Slide 8: More about Unbiased Estimator and MLE

STATS 511: Statistical Inference

Kean Ming Tan

Unbiased Estimator

Definition: An estimator $\hat{\theta}$ for θ is bias if $E(\hat{\theta}) \neq \theta$. The **bias** can be expressed by

$$Bias(\theta) = E_\theta(\hat{\theta}) - \theta.$$

If $Bias(\hat{\theta}) = 0$, then we say that $\hat{\theta}$ is an **unbiased estimator** of θ.

Example: Let $X_1, \ldots, X_n \sim \mathrm{Poisson}(\lambda)$. Then $\bar{X}$ and S^2 are both unbiased estimators of λ.

Question: Is there a principled way to find the best unbiased estimator for a parameter?

Improving an Unbiased Estimator

Rao-Blackwell Theorem (Theorem 7.3.17): Let $T(\mathrm{X})$ be a sufficient statistic for θ. Let $U(\mathrm{X})$ be an unbiased estimator of θ. Define

$$\hat{\theta} = E_\theta[U(\mathrm{X}) \mid T(\mathrm{X})].$$

Then $E_\theta[\hat{\theta}] = \theta$ and $Var_\theta(\hat{\theta}) \leq Var_\theta[U(\mathrm{X})]$ for all θ. That is $\hat{\theta}$ is a ~~uniformly~~ better unbiased estimator of θ. compare to $U(x)$

$$\mathbb{E}[U(x)] = \theta$$

$$\mathbb{E}[\hat{\theta}] = \mathbb{E}[\mathbb{E}[U(x) \mid T(x)]] = \mathbb{E}[U(x)] = \theta$$

$$\text{var}(U(x)) = \text{var}(\mathbb{E}[U(x) \mid T(x)]) + \mathbb{E}[\text{var}(U(x) \mid T(x))]$$

$$= \text{var}(\hat{\theta}) + \mathbb{E}[\text{var}(U(x) \mid T(x))] \text{ (NON zero)}$$

$$\geq \text{var}(\hat{\theta})$$

Equality hold's only if $U(x)$ is a function of $T(x)$

Stronger Version of Theorem

Theorem: Let $T(\mathrm{X})$ be a **complete sufficient statistic** for θ. Let $U(\mathrm{X})$ be an unbiased estimator of θ. Define

$$\hat{\theta} = E_\theta[U(\mathrm{X}) \mid T(\mathrm{X})].$$

Then $\hat{\theta}$ is the **best unbiased estimator** of θ.

indicates that $Var(\hat{\theta})$ is the smallest possible among all possible unbaised estimates of θ

Example: Uniform Distribution

Let $X_1, \ldots, X_n \sim Unif(0, \theta]$. From previous lecture, we know that $X_{(n)}$ is a complete sufficient statistic. How do we find the best unbiased estimator for θ?

$\frac{X_{(n)}}{\theta} \sim Beta(n, 1)$ $\frac{X_{(1)}}{\theta} \sim Beta(1, n)$

$E[X_{(n)}/\theta] = \frac{n}{n+1} \Rightarrow E[X_{(n)}] = \frac{n\theta}{n+1}$

$X_{(n)}$ is C.S.S

By Rao-Blackwell: $\hat{\theta} = E[\underset{}{u(x)} \mid \underbrace{T(x)}_{\text{C.S.S}}]$ is the BUE of θ

Example: Poisson Distribution

Let $X_1, \ldots, X_n \sim Poisson(\lambda)$. We want to estimate

$$\theta = e^{-\lambda} = P(X_1 = 0).$$

What is the best unbiased estimator of θ?

Easy way to find BUE

Theorem 7.3.23: Let $T(\mathrm{X})$ be a complete sufficient statistic for θ. Let $\phi[T(\mathrm{X})]$ be any estimator based only on $T(\mathrm{X})$. Then, $\phi[T(\mathrm{X})]$ is the unique best unbiased estimator of its expected value, i.e., $E_\theta[\phi(T(\mathrm{X}))] = \tau(\theta)$.

Example: Gaussian

Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. So $T = (\bar{X}, S^2)$ are complete and sufficient for (μ, σ^2). What are the BUE for μ, σ^2, μ^2, respectively?

Fisher Information

Introduction to Fisher Information

Method for measuring the amount of information that a sample of data contains about an unknown parameter. This measure has the intuitive properties that more data provide more information, and more precise data provide more information. The information measure can be used to find bounds on the variances of estimators, and it can be used to approximate the variances of estimators obtained from large samples. More data $\longrightarrow$ Higher F.I

F.I Used to Calculate lower Bound

Fisher Information

Fisher Information: Let X be a random variable whose distribution depends on a parameter θ. Let $f(x \mid \theta)$ be the pdf of X. Under some regularity conditions, the Fisher information $I(\theta)$ of X is defined as

$$I_X(\theta) = E_\theta \left\{ \left[\frac{\partial}{\partial \theta} \log f_\theta(X) \right]^2 \right\} \geq 0$$

Theorem: Under some regularity condition, the Fisher Information can also be calculated as

$$I_X(\theta) = -E_\theta \left\{ \frac{\partial^2}{\partial \theta^2} \log f_\theta(X) \right\}$$

Some Properties and Proof

- For nearly almost any distribution, we have

$$E_\theta\left(\frac{\partial}{\partial\theta}\log f_\theta(X)\right)=0$$

- Let $f_n(\boldsymbol{x}\mid\theta)$ be the joint distribution of $X_1,\ldots,X_n$. Then we have

$$I_n(\theta)=-E_\theta\left\{\frac{\partial^2}{\partial\theta^2}\log f_n(\boldsymbol{x}\mid\theta)\right\}$$

- Suppose that $X_1,\ldots,X_n$ are a set of n i.i.d. observations wtih $X_i\sim f(x\mid\theta)$, a regular 1-parameter family. Then the information number for the data $X_1,\ldots,X_n$ is

$$I_n(\theta)=nI_X(\theta)$$

Proof

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

$$= \int_{\mathcal{X}} \frac{f'(x|\theta)}{\cancel{f(x|\theta)}}\, \cancel{f(x|\theta)}\, dx$$

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

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

$$\mathbb{E}\left[\frac{\partial}{\partial\theta}\log f(\theta|x)\right] = 0$$

② WTS $I_x(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f(x|\theta)\right)^2\right]$

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

Proof:

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

$$\frac{\partial^2}{\partial\theta^2}\log f(x|\theta) = \frac{\partial}{\partial\theta}\left[\frac{\frac{\partial}{\partial\theta} f(x|\theta)}{f(x|\theta)}\right]$$

$$= \frac{f(x|\theta)\frac{\partial^2}{\partial\theta^2} f(x|\theta) - \left(\frac{\partial}{\partial\theta} f(x|\theta)\right)^2}{f(x|\theta)^2}$$

$$= \frac{\frac{\partial^2}{\partial\theta^2} f(x|\theta)}{f(x|\theta)} - \left[\frac{\frac{\partial}{\partial\theta} f(x|\theta)}{f(x|\theta)}\right]^2$$

$$\mathbb{E}\left[\frac{\partial^2}{\partial\theta}\log f(x|\theta)\right] = \int_x \left(\frac{\frac{\partial^2}{\partial\theta} f(x|\theta)}{f(x|\theta)} - \left(\frac{\frac{\partial}{\partial\theta} f(x|\theta)}{f(x|\theta)}\right)^2\right) f(x|\theta)\,dx$$

$$= \cancelto{0}{\int_{\mathcal{X}} \frac{\partial^2}{\partial\theta^2} f(x|\theta)\, dx} - \int_{\mathcal{X}} \left(\frac{\frac{\partial}{\partial\theta} f(x|\theta)}{f(x|\theta)} \right)^2 f(x|\theta)\, dx$$

$$= -\int_{\mathcal{X}} \left(\frac{\partial}{\partial\theta} \log f(x|\theta) \right)^2 f(x|\theta)\, dx$$

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

$$\Longrightarrow$$

$$\boxed{\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2} \log f(x|\theta) \right] + \mathbb{E}\left[\left(\frac{\partial}{\partial\theta} \log f(x|\theta) \right)^2 \right] = 0}$$

③ $I_n(\theta) = n \cdot I_x(\theta)$

$$f_n(x|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

$$I_n(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta} \log f_n(x|\theta) \right)^2 \right]$$

$$= \mathrm{var}\left(\frac{\partial}{\partial\theta} \log f_n(x|\theta) \right)$$

Example: Poisson Distribution

Suppose that $X_1, \ldots, X_n \sim \text{Poisson}(\lambda)$. What is the Fisher Information $I(\lambda)$ of X?

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

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

$$\frac{d}{d\lambda}\log f(x|\lambda) = -1 + \frac{x}{\lambda}$$

$\underline{\underline{1^{st}}}$:

$$I_x(\lambda) = \mathbb{E}\left[\left(\frac{d}{d\lambda}f(x|\lambda)\right)^2\right]$$

$$= \mathbb{E}\left[\left(-1+\frac{x}{\lambda}\right)^2\right] = \mathbb{E}\left[\frac{x^2}{\lambda^2} - \frac{2x}{\lambda} + 1\right]$$

$E[x] = \lambda \quad var(X) = \lambda \quad \Rightarrow \quad \mathbb{E}[x^2] = var(x) + \mathbb{E}[x]^2$

$= \lambda + \lambda^2$

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

$$I_x(\lambda) = -\mathbb{E}\left[\frac{\partial^2 \log f(x|\theta)}{\partial \lambda^2}\right] = -\mathbb{E}\left[\frac{-x}{\lambda^2}\right]$$

$$= \frac{1}{\lambda^2}\mathbb{E}[x] = \frac{1}{\lambda}$$

Fisher Information Matrix

Fisher Information for a vector of parameter: Suppose that $X_1, \ldots, X_n$ form a random sample from a distribution for which the p.d.f. is $f(x \mid \boldsymbol{\theta})$ where the value of the parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ must lie in an open subset of a k-dimensional real space. Let $f_n(\mathrm{x} \mid \boldsymbol{\theta})$ be the joint pdf and let $\log f_n(\mathrm{x} \mid \boldsymbol{\theta})$. Under some conditions, the Fisher information matrix $I_n(\boldsymbol{\theta})$ for X is a $k \times k$ matrix with i, j element equal to

$$I_{ij}(\boldsymbol{\theta}) = E_\theta \left\{ \left[\frac{\partial}{\partial \theta_i} \log f_\theta(X \mid \boldsymbol{\theta}) \cdot \frac{\partial}{\partial \theta_j} \log f_\theta(X \mid \boldsymbol{\theta}) \right] \right\}.$$

$$I_{n,ij}(\boldsymbol{\theta}) = n \cdot I_{ij}(\boldsymbol{\theta})$$

Example: Normal Distribution

Suppose that $X \sim N(\mu, \tau)$ with μ and $\tau = \sigma^2$ being unknown.
What is the Fisher Information matrix $I(\mu, \tau)$ of X?

Cramer Rao Lower Bound

Suppose that $T(\mathrm{X})$ is an unbiased estimator of $b(\theta)$. That is, $E(T(\mathrm{X})) = b(\theta)$ for all θ. Assume that $b(\theta)$ is differentiable. Then

$$Var(T(\mathrm{X})) \geq \frac{(b'(\theta))^2}{nI_X(\theta)}.$$

Equality holds if and only if $f(X \mid \theta)$ is a 1-parameter exponential family distribution.

Specifically, the equality holds if and only if

$$\frac{\partial}{\partial\theta}\log f(\mathrm{X} \mid \theta) = a(\theta) + g(\theta)T(\mathrm{X})$$

Special Case

Let $T(\mathrm{X})$ be an unbiased estimator for θ. Then

$$Var(T(\mathrm{X})) \geq \frac{1}{nI_X(\theta)}.$$

Interpretation: The variance of an unbiased estimator of θ cannot be smaller than the reciprocal of the Fisher information in the sample.

Example: Exponential Distribution

Let $X_1, \ldots, X_n$ be random sample from an exponential distribution with parameter λ and density $f(x \mid \lambda) = \lambda \exp(-\lambda x)$. Calculate the Fisher Information. Consider the estimator $T = (n-1)/\sum X_i$. Does this estimator achieves the smallest variance using the Cramer Rao Inequality? What if we want to estimate $m(\lambda) = 1/\lambda$? Is $\bar{X}$ a good estimator for $m(\lambda) = 1/\lambda$?

Example: Binomial Distribution

Suppose that $X \sim \mathrm{Binomial}(n, \theta)$. What is the Fisher Information $I(\theta)$ of X? Suppose that we are interested in estimating $\tau(\theta) = \theta(1 - \theta)$. What is the BUE? Does it achieves the CRLB?

Example: Sampling from Poisson

Let $X_1, \ldots, X_n$ be random sample from Poisson distribution with parameter θ. Remember that we have a lot of different unbiased estimators for estimating θ. Which estimator is the best estimator for θ?

multivariable CRLB

Paratemeter of interest.

let $(\theta_1, \dots, \theta_k) \in \mathbb{R}^k$ $g(\theta)$ is smooth function

let $\nabla_\theta g(\theta) = \begin{bmatrix} \frac{\partial g}{\partial \theta_1} \\ \frac{\partial g}{\partial \theta_2} \\ \vdots \\ \frac{\partial g}{\partial \theta_k} \end{bmatrix} = \left(\frac{\partial g}{\partial \theta_1}, \dots \frac{\partial g}{\partial \theta_k} \right)^T$

let $I(\theta) = \{ I_{ij}(\theta) \mid i, j = 1, 2, \dots k \} \in \mathbb{R}^{k \times k}$

Fisher information matrix

CRLB: Assume that $I_n(\theta) > 0$

PD matrix

for any real valued $T(x)$ in t

$\mathbb{E}[T(x)] < \infty$,

$$\mathrm{var}(T(x)) \geq \left(\nabla_\theta E_\theta[T(x)] \right)^T \left(I_n(\theta) \right)^{-1} \left(\nabla_\theta E_\theta[T(x)] \right)$$

$\nabla_\theta E_\theta[T(x)] = \nabla_\theta g(\theta)$

* Suppose we are only interested in $T(\theta_1)$. This case $(\theta_2, \theta_3, \ldots \theta_k)$ is nuisance parameter.

* Suppose that $(\theta_2, \theta_3, \ldots \theta_k)$ are known.

$\Downarrow$ Univariate case

By CRLB is $\operatorname{var}(T(x)) \geq \dfrac{\left(\frac{d\gamma}{d\theta_1}\right)^2}{I_{11}(\theta)}$

---

if $(\theta_2, \ldots \theta_k)$ are unknown, we need to use the multivariate case.

Still interested in $T(\theta_1)$, Not interested in $T(\theta_2), \ldots T(\theta_k)$ even though they are unknown.

$\operatorname{Var}(T(x)) \geq$

$$\left(\frac{\partial T(\theta_1)}{\partial \theta_1}, 0, 0 \ldots\right) I_n(\theta)^{-1} \left(\frac{dT(\theta_1)}{d\theta_1}, 0, 0 \ldots 0\right)^T$$

$$= \frac{\left(\frac{\partial \gamma(\theta_1)}{\partial \theta_1}\right)^2}{I_n(\theta) - I_{12}(\theta) I_{22}^{-1}(\theta) I_{21}(\theta)}$$

$\Rightarrow I_{11.2}(\theta) \geq 0$

$$I_n(\theta) = \begin{array}{cc} 1\times 1 & 1\times k-1 \\ \left[\begin{array}{cc} I_{11}(\theta) & I_{12}(\theta) \\ I_{21}(\theta) & I_{22}(\theta) \end{array}\right] & \\ k-1\times 1 & (k-1)\times(k-1) \end{array}$$

$$I_{11}(\theta) \geq I_{11.2}(\theta)$$

Therefore $\dfrac{(\tau'(\theta_1))^2}{I_{11}(\theta)} \leq \dfrac{(\tau'(\theta_1))^2}{I_{11.2}(\theta)}$

(univariate case) $(\theta_2, \theta_3, \ldots \theta_k)$ are known

$(\theta_2, \theta_3, \ldots \theta_k)$ are unknown

<u>Example:</u>

$X_1, X_2, \ldots X_n \sim N(\mu, \sigma^2)$

$\theta_1 \quad \theta_2$

$$I_x(\theta_1, \theta_2) = \begin{bmatrix} \frac{1}{\theta_2} & 0 \\ 0 & \frac{1}{2\theta_2^2} \end{bmatrix}$$

Because $I_{12} = 0$, $\theta_1 = \mu$, $\theta_1 = \sigma^2$ are orthogonal

So the CRLB for each of them is the same regardless of where one of them is known or unknown.

CRLB for μ is $\frac{1}{nI_{11}} = \frac{\theta_1}{n} = \frac{\sigma^2}{n}$

" " σ^2 is $\frac{1}{nI_{22}} = \frac{2\theta_2^2}{n} = \frac{2\sigma^4}{n}$

S^2 is unbaised estimator of σ^2

$Var(S^2) = \frac{2\sigma^4}{n-1}$ (obtained using $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$)

$\Rightarrow Var(S^2) = \frac{2\sigma^4}{n-1} > \frac{2\sigma^4}{n}$

So, S^2 does not achieve CRLB.

Q:- Is S^2 the BLE of σ^2?

Summary of CRLB

Suppose that in a given problem a particular estimator T is an efficient estimator of its expectation $m(\theta)$, and let T_1 denote any other unbiased estimator of $m(\theta)$. Then for every value of θ in the parameter space, Var (T) will be equal to the lower bound provided by the information inequality, and Var (T_1) will be at least as large as that lower bound. Hence, $\text{Var}(T) \leq \text{Var}(T_1)$. In other words, if T is an efficient estimator of $m(\theta)$, then among all unbiased estimators of $m(\theta)$, T will have the smallest variance for every possible value of θ.

Asymptotic Properties of MLE

Asymptotic Properties of MLE

what happen's to our MLE when $n \longrightarrow \infty$.

Theorem: Suppose that in an arbitrary problem the M.L.E. $\hat{\theta}$ is determined by solving the equation $\partial log f_n(\mathrm{x}|\theta)/\partial\theta = 0$. Under certain regularity conditions, the asymptotic distribution of the MLE is

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N\left(0, (I(\theta))^{-1}\right)$$

In this case, the MLE is an asymptotically efficient estimator.

so, for the problem's that we are unable to get MLE this way, this theorem doesn't hold. Ex: unif

Interpretation: The MLE always achieves the CRLB (minimum variance) asymptotically as $n \to \infty$.

* The above Theorem cannot be used for the Problem's where we cannot derive MLE using $\frac{\partial}{\partial \theta} \log f(x|\theta) = 0$. Ex: uniform

* we need to be able to interchange the integral and differentiation order, for CRLB computation. and that actually requires that our range doesn't depend on θ. So uniform Range of x depend on θ.

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{(d)} N(0, I(\theta)^{-1})$$

$\frac{1}{I(\theta)}$ fisher info for one sample.

$\Rightarrow$ if $n \to \infty$, The variance of $\hat{\theta}_{MLE}$ goes to CRLB.

$$Var(T(x)) \geq \frac{1}{n I_x(\theta)}$$

$$\Rightarrow \quad Var(T(x) - \mathbb{E}[T(x)]) \geq \frac{1}{n I_x(\theta)}$$

$$\Rightarrow \quad n \operatorname{var}\left(T(x) - E[T(x)]\right) \geq \frac{1}{I_x(\theta)}$$

$$\Rightarrow \quad \operatorname{Var}\left(\sqrt{n}\left(T(x) - E[T(x)]\right)\right) \geq \frac{1}{I_x(\theta)}$$

$\Rightarrow$ The variance $\hat{\theta}$ attain's CRLB asymptotically. even though its biased.

$$\hat{\theta} \xrightarrow{P} \theta \qquad \operatorname{var}(\hat{\theta}) \longrightarrow \frac{1}{I_x(\theta)}$$

* Asymptotically MLE is the most efficient Estimator. (efficient when n is large)

Example: Poisson

Let $X_1, \ldots, X_n$ be random sample from Poisson distribution with parameter θ. What is the MLE of θ? What is the asymptotic distribution of θ?

$\hat{\theta}_{MLE} = \bar{X}$ is the MLE of θ

$$I_X(\theta) = \frac{1}{\theta}$$

CLT: $\sqrt{n}(\bar{X} - \theta) \xrightarrow{(d)} N(0, \theta)$

Asymptotic distribution of MLE:-

$$\sqrt{n}(\bar{X} - \theta) \xrightarrow{d} N\left(0, \frac{1}{I_X(\theta)}\right)$$

$$\Rightarrow \sqrt{n}(\bar{X} - \theta) \xrightarrow{d} N(0, \theta)$$

same

However, Not all example's are like this, It's Just coincidence that CLT is exactly same as Asymptotic distit

Some Important Regularity Conditions

We are really talking about a regular family of distributions:

range of x

- The support of the density cannot depend on θ.
- The density function is differentiable everywhere.

Proof:

we want to show:

$$\sqrt{n}\left(\hat{\theta}_{MLE} - \theta\right) \xrightarrow{(d)} N\left(0, \frac{1}{I_X(\theta)}\right)$$

FACT: $\hat{\theta}_{MLE}$ is the root of the log-likelihood function

$$\frac{\partial l(\hat{\theta}_{MLE})}{\partial \theta} = 0$$

(Apply 2nd order taylor expansion)

$$0 = \frac{\partial l(\hat{\theta})}{\partial \theta} \qquad \text{(Fact for } \hat{\theta}_{MLE}\text{)}$$

$$\Rightarrow 0 = \frac{\partial l(\hat{\theta})}{\partial \theta} = \frac{\partial l(\theta_0)}{\partial \theta} + (\hat{\theta} - \theta_0)\frac{\partial^2 l(\theta_0)}{\partial \theta_0^2} + \frac{(\hat{\theta} - \theta_0)^2}{2}\frac{\partial^3 l(\theta^*)}{\partial \theta_0^3}$$

$$\theta^* \in [\theta_0, \hat{\theta}]$$

$$(\hat{\theta}-\theta_0) = \frac{-\dfrac{\partial l(\theta_0)}{\partial \theta}}{\dfrac{\partial^2 l(\theta_0)}{\partial \theta^2} + \dfrac{(\hat{\theta}-\theta_0)}{2}\dfrac{\partial^3 l(\theta^*)}{\partial \theta^3}}$$

Now we have $\hat{\theta}-\theta_0$, Now multiple Both side by $\sqrt{n}$ and study what happen's to each term on R.H.S.

$$\sqrt{n}(\hat{\theta}-\theta_0) = \frac{-\sqrt{n}\left[\dfrac{1}{n}\sum_{i=1}^{n}\dfrac{\partial \log f(x_i|\theta)}{\partial \theta}\Big|_{\theta=\theta_0}\right]}{\left[\dfrac{1}{n}\sum_{i=1}^{n}\dfrac{\partial^2 \log f(x_i|\theta)}{\partial \theta^2}\Big|_{\theta=\theta_0}\right] + \dfrac{\hat{\theta}-\theta_0}{2}\left[\dfrac{1}{n}\sum_{i=1}^{n}\dfrac{\partial^3 \log f(x_i|\theta)}{\partial \theta^3}\Big|_{\theta_0}\right]}$$

Apply CLT for Numerator..

$$\mathbb{E}\left[\frac{\partial \log f(x_i|\theta)}{\partial \theta}\right] = 0$$

$$\mathrm{Var}\left(\frac{\partial \log f(x_i|\theta)}{\partial \theta}\right) = \mathbb{E}\left[\left(\frac{\partial \log f(x_i|\theta)}{\partial \theta}\right)^2\right]$$

$$= I_x(\theta)$$

so by CLT

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{d\log f(x_i|\theta)}{\partial\theta}\Big|_{\theta=\theta_0}\right) \longrightarrow N(0, I_x(\theta_0))$$

By WLLN :

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2\log f(x_i|\theta)}{\partial\theta^2}\Big|_{\theta=\theta_0} \longrightarrow \mathbb{E}\left[\frac{\partial^2 l}{\partial\theta^2}\right]$$

$$= -I_x(\theta)$$

for third term : we know $\hat{\theta} \xrightarrow{P} \theta_0$,
assume that $\mathbb{E}\left[\frac{\partial^3\log f(x|\theta)}{\partial\theta^3}\right] < \infty$

$$\Rightarrow (\hat{\theta} - \theta^*)(\text{fixed quantity}) = 0$$

$$\Rightarrow \quad \sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \frac{-N(0, I_x(\theta))}{-I_x(\theta) + 0}$$

$$\xrightarrow{(d)} N\left(0, \frac{1}{I_x(\theta)}\right)$$

Summary

Fisher information attempts to measure the amount of information about a parameter that a random variable or sample contains. Fisher information from independent random variables adds together to form the Fisher information in the sample. The information inequality (Cramer-Rao lower bound) provides lower bounds on the variances of all estimators. An estimator is efficient if its variance equals the lower bound. The asymptotic distribution of a maximum likelihood estimator of θ is (under regularity conditions) normal with mean θ and variance equal to 1 over the Fisher information in the sample.

Up Next - Hypothesis Testing