Estimations of Distribution

Zhijian He

2020-10-26

Estimations of Distribution

`Setting`: $X_1, \ldots, X_n \overset{iid}{\sim} F(x)$

- How to estimate $F(x)$ for a given x without any assumption on F?
- If $F(x)$ has a PDF $f(x)$, how to estimate $f(x)$?

In this part, we do not assume a parametric family for the population. The methods we introduce are non-parametric methods.

Empirical CDF (ECDF)

Let $X \sim F(x)$. Notice that

$$F(x) = \mathbb{P}(X \leq x) = \mathbb{E}[1\{X \leq x\}].$$

Applying the method of moments, one gets an estimator for $F(x)$,

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1\{X_i \leq x\}.$$

If X_is are observed, the estimator $\hat{F}_n(x)$ is a function of $x \in \mathbb{R}$. It is a CDF function, which is called the ECDF.

ECDF is a stepwise function.

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1\{x_i \leq x\} = \begin{cases} 0, & x < x_{(1)} \\ 1/n, & x_{(1)} \leq x < x_{(2)} \\ 2/n, & x_{(2)} \leq x < x_{(3)} \\ & \vdots \\ k/n, & x_{(k)} \leq x < x_{(k+1)} \\ & \vdots \\ 1, & x > x_{(n)} \end{cases}$$

Consistency of ECDF: $\hat{F}_n(x) \overset{w.p.1}{\to} F(x)$ for any $x \in \mathbb{R}$.

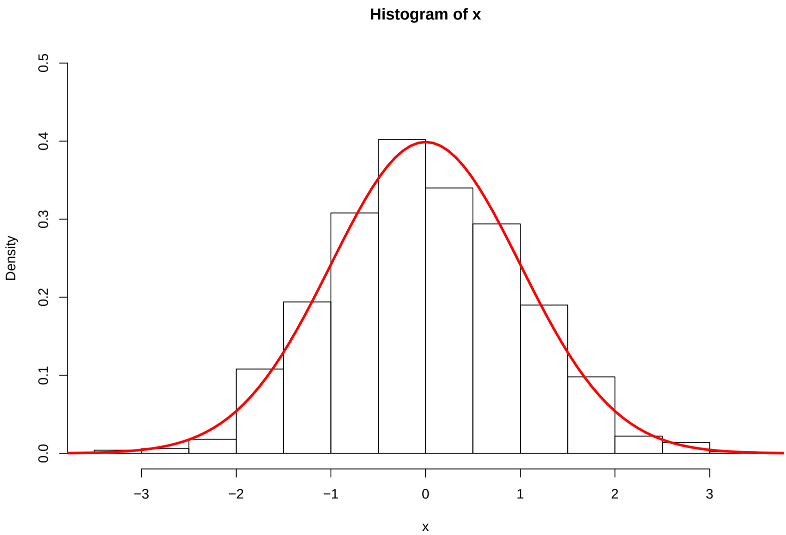
Stronger result (Glivenko–Cantelli theorem):

$$\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| \overset{w.p.1}{\to} 0$$

Estimation of PDF

1. The histogram estimation method
2. The kernel density estimation method

The histogram estimation method



Suppose that $-\infty < t_0 < t_1 < \cdots < t_m < \infty,\ t_{i+1} - t_i = h > 0$.

$$\hat{f}_n(x) := \begin{cases} \frac{\hat{F}_n(t_{i+1}) - \hat{F}_n(t_i)}{h}, \ x \in (t_i, t_{i+1}], i = 0, \ldots, m-1 \\ 0, x \leq t_0, x > t_m \end{cases}$$

- $t_0 < x_{(1)},\ t_m > x_{(n)}$
- rule of thumb: $m \approx 1 + 3.322 \log_{10} n$

Under some conditions, particularly $\lim_n h_n = 0, \lim_n nh_n = \infty$,

Consistency: $\hat{f}_n(x) \stackrel{w.p.1}{\to} f(x)$ for any $x \in \mathbb{R}$.

Stronger result:

$$\sup_{x \in \mathbb{R}} |\hat{f}_n(x) - f(x)| \stackrel{w.p.1}{\to} 0$$

NB: $\hat{f}_n(x)$ is stepwise (discontinuous).

The kernel density estimation

The idea: central difference

$$f(x) \approx \frac{F(x+h) - F(x-h)}{2h} \approx \frac{\hat{F}_n(x+h) - \hat{F}_n(x-h)}{2h}$$

$$\hat{f}_n(x) := \frac{1}{2hn} \sum_{i=1}^{n} 1\{x - h < X_i \leq x + h\} = \frac{1}{hn} \sum_{i=1}^{n} K_0\left(\frac{x - X_i}{h}\right)$$

- $K_0(x) = \frac{1}{2} 1\{-1 \leq x < 1\}$, a PDF of $U[-1, 1]$.

Generalization: use a kernel function $K(x)$ which is a CDF

$$\hat{f}_n(x) := \frac{1}{hn} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

Some common kernel functions:

- uniform kernels

$$K_0(x) = \frac{1}{2} 1\{-1 \leq x \leq 1\}$$

$$K_1(x) = 1\{-1/2 \leq x \leq 1/2\}$$

- Gaussian kernel (default setting for R/Matlab)

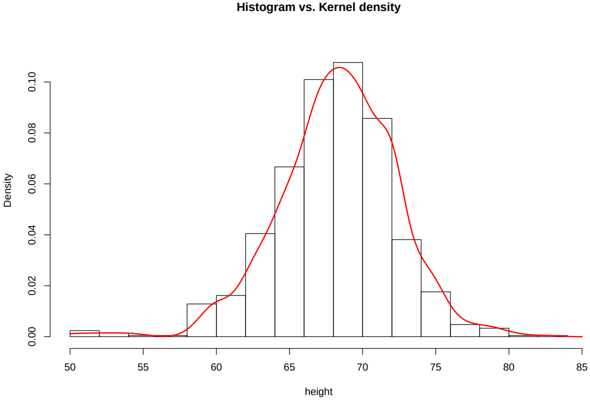
$$K_2(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

Similar consistency results also hold for kernel density estimation if the kernel function $K(x)$ satisfies

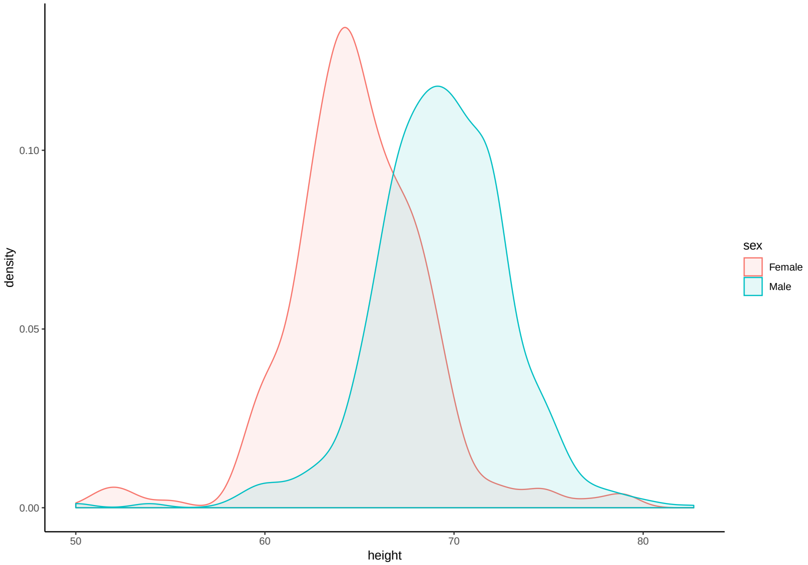
$$\int_{-\infty}^{\infty} K(x)^2 dx < \infty, \ \lim_{|x| \to \infty} |x| K(x) = 0.$$

Heights data

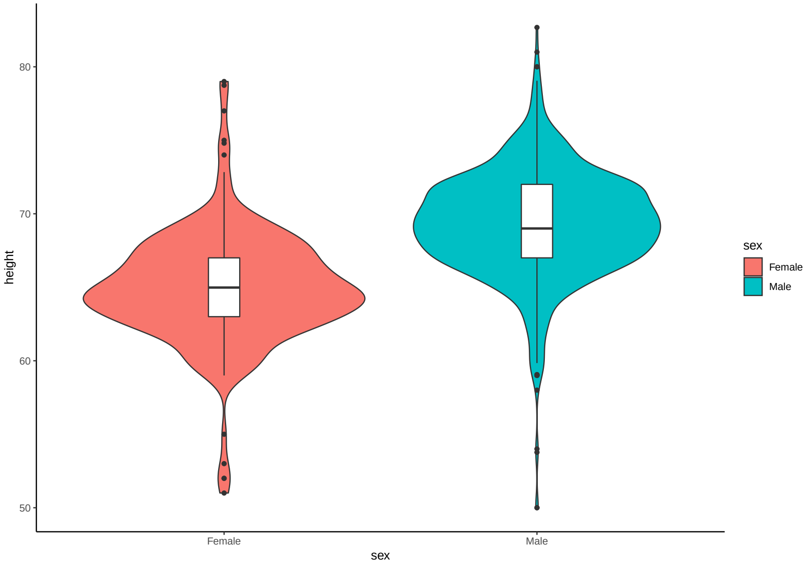
We now look at the data “heights” (in inches) from the R package `dslabs`.



Using R package ggplot2



Violin plot + box plot



Comments

- `rule of thumb`: $h_n \approx 1.06 S_n n^{-1/5}$ for Gaussian kernel
- MSE rates comparison

histogram estimation	$O(n^{-2/3})$
kernel density estimation	$O(n^{-4/5})$

- kernel density estimation is faster!