Modern Statistics

Xiangyu Chang

April 15, 2026

Abstract

To be undated.

1 Lecture 11: Parametric Inference

In Lecture 9 we introduced the three pillars of classical inference—point estimation, confidence intervals, and hypothesis testing—and distinguished *parametric* from *non-parametric* approaches. This lecture focuses exclusively on the parametric setting: we assume the data are generated by a distribution F_θ indexed by a finite-dimensional parameter $\theta \in \Theta \subseteq \mathbb{R}^d$, and our goal is to estimate θ from the observed data. We study two principled estimation methods—the **method of moments** and **maximum likelihood estimation (MLE)**—and then develop the asymptotic theory of the MLE, including consistency (via Kullback–Leibler divergence) and the score function and Fisher information that underpin its asymptotic normality.

1.1 The Parametric Framework

In parametric inference, we assume the data come from a family of distributions $\{F_\theta : \theta \in \Theta\}$. The workflow is:

- **Population**: characterized by a distribution F_θ, where θ is the unknown parameter vector.
- **Sample**: i.i.d. observations $X_1, \ldots, X_n \sim F_\theta$.
- **Inference**: use the sample to produce an estimator $\widehat{\theta}_n$ that approximates the true θ.

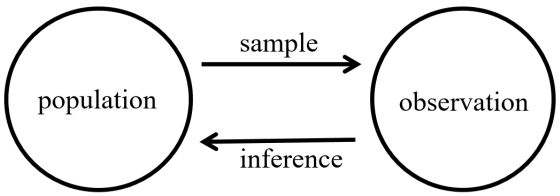

Figure 1: The parametric inference pipeline: population → sample → estimate.

The following examples illustrate the breadth of problems that fit this framework.

Example 1.1 (Bernoulli Model)**.** Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Ber}(p)$, where $p \in (0,1)$ is the unknown success probability. Here $\theta = p$ is a single scalar parameter, and the inference task is to estimate p from the observed successes and failures.

Example 1.2 (Linear Regression Model)**.** Given data pairs $\{(x_i, y_i)\}_{i=1}^n$, the simple linear model assumes

$$y_i = \theta^\top x_i + \varepsilon_i,$$

where $\theta \in \mathbb{R}^p$ is the unknown coefficient vector and ε_i are i.i.d. errors. The inference task is to estimate θ.

Example 1.3 (Large Language Models)**.** In autoregressive language models, the goal is to predict the next token given the context. The joint probability of a sequence $w_1, w_2, \ldots, w_n$ factorizes by the chain rule of probability:

$$\mathbb{P}\text{r}(w_1, w_2, \ldots, w_n) = \prod_{i=1}^n \mathbb{P}\text{r}(w_i \mid w_1, \ldots, w_{i-1}).$$

Each conditional probability is modeled by a neural network with parameter vector θ. Training the model amounts to finding the θ that best fits the observed text corpus—a parametric estimation problem at massive scale.

1.2 Method of Moments

The **method of moments** (MoM) is a simple, general estimation strategy: equate the theoretical moments of the distribution (which depend on θ) to the corresponding sample moments (which can be computed from data), then solve for θ.

Definition 1.4 (Method of Moments Estimator)**.** Let $\theta = (\theta_1, \ldots, \theta_K)^\top \in \mathbb{R}^K$. Define the k-th theoretical and sample moments:

$$\alpha_k(\theta) \overset{\text{def}}{=} \mathop{\mathbb{E}}_{\theta}[X^k] = \int x^k \, \mathrm{d}F_\theta(x), \qquad \widehat{\alpha}_k \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n X_i^k.$$

The **method of moments estimator** $\widehat{\theta}_n^{\text{MoM}}$ is obtained by solving the system

$$\alpha_k(\widehat{\theta}_n) = \widehat{\alpha}_k, \quad k = 1, 2, \ldots, K.$$

By the WLLN, $\widehat{\alpha}_k \xrightarrow{P} \alpha_k(\theta)$ for each k. Under smoothness of the map $\theta \mapsto \alpha_k(\theta)$, the MoM estimator is consistent.

Example 1.5 (MoM for $\text{Ber}(p)$)**.** The first moment is $\alpha_1(p) = \mathbb{E}[X] = p$. Setting $\alpha_1(\widehat{p}) = \widehat{\alpha}_1$:

$$\widehat{p} = \frac{1}{n}\sum_{i=1}^n X_i = \bar{X}_n.$$

The sample mean is both the MoM estimator and an unbiased estimator of p.

Example 1.6 (MoM for $N(\mu, \sigma^2)$)**.** The first two theoretical moments are

$$\alpha_1(\theta) = \mathbb{E}[X] = \mu, \qquad \alpha_2(\theta) = \mathbb{E}[X^2] = \mu^2 + \sigma^2.$$

Matching to sample moments $\widehat{\alpha}_1 = \bar{X}_n$ and $\widehat{\alpha}_2 = \frac{1}{n}\sum X_i^2$ gives

$$\widehat{\mu} = \bar{X}_n, \qquad \widehat{\sigma}^2 = \widehat{\alpha}_2 - \widehat{\alpha}_1^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

Note that $\widehat{\sigma}^2$ is *biased*: $\mathbb{E}[\widehat{\sigma}^2] = \frac{n-1}{n}\sigma^2 \neq \sigma^2$. The unbiased estimator uses denominator $n-1$ (the sample variance S_n^2 from Lecture 5).

1.3 Maximum Likelihood Estimation

The MoM is intuitive but does not always produce the most efficient estimator. **Maximum likelihood estimation** (MLE) provides a principled alternative that, under regularity conditions, achieves the smallest possible asymptotic variance among all consistent estimators.

Definition 1.7 (Likelihood and Log-Likelihood)**.** Given i.i.d. observations $X_1, \ldots, X_n$ with common density or mass function f_θ, the **likelihood function** is

$$L_n(\theta) = \prod_{i=1}^{n} f_\theta(X_i),$$

and the **log-likelihood** is

$$\ell_n(\theta) = \sum_{i=1}^{n} \log f_\theta(X_i).$$

The **maximum likelihood estimator** is

$$\widehat{\theta}_n = \operatorname*{argmax}_{\theta \in \Theta} \ \ell_n(\theta).$$

Since log is monotone, maximizing $L_n(\theta)$ is equivalent to maximizing $\ell_n(\theta)$. Working with the log-likelihood converts a product into a sum, which is algebraically easier and numerically more stable.

Remark 1.8 (General Recipe for MLE)**.**

1. *Write down the log-likelihood $\ell_n(\theta) = \sum_{i=1}^{n} \log f_\theta(X_i)$.*
2. *Take the derivative (or gradient) with respect to θ and set it to zero: $\nabla_\theta \ell_n(\theta) = 0$.*
3. *Solve for $\widehat{\theta}_n$. For convex problems this yields a closed form; in general, use gradient descent or Newton's method.*
4. *Verify that the solution is a maximum (check the second-order condition).*

1.3.1 MLE Examples

Example 1.9 (MLE for Ber(p))**.** The PMF is $f_p(x) = p^x(1-p)^{1-x}$ for $x \in \{0,1\}$.

Log-likelihood:

$$\ell_n(p) = \sum_{i=1}^{n} \big[X_i \log p + (1 - X_i)\log(1-p)\big].$$

Score equation: Setting $\frac{d}{dp}\ell_n(p) = 0$:

$$\frac{\sum_{i=1}^n X_i}{p} - \frac{\sum_{i=1}^n (1 - X_i)}{1-p} = 0.$$

Solution: Cross-multiplying and simplifying yields $\widehat{p} = \frac{1}{n}\sum_{i=1}^n X_i = \bar{X}_n$. For Ber$(p)$, the MLE and MoM estimators coincide.

Example 1.10 (MLE for $N(\mu, \sigma^2)$)**.** The PDF is $f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\big(-\frac{(x-\mu)^2}{2\sigma^2}\big)$.

Log-likelihood:

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

Score equations:

$$\frac{\partial \ell_n}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \mu) = 0 \quad \Rightarrow \quad \widehat{\mu} = \bar{X}_n,$$

$$\frac{\partial \ell_n}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^n (X_i - \mu)^2 = 0 \quad \Rightarrow \quad \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

The MLE for μ is unbiased; the MLE for σ^2 is biased (denominator n instead of $n-1$), as seen in Example **??**.

Example 1.11 (MLE for Linear Regression)**.** Assume $Y_i = \beta^\top X_i + \varepsilon_i$ with $\varepsilon_i \mid X_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$. The conditional likelihood gives

$$L_n(\beta) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(Y_i - \beta^\top X_i)^2}{2\sigma^2}\right).$$

Maximizing the log-likelihood with respect to β is equivalent to minimizing the sum of squared residuals:

$$\widehat{\beta} = \operatorname*{argmin}_\beta \sum_{i=1}^n (Y_i - \beta^\top X_i)^2 = \operatorname*{argmin}_\beta \|Y - X\beta\|^2,$$

where $Y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n \times p}$ stack the observations. This is the **ordinary least squares (OLS)** estimator, which will be derived in detail in Lecture 10.

1.4 Properties of MLE

Having seen how to compute the MLE in specific models, we now study its theoretical properties. The key results are: (i) consistency, understood through Kullback–Leibler divergence; and (ii) asymptotic normality, characterized through the score function and Fisher information.

1.4.1 Consistency via Kullback–Leibler Divergence

The Kullback–Leibler (KL) divergence measures how much one distribution differs from another. It provides the link between the MLE (maximizing the empirical log-likelihood) and consistency (convergence to the truth).

Definition 1.12 (KL Divergence)**.** The **Kullback–Leibler divergence** from F_{θ^*} to F_θ is

$$D_{\mathrm{KL}}(\theta^* \| \theta) \stackrel{\text{def}}{=} \mathop{\mathbb{E}}_{\theta^*}\left[\log \frac{f_{\theta^*}(X)}{f_\theta(X)}\right].$$

Its empirical counterpart for n observations is

$$D_n(\theta^*, \theta) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{f_{\theta^*}(X_i)}{f_\theta(X_i)} = \frac{1}{n}\ell_n(\theta^*) - \frac{1}{n}\ell_n(\theta).$$

Key properties: $D_{\mathrm{KL}}(\theta^* \| \theta) \geqslant 0$ with equality if and only if $\theta = \theta^*$; and D_{KL} is asymmetric.

The connection to MLE is immediate: maximizing $\ell_n(\theta)$ is equivalent to minimizing $D_n(\theta^*, \theta)$ (since $\ell_n(\theta^*)$ does not depend on θ). By the WLLN, $D_n(\theta^*, \theta) \xrightarrow{P} D_{\mathrm{KL}}(\theta^* \| \theta)$, and if θ^* is the unique minimizer of D_{KL}, the MLE should converge to θ^*.

Theorem 1.13 (Consistency of MLE)**.** *Under the following regularity conditions:*

A1 *(Uniform convergence)* $\sup_{\theta\in\Theta} \left|\frac{1}{n}\ell_n(\theta) - \mathbb{E}[\log f_\theta(X)]\right| \xrightarrow{P} 0.$

A2 *(Identifiability) For any* $\varepsilon > 0$, $\sup_{\|\theta-\theta^*\|\geqslant\varepsilon} \mathbb{E}[\log f_\theta(X)] < \mathbb{E}[\log f_{\theta^*}(X)]$.

the MLE satisfies $\widehat{\theta}_n \xrightarrow{P} \theta^*$ *as* $n \to \infty$.

Sketch. By the WLLN, for each fixed θ:

$$\frac{1}{n}\ell_n(\theta) \xrightarrow{P} \mathbb{E}[\log f_\theta(X)].$$

Assumption A2 ensures θ^* is the unique maximizer of the population objective $\mathbb{E}[\log f_\theta(X)]$. Assumption A1 (uniform convergence) then ensures the maximizer of the empirical objective $\frac{1}{n}\ell_n(\theta)$ converges to θ^*, giving $\widehat{\theta}_n \xrightarrow{P} \theta^*$. ■

1.4.2 Score Function and Fisher Information

The **score function** and **Fisher information** are the central objects governing the asymptotic distribution of the MLE. They measure how sensitive the log-likelihood is to the parameter θ.

Definition 1.14 (Score Function and Fisher Information)**.** Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} f_\theta$, where $\theta \in \Theta \subseteq \mathbb{R}^d$.

- The **score function** is the gradient of the log-likelihood with respect to θ:

$$S_\theta(x) \stackrel{\text{def}}{=} \nabla_\theta \log f_\theta(x) = \frac{\nabla_\theta f_\theta(x)}{f_\theta(x)}.$$

- The **Fisher information matrix** is

$$I(\theta) \stackrel{\text{def}}{=} \operatorname{Var}[S_\theta(X)] = -\mathbb{E}\left[\nabla_\theta^2 \log f_\theta(X)\right].$$

The two expressions for $I(\theta)$ (variance of score, and negative expected Hessian) are equivalent under regularity conditions. The following properties explain why.

Theorem 1.15 (Properties of Score and Fisher Information)**.** *Under standard regularity conditions (interchange of differentiation and integration):* (i)(i)

1. **Score has mean zero**: $\mathbb{E}_\theta[S_\theta(X)] = 0$.
2. **Dual representation**: $I(\theta) = \mathbb{E}[S_\theta(X) S_\theta(X)^\top] = -\mathbb{E}[\nabla_\theta^2 \log f_\theta(X)]$.
3. **Additivity**: For n i.i.d. observations, the total Fisher information is $I_n(\theta) = nI(\theta)$.

Proof. **Proof of (i).** Since $\int f_\theta(x)\, \mathrm{d}x = 1$, differentiating both sides with respect to θ and interchanging differentiation and integration:

$$\mathop{\mathbb{E}}_\theta[S_\theta(X)] = \int \nabla_\theta \log f_\theta(x) \cdot f_\theta(x)\, \mathrm{d}x = \int \nabla_\theta f_\theta(x)\, \mathrm{d}x = \nabla_\theta \int f_\theta(x)\, \mathrm{d}x = \nabla_\theta(1) = 0.$$

Proof of (ii). Starting from $S_\theta(x) = \nabla_\theta \log f_\theta(x)$, compute the Hessian:

$$\nabla_\theta^2 \log f_\theta(x) = \frac{\nabla_\theta^2 f_\theta(x)}{f_\theta(x)} - \frac{\nabla_\theta f_\theta(x)\, \nabla_\theta f_\theta(x)^\top}{f_\theta(x)^2}.$$

Taking expectations over $X \sim f_\theta$ and using $\int \nabla_\theta^2 f_\theta(x)\, \mathrm{d}x = \nabla_\theta^2(1) = 0$:

$$\mathbb{E}\left[\nabla_\theta^2 \log f_\theta(X)\right] = \int \nabla_\theta^2 f_\theta(x)\, \mathrm{d}x - \mathbb{E}\left[S_\theta(X) S_\theta(X)^\top\right] = 0 - I(\theta),$$

which gives $I(\theta) = -\mathbb{E}[\nabla_\theta^2 \log f_\theta(X)]$.

Proof of (iii). Since $\ell_n(\theta) = \sum_{i=1}^n \log f_\theta(X_i)$ is a sum of i.i.d. terms, $\operatorname{Var}[\nabla \ell_n(\theta)] = \sum_{i=1}^n \operatorname{Var}[S_\theta(X_i)] = nI(\theta)$. ■

The Fisher information $I(\theta)$ quantifies how much information a single observation carries about θ: a sharper (more peaked) likelihood surface means higher $I(\theta)$. This has a fundamental implication for inference, formalized in the **Cramér–Rao lower bound**: no unbiased estimator can have variance smaller than $[nI(\theta)]^{-1}$. The MLE achieves this bound asymptotically, making it **asymptotically efficient**.

Theorem 1.16 (Asymptotic Normality of MLE)**.** *Under regularity conditions, the MLE satisfies*

$$\sqrt{n}\,(\widehat{\theta}_n - \theta^*) \xrightarrow{d} N\Big(0,\, I(\theta^*)^{-1}\Big),$$

or equivalently, $\frac{\widehat{\theta}_n - \theta^*}{\widehat{\mathrm{se}}} \xrightarrow{d} N(0,1)$*, where* $\widehat{\mathrm{se}} = \left[n\, I(\widehat{\theta}_n)\right]^{-1/2}$.

This theorem—proved by a Taylor expansion of the score equation $\nabla \ell_n(\widehat{\theta}_n) = 0$ around θ^*, combined with the CLT and the dual representation of $I(\theta)$—justifies the standard z-based confidence intervals for MLE-derived estimates.

1.5 Cramér–Rao Lower Bound (CRLB)

The **Cramér–Rao lower bound** (CRLB) provides a fundamental limit on the variance of any unbiased estimator of a parameter. It states that the variance of any unbiased estimator cannot be smaller than the reciprocal of the Fisher information, thus establishing a benchmark for estimator performance.

Theorem 1.17 (Cramér–Rao Lower Bound)**.** *Let $X_1, \ldots, X_n$ be i.i.d. samples from a parametric family f_θ, and let $\widehat{\theta}_n$ be any unbiased estimator of θ. Under standard regularity conditions,*

$$\operatorname{Var}(\widehat{\theta}_n) \geqslant \frac{1}{nI(\theta)},$$

where $I(\theta)$ is the Fisher information for a single observation.

The CRLB quantifies the minimal achievable variance for unbiased estimators. Estimators that attain the CRLB are said to be *efficient*.

The CRLB is a cornerstone of statistical theory: it tells us how close an unbiased estimator can be to the true value in terms of its variance. Notably, the MLE is asymptotically efficient, achieving the CRLB in large samples under regularity conditions.

1.6 Summary and Outlook

This lecture developed two estimation methods for parametric models:

1. **Method of Moments**: match K sample moments to theoretical moments and solve for θ. Simple to apply, consistent by the WLLN, but may not be efficient.

2. **MLE**: maximize $\ell_n(\theta) = \sum \log f_\theta(X_i)$. Consistent under A1–A2 (via KL divergence), and asymptotically normal with variance $[nI(\theta^*)]^{-1}$, achieving the Cramér–Rao lower bound.

The MLE's asymptotic normality $\frac{\widehat{\theta}_n - \theta^*}{\widehat{\mathrm{se}}} \xrightarrow{d} N(0,1)$ immediately yields confidence intervals $\widehat{\theta}_n \pm z_{\alpha/2}\widehat{\mathrm{se}}$ and hypothesis tests, closing the loop with the inference framework of Lecture 8.

What comes next. Lecture 10 applies these ideas to **simple linear regression** (SLR). The OLS estimator (Example **??**) turns out to be the MLE under Gaussian errors, and we will derive its exact distribution, confidence intervals for the slope and intercept, and prediction intervals—all building directly on the variance formulas and CLT results developed in this lecture.