# SDS8102: Advanced Mathematical Statistics

## Lecture 3: Point estimation, MoM, M/Z estimators, MLE

Saptarshi Roy

Fall 2026

Department of Statistics and Data Science
MBZUAI

# Continuation of admissibility/ minimax rule

$$\theta \sim \pi$$

**Optimal Bayes decision:** $\delta_\pi = \arg\min_\delta r(\pi, \delta) := \arg\min_\delta \int R(\theta, \delta)\, \pi(d\theta)$.

**Theorem**

*If $\delta_\pi$ is the unique Bayes rule with finite Bayes risk, then $\delta_\pi$ is* **admissible**.

*If $\delta_\pi$ is such that $r(\pi, \delta_\pi) = \sup_\theta R(\theta, \delta_\pi)$, then $\delta_\pi$ is a* **minimax rule**.

$$\sup_\theta R(\theta, \delta) \geq \int R(\theta, \delta)\pi(d\theta) \geq r_\pi.$$

But, $r_\pi = \sup_\theta R(\theta, \delta_\pi) \Rightarrow \sup_\theta R(\theta, \delta) \geq \sup_\theta R(\theta, \delta_\pi) \quad \forall\, \delta$

# Lecture 3: Point Estimation

$X_1, \ldots, X_n \sim P_\theta, \qquad \theta \in \Theta$, where $\theta$ is unknown.

The goal of point estimation is simple:

Use the observed data to guess $\theta$

or, more generally, some quantity $\tau(\theta)$.

Examples include estimating

$$\mu, \qquad \sigma^2, \qquad p, \qquad P_\theta(X > c).$$

The central question is:

How should we construct a good estimator?

# Learning Objectives

By the end of this lecture, you should be able to

1. Define and evaluate point estimators.
2. Understand bias, variance, and quadratic risk.
3. Construct estimators using the method of moments.
4. Formulate estimators as M-estimators and Z-estimators.
5. Derive maximum likelihood estimators.
6. Compare different principles for constructing estimators.

# Point Estimators

**Definition: Point Estimator**

A **point estimator** of $\tau(\theta)$ is a statistic

$$\widehat{\tau} = \delta(X_1, \ldots, X_n)$$

used to estimate the unknown quantity $\tau(\theta)$.

Before observing the data, $\widehat{\tau}$ is a random variable.

After observing

$$X_1 = x_1, \ldots, X_n = x_n,$$

the realized value

$$\widehat{\tau}(x_1, \ldots, x_n) \quad \text{is called an } \textbf{estimate}.$$

# There Are Many Possible Estimators

Suppose

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2),$$

and we want to estimate $\mu$.

Possible estimators include

$$\widehat{\mu}_1 = \overline{X}, \qquad \widehat{\mu}_2 = X_1, \qquad \widehat{\mu}_3 = \text{median}(X_1, \ldots, X_n),$$

or even

$$\widehat{\mu}_4 = 0.$$

All are legitimate statistics.

# There Are Many Possible Estimators

Suppose

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2),$$

and we want to estimate $\mu$.

Possible estimators include

$$\widehat{\mu}_1 = \overline{X}, \qquad \widehat{\mu}_2 = X_1, \qquad \widehat{\mu}_3 = \text{median}(X_1, \ldots, X_n),$$

or even

$$\widehat{\mu}_4 = 0.$$

All are legitimate statistics.

**All estimators are not good. We need criteria for comparison.**

# Bias

A natural requirement is that an estimator should be correct *on average*.

**Definition: Bias**

The bias of an estimator $\widehat{\tau}$ is

$$\text{Bias}_\theta(\widehat{\tau}) = \mathbb{E}_\theta[\widehat{\tau}] - \tau(\theta).$$

The estimator is **unbiased** if

$$\mathbb{E}_\theta[\widehat{\tau}] = \tau(\theta) \qquad \text{for every } \theta \in \Theta.$$

Unbiasedness concerns the behavior of the estimator under repeated sampling.

## Example: Two Unbiased Estimators

Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$.

Both

$$\widehat{\mu}_1 = X_1 \qquad \text{and} \qquad \widehat{\mu}_2 = \overline{X}$$

are unbiased:

$$\mathbb{E}_\mu[X_1] = \mathbb{E}_\mu[\overline{X}] = \mu.$$

But their variances are

$$\text{Var}(X_1) = \sigma^2, \qquad \text{Var}(\overline{X}) = \frac{\sigma^2}{n}.$$

Thus unbiasedness alone does not tell us which estimator is preferable.

# Quadratic Risk

Recall from Week 1 the squared-error loss

$$L(\theta, a) = \{a - \tau(\theta)\}^2.$$

The corresponding risk is

$$R(\theta, \widehat{\tau}) = \mathbb{E}_\theta \left[\{\widehat{\tau} - \tau(\theta)\}^2\right].$$

This is the **mean squared error** (MSE).

Using the bias–variance decomposition,

$$\boxed{R(\theta, \widehat{\tau}) = \mathrm{Var}_\theta(\widehat{\tau}) + \mathrm{Bias}_\theta(\widehat{\tau})^2.}$$

Thus a small amount of bias may be worthwhile if it substantially reduces variance.

# Bias–Variance Tradeoff

Suppose $X \sim N(\theta, 1)$.

and consider the shrinkage estimator

$$\widehat{\theta}_a = aX, \qquad 0 \leq a \leq 1.$$

Its bias and variance are

$$\text{Bias}_\theta(\widehat{\theta}_a) = -(1-a)\theta, \qquad \text{Var}_\theta(\widehat{\theta}_a) = a^2.$$

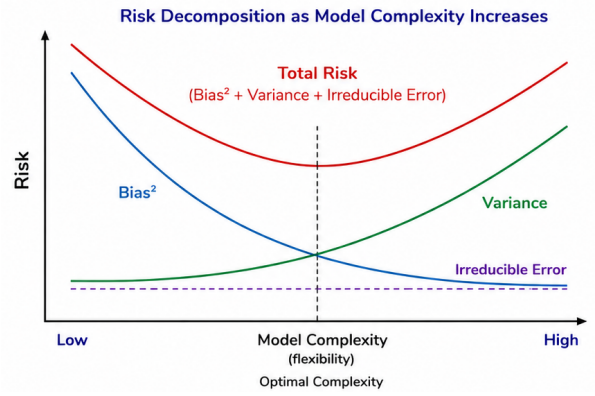
Hence

$$R(\theta, \widehat{\theta}_a) = a^2 + (1-a)^2\theta^2.$$

For $a < 1$, we introduce bias but reduce variance.

We have to find a good balance through bias-variance tradeoff

# Bias-Variance Tradeoff




**High Bias** (underfitting)

- Model too simple to capture the underlying pattern.
- Both training and test risk are high.

**High Variance** (overfitting)

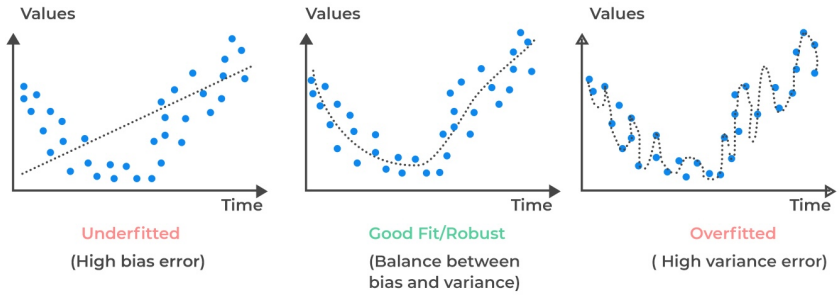
- Model too flexible; fits the noise in the training data.
- Training risk is low, but test risk is high.

**Sweet Spot** (good generalization)

- Balanced bias and variance.
- Minimizes total risk, giving the best out-of-sample performance.

**Figure 1:** Bias-Variance tradeoff (Figure generated by ChatGPT)

# Bias-Variance Tradeoff: generalization and overfitting




**High bias:** Usually comes from simpler models. Leads to underfitting.

**High variance:** Comes from more complex models. Leads to overfitting and generalization ability decreases.

Figure collected from `https://analystprep.com/study-notes/cfa-level-2/quantitative-method/overfitting-methods-addressing/`

# How Do We Construct an Estimator?

So far we have discussed how to *evaluate* an estimator.

But where do estimators come from?

Suppose

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} P_\theta.$$

The distribution $P_\theta$ implies relationships between $\theta$ and population quantities such as

$$\mathbb{E}_\theta[X], \qquad \mathbb{E}_\theta[X^2], \qquad \mathbb{E}_\theta[X^3], \ldots$$

The basic idea of the **method of moments (MoM)** is:

| Replace population moments by their empirical counterparts. |
|---|

## Population and Sample Moments

The $k$-th population moment is

$$\mu_k(\theta) = \mathbb{E}_\theta[X^k].$$

Its empirical analogue is the $k$-th sample moment

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

By the law of large numbers,

$$\widehat{\mu}_k \longrightarrow \mu_k(\theta)$$

under suitable conditions.

This suggests estimating $\theta$ by solving

$$\widehat{\mu}_k = \mu_k(\theta).$$

# Method of Moments

Suppose

$$\theta = (\theta_1, \ldots, \theta_d)$$

contains $d$ unknown parameters.

Choose $d$ population moments

$$\mu_j(\theta) = \mathbb{E}_\theta[X^j], \qquad j = 1, \ldots, d.$$

The **MoM** solves

$$\boxed{\frac{1}{n}\sum_{i=1}^{n} X_i^j = \mu_j(\widehat{\theta}_{\mathrm{MoM}}), \qquad j = 1, \ldots, d.}$$

Thus estimation becomes a system of equations obtained by **matching empirical and theoretical moments**.

# Example: Bernoulli Model

Suppose

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(p).$$

The first population moment is

$$\mathbb{E}_p[X] = p.$$

The corresponding sample moment is

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}.$$

Matching the two gives

$$\overline{X} = \widehat{p}_{\text{MoM}}.$$

Therefore,

$$\boxed{\widehat{p}_{\text{MoM}} = \overline{X}.}$$

# Example: Uniform Model

Suppose

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Unif}(0, \theta).$$

Since

$$\mathbb{E}_\theta[X] = \frac{\theta}{2},$$

the method of moments sets

$$\overline{X} = \frac{\widehat{\theta}_{\text{MoM}}}{2}.$$

Hence

$$\boxed{\widehat{\theta}_{\text{MoM}} = 2\overline{X}.}$$

Notice that the sample maximum $X_{(n)}$ also contains important information about $\theta$.

This already suggests that moment matching need not use all the information available in the likelihood.

# Example: Normal Mean and Variance

Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$.

There are two unknown parameters, so we use two moments:

$$\mathbb{E}[X] = \mu, \qquad \mathbb{E}[X^2] = \mu^2 + \sigma^2.$$

Matching population and sample moments gives

$$\widehat{\mu}_{\text{MoM}} = \overline{X}$$

and

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

Therefore,

$$\widehat{\sigma}^2_{\text{MoM}} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

# Why Does Method of Moments Work?

The method of moments relies on the law of large numbers:

$$\frac{1}{n}\sum_{i=1}^{n} X_i^k \xrightarrow{P} \mathbb{E}_\theta[X^k].$$

If the moment equations uniquely identify $\theta$, then we expect

$$\widehat{\theta}_{\mathrm{MoM}} \xrightarrow{P} \theta.$$

Thus the basic mechanism is

$$\boxed{\text{LLN} \quad + \quad \text{identifiability} \quad \Longrightarrow \quad \text{consistency.}}$$

We will make this argument more rigorous when studying the general theory of estimators.

# Limitations of Method of Moments

Method of moments is simple and often computationally convenient.

But it has limitations.

- Different choices of moments can produce different estimators.
- Moments may not exist.
- Moment equations may have multiple solutions or no valid solution.
- It may discard information contained in the full distribution.
- The resulting estimator need not be statistically efficient.

This raises a broader question:

Can we construct estimators from more general equations or objectives?

# Beyond Moment Matching

The method of moments constructs an estimator by solving equations.

But there are two broader ways to construct estimators:

1. Optimize an empirical objective.
2. Solve an empirical estimating equation.

These lead to two important classes:

| | | |
|---|---|---|
| M-estimators | $\longleftrightarrow$ | optimization |
| Z-estimators | $\longleftrightarrow$ | estimating equations |

Many familiar estimators—including least squares, maximum likelihood, and method of moments—fit into these frameworks.

# M-Estimators

Let $m_\theta(x)$ be a criterion function depending on $\theta$.

**Definition: M-Estimator**

An **M-estimator** is an estimator obtained by optimizing an empirical criterion:

$$\widehat{\theta} \in \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} m_\theta(X_i).$$

Equivalently, one may formulate the problem as maximization.

The population counterpart is

$$M(\theta) = \mathbb{E}[m_\theta(X)].$$

The empirical criterion approximates this population objective.

# Example: Least Squares as an M-Estimator

Suppose

$$Y_i = \theta + \varepsilon_i, \qquad i = 1, \ldots, n.$$

Choose the squared-error criterion

$$m_\theta(Y_i) = (Y_i - \theta)^2.$$

The corresponding M-estimator solves

$$\widehat{\theta} = \arg\min_\theta \frac{1}{n} \sum_{i=1}^{n} (Y_i - \theta)^2.$$

Differentiating,

$$\frac{\partial}{\partial \theta} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \theta)^2 = -\frac{2}{n} \sum_{i=1}^{n} (Y_i - \theta).$$

Setting this equal to zero gives

$$\boxed{\widehat{\theta} = \overline{Y}.}$$

# Changing the Loss Changes the Estimator

Instead of squared error, consider absolute loss:

$$m_\theta(Y_i) = |Y_i - \theta|.$$

Then

$$\widehat{\theta} \in \arg\min_\theta \sum_{i=1}^{n} |Y_i - \theta|.$$

The solution is

$$\boxed{\widehat{\theta} = \text{median}(Y_1, \ldots, Y_n).}$$

Thus

$$\text{squared loss} \quad \longrightarrow \quad \text{sample mean},$$

$$\text{absolute loss} \quad \longrightarrow \quad \text{sample median}.$$

The choice of objective determines which aspects of the data are emphasized.

# Why M-Estimation Is Useful

The M-estimation framework is extremely broad.

It includes:

- least-squares estimation,
- robust estimation,
- maximum likelihood estimation,
- many machine-learning objectives.

The general recipe is simple:

$$\boxed{\text{Choose an empirical objective} \quad \longrightarrow \quad \text{optimize it.}}$$

Different objectives encode different statistical goals and robustness properties.

# Z-Estimators

Instead of optimizing an objective, we can directly solve an equation.

Let

$$\psi_\theta(x) \in \mathbb{R}^d.$$

**Definition: Z-Estimator**

A **Z-estimator** $\widehat{\theta}$ solves

$$\boxed{\frac{1}{n}\sum_{i=1}^{n} \psi_{\widehat{\theta}}(X_i) = 0.}$$

The corresponding population equation is

$$\mathbb{E}_{\theta_0}[\psi_{\theta_0}(X)] = 0,$$

where $\theta_0$ denotes the true parameter.

# Method of Moments as a Z-Estimator

Suppose

$$\mathbb{E}_\theta[X] = \mu(\theta).$$

The method of moments solves

$$\overline{X} - \mu(\theta) = 0.$$

Define

$$\psi_\theta(x) = x - \mu(\theta).$$

Then the estimating equation becomes

$$\frac{1}{n}\sum_{i=1}^{n} \psi_\theta(X_i) = 0.$$

Therefore,

Method of moments is a special case of Z-estimation.

# M-Estimators and Z-Estimators

The two frameworks are closely related.

Suppose an M-estimator minimizes

$$M_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} m_\theta(X_i).$$

If the objective is differentiable, an interior minimizer often satisfies

$$\nabla_\theta M_n(\widehat{\theta}) = 0.$$

Defining

$$\psi_\theta(x) = \nabla_\theta m_\theta(x),$$

we obtain the Z-estimating equation

$$\frac{1}{n}\sum_{i=1}^{n} \psi_{\widehat{\theta}}(X_i) = 0.$$

Hence many M-estimators can also be viewed as Z-estimators.

## But the Two Concepts Are Not Identical

An M-estimator is defined through an **optimization problem**:

$$\widehat{\theta} \in \arg\min_{\theta} M_n(\theta).$$

A Z-estimator is defined through an **estimating equation**:

$$\Psi_n(\widehat{\theta}) = 0.$$

A first-order condition can turn an M-estimation problem into a Z-estimation problem, but:

- an estimating equation need not arise from an objective;
- a root of the gradient need not be a minimizer;
- the objective may not be differentiable.

Thus the two frameworks overlap, but neither definition should be identified with the other.

# Maximum Likelihood Estimation

Suppose the observed data are

$$X_1 = x_1, \ldots, X_n = x_n.$$

Different values of $\theta$ assign different probabilities to the observed data.

A natural principle is:

> Choose the parameter value under which the observed data are most plausible.

This leads to the **maximum likelihood principle**.

# The Likelihood Function

Suppose $X = (X_1, \ldots, X_n)$ has joint density or probability mass function $f_\theta(x)$.

After observing $X = x$, define the likelihood function by

$$L(\theta; x) = f_\theta(x).$$

The crucial distinction is:

$$\begin{aligned} f_\theta(x) &: \text{ function of } x \text{ for fixed } \theta, \\ L(\theta; x) &: \text{ function of } \theta \text{ for fixed } x. \end{aligned}$$

The likelihood is **not** a probability distribution over $\theta$.

# Maximum Likelihood Estimator

**Definition: Maximum Likelihood Estimator**

A maximum likelihood estimator (MLE) is any value

$$\boxed{\widehat{\theta}_{\mathrm{MLE}} \in \arg\max_{\theta \in \Theta} L(\theta; X).}$$

Since the logarithm is strictly increasing, maximizing $L$ is equivalent to maximizing the log-likelihood

$$\ell(\theta; X) := \log L(\theta; X).$$

Thus

$$\widehat{\theta}_{\mathrm{MLE}} \in \arg\max_{\theta \in \Theta} \ell(\theta; X).$$

Log-likelihoods are usually easier to work with analytically and numerically.

# IID Data

If

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} f_\theta,$$

then independence gives

$$L(\theta; X) = \prod_{i=1}^{n} f_\theta(X_i).$$

Therefore the log-likelihood is

$$\boxed{\ell(\theta; X) = \sum_{i=1}^{n} \log f_\theta(X_i).}$$

Likelihood estimation therefore turns a product of densities into an additive empirical objective.

# Example: Bernoulli MLE

Suppose

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(p).$$

The likelihood is

$$L(p) = p^{\sum_i X_i}(1-p)^{n-\sum_i X_i}.$$

Hence

$$\ell(p) = \left(\sum_i X_i\right) \log p + \left(n - \sum_i X_i\right) \log(1-p).$$

Differentiating,

$$\ell'(p) = \frac{\sum_i X_i}{p} - \frac{n - \sum_i X_i}{1-p}.$$

Setting $\ell'(p) = 0$ gives

$$\boxed{\widehat{p}_{\text{MLE}} = \overline{X}.}$$

# Example: Normal Mean

Suppose

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2),$$

where $\sigma^2$ is known.

Ignoring terms that do not depend on $\mu$,

$$\ell(\mu) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2 + C.$$

Therefore maximizing likelihood is equivalent to minimizing

$$\sum_{i=1}^{n} (X_i - \mu)^2.$$

Hence

$$\boxed{\widehat{\mu}_{\text{MLE}} = \overline{X}.}$$

In this example, maximum likelihood and least squares give exactly the same estimator.

# Example: Uniform Model

Recall

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Unif}(0, \theta).$$

The likelihood is

$$L(\theta) = \theta^{-n} \mathbf{1}\{X_{(n)} \leq \theta\},$$

where

$$X_{(n)} = \max_{1 \leq i \leq n} X_i.$$

For $\theta \geq X_{(n)}$, the function $\theta^{-n}$ is decreasing. Therefore the likelihood is maximized at the smallest admissible value:

$$\boxed{\widehat{\theta}_{\text{MLE}} = X_{(n)}.}$$

# MoM versus MLE: Uniform Model

For $X_i \sim \text{Unif}(0, \theta)$, we obtained

$$\widehat{\theta}_{\text{MoM}} = 2\overline{X}$$

but

$$\widehat{\theta}_{\text{MLE}} = X_{(n)}.$$

The two methods use the data differently:

MoM : matches a population moment,

MLE : uses the full likelihood.

Different principles can therefore lead to very different estimators.

Which one is better?

# MLE as an M-Estimator

For iid observations,

$$\widehat{\theta}_{\mathrm{MLE}} \in \arg\max_{\theta} \frac{1}{n} \sum_{i=1}^{n} \log f_\theta(X_i).$$

Equivalently,

$$\widehat{\theta}_{\mathrm{MLE}} \in \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \{-\log f_\theta(X_i)\}.$$

Therefore maximum likelihood is an M-estimator with criterion

$$\boxed{m_\theta(x) = -\log f_\theta(x).}$$

MLE is thus one particularly important member of the much broader M-estimation framework.

# MLE as an Z-Estimator

Suppose the log-likelihood is differentiable.

The **score function** is

$$S_n(\theta) = \nabla_\theta \ell(\theta; X) = \sum_{i=1}^{n} \nabla_\theta \log f_\theta(X_i).$$

An interior MLE typically satisfies the score equation

$$\boxed{S_n(\widehat{\theta}_{\mathrm{MLE}}) = 0.}$$

Thus, under differentiability, the MLE can also be viewed as a Z-estimator with

$$\psi_\theta(x) = \nabla_\theta \log f_\theta(x).$$

# One Estimator, Several Perspectives

Many familiar estimators fit several frameworks simultaneously.

For the Bernoulli model,

$$\widehat{p} = \overline{X}$$

can be viewed as

- a method-of-moments estimator;
- an M-estimator obtained by maximizing likelihood;
- a Z-estimator obtained by solving the score equation.

These are not competing labels.

They are different ways of understanding how an estimator is constructed.

# Invariance of Maximum Likelihood

Suppose $\widehat{\theta}$ is an MLE of $\theta$ and we are interested in a transformed parameter

$$\eta = g(\theta).$$

The invariance property of maximum likelihood says that the MLE of $\eta$ is

$$\boxed{\widehat{\eta}_{\mathrm{MLE}} = g(\widehat{\theta}_{\mathrm{MLE}}).}$$

Thus we do not need to solve a new optimization problem whenever the parameter of interest is transformed.

For a one-to-one transformation, this follows immediately by reparameterizing the likelihood.

# Profile Likelihood for a Transformed Parameter

Suppose $\theta \in \Theta$ and the parameter of interest is

$$\eta = g(\theta), \qquad \eta \in \mathcal{H}.$$

For a fixed value of $\eta$, many values of $\theta$ may satisfy

$$g(\theta) = \eta.$$

Define the **profile likelihood** for $\eta$ by

$$\boxed{L_p(\eta; x) = \sup_{\theta:\, g(\theta)=\eta} L(\theta; x).}$$

Thus $L_p(\eta; x)$ assigns to $\eta$ the largest likelihood among all parameter values compatible with $\eta$.

# Invariance Property of the MLE

**Theorem (Invariance of the MLE)**

*Suppose*

$$\widehat{\theta} \in \arg\max_{\theta \in \Theta} L(\theta; x).$$

*For any function*

$$g : \Theta \to \mathcal{H},$$

*an MLE of* $\eta = g(\theta)$ *is*

$$\boxed{\widehat{\eta} = g(\widehat{\theta}).}$$

Importantly, $g$ need not be one-to-one.

# Proof Using the Profile Likelihood

Let

$$\widehat{\eta} = g(\widehat{\theta}).$$

Since $\widehat{\theta}$ maximizes the likelihood,

$$L(\theta; x) \leq L(\widehat{\theta}; x) \qquad \text{for every } \theta \in \Theta.$$

Therefore, for every $\eta \in \mathcal{H}$,

$$L_p(\eta; x) = \sup_{\theta : g(\theta) = \eta} L(\theta; x) \leq L(\widehat{\theta}; x).$$

But $\widehat{\theta}$ belongs to the fiber $\{\theta : g(\theta) = \widehat{\eta}\}$, so

$$L_p(\widehat{\eta}; x) \geq L(\widehat{\theta}; x).$$

Hence

$$L_p(\widehat{\eta}; x) = L(\widehat{\theta}; x) \geq L_p(\eta; x) \quad \text{for every } \eta \Longrightarrow \boxed{\widehat{\eta} = g(\widehat{\theta}) \text{ maximizes the profile likelihood.}}$$

# Example: Invariance of the MLE

Suppose

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Exponential}(\lambda),$$

with density

$$f_\lambda(x) = \lambda e^{-\lambda x}, \qquad x \geq 0.$$

The MLE of the rate is

$$\widehat{\lambda}_{\text{MLE}} = \frac{1}{\overline{X}}.$$

Suppose instead we want to estimate the mean

$$\mu = \frac{1}{\lambda}.$$

By invariance,

$$\boxed{\widehat{\mu}_{\text{MLE}} = \frac{1}{\widehat{\lambda}_{\text{MLE}}} = \overline{X}.}$$

# Three Principles for Constructing Estimators

We have seen three broad approaches.

| Method | Basic idea | Mathematical form |
|---|---|---|
| MoM | Match moments | $\widehat{\mu} = \mu(\theta)$ |
| M-estimation | Optimize | $\arg\min_\theta M_n(\theta)$ |
| Z-estimation | Solve equations | $\Psi_n(\theta) = 0$ |

Maximum likelihood is an especially important M-estimator:

$$\widehat{\theta}_{\mathrm{MLE}} \in \arg\max_\theta \sum_{i=1}^{n} \log f_\theta(X_i).$$

Under differentiability, its score equation also gives a Z-estimation perspective.

# What Have We Not Answered Yet?

Maximum likelihood assumes that we have specified a full probability model

$$\{P_\theta : \theta \in \Theta\}.$$

But several questions remain.

- What if specifying the full likelihood is difficult?
- What if only part of the model is trustworthy?
- Why does maximum likelihood work so well?
- How can we compare nearby probability distributions?
- What if maximizing the likelihood is computationally difficult?

These questions lead to

| Quasi- and conditional likelihood |
| --- |
| KL and Hellinger geometry |
| EM algorithm |