

# Lesson 2 — Statistical Learning: Parameter Estimation

MLE, Method of Moments, Fisher Information, Uncertainty

Applied Statistics Course

October 15, 2025

- Derive estimators using **maximum likelihood** and **method of moments**.
- Compute **standard errors** via Fisher information and the **delta method**.
- Assess estimators: **bias, variance, MSE**; use **likelihood profiles** and **bootstrap**.
- Implement simulation studies to compare procedures.

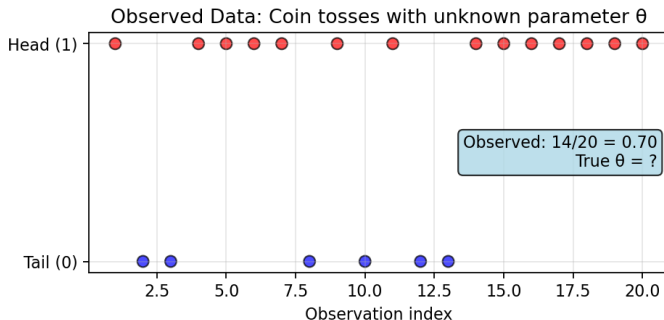
Recall Lesson 1: random variables, laws (PMF/PDF/CDF), LLN/CLT, and notation ( $\mathbb{P}$ ,  $\mathbb{E}$ ,  $\text{Var}$ ).

## 1 Maximum Likelihood Estimation (MLE)

- Likelihood
- Maximum Likelihood Estimate

## 2 Method of Moments

# Parameter Estimation: Why Do We Care?

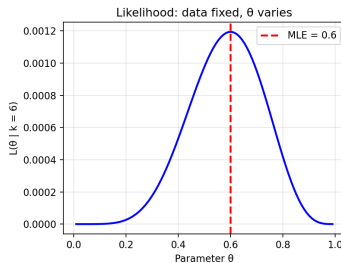
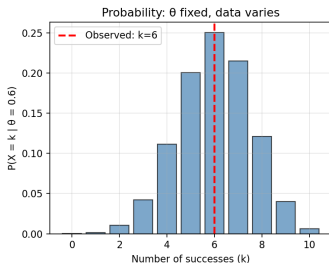


## The Central Question

Given observed data, what parameter values make this data most plausible under our model?

- We have data  $x_1, x_2, \dots, x_n$
- We assume a probabilistic model  $f(x \mid \theta)$
- We want to find the “best” estimate  $\hat{\theta}$

# Probability vs. Likelihood: The Key Duality



## Probability

Fixed  $\theta$ , varying data  $x$   
 $P(X = x | \theta)$

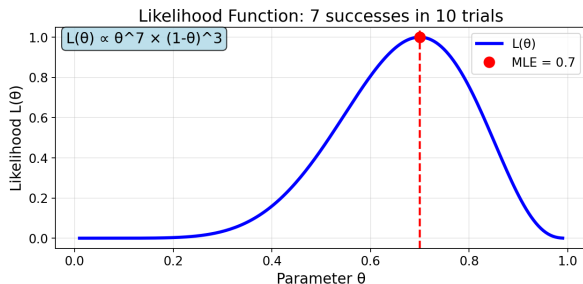
## Likelihood

Fixed data  $x$ , varying  $\theta$   
 $L(\theta | x) \propto P(X = x | \theta)$

## Key Insight

Same mathematical function, but we're asking different questions!

# What is the Likelihood Function?



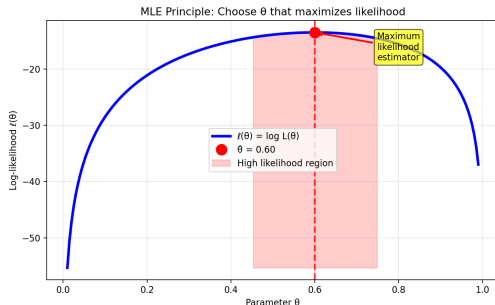
## Definition

For observed data  $x_1, \dots, x_n$  and model  $f(x | \theta)$ :

$$L(\theta) = \prod_{i=1}^n f(x_i | \theta)$$

- Measures how well parameter  $\theta$  explains the observed data
- Higher likelihood  $\Rightarrow$  parameter is more supported by data
- Often work with log-likelihood:  $\ell(\theta) = \sum_{i=1}^n \log f(x_i | \theta)$

# The Maximum Likelihood Principle



## MLE Definition

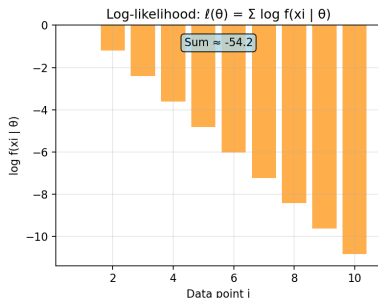
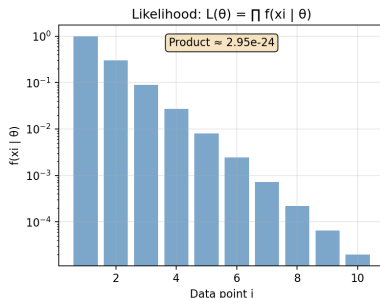
The Maximum Likelihood Estimator (MLE) is:

$$\hat{\theta}_{\text{MLE}} = \arg \max_{\theta} L(\theta) = \arg \max_{\theta} \ell(\theta)$$

## Intuition

Choose the parameter value that makes our observed data as likely as possible.

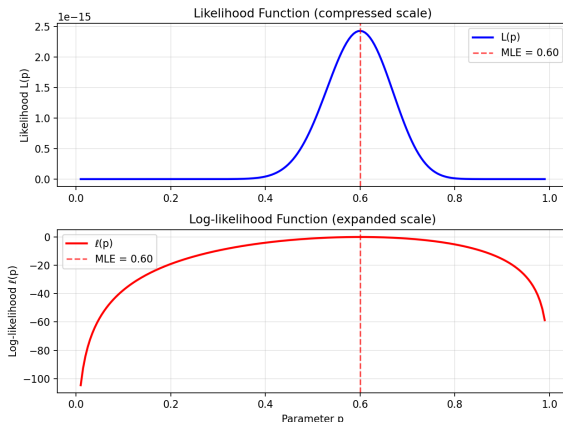
# From Likelihood to Log-Likelihood



- Likelihood is a product of terms:  $L(\theta) = \prod_{i=1}^n f_{\theta}(x_i)$
- Products are difficult to optimize when  $n$  is large
- Take the log to simplify:  $\ell(\theta) = \log L(\theta) = \sum_{i=1}^n \log f_{\theta}(x_i)$
- Since log is strictly increasing, maximizing  $L$  or  $\ell$  is equivalent

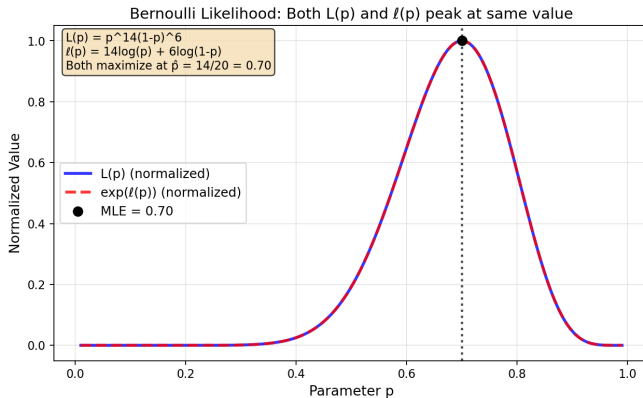


# Why use log-likelihood in practice?



- **Numerical stability:** avoids underflow when multiplying many small numbers
- **Simplifies optimization:** turns products into sums, easier differentiation
- **Reveals structure:** concavity/convexity properties often clearer
- **Same maximum:**  $\arg \max L(\theta) = \arg \max \ell(\theta)$

# Log-Likelihood for the Bernoulli Model



## Bernoulli Case

Likelihood:  $L(p) = p^{\sum x_i} (1-p)^{n-\sum x_i}$

Log-likelihood:  $\ell(p) = (\sum x_i) \log p + (n - \sum x_i) \log(1-p)$

Easier to differentiate, leads to closed-form solution:  $\hat{p} = \frac{1}{n} \sum x_i$

## Example: Coin Flips (Bernoulli Model)

**Setup:**  $n$  coin flips,  $k$  heads observed. Model:  $X_i \sim \text{Bernoulli}(p)$

**Likelihood:**  $L(p) = p^k(1 - p)^{n-k}$

**Log-likelihood:**  $\ell(p) = k \log p + (n - k) \log(1 - p)$

**Find MLE:**  $\frac{d\ell}{dp} = \frac{k}{p} - \frac{n-k}{1-p} = 0$

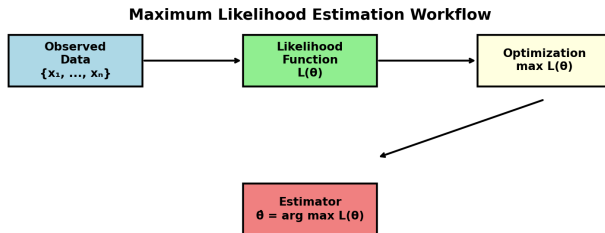
### Result

$$\hat{p}_{\text{MLE}} = \frac{k}{n} = \bar{x}$$

### Makes Sense!

The proportion of heads in our sample is the most likely value for the coin's bias.

# MLE: Key Takeaways

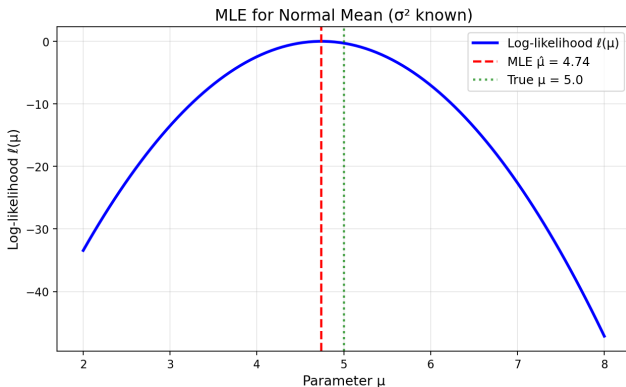


- **Intuitive:** Choose parameters that best explain observed data
- **General:** Works for any probabilistic model
- **Principled:** Solid theoretical foundation
- **Practical:** Often gives closed-form solutions

## Coming Up

More examples, properties, and when MLE works well (or doesn't!)

# MLE for the Normal Mean ( $\sigma^2$ known)



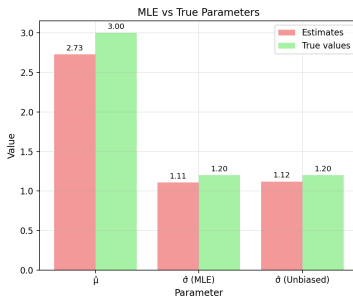
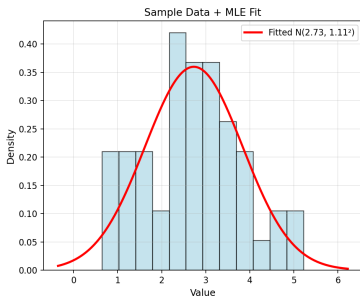
## Model & Solution

Model:  $X_i \sim \mathcal{N}(\mu, \sigma^2)$  with  $\sigma^2$  known

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

MLE:  $\hat{\mu}_{MLE} = \bar{X}_n$

# MLE for Normal ( $\mu, \sigma^2$ unknown)



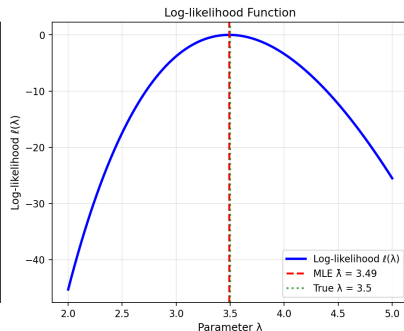
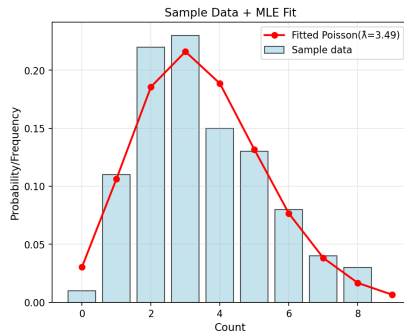
## Joint Estimation

Model:  $X_i \sim \mathcal{N}(\mu, \sigma^2)$

MLEs:  $\hat{\mu}_{MLE} = \bar{X}_n$ ,  $\hat{\sigma}_{MLE}^2 = \frac{1}{n} \sum (x_i - \bar{X}_n)^2$

Note: MLE uses  $n$ , not  $n - 1$

# MLE for Poisson Parameter $\lambda$



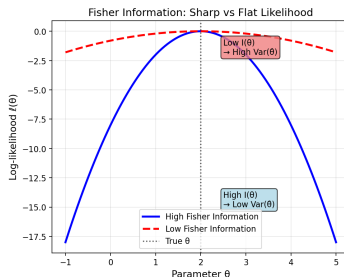
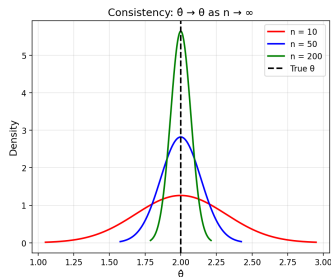
## Discrete Distribution MLE

Model:  $X_i \sim \text{Poisson}(\lambda)$

Log-likelihood:  $\ell(\lambda) = \sum (x_i \log \lambda - \lambda - \log(x_i!))$

MLE:  $\hat{\lambda}_{MLE} = \bar{X}_n$

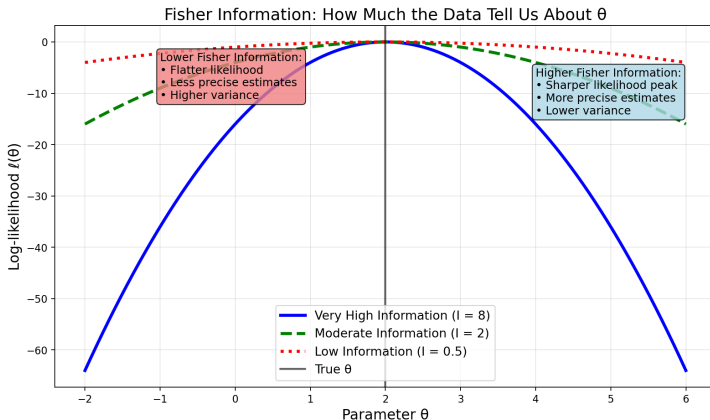
# Theoretical Properties of MLE



- **Consistency:**  $\hat{\theta}_{MLE} \rightarrow \theta$
- **Asymptotic Normality:**  $\sqrt{n}(\hat{\theta}_{MLE} - \theta) \xrightarrow{d} \mathcal{N}(0, I(\theta)^{-1})$
- **Efficiency:** achieves the CramérRao lower bound asymptotically



# Fisher Information

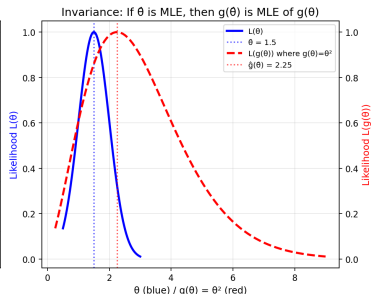
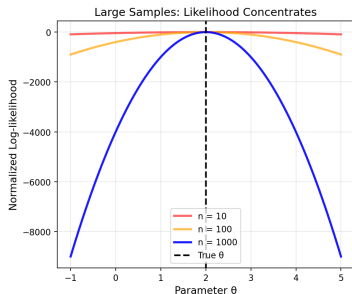


## Definition & Intuition

Fisher Information:  $I(\theta) = -\mathbb{E} \left[ \frac{\partial^2}{\partial \theta^2} \ell(\theta) \right]$

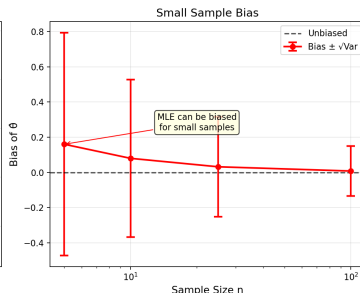
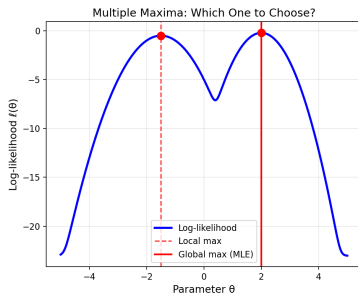
High information sharper peak lower variance

# Strengths of MLE



- Works well for large  $n$  (asymptotic guarantees)
- Very flexible, applicable to many models
- **Invariance property:** if  $\hat{\theta}$  is MLE of  $\theta$ , then  $g(\hat{\theta})$  is MLE of  $g(\theta)$

# Limitations of MLE



- Small  $n$  bias and instability
- Non-identifiability multiple maxima
- Likelihood surface can be flat or multimodal
- Sensitive to model misspecification

# MLE: Strengths and Limitations

## ✓ MLE Works Well When...

✓ Sufficient data

✓ Regular likelihood surface

✓ Identifiable parameters

✓ Correct model specification

✓ Large sample size ( $n \rightarrow \infty$ )

## ⚠ MLE May Struggle When...

⚠ Insufficient/poor data

⚠ Multimodal likelihood

⚠ Non-identifiable parameters

⚠ Model misspecification

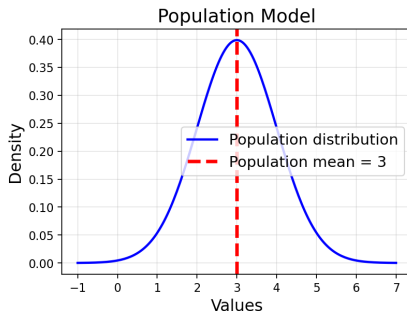
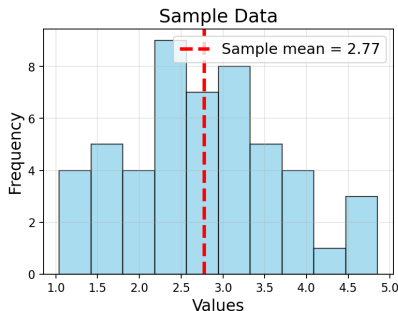
⚠ Small sample size (finite  $n$ )

- MLE is powerful and widely used
- Asymptotically consistent, normal, and efficient
- Must be cautious with small samples or misspecified models

- 1 Maximum Likelihood Estimation (MLE)
- 2 Method of Moments

# Why another estimation method?

- MLE is powerful, but sometimes hard to compute.
- Method of Moments (MoM) offers a simpler alternative.
- Idea: match sample moments with theoretical moments.



# The Method of Moments

- For model parameter  $\theta$ , theoretical moment:

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

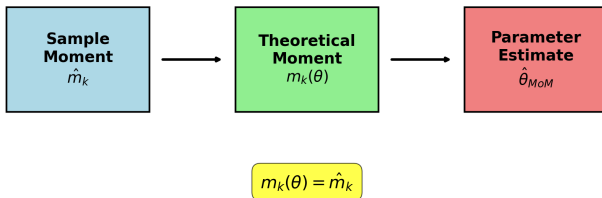
- Empirical moment:

$$\hat{m}_k = \frac{1}{n} \sum_{i=1}^n X_i^k.$$

- Solve equations:

$$m_k(\theta) = \hat{m}_k.$$

## Method of Moments Principle



## Setting up the system of equations:

Theoretical first moment:  $E[X] = p$  (1)

Sample first moment:  $\hat{m}_1 = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}_n$  (2)

## Method of Moments equation:

$$E[X] = \hat{m}_1 \quad (3)$$

$$p = \bar{X}_n \quad (4)$$

## Solving for the parameter:

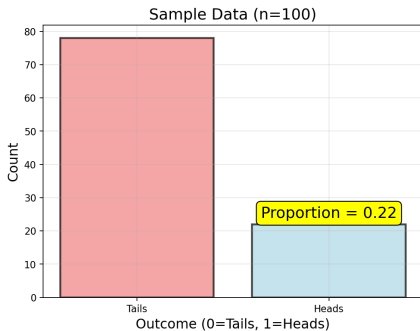
$$\hat{p}_{MoM} = \bar{X}_n \quad (5)$$

*This shows that MoM reduces to solving a simple algebraic equation!*



# MoM for Bernoulli( $p$ )

- Theoretical mean:  $E[X] = p$ .
- Sample mean:  $\hat{m}_1 = \bar{X}_n$ .
- Solve:  $\hat{p}_{MoM} = \bar{X}_n$ .
- Note: coincides with MLE.



## Method of Moments:

$$E[X] = p$$

$$\hat{m}_1 = \bar{X}_n$$

$$\hat{p}_{MoM} = \bar{X}_n$$

$$\hat{p}_{MoM} = 0.22$$

## Setting up the system of equations:

Theoretical first moment:  $E[X] = \lambda$  (6)

Sample first moment:  $\hat{m}_1 = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}_n$  (7)

## Method of Moments equation:

$$E[X] = \hat{m}_1 \quad (8)$$

$$\lambda = \bar{X}_n \quad (9)$$

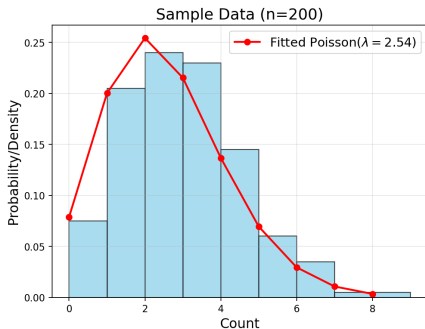
## Solving for the parameter:

$$\hat{\lambda}_{MoM} = \bar{X}_n \quad (10)$$

*Again, we solve a simple equation: parameter = sample mean!*

# MoM for Poisson( $\lambda$ )

- Theoretical mean:  $E[X] = \lambda$ .
- Sample mean:  $\hat{m}_1 = \bar{X}_n$ .
- Solve:  $\hat{\lambda}_{MoM} = \bar{X}_n$ .
- Note: coincides with MLE.



## Method of Moments:

$$E[X] = \lambda$$

$$\hat{m}_1 = \bar{X}_n$$

$$\hat{\lambda}_{MoM} = \bar{X}_n$$

$$\hat{\lambda}_{MoM} = 2.54$$

## Setting up the system of equations (2 parameters $\Rightarrow$ 2 moments):

Theoretical moments:  $E[X] = \mu, \quad E[X^2] = \mu^2 + \sigma^2 \quad (11)$

Sample moments:  $\hat{m}_1 = \bar{X}_n, \quad \hat{m}_2 = \frac{1}{n} \sum_{i=1}^n X_i^2 \quad (12)$

## Method of Moments equations:

$$\mu = \bar{X}_n \quad (13)$$

$$\mu^2 + \sigma^2 = \hat{m}_2 \quad (14)$$

## Solving the system:

$$\hat{\mu}_{MoM} = \bar{X}_n \quad (15)$$

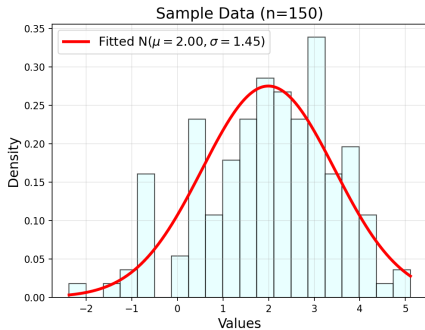
$$\hat{\sigma}_{MoM}^2 = \hat{m}_2 - (\hat{\mu}_{MoM})^2 = \hat{m}_2 - (\bar{X}_n)^2 \quad (16)$$

*Two parameters require solving a system of two equations!*

# MoM for Normal( $\mu, \sigma^2$ )

- $E[X] = \mu, E[X^2] = \mu^2 + \sigma^2$ .
- Empirical moments:  $\hat{m}_1 = \bar{X}_n, \hat{m}_2 = \frac{1}{n} \sum X_i^2$ .
- Solve:

$$\hat{\mu}_{MoM} = \bar{X}_n, \quad \hat{\sigma}_{MoM}^2 = \hat{m}_2 - (\bar{X}_n)^2.$$



## Method of Moments:

$$E[X] = \mu, \quad E[X^2] = \mu^2 + \sigma^2$$

$$\hat{m}_1 = \bar{X}_n, \quad \hat{m}_2 = \frac{1}{n} \sum X_i^2$$

$$\hat{\mu}_{MoM} = \bar{X}_n$$

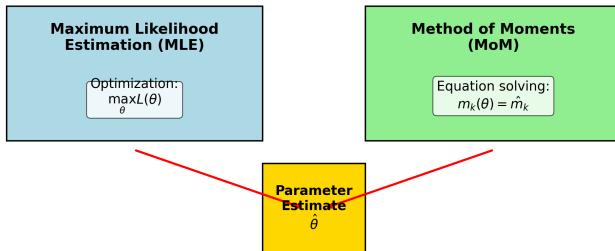
$$\hat{\sigma}_{MoM}^2 = \hat{m}_2 - (\bar{X}_n)^2$$

$$\hat{\mu}_{MoM} = 2.00, \hat{\sigma}_{MoM} = 1.45$$

# MLE vs MoM: Similarities and Differences

- Both provide consistent estimators (under conditions).
- MLE is asymptotically efficient; MoM is not guaranteed to be.
- MoM often easier to compute (simple equations).
- MoM can give nonsensical estimates (e.g., negative variance).
- In simple models,  $\text{MLE} = \text{MoM}$ .

## Two Roads to Parameter Estimation



# Normal Variance Estimate (MLE vs MoM)

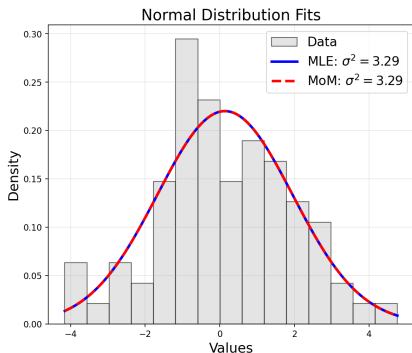
- MLE:

$$\hat{\sigma}_{MLE}^2 = \frac{1}{n} \sum (X_i - \bar{X}_n)^2.$$

- MoM:

$$\hat{\sigma}_{MoM}^2 = \hat{m}_2 - (\bar{X}_n)^2.$$

- Here they are equal, but in other models they may differ.



## Variance Estimators:

### MLE:

$$\hat{\sigma}_{MLE}^2 = \frac{1}{n} \sum (X_i - \bar{X}_n)^2$$

### MoM:

$$\hat{\sigma}_{MoM}^2 = \hat{m}_2 - (\bar{X}_n)^2$$

In this case: MLE = 3.288, MoM = 3.288

*Note: These are equal for Normal distribution!*

## MLE vs MoM: Summary Comparison

Property	MLE	MoM
<b>Computation</b>	Requires optimization	Simple equations
<b>Large-sample behavior</b>	Consistent, efficient	Consistent, less efficient
<b>Small-sample behavior</b>	Biased but systematic	Can be unstable, nonsensical
<b>Flexibility</b>	Works for many models	Requires finite moments
<b>Coincidence</b>	Often equals MoM in simple cases	Same as MLE sometimes



# Key Takeaways: Method of Moments

- MoM: intuitive, simple, often first attempt.
- MLE: more powerful, statistically optimal in large samples.
- Both are fundamental tools in parameter estimation.

## Method of Moments: Key Takeaways

