

Stat 710: Mathematical Statistics

Lecture 10

Jun Shao

Department of Statistics
University of Wisconsin
Madison, WI 53706, USA

Lecture 10: Likelihood and maximum likelihood estimator (MLE)

The *maximum likelihood method* is the most popular method for deriving estimators in statistical inference that does not use any loss function.

Example 4.28

Let X be a single observation taking values from $\{0, 1, 2\}$ according to P_θ , where $\theta = \theta_0$ or θ_1 and the values of $P_{\theta_j}(\{i\})$ are given by the following table:

	$x = 0$	$x = 1$	$x = 2$
$\theta = \theta_0$	0.8	0.1	0.1
$\theta = \theta_1$	0.2	0.3	0.5

If $X = 0$ is observed, it is more plausible that it came from P_{θ_0} , since $P_{\theta_0}(\{0\})$ is much larger than $P_{\theta_1}(\{0\})$.

We then estimate θ by θ_0 .

Lecture 10: Likelihood and maximum likelihood estimator (MLE)

The *maximum likelihood method* is the most popular method for deriving estimators in statistical inference that does not use any loss function.

Example 4.28

Let X be a single observation taking values from $\{0, 1, 2\}$ according to P_θ , where $\theta = \theta_0$ or θ_1 and the values of $P_{\theta_j}(\{i\})$ are given by the following table:

	$x = 0$	$x = 1$	$x = 2$
$\theta = \theta_0$	0.8	0.1	0.1
$\theta = \theta_1$	0.2	0.3	0.5

If $X = 0$ is observed, it is more plausible that it came from P_{θ_0} , since $P_{\theta_0}(\{0\})$ is much larger than $P_{\theta_1}(\{0\})$.

We then estimate θ by θ_0 .

Example 4.28 (continued)

On the other hand, if $X = 1$ or 2 , it is more plausible that it came from P_{θ_1} , although in this case the difference between the probabilities is not as large as that in the case of $X = 0$.

This suggests the following estimator of θ :

$$T(X) = \begin{cases} \theta_0 & X = 0 \\ \theta_1 & X \neq 0. \end{cases}$$

The idea in Example 4.28 can be easily extended to the case where P_θ is a discrete distribution and $\theta \in \Theta \subset \mathcal{R}^k$.

If $X = x$ is observed, θ_1 is more plausible than θ_2 if and only if $P_{\theta_1}(\{x\}) > P_{\theta_2}(\{x\})$.

We then estimate θ by a $\hat{\theta}$ that maximizes $P_\theta(\{x\})$ over $\theta \in \Theta$, if such a $\hat{\theta}$ exists.

Under the Bayesian approach with a prior that is the discrete uniform distribution on $\{\theta_1, \dots, \theta_m\}$, $P_\theta(\{x\})$ is proportional to the posterior probability and we can say that θ_1 is more probable than θ_2 if $P_{\theta_1}(\{x\}) > P_{\theta_2}(\{x\})$.

Example 4.28 (continued)

On the other hand, if $X = 1$ or 2 , it is more plausible that it came from P_{θ_1} , although in this case the difference between the probabilities is not as large as that in the case of $X = 0$.

This suggests the following estimator of θ :

$$T(X) = \begin{cases} \theta_0 & X = 0 \\ \theta_1 & X \neq 0. \end{cases}$$

The idea in Example 4.28 can be easily extended to the case where P_{θ} is a discrete distribution and $\theta \in \Theta \subset \mathcal{R}^k$.

If $X = x$ is observed, θ_1 is more plausible than θ_2 if and only if $P_{\theta_1}(\{x\}) > P_{\theta_2}(\{x\})$.

We then estimate θ by a $\hat{\theta}$ that maximizes $P_{\theta}(\{x\})$ over $\theta \in \Theta$, if such a $\hat{\theta}$ exists.

Under the Bayesian approach with a prior that is the discrete uniform distribution on $\{\theta_1, \dots, \theta_m\}$, $P_{\theta}(\{x\})$ is proportional to the posterior probability and we can say that θ_1 is more probable than θ_2 if $P_{\theta_1}(\{x\}) > P_{\theta_2}(\{x\})$.

Note that $P_\theta(\{x\})$ is the p.d.f. w.r.t. the counting measure. Hence, it is natural to extend the idea to the case of continuous (or arbitrary) X by using the p.d.f. of X w.r.t. some σ -finite measure on the range \mathcal{X} of X .

Definition 4.3

Let $X \in \mathcal{X}$ be a sample with a p.d.f. f_θ w.r.t. a σ -finite measure ν , where $\theta \in \Theta \subset \mathcal{R}^k$.

(i) For each $x \in \mathcal{X}$, $f_\theta(x)$ considered as a function of θ is called the *likelihood function* and denoted by $\ell(\theta)$.

(ii) Let $\bar{\Theta}$ be the closure of Θ . A $\hat{\theta} \in \bar{\Theta}$ satisfying $\ell(\hat{\theta}) = \max_{\theta \in \bar{\Theta}} \ell(\theta)$ is called a *maximum likelihood estimate* (MLE) of θ . If $\hat{\theta}$ is a Borel function of X a.e. ν , then $\hat{\theta}$ is called a *maximum likelihood estimator* (MLE) of θ .

(iii) Let g be a Borel function from Θ to \mathcal{R}^p , $p \leq k$. If $\hat{\theta}$ is an MLE of θ , then $\hat{\vartheta} = g(\hat{\theta})$ is defined to be an MLE of $\vartheta = g(\theta)$.

Note that $P_\theta(\{x\})$ is the p.d.f. w.r.t. the counting measure. Hence, it is natural to extend the idea to the case of continuous (or arbitrary) X by using the p.d.f. of X w.r.t. some σ -finite measure on the range \mathcal{X} of X .

Definition 4.3

Let $X \in \mathcal{X}$ be a sample with a p.d.f. f_θ w.r.t. a σ -finite measure ν , where $\theta \in \Theta \subset \mathcal{R}^k$.

(i) For each $x \in \mathcal{X}$, $f_\theta(x)$ considered as a function of θ is called the *likelihood function* and denoted by $\ell(\theta)$.

(ii) Let $\bar{\Theta}$ be the closure of Θ . A $\hat{\theta} \in \bar{\Theta}$ satisfying $\ell(\hat{\theta}) = \max_{\theta \in \bar{\Theta}} \ell(\theta)$ is called a *maximum likelihood estimate* (MLE) of θ . If $\hat{\theta}$ is a Borel function of X a.e. ν , then $\hat{\theta}$ is called a *maximum likelihood estimator* (MLE) of θ .

(iii) Let g be a Borel function from Θ to \mathcal{R}^p , $p \leq k$. If $\hat{\theta}$ is an MLE of θ , then $\hat{\vartheta} = g(\hat{\theta})$ is defined to be an MLE of $\vartheta = g(\theta)$.

Remarks

- Note that $\bar{\Theta}$ instead of Θ is used in the definition of an MLE. This is because a maximum of $\ell(\theta)$ may not exist when Θ is an open set.
In some textbooks, Θ is used, instead of $\bar{\Theta}$
- Part (iii) of Definition 4.3 is motivated by a fact given in Exercise 95 of §4.6.
- An MLE may not exist, or there are many MLE's.
- An MLE may not have an explicit form.
- In terms of their mse's, MLE's are not necessarily better than UMVUE's or Bayes estimators.
- MLE's are frequently inadmissible.
This is not surprising, since MLE's are not derived under any given loss function.
- The main theoretical justification for MLE's is provided in the theory of asymptotic efficiency considered in §4.5.

Computation of MLE

If Θ contains finitely many points, then $\bar{\Theta} = \Theta$ and an MLE exists and can always be obtained by comparing finitely many values $\ell(\theta)$, $\theta \in \Theta$. Since $\log x$ is a strictly increasing function, $\hat{\theta}$ is an MLE if and only if it maximizes the log-likelihood function $\log \ell(\theta)$.

It is often more convenient to work with $\log \ell(\theta)$.

If $\ell(\theta)$ is differentiable on Θ° , the interior of Θ , then possible candidates for MLE's are the values of $\theta \in \Theta^\circ$ satisfying

$$\frac{\partial \log \ell(\theta)}{\partial \theta} = 0,$$

which is called the *likelihood equation* or *log-likelihood equation*.

A root of the likelihood equation may be local or global minima, local or global maxima, or simply stationary points.

Also, extrema may occur at the boundary of Θ or when $\|\theta\| \rightarrow \infty$.

Furthermore, if $\ell(\theta)$ is not always differentiable, then extrema may occur at nondifferentiable or discontinuity points of $\ell(\theta)$.

Hence, it is important to analyze the entire likelihood function to find its maxima.

Example 4.29

Let X_1, \dots, X_n be i.i.d. binary random variables with

$$P(X_1 = 1) = p \in \Theta = (0, 1).$$

When $(X_1, \dots, X_n) = (x_1, \dots, x_n)$ is observed, the likelihood function is

$$\ell(p) = \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i} = p^{n\bar{x}} (1-p)^{n(1-\bar{x})},$$

where $\bar{x} = n^{-1} \sum_{i=1}^n x_i$.

Note that $\bar{\Theta} = [0, 1]$ and $\Theta^\circ = \Theta$.

The likelihood equation is

$$\frac{n\bar{x}}{p} - \frac{n(1-\bar{x})}{1-p} = 0.$$

If $0 < \bar{x} < 1$, then this equation has a unique solution \bar{x} .

The second-order derivative of $\log \ell(p)$ is

$$-\frac{n\bar{x}}{p^2} - \frac{n(1-\bar{x})}{(1-p)^2},$$

which is always negative.

Example 4.29 (continued)

Also, when p tends to 0 or 1 (the boundary of Θ), $\ell(p) \rightarrow 0$.

Thus, \bar{x} is the unique MLE of p .

When $\bar{x} = 0$, $\ell(p) = (1 - p)^n$ is a strictly decreasing function of p and, therefore, its unique maximum is 0.

Similarly, the MLE is 1 when $\bar{x} = 1$.

Combining these results, we conclude that the MLE of p is \bar{x} .

When $\bar{x} = 0$ or 1, a maximum of $\ell(p)$ does not exist on $\Theta = (0, 1)$, although $\sup_{p \in (0, 1)} \ell(p) = 1$; the MLE takes a value outside of Θ and, hence, is not a reasonable estimator.

However, if $p \in (0, 1)$, the probability that $\bar{x} = 0$ or 1 tends to 0 quickly as $n \rightarrow \infty$.

Discussion

Example 4.29 indicates that, for small n , a maximum of $\ell(\theta)$ may not exist on Θ and an MLE may be an unreasonable estimator; however, this is unlikely to occur when n is large.

A rigorous result of this sort is given in §4.5.2, where we study asymptotic properties of MLE's.

Example 4.29 (continued)

Also, when p tends to 0 or 1 (the boundary of Θ), $\ell(p) \rightarrow 0$.

Thus, \bar{x} is the unique MLE of p .

When $\bar{x} = 0$, $\ell(p) = (1 - p)^n$ is a strictly decreasing function of p and, therefore, its unique maximum is 0.

Similarly, the MLE is 1 when $\bar{x} = 1$.

Combining these results, we conclude that the MLE of p is \bar{x} .

When $\bar{x} = 0$ or 1, a maximum of $\ell(p)$ does not exist on $\Theta = (0, 1)$, although $\sup_{p \in (0, 1)} \ell(p) = 1$; the MLE takes a value outside of Θ and, hence, is not a reasonable estimator.

However, if $p \in (0, 1)$, the probability that $\bar{x} = 0$ or 1 tends to 0 quickly as $n \rightarrow \infty$.

Discussion

Example 4.29 indicates that, for small n , a maximum of $\ell(\theta)$ may not exist on Θ and an MLE may be an unreasonable estimator; however, this is unlikely to occur when n is large.

A rigorous result of this sort is given in §4.5.2, where we study asymptotic properties of MLE's.

Example 4.30

Let X_1, \dots, X_n be i.i.d. from $N(\mu, \sigma^2)$ with unknown $\theta = (\mu, \sigma^2)$, $n \geq 2$. Consider first the case where $\Theta = \mathcal{R} \times (0, \infty)$.

$$\log \ell(\theta) = -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 - \frac{n}{2} \log \sigma^2 - \frac{n}{2} \log(2\pi).$$

The likelihood equation is

$$\frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu) = 0 \quad \text{and} \quad \frac{1}{\sigma^4} \sum_{i=1}^n (x_i - \mu)^2 - \frac{n}{\sigma^2} = 0.$$

Solving the first equation for μ , we obtain a unique solution $\bar{x} = n^{-1} \sum_{i=1}^n x_i$, and substituting \bar{x} for μ in the second equation, we obtain a unique solution $\hat{\sigma}^2 = n^{-1} \sum_{i=1}^n (x_i - \bar{x})^2$.

To show that $\hat{\theta} = (\bar{x}, \hat{\sigma}^2)$ is an MLE, first note that Θ is an open set and $\ell(\theta)$ is differentiable everywhere; as θ tends to the boundary of Θ or $\|\theta\| \rightarrow \infty$, $\ell(\theta)$ tends to 0; and

$$\frac{\partial^2 \log \ell(\theta)}{\partial \theta \partial \theta^\tau} = - \begin{pmatrix} \frac{n}{\sigma^2} & \frac{1}{\sigma^4} \sum_{i=1}^n (x_i - \mu) \\ \frac{1}{\sigma^4} \sum_{i=1}^n (x_i - \mu) & \frac{1}{\sigma^6} \sum_{i=1}^n (x_i - \mu)^2 - \frac{n}{2\sigma^4} \end{pmatrix}$$

Example 4.30 (continued)

This matrix is negative definite when $\mu = \bar{x}$ and $\sigma^2 = \hat{\sigma}^2$.

Hence $\hat{\theta}$ is the unique MLE.

Sometimes we can avoid the calculation of the second-order derivatives.

For instance, in this example we know that $\ell(\theta)$ is bounded and $\ell(\theta) \rightarrow 0$ as $\|\theta\| \rightarrow \infty$ or θ tends to the boundary of Θ ; hence the unique solution to the likelihood equation must be the MLE.

Another way to show that $\hat{\theta}$ is the MLE is indicated by the following discussion.

Consider next the case where $\Theta = (0, \infty) \times (0, \infty)$, i.e., μ is known to be positive.

The likelihood function is differentiable on $\Theta^\circ = \Theta$ and $\bar{\Theta} = [0, \infty) \times [0, \infty)$.

If $\bar{x} > 0$, then the same argument for the previous case can be used to show that $(\bar{x}, \hat{\sigma}^2)$ is the MLE.

If $\bar{x} \leq 0$, then the first equation in the likelihood equation does not have a solution in Θ .

Example 4.30 (continued)

However, the function $\log \ell(\theta) = \log \ell(\mu, \sigma^2)$ is strictly decreasing in μ for any fixed σ^2 .

Hence, a maximum of $\log \ell(\mu, \sigma^2)$ is $\mu = 0$, which does not depend on σ^2 .

Then, the MLE is $(0, \tilde{\sigma}^2)$, where $\tilde{\sigma}^2$ is the value maximizing $\log \ell(0, \sigma^2)$ over $\sigma^2 \geq 0$.

Maximizing $\log \ell(0, \sigma^2)$ leads to $\tilde{\sigma}^2 = n^{-1} \sum_{i=1}^n x_i^2$.

Thus, the MLE is

$$\hat{\theta} = \begin{cases} (\bar{x}, \hat{\sigma}^2) & \bar{x} > 0 \\ (0, \tilde{\sigma}^2) & \bar{x} \leq 0. \end{cases}$$

Again, the MLE in this case is not in Θ if $\bar{x} \leq 0$.

One can show that a maximum of $\ell(\theta)$ does not exist on Θ when $\bar{x} \leq 0$.

Example 4.31

Let X_1, \dots, X_n be i.i.d. from the uniform distribution on an interval \mathcal{I}_θ with an unknown θ .

First, consider the case where $\mathcal{I}_\theta = (0, \theta)$ and $\theta > 0$, $\Theta^\circ = (0, \infty)$.

The likelihood function is $\ell(\theta) = \theta^{-n} I_{(x_{(n)}, \infty)}(\theta)$, $x_{(n)} = \max(x_1, \dots, x_n)$.

On $(0, x_{(n)})$, $\ell \equiv 0$ and on $(x_{(n)}, \infty)$, $\ell'(\theta) = -n\theta^{n-1} < 0$ for all θ .

$\ell(\theta)$ is not differentiable at $x_{(n)}$ and the method of using the likelihood equation is not applicable.

Since $\ell(\theta)$ is strictly decreasing on $(x_{(n)}, \infty)$ and is 0 on $(0, x_{(n)})$, a unique maximum of $\ell(\theta)$ is $x_{(n)}$, which is a discontinuity point of $\ell(\theta)$.

This shows that the MLE of θ is the largest order statistic $X_{(n)}$.

Next, consider the case where $\mathcal{I}_\theta = (\theta - \frac{1}{2}, \theta + \frac{1}{2})$ with $\theta \in \mathcal{R}$.

The likelihood function is $\ell(\theta) = I_{(x_{(n)} - \frac{1}{2}, x_{(1)} + \frac{1}{2})}(\theta)$, $x_{(1)} = \min(x_1, \dots, x_n)$

Again, the method of using the likelihood equation is not applicable.

However, it follows from Definition 4.3 that any statistic $T(X)$ satisfying

$x_{(n)} - \frac{1}{2} \leq T(\mathbf{x}) \leq x_{(1)} + \frac{1}{2}$ is an MLE of θ .

This example indicates that MLE's may not be unique and can be unreasonable.