

# Stochastic Processes



**Week 07 (Version 01)**

**Estimation Theory 02**

Hamid R. Rabiee

Fall 2024

# Outline of Week 7 Lectures

- Introduction to Optimal Frequentist Estimator
- Score and Fisher Information
- Cramer-Rao Lower Bound
- Rao-Blackwell Theorem
- UMVUE
- Bayesian Estimation
- Conjugate Prior
- Consistency
- Efficiency
- Estimator Comparison
- Summary

# Introduction to Optimal Frequentist Estimator

- In the Frequentist's point of view, an optimal estimator is both **unbiased** and **minimum variance**.
- How can we obtain an estimator  $\hat{\theta}$  that is unbiased?
  - Given any biased estimator  $\hat{\theta}_b$  with bias  $b$ , then we can remove the bias to obtain an unbiased estimator  $\hat{\theta}$  from  $\hat{\theta}_b$ , i. e.  $\hat{\theta} = \hat{\theta}_b - b$ .
- How can we obtain a minimum variance estimator  $\hat{\theta}_{mv}$  from an unbiased estimator?
  - We need to obtain a lower bound for an unbiased estimator and make sure  $\hat{\theta}_{mv}$  achieves that bound.

# Outline of Week 7 Lectures

- Introduction to Optimal Frequentist Estimator
- Score and Fisher Information
- Cramer-Rao Lower Bound
- Rao-Blackwell Theorem
- UMVUE
- Bayesian Estimation
- Conjugate Prior
- Consistency
- Efficiency
- Estimator Comparison
- Summary

# Score and Fisher Information

- The **score**  $s(\theta)$  is defined as the gradient of the log-likelihood function with respect to the parameter  $\theta$ .

$$s(\theta) = \frac{\partial \log L(\theta|x)}{\partial \theta} = \frac{\partial \log f(x|\theta)}{\partial \theta}$$

- When evaluated at a particular value of the parameter vector, the score indicates the sensitivity of the log-likelihood function to infinitesimal changes to the parameter values.

# Score and Fisher Information

- The mean of score  $s(\theta)$
- Although  $s(\theta)$  is a function of  $\theta$ , it also depends on the observations  $X$ , at which the likelihood function is evaluated, and the expected value of the score, evaluated at the parameter value  $\theta$ , is zero.

$$\begin{aligned} E(s \mid \theta) &= \int_{\mathcal{X}} f(x \mid \theta) \frac{\partial}{\partial \theta} \log \mathcal{L}(\theta \mid x) dx \\ &= \int_{\mathcal{X}} f(x \mid \theta) \frac{1}{f(x \mid \theta)} \frac{\partial f(x \mid \theta)}{\partial \theta} dx = \int_{\mathcal{X}} \frac{\partial f(x \mid \theta)}{\partial \theta} dx \end{aligned}$$

# Score and Fisher Information

- We can interchange the derivative and integral by using Leibniz integral rule:

$$\frac{\partial}{\partial \theta} \int_{\mathcal{X}} f(x | \theta) dx = \frac{\partial}{\partial \theta} 1 = 0$$

- If we repeatedly sample from some distribution, and repeatedly calculate its score, then the mean value of the scores would tend to zero asymptotically.

# Score and Fisher Information

- The **Fisher Information** is defined as the **variance of score**. It is a way of measuring the amount of information that an observable random variable  $X$  carries about an unknown parameter  $\theta$  of a distribution that models  $X$ .

$$\mathcal{I}(\theta) = \mathbb{E} \left[ \left( \frac{\partial}{\partial \theta} \log f(X | \theta) \right)^2 \middle| \theta \right] = \int_{\mathbb{R}} \left( \frac{\partial}{\partial \theta} \log f(x | \theta) \right)^2 f(x | \theta) dx$$

- The Fisher information is not a function of a particular observation, as the random variable  $X$  has been averaged out.



## Score and Fisher Information

- If  $\log f(x|\theta)$  is twice differentiable with respect to  $\theta$ , and under certain regularity conditions, the Fisher information may also be written as:

$$\mathcal{I}(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2} \log f(X|\theta) \middle| \theta\right]$$

- The regularity conditions are as follows:
  - The partial derivative of  $f(X|\theta)$  with respect to  $\theta$  exists.
  - The integral of  $f(X|\theta)$  can be differentiated under the integral sign with respect to  $\theta$ .
  - The support of  $f(X|\theta)$  does not depend on  $\theta$ .

Why the two equations to compute Fisher Information are Equal?

$$\begin{aligned}
 \mathcal{I}(\theta) &= \mathbb{E} \left[ \left( \frac{\partial}{\partial \theta} \log f(X|\theta) \right)^2 \middle| \theta \right] = - \mathbb{E} \left[ \frac{\partial^2}{\partial \theta^2} \log f(X|\theta) \middle| \theta \right] \\
 \text{Let } \frac{\partial}{\partial \theta} &= \nabla_{\theta} \\
 \nabla_{\theta}[s(X; \theta)] &= \nabla_{\theta}^2 [\ln(f(X; \theta))] \\
 &= \nabla_{\theta} [\nabla_{\theta} [\ln(f(X; \theta))]] \\
 &= \nabla_{\theta} \left[ \frac{\nabla_{\theta}[f(X; \theta)]}{f(X; \theta)} \right] \\
 &= \frac{(f(X; \theta) \nabla_{\theta}^2[f(X; \theta)]) - (\nabla_{\theta}[f(X; \theta)] \nabla_{\theta}[f(X; \theta)])}{(f(X; \theta))^2} \\
 &= \frac{\nabla_{\theta}^2[f(X; \theta)]}{f(X; \theta)} - \frac{(\nabla_{\theta}[f(X; \theta)] \nabla_{\theta}[f(X; \theta)])}{(f(X; \theta))^2} \\
 &= \frac{\nabla_{\theta}^2[f(X; \theta)]}{f(X; \theta)} - (\nabla_{\theta}[\ln(f(X; \theta))] \nabla_{\theta}[\ln(f(X; \theta))]) \\
 &= \frac{\nabla_{\theta}^2[f(X; \theta)]}{f(X; \theta)} - (s(X; \theta))^2
 \end{aligned}$$

$$\begin{aligned}
E \left[ \frac{\nabla_{\theta}^2 [f(X; \theta)]}{f(X; \theta)} \right] &= \int \frac{\nabla_{\theta}^2 [f(X; \theta)]}{f(X; \theta)} f(X; \theta) dx \\
&= \int \nabla_{\theta}^2 [f(X; \theta)] dx \\
&= \nabla_{\theta}^2 \left[ \int f(X; \theta) dx \right] \\
&= \nabla_{\theta}^2 [1] \\
&= 0
\end{aligned}$$

$$\begin{aligned}
E \left[ \nabla_{\theta}^2 [\ln(f(X; \theta))] \right] &= E \left[ \frac{\nabla_{\theta}^2 [f(X; \theta)]}{f(X; \theta)} \right] - E \left[ (s(X; \theta))^2 \right] \\
&= 0 - E \left[ (s(X; \theta))^2 \right] \\
&= 0 - I(\theta)
\end{aligned}$$

# Outline of Week 7 Lectures

- Introduction to Optimal Frequentist Estimator
- Score and Fisher Information
- Cramer-Rao Lower Bound
- Rao-Blackwell Theorem
- UMVUE
- Bayesian Estimation
- Conjugate Prior
- Consistency
- Efficiency
- Estimator Comparison
- Summary

## Cramer-Rao Lower Bound

- The **Cramer–Rao bound (CRB)** expresses a lower bound on the variance of unbiased estimators of a deterministic (fixed, though unknown) parameter  $\theta$ , stating that the **variance** of any such estimator is **at least as high as the inverse of the Fisher information**.
- An unbiased estimator which achieves this lower bound is said to be **efficient**.
- Suppose  $\theta$  is an unknown deterministic parameter which is to be estimated from  $n$  independent observations of  $x$ , each from a distribution according to some probability density function  $f(x|\theta)$ .

## Cramer-Rao Lower Bound

- The variance of any *unbiased* estimator  $\hat{\theta}$  of  $\theta$  is then bounded by the reciprocal of the Fisher information  $I(\theta)$ :

$$\text{var}(\hat{\theta}) \geq \frac{1}{I(\theta)}$$

- The **efficiency** of an **unbiased estimator**  $\hat{\theta}$  measures how close this estimator's variance comes to this lower bound; estimator efficiency is defined as:

$$e(\hat{\theta}) = \frac{I(\theta)^{-1}}{\text{var}(\hat{\theta})}$$

- The Cramer–Rao lower bound gives:  $e(\hat{\theta}) \leq 1$

$$X_1, \dots, X_n \stackrel{iid}{\sim} f(X_i|\theta)$$

$$f(X|\theta) = \prod_{i=1}^n f(X_i|\theta)$$

$$\begin{aligned} S(\theta) &= \frac{\partial}{\partial \theta} \log f(X|\theta) \\ &= \frac{\partial}{\partial \theta} \log \left( \prod_{i=1}^n f(X_i|\theta) \right) \end{aligned}$$

$$\begin{aligned} I(\theta) &= E \left[ \left( \frac{\partial}{\partial \theta} \log \left( \prod_{i=1}^n f(X_i|\theta) \right) \right)^2 \right] = \\ &= n E \left[ \left( \frac{\partial}{\partial \theta} \log f(X_i|\theta) \right)^2 \right] = n I_{X_i}(\theta) \end{aligned}$$

$$X_1, \dots, X_n \stackrel{iid}{\sim} P(\lambda) \rightarrow I(\theta) = ?$$

$$f(X|\lambda) = \frac{e^{-\lambda} \lambda^x}{x!}$$

$$\log f(X|\lambda) = -\lambda + x \log \lambda - \log x!$$

$$\frac{\partial}{\partial \lambda} \log f(X|\lambda) = -1 + \frac{X}{\lambda}$$

$$\frac{\partial^2}{\partial \lambda^2} \log f(X|\lambda) = -\frac{X}{\lambda^2}$$

$$\Rightarrow I(\theta) = -nE\left[-\frac{X}{\lambda^2}\right] = \frac{n}{\lambda}$$



$$\hat{\lambda}_{ML} = \frac{1}{n} \sum_{i=1}^n X_i$$

Bias:

$$E[\hat{\lambda}_{ML}] = \frac{1}{n} \sum_{i=1}^n E[X_i|\lambda] = n \frac{\lambda}{n} = \lambda$$

Variance:

$$\stackrel{(1)}{\Rightarrow} \text{var}[\hat{\lambda}_{ML}] = \frac{1}{n^2} \sum_{i=1}^n \text{var}(X_i|\lambda) = \frac{\lambda}{n}$$

We Knew:

$$\stackrel{(2)}{\Rightarrow} \text{var}(\hat{\lambda}) \geq \frac{1}{I(\lambda)} = \frac{n}{\lambda}$$

Thus:

$$\stackrel{(1),(2)}{\Rightarrow} \hat{\lambda}_{ML} \equiv UMVE$$

$$X_1, \dots, X_n \stackrel{iid}{\sim} U(0, \theta)$$

$$f(X|\theta) = \frac{1}{\theta}$$

$$\log f(X|\theta) = -\log \theta$$

$$I(\theta) = nE\left[\left(\frac{\partial}{\partial \theta} \log f(X|\theta)\right)^2\right] = \frac{n}{\theta^2}$$

According to CRB:

$$\text{var}(\theta) \geq \frac{\theta^2}{n}$$

Recall:

$$y = \max_i (X_i)$$

Bias Analysis:

$$f_y(y|\theta) = \frac{n y^{n-1}}{\theta^n}$$

$$E[y] = \int_0^\theta y f(y|\theta) dy = \int_0^\theta \frac{n}{\theta^n} y^n dy = \frac{n}{n+1} \theta$$

$$\Rightarrow E[y] \neq \theta$$

$$\hat{\theta} = \frac{n+1}{n}y$$

$$E[\hat{\theta}] = \theta$$

Variance Analysis (why?):

$$\text{var}(\hat{\theta}) = \text{var}\left(\frac{n+1}{n}y\right) = \frac{\theta^2}{n(n+2)}$$

$$\text{var}(\hat{\theta}) = \frac{\theta^2}{n(n+2)} \leq \frac{\theta^2}{n} = \frac{1}{I(\hat{\theta})}$$

# Rao-Blackwell Theorem

- The Rao-Blackwell theorem uses sufficiency to characterizes the transformation of an arbitrarily estimator into an estimator that is optimal by the mean-squared-error (MSE) criterion.
- Recall:  $x$  and  $y$  are random variables:

$$E[X] = E[E[X|Y]]$$

$$\text{var}(X) = \text{var}(E[X|Y]) + E[\text{var}(X|Y)]$$

---

## Rao-Blackwell Theorem:

Let  $w$  be unbiased for  $\theta$ , and let  $T$  be a sufficient statistic for  $\theta$ :

Define  $\phi(T) = E[w|T]$ , then:

$$E[\phi(T)] = \theta$$

$$\text{and } \text{var}(\phi(T)) \leq \text{var}_{\theta}(w).$$

# Rao-Blackwell Theorem

## Proof:

(1)  $\phi(T) = E_{\theta}(w|T)$  is an estimator because  $T$  is sufficient  
 $\Rightarrow$  conditional dist. of  $\underline{X}$  given  $T$  does not depend on  $\theta$

and  $w$  is a function of  $\underline{X}$  only:

$$E_{\theta}(\phi(T)) = E_{\theta}(E(w|T)) = E_{\theta}(w) = \theta$$

$$(2) \text{Var}_{\theta}(w) = \text{Var}_{\theta}[E(w|T)] + E_{\theta}[\text{Var}(w|T)]$$

$$= \text{Var}_{\theta}(\phi(T)) + \underbrace{E_{\theta}(\text{Var}(w|T))}_{\text{positive}} \geq \text{Var}_{\theta}(\phi(T))$$

# Outline of Week 7 Lectures

- Introduction to Optimal Frequentist Estimator
- Score and Fisher Information
- Cramer-Rao Lower Bound
- Rao-Blackwell Theorem
- **UMVUE**
- Bayesian Estimation
- Conjugate Prior
- Consistency
- Efficiency
- Estimator Comparison
- Summary

# UMVUE

**Example:**  $x_1, \dots, x_n \text{ iid } N(\mu, 1)$

Median  $(x_1, \dots, x_n)$  is unbiased.

However, it can't be UMVUE since it is not sufficient statistics (i.e. sufficient statistics is  $\bar{X}$ ).

---

## Theorem:

If  $w$  is an UMVUE of  $\theta$ , then  $w$  is unique.

$$(1) : W \leftarrow UMVE$$

$$(2) : W' \leftarrow UMVE$$

$$\Rightarrow W^* := \frac{W + W'}{2}$$

# UMVUE

**Proof:**

$$E[W^*] = E\left[\frac{w + w'}{2}\right] = \theta$$

$$\begin{aligned} \text{var}(W^*) &= \frac{1}{4}\text{var}(W) + \frac{1}{4}\text{var}(W') + \frac{1}{2}\text{cov}(W, W') \\ &\leq \frac{1}{4}\text{var}(W) + \frac{1}{4}\text{var}(W') + \frac{1}{2}\sqrt{\text{var}(W) \text{var}(W')} \\ &\leq \text{var}(W) \end{aligned}$$



# UMVUE

## **Theorem:**

Let  $T$  be a complete sufficient statistic for a parameter  $\theta$  and let  $\phi(T)$  be any unbiased estimator based only on  $T$ .

Then  $\phi(T)$  is the unique *UMVUE* for  $\theta$ .

## **2 strategies for finding *UMVUE*'s:**

(1) Let  $T$  be a complete sufficient statistics for  $\theta$ , find a function of  $T$ ,  $\phi(T)$ , such that  $E_{\theta}[\phi(T)] = \theta$ .

(2) Let  $T$  be a sufficient statistics and  $w$  be any unbiased estimator for  $\theta$ , compute  $\phi(T) = E(w|T)$

# UMVUE

**Example:**  $x_1, \dots, x_n$  iid  $Bern(\theta)$

We know  $\bar{X}$  is the *UMVUE* (CRB attained)

Showed  $T = \sum X_i$  is a complete suff. Stat. for  $\theta$ .

$$E(T) = n\theta \Rightarrow \phi(T) = \frac{T}{n}$$

---

**Example:**  $x_1, \dots, x_n$  iid  $N(\mu, \delta^2)$

Showed  $T = (T_1, T_2) = (\sum X_i, \sum X_i^2)$  is a complete suff. stat. for  $N(\mu, \delta^2)$

$$\text{Consider } (\bar{X}, S^2) = \left( \frac{T_1}{n}, \frac{1}{n-1} \left( T_2 - \frac{T_1^2}{n} \right) \right)$$

# UMVUE

**Example:**  $x_1, \dots, x_n$  iid  $p(\lambda)$

Interested in estimating  $\theta = e^{-\lambda} = P_\lambda(X = 0)$

$\sum x_i \sim p(n, \lambda)$  is a complete sufficient statistic and:

$\frac{\sum x_i}{n}$  is the UMUVE for  $\lambda$ .

---

# UMVUE

*Guess  $e^{-\bar{X}}$*

$$W(\underline{X}) = \begin{cases} 1 & X = 0 \\ 0 & X > 0 \end{cases}$$

$$E_{\lambda}(w) = e^{-\lambda} \rightarrow \text{unbiased}$$

*Compute  $E_{\lambda}(w|T)$ :*

$$\begin{aligned} \phi(t) &= E(w|T = t) = P_{\lambda} \left( X_1 = 0 \mid \sum_i^n X_i = t \right) \\ &= \frac{P_{\lambda}(X_1 = 0, \sum_i^n X_i = t)}{P_{\lambda}(\sum_i^n X_i = t)} = \frac{P_{\lambda}(X_1 = 0)P_{\lambda}(\sum_i^n X_i = t)}{P_{\lambda}(\sum_i^n X_i = t)} \end{aligned}$$

$$X_i \sim P(\lambda) \qquad \sum_{i=2}^n X_i \sim P((n-1)\lambda) \qquad \sum_{i=1}^n X_i \sim P(n\lambda)$$

# UMVUE

$$\Rightarrow \phi(t) = \frac{[e^{-\lambda}] \left[ e^{-(n-1)\lambda} \times \frac{[(n-1)\lambda]^t}{t!} \right]}{e^{-n\lambda} \times \frac{[n\lambda]^t}{t!}}$$

$$\therefore \phi(t) = \left( \frac{n-1}{n} \right)^t = \left( 1 - \frac{1}{n} \right)^t \text{ is UMUVE of } e^{-\lambda}$$

$$\text{We can write: } \phi(t) = \left( \frac{n-1}{n} \right)^t = \left( \left( 1 - \frac{1}{n} \right)^n \right)^{\frac{1}{n} \sum x_i}$$

$$\text{as } n \rightarrow \infty, \phi(t) \rightarrow e^{-\bar{X}}$$

# Outline of Week 7 Lectures

- Introduction to Optimal Frequentist Estimator
- Score and Fisher Information
- Cramer-Rao Lower Bound
- Rao-Blackwell Theorem
- UMVUE
- **Bayesian Estimation**
- Conjugate Prior
- Consistency
- Efficiency
- Estimator Comparison
- Summary

# Bayes estimation

## Bayes estimation

- Frequentists or classical estimation regards the parameter  $\theta$  as an unknown but fixed.

- Bayes: regards  $\theta$  as random variable, with prior distribution  $\pi(\theta)$ .

- Observe data  $x_1, \dots, x_n$
- Update the prior into a posterior distribution;  $\pi(\theta|X)$ .

- $$\pi(\theta|X) = \frac{f(X, \theta)}{m(X)} = \frac{f(X|\theta)\pi(\theta)}{m(X)}$$

$$m(x) = \int f(X|\theta)\pi(\theta)d\theta = \text{marginal dist. of } X$$

# Bayes estimation

**Example:**  $x_1, \dots, x_n$  iid  $Bernoulli(\theta)$ ,  $\theta \sim \text{beta}(\alpha, \beta)$

$$\pi(\theta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1 - \theta)^{\beta-1}$$

$$f(x|\theta) = \theta^{\sum x_i} (1 - \theta)^{n - \sum x_i}$$

$$m(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 \theta^{\sum x_i + \alpha - 1} (1 - \theta)^{n - \sum x_i + \beta - 1} d\theta$$

$$\begin{aligned} & \text{beta}\left(\sum x_i + \alpha, n - \sum x_i + \beta\right) \\ &= \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\sum x_i + \alpha) \Gamma(n - \sum x_i + \beta)}{\Gamma(n + \alpha + \beta)} \end{aligned}$$

$$\begin{aligned} \Gamma(\theta | x) &= \frac{f(x | \theta) \pi(\theta)}{m(x)} \\ &= \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \sum \theta^{\sum x_i + \alpha - 1} (1 - \theta)^{n - \sum x_i + \beta - 1} \times \frac{1}{m(\alpha)} \end{aligned}$$

$$\pi(\theta|X) \sim \text{beta}(\sum X_i + \alpha, n - \sum X_i + \beta)$$



# Bayes estimation

## Finding the posterior:

- (a) Calculate  $\pi(\theta)f(X|\theta)$
- (b) Factor into piece depending on  $\theta$  and piece not depending on  $\theta$ .
- (c) Drop piece not depending on  $\theta$ , multiply and divide by constants.
- (d)  $\pi(\theta|X)$  is  $k(X)$  times what is left.  
choose  $k(X)$  s.t.  $\int \pi(\theta|X) d\theta = 1$

# Bayes estimation

**Example:**  $x_1, \dots, x_n$  iid  $N(\mu, \delta^2)$ ,  $\delta^2$  known

$$f(x | \mu) = (2\pi\delta^2)^{-\frac{n}{2}} e^{-\frac{1}{2\delta^2} \sum (x_i - \mu)^2}$$

$$\Pi(\mu) = N(\mu_0, \delta_0^2)$$

$$\pi(\mu)f(x | \mu) = \left(\frac{1}{\sqrt{2\pi\delta^2}}\right)^n \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2\delta^2} \sum (x_i - \mu)^2} e^{-\frac{1}{2\delta_0^2} (\mu - \mu_0)^2}$$

$$\propto \exp \left[ -\frac{1}{2\delta_0^2} (\mu - \mu_0)^2 - \frac{1}{2\delta^2} \sum (x_i - \bar{x})^2 - \frac{1}{2\delta^2} n(\bar{x} - \mu)^2 \right]$$

$$= \exp \left[ -\frac{1}{2} \left( \frac{(\mu - \mu_0)^2}{\delta_0^2} + \frac{n(\bar{x} - \mu)^2}{\delta^2} \right) \right]$$

# Bayes estimation

$$\begin{aligned} &= \exp \left[ -\frac{1}{2} \left( \frac{(\mu - \mu_0)^2}{\delta_0^2} + \frac{n(\bar{x} - \mu)^2}{\delta^2} \right) \right] \\ &= \exp \left[ -\frac{1}{2} \left( \left( \frac{1}{\delta_0^2} + \frac{n}{\delta^2} \right) \mu^2 - 2\mu \left( \frac{\mu_0}{\delta_0^2} + \frac{n\bar{x}}{\delta^2} \right) + \frac{\mu_0^2}{\delta_0^2} + \frac{n\bar{x}^2}{\delta^2} \right) \right] \\ &= \frac{-1}{2} a \mu^2 - 2b\mu = \frac{-1}{2} a \left( \mu - \frac{b}{a} \right)^2 \\ &a = \frac{1}{\delta_0^2} + \frac{n}{\delta^2} \\ &b = \frac{\mu_0}{\delta_0^2} + \frac{n\bar{x}}{\delta^2} \end{aligned}$$

# Bayes estimation

$$\begin{aligned}\pi(\mu)f(x \mid \mu) &\propto \exp\left[-\frac{1}{2}a\left(\mu - \frac{b}{a}\right)^2\right] \\ &= N\left(\frac{b}{a}, \frac{1}{a}\right) \sim \pi(\mu \mid \underline{x})\end{aligned}$$

# Bayes estimation

**Bayes estimator:**

**(1) Maximum A Posteriori (MAP) Estimator:**

In Bayesian statistics, a maximum a posteriori probability (MAP) estimate is an estimate of an unknown quantity, that equals the mode of the posterior distribution.

# Bayes estimation

**Bayes estimator:**

**(1) Maximum A posteriori (MAP) Estimator:**

$$\hat{\theta}_{ML} = \underset{\theta}{\operatorname{argmax}} f(X|\theta)$$

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}} \Pi(X|\theta)$$

$$\Pi(X|\theta) \propto f(X|\theta)\pi(\theta)$$

# Bayes estimation

## (2) Bayes Minimum Loss (Risk) Estimator:

- Define a loss function  $L(\theta, \hat{\theta})$

$L(\theta, \hat{\theta}) = \text{loss of estimation } \theta \text{ by } \hat{\theta}$

- Minimize expected loss:

$$\min \int_{\Theta} L(\theta, \hat{\theta}) \pi(\theta|X) d\theta$$

then  $\hat{\theta} \sim$  Bayes minimum loss estimator.

# Bayes estimation

(1)  $L(\theta - \hat{\theta}) = (\theta - \hat{\theta})^2$  squared error loss

$$\Rightarrow E(\theta|X) = \hat{\theta}$$

(2)  $L(\theta - \hat{\theta}) = |\theta - \hat{\theta}|$  absolute error loss

$$\Rightarrow \hat{\theta} = \text{Median of } \pi(\theta|X)$$

---

**Example:**  $x_1, \dots, x_n$  iid  $N(\mu, \delta^2)$

Posterior is normal with mean:  $\left(\frac{\mu_0}{\delta_0^2} + \frac{n\bar{x}}{\delta^2}\right) / \left(\frac{1}{\delta_0^2} + \frac{n}{\delta^2}\right)$

And variance:  $1 / \left(\frac{1}{\delta_0^2} + \frac{n}{\delta^2}\right)$  using squared loss criterion.

$$\hat{\mu} = E(\mu | x) = \alpha \bar{x} + (1 - \alpha) \mu_0$$

$$\alpha = n/\delta^2 / \left(\frac{n}{\delta^2} + \frac{1}{\delta_0^2}\right) = \frac{n}{n + \frac{\delta^2}{\delta_0^2}}$$



# Bayes estimation

Note:

$$(1) \text{ as } n \rightarrow \infty, \alpha \rightarrow 1 \\ \Rightarrow E(\mu | x) \rightarrow \bar{x}$$

(2) prior information:

Let  $\delta_0^2 \rightarrow \infty$

$$\mu \sim N(\mu_0, \infty) \Rightarrow E(\mu | x) \rightarrow \bar{x}$$

(3) good prior info:

$$\text{Let } \delta_0^2 \rightarrow 0 \Rightarrow E(\mu | x) \rightarrow \mu_0$$

# Outline of Week 7 Lectures

- Introduction to Optimal Frequentist Estimator
- Score and Fisher Information
- Cramer-Rao Lower Bound
- Rao-Blackwell Theorem
- UMVUE
- Bayesian Estimation
- Conjugate Prior
- Consistency
- Efficiency
- Estimator Comparison
- Summary

# Conjugate Prior

In Bayesian probability theory, if the posterior distribution  $p(\theta | x)$  is in the same probability distribution family as the prior probability distribution  $\pi(\theta)$ , the prior and posterior are then called conjugate distributions, and the prior is called a conjugate prior for the likelihood function  $p(x | \theta)$ .

## Examples:

Conjugate Prior	Likelihood	Posterior
Beta	Bernoulli	Beta
Gamma	Poisson	Gamma
Normal	Normal	Normal

# Problems with Bayes Estimator

choice of prior:

- subjective
- non informative priors

Prior:  $\pi(\gamma) = 1 \quad \forall \gamma$

Posterior:  $N(\bar{X}, \frac{\sigma^2}{n})$

What can we do when we do not have the prior?

# Jeffreys Prior

**Jeffreys Prior:** is a non-informative (objective) prior distribution for a parameter space; its density function is proportional to the square root of the determinant of the Fisher information matrix:

---

Example:  $x_1, \dots, x_n$  iid  $Bern(\theta)$

$$\log(f(X|\theta)) = x \log \theta + (1 - x) \log(1 - \theta)$$

$$\frac{\partial}{\partial \theta} \log(f(X|\theta)) = \frac{x}{\theta} - \frac{1 - x}{1 - \theta} \rightarrow \frac{\partial^2}{\partial \theta^2} \log(f(X|\theta)) = \frac{-x}{\theta^2} + \frac{1 - x}{(1 - \theta)^2}$$

$$E_{\theta} \left[ \frac{\partial^2}{\partial \theta^2} \log(f(X|\theta)) \right] = -\frac{1}{\theta} - \frac{1}{1 - \theta} = -\frac{1}{\theta(1 - \theta)}$$

$$\pi(\theta) \propto \frac{1}{\theta(1 - \theta)^{\frac{1}{2}}} \text{ i.e. } \text{beta}\left(\frac{1}{2}, \frac{1}{2}\right)$$

# Outline of Week 7 Lectures

- Introduction to Optimal Frequentist Estimator
- Score and Fisher Information
- Cramer-Rao Lower Bound
- Rao-Blackwell Theorem
- UMVUE
- Bayesian Estimation
- Conjugate Prior
- Consistency
- Efficiency
- Estimator Comparison
- Summary

# Consistency

Why do frequentists use MLE's?

- MLE's have nice asymptotic properties

**Def:** a sequence of estimators:

$w_n = w_n(x_1, \dots, x_n)$  is a consistent sequence of estimators of the parameter  $\theta$  if for any  $\epsilon > 0$ ,  $\theta \in \Theta$ :

$$\lim_{n \rightarrow \infty} P_{\theta}(|w_n - \theta| < \epsilon) = 1$$

or: 
$$\lim_{n \rightarrow \infty} P_{\theta}(|w_n - \theta| \geq \epsilon) = 0$$

(it means  $w_n$  converges to  $\theta$  in probability)

# Consistency

## Theorem:

If  $w_n$  is a sequence of estimators of a parameter  $\theta$  with:

- (a)  $\lim_{n \rightarrow \infty} \text{Var}_{\theta}(w_n) = 0$  and
- (b)  $w_n$  unbiased estimator of  $\theta$

Then  $w_n$  is a consistent sequence of estimators of  $\theta$ .

Proof:

$$\text{Chebychev} \Rightarrow P_{\theta}(|w_n - \theta| \geq \varepsilon) \leq \frac{E_{\theta}(w_n - \theta)^2}{\varepsilon^2}$$

$$E_{\theta}(w_n - \theta)^2 = E_{\theta}(w_n + Ew_n - Ew_n - \theta)^2$$

$$= \text{Var}_{\theta} w_n + (\text{Bias}_{\theta} w_n)^2$$



# Consistency

- MLE's are consistent
- MLE's are asymptotically unbiased

## Theorem:

Let  $x_1, \dots, x_n$  iid  $f(X|\theta)$ .

Let  $L(\theta|X) = \prod f(X_i|\theta)$

$\hat{\theta} = \text{MLE of } \theta$

Then with some regularity conditions on  $f(X|\theta)$  we have:

$\hat{\theta}_n$  is a consistent estimator of  $\theta$ .

Condition: support of pdf does not depend on parameters and rules out  $U(0, \theta)$

# Outline of Week 7 Lectures

- Introduction to Optimal Frequentist Estimator
- Score and Fisher Information
- Cramer-Rao Lower Bound
- Rao-Blackwell Theorem
- UMVUE
- Bayesian Estimation
- Conjugate Prior
- Consistency
- **Efficiency**
- Estimator Comparison
- Summary

# Efficiency

- Let  $I(\theta) = E_{\theta} \left( \frac{\partial}{\partial \theta} \log f(X|\theta) \right)^2$  and  $X$  is not a vector.

**Def:**

Let  $w$  be an unbiased estimator of  $\theta$ . The efficiency of  $w$  is:

$$eff(w) = \frac{\left[ \frac{1}{n} I(\theta) \right]}{var(w)} \longrightarrow \text{CRB lower bound}$$

---

# Efficiency

## Definition:

A sequence of estimators  $w$  is said to be asymptotically efficient if:

$$\lim_{n \rightarrow \infty} \text{eff}(w_n) \rightarrow 1$$

As  $n \rightarrow \infty$ ,  $\text{var } w_n$  attains CR lower bound.

- 
- MLE's are asymptotically efficient.
  - MLE's are asymptotically normal.

i.e.  $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{D} N\left(0, \frac{1}{I(\theta)}\right)$

❖ MLE's are:

- (1) Consistent
- (2) asymptotically unbiased
- (3) asymptotically efficient
- (4) asymptotically normal

# Outline of Week 7 Lectures

- Introduction to Optimal Frequentist Estimator
- Score and Fisher Information
- Cramer-Rao Lower Bound
- Rao-Blackwell Theorem
- UMVUE
- Bayesian Estimation
- Conjugate Prior
- Consistency
- Efficiency
- Estimator Comparison
- Summary

# Asymptotic variance of MLE

## Asymptotic variance of MLE

$$eff(\hat{\theta}_n) = \frac{1/n I(\theta)}{var(\hat{\theta}_n)} \longrightarrow 1$$

Approximate  $var(\hat{\theta}_n)$  by  $nI(\theta) \leftrightarrow$  expected information

$nI(\theta)|_{\theta=\hat{\theta}} \leftarrow$  observed info.

Better approximation for finite sample sizes.

# Asymptotic variance of MLE

**Expected information:**

$$\begin{aligned} nI(\theta) &= nE_{\theta} \left( \frac{\partial}{\partial \theta} \log f(X|\theta) \right)^2 \\ &= E_{\theta} \left( \frac{\partial}{\partial \theta} \log \prod f(X_i|\theta) \right)^2 = E_{\theta} \left( \frac{\partial}{\partial \theta} \log L(\theta|X) \right)^2 \end{aligned}$$

**Approximation:** if  $x_1, \dots, x_n$  are iid  $f(X|\theta)$ ,  $\hat{\theta}$  is the MLE of  $\theta$ .

$$\text{var}_{\theta}(\hat{\theta}) \simeq \frac{1}{E_{\theta} \left[ \frac{\partial}{\partial \theta} \log L(\theta|X) \right]^2} \simeq \frac{1}{-\frac{\partial^2}{\partial \theta^2} \log L(\theta|X) |_{\theta=\hat{\theta}}} \quad (*)$$

# Asymptotic variance of MLE

**Example:**  $x_1, \dots, x_n$  are iid from  $\text{Bern}(\theta)$

MLE is  $\hat{p} = \bar{X}$

$$\text{Var}(\hat{p}) = \frac{p(1-p)}{n}$$

$$\widehat{\text{Var}} \hat{p} = \frac{\hat{p}(1 - \hat{p})}{n} \quad \text{an approximated variance}$$

$$\text{Use } (*) \rightarrow \widehat{\text{Var}} \hat{p} \approx \frac{1}{-\frac{\partial^2}{\partial \theta^2} \log L(p|x)|_{p=\hat{p}}}$$

$$\log L = \sum x_i \log p + \left( n - \sum x_i \right) \log(1 - p)$$

$$\frac{\partial^2}{\partial \theta^2} \log L = -\frac{n\bar{X}}{p^2} - \frac{n(1 - \bar{X})}{(1 - p)^2}$$



## Asymptotic variance of MLE

$$\Rightarrow \frac{\partial^2}{\partial \theta^2} \log L|_{p=\hat{p}} = -\frac{n\bar{X}}{\bar{X}^2} - \frac{n(1-\bar{X})}{(1-\bar{X})^2} = -\frac{n}{\bar{X}(1-\bar{X})}$$

$$(*) \text{ also gives: } \widehat{Var} \hat{p} = \frac{\bar{X}(1-\bar{X})}{n}$$

# Estimator Comparison

- Frequentists:  $\min E_{\theta}(\hat{\theta} - \theta)^2$

**Example:**  $x_1, \dots, x_n$  iid  $N(\mu, \delta^2)$ , want to estimate  $\delta^2$

**MLE**  $\widehat{\delta}_1^2 = \frac{s}{n}$  when  $s = \sum (x_i - \bar{x})^2$

**Bayes(Jeffery's prior)**  $\pi(\delta^2) \propto \frac{1}{s^2}$   $\widehat{\delta}_2^2 = \frac{s}{n-2}$

**UMVUE**  $\widehat{\delta}_3^2 = \frac{s}{n-1}$

# Estimator Comparison

$$E[aS - \delta^2]^2 = a^2 E(s^2) - 2a\delta^2 ES + \delta^4$$

$$= a^2 \text{Var}(s) + a^2 [E(s)]^2 - 2a\delta^2 ES + \delta^4$$

$$\frac{s}{\delta^2} \sim X_{n-1}^2 \Rightarrow E(S) = (n-1)\delta^2$$

$$\text{Var}(S) = 2(n-1)\delta^4$$

$$E[as - \delta^2]^2 = \delta^4 [a^2(n-1)(n+1) - 2a(n-1) + 1]$$

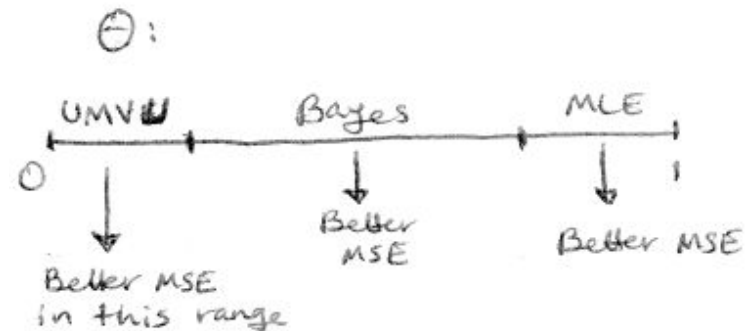
$$\text{Minimized by: } a = \frac{1}{n+1}, \quad \widehat{\delta^4} = \frac{s}{n+1}$$

# Estimator Comparison

estimator				
MSE				

theta	k1	MLE	Bayes	UMVUE
0.10	2	0.0258	0.0250	0.0148
0.20	4	0.0171	0.0169	0.0125
0.30	6	0.0159	0.0151	0.0134
0.40	8	0.0154	0.0140	0.0141
0.50	10	0.0142	0.0126	0.0138
0.60	12	0.0127	0.0110	0.0128
0.70	14	0.0105	0.0090	0.0109
0.80	16	0.0077	0.0067	0.0082
0.90	18	0.0042	0.0038	0.0045
0.95	19	0.0021	0.0022	0.0023

\* Mean squared error



# Estimator Comparison

**Example:** let  $R$  = #of tosses needed to reach  $k$  heads,  $\theta = p(\text{head})$

$$P[R = r] = \binom{r-1}{k-1} \theta^k (1-\theta)^{r-k} \quad r = k, k+1, \dots$$

$R$  has negative binomial distribution.

(1) **MLE**      $\widehat{\theta}_1 = \frac{k}{r}$

(2) **Bayes**      $\pi(\theta) \propto [\theta(1-\theta)]^{-\frac{1}{2}}$

$$\Rightarrow \pi(\theta|R) \propto \theta^{k-\frac{1}{2}}(1-\theta)^{r-k-\frac{1}{2}}$$

$$\Rightarrow \widehat{\theta}_2 = E(\theta|R) = \frac{k + \frac{1}{2}}{r + 1}$$

# Estimator Comparison

(3) **UMVUE:**  $r$  is complete and sufficient for  $\theta$ :

$$E \left[ \frac{1}{r-1} \right] = \frac{\theta}{k-1}$$

$$\Rightarrow \widehat{\theta}_3 = \frac{k-1}{r-1} \quad \text{which is the UMVUE of } \theta$$

# Outline of Week 7 Lectures

- Introduction to Optimal Frequentist Estimator
- Score and Fisher Information
- Cramer-Rao Lower Bound
- Rao-Blackwell Theorem
- UMVUE
- Bayesian Estimation
- Conjugate Prior
- Consistency
- Efficiency
- Estimator Comparison
- Summary

# Summary

## (1) *Likelihood:*

Estimate  $\theta$  by the value  $\hat{\theta}$  which maximizes the likelihood

## (2) *Bayes:*

Let  $\pi(\theta)$  be a prior distribution for  $\theta$  leading to a posterior  $\pi(\theta|\underline{X})$

Let  $L(\theta, \hat{\theta})$  be a loss function. Choose  $\hat{\theta}$  to minimize:  $\int_{\theta} L(\theta, \hat{\theta}) \pi(\theta|X) d\theta$

$$L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2 \Rightarrow \hat{\theta} = E[\theta|X]$$

$$L(\theta, \hat{\theta}) = |\theta - \hat{\theta}| \Rightarrow \hat{\theta} = \text{median of } \pi(\theta|X)$$



# Summary

## **(3)Frequentist:**

(a) If possible, find the UMVUE of  $\theta$

(b) If (a) hard, use the MLE  $\hat{\theta}$  which is asymptotically unbiased and whose efficiency  $\rightarrow 1$  as  $n \rightarrow \infty$

---

(1), (2) and (3) may not exist!

## **Example:**

*UMVUE: Bern( $p$ ). Then  $\theta = \frac{p}{1-p} \Rightarrow$  UMVUE of  $\theta$  does not exist*

# Summary

- *MLE* and *Bayes* may not be unique, but *UMVUE* is unique.
- *MLE* has invariance property, *UMVUE* and *Bayes* do not.
- *Bayes*: incorporate prior information, but *MLE* and *UMVUE* don't.