Modern Statistics

Xiangyu Chang

April 7, 2026

Abstract

To be undated.

1 Lecture 9: Statistical Inference

Lectures 1–8 developed the probabilistic machinery we need: probability spaces, random variables, distributions, expectation, and the key limit theorems (WLLN, CLT, Slutsky, Delta Method). We now use these tools to do **statistics**. The central question changes from "given a distribution, what are the properties of random variables drawn from it?" to "given observed data, what can we infer about the distribution that generated it?" This lecture introduces the three pillars of classical inference: **point estimation**, **confidence sets**, and **hypothesis testing**.

1.1 What Is Statistical Inference?

Statistical inference is the process of using observed data $\{Z_i\}_{i=1}^n$ to learn about the distribution F that generated the data. The data are treated as realizations of random variables; the goal is to draw conclusions about the underlying population. Inference problems appear in almost every quantitative field.

A central organizing distinction is between **parametric** and **non-parametric** inference.

- **Parametric inference**: We assume the data come from a family of distributions $\{F_\theta : \theta \in \Theta\}$ indexed by a finite-dimensional parameter $\theta \in \mathbb{R}^d$. The inference problem reduces to estimating θ.
- **Non-parametric inference**: No such parametric form is assumed; the goal is to estimate the entire distribution F or a functional of it (e.g., a regression function $r(\boldsymbol{X}) = \mathbb{E}[Y \mid \boldsymbol{X}]$).

The examples below illustrate both approaches. Lectures 9 onward focus on parametric inference, culminating in the theory of maximum likelihood estimation.

Example 1.1 (Estimating a Population Mean)**.** Suppose a population X has unknown mean μ. Given a random sample $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} F$, we estimate μ by the sample mean

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

By the WLLN (Lecture 7), $\bar{X}_n \xrightarrow{P} \mu$, justifying this choice.

Example 1.2 (Estimating Parameters of a Normal Distribution)**.** Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$. Both parameters are unknown. The maximum likelihood estimators are

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

Example 1.3 (Linear Regression)**.** Consider i.i.d. data pairs $(\boldsymbol{X}_1, Y_1), \dots, (\boldsymbol{X}_n, Y_n)$, where $\boldsymbol{X}_i \in \mathbb{R}^p$ and $Y_i \in \mathbb{R}$. Suppose

$$Y_i = \boldsymbol{X}_i^\top \boldsymbol{\beta} + \varepsilon_i,$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ is an unknown coefficient vector and ε_i are i.i.d. errors with mean zero. The inference problem reduces to estimating $\boldsymbol{\beta}$: a finite-dimensional parameter in place of the infinite-dimensional function $r(\boldsymbol{X}) = \mathbb{E}[Y \mid \boldsymbol{X}]$.

Example 1.4 (k-Nearest Neighbors Regression)**.** In the same setup, one can also estimate $r(\boldsymbol{X}) = \mathbb{E}[Y \mid \boldsymbol{X}]$ non-parametrically. The k-nearest neighbors (kNN) estimator at a new point $\boldsymbol{X}^*$ finds the k training points closest to $\boldsymbol{X}^*$ in Euclidean distance and averages their responses:

$$\widehat{r}(\boldsymbol{X}^*) = \frac{1}{k}\sum_{i=1}^{k} Y_{(i)}.$$

This requires no parametric assumption on r, at the cost of higher variance when p is large.

1.2 Point Estimation

A **point estimator** $\widehat{\theta}_n$ summarizes the data by a single value meant to approximate an unknown parameter θ.

Definition 1.5 (Point Estimator)**.** Let $X_1, \dots, X_n$ be i.i.d. from some distribution F. A **point estimator** of a parameter $\theta = \theta(F)$ is any function

$$\widehat{\theta}_n = g(X_1, \dots, X_n).$$

Since $\widehat{\theta}_n$ is a function of random variables, it is itself a random variable with its own distribution, expectation, and variance.

1.2.1 Performance Metrics

How do we judge whether an estimator is good? Three criteria—bias, variance, and consistency—capture different aspects of estimator quality.

Definition 1.6 (Bias and Unbiasedness)**.** The **bias** of $\widehat{\theta}_n$ is

$$\text{Bias}(\widehat{\theta}_n) \overset{\text{def}}{=} \mathbb{E}[\widehat{\theta}_n] - \theta.$$

We say $\widehat{\theta}_n$ is **unbiased** if $\mathbb{E}[\widehat{\theta}_n] = \theta$, i.e., $\text{Bias}(\widehat{\theta}_n) = 0$.

Definition 1.7 (Mean Squared Error)**.** The **mean squared error (MSE)** of $\widehat{\theta}_n$ is

$$\text{MSE}(\widehat{\theta}_n) \overset{\text{def}}{=} \mathbb{E}\left[(\widehat{\theta}_n - \theta)^2\right].$$

The MSE decomposes into bias and variance, making explicit the trade-off between them:

Theorem 1.8 (Bias-Variance Decomposition)**.**

$$\mathrm{MSE}(\widehat{\theta}_n) = \mathrm{Bias}^2(\widehat{\theta}_n) + \mathrm{Var}(\widehat{\theta}_n).$$

Proof. Let $\bar{\theta} = \mathbb{E}[\widehat{\theta}_n]$. Then:

$$\begin{aligned}
\mathrm{MSE}(\widehat{\theta}_n) &= \mathbb{E}\left[(\widehat{\theta}_n - \theta)^2\right] \\
&= \mathbb{E}\left[(\widehat{\theta}_n - \bar{\theta} + \bar{\theta} - \theta)^2\right] \\
&= \mathbb{E}\left[(\widehat{\theta}_n - \bar{\theta})^2\right] + (\bar{\theta} - \theta)^2 + 2(\bar{\theta} - \theta)\ \mathbb{E}[\widehat{\theta}_n - \bar{\theta}] \\
&= \mathrm{Var}(\widehat{\theta}_n) + \mathrm{Bias}^2(\widehat{\theta}_n) + 0,
\end{aligned}$$

where the cross term vanishes because $\mathbb{E}[\widehat{\theta}_n - \bar{\theta}] = \bar{\theta} - \bar{\theta} = 0$. ■

Definition 1.9 (Standard Error)**.** The **standard error** of $\widehat{\theta}_n$ is

$$\mathrm{se}(\widehat{\theta}_n) = \sqrt{\mathrm{Var}(\widehat{\theta}_n)}.$$

Definition 1.10 (Consistency)**.** An estimator $\widehat{\theta}_n$ is **consistent** for θ if $\widehat{\theta}_n \xrightarrow{P} \theta$ as $n \to \infty$.

Bias and variance going to zero is a sufficient condition for consistency, via the MSE:

Theorem 1.11 (Sufficient Condition for Consistency)**.** *If* $\mathrm{Bias}(\widehat{\theta}_n) \to 0$ *and* $\mathrm{Var}(\widehat{\theta}_n) \to 0$ *as* $n \to \infty$*, then* $\widehat{\theta}_n \xrightarrow{P} \theta$*.*

Proof. By Markov's inequality applied to $(\widehat{\theta}_n - \theta)^2$:

$$\Pr(|\widehat{\theta}_n - \theta| > \varepsilon) \leqslant \frac{\mathbb{E}[(\widehat{\theta}_n - \theta)^2]}{\varepsilon^2} = \frac{\mathrm{Bias}^2(\widehat{\theta}_n) + \mathrm{Var}(\widehat{\theta}_n)}{\varepsilon^2} \to 0.$$

■

Example 1.12 (Sample Mean as an Estimator of μ)**.** Let $\widehat{\mu}_n = \bar{X}_n$ estimate $\mu = \mathbb{E}[X]$. From Lecture 5:

$$\mathbb{E}[\widehat{\mu}_n] = \mu \quad \text{(unbiased)}, \qquad \mathrm{Var}(\widehat{\mu}_n) = \frac{\sigma^2}{n} \to 0.$$

By Theorem 1.11, $\widehat{\mu}_n$ is consistent. The standard error is $\mathrm{se}(\widehat{\mu}_n) = \sigma/\sqrt{n}$.

To build a confidence interval without the CLT, Chebyshev's inequality gives, for any $\varepsilon > 0$:

$$\Pr(|\widehat{\mu}_n - \mu| > \varepsilon) \leqslant \frac{\mathrm{Var}(\widehat{\mu}_n)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2}.$$

Setting $\alpha = \sigma^2/(n\varepsilon^2)$ and solving for $\varepsilon = \sigma/\sqrt{n\alpha}$, a valid (but conservative) $1 - \alpha$ confidence interval is

$$\left[\widehat{\mu}_n - \frac{\sigma}{\sqrt{n\alpha}},\ \widehat{\mu}_n + \frac{\sigma}{\sqrt{n\alpha}}\right].$$

This interval is distribution-free but wider than the CLT-based interval, which we derive next.

1.3 Confidence Sets

A point estimate gives our best single guess for θ, but provides no information about uncertainty. A **confidence set** quantifies the precision of the estimate by constructing a random set that contains θ with guaranteed probability.

Definition 1.13 (Confidence Interval)**.** A $1-\alpha$ **confidence interval** for θ is a random interval $C_n = (a, b)$, where $a = a(X_1, \ldots, X_n)$ and $b = b(X_1, \ldots, X_n)$ are data-dependent endpoints, such that

$$\inf_{\theta} \mathbb{P}\mathrm{r}(\theta \in C_n) \geqslant 1-\alpha \quad \text{for all } \theta \in \Theta.$$

The probability $1-\alpha$ is called the **coverage** of the interval.

Remark 1.14 (Common Misconception)**.** *C_n is random and θ is fixed. A 95% confidence interval does* not *mean "there is a 95% probability that θ lies in this specific interval." Rather, if we were to repeat the experiment many times, 95% of the realized intervals would contain the true θ.*

The Chebyshev-based interval in Example 1.12 is valid for any distribution but is conservative. The CLT provides a sharper interval by leveraging asymptotic normality.

Theorem 1.15 (Normal Confidence Interval)**.** *Suppose an estimator $\widehat{\theta}_n$ satisfies the asymptotic normality condition*

$$\frac{\widehat{\theta}_n - \theta}{\mathrm{se}(\widehat{\theta}_n)} \xrightarrow{d} N(0,1).$$

Then the interval

$$C_n = \left[\widehat{\theta}_n - z_{\alpha/2}\,\mathrm{se}(\widehat{\theta}_n),\ \ \widehat{\theta}_n + z_{\alpha/2}\,\mathrm{se}(\widehat{\theta}_n)\right]$$

satisfies $\mathbb{P}\mathrm{r}(\theta \in C_n) \to 1-\alpha$, where $z_{\alpha/2}$ is the $1-\alpha/2$ quantile of $N(0,1)$, i.e., $\mathbb{P}\mathrm{r}(Z > z_{\alpha/2}) = \alpha/2$.

Proof.

$$\begin{aligned}
\mathbb{P}\mathrm{r}(\theta \in C_n) &= \mathbb{P}\mathrm{r}\left(-z_{\alpha/2} \leqslant \frac{\widehat{\theta}_n - \theta}{\mathrm{se}(\widehat{\theta}_n)} \leqslant z_{\alpha/2}\right) \\
&\to \mathbb{P}\mathrm{r}(-z_{\alpha/2} \leqslant Z \leqslant z_{\alpha/2}), \quad Z \sim N(0,1) \\
&= 1 - 2\Phi(-z_{\alpha/2}) = 1-\alpha.
\end{aligned}$$

■

The CLT (Lecture 7) tells us that for i.i.d. data with mean μ and variance σ^2,

$$\frac{\sqrt{n}(\widehat{\mu}_n - \mu)}{\sigma} \xrightarrow{d} N(0,1),$$

so the standard error is $\mathrm{se}(\widehat{\mu}_n) = \sigma/\sqrt{n}$. The resulting $1-\alpha$ confidence interval for μ is

$$C_n = \left[\bar{X}_n - z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}},\ \ \bar{X}_n + z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}}\right].$$

When σ is unknown, we substitute the sample standard deviation $S_n \xrightarrow{P} \sigma$ (proved in Lecture 7), giving the **studentized interval**

$$C_n = \left[\bar{X}_n - z_{\alpha/2}\frac{S_n}{\sqrt{n}},\ \ \bar{X}_n + z_{\alpha/2}\frac{S_n}{\sqrt{n}}\right].$$

By Slutsky's theorem, the coverage still converges to $1-\alpha$.

1.4 Hypothesis Testing

Confidence intervals answer the question "what values of θ are consistent with the data?" Hypothesis testing answers a complementary question: "is there sufficient evidence in the data to reject a specific claim about θ?"

Definition 1.16 (Hypothesis Testing Framework)**.** A **hypothesis test** specifies:

- **Null hypothesis** H_0: a statement representing the default assumption (e.g., $\theta = \theta_0$).
- **Alternative hypothesis** H_1: a statement that contradicts H_0 (e.g., $\theta \neq \theta_0$).
- **Test statistic** T_n: a function of the data that summarizes evidence against H_0.
- **Significance level** α: the tolerated probability of falsely rejecting H_0 (Type I error). Common choices are $\alpha = 0.05$ or 0.01.
- **Rejection region**: the set of values of T_n for which H_0 is rejected.
- p**-value**: the probability, under H_0, of observing a test statistic as extreme or more extreme than the observed value. We reject H_0 when the p-value $\leqslant \alpha$.

Two types of error arise in any hypothesis test:

	H_0 true	H_0 false
Reject H_0	**Type I error** (probability α)	Correct (power $= 1-\beta$)
Fail to reject H_0	Correct	**Type II error** (probability β)

The significance level α controls the Type I error rate. The **power** $1-\beta$ measures the probability of correctly detecting a true effect; it depends on the true value of θ, the sample size n, and the effect size. For a fixed α, power increases with n.

The duality between hypothesis testing and confidence intervals is a key structural insight: at level α, we reject $H_0 : \theta = \theta_0$ if and only if $\theta_0 \notin C_n$, where C_n is a $1-\alpha$ confidence interval.

Example 1.17 (Testing a Bernoulli Parameter)**.** Let $x_1, \ldots, x_n \overset{\text{i.i.d.}}{\sim} \text{Ber}(p)$. We wish to test whether $p = 1/2$.

Hypotheses.

$$H_0 : p = \tfrac{1}{2} \qquad \text{vs.} \qquad H_1 : p \neq \tfrac{1}{2}.$$

Test statistic. The sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ estimates p. Under H_0, by the CLT:

$$Z_n = \frac{\bar{x} - p_0}{\sqrt{p_0(1-p_0)/n}} \xrightarrow{d} N(0,1), \quad \text{where } p_0 = \tfrac{1}{2}.$$

Decision rule. At significance level α, reject H_0 if $|Z_n| > z_{\alpha/2}$; otherwise fail to reject.

Interpretation.

- $|Z_n| > z_{\alpha/2}$: the data are unlikely under H_0; we conclude there is sufficient evidence that $p \neq 1/2$.
- $|Z_n| \leqslant z_{\alpha/2}$: the data are consistent with H_0; we do not have enough evidence to conclude $p \neq 1/2$ at level α.

Equivalence with confidence interval. Failing to reject H_0 is equivalent to $p_0 = 1/2$ lying inside the $1-\alpha$ confidence interval $[\bar{x} \pm z_{\alpha/2}\sqrt{p_0(1-p_0)/n}]$.

1.5 Summary and Outlook

This lecture established the three pillars of classical inference for a parameter θ:

1. **Point estimation** ($\widehat{\theta}_n$): summarizes the data by a single value. Key quality criteria are bias, variance, MSE (= Bias^2 + Var), and consistency.
2. **Confidence intervals** (C_n): random intervals that trap θ with probability $\geqslant 1-\alpha$. The CLT-based normal CI $\widehat{\theta}_n \pm z_{\alpha/2}\,\mathrm{se}(\widehat{\theta}_n)$ applies whenever $(\widehat{\theta}_n - \theta)/\mathrm{se}(\widehat{\theta}_n) \xrightarrow{d} N(0,1)$.
3. **Hypothesis testing** (H_0 vs. H_1): rejects the null when the test statistic falls in the rejection region. Controlled at significance level α (Type I error), with power $1-\beta$ (probability of detecting a true effect).

All three procedures rest on the limit theorems of Lecture 7: the WLLN guarantees consistency, the CLT yields asymptotic normality, and Slutsky's theorem allows unknown nuisance parameters (like σ) to be estimated without affecting the limiting distribution.

What comes next. So far we have treated $\widehat{\theta}_n$ as a generic estimator. In Lecture 9 we study two principled methods for *constructing* estimators in parametric models:

- **Method of moments**: match theoretical moments $\mathbb{E}_\theta[X^k]$ to sample moments $\frac{1}{n}\sum X_i^k$ and solve for θ.
- **Maximum likelihood estimation (MLE)**: choose $\widehat{\theta}_n$ to maximize the likelihood $\prod_{i=1}^n f_\theta(X_i)$. The MLE is consistent, asymptotically normal, and—under regularity conditions—achieves the **Cramér–Rao lower bound**, making it the most efficient estimator in its class.

2 Lecture 10: 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.

2.1 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.

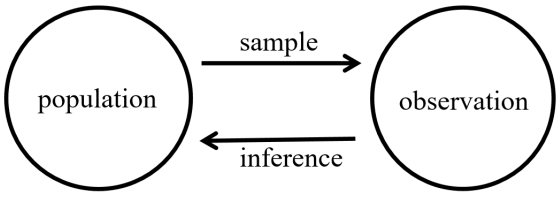

Figure 1: The parametric inference pipeline: population → sample → estimate.

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

Example 2.1 (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 2.2 (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 2.3 (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.

2.2 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 2.4 (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 2.5 (MoM for 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 2.6 (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).

2.3 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 2.7 (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

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

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 2.8 (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).*

2.3.1 MLE Examples

Example 2.9 (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 2.10 (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 2.6.

Example 2.11 (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.

References