Maximum likelihood estimation

- Suppose that $(X_1, \ldots, X_n)$ is a sample with joint PMF (or joint PDF) f_θ, where the parameter vector θ is in a known space Θ. For $\eta \in \Theta$, define

$$L(\eta) = f_\eta(X_1, \ldots, X_n), \tag{1}$$

then L is called the likelihood function (概似函數) based on the sample $(X_1, \ldots, X_n)$. Suppose that L attains it maximum on Θ at $\hat{\theta}$, then $\hat{\theta}$ is called the maximum likelihood estimator (MLE; 最大概似估計量) of θ (assuming the maximizer is unique). That is,

$$\hat{\theta} = \arg\max_{\eta \in \Theta} L(\eta).$$

  - Suppose that we observe $(X_1, \ldots, X_n) = (x_1, \ldots, x_n)$ and define L according to (1) with $(X_1, \ldots, X_n)$ replaced by $(x_1, \ldots, x_n)$, then the function L is called the observed likelihood function, and the maximizer of L is called the (observed) MLE of θ.

- Example 1. Suppose that $(X_1, \ldots, X_n)$ is a random sample and $X_1 \sim Bin(1, p)$ and $p \in [0, 1]$. Find the MLE of p based on the sample $(X_1, \ldots, X_n)$.

Sol. Let p_X be the PMF of X_1, then for $x \in \{0, 1\}$,

$$\begin{aligned} p_X(x) &= \begin{cases} p & \text{if } x = 1; \\ 1 - p & \text{if } x = 0. \end{cases} \\ &= p^x (1-p)^{1-x}. \end{aligned}$$

Let f_p be the joint PMF of $(X_1, \ldots, X_n)$, then for $x_1, \ldots, x_n \in \{0, 1\}$,

$$f_p(x_1, \ldots, x_n) = \prod_{i=1}^n p_X(x_i) = p^{\sum_{i=1}^n x_i} (1-p)^{n - \sum_{i=1}^n x_i}.$$

Let L be the likelihood function based on $(X_1, \ldots, X_n)$, then

$$\begin{aligned} L(q) &= f_q(X_1, \ldots, X_n) \\ &= q^{\sum_{i=1}^n X_i} (1-q)^{n - \sum_{i=1}^n X_i} \\ &= q^{n\bar{X}} (1-q)^{n - n\bar{X}} \end{aligned}$$

for $q \in [0, 1]$, where $\bar{X}$ is the sample mean $\sum_{i=1}^n X_i/n$. Computing the derivative of $\log L$ and we have for $q \in (0, 1)$,

$$\begin{aligned} \frac{d}{dq} \log(L(q)) &= \frac{n(\bar{X} - q)}{q(1-q)} \\ & \begin{cases} > 0 & \text{if } 0 < q < \bar{X}; \\ = 0 & \text{if } q = \bar{X}; \\ < 0 & \text{if } \bar{X} < q < 1. \end{cases} \end{aligned}$$

Thus $\log L$ attains its maximum on $(0, 1)$ at $\bar{X}$ and so is L. It is clear that L does not attain its maximum at 0 or 1, so L attains its maximum on $[0, 1]$ at $\bar{X}$. The MLE of p is $\bar{X}$.

- Example 2. Suppose that $(X_1, \ldots, X_{10})$ is a random sample and $X_1 \sim Bin(1, p)$ and $p \in \{0.1, 0.5, 0.8\}$. Suppose that we observe that $(X_1, \ldots, X_6) = (0, \ldots, 0)$ and $(X_7, \ldots, X_{10}) = (1, \ldots, 1)$. Find the observed MLE of p. The output after running the following R commands

```
q <- c(0.1, 0.5, 0.8)
q^4*(1-q)^6
```

is

```
[1] 0.0000531441 0.0009765625 0.0000262144
```

Sol. The observed likelihood function L is given by

$$L(q) = q^4(1-q)^6$$

for $q \in \{0.1, 0.5, 0.8\}$. From the R output, $L(0.5) > L(0.1) > L(0.8)$, so the observed MLE of p is 0.5.

- Example 3. Suppose that $(X_1, \ldots, X_n)$ is a random sample and $X_1 \sim N(0, \theta)$, where $\theta > 0$. Find the MLE of θ based on $(X_1, \ldots, X_n)$.

Sol. A PDF of $(X_1, \ldots, X_n)$ is f_θ, where

$$f_\theta(x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\theta}} e^{-x_i^2/(2\theta)}$$

for $(x_1, \ldots, x_n) \in R^n$. Let L be the likelihood function based on $(X_1, \ldots, X_n)$, then

$$\begin{aligned} L(\eta) &= f_\eta(X_1, \ldots, X_n) \\ &= \left(\frac{1}{\sqrt{2\pi\eta}}\right)^n e^{-\sum_{i=1}^n X_i^2/(2\eta)} \end{aligned}$$

for $\eta > 0$. Computing the derivative of $\log L$ and we have for $\eta > 0$,

$$\begin{aligned} \frac{d}{d\eta} \log(L(\eta)) &= \frac{n}{2\eta^2}\left(\frac{\sum_{i=1}^n X_i^2}{n} - \eta\right) \\ & \begin{cases} > 0 & \text{if } \eta < \sum_{i=1}^n X_i^2/n; \\ = 0 & \text{if } \eta = \sum_{i=1}^n X_i^2/n; \\ < 0 & \text{if } \sum_{i=1}^n X_i^2/n < \eta. \end{cases} \end{aligned}$$

Thus $\log L$ attains its maximum on $(0, \infty)$ at $\sum_{i=1}^n X_i^2/n$ and $\sum_{i=1}^n X_i^2/n$ is the MLE of θ.

- To see why MLE is a reasonable estimator, note that for $L(\eta)$ defined in (1), $\log L(\eta)/n$ converges to $E \log f_\eta(X_1)$ by SLLN. It can be shown by Jensen's inequality that

$$E \log f_\eta(X_1) \leq E \log f_\theta(X_1) \tag{2}$$

for $\eta \in \Theta$ when X_1 has PDF f_θ. Since

$$\theta = \arg\max_{\eta \in \Theta} \; E \log f_\eta(X_1)$$

and $E \log f_\eta(X_1) \approx \log L(\eta)/n$, it makes sense to estimate θ using

$$\hat{\theta} = \arg\max_{\eta \in \Theta} \; \log L(\eta).$$

- Convex function. A function φ defined on an open interval I is convex means that for any two points a, $b \in I$ such that $a < b$,

  $$\varphi(ta + (1-t)b) \leq t\varphi(a) + (1-t)\varphi(b)$$

  for $t \in [0, 1]$.

  - If for any two points a, $b \in I$ such that $a < b$,

    $$\varphi(ta + (1-t)b) < t\varphi(a) + (1-t)\varphi(b)$$

    for $t \in (0, 1)$, φ is strictly convex on I.
  - If $\varphi'' > 0$ on I, then φ is strictly convex on I.

- Jensen's inequality. Suppose that φ is a convex function on an open interval I and X is a random variable that takes values in I and $E(X)$ is finite. Then

  $$E\varphi(X) \geq \varphi(E(X)).$$

  Moreover, if φ is stricly convex on I, then $E\varphi(X) = \varphi(E(X))$ only if $P(X = E(X)) = 1$.

- Apply Jensen's inequality with $\varphi = -\log$ and we have

  $$-E \log\left(\frac{f_\eta(X_1)}{f_\theta(X_1)}\right) \geq -\log E\left(\frac{f_\eta(X_1)}{f_\theta(X_1)}\right) = 0,$$

  so $E \log f_\eta(X_1) \leq E \log f_\theta(X_1)$ and (2) holds.