Modern Statistics

Xiangyu Chang

March 31, 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.

References