Modern Statistics **Lecture 14 - 04/07/2025**

Nonparametric Inference

Lecturer:Xiangyu Chang *Scribe: LAI JIA XUAN, ARTEM ROMANOV*

Edited by: Zhihong Liu

1 Empirical Distribution Function (EDF)

Definition 1 *Given distribution function F, its empirical distribution function is:*

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x).$$

Proposition 1 *The EDF estimator is unbiased.*

Proof 1

$$\mathbb{E}[F_n(x)] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[I(X_i \le x)] = \frac{1}{n}\sum_{i=1}^{n} P(X_i \le x) = \frac{1}{n}\cdot nF(x) = F(x).$$

2 Histogram Density Estimator

Definition 2 *Histogram density estimator for density function f is defined as*

$$\hat{f}_n(x) = \frac{1}{h}\sum_{j=1}^{m}\hat{P}_j I(x \in B_j), \quad \hat{P}j = \frac{n_j}{n}, \quad n_j = \sum_{i=1}^{n} I(X_i \in B_j).$$

Let $B_j = [b_{j-1}, b_j)$,

$$h = b_j - b_{j-1}.$$

Bias:

For $x \in B_j$,

$$\mathbb{E}[\hat{f}_n(x)] = \frac{1}{h}\mathbb{E}[\hat{P}j] = \frac{1}{h}P(X_i \in B_j) = \frac{1}{h}\int B_j f(u)du.$$

Let x^* be the midpoint of B_j, use Taylor expansion:

$$f(u) = f(x^*) + f'(x^*)(u - x^*) + \cdots \Rightarrow \int_{B_j} f(u)du \approx hf(x^*),$$

Thus,

$$\mathbb{E}[\hat{f}_n(x)] \approx f(x) \Rightarrow |\mathbb{E}[\hat{f}_n(x)] - f(x)| \le |f(x) - f(x)| \le L|x - x^*| \le Lh.$$

Variance:

The variance can be calculated through

$$\hat{f}_n(x) = \frac{n_j}{nh},$$

$$\mathrm{Var}(\hat{f}_n(x)) = \frac{1}{n^2h^2}\mathrm{Var}(n_j).$$

Since $n_j \sim \mathrm{Binomial}(n, p_j)$, with $p_j = P(X_i \in B_j) \approx hf(x^*), \mathrm{Var}(n_j) = np_j(1-p_j) \approx nhf(x^*)(1 - hf(x^*))$, Therefore,

$$\mathrm{Var}(\hat{f}_n(x)) = \frac{hf(x)}{nh^2} = \frac{f(x)}{nh}.$$

MISE:

Proposition 2 (Mean Integrated Squared Error (MISE)) *MISE* $= \mathcal{O}(n^{-\frac{2}{3}})$, *where* $MISE = \mathbb{E}\left[\int(\hat{f}_n(x) - f(x))^2 dx\right] = \int b^2(x)dx + \int Var[\hat{f}_n(x)]dx$.

Proof 2

$$bias(x) = \mathbb{E}[\hat{f}_n(x)] - f(x),$$

$$Var(x) = \mathbb{E}\left[\hat{f}_n(x) - \mathbb{E}[\hat{f}_n(x)]\right]^2,$$

$\hat{f}_n(x)$ *is a histogram*

$$\begin{aligned} |bias(x)| &= |\mathbb{E}[\hat{f}_n(x)] - f(x)| \text{ or } |f(x) - f(y)| \le L|x-y| \\ &= |f(x^*) - f(x)| \le L|x^* - x| \\ &\le Lh. \end{aligned}$$

f *is L-Lipschitz:* $\max_{x\in[0,1]} |f(x)| \le M$, $\max_{x\in[0,1]} |f'(x)| \le L$, *for if* $x \in B_j$,

$$\begin{aligned} Var[\hat{f}_n(x)] &= Var\left(\frac{\hat{p}_j}{n}\right) \\ &= \frac{1}{h^2} Var(\hat{p}_j) = \frac{1}{h^2} Var\left(\frac{n_j}{n}\right) \\ &= \frac{1}{n^2h^2} Var\left[\sum_{i=1}^{n} \mathbb{I}(x_i \in B_j)\right] \\ &= \frac{1}{n^2h^2} \sum_{i=1}^{n} Var[\mathbb{I}(x_i \in B_j)] \\ &= \frac{1}{n^2h^2} np_j(1-p_j) = \frac{p_j(1-p_j)}{nh^2} = \frac{h \cdot f(x^*)(1 - hf(x^*))}{nh^2} \\ &= \frac{f(x^*)(1-hf(x^*))}{nh} \le \frac{M}{nh} + \frac{M^2}{n}. \end{aligned}$$

Then,

$$MISE = L^2h^2 + \frac{M}{nh} + \frac{M^2}{n} = L^2h^2 + \frac{M}{2nh} + \frac{M}{2nh} + \frac{M^2}{n} \geq 3\sqrt[3]{L^2h^2\left(\frac{M}{2nh}\right)^2} + \frac{M^2}{n},$$

where $L^2h^2 = \frac{M}{2nh}$, $h_{opt} = \left(\frac{M}{2nL^2}\right)^{\frac{1}{3}} = \mathcal{O}(n^{-\frac{1}{3}})$, *and* $MISE = \mathcal{O}(n^{-\frac{2}{3}})$.

3 Kernel Density Estimation (KDE)

Definition 3 (Kernel Density Estimation) *Kernel density estimatior is defined as*

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right),$$

where K *satisfies*

- $\int_{\mathbb{R}} K(x)dx = 1$.
- $K(x) = K(-x)$.
- $\lim_{x\to\infty} K(x) = \lim_{x\to-\infty} K(x) = 0$.

We also define some quantity for kernel density estimation, $\mu_k^2 \triangleq \int x^2K(x)dx < +\infty, \sigma_k^2 \triangleq \int k^2(x)dx < +\infty$.

Bias:

The bias at x_0 is

$$\begin{aligned}
\text{Bias}(x_0) &= \mathbb{E}[f_n(x_0)] - f(x_0) = \frac{1}{nh}\sum_{i=1}^{n}\mathbb{E}[K(\frac{x_i - x_0}{h})] - f(x_0) \\
&= \frac{1}{h}\mathbb{E}[K(\frac{x - x_0}{h})] - f(x_0) \\
&= \frac{1}{h}\int K(\frac{x - x_0}{h})f(x)dx - f(x_0) \\
&\overset{y=\frac{x-x_0}{h}}{=} \frac{1}{h}\int K(y)f(x_0 + hy)h \cdot dy \\
&= \int K(y)f(x_0 + hy)dy - \int K(y)f(x)dy \\
&= \int K(y)[f(x_0 + hy) - f(x_0)]dy \\
&= \int K(y)[f'(x_0)hy + \frac{f''(x_0)h^2y^2}{2} + \mathcal{O}(h^3)]dy \\
&= \frac{f''(x_0)h^2}{2}\int y^2K(y)dy + \mathcal{O}(h^3) \\
&= \frac{f''(x_0)h^2\mu_k^2}{2}.
\end{aligned}$$

Variance:

The variance of $f_n(x_0)$ is

$$\begin{aligned}
\mathrm{Var}(f_n(x_0)) &= \frac{1}{n^2h^2}\sum_{i=1}^{n}\mathrm{Var}(K(\frac{x_i-x_0}{h})),\\
&= \frac{1}{nh^2}\mathrm{Var}(K(\frac{x-x_0}{h}))\\
&\le \frac{1}{nh^2}\mathbb{E}[k^2(\frac{x-x_0}{h})]\\
&= \frac{1}{nh^2}\int k^2(\frac{x-x_0}{h})f(x)dx\\
&= \frac{1}{nh}\int k^2(y)f(x_0+hy)dy\\
&= \frac{1}{nh}\int k^2(y)[f(x_0)+f'(x_0)hy+\mathcal{O}(h)]dy\\
&= \frac{f(x_0)\sigma_k^2}{nh}+\frac{f'(x_0)}{n}\int yk^2(y)dy+\mathcal{O}(\frac{1}{nh})\\
&= \frac{f(x_0)\sigma_k^2}{nh}.
\end{aligned}$$

MISE:

$$\begin{aligned}
\mathrm{MISE} &= \frac{[f''(x_0)]^2h^2\mu_k^4}{4}+\frac{f(x_0)\sigma_k^2}{nh}\\
&= \frac{[f''(x_0)]^2h^2\mu_k^4}{4}+\frac{f(x_0)\sigma_k^2}{4nh}\cdot 4 \ge 5\cdot\sqrt[5]{\frac{[f''(x_0)]^2h^2\mu_k^4}{4}\left(\frac{f(x_0)\sigma_k^2}{4nh}\right)^4}.
\end{aligned}$$

The condition for the equality is $h^4=\frac{f(x_0)\sigma_k^2}{n\mu_k^4[f''(x_0)]^2}$, which is $h_{\mathrm{opt}}=\left[\frac{f(x_0)\sigma_k^2}{n\mu_k^4[f''(x_0)]^2}\right]^{\frac{1}{5}}=\mathcal{O}(n^{\frac{1}{5}})$, which implies $\mathrm{MISE}=\mathcal{O}(n^{-\frac{4}{5}})$.

4 Nonparametric Regression

Suppose the data is $\{(x_i,y_i)\}_{i=1}^n$, the nonparametric regression function r is defined $r(x):\mathbb{R}^d\to\mathbb{R}$, where $y=r(x)+\epsilon$ and r is actually

$$r(x)=\mathbb{E}[y|x]=\begin{cases}r(x)=\beta_0+\beta_1x \text{ or } r(x)=\beta^{\mathrm{T}}x, \text{ for linear regression.}\\ r(x)=\frac{\sum_{i=1}^n w_{n,i}(x)y_i}{\sum_{i=1}^n w_{n,i}}, \text{ for nonparametric regression.}\end{cases}\tag{1}$$

Define $w_{n,i}(x)\triangleq w(x,x_1,\cdots,x_n)$, $v_{n,j}(x)\triangleq\frac{w_{n,j}(x)}{\sum_{i=1}^n w_{n,i}(x)}$, then $r(x)=\sum_{i=1}^n w_{n,i}(x)y_i$ and $\sum_{i=1}^n v_{n,i}(x)=1$.

Example 1 (Partition Estimation) *For $R = \bigcup_{j=1}^{M} B_j$, where $B_i \cap B_j = \varnothing$, define the partition estimator $r_n(x)$ is*

$$r_n(x) = \frac{\sum_{i=1}^{n} \mathbb{I}(x_i \in B_j) y_i}{\sum_{i=1}^{n} \mathbb{I}(x_i \in B_j)}, x \in B_j. \tag{2}$$

Example 2 (KNN) *For fixed x, $x_{(i)}$ is defined as the i-th nearest neighbor with respect to norm $||\cdot||$, $i = 1, 2, \cdots, n$, that is,*

$$||x_{(1)} - x|| \leq ||x_{(2)} - x|| \leq \cdots \leq ||x_{(n)} x||.$$

Then, for some constant K, the KNN estimator is defined as

$$r_n(x) = \frac{1}{K} \sum_{i=1}^{k} y_{(i)}. \tag{3}$$

Example 3 (Kernel Estimation) *For some kernel function K, the kernel estimator is given as*

$$r_n(x) = \frac{\sum_{i=1}^{n} K(\frac{x_i - x}{h}) y_i}{\sum_{i=1}^{n} K(\frac{x_i - x}{h})}. \tag{4}$$

For instance, if $K(x) = \mathbb{I}[-\frac{1}{2}, \frac{1}{2}]$, then only $x \in [x_i - \frac{h}{2}, x_i + \frac{h}{2}]$ can have no-zero weights.

5 Bootstrap Method

5.1 Confidence Interval for the Mean: Classical vs. Bootstrap

1. Classical Confidence Interval via Central Limit Theorem

Let $X_1, X_2, \ldots, X_n \overset{\text{i.i.d.}}{\sim} F$ with $\mu = \mathbb{E}[X_i]$ and $\sigma^2 = \mathrm{Var}(X_i) < \infty$. Define the sample mean and sample variance:

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

Then, by the Central Limit Theorem (CLT),

$$\frac{\bar{X}_n - \mu}{\hat{\sigma}/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, 1),$$

so an approximate $100(1-\alpha)\%$ confidence interval for μ is

$$\left(\bar{X}_n - z_{\alpha/2} \cdot \frac{\hat{\sigma}}{\sqrt{n}},\ \bar{X}_n + z_{\alpha/2} \cdot \frac{\hat{\sigma}}{\sqrt{n}} \right).$$

2. Bootstrap Confidence Interval (Normal Approximation)

The classical CLT-based method relies on analytic variance estimation and asymptotic normality. But what if σ^2 is hard to estimate, or the sampling distribution of $\bar{X}_n$ is complicated?

We can take a different approach by approximating the sampling distribution of $\bar{X}_n$ using **bootstrap**, which replaces the unknown distribution F by the empirical distribution F_n.

1. Resample: Draw B bootstrap samples $(X_1^{*(b)}, \ldots, X_n^{*(b)})$ from the empirical distribution F_n (i.e., sample with replacement from $\{X_1, \ldots, X_n\}$).
2. For each resample, compute the sample mean $\bar{X}_n^{*(b)}$.
3. Estimate the standard error of $\bar{X}_n$ by the sample standard deviation of these bootstrap means:
$$\hat{se}_{\text{boot}} = \sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(\bar{X}_n^{*(b)} - \bar{X}_n^{*}\right)^2}, \quad \text{where } \bar{X}_n^{*} = \frac{1}{B}\sum_{b=1}^{B}\bar{X}_n^{*(b)}.$$
4. Construct the confidence interval using a normal approximation:
$$\left(\bar{X}_n - z_{\alpha/2}\cdot \hat{se}_{\text{boot}},\ \bar{X}_n + z_{\alpha/2}\cdot \hat{se}_{\text{boot}}\right).$$

3. Comparison of the Two Methods

- **Classical CLT method:** Uses theoretical results (LLN + CLT + Slutsky) to derive asymptotic normality and estimate the variance analytically via $\hat{\sigma}^2$.
- **Bootstrap method:** Bypasses analytical variance estimation and instead estimates the standard error directly via simulation from the empirical distribution. Assumes the bootstrap distribution is approximately normal around the original estimator.
- **Key difference:** Classical method relies on known asymptotic theory; bootstrap relies on resampling from the observed data, assuming $F_n \approx F$.
- **Heuristic nature:** The bootstrap confidence interval via normal approximation is heuristic—it works well when the sampling distribution is roughly symmetric and the sample size is large, but it doesn't follow from a strict limit theorem like the classical method.

Example 4 (Bootstrap Sampling Example) *Suppose we already have one sample set* $\{X_i\}_{i=1}^n$, *with* $X_i \overset{i.i.d.}{\sim} F$.

$$\textit{Step 1: } \{X_i^{(1)}\}_{i=1}^n \to T_M^{(1)}.$$
$$\vdots$$
$$\textit{Step B: } \{X_i^{(B)}\}_{i=1}^n \to T_M^{(1)}.$$

Then,

$$\bar{T}_M = \frac{1}{B}\sum_{b=1}^{B} T_M^{(b)}, \hat{\sigma}_M^2 = \frac{1}{n}\sum_{b=1}^{B}(T_M^{(b)} - \bar{T}_M)^2.$$

Specifically, bootstrap for linear regression can be demonstrated as

Linear Model: *$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, and the least square estimator for β_1 is $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$, where* ***Var***$(\hat{\beta}_1) = \frac{\sigma^2}{nS_{xx}}$.

$$\textbf{\textit{Step 1:}}\ \{X_i^{(1)}\}_{i=1}^n \to \hat{\beta}_1^{(1)} = \frac{S_{xy}^{(1)}}{S_{xx}^{(1)}}.$$

$$\vdots$$

$$\textbf{\textit{Step B:}}\ \{X_i^{(B)}\}_{i=1}^n \to \hat{\beta}_1^{(B)} = \frac{S_{xy}^{(B)}}{S_{xx}^{(B)}}.$$

Then,

$$\bar{\hat{\beta}}_1 = \frac{1}{B}\sum_{b=1}^{B} \hat{\beta}_1^{(b)},\ \hat{\sigma}^2 = \frac{1}{n}\sum_{b=1}^{B}(\hat{\beta}_1^{(b)} - \bar{\hat{\beta}}_1)^2.$$

We have $\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}} \xrightarrow{d} \mathcal{N}(0,1)$, so the confidence intervel for β_1 is $\left(\hat{\beta}_1 - \hat{\sigma} z_{\alpha/2},\ \hat{\beta}_1 + \hat{\sigma} z_{\alpha/2}\right)$.