal{X}} \int_{\Theta} f_{\theta}(x) d\Pi d\nu = \int_{\Theta} \int_{\mathcal{X}} f_{\theta}(x) d\nu d\Pi = 1$$

The second equality follows from Fubini's theorem

$m(x)$  is integrable w.r.t.  $\nu$  and  $m(x) < \infty$  a.e.  $\nu$

Because of this,  $m(x)$  is called the marginal p.d.f. of  $X$  w.r.t.  $\nu$

Without loss of generality we may assume  $m(x) > 0$

If  $m(x) = 0$  for an  $x \in \mathcal{X}$ , then  $f_{\theta}(x) = 0$  a.s.  $\Pi$

Either  $x$  should be eliminated from  $\mathcal{X}$  or the prior  $\Pi$  is incorrect and a new prior should be specified

For  $x \in \mathcal{X}$  with  $m(x) < \infty$ , define

$$P(B, x) = \frac{1}{m(x)} \int_B f_{\theta}(x) d\Pi, \quad B \in \mathcal{B}_{\Theta}$$

Then  $P(\cdot, x)$  is a probability measure on  $\Theta$  a.e.  $\nu$ .

## Proof for (i)

$$\int_{\mathcal{X}} m(x) d\nu = \int_{\mathcal{X}} \int_{\Theta} f_{\theta}(x) d\Pi d\nu = \int_{\Theta} \int_{\mathcal{X}} f_{\theta}(x) d\nu d\Pi = 1$$

The second equality follows from Fubini's theorem

$m(x)$  is integrable w.r.t.  $\nu$  and  $m(x) < \infty$  a.e.  $\nu$

**Because of this,  $m(x)$  is called the marginal p.d.f. of  $X$  w.r.t.  $\nu$**

Without loss of generality we may assume  $m(x) > 0$

If  $m(x) = 0$  for an  $x \in \mathcal{X}$ , then  $f_{\theta}(x) = 0$  a.s.  $\Pi$

Either  $x$  should be eliminated from  $\mathcal{X}$  or the prior  $\Pi$  is incorrect and a new prior should be specified

For  $x \in \mathcal{X}$  with  $m(x) < \infty$ , define

$$P(B, x) = \frac{1}{m(x)} \int_B f_{\theta}(x) d\Pi, \quad B \in \mathcal{B}_{\Theta}$$

Then  $P(\cdot, x)$  is a probability measure on  $\Theta$  a.e.  $\nu$ .

## Proof for (i)

$$\int_{\mathcal{X}} m(x) d\nu = \int_{\mathcal{X}} \int_{\Theta} f_{\theta}(x) d\Pi d\nu = \int_{\Theta} \int_{\mathcal{X}} f_{\theta}(x) d\nu d\Pi = 1$$

The second equality follows from Fubini's theorem

$m(x)$  is integrable w.r.t.  $\nu$  and  $m(x) < \infty$  a.e.  $\nu$

Because of this,  $m(x)$  is called the marginal p.d.f. of  $X$  w.r.t.  $\nu$

Without loss of generality we may assume  $m(x) > 0$

If  $m(x) = 0$  for an  $x \in \mathcal{X}$ , then  $f_{\theta}(x) = 0$  a.s.  $\Pi$

Either  $x$  should be eliminated from  $\mathcal{X}$  or the prior  $\Pi$  is incorrect and a new prior should be specified

For  $x \in \mathcal{X}$  with  $m(x) < \infty$ , define

$$P(B, x) = \frac{1}{m(x)} \int_B f_{\theta}(x) d\Pi, \quad B \in \mathcal{B}_{\Theta}$$

Then  $P(\cdot, x)$  is a probability measure on  $\Theta$  a.e.  $\nu$ .

By Theorem 1.7, it remains to show that

$$P(B, x) = P(\vec{\theta} \in B | X = x)$$

By Fubini's theorem,  $P(B, \cdot)$  is a measurable function of  $x$

Let  $P_{x, \theta}$  denote the "joint" distribution of  $(X, \vec{\theta})$

For any  $A \in \sigma(X)$ ,

$$\begin{aligned} \int_{A \times \Theta} I_B(\theta) dP_{x, \theta} &= \int_A \int_B f_{\theta}(x) dv d\Pi \\ &= \int_A \left[ \int_B \frac{f_{\theta}(x)}{m(x)} d\Pi \right] \left[ \int_{\Theta} f_{\theta}(x) d\Pi \right] dv \\ &= \int_{\Theta} \int_A \left[ \int_B \frac{f_{\theta}(x)}{m(x)} d\Pi \right] f_{\theta}(x) dv d\Pi \\ &= \int_{A \times \Theta} P(B, x) dP_{x, \theta} \end{aligned}$$

where the third equality follows from Fubini's theorem

This completes the proof

## Discrete $X$ and $\vec{\theta}$ : The Bayes formula in elementary probability

$$P(\vec{\theta} = \theta | X = x) = \frac{P(X = x | \vec{\theta} = \theta)P(\vec{\theta} = \theta)}{\sum_{\theta \in \Theta} P(X = x | \vec{\theta} = \theta)P(\vec{\theta} = \theta)}$$

### Remarks on the Bayesian approach

- The posterior  $P_{\theta|x}$  contains all the information we have about  $\theta$
- Statistical decisions and inference should be made based on  $P_{\theta|x}$ , conditional on the observed  $X = x$
- In estimating  $\theta$ ,  $P_{\theta|x}$  can be viewed as a randomized decision rule under the approach discussed in §2.3  
After  $X = x$  is observed,  $P_{\theta|x}$  is a randomized rule, which is a probability distribution on the action space  $\mathcal{A} = \Theta$
- The Bayesian method can be applied iteratively

## Discrete $X$ and $\vec{\theta}$ : The Bayes formula in elementary probability

$$P(\vec{\theta} = \theta | X = x) = \frac{P(X = x | \vec{\theta} = \theta)P(\vec{\theta} = \theta)}{\sum_{\theta \in \Theta} P(X = x | \vec{\theta} = \theta)P(\vec{\theta} = \theta)}$$

### Remarks on the Bayesian approach

- The posterior  $P_{\theta|x}$  contains all the information we have about  $\theta$
- Statistical decisions and inference should be made based on  $P_{\theta|x}$ , conditional on the observed  $X = x$
- In estimating  $\theta$ ,  $P_{\theta|x}$  can be viewed as a randomized decision rule under the approach discussed in §2.3  
After  $X = x$  is observed,  $P_{\theta|x}$  is a randomized rule, which is a probability distribution on the action space  $\mathcal{A} = \Theta$
- The Bayesian method can be applied iteratively