Modern Statistics

Xiangyu Chang

April 21, 2026

Abstract

To be undated.

1 Lecture 15: Nonparametric Inference I

In the previous lectures we developed a comprehensive toolkit for **parametric inference**: given data, we assumed a parametric form for the distribution (e.g., $N(\mu, \sigma^2)$) or the regression function (e.g., $r(x) = \boldsymbol{x}^\top \boldsymbol{\beta}$), and used methods such as MLE and OLS to estimate the unknown parameters. However, in many applications the true distribution or regression function may not conform to any simple parametric family. This lecture introduces **nonparametric inference**, where we estimate quantities of interest—the distribution function F and the density f—without assuming a parametric model. We begin with the empirical distribution function (EDF) and its statistical properties, then develop histogram and kernel density estimators, and conclude by analyzing their performance through the mean integrated squared error (MISE).

1.1 Recall: Parametric Inference

We briefly recall the parametric paradigm to motivate the nonparametric approach.

Distribution Estimation

Given i.i.d. data $\{X_i\}_{i=1}^n$, we assume the distribution belongs to a known parametric family (e.g., $N(\mu, \sigma^2)$, $\text{Unif}(0, \theta)$, $\text{Ber}(p)$) with an unknown parameter vector θ. We then apply maximum likelihood estimation:

$$\text{MLE} \implies \widehat{\theta} \implies F_{\widehat{\theta}}.$$

Regression

Given paired data $\{(x_i, y_i)\}_{i=1}^n$, we seek the regression function $r(x) = \mathbb{E}[Y \mid X = x]$ that minimizes the mean squared error:

$$\min_r \ \mathbb{E}\left[(Y - r(X))^2\right].$$

In parametric regression, we assume a functional form such as $r_\beta(x) = \beta_0 + \beta_1 x$ (simple linear regression) or $r_\beta(x) = \boldsymbol{x}^\top \boldsymbol{\beta}$ (multiple linear regression), and apply MLE or least squares to obtain $r_{\hat{\beta}}(x)$.

All of the above are **parametric (generative) models**: they impose a specific structural assumption on the data-generating process. We now ask: what can we learn when no such assumption is made?

1.2 Nonparametric Inference

Definition 1.1 (Nonparametric Inference)**.** Nonparametric inference refers to statistical techniques that use data to infer unknown quantities of interest while making as few assumptions as possible. Given i.i.d. samples $\{X_i\}_{i=1}^n \sim F$, we aim to estimate the distribution function $F(x)$ or the density $f(x)$ without assuming a parametric form for F.

The simplest and most fundamental nonparametric estimator is the empirical distribution function.

1.2.1 Empirical Distribution Function (EDF)

Definition 1.2 (Empirical Distribution Function)**.** For i.i.d. samples $\{X_i\}_{i=1}^n$ from a distribution F, the **empirical distribution function** (EDF) is

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(X_i \leqslant x),$$

where $\mathbb{I}(X_i \leqslant x) = \begin{cases} 1 & \text{if } X_i \leqslant x, \\ 0 & \text{otherwise.} \end{cases}$

The EDF is motivated by the observation that

$$F(x) = \mathbb{P}\mathrm{r}(X \leqslant x) = \mathbb{P}\mathrm{r}(\mathbb{I}(X \leqslant x) = 1) = \mathbb{E}[\mathbb{I}(X \leqslant x)],$$

so $F_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}(X_i \leqslant x)$ is the sample average of the indicator random variables, and by the law of large numbers it should converge to $F(x)$.

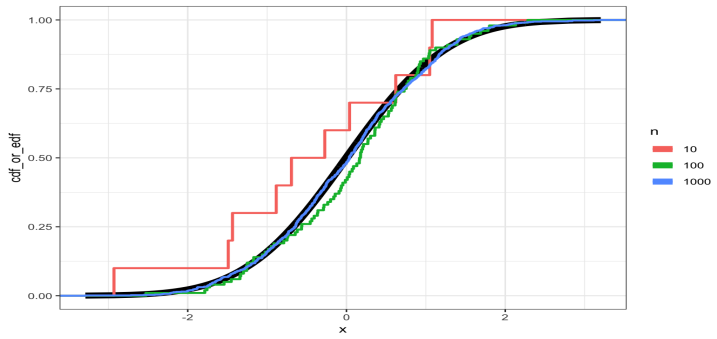

Figure 1: The empirical distribution function F_n (staircase) approximates the true CDF F (smooth curve).

Theorem 1.3 (Properties of the EDF)**.** *Let F_n be the EDF of i.i.d. samples $\{X_i\}_{i=1}^n \sim F$. For each fixed x:*

1. ***Unbiasedness:*** $\mathbb{E}[F_n(x)] = F(x)$.
2. ***Variance:*** $\mathrm{Var}(F_n(x)) = \dfrac{F(x)(1-F(x))}{n}$.
3. ***Consistency:*** $F_n(x) \xrightarrow{P} F(x)$ *as* $n \to \infty$.

Proof. 1. **Unbiasedness.** Since the X_i are i.i.d.,

$$\mathbb{E}[F_n(x)] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\mathbb{I}(X_i \leqslant x)] = \mathbb{E}[\mathbb{I}(X \leqslant x)] = \mathbb{P}\mathrm{r}(X \leqslant x) = F(x).$$

2. **Variance.** Each $\mathbb{I}(X_i \leqslant x) \sim \mathrm{Ber}(F(x))$, so

$$\begin{aligned}\mathrm{Var}(\mathbb{I}(X_i \leqslant x)) &= \mathbb{E}[\mathbb{I}(X_i \leqslant x)^2] - (\mathbb{E}[\mathbb{I}(X_i \leqslant x)])^2 \\ &= F(x) - F(x)^2 = F(x)(1-F(x)).\end{aligned}$$

   By independence,

$$\mathrm{Var}(F_n(x)) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(\mathbb{I}(X_i \leqslant x)) = \frac{F(x)(1-F(x))}{n}.$$

3. **Consistency.** By Chebyshev's inequality, for any $\epsilon > 0$:

$$\mathbb{P}\mathrm{r}(|F_n(x) - F(x)| \geqslant \epsilon) \leqslant \frac{\mathrm{Var}(F_n(x))}{\epsilon^2} = \frac{F(x)(1-F(x))}{n\epsilon^2} \xrightarrow{n\to\infty} 0.$$

   In fact, the Glivenko–Cantelli theorem gives the stronger *uniform* result: $\sup_x |F_n(x) - F(x)| \xrightarrow{\text{a.s.}} 0$.

■

The EDF provides a nonparametric estimate of the distribution function. To estimate the *density* $f = F'$ (when it exists), we turn to density estimation methods.

1.3 Density Estimation

While the EDF estimates the distribution function F directly, many applications require an estimate of the probability density function f. Since f is a continuous object, estimating it from discrete data is fundamentally harder than estimating F. We develop two approaches: the histogram and the kernel density estimator.

1.3.1 Histogram Density Estimation

Definition 1.4 (Histogram Density Estimator)**.** For i.i.d. samples $\{X_i\}_{i=1}^n$ with density f supported on $[0,1]$, the histogram density estimator is constructed as follows:

1. **Bin construction**: Partition $[0,1]$ into m bins of equal width $h = 1/m$:
$$B_1 = [0, \tfrac{1}{m}),\ B_2 = [\tfrac{1}{m}, \tfrac{2}{m}),\ \ldots,\ B_m = [\tfrac{m-1}{m}, 1].$$
2. **Count**: Let $n_j = \sum_{i=1}^n \mathbb{I}(X_i \in B_j)$ be the number of samples in bin B_j.
3. **Probability estimate**: $\widehat{p}_j = n_j/n$.
4. **Density estimate**: For $x \in B_j$,
$$\widehat{f}_n(x) = \frac{\widehat{p}_j}{h} = \frac{1}{h}\sum_{j=1}^m \widehat{p}_j\, \mathbb{I}(x \in B_j).$$

Remark 1.5 (Motivation)**.** *The construction is justified by the following observations:*

1. *By the mean value theorem,* $p_j = \int_{B_j} f(x)\,\mathrm{d}x = f(x^*)\,h$ *for some* $x^* \in B_j$.
2. *The probability estimate is unbiased:* $\mathbb{E}[\widehat{p}_j] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\mathbb{I}(X_i \in B_j)] = \mathbb{P}\mathrm{r}(X \in B_j) = p_j$.
3. *Therefore,* $\mathbb{E}[\widehat{f}_n(x)] = \frac{p_j}{h} = f(x^*)$*, which approximates* $f(x)$ *when* h *is small (so that* $x \approx x^*$*).*

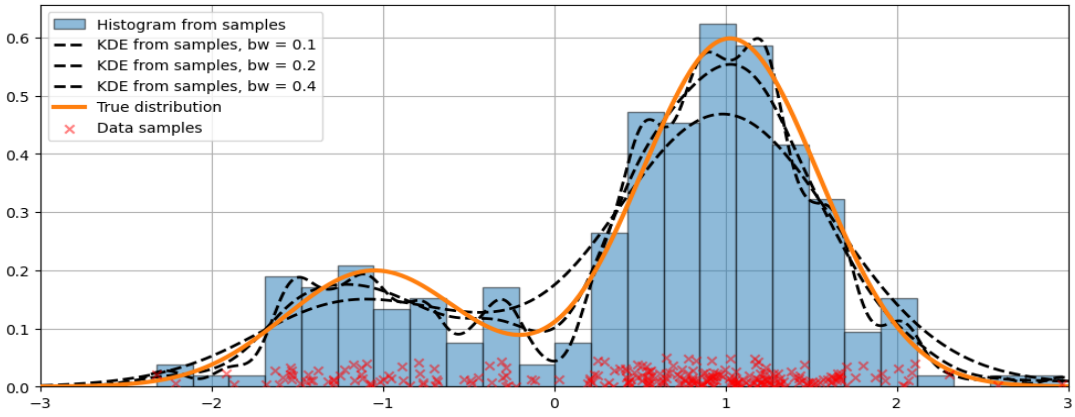

Figure 2: Histogram density estimation with different bin widths h.

The histogram is simple but produces a discontinuous density estimate. A natural question is whether we can obtain a smoother estimator. This motivates kernel density estimation.

1.4 Summary and Outlook

This lecture introduced the foundations of nonparametric inference:

1. **Empirical distribution function**: $F_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}(X_i \leqslant x)$ is an unbiased, consistent estimator of $F(x)$ with variance $F(x)(1-F(x))/n$.

2. **Histogram density estimation**: partitioning the domain into bins of width h and estimating the density as $\widehat{f}_n(x) = \widehat{p}_j/h$. The optimal bandwidth is $h = O(n^{-1/3})$, giving MISE $= O(n^{-2/3})$.

References