Lecture 5: Point Estimation

Method of Moments · Maximum Likelihood · Why MLE Works

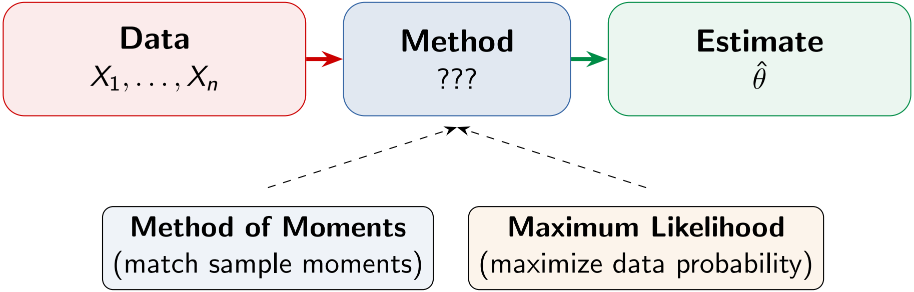
The Estimation Problem

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?

In Lectures 3–4 we learned how to **judge** estimators (bias, variance, MSE, efficiency). Today: how to **construct** them systematically.

From Data to Parameters



Method of Moments (MoM)

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

$$\begin{aligned} \mathbb{E}[X] &= g_1(\theta) \\ \mathbb{E}[X^2] &= g_2(\theta) \\ &\vdots \end{aligned} \quad \xrightarrow[\text{replace with}]{} \quad \begin{aligned} \bar{X} &= g_1(\hat{\theta}) \\ \frac{1}{n}\sum X_i^2 &= g_2(\hat{\theta}) \\ &\vdots \end{aligned}$$

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 $n-1$ — **biased!** Recall Bessel's correction from Lecture 3.)

The Likelihood Function

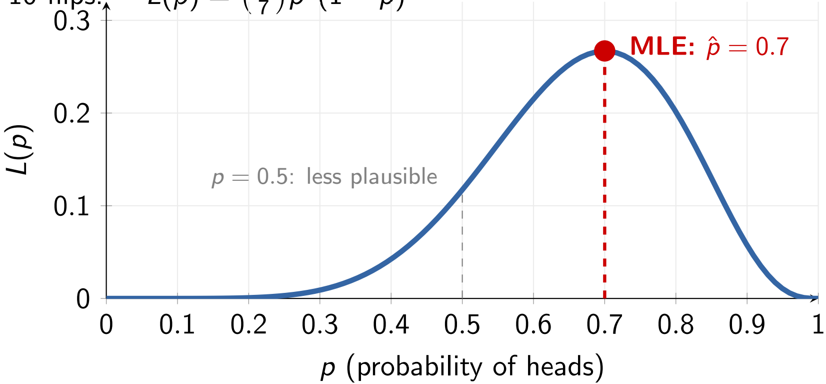
Given the data I observed, how plausible is each parameter value?

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

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

Likelihood

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

Product of n small numbers

$\Rightarrow$ numerical underflow

$\xrightarrow{\text{log}}$

Log-Likelihood

$\ell(\theta) = \sum_{i=1}^{n} \log f(X_i \mid \theta)$

Sum: numerically stable

Same maximizer as L

Score function (from Lecture 4): $s(\theta) = \frac{\partial \ell(\theta)}{\partial \theta}$ — the gradient of the log-likelihood.

At the MLE: $s(\hat{\theta}) = 0$ (first-order condition). The **curvature** of ℓ at this point $\to$ Fisher information.

Maximum Likelihood Estimation

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

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

Write $\ell(\theta)$ → Differentiate → Set $\frac{d\ell}{d\theta} = 0$ → Solve for $\hat{\theta}$ → Check $\ell'' < 0$

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** (Lecture 3: Bessel's correction).

But recall: the biased $\hat{\sigma}^2_n$ has **lower MSE** than the unbiased S^2!

MLE: Poisson (Rare Events)

Model: $X_i \sim \mathrm{Pois}(\lambda)$

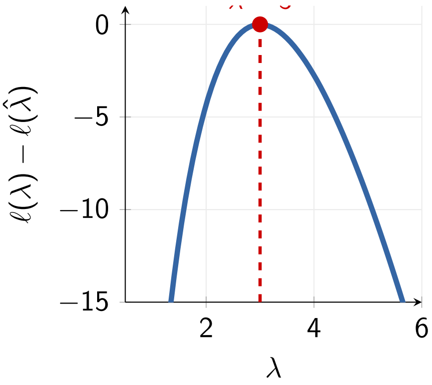
(goals/match, earthquakes/yr, typos/pg)

$\ell(\lambda) = (\sum X_i) \log \lambda - n\lambda + \mathrm{c}$

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

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

Sample mean estimates the *rate*.



Example: $n = 20$, $\sum X_i = 60$

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. Next: we'll see that these MLEs are **efficient** (hit the Cramér–Rao bound from Lecture 4).

Invariance Property

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

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

Example:

- MLE of λ for Exp is $\hat{\lambda} = 1/\bar{X}$
- Want MLE of mean $\mu = 1/\lambda$?
- Apply $g(\lambda) = 1/\lambda$: $\hat{\mu} = \bar{X}$ ✓

This doesn't hold for MoM or other estimators in general.

$$\begin{tikzcd}
\theta \arrow[r, "g"] \arrow[d, "\text{MLE}"'] \arrow[dr, phantom, "="] & g(\theta) \arrow[d, "\text{MLE}"] \\
\hat{\theta}_{\text{MLE}} \arrow[r, "g"'] & g(\hat{\theta}_{\text{MLE}})
\end{tikzcd}$$

Identifiability

Can we even hope to recover θ?

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

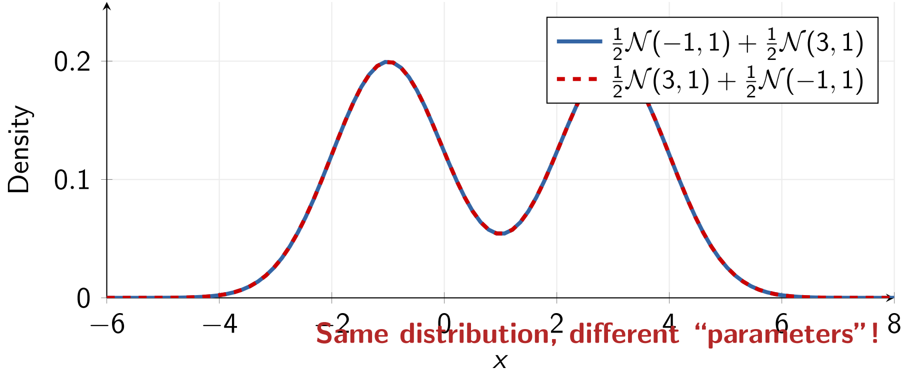
$$\theta_1 \neq \theta_2 \quad \Rightarrow \quad f(\cdot \mid \theta_1) \neq f(\cdot \mid \theta_2)$$

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



MLE and Sufficient Statistics

In Lecture 3 we learned: a **sufficient statistic** $T(\mathbf{X})$ captures everything about θ.

Key fact: The MLE depends on the data **only through** the sufficient statistic.

If $T(\mathbf{X})$ is sufficient for θ, then the MLE $\hat{\theta}$ is a function of T.

Check our examples:

Model	**Suff. stat T**	**MLE**	**Function of T?**
Bern(p)	$\sum X_i$	$\bar{X} = T/n$	✓
$N(\mu, \sigma_0^2)$	$\sum X_i$	$\bar{X} = T/n$	✓
Pois(λ)	$\sum X_i$	$\bar{X} = T/n$	✓
Exp(λ)	$\sum X_i$	$1/\bar{X} = n/T$	✓

No coincidence — MLE **always** uses sufficient statistics. No information is wasted.

MLE in Exponential Families

Recall the **exponential family** form from Lecture 3: $f(x \mid \theta) = h(x)\exp\big(\eta(\theta)\,T(x) - A(\theta)\big)$

For n i.i.d. observations, the log-likelihood is:

$$\ell(\theta) = \eta(\theta)\sum_{i=1}^{n} T(X_i) - nA(\theta) + \text{const}$$

Setting the derivative to zero:

$$\eta'(\theta)\sum_{i=1}^{n} T(X_i) = nA'(\theta)$$

For natural exponential families ($\eta = \theta$):

$A'(\hat{\theta}_{\text{MLE}}) = \frac{1}{n}\sum_{i=1}^{n} T(X_i)$ (match population mean to sample mean)

The MLE is **always** a function of the sufficient statistic $\sum T(X_i)$, and it **equals the MoM estimator** in natural form!

Why MLE Works: The Big Theoretical Guarantees

Under regularity conditions (Lecture 4), MLE has remarkable properties:

1. Consistent: $\hat{\theta}_{\text{MLE}} \xrightarrow{P} \theta_0$ as $n \to \infty$ (gets the right answer eventually)

2. Asymptotically Normal: $\sqrt{n}(\hat{\theta}_{\text{MLE}} - \theta_0) \xrightarrow{d} N\left(0, \frac{1}{I(\theta_0)}\right)$

3. Asymptotically Efficient: achieves the **Cramér–Rao bound** as $n \to \infty$

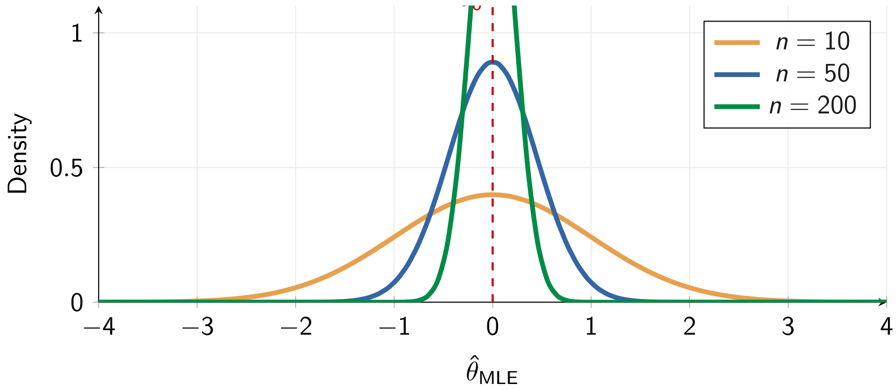
4. Invariant: MLE of $g(\theta)$ is $g(\hat{\theta}_{\text{MLE}})$ for any function g

Translation: With enough data, MLE is approximately unbiased, approximately normal, and **no other estimator can do better**.

This is why MLE is the default method in statistics and machine learning.

Asymptotic Normality: Seeing It

As n grows, the sampling distribution of the MLE converges to a Normal centered at the truth:



Variance shrinks as $\frac{1}{n\,I(\theta_0)}$: more data $\Rightarrow$ tighter bell $\Rightarrow$ more precise estimate.
With $n = 200$ observations, MLE is practically pinpointed at the truth.

MLE Achieves the Cramér–Rao Bound

From Lecture 4, the **CR bound**: $\mathrm{Var}(\hat{\theta}) \geq \frac{1}{n\,I(\theta)}$ for unbiased estimators.

Does MLE hit this bound?

Model	**MLE**	$\mathrm{Var}(\hat{\theta}_{\mathrm{MLE}})$	**CR bound**	**Efficient?**
$\mathrm{Bern}(p)$	$\bar{X}$	$\frac{p(1-p)}{n}$	$\frac{p(1-p)}{n}$	**Yes**
$N(\mu, \sigma_0^2)$	$\bar{X}$	$\frac{\sigma_0^2}{n}$	$\frac{\sigma_0^2}{n}$	**Yes**
$\mathrm{Pois}(\lambda)$	$\bar{X}$	$\frac{\lambda}{n}$	$\frac{\lambda}{n}$	**Yes**

For **exponential families**, the MLE of the natural parameter is efficient (hits the CR bound exactly). For other models, MLE is **asymptotically** efficient — it approaches the bound as $n \to \infty$.

Practical: Implement MLE

1. Implement MLE for a Gaussian **from scratch**:
   - Write the log-likelihood function
   - Optimize numerically (`scipy.optimize`) and compare with closed form
2. Compare $\hat{\sigma}^2_{\text{MLE}}$ (divides by n) with S^2 (divides by $n-1$). Verify the bias from Lecture 3 empirically with simulation.
3. Fit a Poisson to real count data. Check: is the MLE efficient? Compute the CR bound and compare with the observed variance.
4. Plot the log-likelihood surface — observe the peak at the MLE and relate its **curvature** to Fisher information.

Homework

1. Derive the MLE for Geometric(p): $f(x \mid p) = (1-p)^{x-1}p, \;\; x = 1, 2, \ldots$
   Is this MLE unbiased? Is it efficient (check against the CR bound)?
2. For $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, show that $\hat{\sigma}^2_{\text{MLE}} = \frac{1}{n}\sum(X_i - \bar{X})^2$
   equals the MoM estimator. Why is this not a coincidence? (Hint: exponential family.)
3. Show that the MLE for Uniform$(0, \theta)$ is $\hat{\theta} = X_{(n)} = \max(X_1, \ldots, X_n)$.
   Is this unbiased? Is it consistent? (Note: this is **not** an exponential family!)
4. Simulate $n = 50$ samples from Poisson$(\lambda = 3)$ and compute the MLE.
   Repeat 10,000 times. Verify: (a) $\hat{\lambda}$ is approximately unbiased, (b) $\text{Var}(\hat{\lambda}) \approx \lambda/n$.

Questions?

Next lecture: MAP estimation, priors, and the Bayesian perspective