Lecture 3

Properties of MLE: consistency, asymptotic normality. Fisher information.

In this section we will try to understand why MLEs are 'good'.

Let us recall two facts from probability that we be used often throughout this course.

- **Law of Large Numbers (LLN):**

  If the distribution of the i.i.d. sample $X_1, \dots, X_n$ is such that X_1 has finite expectation, i.e. $|\mathbb{E}X_1| < \infty$, then the sample average

  $$\bar{X}_n = \frac{X_1 + \ldots + X_n}{n} \to \mathbb{E}X_1$$

  converges to its expectation *in probability* , which means that for any arbitrarily small $\varepsilon > 0$,

  $$\mathbb{P}(|\bar{X}_n - \mathbb{E}X_1| > \epsilon) \to 0 \text{ as } n \to \infty.$$

  Note. Whenever we will use the LLN below we will simply say that the average converges to its expectation and will not mention in what sense. More mathematically inclined students are welcome to carry out these steps more rigorously, especially when we use LLN in combination with the Central Limit Theorem.

- **Central Limit Theorem (CLT):**

  If the distribution of the i.i.d. sample $X_1, \dots, X_n$ is such that X_1 has finite expectation and variance, i.e. $|\mathbb{E}X_1| < \infty$ and $\sigma^2 = \mathrm{Var}(X) < \infty$, then

  $$\sqrt{n}(\bar{X}_n - \mathbb{E}X_1) \to^d N(0, \sigma^2)$$

  $$\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \rightsquigarrow Z$$ ~ unit gaussian

  converges in distribution to normal distribution with zero mean and variance σ^2, which means that for any interval $[a, b]$,

  $$\mathbb{P}\Big(\sqrt{n}(\bar{X}_n - \mathbb{E}X_1) \in [a, b]\Big) \to \int_a^b \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} dx.$$

In other words, the random variable $\sqrt{n}(\bar{X}_n - \mathbb{E}X_1)$ will behave like a random variable from normal distribution when n gets large.

Exercise. Illustrate CLT by generating 100 Bernoulli random varibles $B(p)$ (or one Binomial r.v. $B(100, p)$) and then computing $\sqrt{n}(\bar{X}_n - \mathbb{E}X_1)$. Repeat this many times and use 'dfittool' to see that this random quantity will be well approximated by normal distribution.

We will prove that MLE satisfies (usually) the following two properties called *consistency* and *asymptotic normality*.

Converge in Prob.

1. **Consistency.** We say that an estimate $\hat{\theta}$ is consistent if $\hat{\theta} \to \theta_0$ in probability as $n \to \infty$, where θ_0 is the 'true' unknown parameter of the distribution of the sample.

2. **Asymptotic Normality.** We say that $\hat{\theta}$ is asymptotically normal if

$$\sqrt{n}(\hat{\theta} - \theta_0) \to^d N(0, \sigma^2_{\theta_0})$$

where $\sigma^2_{\theta_0}$ is called the asymptotic variance of the estimate $\hat{\theta}$. Asymptotic normality says that the estimator not only converges to the unknown parameter, but it converges fast enough, at a rate $1/\sqrt{n}$.

Consistency of MLE.

To make our discussion as simple as possible, let us assume that a likelihood function is smooth and behaves in a nice way like shown in figure 3.1, i.e. its maximum is achieved at a unique point $\hat{\theta}$.

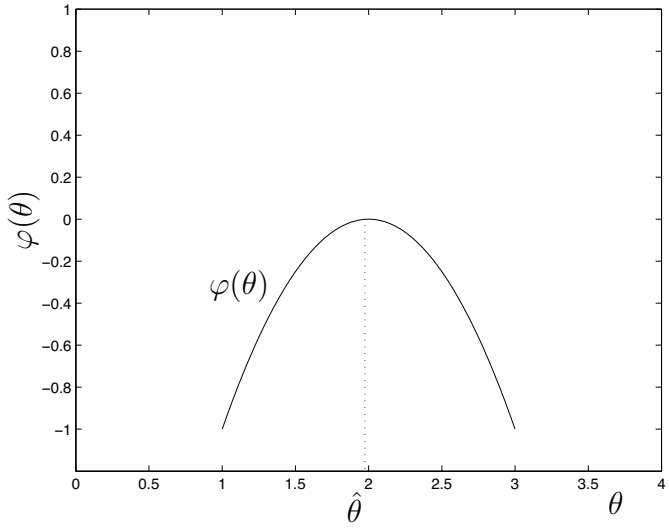

Figure 3.1: Maximum Likelihood Estimator (MLE)

Suppose that the data $X_1, \ldots, X_n$ is generated from a distribution with unknown parameter θ_0 and $\hat{\theta}$ is a MLE. Why $\hat{\theta}$ converges to the unknown parameter θ_0? This is not immediately obvious and in this section we will give a sketch of why this happens.

First of all, MLE $\hat{\theta}$ is the maximizer of

$$L_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log f(X_i|\theta)$$

which is a log-likelihood function normalized by $\frac{1}{n}$ (of course, this does not affect maximization). Notice that function $L_n(\theta)$ depends on data. Let us consider a function $l(X|\theta) = \log f(X|\theta)$ and define

$$L(\theta) = \mathbb{E}_{\theta_0} l(X|\theta),$$

where $\mathbb{E}_{\theta_0}$ denotes the expectation with respect to the true uknown parameter θ_0 of the sample $X_1, \ldots, X_n$. If we deal with continuous distributions then

$$L(\theta) = \int (\log f(x|\theta)) f(x|\theta_0) dx.$$

By law of large numbers, for any θ,

$$L_n(\theta) \to \mathbb{E}_{\theta_0} l(X|\theta) = L(\theta).$$

Note that $L(\theta)$ does not depend on the sample, it only depends on θ. We will need the following

Lemma. *We have that for any* θ,

$$L(\theta) \le L(\theta_0).$$

Moreover, the inequality is strict, $L(\theta) < L(\theta_0)$, *unless*

$$\mathbb{P}_{\theta_0}(f(X|\theta) = f(X|\theta_0)) = 1.$$

which means that $\mathbb{P}_\theta = \mathbb{P}_{\theta_0}$.

Proof. Let us consider the difference

$$L(\theta) - L(\theta_0) = \mathbb{E}_{\theta_0}(\log f(X|\theta) - \log f(X|\theta_0)) = \mathbb{E}_{\theta_0} \log \frac{f(X|\theta)}{f(X|\theta_0)}.$$

Since $\log t \le t - 1$, we can write

$$\begin{aligned} \mathbb{E}_{\theta_0} \log \frac{f(X|\theta)}{f(X|\theta_0)} &\le \mathbb{E}_{\theta_0}\Big(\frac{f(X|\theta)}{f(X|\theta_0)} - 1\Big) = \int \Big(\frac{f(x|\theta)}{f(x|\theta_0)} - 1\Big) f(x|\theta_0) dx \\ &= \int f(x|\theta) dx - \int f(x|\theta_0) dx = 1 - 1 = 0. \end{aligned}$$

Both integrals are equal to 1 because we are integrating the probability density functions. This proves that $L(\theta) - L(\theta_0) \le 0$. The second statement of Lemma is also clear.

□

We will use this Lemma to sketch the consistency of the MLE.

Theorem: *Under some regularity conditions on the family of distributions, MLE $\hat{\theta}$ is consistent, i.e. $\hat{\theta} \to \theta_0$ as $n \to \infty$.*

The statement of this Theorem is not very precise but but rather than proving a rigorous mathematical statement our goal here is to illustrate the main idea. Mathematically inclined students are welcome to come up with some precise statement.

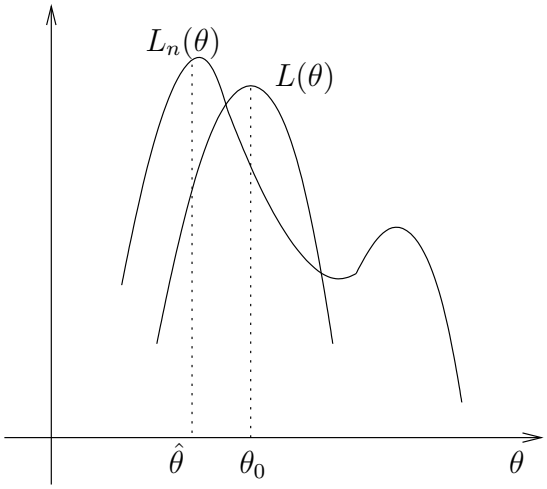

Figure 3.2: Illustration to Theorem.

Proof. We have the following facts:

1. $\hat{\theta}$ is the maximizer of $L_n(\theta)$ (by definition).

2. θ_0 is the maximizer of $L(\theta)$ (by Lemma).

3. $\forall \theta$ we have $L_n(\theta) \to L(\theta)$ by LLN.

This situation is illustrated in figure 3.2. Therefore, since two functions L_n and L are getting closer, the points of maximum should also get closer which exactly means that $\hat{\theta} \to \theta_0$.

□

Asymptotic normality of MLE. Fisher information.

We want to show the asymptotic normality of MLE, i.e. to show that

$$\sqrt{n}(\hat{\theta} - \theta_0) \to^d N(0, \sigma^2_{MLE}) \text{ for some } \sigma^2_{MLE}$$

and compute σ^2_{MLE}. This asymptotic variance in some sense measures the quality of MLE. First, we need to introduce the notion called Fisher Information.

Let us recall that above we defined the function $l(X|\theta) = \log f(X|\theta)$. To simplify the notations we will denote by $l'(X|\theta), l''(X|\theta)$, etc. the derivatives of $l(X|\theta)$ **with respect to** θ.

Definition. *(Fisher information.)* Fisher information of a random variable X with distribution $\mathbb{P}_{\theta_0}$ from the family $\{\mathbb{P}_\theta : \theta \in \Theta\}$ is defined by

$$I(\theta_0) = \mathbb{E}_{\theta_0}(l'(X|\theta_0))^2 \equiv \mathbb{E}_{\theta_0}\Big(\frac{\partial}{\partial\theta}\log f(X|\theta)\Big|_{\theta=\theta_0}\Big)^2.$$

Remark. Let us give a very informal interpretation of Fisher information. The derivative

$$l'(X|\theta_0) = (\log f(X|\theta_0))' = \frac{f'(X|\theta_0)}{f(X|\theta_0)}$$

can be interpreted as a measure of how quickly the distribution density or p.f. will change when we slightly change the parameter θ near θ_0. When we square this and take expectation, i.e. average over X, we get an averaged version of this measure. So if Fisher information is large, this means that the distribution will change quickly when we move the parameter, so the distribution with parameter θ_0 is 'quite different' and 'can be well distinguished' from the distributions with parameters not so close to θ_0. This means that we should be able to estimate θ_0 well based on the data. On the other hand, if Fisher information is small, this means that the distribution is 'very similar' to distributions with parameter not so close to θ_0 and, thus, more difficult to distinguish, so our estimation will be worse. We will see precisely this behavior in Theorem below.

Next lemma gives another often convenient way to compute Fisher information.

Lemma. *We have,*

$$\mathbb{E}_{\theta_0} l''(X|\theta_0) \equiv \mathbb{E}_{\theta_0} \frac{\partial^2}{\partial\theta^2} \log f(X|\theta_0) = -I(\theta_0).$$

second derivative의 expectation = $-I(\theta_0)$

Proof. First of all, we have

$$l'(X|\theta) = (\log f(X|\theta))' = \frac{f'(X|\theta)}{f(X|\theta)}$$

and

$$(\log f(X|\theta))'' = \frac{f''(X|\theta)}{f(X|\theta)} - \frac{(f'(X|\theta))^2}{f^2(X|\theta)}.$$

Also, since p.d.f. integrates to 1,

$$\int f(x|\theta)dx = 1,$$

if we take derivatives of this equation with respect to θ (and interchange derivative and integral, which can usually be done) we will get,

$$\int \frac{\partial}{\partial\theta} f(x|\theta)dx = 0 \text{ and } \int \frac{\partial^2}{\partial\theta^2} f(x|\theta)dx = \int f''(x|\theta)dx = 0.$$

To finish the proof we write the following computation

θ_0는 true distribution의 parameter

$$\begin{aligned}
\mathbb{E}_{\theta_0} l''(X|\theta_0) &= \mathbb{E}_{\theta_0} \frac{\partial^2}{\partial\theta^2} \log f(X|\theta_0) = \int (\log f(x|\theta_0))'' f(x|\theta_0)dx \\
&= \int \Big(\frac{f''(x|\theta_0)}{f(x|\theta_0)} - \Big(\frac{f'(x|\theta_0)}{f(x|\theta_0)} \Big)^2 \Big) f(x|\theta_0)dx \\
&= \int f''(x|\theta_0)dx - \mathbb{E}_{\theta_0}(l'(X|\theta_0))^2 = 0 - I(\theta_0 = -I(\theta_0).
\end{aligned}$$

ok. 여기까지 이해함.

We are now ready to prove the main result of this section.

Theorem. *(Asymptotic normality of MLE.) We have,*

$$\sqrt{n}(\hat{\theta} - \theta_0) \to N\Big(0, \frac{1}{I(\theta_0)}\Big).$$

θ_0는 true distribution의 parameter

As we can see, the asymptotic variance/dispersion of the estimate around true parameter will be smaller when Fisher information is larger. 역의관계에서까.

Proof. Since MLE $\hat{\theta}$ is maximizer of $L_n(\theta) = \frac{1}{n}\sum_{i=1}^n \log f(X_i|\theta)$, we have

$$L_n'(\hat{\theta}) = 0.$$

maximum에서가 미분값 0

Let us use the Mean Value Theorem

$$\frac{f(a) - f(b)}{a - b} = f'(c) \text{ or } f(a) = f(b) + f'(c)(a - b) \text{ for } c \in [a, b]$$

delta method ?

with $f(\theta) = L_n'(\theta), a = \hat{\theta}$ and $b = \theta_0$. Then we can write,

$\hat{\theta} - \theta_0 = \frac{-L_n'(\theta_0)}{L_n''(\hat{\theta}_1)}$

$$0 = L_n'(\hat{\theta}) = L_n'(\theta_0) + L_n''(\hat{\theta}_1)(\hat{\theta} - \theta_0)$$

for some $\hat{\theta}_1 \in [\hat{\theta}, \theta_0]$. From here we get that

양변에 $\sqrt{n}$을 곱해줌

$$\hat{\theta} - \theta_0 = -\frac{L_n'(\theta_0)}{L_n''(\hat{\theta}_1)} \text{ and } \sqrt{n}(\hat{\theta} - \theta_0) = -\frac{\sqrt{n}L_n'(\theta_0)}{L_n''(\hat{\theta}_1)}. \quad (3.0.1)$$

Since by Lemma in the previous section we know that θ_0 is the maximizer of $L(\theta)$, we have

$$L'(\theta_0) = \mathbb{E}_{\theta_0} l'(X|\theta_0) = 0. \quad (3.0.2)$$

Therefore, the numerator in (3.0.1)

numerator / denominator

$$\begin{aligned} \sqrt{n}L_n'(\theta_0) &= \sqrt{n}\Big(\frac{1}{n}\sum_{i=1}^n l'(X_i|\theta_0) - 0\Big) \quad (3.0.3) \\ &= \sqrt{n}\Big(\frac{1}{n}\sum_{i=1}^n l'(X_i|\theta_0) - \mathbb{E}_{\theta_0} l'(X_1|\theta_0)\Big) \to N\Big(0, \mathrm{Var}_{\theta_0}(l'(X_1|\theta_0))\Big) \end{aligned}$$

converges in distribution by Central Limit Theorem.

Next, let us consider the denominator in (3.0.1). First of all, we have that for all θ,

$$L_n''(\theta) = \frac{1}{n}\sum l''(X_i|\theta) \to \mathbb{E}_{\theta_0} l''(X_1|\theta) \text{ by LLN.} \quad (3.0.4)$$

Also, since $\hat{\theta}_1 \in [\hat{\theta}, \theta_0]$ and by consistency result of previous section, $\hat{\theta} \to \theta_0$, we have $\hat{\theta}_1 \to \theta_0$. Using this together with (10.0.3) we get

$$L_n''(\hat{\theta}_1) \to \mathbb{E}_{\theta_0} l''(X_1|\theta_0) = -I(\theta_0) \text{ by Lemma above.}$$

Combining this with (3.0.3) we get

$$-\frac{\sqrt{n}L_n'(\theta_0)}{L_n''(\hat{\theta}_1)} \to^d N\Big(0, \frac{\mathrm{Var}_{\theta_0}(l'(X_1|\theta_0))}{(I(\theta_0))^2}\Big).$$

Finally, the variance,

$$\mathrm{Var}_{\theta_0}(l'(X_1|\theta_0)) = \mathbb{E}_{\theta_0}(l'(X|\theta_0))^2 - (\mathbb{E}_{\theta_0}l'(x|\theta_0))^2 = I(\theta_0) - 0 \Rightarrow N\Big(0, \frac{1}{I(\theta_0)}\Big)$$

where in the last equality we used the definition of Fisher information and (3.0.2).

□

Let us compute Fisher information for some particular distributions.

Example 1. The family of Bernoulli distributions $B(p)$ has p.f.

$$f(x|p) = p^x(1-p)^{1-x}$$

and taking the logarithm

$$\log f(x|p) = x\log p + (1-x)\log(1-p).$$

The second derivative with respect to parameter p is

$$\frac{\partial}{\partial p}\log f(x|p) = \frac{x}{p} - \frac{1-x}{1-p}, \quad \frac{\partial^2}{\partial p^2}\log f(x|p) = -\frac{x}{p^2} - \frac{1-x}{(1-p)^2}.$$

Then the Fisher information can be computed as

$$I(p) = -\mathbb{E}\frac{\partial^2}{\partial p^2}\log f(X|p) = \frac{\mathbb{E}X}{p^2} + \frac{1-\mathbb{E}X}{(1-p)^2} = \frac{p}{p^2} + \frac{1-p}{(1-p)^2} = \frac{1}{p(1-p)}.$$

The MLE of p is $\hat{p} = \bar{X}$ and the asymptotic normality result states that

$$\sqrt{n}(\hat{p} - p_0) \to N(0, p_0(1-p_0))$$

which, of course, also follows directly from the CLT.

Example. The family of exponential distributions $E(\alpha)$ has p.d.f.

$$f(x|\alpha) = \begin{cases} \alpha e^{-\alpha x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

and, therefore,

$$\log f(x|\alpha) = \log\alpha - \alpha x \Rightarrow \frac{\partial^2}{\partial\alpha^2}\log f(x|\alpha) = -\frac{1}{\alpha^2}.$$

This does not depend on X and we get

$$I(\alpha) = -\mathbb{E}\frac{\partial^2}{\partial\alpha^2}\log f(X|\alpha) = \frac{1}{\alpha^2}.$$

Therefore, the MLE $\hat{\alpha} = 1/\bar{X}$ is asymptotically normal and

$$\sqrt{n}(\hat{\alpha} - \alpha_0) \to N(0, \alpha_0^2).$$

□