Chapter 7. Point Estimation
Lecture 1: Maximum likelihood and moment methods

Point estimation

We consider a sample $X$ (a random sample for most discussions) from a population indexed by unknown $\theta \in \Theta$.
At least in the following two situations we want to estimate $\theta$:

- We want to make inference about the population (or $\theta$), and it is reasonable to start with the estimation of $\theta$.
- A function $g(\theta)$ has a meaningful physical interpretation (such as a population mean or median) so there is a direct interest to estimate it.

Definition 7.1.1.

Let $X$ be a sample from a population indexed by unknown $\theta \in \Theta$. A point estimator (or estimator) of a function of $\theta$, $g(\theta)$, is any statistic $T(X)$. We call $T(x)$ an estimate after we observe $X = x$.

This is a vague definition: it is not even required that $T(X) \in \Theta$. 
Maximum likelihood estimators

The maximum likelihood estimation is the most popular technique.

Example.

Let $X$ be a single observation taking values either 0 or 1, with a pmf $f_{\theta}$, where $\theta = \theta_0$ or $\theta_1$ and the values of $f_{\theta_j}(i)$ are given as follows:

<table>
<thead>
<tr>
<th></th>
<th>$x = 0$</th>
<th>$x = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_{\theta_0}$</td>
<td>0.9</td>
<td>0.1</td>
</tr>
<tr>
<td>$f_{\theta_1}$</td>
<td>0.4</td>
<td>0.6</td>
</tr>
</tbody>
</table>

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

If $X = 1$, it is more plausible that it came from $f_{\theta_1}$, although the difference between the probabilities is not as large as that for $X = 0$.

This suggests the following estimator of $\theta$:

$$T(X) = \begin{cases} 
\theta_0 & X = 0 \\
\theta_1 & X \neq 0
\end{cases}$$
The idea in the example can be easily extended to the case where $f_\theta$ is any pmf and $\theta \in \Theta \subset \mathbb{R}^k$.

If $X = x$ is observed, $\theta_1$ is more plausible than $\theta_2$ iff $f_{\theta_1}(x) > f_{\theta_2}(x)$.

We then estimate $\theta$ by a $\hat{\theta}$ maximizing $f_\theta(x)$ over $\theta \in \Theta$.

For continuous variables, it is natural to extend this idea with pmf replaced by pdf.

Recall that, given $X = x$, the likelihood function $L(\theta | x)$ is the joint pmf for the discrete case or the joint pdf for the continuous case.

**Definition 7.2.4.**

For each sample point $x$, the maximum likelihood estimate (MLE) of $\theta$ is any $\hat{\theta}(x) \in \Theta$ (if it exists) such that

$$L(\hat{\theta}(x) | x) = \sup_{\theta \in \Theta} L(\theta | x)$$

A maximum likelihood estimator (MLE) of $\theta$ is defined to be $\hat{\theta}(X)$, when $\hat{\theta}(x)$ is a well-defined function.

- Note that the MLE is defined for vector $\theta$ when $k > 1$. If $g$ is a function on $\Theta$, then an MLE of $g(\theta)$ is defined as $g(\hat{\theta})$ (see Theorem 7.2.10).
In some cases, an MLE exists if we add boundary points of $\Theta$ into $\Theta$ to form a new parameter space.

If the parameter space $\Theta$ contains finitely many points, then an MLE can always be obtained by comparing finitely many values $L(\theta|x)$, $\theta \in \Theta$.

If $L(\theta|x)$ is differentiable on $\Theta$, then possible candidates for MLEs are the values of $\theta \in \Theta$ satisfying

$$\frac{\partial L(\theta|x)}{\partial \theta} = 0 \quad \text{or} \quad \frac{\partial \log L(\theta|x)}{\partial \theta} = 0$$

which is called the likelihood equation or log likelihood equation.

It is important to analyze the entire likelihood function to find its maxima.

- Note that $\theta$ satisfying likelihood equation may be local or global minima, local or global maxima, or simply stationary points.
- Extrema may occur at the boundary of $\Theta$ or when $\theta$ diverges to $\infty$ in some way.
- If $L(\theta|x)$ is not always differentiable, then extrema may occur at non-differentiable or discontinuity points of $L(\theta|x)$.
Example 7.2.7.

Let \( X_1, \ldots, X_n \) be iid Bernoulli random variables with 
\[ P(X_1 = 1) = p \in \Theta = (0, 1). \]

When \( X = (X_1, \ldots, X_n) = (x_1, \ldots, x_n) \) is observed, the likelihood function is 
\[ L(p|x) = \prod_{i=1}^{n} p^{x_i} (1 - p)^{1-x_i} = p^{n\bar{x}} (1 - p)^{n(1 - \bar{x})}, \]
where \( \bar{x} \) is the sample mean.

The likelihood equation is 
\[ \frac{d \log L(p|x)}{dp} = \frac{n\bar{x}}{p} - \frac{n(1 - \bar{x})}{1 - p} = 0. \]

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

The second-order derivative of \( \log L(p|x) \) is 
\[ \frac{d^2 \log L(p|x)}{dp^2} = -\frac{n\bar{x}}{p^2} - \frac{n(1 - \bar{x})}{(1 - p)^2}, \]
which is always negative.

Also, when \( p \) tends to 0 or 1 (the boundary of \( \Theta \)), \( L(p|x) \rightarrow 0 \).
Thus, \( \bar{x} \) is the unique MLE of \( p \).
When $\bar{x} = 0$, $L(p|x) = (1 - p)^n$ is a strictly decreasing function of $p$ and, therefore, an MLE does not exist unless we add 0 to the parameter space. If $p \in (0, 1)$, an MLE $= 0$ is not reasonable; however, the probability that $\bar{x} = 0$ is $p^n$, which tends to 0 quickly as $n \to \infty$. A similar discussion can be made when $\bar{x} = 1$.

This example indicates that, for small $n$, an MLE may not exist on $\Theta$ or an MLE may be an unreasonable estimator; however, this is unlikely to occur when $n$ is large.

**Example 7.2.11.**

Let $X_1, \ldots, X_n$ be iid from $N(\mu, \sigma^2)$ with unknown $\theta = (\mu, \sigma^2)$. Consider first the case where $\Theta = \mathbb{R} \times (0, \infty)$. When $x = (x_1, \ldots, x_n)$ is observed, the log-likelihood function is

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

The likelihood equation becomes
\[
\frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) = 0 \quad \text{and} \quad \frac{1}{\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2 - \frac{n}{\sigma^2} = 0.
\]

Solving the first equation for \( \mu \), we obtain a unique solution \( \bar{x} = n^{-1} \sum_{i=1}^{n} x_i \), and substituting \( \bar{x} \) for \( \mu \) 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 \( \Theta \) is an open set and \( L(\theta|x) \) is differentiable everywhere; as \( \theta \) tends to the boundary of \( \Theta \) or diverges to \( \infty \), \( L(\theta|x) \) tends to 0; and

\[
\frac{\partial^2 \log L(\theta|x)}{\partial \theta \partial \theta^T} = -\begin{pmatrix}
\frac{n}{\sigma^2} & \frac{1}{\sigma^4} \sum_{i=1}^{n} (x_i - \mu) \\
\frac{1}{\sigma^4} \sum_{i=1}^{n} (x_i - \mu) & \frac{1}{\sigma^6} \sum_{i=1}^{n} (x_i - \mu)^2 - \frac{n}{2\sigma^4}
\end{pmatrix}
\]

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; e.g., in this example we know that \( L(\theta|x) \) is bounded and \( L(\theta|x) \rightarrow 0 \) as \( \theta \) diverges to \( \infty \) or \( \theta \) tends to the boundary of \( \Theta \); hence the unique solution to the likelihood equation must be the MLE.
Consider next the case where $\Theta = [0, \infty) \times (0, \infty)$, i.e., we know $\mu \geq 0$. The likelihood function is differentiable on $\Theta$.

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 likelihood equation does not have a solution in $\Theta$. However, the function $\log L(\theta | x) = \log L(\mu, \sigma^2 | x)$ is strictly decreasing in $\mu$ for any fixed $\sigma^2$.

Hence, a maximum of $\log L(\mu, \sigma^2 | x)$ is $\mu = 0$ not depending on $\sigma^2$.

Then, the MLE of $\sigma^2$ is the value maximizing $\log L(0, \sigma^2 | x)$ over $\sigma^2 \geq 0$. Solving
\[
\frac{d}{d\sigma^2} \log L(0, \sigma^2 | x) = -\frac{1}{\sigma^4} \sum_{i=1}^{n} x_i^2 - \frac{n}{2\sigma^2} = 0
\]
gives $\hat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} x_i^2$, i.e., the MLE of $\theta$ is
\[
\hat{\theta} = \begin{cases} 
(\bar{x}, \hat{\sigma}^2) & \bar{x} > 0 \\
(0, \hat{\sigma}^2) & \bar{x} \leq 0.
\end{cases}
\]

Example.

Let $X_1, \ldots, X_n$ be iid from $\text{uniform}(\theta, \theta + 1)$ with $\theta \in \mathbb{R}$. 
The likelihood function is

\[
L(\theta|x) = \begin{cases} 
1 & x(n) - 1 < \theta < x(1) \\
0 & \text{otherwise}
\end{cases}
\]

The method of using the likelihood equation is not applicable. However, it follows from the definition that any statistic \( T(X) \) satisfying

\[
x(n) - 1 \leq T(x) \leq x(1)
\]

is an MLE of \( \theta \).

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

Example

Let \( X \) be an observation from the hypergeometric distribution

\[
P(X = x) = \binom{n}{x} \frac{\theta-n}{\theta} \binom{r-x}{\theta}, \quad x = 0, 1, ..., r < \min(n, \theta - n)
\]

with known \( r, n \), and an unknown \( \theta = n + 1, n + 2, ... \)

In this case, the likelihood function is defined on integers and the method of using the likelihood equation is certainly not applicable.
Note that
\[
\frac{L(\theta|x)}{L(\theta-1|x)} = \frac{(\theta-r)(\theta-n)}{\theta(\theta-n-r+x)},
\]
which is larger than 1 iff \(\theta < rn/x\) and is smaller than 1 iff \(\theta > rn/x\). Thus, \(L(\theta|x)\) has a maximum \(\theta = \) the integer part of \(rn/x\), which is the MLE of \(\theta\).

Example.

Let \(X_1,...,X_n\) be iid from \(\text{gamma}(\alpha, \gamma)\) with unknown \(\alpha > 0\) and \(\gamma > 0\). The log-likelihood function is
\[
\log L(\theta|x) = -n\alpha \log \gamma - n\log \Gamma(\alpha) + (\alpha - 1) \sum_{i=1}^{n} \log x_i - \frac{1}{\gamma} \sum_{i=1}^{n} x_i
\]
and the likelihood equation is
\[
-n\log \gamma - \frac{n\Gamma'(\alpha)}{\Gamma(\alpha)} + \sum_{i=1}^{n} \log x_i = 0 \quad \text{and} \quad -\frac{n\alpha}{\gamma} + \frac{1}{\gamma^2} \sum_{i=1}^{n} x_i = 0.
\]
The second equation yields \(\gamma = \bar{x}/\alpha\).
Substituting \(\gamma = \bar{x}/\alpha\) into the first equation we obtain that
\[
\log \alpha - \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} + \frac{1}{n} \sum_{i=1}^{n} \log x_i - \log \bar{x} = 0.
\]

In this case, the likelihood equation does not have an explicit solution. A numerical method has to be applied to compute the MLE for any given observations \( x_1, \ldots, x_n \).

**Exponential families**

Suppose that \( X \) has a pmf or pdf from an exponential family so that

\[
L(\eta | x) = \exp(\eta' T(x) - \zeta(\eta)) h(x),
\]

where \( \eta \in \Xi \) is a vector of unknown parameters.

The likelihood equation is then

\[
\frac{\partial \log L(\eta | x)}{\partial \eta} = T(x) - \frac{\partial \zeta(\eta)}{\partial \eta} = 0,
\]

which has a unique solution \( T(x) = \partial \zeta(\eta) / \partial \eta \), assuming that \( T(x) \) is in the range of \( \partial \zeta(\eta) / \partial \eta \).

Consider the second order derivative matrix:
\[
\frac{\partial^2 \log L(\eta|x)}{\partial \eta \partial \eta'} = - \frac{\partial^2 \zeta(\eta)}{\partial \eta \partial \eta'} = -\text{Var}(T)
\]

Since \(\text{Var}(T)\) is positive definite, \(-\log L(\eta|x)\) is convex in \(\eta\) and \(T(x)\) is the unique MLE of the parameter \(\mu(\eta) = \partial \zeta(\eta)/\partial \eta\).

The function \(\mu(\eta)\) is one-to-one so that \(\mu^{-1}\) exists.

Then the MLE of \(\eta\) is \(\hat{\eta} = \mu^{-1}(T(x))\).

If the likelihood function is

\[
L(\theta|x) = \exp (\eta(\theta)' T(x) - \xi(\theta)) h(x),
\]

then the MLE of \(\theta\) is \(\hat{\theta} = \eta^{-1}(\hat{\eta})\), if \(\eta^{-1}\) exists and \(\hat{\eta}\) is in the range of \(\eta(\theta)\).

Of course, \(\hat{\theta}\) is also the solution of the likelihood equation

\[
\frac{\partial \log L(\theta|x)}{\partial \theta} = \frac{\partial \eta(\theta)}{\partial \theta} T(x) - \frac{\partial \xi(\theta)}{\partial \theta} = 0.
\]
Method of moments

This is the oldest of the three main methods of finding point estimators. Let $X_1, \ldots, X_n$ be a random sample from a pdf or pmf $f_\theta(x)$ with finite $k$th moments, where $\theta = (\theta_1, \ldots, \theta_k) \in \mathbb{R}^k$ is unknown.

Define

$$m_j = \frac{1}{n} \sum_{i=1}^{n} X_i^j \quad \mu_j = E(X_1^j) = h_j(\theta_1, \ldots, \theta_k), \quad j = 1, \ldots, k$$

where each $h_j$ is a known function of $\theta = (\theta_1, \ldots, \theta_k)$.

The method of moment defines an estimator $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_k)$ as a solution to the following system of $k$ equations with $k$ variables:

$$m_j = h_j(\theta_1, \ldots, \theta_k), \quad j = 1, \ldots, k$$

- Let $m = (m_1, \ldots, m_k) = h(\hat{\theta})$, $h = (h_1, \ldots, h_k)$. If the inverse function $h^{-1}$ exists, then the unique moment estimator of $\theta$ is $\hat{\theta} = h^{-1}(m)$.

- When $h^{-1}$ does not exist (i.e., $h$ is not one-to-one), any solution to $m = h(\hat{\theta})$ is a moment estimator of $\theta$; if possible, we always choose a solution so that $\hat{\theta} \in \Theta$. 
Two important statistical principles, the moment matching and substitution principle, are applied in this method.

Moment estimators may not be unique; in some cases, a moment estimator does not exist, or no solution to \( m = h(\hat{\theta}) \) is in \( \Theta \).

When the \( k \) equations involving the first \( k \) moments do not provide a solution, we may consider more equations with moments higher than \( k \).

Moment estimators may not be efficient, but they are simple and can be building blocks for more efficient estimators.

**Example 7.2.1.**

Let \( X_1, \ldots, X_n \) be iid from \( N(\mu, \sigma^2) \), \( \theta = (\mu, \sigma^2) \), \( \mu \in \mathbb{R} \) and \( \sigma^2 \in (0, \infty) \).

Since

\[
E(X_1) = \mu \quad \text{and} \quad E(X_1^2) = \text{Var}(X_1) + [E(X_1)]^2 = \sigma^2 + \mu^2
\]

setting \( m_1 = \mu \) and \( m_2 = \sigma^2 + \mu^2 \) we obtain the moment estimator

\[
\hat{\theta} = \left( \bar{X}, \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2 \right) = \left( \bar{X}, \frac{n-1}{n} S^2 \right).
\]
Example.

Let $X_1, \ldots, X_n$ be iid from $\text{uniform}(\theta_1, \theta_2)$, $-\infty < \theta_1 < \theta_2 < \infty$. Note that

$$ E(X_1) = \frac{\theta_1 + \theta_2}{2} \quad \text{and} \quad E(X_1^2) = \frac{\theta_1^2 + \theta_2^2 + \theta_1 \theta_2}{3} $$

Setting $m_1 = EX_1$ and $m_2 = EX_1^2$ and substituting $\theta_1$ in the second equation by $2m_1 - \theta_2$ (the first equation), we obtain that

$$(2m_1 - \theta_2)^2 + \theta_2^2 + (2m_1 - \theta_2)\theta_2 = 3m_2,$$

which is the same as

$$(\theta_2 - m_1)^2 = 3(m_2 - m_1^2).$$

Since $\theta_2 > EX_1$, we obtain that

$$ \hat{\theta}_2 = m_1 + \sqrt{3(m_2 - m_1^2)} = \bar{X} + \sqrt{\frac{3(n-1)}{n}}S^2 $$

$$ \hat{\theta}_1 = m_1 - \sqrt{3(m_2 - m_1^2)} = \bar{X} - \sqrt{\frac{3(n-1)}{n}}S^2 $$

These estimators are not functions of the sufficient and complete statistic $(X_{(1)}, X_{(n)})$ when $n > 2$. 
Generalized Method of Moments (GMM)

The method of moments tries to solve

\[ m_j - h_j(\theta) = 0, \quad j = 1, \ldots, k. \]

Sometimes, it has no solution.

Moreover, sometimes \( E(X_1^m) \) exists for an \( m > k \) and we may wonder which of the \( k \) moments out of \( m \) moments we should use.

If we consider all \( m \) equations

\[ m_j - h_j(\theta) = 0, \quad j = 1, \ldots, m, \]

then there is typically no solution, since there are more equations than variables.

The generalized method of moments (GMM) can be applied: Let

\[
G(\theta | X) = \left( \frac{1}{n} \sum_{i=1}^{n} \psi_1(X_i, \theta), \ldots, \frac{1}{n} \sum_{i=1}^{n} \psi_m(X_i, \theta) \right)', \quad \theta \in \Theta
\]

where \( \psi_j(x, \theta) = x^j - h_j(\theta) \).
The method of moments tries to solve $G(\theta|x) = 0$, which may not exist. Instead of getting a $\hat{\theta}$ such that $G(\hat{\theta}|X) = 0$, the GMM tries to find a $\hat{\theta}$ that minimizes $G(\theta|X)'G(\theta|X)$ over $\theta \in \Theta$.

Of course, when there is a $\hat{\theta}$ such that $G(\hat{\theta}|X) = 0$, then $\hat{\theta}$ that minimizes $G(\theta|X)'G(\theta|X)$ over $\theta \in \Theta$, since $G(\hat{\theta}|X)'G(\hat{\theta}|X) = 0$.

In general, a GMM estimator of $\theta$ is obtained using the following two-step algorithm.

1. Obtain $\hat{\theta}^{(1)}$ by minimizing $G'(\theta|X)G(\theta|X)$ over $\theta \in \Theta$.

2. Let $\hat{W}$ be the inverse matrix of the $m \times m$ matrix whose $(j, l)$th element is equal to

$$
\frac{1}{n} \sum_{i=1}^{n} \psi_j(X_i, \hat{\theta}^{(1)}) \psi_l(X_i, \hat{\theta}^{(1)})
$$

The GMM estimator $\hat{\theta}$ is obtained by minimizing

$$
G'(\theta|x)\hat{W}G(\theta|x) \quad \text{over } \theta \in \Theta
$$

Note that the solution of a GMM is always in $\Theta$. 