Mathematical Statistics

MAS 713

Chapter 6

Previous lecture

- Point estimators
  - Estimation and sampling distribution
  - Point estimation
  - Properties of estimators
- Sufficient Statistics
  - Factorization Theorem

Any questions?

This lecture

1. 6.1 Maximum Likelihood Estimation
   - 6.1.1 Introduction
   - 6.1.3 Maximum Likelihood Principle
2. 6.2 Cramér-Rao Lower Bound
   - 6.2.1 Introduction
   - 6.2.2 Examples
3. 6.3 Method of Moments
4. 6.4 Examples: MLE and Methods of Moments

Additional reading : Chapter 7

Intuition of MLE

A patient visits a physican and complains about the following sympthoms:

"I have a headache, I'm feeling weak and have no appetite."

The Doctor's diagnostics options:

1. You have a brain tumor.
2. You broke your foot.
3. You have a cold.

The Doctor's job is to determine the most likely illness.

We'll revisit this example later.

Intuition of MLE

A patient visits a physican and complains about the following sympthoms:

"I have a headache, I'm feeling weak and have no appetite."

The Doctor's diagnostics options:

1. You have a brain tumor.
2. You broke your foot.
3. You have a cold.

The Doctor's job is to determine the most likely illness.

We'll revisit this example later.

Intuition of MLE

A patient visits a physican and complains about the following sympthoms:

“I have a headache, I’m feeling weak and have no appetite."

The Doctor’s diagnostics options:

1. You have a brain tumor.
2. You broke your foot.
3. You have a cold.

The Doctor’s job is to determine the most likely illness.

We’ll revisit this example later.

Intuition of MLE

A patient visits a physican and complains about the following sympthoms:

“I have a headache, I’m feeling weak and have no appetite."

The Doctor’s diagnostics options:

1. You have a brain tumor.
2. You broke your foot.
3. You have a cold.

The Doctor’s job is to determine the most likely illness.

We’ll revisit this example later.

Intuition of MLE

A patient visits a physican and complains about the following sympthoms:

“I have a headache, I’m feeling weak and have no appetite."

The Doctor’s diagnostics options:

1. You have a brain tumor.
2. You broke your foot.
3. You have a cold.

The Doctor’s job is to determine the most likely illness.

We’ll revisit this example later.

Intuition of MLE

A patient visits a physican and complains about the following sympthoms:

“I have a headache, I’m feeling weak and have no appetite."

The Doctor’s diagnostics options:

1. You have a brain tumor.
2. You broke your foot.
3. You have a cold.

The Doctor’s job is to determine the most likely illness.

We’ll revisit this example later.

Intuition of MLE

A patient visits a physican and complains about the following sympthoms:

“I have a headache, I’m feeling weak and have no appetite."

The Doctor’s diagnostics options:

1. You have a brain tumor.
2. You broke your foot.
3. You have a cold.

The Doctor’s job is to determine the most likely illness.

We’ll revisit this example later.

Intuition of MLE

A patient visits a physican and complains about the following sympthoms:

“I have a headache, I’m feeling weak and have no appetite."

The Doctor’s diagnostics options:

1. You have a brain tumor.
2. You broke your foot.
3. You have a cold.

The Doctor’s job is to determine the most likely illness.

We’ll revisit this example later.

- We have seen that there are plenty of choices for an estimator $\widehat{\theta}$ of an unknown parameter θ

$\Longrightarrow$ **How to choose $\widehat{\theta}$?**

One possible approach:

Given observations $x_1, x_2, \ldots, x_n$, choose unknown parameter $\widehat{\theta} = \widehat{\theta}(x_1, \ldots, x_n)$ in such a way that it maximizes the probability of the occurrence of our observed values $x_1, x_2, \ldots, x_n$.

$\Longrightarrow$ choose $\widehat{\theta}$ such that

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n \,|\, \widehat{\theta}) = \max_{\theta} \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n \,|\, \theta)$$

- It is the intuition behind the **Maximum Likelihood estimator (MLE)**.

The Maximum Likelihood Principle

The main ingredients:

1. X: a random variable.
2. θ: parameter to estimate (restricted to a parameter space S_θ).
3. $p(X;\theta)$ (or $p(X|\theta)$): a statistical model (pmf or pdf)
4. $X_1, \ldots, X_n$: a random sample from X.

We want to construct good estimators for θ

Notation: Given observation $x_1, \ldots, x_n$, we write

$$p(\mathbf{x}|\theta) = \begin{cases} \text{joint probability mass function} & \text{if } X \text{ is discrete} \\ \text{joint probability density function} & \text{if } X \text{ is continuous} \end{cases}$$

The Maximum Likelihood Principle

Definition

Let $\mathbf{X} = (X_1, \ldots, X_n)$ have joint pdf/pmf $p(\mathbf{x}; \theta)$ where $\theta \in S_\theta$. The likelihood function (or simply likelihood) is defined by

$$S_\theta \ni \theta \mapsto L(\theta) := L(\theta; \mathbf{x}) = p(\mathbf{x}; \theta)$$

Note: $\mathbf{x}$ is fixed and θ varies in S_θ.

- The likelihood is a function of θ.
- The likelihood is not a pdf/pmf (as function of θ, for fixed $\mathbf{x}$).
- If the data is i.i.d then

$$L(\theta; \mathbf{x}) = \prod_{i=1}^{n} p(x_i; \theta)$$

The Maximum Likelihood Principle

- Choose $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\mathbf{x})$ which maximizes the likelihood function, i.e.

$$L\left(\widehat{\boldsymbol{\theta}}; \mathbf{x}\right) = \max_{\boldsymbol{\theta} \in S_{\boldsymbol{\theta}}} L\left(\boldsymbol{\theta}; \mathbf{x}\right)$$

- by definition of the $\arg\max$, this means

$$\widehat{\boldsymbol{\theta}}(\mathbf{x}) \in \arg\max_{\boldsymbol{\theta} \in S_{\boldsymbol{\theta}}} L\left(\boldsymbol{\theta}; \mathbf{x}\right)$$

Definition of Maximum Likelihood Estimator (MLE)

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random sample. If

$$\widehat{\boldsymbol{\theta}}(\mathbf{X}) \in \arg\max_{\boldsymbol{\theta} \in S_{\boldsymbol{\theta}}} L\left(\boldsymbol{\theta}; \mathbf{X}\right)$$

Then we call $\widehat{\boldsymbol{\theta}}(\mathbf{X})$ a Maximum Likelihood Estimator (MLE) for θ.
Note: MLE may not be unique or may not exist.

Remark: $\arg\max_\theta f(\theta)$ is the set of points, θ, for which $f(\theta)$ attains the function's largest value.

Intuition of MLE

The data:

$x =$"I have a headache, I'm feeling weak and have no appetite."

The (discrete) parameter space θ:

1. You have a brain tumor.
2. You broke your foot.
3. You have a cold.

The likelihood under each parameter:

$$\mathbb{P}(\text{"headache, weakness,no appetite"}|\theta = \text{brain tumor}) = 0.2$$
$$\mathbb{P}(\text{"headache, weakness,no appetite"}|\theta = \text{broken foot}) = 0.05$$
$$\mathbb{P}(\text{"headache, weakness,no appetite"}|\theta = \text{cold}) = 0.4$$

.

The ML estimate

The likelihood of having a cold is the highest.

$\widehat{\theta} =$ **cold**

Intuition of MLE

The data:

$x =$"I have a headache, I'm feeling weak and have no appetite."

The (discrete) parameter space θ:

1. You have a brain tumor.
2. You broke your foot.
3. You have a cold.

The likelihood under each parameter:

$$\mathbb{P}(\text{"headache, weakness,no appetite"}|\theta = \text{brain tumor}) = 0.2$$
$$\mathbb{P}(\text{"headache, weakness,no appetite"}|\theta = \text{broken foot}) = 0.05$$
$$\mathbb{P}(\text{"headache, weakness,no appetite"}|\theta = \text{cold}) = 0.4$$

.

The ML estimate

The likelihood of having a cold is the highest.

$\widehat{\theta} = \textbf{cold}$

Intuition of MLE

The data:

$x =$"I have a headache, I'm feeling weak and have no appetite."

The (discrete) parameter space θ:

1. You have a brain tumor.
2. You broke your foot.
3. You have a cold.

The likelihood under each parameter:

$$\mathbb{P}(\text{"headache, weakness,no appetite"}|\theta = \text{brain tumor}) = 0.2$$
$$\mathbb{P}(\text{"headache, weakness,no appetite"}|\theta = \text{broken foot}) = 0.05$$
$$\mathbb{P}(\text{"headache, weakness,no appetite"}|\theta = \text{cold}) = 0.4$$

.

The ML estimate

The likelihood of having a cold is the highest.

$\widehat{\theta} =$ **cold**

Intuition of MLE

The data:

$x =$"I have a headache, I'm feeling weak and have no appetite."

The (discrete) parameter space θ:

1. You have a brain tumor.
2. You broke your foot.
3. You have a cold.

The likelihood under each parameter:

$$\mathbb{P}(\text{"headache, weakness,no appetite"}|\theta = \text{brain tumor}) = 0.2$$
$$\mathbb{P}(\text{"headache, weakness,no appetite"}|\theta = \text{broken foot}) = 0.05$$
$$\mathbb{P}(\text{"headache, weakness,no appetite"}|\theta = \text{cold}) = 0.4$$

.

The ML estimate

The likelihood of having a cold is the highest.

$\widehat{\theta} =$ **cold**

The Maximum Likelihood Principle

We may apply any monotone increasing function, and still achieve maximization. Very often it is more convenient to consider the logarithm of the likelihood function (log-likelihood function)

$$\log L(\boldsymbol{\theta}, \mathbf{x}) = \log p(\mathbf{x}|\boldsymbol{\theta})$$

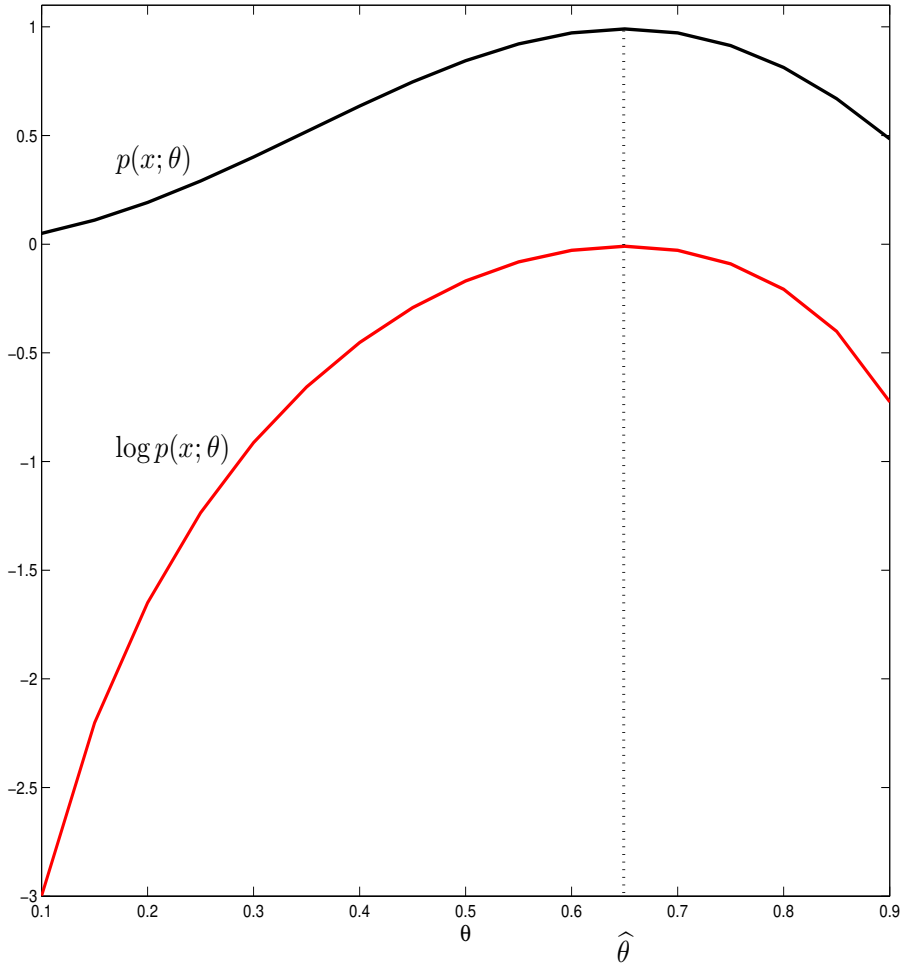
Since the logarithm is a monotonic function, the maximization of the likelihood and *log*-likelihood functions is equivalent, that is, $\widehat{\boldsymbol{\theta}}$ maximizes the likelihood function if and only if it also maximizes the *log*-likelihood function.

$$\arg\max_{\boldsymbol{\theta}\in S_{\boldsymbol{\theta}}} L(\boldsymbol{\theta}; \mathbf{X}) = \arg\max_{\boldsymbol{\theta}\in S_{\boldsymbol{\theta}}} \log L(\boldsymbol{\theta}; \mathbf{X})$$

or in other words

$$\widehat{\boldsymbol{\theta}} \in \arg\max_{\boldsymbol{\theta}\in S_{\boldsymbol{\theta}}} L(\boldsymbol{\theta}; \mathbf{X}) \iff \widehat{\boldsymbol{\theta}} \in \arg\max_{\boldsymbol{\theta}\in S_{\boldsymbol{\theta}}} \log L(\boldsymbol{\theta}; \mathbf{X})$$

Maximum Likelihood Estimation



Example

Suppose that X is a discrete random variable with the following probability mass function:

X	0	1	2	3
p(X)	$2\theta/3$	$\theta/3$	$2(1-\theta)/3$	$(1-\theta)/3$

where $0 < \theta < 1$ is a parameter. The following 10 independent observations were taken from such a distribution:
$\mathbf{x} = (x_1, \ldots, x_{10}) = (3, 0, 2, 1, 3, 2, 1, 0, 2, 1)$.

Find a point estimate of θ using the MLE.

Solution:

- The likelihood function given the observations

$\mathbf{x} = (x_1, \ldots, x_{10}) = (3, 0, 2, 1, 3, 2, 1, 0, 2, 1)$ is given by

$$\begin{aligned} L(\theta; \mathbf{x}) &= \prod_{i=1}^{n} p(x_i|\theta) \\ &= p(X=3|\theta)p(X=0|\theta)p(X=2|\theta)p(X=1|\theta)p(X=3|\theta) \\ &\quad \times p(X=2|\theta)p(X=1|\theta)p(X=0|\theta)p(X=2|\theta)p(X=1|\theta) \\ &= \left(\frac{2\theta}{3}\right)^2 \left(\frac{\theta}{3}\right)^3 \left(\frac{2(1-\theta)}{3}\right)^3 \left(\frac{1-\theta}{3}\right)^2. \end{aligned}$$

$$\Longrightarrow \widehat{\theta} \in \underset{\theta \in (0,1)}{\arg\max} \left(\frac{2\theta}{3}\right)^2 \left(\frac{\theta}{3}\right)^3 \left(\frac{2(1-\theta)}{3}\right)^3 \left(\frac{1-\theta}{3}\right)^2$$

Clearly, the likelihood function is not easy to maximize.
Let's look at the log-likelihood

- The log-likelihood function given the observations $\mathbf{x} = (x_1, \ldots, x_{10}) = (3, 0, 2, 1, 3, 2, 1, 0, 2, 1)$ is

$$\begin{aligned}
\log L(\theta; \mathbf{x}) &= \log \prod_{i=1}^{n} p(x_i|\theta) \\
&= 2\left(\log \frac{2}{3} + \log\theta\right) + 3\left(\log \frac{1}{3} + \log\theta\right) + 3\left(\log \frac{2}{3} + \log(1-\theta)\right) \\
&+ 2\left(\log(\frac{1}{3} - \log(1-\theta)\right) \\
&= \text{Constant} + 5\log\theta + 5\log(1-\theta)
\end{aligned}$$

- Setting the derivative to 0 and solving

$$\frac{\mathrm{d}\log L(\theta)}{\mathrm{d}\theta} = 5\left(\frac{1}{\theta} - \frac{1}{1-\theta}\right) = 0$$

$\widehat{\theta} = \widehat{\theta}(\mathbf{x}) = 0.5$

Example: Estimating mean and variance in a normal population

Given a random sample $\mathbf{X} = (X_1, \ldots, X_n)$ of size n where

$$X_i \overset{\text{i.i.d}}{\sim} N(\mu, \sigma)$$

Derive the Maximum Likelihood estimator for the mean and variance of a Normal random variable

Solution:

- $\theta = (\mu, \sigma^2), \quad S_\theta = \mathbb{R} \times (0, \infty)$

We need to find

$$\left(\widehat{\mu}, \widehat{\sigma}^2\right) \in \underset{(\mu,\sigma^2)}{\arg\max}\, p\left(\mathbf{x}|\mu, \sigma^2\right)$$

Notation: We write $\phi(x|\mu, \sigma)$ for the pdf of a $N(\mu, \sigma)$-distributed random variable, i.e.

$$\phi(x|\mu, \sigma) := \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Example: Estimating mean and variance in a normal population

Given a random sample $\mathbf{X} = (X_1, \ldots, X_n)$ of size n where

$$X_i \overset{\text{i.i.d}}{\sim} N(\mu, \sigma)$$

Derive the Maximum Likelihood estimator for the mean and variance of a Normal random variable

Solution:

- $\theta = (\mu, \sigma^2), \quad S_\theta = \mathbb{R} \times (0, \infty)$

We need to find

$$\left(\widehat{\mu}, \widehat{\sigma}^2\right) \in \underset{(\mu,\sigma^2)}{\arg\max}\, p\left(\mathbf{x}|\mu, \sigma^2\right)$$

Notation: We write $\phi(x|\mu, \sigma)$ for the pdf of a $N(\mu, \sigma)$-distributed random variable, i.e.

$$\phi(x|\mu, \sigma) := \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\widehat{\theta} := \left(\widehat{\mu}, \widehat{\sigma}^2\right) \in \arg\max_{\mu,\sigma^2} p\left(\mathbf{x}|\mu,\sigma^2\right)$$

$$\overset{\text{i.i.d}}{=} \arg\max_{\mu,\sigma^2} \prod_{i=1}^{n} p\left(x_i|\mu,\sigma^2\right)$$

$$= \arg\max_{\mu,\sigma^2} \prod_{i=1}^{n} \phi\left(x_i|\mu,\sigma\right)$$

$$= \arg\max_{\mu,\sigma^2} \sum_{i=1}^{n} \log \phi\left(x_i|\mu,\sigma\right)$$

$$= \arg\max_{\mu,\sigma^2} \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left(x_i-\mu\right)^2\right)\right)$$

$$= \arg\max_{\mu,\sigma^2} \frac{-n}{2}\left(\log(2\pi) + \log(\sigma^2)\right) - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}$$

$$\widehat{\theta} := \left(\widehat{\mu}, \widehat{\sigma}^2\right) \in \arg\max_{\mu,\sigma^2} p\left(\mathbf{x}|\mu,\sigma^2\right)$$

$$\overset{\text{i.i.d}}{=} \arg\max_{\mu,\sigma^2} \prod_{i=1}^{n} p\left(x_i|\mu,\sigma^2\right)$$

$$= \arg\max_{\mu,\sigma^2} \prod_{i=1}^{n} \phi\left(x_i|\mu,\sigma\right)$$

$$= \arg\max_{\mu,\sigma^2} \sum_{i=1}^{n} \log \phi\left(x_i|\mu,\sigma\right)$$

$$= \arg\max_{\mu,\sigma^2} \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left(x_i-\mu\right)^2\right)\right)$$

$$= \arg\max_{\mu,\sigma^2} \frac{-n}{2}\left(\log(2\pi) + \log(\sigma^2)\right) - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}$$

$$\widehat{\theta} := \left(\widehat{\mu}, \widehat{\sigma}^2\right) \in \arg\max_{\mu,\sigma^2} p\left(\mathbf{x}|\mu,\sigma^2\right)$$

$$\overset{\text{i.i.d}}{=} \arg\max_{\mu,\sigma^2} \prod_{i=1}^{n} p\left(x_i|\mu,\sigma^2\right)$$

$$= \arg\max_{\mu,\sigma^2} \prod_{i=1}^{n} \phi\left(x_i|\mu,\sigma\right)$$

$$= \arg\max_{\mu,\sigma^2} \sum_{i=1}^{n} \log \phi\left(x_i|\mu,\sigma\right)$$

$$= \arg\max_{\mu,\sigma^2} \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left(x_i-\mu\right)^2\right)\right)$$

$$= \arg\max_{\mu,\sigma^2} \frac{-n}{2}\left(\log(2\pi) + \log(\sigma^2)\right) - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}$$

$$\widehat{\theta} := \left(\widehat{\mu}, \widehat{\sigma}^2\right) \in \arg\max_{\mu,\sigma^2} p\left(\mathbf{x}|\mu,\sigma^2\right)$$

$$\overset{\text{i.i.d}}{=} \arg\max_{\mu,\sigma^2} \prod_{i=1}^{n} p\left(x_i|\mu,\sigma^2\right)$$

$$= \arg\max_{\mu,\sigma^2} \prod_{i=1}^{n} \phi\left(x_i|\mu,\sigma\right)$$

$$= \arg\max_{\mu,\sigma^2} \sum_{i=1}^{n} \log \phi\left(x_i|\mu,\sigma\right)$$

$$= \arg\max_{\mu,\sigma^2} \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left(x_i-\mu\right)^2\right)\right)$$

$$= \arg\max_{\mu,\sigma^2} \frac{-n}{2}\left(\log(2\pi) + \log(\sigma^2)\right) - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}$$

$$
\begin{aligned}
\widehat{\theta} := \left(\widehat{\mu}, \widehat{\sigma}^2\right) &\in \arg\max_{\mu,\sigma^2} p\left(\mathbf{x}|\mu,\sigma^2\right) \\
&\overset{\text{i.i.d}}{=} \arg\max_{\mu,\sigma^2} \prod_{i=1}^{n} p\left(x_i|\mu,\sigma^2\right) \\
&= \arg\max_{\mu,\sigma^2} \prod_{i=1}^{n} \phi\left(x_i|\mu,\sigma\right) \\
&= \arg\max_{\mu,\sigma^2} \sum_{i=1}^{n} \log \phi\left(x_i|\mu,\sigma\right) \\
&= \arg\max_{\mu,\sigma^2} \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left(x_i-\mu\right)^2\right)\right) \\
&= \arg\max_{\mu,\sigma^2} \frac{-n}{2}\left(\log(2\pi) + \log(\sigma^2)\right) - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}
\end{aligned}
$$

$$
\begin{aligned}
\widehat{\theta} := \left(\widehat{\mu}, \widehat{\sigma}^2\right) &\in \arg\max_{\mu,\sigma^2} p\left(\mathbf{x}|\mu,\sigma^2\right) \\
&\overset{\text{i.i.d}}{=} \arg\max_{\mu,\sigma^2} \prod_{i=1}^{n} p\left(x_i|\mu,\sigma^2\right) \\
&= \arg\max_{\mu,\sigma^2} \prod_{i=1}^{n} \phi\left(x_i|\mu,\sigma\right) \\
&= \arg\max_{\mu,\sigma^2} \sum_{i=1}^{n} \log \phi\left(x_i|\mu,\sigma\right) \\
&= \arg\max_{\mu,\sigma^2} \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left(x_i-\mu\right)^2\right)\right) \\
&= \arg\max_{\mu,\sigma^2} \frac{-n}{2}\left(\log(2\pi) + \log(\sigma^2)\right) - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}
\end{aligned}
$$

- To find the maximizer, we calculate

$$\frac{\partial}{\partial \mu}\Big(- n\tfrac{1}{2}\big(\log(2\pi) + \log(\sigma^2)\big) - \sum_{i=1}^{n} \tfrac{(x_i-\mu)^2}{2\sigma^2}\Big) = \sum_{i=1}^{n} \tfrac{x_i-\mu}{\sigma^2}.$$

- Similarly, setting $v := \sigma^2$ and taking the derivatives yields

$$\begin{aligned}
&\frac{\partial}{\partial \sigma^2}\Big(\tfrac{-n}{2}\big(\log(2\pi) + \log(\sigma^2)\big) - \sum_{i=1}^{n} \tfrac{(x_i-\mu)^2}{2\sigma^2}\Big) \\
&= \frac{\partial}{\partial v}\Big(\tfrac{-n}{2}\big(\log(2\pi) + \log(v)\big) - \sum_{i=1}^{n} \tfrac{(x_i-\mu)^2}{2v}\Big) \\
&= \tfrac{-n}{2}\tfrac{1}{v} + \tfrac{1}{2v^2}\sum_{i=1}^{n}(x_i-\mu)^2 \\
&= \tfrac{-n}{2}\tfrac{1}{\sigma^2} + \tfrac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2
\end{aligned}$$

Setting both derivatives equal to 0 implies

$$\sum_{i=1}^{n} \frac{x_i-\mu}{\sigma^2} = 0 \quad \Longrightarrow \quad \mu = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}_n$$

$$\frac{-n}{2}\frac{1}{v}+\frac{1}{2v^2}\sum_{i=1}^{n}(x_i-\mu)^2 = 0 \quad \Longrightarrow \quad v = \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x}_n)^2$$

Therefore, we obtained the estimators

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}_n$$

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X}_n)^2$$

Note: Don't forget, estimators are random variables!

Note:

$$\mathbb{E}[\widehat{\mu}] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \mu \quad \Longrightarrow \widehat{\mu} \text{ unbiased}$$

But, one can show that

$$\mathbb{E}[\widehat{\sigma}^2] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\right] = \frac{n-1}{n}\sigma^2 \quad \Longrightarrow \widehat{\sigma}^2 \text{ biased}$$

Observe:

In this setting $S^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ is unbiased estimator for σ^2.

Some issues to consider:

1. How do we guarantee that MLE exists?
2. How do we guarantee that MLE is unique?
3. How do we guarantee that calculation of MLE is tractable?
4. Is the Likelihood function convex (related to uniqueness)?
5. Boundary conditions?
6. Numerical sensitivity: in many cases the likelihood function is flat...

These are not statistical questions, but mathematical ones, namely *functional analysis, convex analysis....*

Cramér-Rao Bound (CRLB)

Cramér-Rao Lower Bound (CRLB)

The Cramér-Rao Lower Bound (CRLB) sets a lower bound on the variance of any **unbiased estimator**. This can be extremely useful in several ways:

1. If we find an estimator that achieves the CRLB, then we know that we have found an Minimum Variance Unbiased estimator (MVUE)!
2. The CRLB can provide a benchmark against which we can compare the performance of any unbiased estimator (We know we're doing very well if our estimator is "close" to the CRLB)
3. The CRLB enables us to rule-out impossible estimators. That is, we know that it is physically impossible to find an unbiased estimator that beats the CRLB. This is useful in feasibility studies.
4. The theory behind the CRLB can tell us if an estimator exists which achieves the bound.

Cramér-Rao Lower Bound (CRLB)

The Cramér-Rao Lower Bound (CRLB) sets a lower bound on the variance of any **unbiased estimator**. This can be extremely useful in several ways:

1. If we find an estimator that achieves the CRLB, then we know that we have found an Minimum Variance Unbiased estimator (MVUE)!
2. The CRLB can provide a benchmark against which we can compare the performance of any unbiased estimator (We know we're doing very well if our estimator is "close" to the CRLB)
3. The CRLB enables us to rule-out impossible estimators. That is, we know that it is physically impossible to find an unbiased estimator that beats the CRLB. This is useful in feasibility studies.
4. The theory behind the CRLB can tell us if an estimator exists which achieves the bound.

Cramér-Rao Lower Bound (CRLB)

The Cramér-Rao Lower Bound (CRLB) sets a lower bound on the variance of any **unbiased estimator**. This can be extremely useful in several ways:

1. If we find an estimator that achieves the CRLB, then we know that we have found an Minimum Variance Unbiased estimator (MVUE)!
2. The CRLB can provide a benchmark against which we can compare the performance of any unbiased estimator (We know we're doing very well if our estimator is "close" to the CRLB)
3. The CRLB enables us to rule-out impossible estimators. That is, we know that it is physically impossible to find an unbiased estimator that beats the CRLB. This is useful in feasibility studies.
4. The theory behind the CRLB can tell us if an estimator exists which achieves the bound.

Cramér-Rao Lower Bound (CRLB)

The Cramér-Rao Lower Bound (CRLB) sets a lower bound on the variance of any **unbiased estimator**. This can be extremely useful in several ways:

1. If we find an estimator that achieves the CRLB, then we know that we have found an Minimum Variance Unbiased estimator (MVUE)!
2. The CRLB can provide a benchmark against which we can compare the performance of any unbiased estimator (We know we're doing very well if our estimator is "close" to the CRLB)
3. The CRLB enables us to rule-out impossible estimators. That is, we know that it is physically impossible to find an unbiased estimator that beats the CRLB. This is useful in feasibility studies.
4. The theory behind the CRLB can tell us if an estimator exists which achieves the bound.

Cramér-Rao Lower Bound (CRLB)

Theorem: Cramér-Rao Lower Bound

If $\widehat{\theta}$ is any **unbiased** estimator of θ based on the random sample $\mathbf{X}$, then the variance of the error in the estimator is bounded by the inverse of the Fisher Information $\mathcal{I}$:

$$\mathbb{E}\left[\left\|\widehat{\theta}-\theta\right\|^2\right] = \mathbb{V}\mathrm{ar}(\widehat{\theta}) \geq \mathcal{I}^{-1},$$

where $\mathcal{I}$ is given by:

$$\mathcal{I} = -\mathbb{E}\left[\frac{\mathrm{d}^2 \log p\left(\mathbf{X}|\theta\right)}{\mathrm{d}\theta^2}\right].$$

Cramér-Rao Bound (CRLB)

Definition: Efficient Estimator

An **unbiased** estimator $\widehat{\theta}$ is called efficient if

$$\mathbb{V}\text{ar}(\widehat{\theta}) = \mathcal{I}^{-1}$$

- Efficient estimator is an unbiased estimator with minimal possible variance.

Theorem: Sufficient condition for efficiency

If $\widehat{\theta}$ is an **unbiased** estimator of θ and

$$\frac{\partial \log p(\mathbf{Y}|\theta)}{\partial \theta} = c(\theta)\left(\widehat{\theta} - \theta\right)$$

then $\widehat{\theta}$ is an efficient estimator.

Example

Suppose that $X \sim \text{Bin(m, p)}$, where m is known. The **pmf** is given by

$$p(x;p) = \left(\begin{array}{c} m \\ x \end{array} \right) p^x (1-p)^{m-x}, \; x = 0, 1, \ldots, m.$$

Find the CRLB.

Note: The range of X depends on m, but not on the unknown parameter p. Also, the sample size equals $n = 1$.

Solution:

- The log-likelihood is given by

$$\log p(x;p) = \log \begin{pmatrix} m \\ x \end{pmatrix} + x \log p + (m-x) \log (1-p)$$

- The first derivative is given by:

$$\frac{\partial \log p(x;p)}{\partial p} = \frac{x}{p} - (m-x) \frac{1}{1-p}$$

- The second derivative is given by:

$$\frac{\partial^2 \log p(x;p)}{\partial p^2} = \frac{-x}{p^2} - (m-x) \frac{1}{(1-p)^2}$$

- Therefore the Fisher Information $\mathcal{I}$ satisfies

$$\begin{aligned}\mathcal{I} := -\mathbb{E}\left[\frac{-X}{p^2} - (m-X)\frac{1}{(1-p)^2}\right] &= \frac{\mathbb{E}[X]}{p^2} + (m - \mathbb{E}[X])\frac{1}{(1-p)^2} \\ &= \frac{mp}{p^2} + (m-mp)\frac{1}{(1-p)^2} \\ &= \frac{m}{p(1-p)}\end{aligned}$$

It follows that the CRLB is given by

$$\mathbb{V}\text{ar}\left(\widehat{p}\right) \geq \mathcal{I}^{-1} = \frac{p(1-p)}{m}$$

Cramér-Rao Bound (CRLB)

Example

Consider n observations, such that

$$Y_k = m + W_k, \quad k = \{1, \ldots, n\}$$

where $W_k \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$

1. Find the MLE for m.
2. Is $\widehat{m}$ and efficient estimator?

Cramér-Rao Bound (CRLB)

Example

Consider n observations, such that

$$Y_k = m + W_k, \quad k = \{1, \dots, n\}$$

where $W_k \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$

1. Find the MLE for m.
2. Is $\widehat{m}$ and efficient estimator?

Cramér-Rao Bound (CRLB)

Solution:

1) As $Y_k \overset{i.i.d}{\sim} \mathcal{N}\left(m, \sigma^2\right)$, we know from Slide 18 that

$$\widehat{m} = \frac{\sum_{i=1}^n Y_i}{n} = \bar{Y}_n$$

2)

- $\widehat{m}$ is unbiased, as $\mathbb{E}\left[\widehat{m}\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[Y_i\right] = m$
- Moreover, from calculation on Slide 16–17

$$\begin{aligned}\frac{\partial \log p\left(\mathbf{Y}|m,\sigma^2\right)}{\partial m} &= \sum_{i=1}^n \frac{\left(Y_i - m\right)}{\sigma^2} \\ &= \frac{n}{\sigma^2}\left(\frac{1}{n}\sum_{i=1}^n Y_i - m\right) \\ &= c\left(\widehat{m} - m\right)\end{aligned}$$

$\rightsquigarrow$ efficient estimator

Properties of MLE

The concept of MLE makes sense, but can we scientifically justify it?

Bad news: no optimum properties for finite samples.

Good news: has a few attractive limiting properties.

The concept of MLE makes sense, but can we scientifically justify it?

Bad news: no optimum properties for finite samples.

Good news: has a few attractive limiting properties.

The concept of MLE makes sense, but can we scientifically justify it?

Bad news: no optimum properties for finite samples.

Good news: has a few attractive limiting properties.

Properties of MLE

What are the criteria for a “good" estimator?

- Unbiased.
- Consistency.
- normality.
- efficiency.

Properties of MLE

What are the criteria for a “good" estimator?

- Unbiased.
- Consistency.
- normality.
- efficiency.

Properties of MLE

What are the criteria for a “good" estimator?

- Unbiased.
- Consistency.
- normality.
- efficiency.

Properties of MLE

What are the criteria for a “good" estimator?

- Unbiased.
- Consistency.
- normality.
- efficiency.

The MLE satisfies the following 4 asymptotic properties:
(under some additional regularity and integrability conditions)

Consistency: the sequence of MLEs converges in probability to the value being estimated.

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\widehat{\theta}^{(n)} - \theta\right| > \epsilon\right) = 0 \qquad \forall \varepsilon > 0.$$

Asymptotically unbiased: The MLE satisfies

$$\lim_{n\to\infty} \mathbb{E}\left(\widehat{\theta}^{(n)} - \theta\right) = 0$$

Asymptotic normality: A consistent estimator is called asymptotically normal if for some $\sigma_\infty^2 > 0$ we have that the limiting distribution of $\sqrt{n}\left(\widehat{\theta}^{(n)} - \theta\right)$ is equal $N\left(0, \sigma_\infty^2\right)$, i.e.

$$\lim_{n\to\infty} \sqrt{n}\left(\widehat{\theta}^{(n)} - \theta\right) \xrightarrow{\mathrm{d}} N\left(0, \sigma_\infty\right)$$

Asymptotic efficiency: Moreover, we call a consistent estimator asymptotically efficient if $\sigma_\infty^2 = \mathcal{I}^{-1}$, meaning that

$$\lim_{n\to\infty} \sqrt{n}\left(\widehat{\theta}^{(n)} - \theta\right) \xrightarrow{\mathrm{d}} N\left(0, \sqrt{\mathcal{I}^{-1}}\right)$$

Method of Moments Estimator

Method of Moments Estimator

Facts :

- Moments give good (but not always full!) information about distribution.
- If the distribution has bounded support then moments uniquely determine the law.

Idea:

$\Longrightarrow$ **match sample moments with population moments**

Theorem: Law of Large Numbers

Let $X_1, \ldots, X_n$ be i.i.d random variable with $\mathbb{E}[|X_1|] < \infty$ and denote the mean $\mu = \mathbb{E}[X_1]$. Then

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to \mu$$

Method of Moments

- Let $X_1, X_2, \ldots, X_n$ be a sample from a population with pdf or pmf $p(x|\theta_1, \theta_2, \ldots, \theta_k)$.
- Let the unknown parameter $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$ be k-dimensional.

The method of moments estimation is found by:
1) equating the first k sample moments to the corresponding k population moments,
2) solving the resulting system of simultaneous equations.

- The k-th theoretical/population moment of this random variable is defined as

$$\mu_k = \mathbb{E}\left[X^k\right] = \int x^k p(x|\theta_1, \theta_2, \ldots, \theta_k)\, dx \qquad \text{if } X \text{ continuous}$$

$$\mu_k = \mathbb{E}\left[X^k\right] = \sum_x x^k p(x|\theta_1, \theta_2, \ldots, \theta_k) \qquad \text{if } X \text{ discrete.}$$

- If $X_1, X_2, \ldots, X_n$ are i.i.d. random variables from that distribution, the k-th sample moment is defined as

$$m_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k,$$

thus m_k can be viewed as an estimator for μ_k. From the law of large number, we have $m_k \to \mu_k$ in probability as $n \to \infty$.

Method of Moments:

$$\begin{cases} \mathbb{E}[X] & = \frac{1}{n}\sum_{i=1}^{n} X_i \\ \mathbb{E}\left[X^2\right] & = \frac{1}{n}\sum_{i=1}^{n} X_i^2 \\ \vdots & \vdots \\ \mathbb{E}\left[X^k\right] & = \frac{1}{n}\sum_{i=1}^{n} X_i^k \end{cases}$$

$\Longrightarrow$ **Solve the set of k equations and find $\theta_1, \ldots, \theta_k$.**

Example

Suppose that X is a discrete random variable with the following probability mass function:

X	0	1	2	3
p(X)	$2\theta/3$	$\theta/3$	$2(1-\theta)/3$	$(1-\theta)/3$

where θ is a parameter in $(0, 1)$. The following 10 independent observations were taken from such a distribution:
$\mathbf{x} = (x_1, \ldots, x_{10}) = (3, 0, 2, 1, 3, 2, 1, 0, 2, 1)$.

Find a point estimate of θ using the **method of moments** and **MLE**.

Solution:

- We have only a single parameter to estimate
$\Longrightarrow$ we need to calculate only the first moment.
- The theoretical mean value is

$$\mathbb{E}[X] = \sum_{x=0}^{3} x p(x;\theta) = 0\frac{2\theta}{3} + 1\frac{\theta}{3} + 2\frac{2(1-\theta)}{3} + 3\frac{(1-\theta)}{3} = \frac{7}{3} - 2\theta$$

- The sample mean is

$$\bar{x} = \frac{1}{n}\sum_{i=0}^{n} x_i = \frac{3+0+2+1+3+2+1+0+2+1}{10} = 1.5$$

- We solve the single equation

$$\frac{7}{3} - 2\theta = 1.5$$

and find that $\widehat{\theta} = \frac{5}{12}$.

The likelihood function of X given the observations $\mathbf{x} = (x_1, \ldots, x_{10}) = (3, 0, 2, 1, 3, 2, 1, 0, 2, 1)$ is

$$\begin{aligned} L(\theta; \mathbf{x}) &= \prod_{i=1}^{n} p(x_i|\theta) \\ &= p(X = 3|\theta)p(X = 0|\theta)p(X = 2|\theta)p(X = 1\theta)p(X = 3\theta) \\ &\times p(X = 2\theta)p(X = 1\theta)p(X = 0\theta)p(X = 2\theta)p(X = 1\theta) \\ &= \left(\frac{2\theta}{3}\right)^2 \left(\frac{\theta}{3}\right)^3 \left(\frac{2(1-\theta)}{3}\right)^3 \left(\frac{1-\theta}{3}\right)^2. \end{aligned}$$

$$\widehat{\theta} = \underset{\theta \in (0,1)}{\arg\max} \left(\frac{2\theta}{3}\right)^2 \left(\frac{\theta}{3}\right)^3 \left(\frac{2(1-\theta)}{3}\right)^3 \left(\frac{1-\theta}{3}\right)^2$$

Clearly, the likelihood function is not easy to maximize.
Let's look at the log-likelihood

The log-likelihood function of X given the observations $\mathbf{x} = (x_1, \ldots, x_{10}) = (3, 0, 2, 1, 3, 2, 1, 0, 2, 1)$ is

$$\begin{aligned}
\log L(\theta) &= \log \prod_{i=1}^{n} p(x_i|\theta) \\
&= 2\left(\log\frac{2}{3} + \log\theta\right) + 3\left(\log\frac{1}{3} + \log\theta\right) + 3\left(\log\frac{2}{3} + \log(1-\theta)\right) \\
&+ 2\left(\log(\frac{1}{3} - \log(1-\theta)\right) \\
&= \text{Constant} + 5\log\theta + 5\log(1-\theta)
\end{aligned}$$

Setting the derivative to 0 and solving

$$\frac{\mathrm{d}\log L(\theta)}{\mathrm{d}\theta} = 5\left(\frac{1}{\theta} - \frac{1}{1-\theta}\right) = 0$$

$\widehat{\theta} = 0.5$

(the Method of Moments yields $\widehat{\theta} = 5/12$, which is different from MLE.)

Example

Use the Method of Moments to estimate the parameters μ and σ^2 for the normal density

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

based on i.i.d. random sample $X_1, \ldots, X_n$.

Solution:

- First and second theoretical moments for the normal distribution are

$$\mu_1 = \mathbb{E}[X] = \mu$$
$$\mu_2 = \mathbb{E}\left[X^2\right] = \sigma^2 + \mu^2.$$

- The first and second sample moments are

$$m_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$$
$$m_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

- Solving the equations

$$\mu = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\sigma^2 + \mu^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

- We have the Method of Moments estimator

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\bar{X}_n)^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$

In this case the MLE and MME yield the same estimators.

Example

Let $X_1, \ldots, X_n$ be i.i.d samples with from a uniform distribution on the interval $[a, b]$, that is

$$p(x|a,b) = \begin{cases} \frac{1}{b-a} & , \ a \leq x \leq b \\ 0 & , \text{otherwise} \end{cases}$$

Find the Method of Moments estimator for a, b.

Solution:

- The first two moments are:

$$\mu_1 = \mathbb{E}[X] = \int_a^b x \frac{1}{b-a} \mathrm{d}\, x = \frac{a+b}{2}$$

$$\mu_2 = \mathbb{E}\left[X^2\right] = \int_a^b x^2 \frac{1}{b-a} \mathrm{d}\, x = \frac{a^2+ab+b^2}{3}.$$

- The corresponding sample moments are:

$$m_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$m_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

- We solve the equations:

$$\mu_1 = m_1$$
$$\mu_2 = m_2.$$

and obtain:

$$\widehat{a} = m_1 - \sqrt{3\left(m_2 - m_1^2\right)}$$
$$\widehat{b} = m_1 + \sqrt{3\left(m_2 - m_1^2\right)}$$

Example

Let $X_1, \ldots, X_n$ be i.i.d samples with from a beta distribution ($X \sim \beta(\theta, 1)$) with pdf

$$p(x|\theta) = \theta x^{\theta-1}, \quad 0 \leq x \leq 1, \quad 0 \leq \theta \leq \infty$$

1. Find the MLE for θ.
2. Find the Method of Moments estimator for θ.

The likelihood function is given by

$$p(x|\theta) = \prod_{i=1}^{n} \theta x_i^{\theta-1} = \theta^n \prod_{i=1}^{n} (x_i)^{\theta-1} = \theta^n \left(\prod_{i=1}^{n} x_i \right)^{\theta-1}$$

Its derivative is given by

$$\begin{aligned} \frac{d}{d\theta} \log p(x|\theta) &= \frac{d}{d\theta} \log \left(\theta^n \left(\prod_{i=1}^{n} x_i \right)^{\theta-1} \right) \\ &= \frac{d}{d\theta} \left(n \log \theta + (\theta - 1) \sum_{i=1}^{n} \log (x_i) \right) \\ &= \frac{n}{\theta} + \sum_{i=1}^{n} \log(x_i) \end{aligned}$$

- Set the derivative equal to zero, solve for θ , and replace x_i by X_i to obtain

$$\widehat{\theta} = -\frac{n}{\sum_{i=1}^{n} \log(X_i)}$$

Is this the maximum?

- Let's calculate the second derivative

$$\begin{aligned}\frac{d}{d\theta^2} \log p(x|\theta) &= \frac{d}{d\theta}\left(\frac{n}{\theta} + \sum_{i=1}^{n} \log(x_i)\right) \\ &= -\frac{n}{\theta^2} \leq 0,\end{aligned}$$

so this is the MLE.

The Method of Moments for θ:

- The first moment of $X \sim \beta(\theta, 1)$

$$\mathbb{E}[X] = \frac{\theta}{\theta + 1}$$

- The first sample moment is

$$m_1 = \frac{1}{n}\sum_{i=1}^{n} x_i$$

We solve the equation

$$\frac{\theta}{\theta + 1} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

which yields $\widehat{\theta} = \dfrac{\sum_{i=1}^{n} x_i}{n - \sum_{i=1}^{n} x_i}$

Objectives

Now you should be able to :

- Understand the likelihood principle □
- Understand how to formulate the MLE procedure □
- Apply the CRLB □
- Understand how to formulate the Method of Moments estimation procedure □

Put yourself to the test ! $\rightsquigarrow$ Q7.1 p.355, Q7.2 p.355, Q7.6 p.355, Q7.8 p.355, Q7.10 p.355, Q7.15 p.355