

# **Likelihood and MLE**

Gov 2001: Quantitative Social Science Methods I

Scott Cunningham

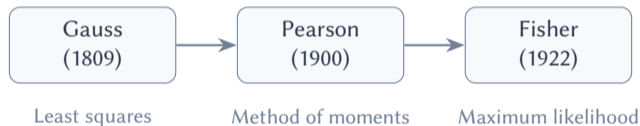
Harvard University

Spring 2026

## Today's reading

- A&M §5.3: Maximum likelihood estimation
- Blackwell Ch. 2: Estimation framework (review)

## Three ideas shaped how we learn from data



Each solved a problem the previous one could not.

Today we arrive at Fisher's answer—but to understand it, we need to see what came before.

## Pearson's method of moments: match the population to the sample

**The idea** (1894/1900): The normal distribution has two parameters,  $\mu$  and  $\sigma^2$ :

$$f(x | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

The population moments are functions of  $\mu$  and  $\sigma^2$ . Set them equal to the sample moments:

$$\underbrace{E[X]}_{\mu} = \bar{X} \quad \text{and} \quad \underbrace{E[X^2]}_{\mu^2 + \sigma^2} = \frac{1}{n} \sum X_i^2$$

Two equations, two unknowns  $\Rightarrow$  solve for  $\hat{\mu}$  and  $\hat{\sigma}^2$ .

## Method of moments is algebra, not optimization

It might *sound* like you're fixing some data-informed numbers and then maximizing an objective function.

**That is not what is happening.**

- There is **no objective function**—no likelihood, no loss, no Lagrangian
- You just set  $k$  population moments equal to  $k$  sample moments and **solve the system**
- This is the **plug-in principle** from last week—Pearson formalized it

Still widely used: GMM in econometrics is a direct descendant. When you have more moment conditions than parameters, *then* you optimize.

## Fisher asked a different question: which parameter makes the data most probable?

**The idea** (1912/1922): Don't just match moments—use *all* the information in the data.

- Pearson: “What parameter gives these sample averages?”
- Fisher: “What parameter would have made *this exact dataset* most likely to occur?”

Fisher showed his method was:

- **Lower variance** than method of moments
- As  $n$  grows, no estimator can do better—we will make this precise on Wednesday

This launched a bitter feud with Pearson's son, Egon. The rivalry shaped 20th-century statistics.

## Least squares targets the conditional mean; MLE targets the full distribution

OLS consistently estimates the BLP—and therefore the CEF—**regardless of the distribution of  $Y$** . Binary, count, continuous: OLS always gives you a conditional mean.

**So why would you ever want something else?**

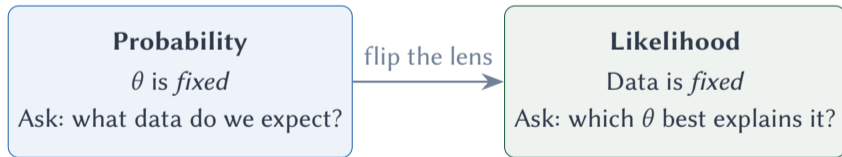
- OLS gives  $E[Y | X]$ . MLE gives the full  $f(Y | X, \theta)$ .
- From the full distribution you get probabilities, quantiles, variances—not just the mean.
- A parametric model can also respect natural constraints (e.g., predicted probabilities in  $[0, 1]$ ).

Under Normal errors, the OLS estimator *is* the MLE. But OLS does not require normality to work—it just requires normality to be the MLE.

## Two ways to read the same equation

We write  $\theta$  for the unknown **parameter** we want to learn. In a Bernoulli model,  $\theta$  is the probability of success:

$$X \sim \text{Bernoulli}(\theta), \quad f(x | \theta) = \theta^x (1 - \theta)^{1-x}$$



Same function  $f(x | \theta)$ , different question.

## You know your sample — you don't know the population

- **What you observe:** a sample — you knock on 200 doors, 68 say they voted
- **What you want:** the population parameter  $\theta$  — the city-wide turnout rate across *all* registered voters
- **Probability** ( $\theta$  fixed): If  $\theta = 0.40$ , how many of 200 would we expect to say yes?
- **Likelihood** (data fixed): We saw 68/200 — which  $\theta$  makes that most plausible?

Next slide: we formalize this as a Bernoulli model.

## Estimating voter turnout in a local election

**Setup:** Survey  $n = 200$  registered voters. Of these,  $k = 68$  voted.

**Model:**  $X_i \sim \text{Bernoulli}(\theta)$ , i.i.d.



We will carry this example through the entire lecture.

## The likelihood function asks: how probable was my data under each $\theta$ ?

### Definition: Likelihood

$$L(\theta; \mathbf{x}) = f(\mathbf{x} | \theta) \quad \text{viewed as a function of } \theta$$

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

**Voter turnout:**

$$L(\theta) \propto \theta^{68} (1 - \theta)^{132}$$

Different values of  $\theta$  make the observed data more or less probable.

## Independence turns joint densities into products

**Recall:** If  $X_1, \dots, X_n$  are independent, the joint density factors:

$$f(x_1, \dots, x_n | \theta) = f(x_1 | \theta) f(x_2 | \theta) \cdots f(x_n | \theta) = \prod_{i=1}^n f(x_i | \theta)$$

A density  $f(x)$  can exceed 1 at a point—it is *not* a probability. What must equal 1 is the integral:  $\int f(x) dx = 1$ . The same holds for joint densities:

$$\int \cdots \int f(x_1, \dots, x_n) dx_1 \cdots dx_n = 1.$$

**Notation:** We write  $L(\theta)$  for the *likelihood function*—the same object as the joint density, but viewed as a function of  $\theta$ :

$$L(\theta; x_1, \dots, x_n) = f(x_1, \dots, x_n | \theta)$$

## For i.i.d. data, the likelihood is a product

Combining independence with the likelihood notation:

$$L(\theta; x_1, \dots, x_n) = \prod_{i=1}^n f(x_i | \theta)$$

**Voter turnout:** Each voter's response is independent, so

$$L(\theta) = \prod_{i=1}^{200} \theta^{x_i} (1 - \theta)^{1-x_i} = \theta^{68} (1 - \theta)^{132}$$

## Logarithms turn products into sums

### Log-likelihood:

$$\ell(\theta; \mathbf{x}) = \log L(\theta; \mathbf{x}) = \sum_{i=1}^n \log f(x_i | \theta)$$

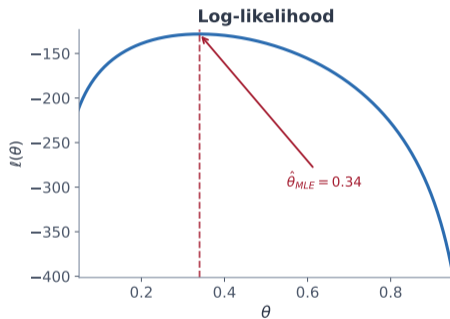
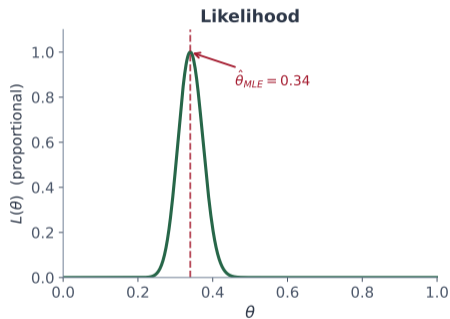
**Same maximizer:** log is monotone, so  $\arg \max_{\theta} L(\theta) = \arg \max_{\theta} \ell(\theta)$

### Voter turnout:

$$\ell(\theta) = 68 \log \theta + 132 \log(1 - \theta)$$

Multiplying 200 numbers below 1 gives  $\approx 10^{-97}$ —representable. With  $n = 2,000$  you get  $\approx 10^{-970}$ , which rounds to exactly 0 (numerical underflow). Log-sums stay manageable regardless of  $n$ .

## The likelihood peaks at $\hat{\theta} = 0.34$



$n = 200$  voters,  $k = 68$  voted. Both functions peak at  $\hat{\theta} = k/n = 0.34$ .

## The MLE is the $\theta$ that maximizes the likelihood

### Definition: Maximum Likelihood Estimator

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

**In practice:** take the derivative and set it to zero.

**Score equation:**

$$\frac{\partial \ell(\theta)}{\partial \theta} = 0 \quad \implies \quad \hat{\theta}_{\text{MLE}}$$

## Solving for the voter turnout MLE

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

**Score:**

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

**Solve:**  $k(1 - \theta) = (n - k)\theta \implies k = n\theta$

$$\hat{\theta}_{\text{MLE}} = \frac{k}{n} = \frac{68}{200} = 0.34$$

The MLE for a Bernoulli is the sample proportion — the plug-in estimator.

## For exponential families, MLE and plug-in give the same answer

Parameter	Plug-in	MLE
Bernoulli $\theta$	$\bar{X}$	$\bar{X}$
Normal $\mu$	$\bar{X}$	$\bar{X}$
Normal $\sigma^2$	$\frac{1}{n} \sum (X_i - \bar{X})^2$	$\frac{1}{n} \sum (X_i - \bar{X})^2$
Poisson $\lambda$	$\bar{X}$	$\bar{X}$

This is not a coincidence—it holds for all exponential families.

So why bother with MLE if plug-in gives the same answer?

## But they can diverge—even for the Normal

**Estimand:** the **median** of  $X \sim N(\mu, \sigma^2)$ . Since median =  $\mu$ :

	Estimator	Why?
Plug-in	sample median	Replace $F$ with $\hat{F}_n$ , take median
MLE	$\bar{X}$	MLE of $\mu$ is $\bar{X}$ ; invariance gives MLE of median = $\bar{X}$

Same estimand, different estimators. Under normality,  $\bar{X}$  is **more efficient**—the sample median uses only  $2/\pi \approx 64\%$  as much information.

MLE leverages the model (median =  $\mu$ ) to get a better answer. Plug-in ignores it.

**The bigger point:** the likelihood machinery you just learned works for *any* parametric model—logit, probit, Poisson regression, mixture models—where there is no plug-in shortcut.

## The Normal MLE for variance divides by $n$ , not $n-1$

**Model:**  $X_i \sim N(\mu, \sigma^2)$ , i.i.d.

**MLE:**

$$\hat{\mu}_{\text{MLE}} = \bar{X} \qquad \hat{\sigma}_{\text{MLE}}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

- Divides by  $n$ , not  $n - 1$
- **Biased:**  $E[\hat{\sigma}_{\text{MLE}}^2] = \frac{n-1}{n} \sigma^2$
- **But the bias shrinks:**  $\frac{n-1}{n} \rightarrow 1$  as  $n$  grows

MLE does not guarantee unbiasedness. It optimizes likelihood, not bias.

## What do “regularity conditions” mean?

MLE’s nice properties require the likelihood to be **well-behaved**. In plain terms:

- **Smooth:** you can take derivatives of  $\ell(\theta)$  — no sharp edges or jumps
- **Identified:** different values of  $\theta$  produce different distributions — the data can tell parameters apart
- **Interior:** the true  $\theta_0$  is not sitting on a boundary of the parameter space
- **Informative:**  $0 < \mathcal{I}(\theta) < \infty$  — the data actually contain information about  $\theta$

When these fail, MLE can still work—but the standard results (speed of convergence, normality) may not hold. Example:  $\text{Uniform}(0, \theta)$ , where the support depends on  $\theta$ .

## Preview: MLE has remarkable properties as $n$ grows

Under regularity conditions, Fisher showed that MLE is:

- **Consistent:** the estimate converges to the truth as  $n$  grows
- **Asymptotically normal:** the sampling distribution becomes Normal for large  $n$
- **Asymptotically efficient:** no other estimator can do better (in a sense we will make precise)
- **Invariant:** MLE of  $g(\theta)$  is  $g(\hat{\theta}_{\text{MLE}})$  — this one holds at *any*  $n$

We need the Law of Large Numbers to prove these. That is Wednesday's lecture. Today we state them and use invariance.

## Invariance means you get transformations for free

### Invariance Property

If  $\hat{\theta}$  is the MLE of  $\theta$ , then  $g(\hat{\theta})$  is the MLE of  $g(\theta)$  for any function  $g$ .

#### Voter turnout example:

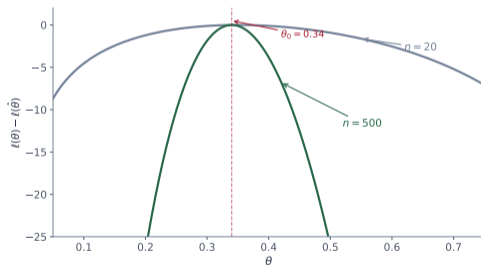
- MLE of turnout rate:  $\hat{\theta} = 0.34$
- MLE of turnout *odds*:  $\frac{\hat{\theta}}{1 - \hat{\theta}} = \frac{0.34}{0.66} = 0.515$
- MLE of log-odds:  $\log\left(\frac{0.34}{0.66}\right) = -0.663$

No need to re-derive — just transform.

## Fisher information measures how much the data tell you about $\theta$

### Definition: Fisher Information

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



Sharp peak = high information. Flat curve = low information.

## More data makes the likelihood more informative

- **Flat peak** ( $n = 20$ ): many values of  $\theta$  are roughly equally plausible — the data cannot tell  $\theta = 0.30$  from  $\theta = 0.40$
- **Sharp peak** ( $n = 500$ ): the data strongly discriminate between nearby values of  $\theta$  — much more precise
- Information grows linearly with  $n$ : total information =  $n \cdot \mathcal{I}(\theta)$ , where  $\mathcal{I}(\theta)$  is the per-observation information
- The second derivative measures curvature at the peak — the negative sign makes sharp curvature  $\Rightarrow$  large positive number

$\mathcal{I}(\theta)$  is a population quantity — a function of the true  $\theta$ , not of one dataset. You can estimate it by plugging in  $\hat{\theta}$ .

## The Cramér-Rao bound sets a floor on estimator variance

### Cramér-Rao Lower Bound

For any **unbiased** estimator  $\hat{\theta}$ :

$$\text{Var}(\hat{\theta}) \geq \frac{1}{n\mathcal{I}(\theta)}$$

- This is a **finite-sample** result — it holds for any  $n$
- No unbiased estimator can beat this floor — but a *biased* estimator can (bias-variance tradeoff)
- More observations  $\Rightarrow$  larger  $n\mathcal{I}(\theta)$   $\Rightarrow$  tighter floor  $\Rightarrow$  more precise estimates

Wednesday we will show that MLE achieves this bound as  $n$  grows. That is what “asymptotic efficiency” means.

## When the model is wrong, MLE still finds something useful

What if  $f(x | \theta)$  is not the true distribution?

**Result:** MLE finds the parameter that makes your model *as close as possible* to the truth (in a precise sense called Kullback-Leibler divergence):

$$\theta^* = \arg \max_{\theta} E_{\text{true}}[\log f(X | \theta)]$$

- $\theta^*$  is the “closest” parameter in the model family
- With more data, MLE hones in on  $\theta^*$ —not the “true”  $\theta_0$ , but the best your model can do

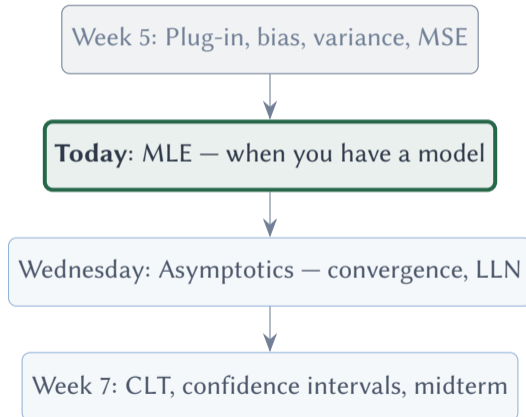
A&M §5.2.3: “All models are wrong, but MLE finds the least wrong.”

## Plug-in vs. MLE: two philosophies of estimation

	Plug-in	MLE
<b>Parametric family?</b>	No — uses $\hat{F}_n$ directly	Yes — you commit to a family (Bernoulli, Normal, ...)
<b>Method</b>	Replace $F$ with $\hat{F}_n$	Maximize $L(\theta) = \prod f(x_i   \theta)$
<b>Large <math>n</math>?</b>	Improves (by LLN)	Improves + reaches variance floor
<b>If family is wrong?</b>	No risk — never chose one	Finds closest $\theta^*$ in the family

**Plug-in** makes no modeling assumptions. **MLE** commits to a parametric family — and extracts more from the data when that family is right.

## Where we are in the course



## This material appears on the second midterm

### What to practice:

- Write a likelihood for a given model (Bernoulli, Normal, Poisson)
- Derive the MLE by maximizing the log-likelihood
- State and interpret MLE properties (consistency, asymptotic normality, efficiency, invariance)
- Compare plug-in vs. MLE for the same estimand

Not on the first midterm. Will appear on the second midterm and the final exam.

## Wednesday: Asymptotics — convergence, inequalities, and the Law of Large Numbers

We now have two ways to construct estimators (plug-in and MLE). Both are consistent.

**But what does “consistent” actually mean, formally?**

- Convergence in probability
- Markov and Chebyshev inequalities
- Weak Law of Large Numbers (with proof)

**Reading:**

- A&M §3.2: Asymptotics
- Blackwell Ch. 3: Large-sample properties