Lecture 2a: Point Estimation

Method of Moments and Maximum Likelihood

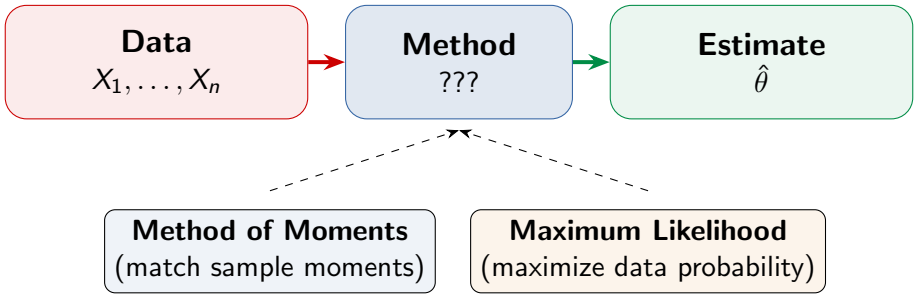
The Estimation Problem

A factory produces lightbulbs. You test 50 and find a mean lifetime of 1,200 hours.

What can you say about the **true** mean lifetime?

And how confident should you be in your *method*?

From Data to Parameters



Method of Moments (MoM)

Idea: Set population moments equal to sample moments, then solve for the parameters.

$$\begin{aligned} \mathbb{E}[X] &= g_1(\theta) \\ \mathbb{E}[X^2] &= g_2(\theta) \\ &\vdots \end{aligned}$$

$\longrightarrow$ replace with

$$\begin{aligned} \bar{X} &= g_1(\hat{\theta}) \\ \tfrac{1}{n}\textstyle\sum X_i^2 &= g_2(\hat{\theta}) \\ &\vdots \end{aligned}$$

Pros:

- Simple, quick to compute
- No distributional assumption needed for computation

Cons:

- Can give impossible values (e.g., $\hat{\sigma}^2 < 0$)
- Generally less efficient than MLE
- Awkward with many parameters

MoM Example: Normal Distribution

Model: $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$. Two unknowns, need two equations.

1st moment: $\mathbb{E}[X] = \mu \quad \Rightarrow \quad \hat{\mu}_{\text{MoM}} = \bar{X}$

2nd moment: $\mathbb{E}[X^2] = \mu^2 + \sigma^2 \quad \Rightarrow \quad \hat{\sigma}^2_{\text{MoM}} = \frac{1}{n}\sum X_i^2 - \bar{X}^2 = \frac{1}{n}\sum (X_i - \bar{X})^2$

(Note: this divides by n, not $n-1$ — the MoM estimator is slightly biased.)

The Likelihood Function

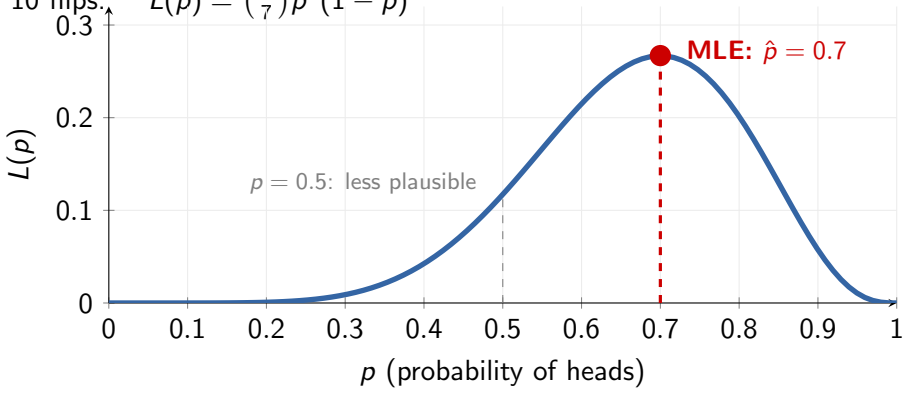
Given the data I observed, how plausible is each parameter value?

$$L(\theta) = P(\text{data} \mid \theta) = \prod_{i=1}^{n} f(X_i \mid \theta)$$

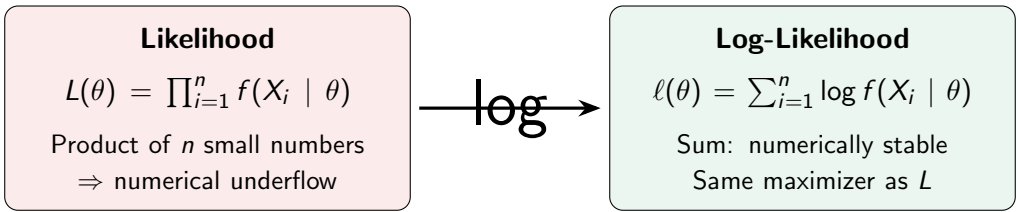
Same formula as the joint density, but now
data is fixed, θ varies (not the other way around!)

Likelihood: Coin Flip Example

Data: 7 heads in 10 flips. $L(p) = \binom{10}{7} p^7 (1-p)^3$



Log-Likelihood: Why We Prefer It



Score function: $s(\theta) = \frac{\partial \ell(\theta)}{\partial \theta}$ — the gradient of the log-likelihood.

At the MLE: $s(\hat{\theta}) = 0$ (first-order condition).

Maximum Likelihood Estimation

$$\hat{\theta}_{\text{MLE}} = \arg\max_\theta \ \ell(\theta) = \arg\max_\theta \sum_{i=1}^{n} \log f(X_i \mid \theta)$$

"Choose the parameter value that makes the observed data **most probable.**"

Recipe:

1. Write down the log-likelihood $\ell(\theta)$
2. Take the derivative and set $\frac{\partial \ell}{\partial \theta} = 0$
3. Solve for $\hat{\theta}$
4. Check second-order condition ($\frac{\partial^2 \ell}{\partial \theta^2} < 0$)

MLE: Bernoulli (Coin Fairness)

Model: $X_i \sim \text{Bernoulli}(p)$, observe k successes in n trials.

$\ell(p) = k \log p + (n-k) \log(1-p)$

$\frac{\partial \ell}{\partial p} = \frac{k}{p} - \frac{n-k}{1-p} = 0$

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

The sample proportion — exactly what you'd guess intuitively.

MLE: Normal (Measurement Error)

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

$\ell(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2$

Setting $\frac{\partial \ell}{\partial \mu} = 0$: $\boxed{\hat{\mu}_{\text{MLE}} = \bar{X}}$

Setting $\frac{\partial \ell}{\partial \sigma^2} = 0$: $\boxed{\hat{\sigma}^2_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$

Note: divides by n, not $n-1$ — the MLE for σ^2 is **biased** (slightly too small).

MLE: Poisson (Rare Events)

Model: $X_i \sim \text{Poisson}(\lambda)$ (goals/match, earthquakes/year, typos/page)

$\ell(\lambda) = (\sum_{i=1}^{n} X_i) \log \lambda - n\lambda - \sum_{i=1}^{n} \log(X_i!)$

$\frac{\partial \ell}{\partial \lambda} = \frac{\sum X_i}{\lambda} - n = 0$

$$\boxed{\hat{\lambda}_{\text{MLE}} = \bar{X}}$$

Again the sample mean — but now it estimates the *rate*, not just a mean.

MLE: Exponential (Waiting Times)

Model: $X_i \sim \text{Exp}(\lambda)$ (time between arrivals, device lifetimes)

$f(x \mid \lambda) = \lambda e^{-\lambda x}$ for $x \geq 0$

$\ell(\lambda) = n \log \lambda - \lambda \sum_{i=1}^{n} X_i$

$\frac{\partial \ell}{\partial \lambda} = \frac{n}{\lambda} - \sum X_i = 0$

$$\boxed{\hat{\lambda}_{\text{MLE}} = \frac{1}{\bar{X}}}$$

The reciprocal of the sample mean — intuitive since $\mathbb{E}[X] = 1/\lambda$.

MLE: Summary of Examples

Distribution	**Parameter**	**MLE**	**Real-world use**
Bernoulli(p)	p	$\bar{X}$	Coin fairness, conversion rates
Normal(μ, σ^2)	μ, σ^2	$\bar{X}$, $\frac{1}{n}\sum(X_i - \bar{X})^2$	Measurement error
Poisson(λ)	λ	$\bar{X}$	Count data, rare events
Exponential(λ)	λ	$1/\bar{X}$	Waiting times, lifetimes

All connect to distributions from Module 19.

Invariance Property

If $\hat{\theta}_{\text{MLE}}$ is the MLE of θ, then for any function g:

$$\widehat{g(\theta)}_{\text{MLE}} \quad = \quad g(\hat{\theta}_{\text{MLE}})$$

Example:

- MLE of λ for Exponential is $\hat{\lambda} = 1/\bar{X}$
- Want the MLE of the *mean* $\mu = 1/\lambda$? Just apply $g(\lambda) = 1/\lambda$:
- $\hat{\mu}_{\text{MLE}} = 1/\hat{\lambda} = \bar{X}$ ✓

This doesn't hold for method of moments or other estimators in general.

Identifiability

Can we even hope to recover θ?

A model is **identifiable** if different parameter values give different distributions:

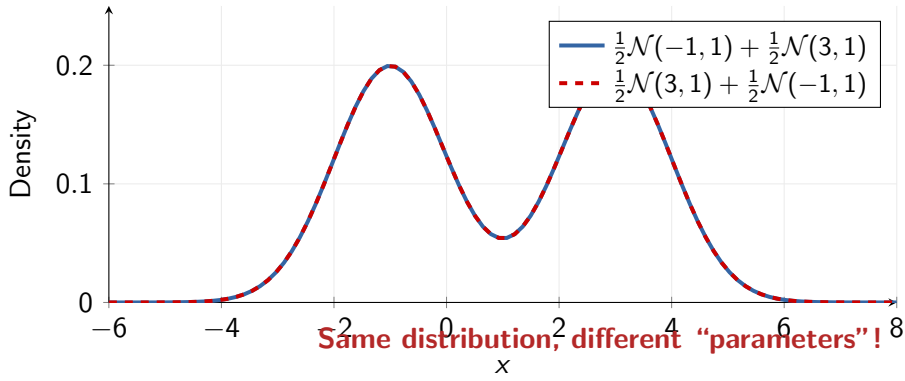
$$\theta_1 \neq \theta_2 \quad \Rightarrow \quad f(\cdot \mid \theta_1) \neq f(\cdot \mid \theta_2)$$

When it fails:

- **Mixture models:** swapping component labels gives the same distribution
- **Overparameterized models:** more parameters than the data can distinguish
- **Symmetric likelihoods:** multiple maxima, MLE is not unique

If the model isn't identifiable, no amount of data will help.

Visualizing Non-Identifiability



Practical: Implement MLE

1. Implement MLE for a Gaussian **from scratch**:
   - Write the log-likelihood function
   - Optimize numerically (scipy) and compare with the closed-form solution
2. Compare with `scipy.stats.norm.fit`
3. Fit a Poisson to real count data (e.g., goals per football match, earthquake counts per year)
4. Plot the log-likelihood surface — observe the peak at the MLE

Questions?

Next lecture: MAP, Priors, and the Bayesian Perspective