Lesson 2 — Statistical Learning: Parameter Estimation

MLE, Method of Moments, Fisher Information, Uncertainty

Applied Statistics Course

September 25, 2025

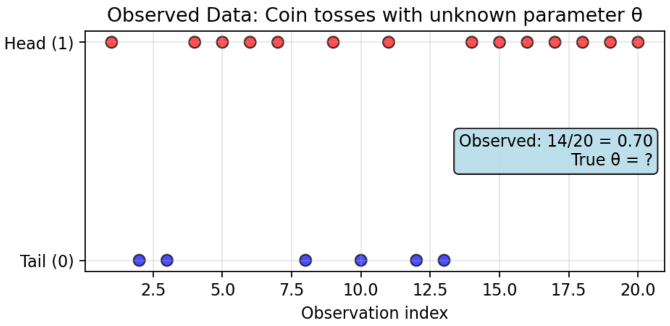
Learning Objectives

- 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 ($\mathbb{P}, \mathbb{E}, \mathrm{Var}$).

Outline

Parameter Estimation: Why Do We Care?

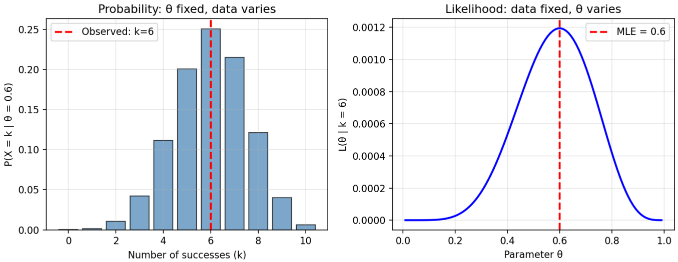


The Central Question

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

- We have data $x_1, x_2, \ldots, x_n$
- 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 θ, varying data x

$P(X = x \mid \theta)$

Likelihood

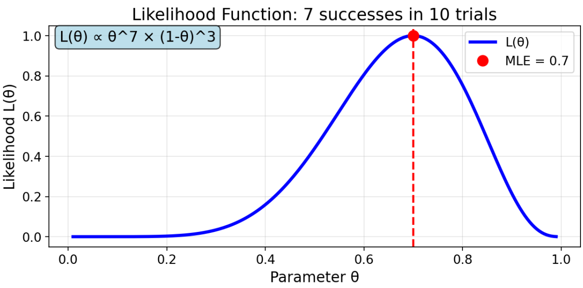
Fixed data x, varying θ

$L(\theta \mid x) \propto P(X = x \mid \theta)$

Key Insight

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

What is the Likelihood Function?



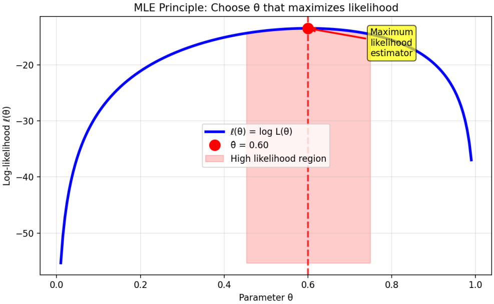
Definition

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

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

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

The Maximum Likelihood Principle



MLE Definition

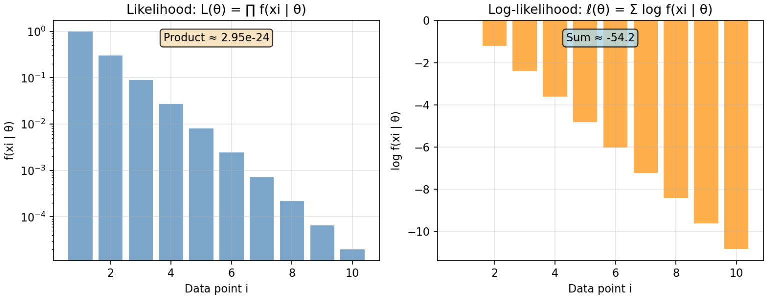
The Maximum Likelihood Estimator (MLE) is:

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

Intuition

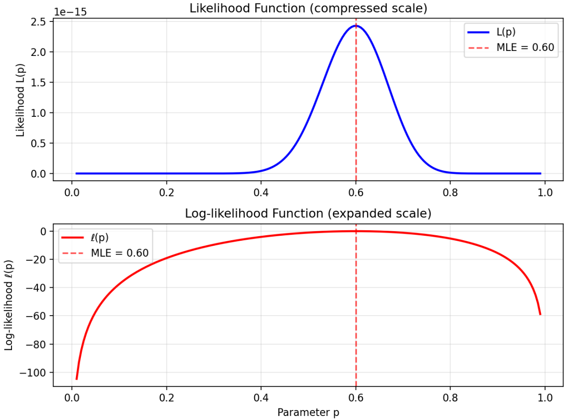
Choose the parameter value that makes our observed data as likely as possible.

From Likelihood to Log-Likelihood



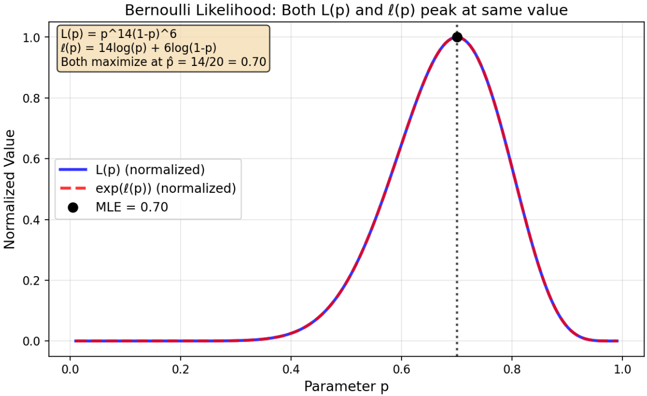
- Likelihood is a product of terms: $L(\theta) = \prod_{i=1}^{n} f_\theta(x_i)$
- Products are difficult to optimize when n is large
- Take the log to simplify: $\ell(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f_\theta(x_i)$
- Since log is strictly increasing, maximizing L or ℓ is equivalent

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

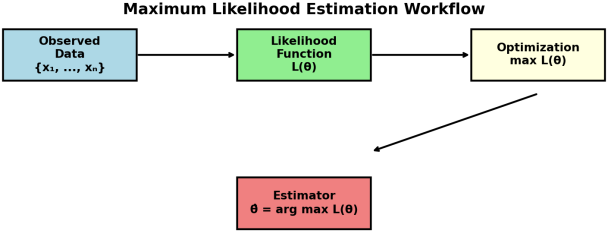
Result

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

Makes Sense!

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

MLE: Key Takeaways

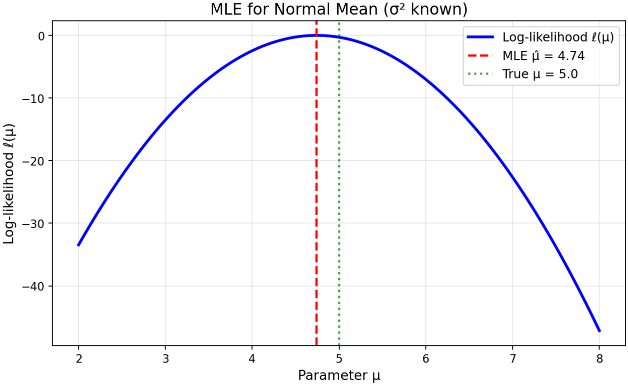


- **Intuitive:** Choose parameters that best explain observed data
- **General:** Works for any probabilistic model
- **Principled:** Solid theoretical foundation
- **Practical:** Often gives closed-form solutions

Coming Up

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

MLE for the Normal Mean (σ^2 known)



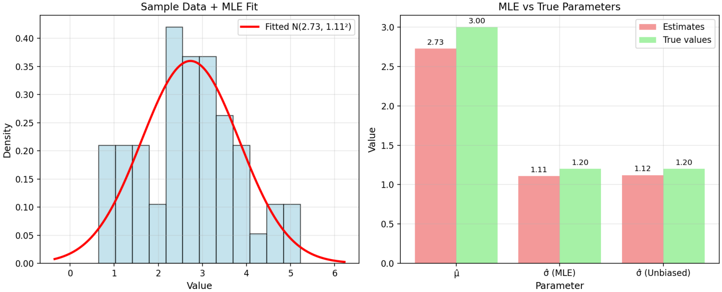
Model & Solution

Model: $X_i \sim \mathcal{N}(\mu, \sigma^2)$ with σ^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: $\hat{\mu}_{MLE} = \overline{X}_n$

MLE for Normal (μ, σ^2 unknown)



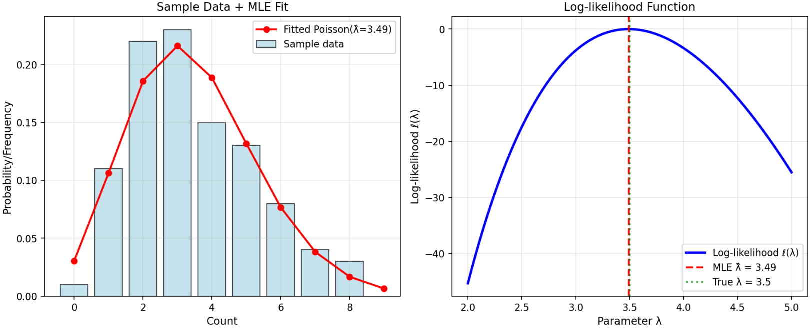
Joint Estimation

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

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

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

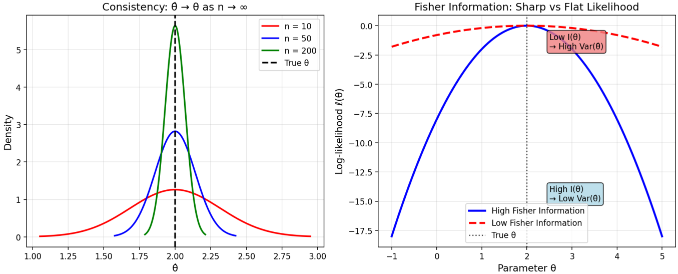
MLE for Poisson Parameter λ



Discrete Distribution MLE

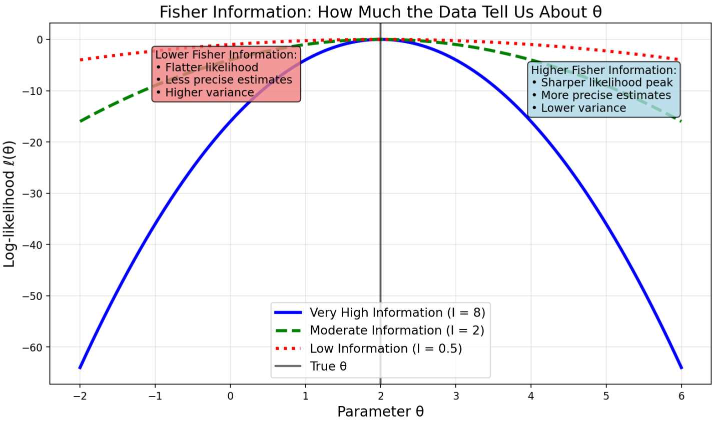
Model: $X_i \sim \text{Poisson}(\lambda)$
Log-likelihood: $\ell(\lambda) = \sum(x_i \log \lambda - \lambda - \log(x_i!))$
MLE: $\hat{\lambda}_{MLE} = \overline{X}_n$

Theoretical Properties of MLE



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

Fisher Information

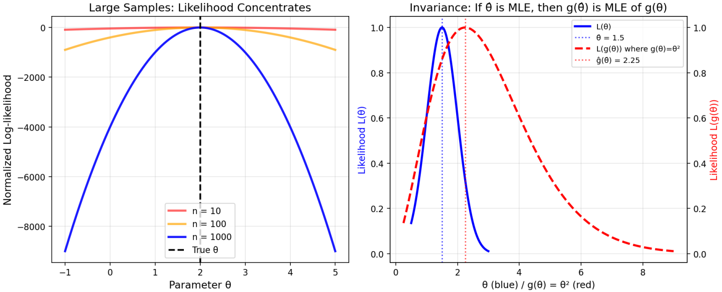


Definition & Intuition

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

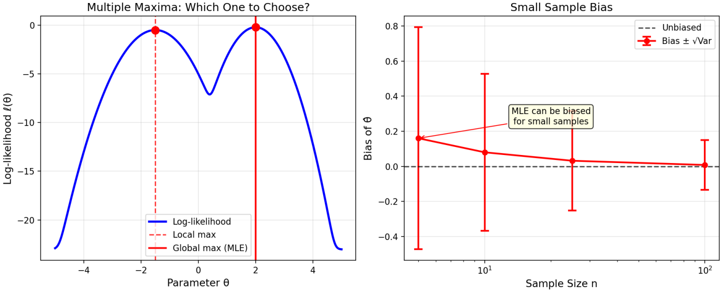
High information $\rightarrow$ sharper peak $\rightarrow$ lower variance

Strengths of MLE



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

Limitations of MLE



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

MLE: Strengths and Limitations

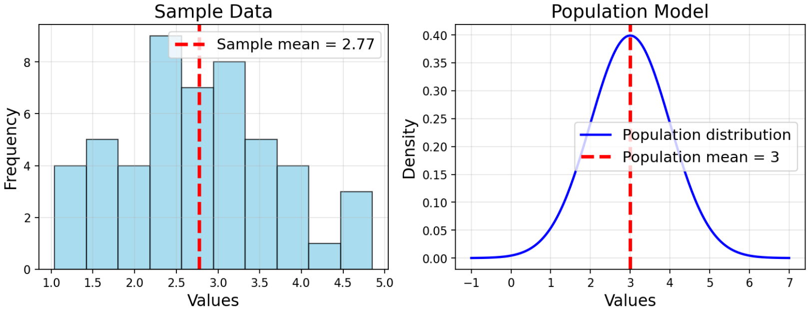
✓ MLE Works Well When...	⚠ MLE May Struggle When...
✓ **Sufficient data**	⚠ **Insufficient/poor data**
✓ **Regular likelihood surface**	⚠ **Multimodal likelihood**
✓ **Identifiable parameters**	⚠ **Non-identifiable parameters**
✓ **Correct model specification**	⚠ **Model misspecification**
✓ **Large sample size ($n \to \infty$)**	⚠ **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

Outline

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.



- For model parameter θ, theoretical moment:

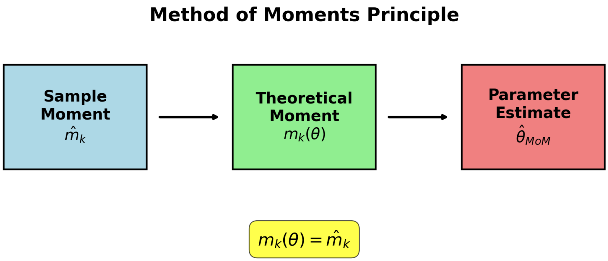
$$m_k(\theta) = \mathbb{E}_\theta[X^k].$$

- Empirical moment:

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

- Solve equations:

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



MoM for Bernoulli: Step-by-Step Derivation

Setting up the system of equations:

$$\text{Theoretical first moment:} \quad E[X] = p \tag{1}$$

$$\text{Sample first moment:} \quad \hat{m}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}_n \tag{2}$$

Method of Moments equation:

$$E[X] = \hat{m}_1 \tag{3}$$

$$p = \overline{X}_n \tag{4}$$

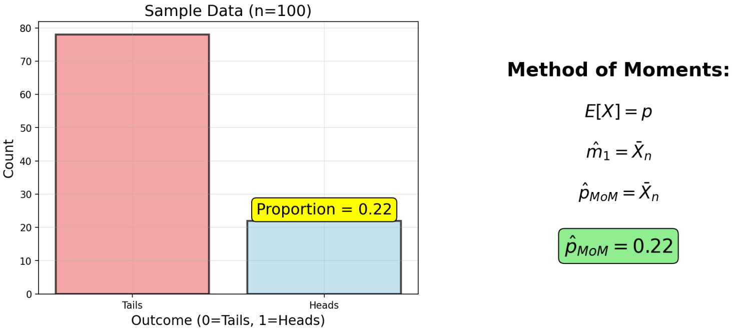
Solving for the parameter:

$$\hat{p}_{MoM} = \overline{X}_n \tag{5}$$

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

MoM for Bernoulli(p)

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



MoM for Poisson: Step-by-Step Derivation

Setting up the system of equations:

$$\text{Theoretical first moment:} \quad E[X] = \lambda \tag{6}$$

$$\text{Sample first moment:} \quad \hat{m}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}_n \tag{7}$$

Method of Moments equation:

$$E[X] = \hat{m}_1 \tag{8}$$

$$\lambda = \overline{X}_n \tag{9}$$

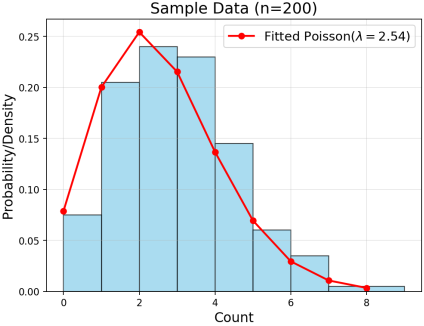
Solving for the parameter:

$$\hat{\lambda}_{MoM} = \overline{X}_n \tag{10}$$

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

MoM for Poisson(λ)

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



Method of Moments:

$$E[X] = \lambda$$

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

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

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

MoM for Normal: Step-by-Step Derivation

Setting up the system of equations (2 parameters ⇒ 2 moments):

$$\text{Theoretical moments:} \quad E[X] = \mu, \quad E[X^2] = \mu^2 + \sigma^2 \tag{11}$$

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

Method of Moments equations:

$$\mu = \overline{X}_n \tag{13}$$

$$\mu^2 + \sigma^2 = \hat{m}_2 \tag{14}$$

Solving the system:

$$\hat{\mu}_{MoM} = \overline{X}_n \tag{15}$$

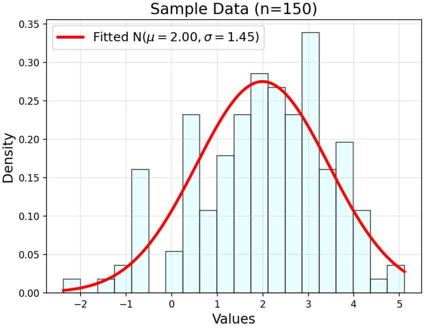
$$\hat{\sigma}^2_{MoM} = \hat{m}_2 - (\hat{\mu}_{MoM})^2 = \hat{m}_2 - (\overline{X}_n)^2 \tag{16}$$

Two parameters require solving a system of two equations!

MoM for Normal(μ, σ^2)

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

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



Method of Moments:

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

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

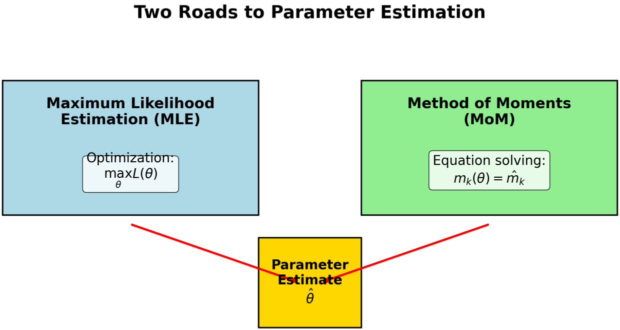
$\hat{\mu}_{MoM} = \bar{X}_n$

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

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

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, MLE = MoM.



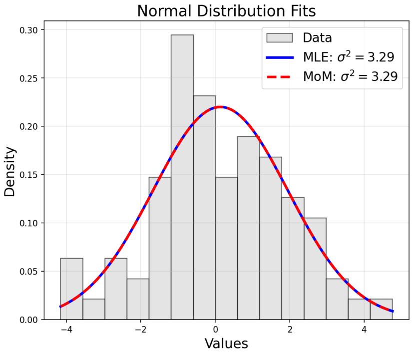
- MLE:

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

- MoM:

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

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



Variance Estimators:

MLE:

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

MoM:

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

In this case: MLE = 3.288, MoM = 3.288

Note: These are equal for Normal distribution!

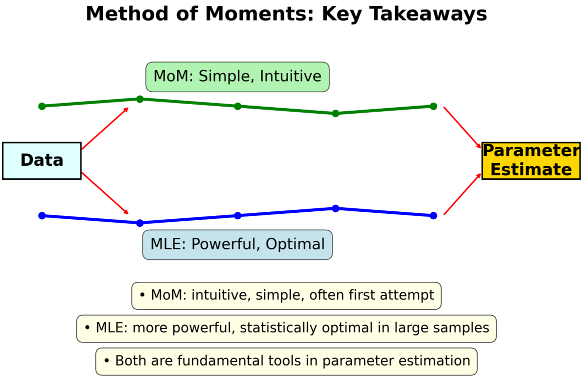
MLE vs MoM: Summary

MLE vs MoM: Summary Comparison

Property	MLE	MoM
Computation	Requires optimization	Simple equations
Large-sample behavior	Consistent, efficient	Consistent, less efficient
Small-sample behavior	Biased but systematic	Can be unstable, nonsensical
Flexibility	Works for many models	Requires finite moments
Coincidence	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.



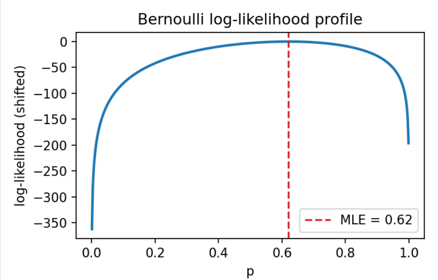
Outline

Definitions

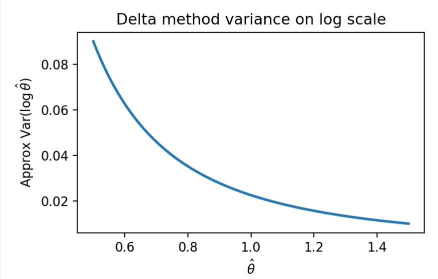
Fisher information: $I(\theta) = \mathbb{E}\left[-\frac{\partial^2}{\partial\theta^2}\ell(\theta)\right]$. SE: $\mathrm{SE}(\hat{\theta}) \approx \sqrt{I(\hat{\theta})^{-1}/n}$.

Delta method

If $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma)$, then $\sqrt{n}(g(\hat{\theta}) - g(\theta)) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \nabla g\, \Sigma\, \nabla g^\top)$.



Likelihood.



Delta method.

Score equations and Newton–Raphson

Score

The score is $U(\theta) = \partial\ell(\theta)/\partial\theta$. MLEs solve the **score equations** $U(\hat{\theta}) = 0$.

Newton–Raphson (scalar)

Initialize $\theta^{(0)}$. Iterate $\theta^{(t+1)} = \theta^{(t)} - \dfrac{U(\theta^{(t)})}{U'(\theta^{(t)})}$. In multiple dimensions, replace derivatives with gradient and Hessian.

Asymptotic normality (proof sketch)

Taylor expand the score around θ_0: $0 = U(\hat{\theta}) \approx U(\theta_0) + U'(\theta_0)(\hat{\theta} - \theta_0)$. Since $U(\theta_0) = \mathcal{O}_p(\sqrt{n})$ and $-n^{-1}U'(\theta_0) \to I(\theta_0)$, it follows that $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, I^{-1})$.

Exponential families

Form

$f(x \mid \eta) = h(x)\exp\{\eta^\top T(x) - A(\eta)\}$, with natural parameter η, sufficient statistic T, log-partition A.

- $\ell(\eta) = \sum_i \eta^\top T(x_i) - nA(\eta) + \text{const}$, concave in η. MLE solves $\nabla A(\hat{\eta}) = \bar{T}$.
- Fisher information: $I(\eta) = n\,\nabla^2 A(\eta) = n\,\mathrm{Var}_\eta[T(X)]$.

Examples

Bernoulli, Poisson, Normal with known variance (for the mean), Gamma with fixed shape, etc.

Cramér–Rao lower bound (CRLB)

Bound

For any unbiased estimator $\tilde{\theta}$ of scalar θ: $\mathrm{Var}(\tilde{\theta}) \geq \dfrac{1}{n\,I(\theta)}$.

- Equality holds for **efficient** estimators; asymptotically, MLE attains the bound under regularity.
- Multivariate version: $\mathrm{Cov}(\tilde{\theta}) \succeq (nI(\theta))^{-1}$.

Implication

The information $I(\theta)$ sets a fundamental precision limit; designs that increase I (e.g., larger sample sizes, informative data) reduce variance.

Outline

- **Bias–Variance–MSE**: MSE $=$ Bias^2 $+$ Variance.
- **Likelihood profiles**: visualize curvature near $\hat{\theta}$ to gauge uncertainty.
- **Parametric bootstrap**: resample from fitted model; approximate SE and CIs.
- Regularization and model selection are covered later in the course.

Parametric bootstrap (algorithm)

Fit $\hat{\theta}$. For $b = 1, \ldots, B$: simulate $x^{(b)} \sim f(\cdot \mid \hat{\theta})$; compute $\hat{\theta}^{(b)}$. Use the empirical sd of $\{\hat{\theta}^{(b)}\}$ as $\widehat{\mathrm{SE}}$; percentiles for CIs.

Outline

Example: Exponential(λ)

- MLE: $\hat{\lambda} = 1/\bar{x}$; asymptotic SE: $\mathrm{SE}(\hat{\lambda}) \approx \hat{\lambda}/\sqrt{n}$.
- Compare with MoM: solve $\mathbb{E}[X] = 1/\lambda \Rightarrow \tilde{\lambda} = 1/\bar{x}$ (same here).

Derivation

$\ell(\lambda) = n \log \lambda - \lambda \sum x_i \Rightarrow \partial\ell/\partial\lambda = \frac{n}{\lambda} - \sum x_i = 0 \Rightarrow \hat{\lambda} = 1/\bar{x}$. Info: $I(\lambda) = n/\lambda^2$.

Joint MLEs: $\hat{\mu} = \bar{x}$, $\widehat{\sigma^2} = \frac{1}{n} \sum (x_i - \bar{x})^2$ (biased). Unbiased variance uses $\frac{1}{n-1}$.

- Asymptotic var: $\mathrm{Var}(\hat{\mu}) \approx \sigma^2/n$; $\mathrm{Var}(\widehat{\sigma^2})$ from information or delta method.

Derivation sketch

$\ell(\mu, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum (x_i - \mu)^2$. Partial derivatives yield the stated MLEs.

Outline

Short Exercises

1. Binomial(n, p), known n: show MLE $\hat{p} = \bar{x}/n$ equals MoM.
2. Uniform(0,θ): derive MLE and discuss the bias of the maximum statistic.
3. Delta method: if $\hat{\theta} \sim \mathcal{N}(\theta, \sigma^2/n)$, derive Var$(\log \hat{\theta})$ to first order.

Hint for 2

Order statistic $X_{(n)} = \max X_i$ has CDF $(x/\theta)^n$ on $[0, \theta]$; compute $\mathbb{E}[X_{(n)}]$ to analyze bias.

Practical Preview

We'll simulate to compare MLE and MoM and use bootstrap for SEs.

- Keep code deterministic (set a RNG seed).
- Use `numpy`, `scipy`, and `pandas` as needed. See course style guide.

Figure generation

Likelihood profile and delta-method visuals are generated by `figures/make_figures.py` (Python).

Summary

- MLE and MoM provide complementary routes to estimation.
- Fisher information and the delta method quantify uncertainty.
- Simulation and bootstrap help validate asymptotics.