Lecture 6: Properties of MLEs

Fisher Information:

Let X be a continuous r.v with p.d.f $f(x|\theta)$. We shall assume that $f(x|\theta)$ is at least second order differentiable w.r.t θ and that the limits of the interval of support of $f(x|\theta)$ does not depend on θ.

Why? So that we can exchange the order of differentiation w.r.t θ and integration w.r.t x of certain functions of $f(x|\theta)$.

Note: $f(x|\theta) = 1/\theta, \; 0 < x \leq \theta$ does not meet these regularity conditions, because the support of $f(x|\theta)$ depends on θ.

The Fisher information is defined as

$$I(\theta) = \int_{-\infty}^{\infty} \left[\frac{d \ln f(x|\theta)}{d\theta}\right]^2 f(x|\theta)\,dx$$

$$= E\left\{\left[\frac{d \ln f(x|\theta)}{d\theta}\right]^2\right\} \quad \text{— (1).}$$

A useful identity concerning $I(\theta)$:

From the equation $\int_{-\infty}^{\infty} f(x|\theta)\,dx = 1$,

$$\int_{-\infty}^{\infty} \frac{d f(x|\theta)}{d\theta}\,dx = \frac{d}{d\theta}(1) = 0$$

i.e. $$\int_{-\infty}^{\infty} \frac{d f(x|\theta)}{d\theta}\,dx = \int_{-\infty}^{\infty} \frac{d f(x|\theta)}{d\theta} \cdot \frac{1}{f(x|\theta)}\, f(x|\theta)\,dx$$

$$= \int_{-\infty}^{\infty} \frac{d \ln f(x|\theta)}{d\theta}\, f(x|\theta)\,dx$$

$$= E\left\{\frac{d \ln f(x|\theta)}{d\theta}\right\} = 0$$

But $$I(\theta) = E\left\{\left[\frac{d \ln f(x|\theta)}{d\theta}\right]^2\right\}$$

$$= \operatorname{var}\left\{\frac{d \ln f(x|\theta)}{d\theta}\right\}$$

Now, $\frac{d}{d\theta}\int_{-\infty}^{\infty}\frac{d\ln f(x|\theta)}{d\theta}f(x|\theta)dx$

$$= \int_{-\infty}^{\infty}\left[\frac{d^2\ln f(x|\theta)}{d\theta^2}f(x|\theta) + \frac{d}{d\theta}\ln f(x|\theta)\frac{df(x|\theta)}{d\theta}\right]dx$$

by the Product Rule

$$= \int_{-\infty}^{\infty}\left[\frac{d^2\ln f(x|\theta)}{d\theta^2} + \frac{d\ln f(x|\theta)}{d\theta}\cdot\frac{df(x|\theta)}{d\theta}\cdot\frac{1}{f(x|\theta)}\right]$$

$$\times f(x|\theta)\,dx$$

$$= \int_{-\infty}^{\infty}\left[\frac{d^2\ln f(x|\theta)}{d\theta^2} + \left\{\frac{d\ln f(x|\theta)}{d\theta}\right\}^2\right]f(x|\theta)dx = 0$$

$$\Rightarrow \int_{-\infty}^{\infty}\left[\frac{d\ln f(x|\theta)}{d\theta}\right]^2 f(x|\theta)dx = -\int_{-\infty}^{\infty}\left[\frac{d^2\ln f(x|\theta)}{d\theta^2}\right]f(x|\theta)dx$$

$$\Rightarrow I(\theta) = -E\left\{\left[\frac{d^2\ln f(x|\theta)}{d\theta^2}\right]\right\} \qquad \text{—— (2)}$$

Example 1: Fisher information for μ: (σ^2 known)

Consider $X \sim N(\mu, \sigma^2)$ i.e

$$f(x|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ \frac{-(x-\mu)^2}{2\sigma^2} \right\}$$

$$\ln f(x|\mu) = -\frac{1}{2}\ln(2\pi\sigma^2) - (x-\mu)^2/2\sigma^2$$

$$\Rightarrow \frac{d \ln f(x|\mu)}{d\mu} = \frac{x-\mu}{\sigma^2}$$

and $\dfrac{d^2 \ln f(x|\mu)}{d\mu^2} = -\dfrac{1}{\sigma^2}$

$$\Rightarrow I(\mu) = -E\left\{ \frac{d^2 \ln f(X|\mu)}{d\mu^2} \right\} = \frac{1}{\sigma^2}$$

This same result can be obtained from

using $I(\mu) = E\left\{ \left[\dfrac{d \ln f(X|\mu)}{d\mu} \right]^2 \right\}$

$$= E\left\{ \frac{(X-\mu)^2}{\sigma^4} \right\} = \frac{var(X)}{\sigma^4}$$

$$= \sigma^2/\sigma^4 = \frac{1}{\sigma^2}$$

Thus, the higher the variance σ^2, the less "information" there is in a single observation

Now, consider an r.s. $X_1, X_2, \ldots, X_n$ from p.d.f $f(x|\theta)$. Then

$$I_n(\theta) = -E\left\{\frac{d^2 \ln f(X_1, \ldots, X_n|\theta)}{d\theta^2}\right\}$$

$$= -E\left\{\frac{d^2}{d\theta^2}\left(\ln f(X_1|\theta) + \ln f(X_2|\theta) + \cdots + \ln f(X_n|\theta)\right)\right\}$$

$$= -E\left\{\frac{d^2}{d\theta}\ln f(X_1|\theta)\right\} - E\left\{\frac{d^2}{d\theta}\ln f(X_2|\theta)\right\} - \cdots - E\left\{\frac{d^2}{d\theta^2}\ln f(X_n|\theta)\right\}$$

$$= I(\theta) + I(\theta) + \cdots + I(\theta) = n\,I(\theta)$$

So, for IID r.v.s, $I_n(\theta) = n\,I(\theta)$

In the previous example, $I_n(\theta) = \frac{n}{\sigma^2}$ for a r.s. of size n from $N(\mu, \sigma^2)$.

<u>Multivariate case:</u> $\underline{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$

$I(\underline{\theta})$ is now the information matrix, whose (i,j)-th element is

$$I_{ij}(\underline{\theta}) = E\left\{\left[\frac{\partial}{\partial\theta_i}\ln f(x|\underline{\theta})\right]\left[\frac{\partial}{\partial\theta_j}\ln f(x|\underline{\theta})\right]\right\}$$

$$= -E\left\{ \frac{\partial^2 \ln f(X|\underline{\theta})}{\partial\theta_i \, \partial\theta_j} \right\} \qquad \text{—} \qquad (3)$$

For IID observations,

$$I_n(\underline{\theta}) = n\, I(\underline{\theta})$$

<u>Example 2:</u> $I(\underline{\theta})$ for $\underline{\theta} = (\mu, \sigma^2)$ in $N(\mu, \sigma^2)$.

Let $X_1, X_2, \ldots, X_n$ be IID $N(\mu, \sigma^2)$. Denote $\theta_1 = \mu$, $\theta_2 = \sigma^2$. Then

$$\ln f(x|\theta_1, \theta_2) = -\frac{1}{2}\ln(2\pi\theta_2) - \frac{(x-\theta)^2}{2\theta_2}$$

It follows that

$$\frac{\partial}{\partial\theta_1} \ln f(x|\theta_1,\theta_2) = \frac{(x-\theta_1)}{\theta_2}$$

$$\frac{\partial}{\partial\theta_2} \ln f(x|\theta_1,\theta_2) = -\frac{1}{2\theta_2} + \frac{(x-\theta_1)^2}{2\theta_2^2}$$

$$\frac{\partial^2 \ln f(x|\theta_1,\theta_2)}{\partial\theta_1\,\partial\theta_2} = -\frac{(x-\theta_1)}{\theta_2^2}$$

$$\frac{\partial^2}{\partial\theta_2^2} \ln f(x|\theta_1,\theta_2) = \frac{1}{2\theta_2^2} - \frac{(x-\theta_1)^2}{\theta_2^3}$$

From eqn (3),

$$I_{11}(\theta_1, \theta_2) = -E\left\{\frac{\partial^2 \ln f(X|\theta_1, \theta_2)}{\partial \theta_1^2}\right\} = \frac{1}{\theta_2}$$

$$I_{12}(\theta_1, \theta_2) = -E\left\{\frac{\partial^2 \ln f(X|\theta_1, \theta_2)}{\partial \theta_1 \partial \theta_2}\right\}$$

$$= E\left\{\frac{(X-\theta_1)}{\theta_2^2}\right\} = 0$$

and

$$I_{22}(\theta_1, \theta_2) = -E\left\{\frac{\partial^2 \ln f(X|\theta_1, \theta_2)}{\partial \theta_2^2}\right\}$$

$$= \frac{-1}{2\theta_2^2} + \frac{E(X-\theta_1)^2}{\theta_2^3}$$

$$= -\frac{1}{2\theta_2^2} + \frac{\theta_2}{\theta_2^3} = \frac{1}{2\theta_2^2}$$

$$\therefore n I(\mu, \sigma^2) = \begin{bmatrix} n/\sigma^2 & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix}$$

Cramer-Rao Lower Bound

Let $\hat{\theta}$ be any estimator of θ with

$E(\hat{\theta}) = \theta + \mathrm{Bias}(\theta)$.

If $\mathrm{Bias}(\theta)$ is differentiable and if certain regularity conditions hold, then

$\mathrm{var}(\hat{\theta}) \geq (1 + \mathrm{Bias}'(\theta))^2 / nI(\theta)$, where

$\mathrm{Bias}'(\theta)$ is the first derivative of $\mathrm{Bias}(\theta)$.

This is the Cramer-Rao inequality.

If $\hat{\theta}$ is unbiased, then

$$\mathrm{var}(\hat{\theta}) \geq \frac{1}{nI(\theta)} \qquad \text{—— (4)}$$

The ratio of the lower bound $\left((1 + \mathrm{Bias}'(\theta))^2 / nI(\theta)\right)$ to the variance of any estimator of θ (biased or unbiased) is called the efficiency of that estimator.

Example 3: Efficiency of S^2.

Consider a r.s. of size n from $N(\mu, \sigma^2)$.

Now, $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$

and so $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}$

$$\Rightarrow \quad S^2 \sim \frac{\sigma^2 \chi^2_{(n-1)}}{n-1}$$

$$E(S^2) = E\left(\frac{\sigma^2 \chi^2_{(n-1)}}{n-1}\right) = \frac{\sigma^2}{n-1} E\left(\chi^2_{(n-1)}\right)$$

$$= \frac{\sigma^2}{n-1} \cdot n-1 = \sigma^2$$

$$var(S^2) = var\left(\frac{\sigma^2 \chi^2_{(n-1)}}{n-1}\right) = \frac{\sigma^4}{(n-1)^2} var\left(\chi^2_{(n-1)}\right)$$

$$= \frac{\sigma^4}{(n-1)^2} \cdot 2(n-1) = \frac{2\sigma^4}{n-1}$$

$\Rightarrow$ S^2 is an unbiased estimator of σ^2.

The Cramer-Rao lower bound for the variance of S^2 is $\frac{1}{nI(\sigma^2)} = \frac{1}{n/2\sigma^4} = \frac{2\sigma^4}{n}$.

$\Rightarrow$ the efficiency of S^2 $= \frac{2\sigma^4}{n} \Big/ \frac{2\sigma^4}{n-1}$

$$= \frac{n-1}{n} = 1 - 1/n.$$

As $n \to \infty$, $1 - 1/n \to 1 \Rightarrow S^2$ is an asymptotically efficient estimator of σ^2.

Now $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$, the MLE of σ^2.

$E(\hat{\sigma}^2) = E\left(\frac{n-1}{n} S^2\right) = \frac{n-1}{n} E(S^2) = \frac{n-1}{n}\sigma^2$.

Thus $\text{Bias}(\hat{\sigma}^2) = -\sigma^2/n$ and $\text{Bias}'(\hat{\sigma}^2) = -1/n$.

$\Rightarrow$ the lower bound for $\hat{\sigma}^2$

$$= \frac{\left(1 + \text{Bias}'(\hat{\sigma}^2)\right)^2}{n I(\sigma^2)} = \frac{\left(1 - 1/n\right)^2}{n/2\sigma^4} = \frac{2(n-1)^2\sigma^4}{n^3}$$

The ratio of this lower bound to

$$\text{Var}(\hat{\sigma}^2) = 2(n-1)\sigma^4/n^2$$

is $(n-1)/n \to 1$ as $n \to \infty$. Thus

$\hat{\sigma}^2$ is also an asymptotically efficient estimator of σ^2.

Asymptotic Normality:

Under certain regularity conditions on $f(x|\theta)$, the MLE of θ based on an r.s. of size n from $f(x|\theta)$ is asymptotically normally-distributed with mean θ and variance $= 1/nI(\theta)$.

So, $E(\hat{\theta}) \to \theta$ and $var(\hat{\theta}) \to \frac{1}{nI(\theta)}$

Indeed, by Chebyshev's inequality

$P\{|\hat{\theta} - \theta| > \varepsilon) \to 0$ as $n \to \infty$ for any $\varepsilon > 0$.

That is, $\hat{\theta}$ is a consistent estimator of θ, as it converges in probability to θ.