The Histogram and Kernel Density Estimation

Probability and Statistics for Data Science

Carlos Fernandez-Granda

NYU COURANT INSTITUTE OF MATHEMATICAL SCIENCES

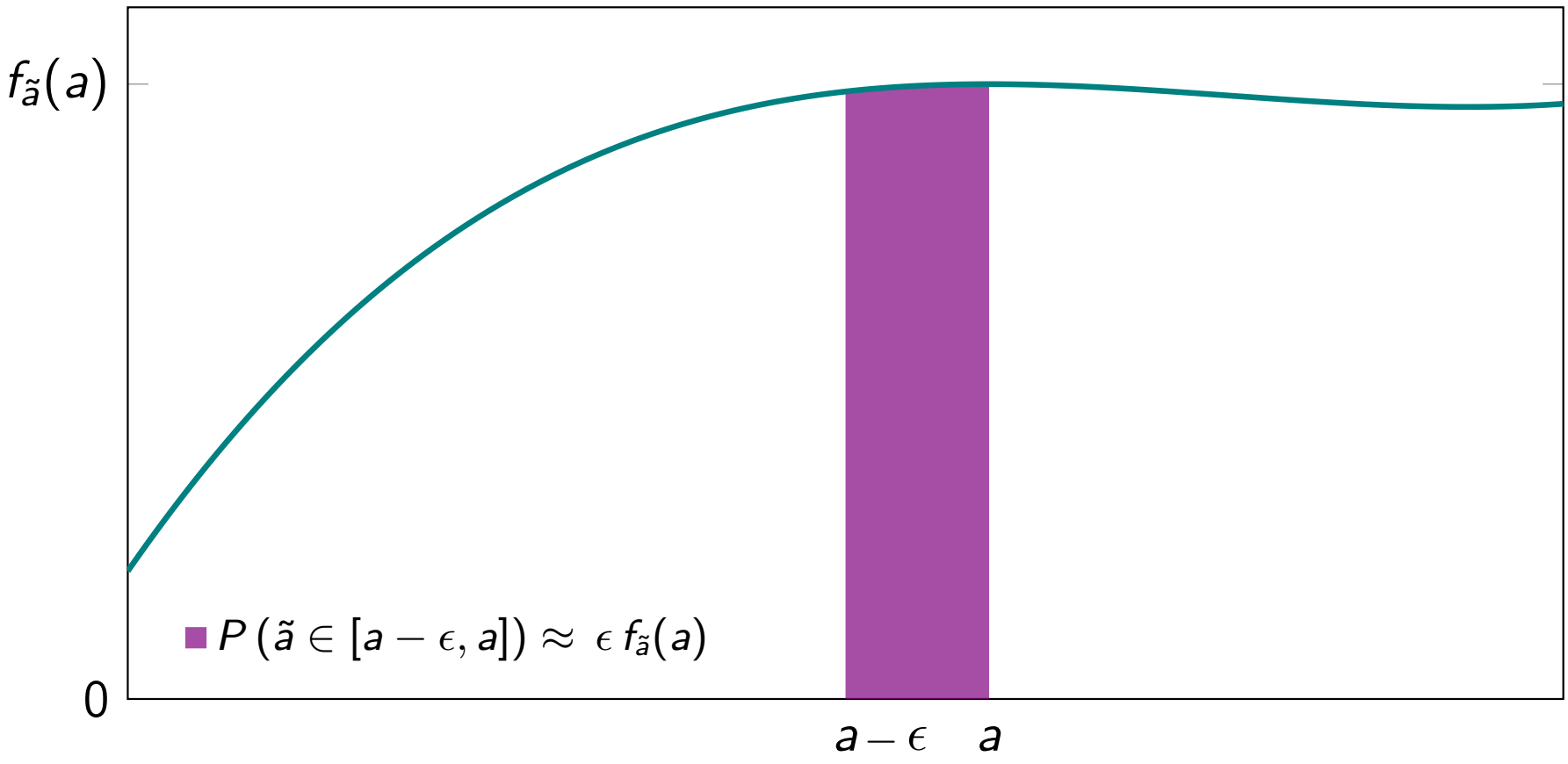
NYU DATA SCIENCE

These slides are based on the book Probability and Statistics for Data Science by Carlos Fernandez-Granda, available for purchase here. A free preprint, videos, code, slides and solutions to exercises are available at https://www.ps4ds.net

Goal

Estimate probability density from data

Probability density function



Definition

Let $\tilde{a} : \Omega \to \mathbb{R}$ be a random variable with cdf $F_{\tilde{a}}$

If $F_{\tilde{a}}$ is differentiable, the pdf of $\tilde{a}$ is

$$f_{\tilde{a}}(a) := \frac{\mathrm{d}F_{\tilde{a}}(a)}{\mathrm{d}a}$$

How to estimate a pdf from data?

Can we just differentiate the empirical cdf?

Empirical cdf

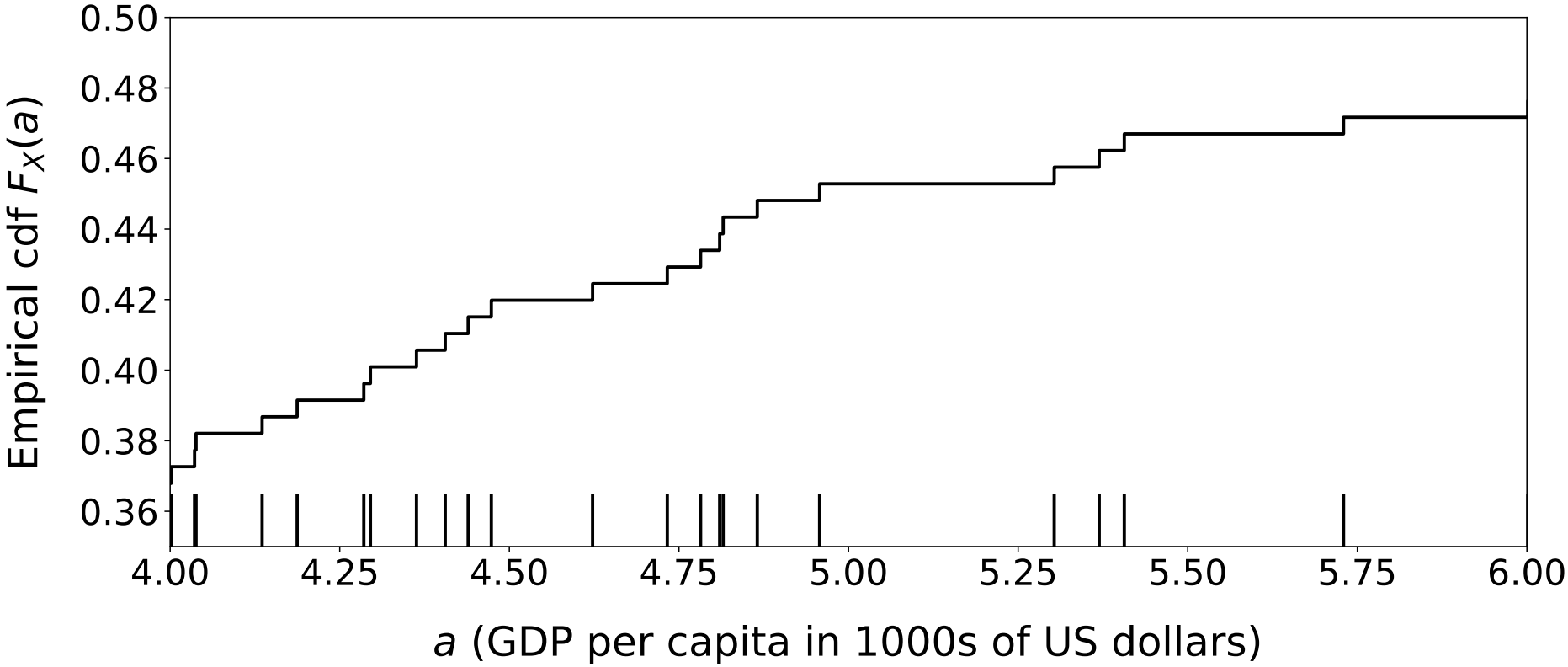
Dataset $X := \{x_1, x_2, \ldots, x_n\}$

The empirical cumulative distribution function $F_X : \mathbb{R} \to [0, 1]$ equals

$$F_X(a) := \frac{1}{n} \sum_{i=1}^{n} 1_{x_i \leq a}$$

where $1_{x_i \leq a}$ equals one if $x_i \leq a$ and zero otherwise

Empirical cdf



How to estimate a pdf from data?

We need it to be nonnegative and integrate to one

We cannot use empirical probabilities, probability of each data point is zero!

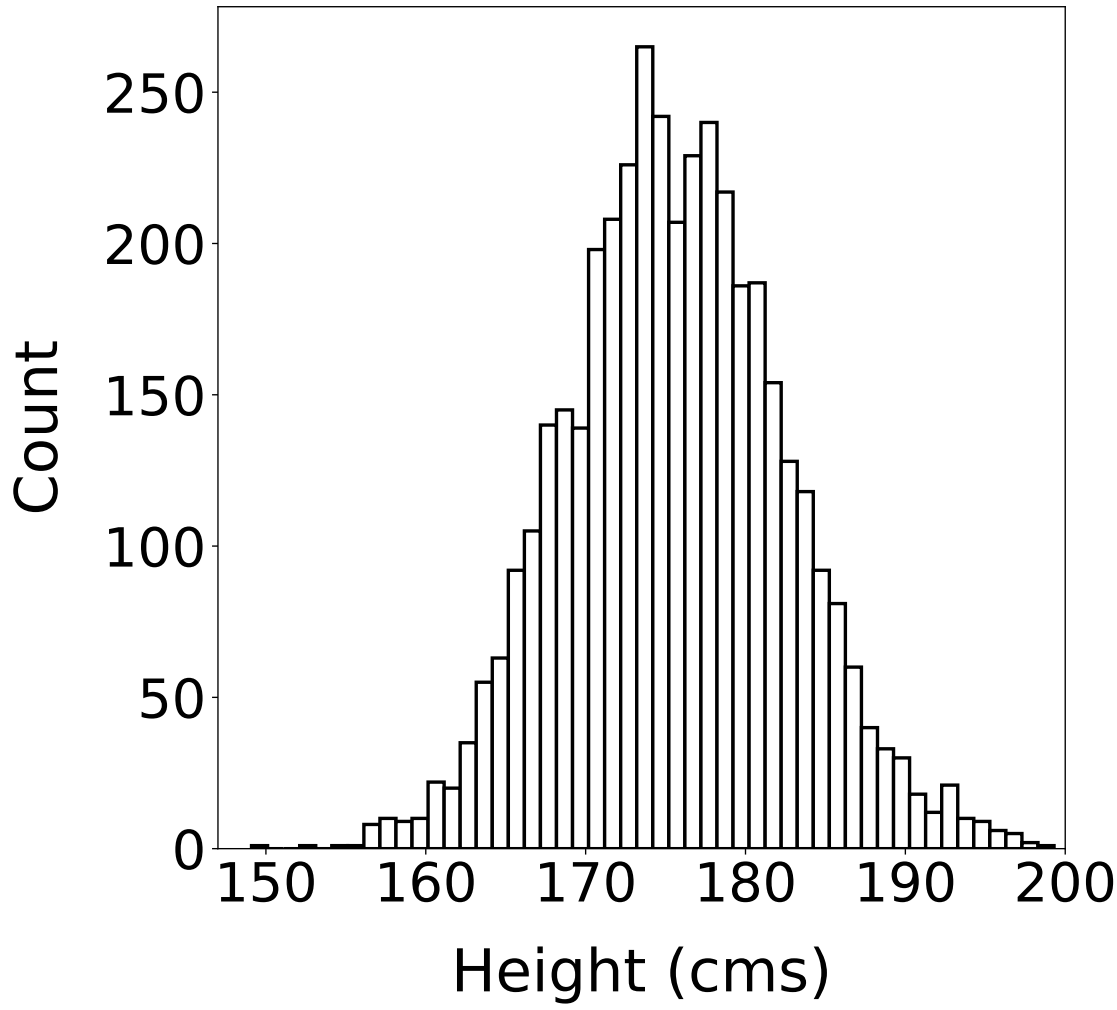
We need to make some assumption about the density

Histogram

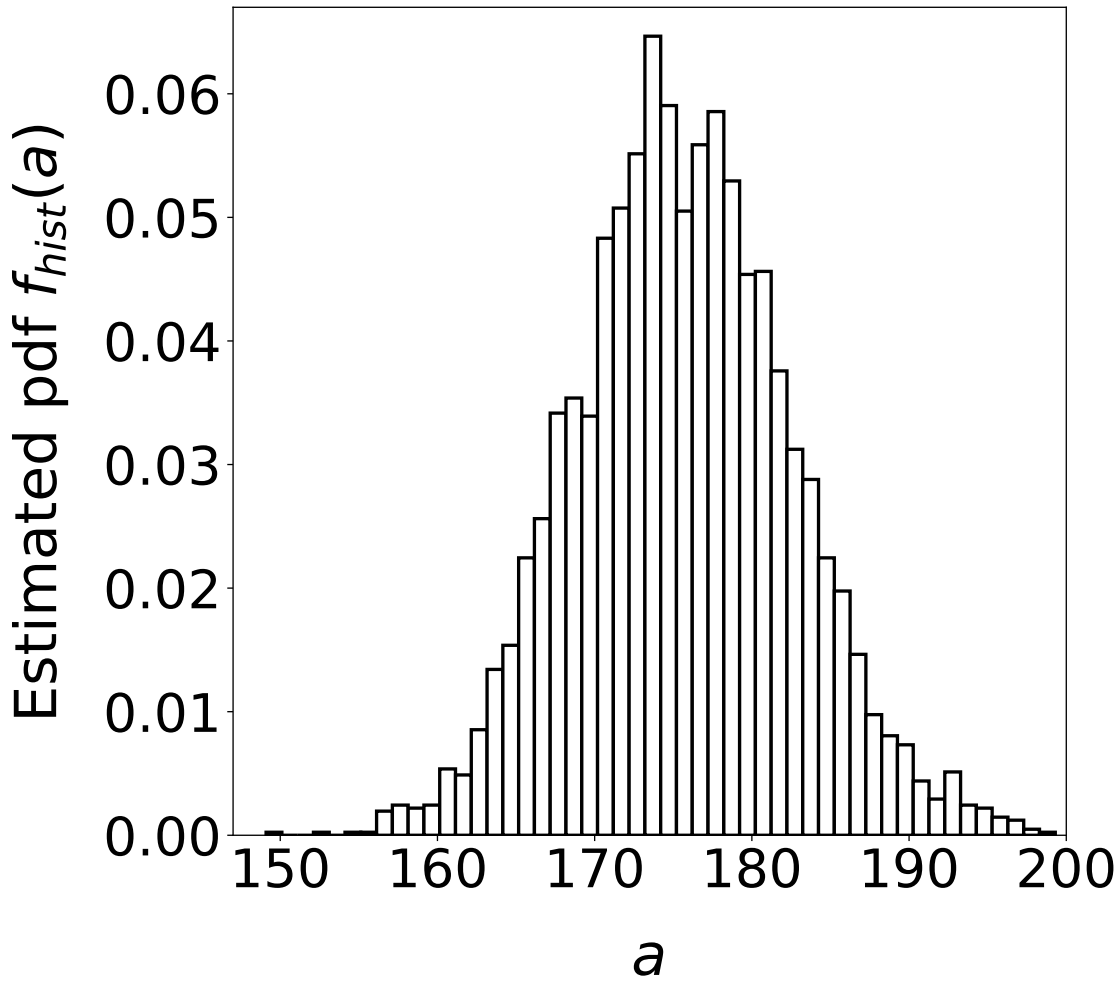
Since $\mathrm{P}(a - \epsilon < \tilde{a} \leq a) \approx \epsilon f_{\tilde{a}}(a)$

1. Divide possible values into bins
2. Approximate the probability of each bin using empirical probability
3. Assume constant density in each bin (divide probability by bin width)

Height in US army



Normalized histogram



Histogram

Let $X := \{x_1, x_2, \ldots, x_n\}$ be data in an interval of length ℓ

1. Divide interval into b bins of length ℓ/b

2. Count fraction of data in each bin

$$c_i := \sum_{j=1}^{n} 1_{x_j \in \mathcal{B}_i}, \quad 1 \leq i \leq b$$

   where $1_{x_j \in \mathcal{B}_i}$ equals one if x_j is in $\mathcal{B}_i$ and zero otherwise

3. Normalize to obtain density estimate for $t \in \mathcal{B}_i$

$$f_{\text{hist}}(t) := \frac{b c_i}{\ell n}$$

Is f_{hist} a valid pdf?

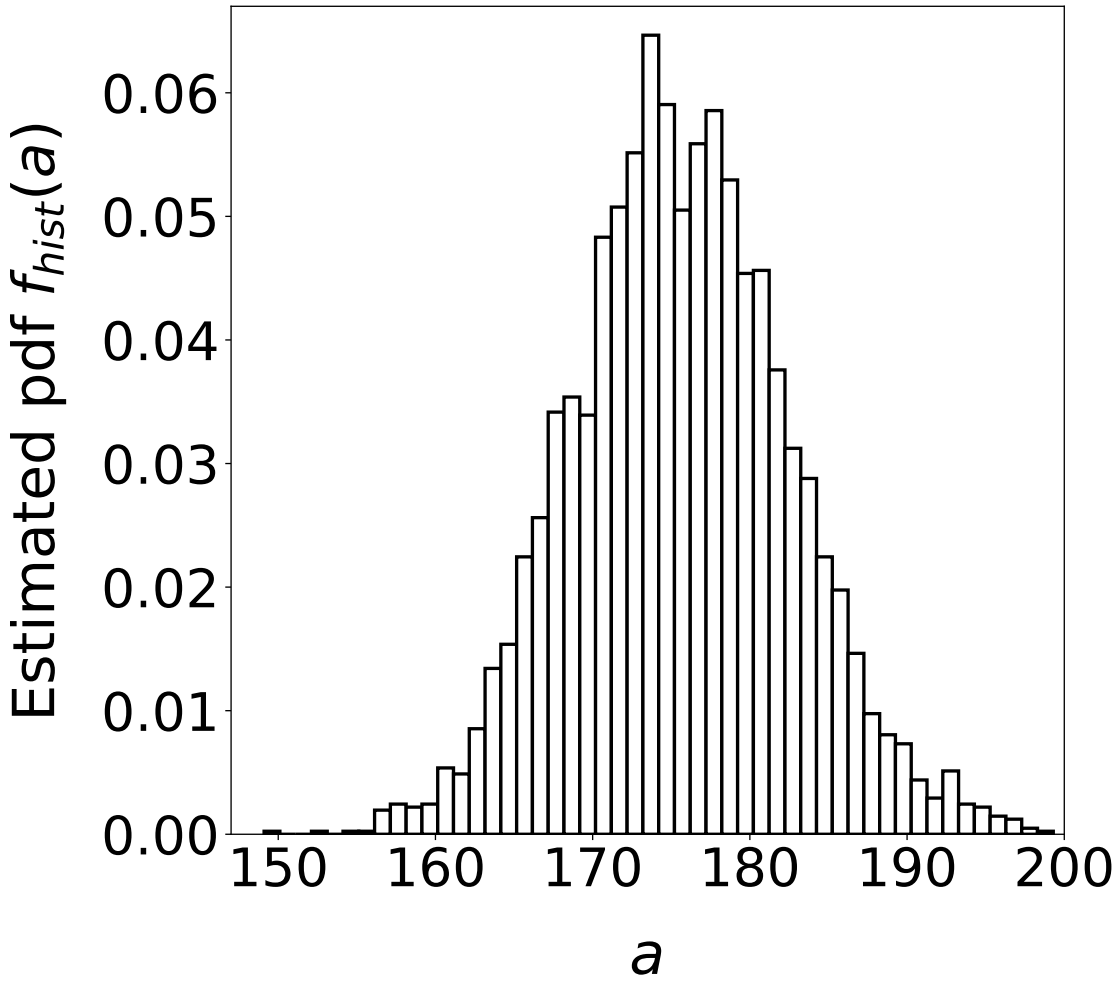
$$f_{\text{hist}}(t) = \sum_{j=1}^{n} \Pi_j(t)$$

Π_j: rectangle of length $\frac{\ell}{b}$ in bin $\mathcal{B}_j$ where x_j is located

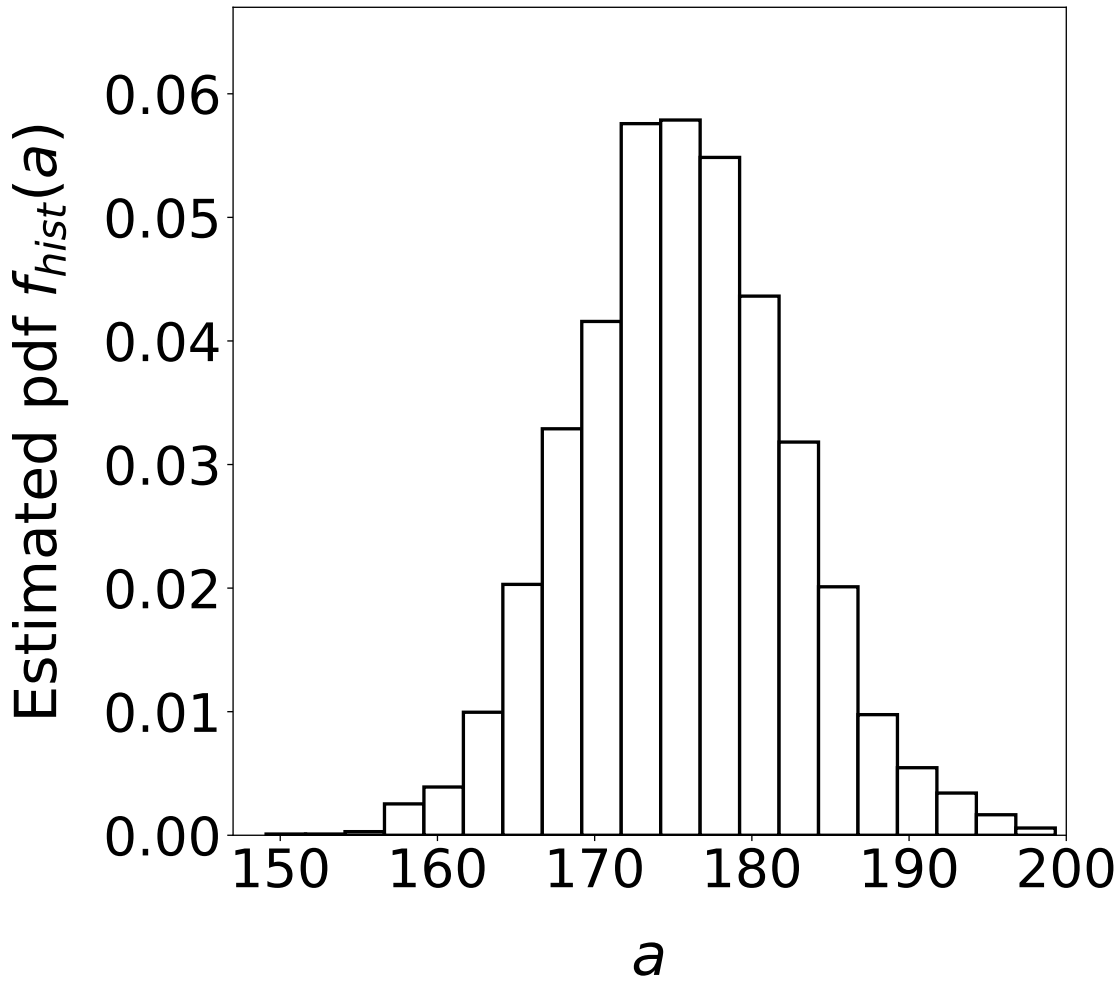
$$\Pi_j(t) = \begin{cases} \frac{b}{n\ell} & \text{for } t \in \mathcal{B}_j \\ 0 & \text{otherwise} \end{cases}$$

Total area? n rectangles with area $\frac{\ell}{b}\frac{b}{n\ell} = \frac{1}{n}$

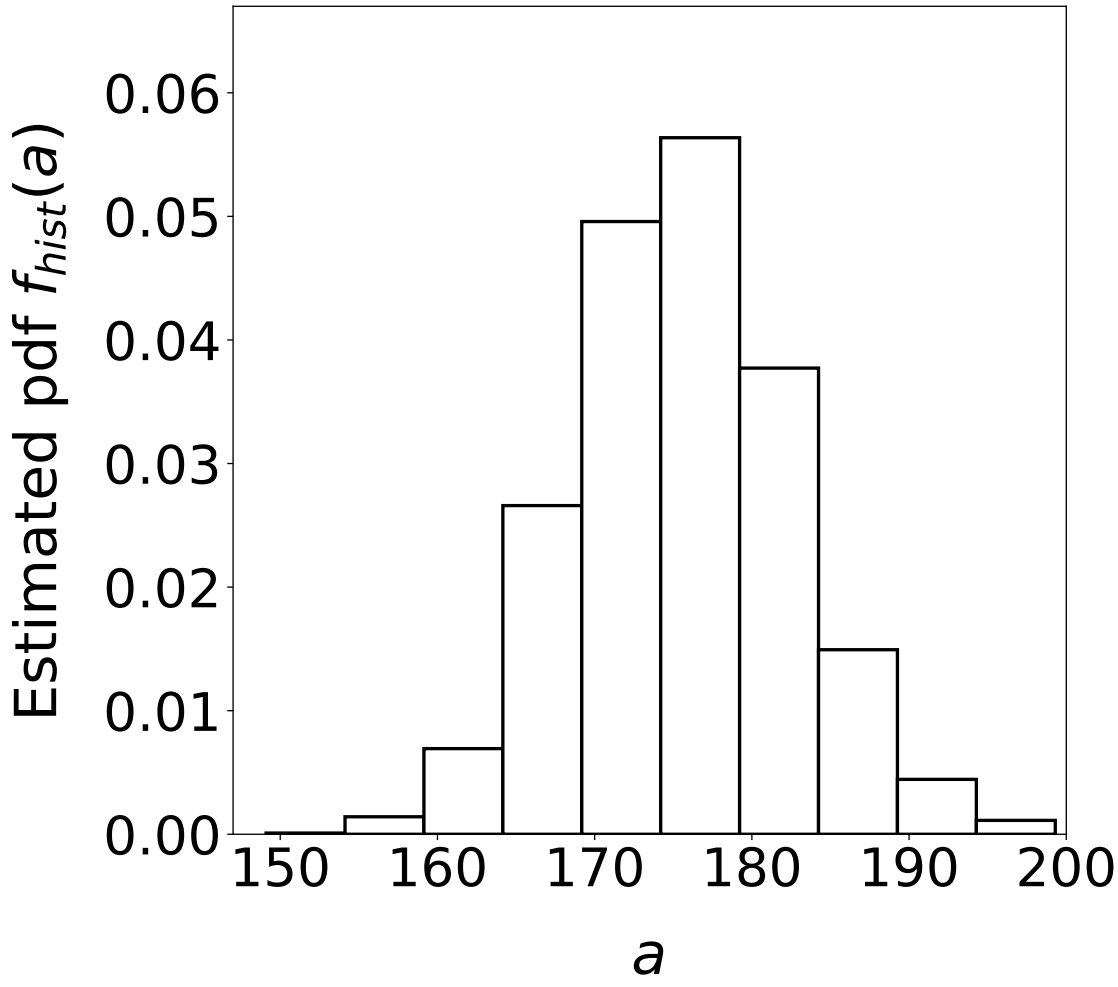
Height in US army: 50 bins



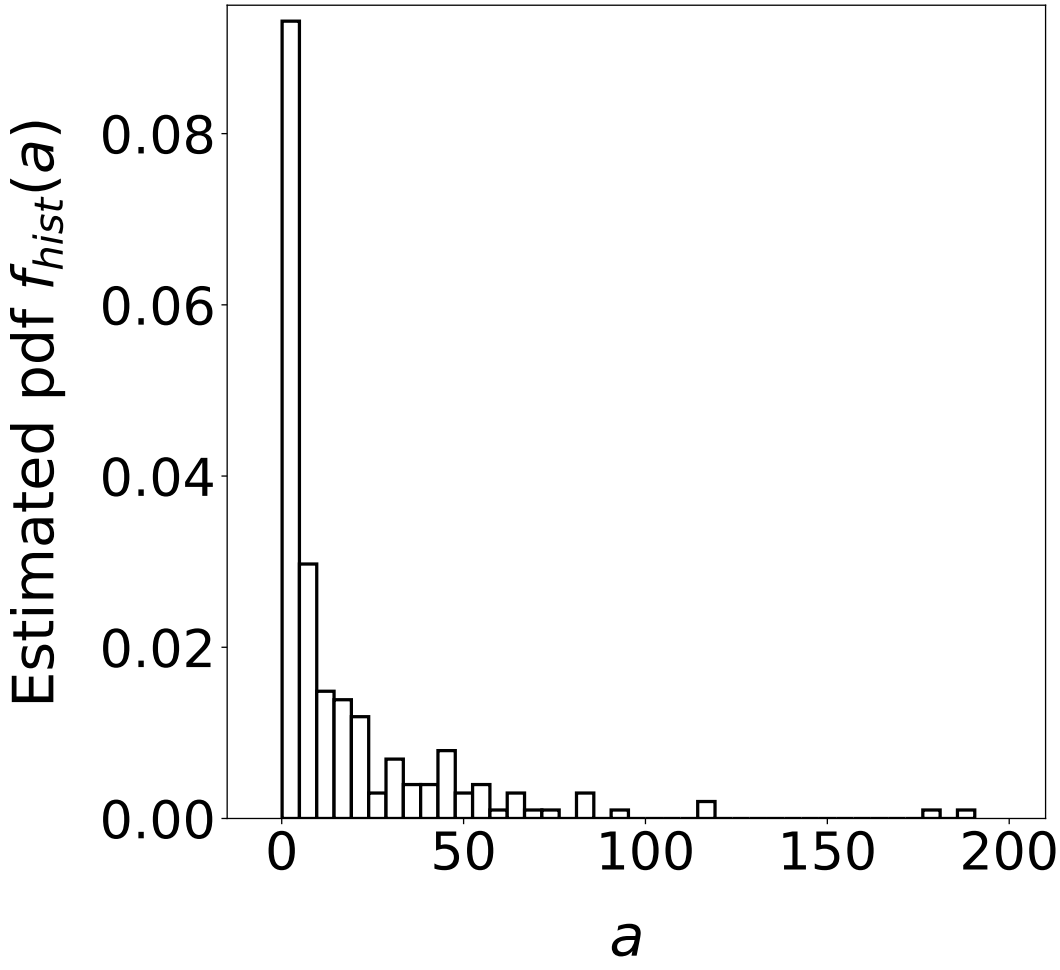
Height in US army: 20 bins



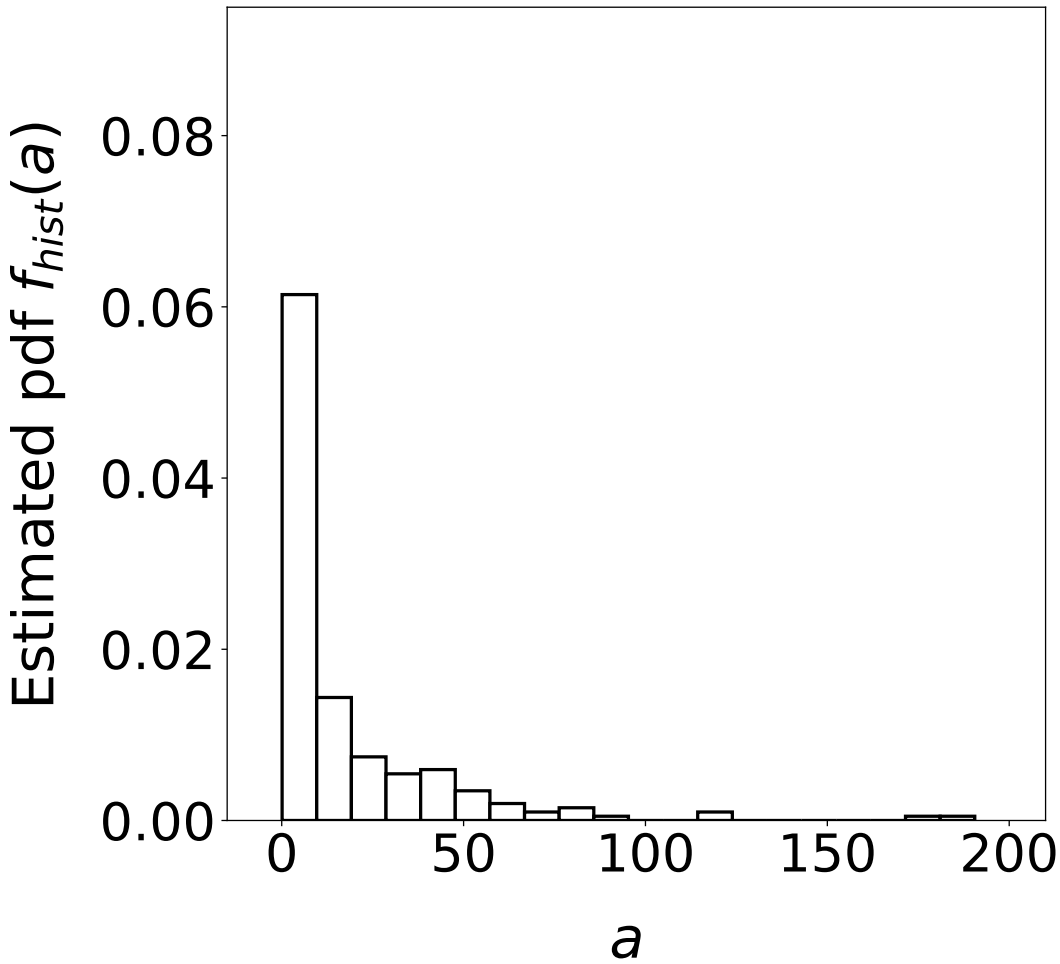
Height in US army: 10 bins



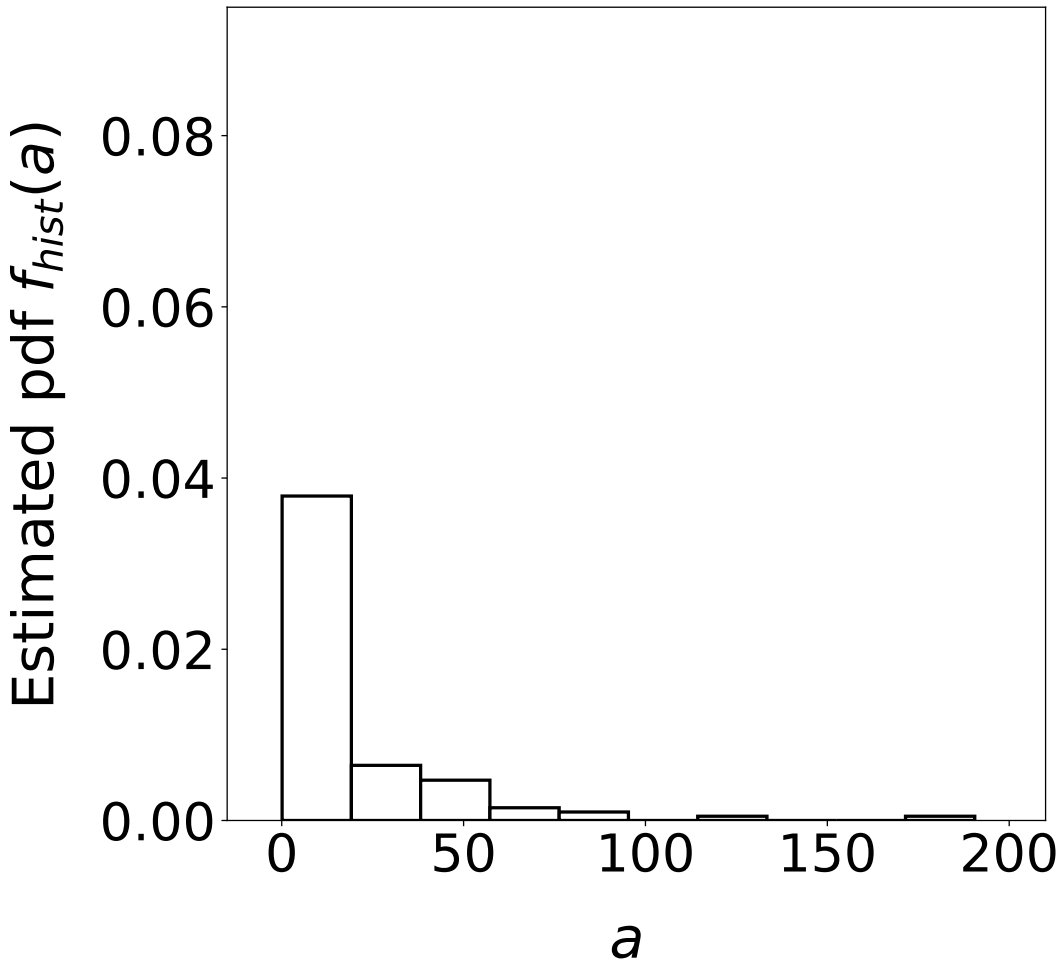
GDP per capita: 40 bins



GDP per capita: 20 bins



GDP per capita: 10 bins



Wait a minute

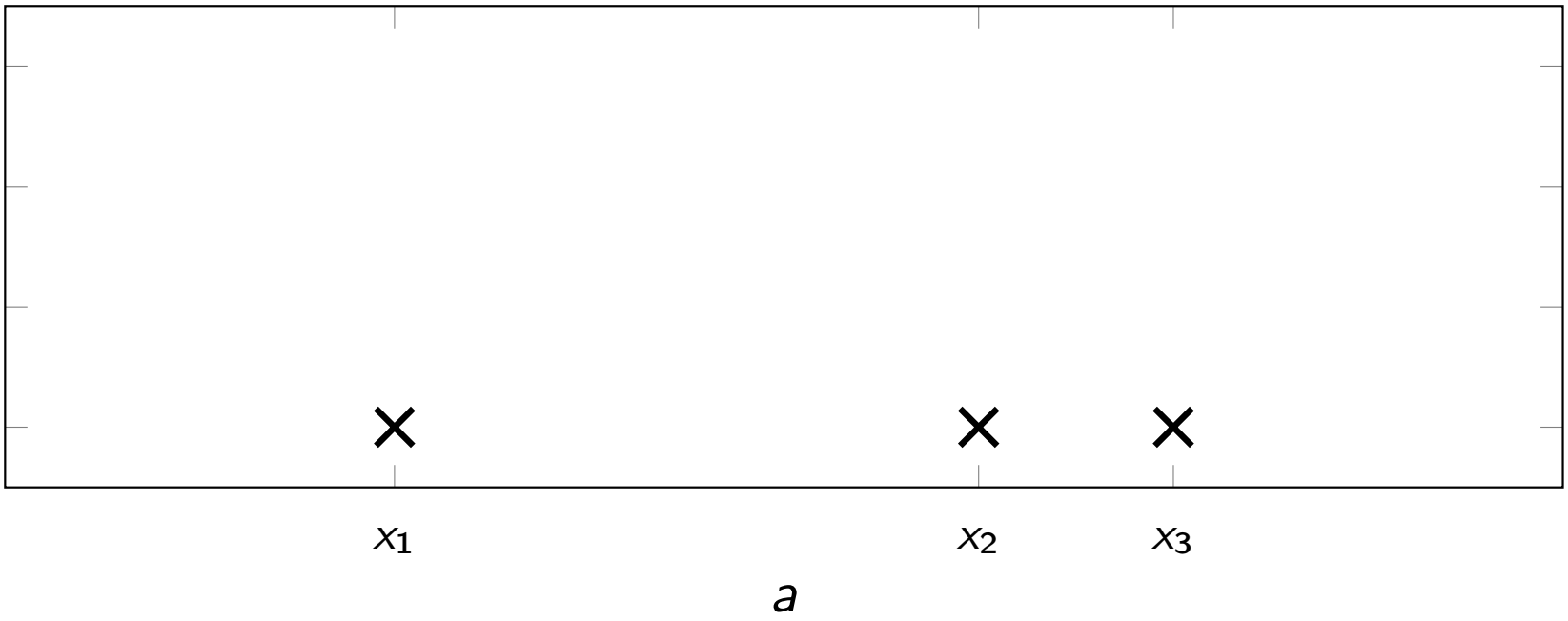
$$f_{\text{hist}}(t) = \sum_{j=1}^{n} \Pi_j(t)$$

Π_j: rectangle of length $\frac{\ell}{b}$ in bin $\mathcal{B}_j$ where x_j is located

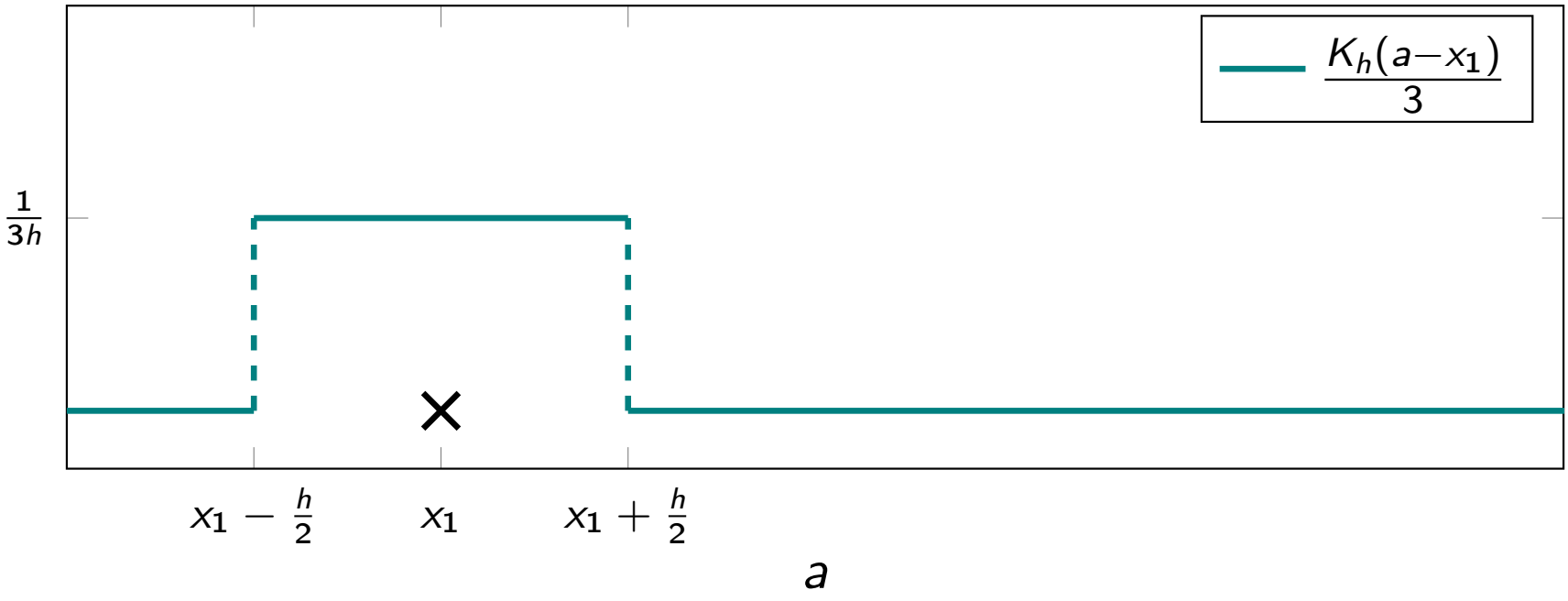
$$\Pi_j(t) = \begin{cases} \frac{b}{n\ell} & \text{for } t \in \mathcal{B}_j \\ 0 & \text{otherwise} \end{cases}$$

Shouldn't Π_j be centered at data point x_j?

Kernel density estimation

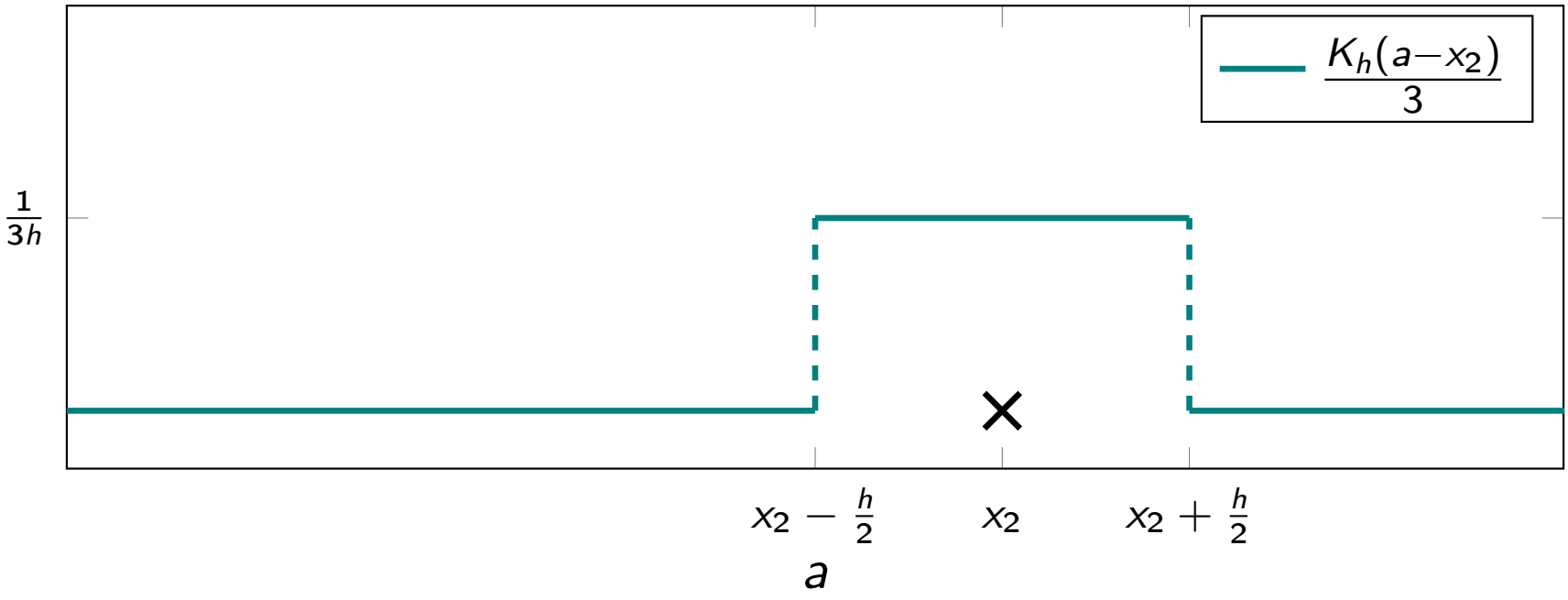


Kernel density estimation



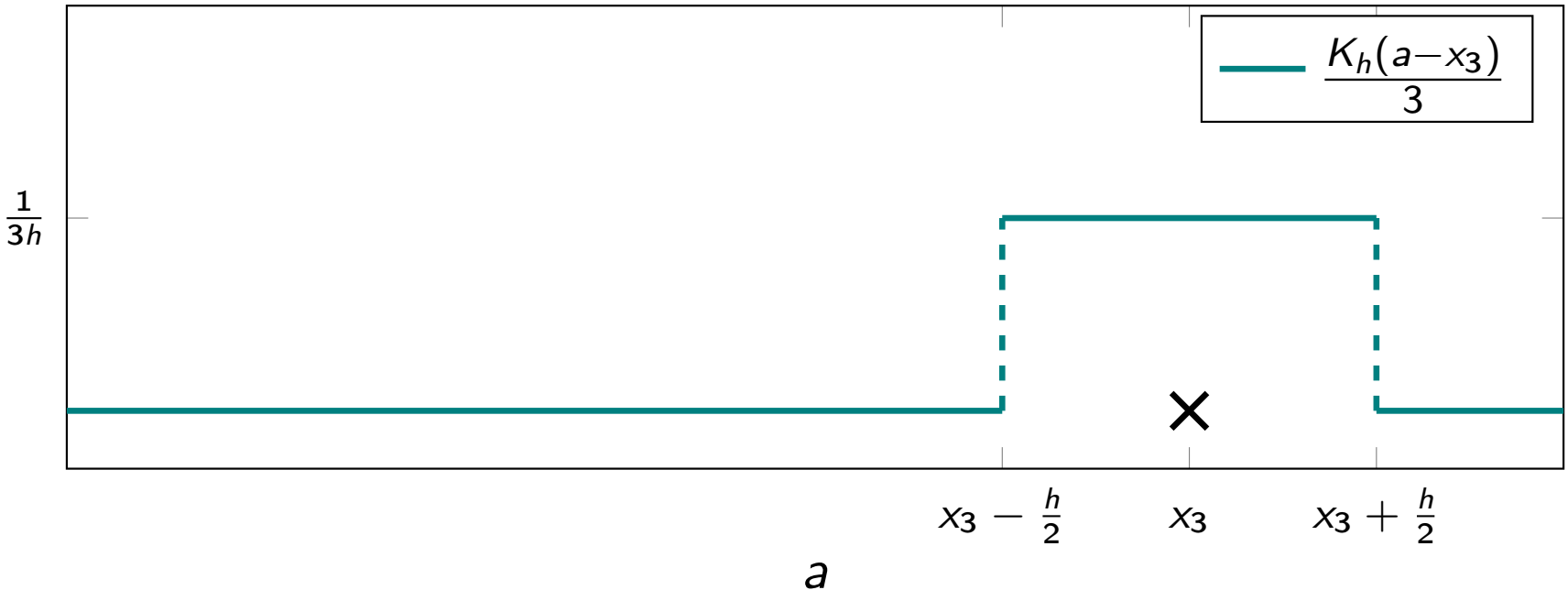
$$f_{X,h}(a) := \frac{1}{3h} K\left(\frac{a - x_1}{h}\right)$$

Kernel density estimation



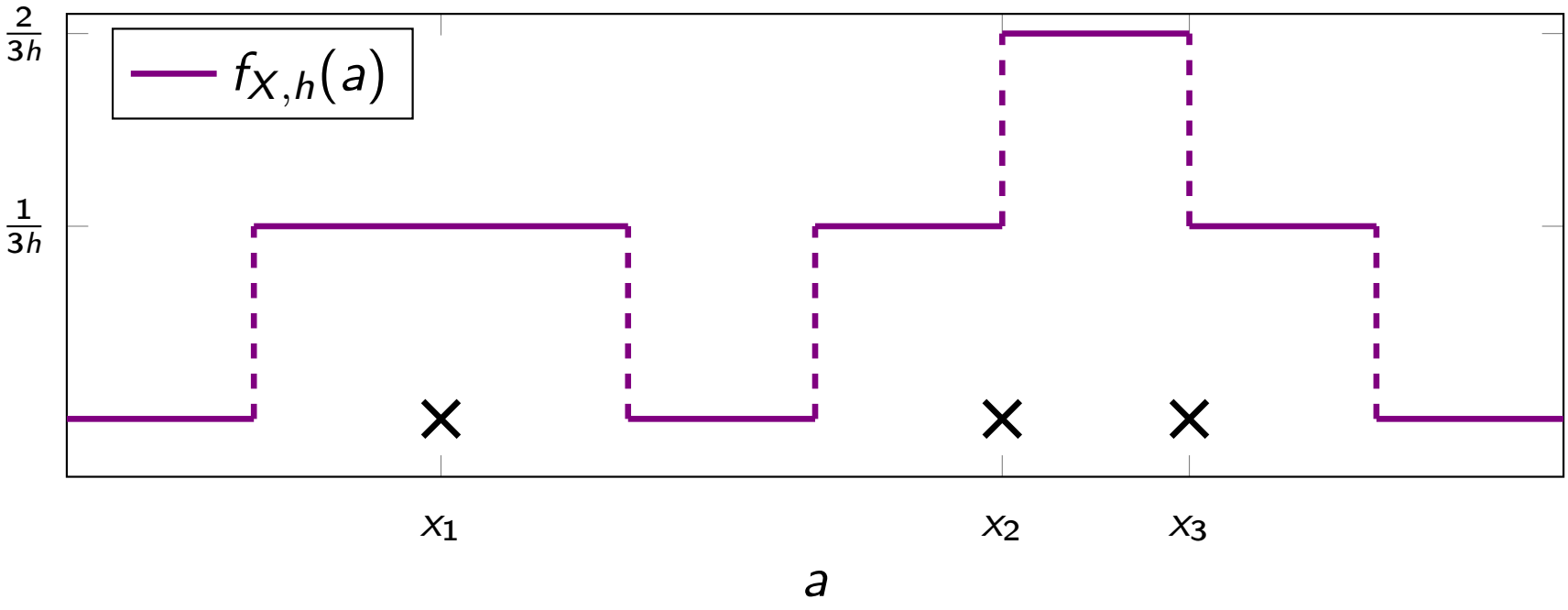
$$f_{X,h}(a) := \frac{1}{3h} K\left(\frac{a - x_1}{h}\right) + \frac{1}{3h} K\left(\frac{a - x_2}{h}\right)$$

Kernel density estimation



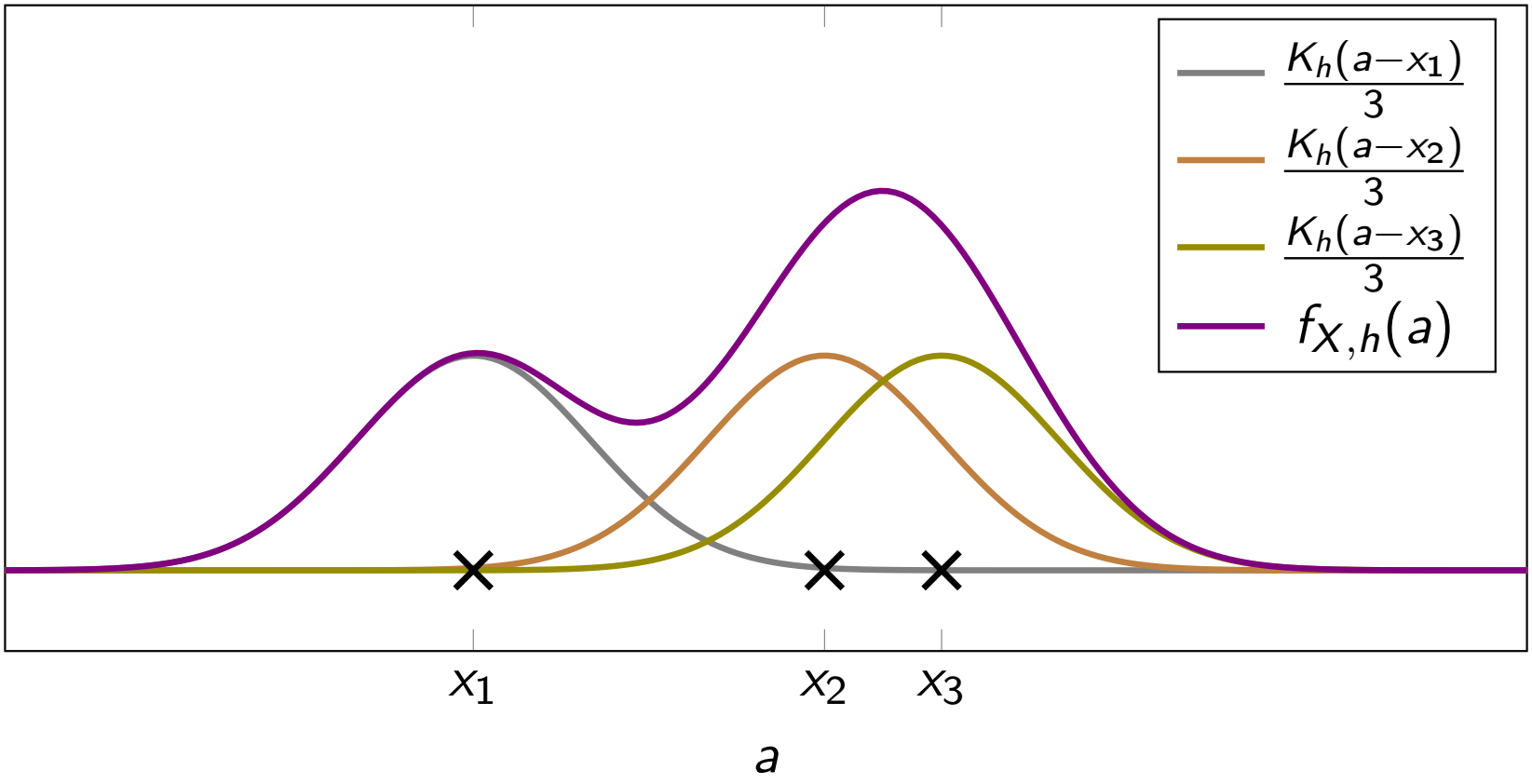
$$f_{X,h}(a) := \frac{1}{3h}K\left(\frac{a-x_1}{h}\right) + \frac{1}{3h}K\left(\frac{a-x_2}{h}\right) + \frac{1}{3h}K\left(\frac{a-x_3}{h}\right)$$

Kernel density estimation



$$f_{X,h}(a) := \frac{1}{3h} K\left(\frac{a - x_1}{h}\right) + \frac{1}{3h} K\left(\frac{a - x_2}{h}\right) + \frac{1}{3h} K\left(\frac{a - x_3}{h}\right)$$

Do we need to use rectangles?



$$f_{X,h}(a) := \frac{1}{3h}K\left(\frac{a-x_1}{h}\right) + \frac{1}{3h}K\left(\frac{a-x_2}{h}\right) + \frac{1}{3h}K\left(\frac{a-x_3}{h}\right)$$

Kernel density estimation

Data $X := \{x_1, x_2, \ldots, x_n\}$

Kernel density estimate with bandwidth h is

$$f_{X,h}(a) := \frac{1}{n\,h} \sum_{i=1}^{n} K\left(\frac{a - x_i}{h}\right)$$

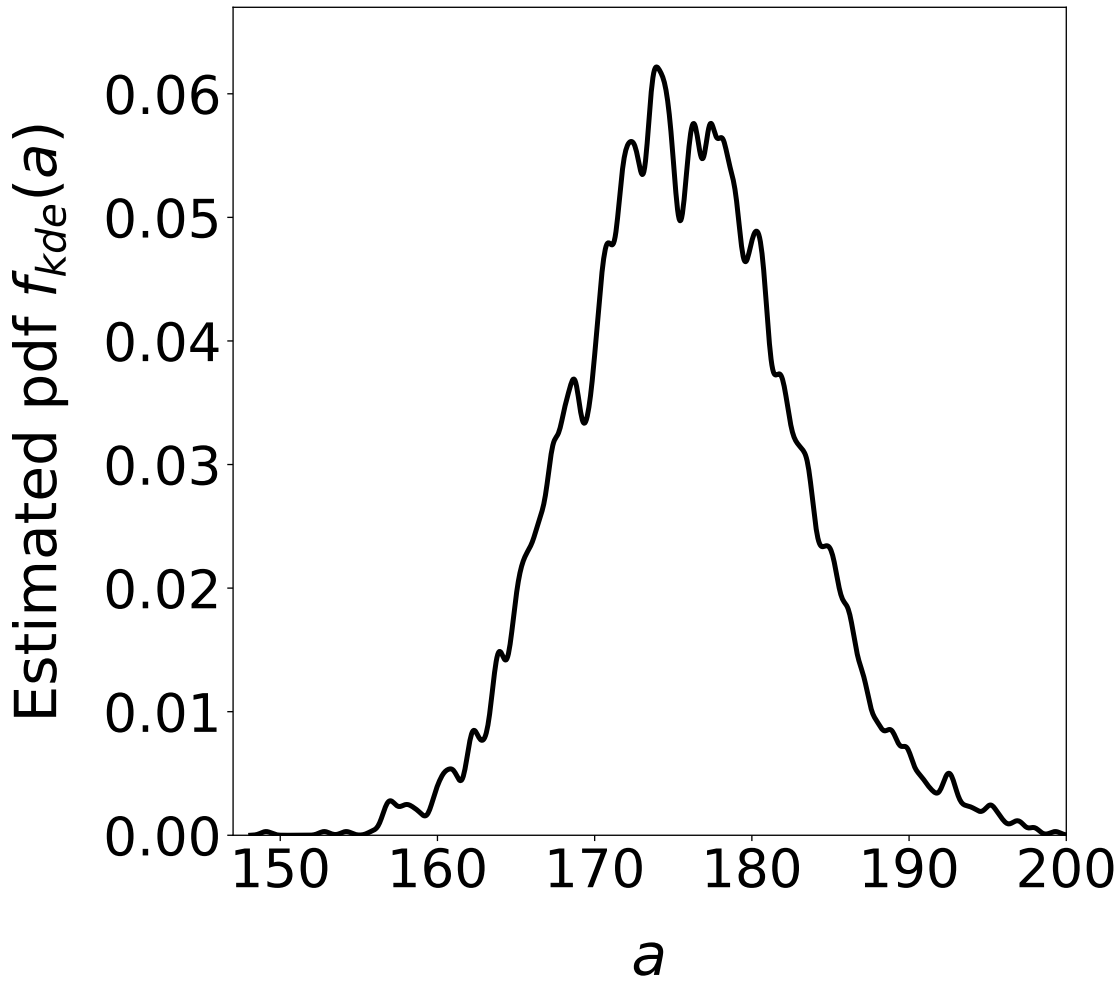
where K is a kernel that satisfies

$$K(a) \geq 0 \quad \text{for all } a \in \mathbb{R},$$
$$\int_{\mathbb{R}} K(a)\, \mathrm{d}x = 1$$

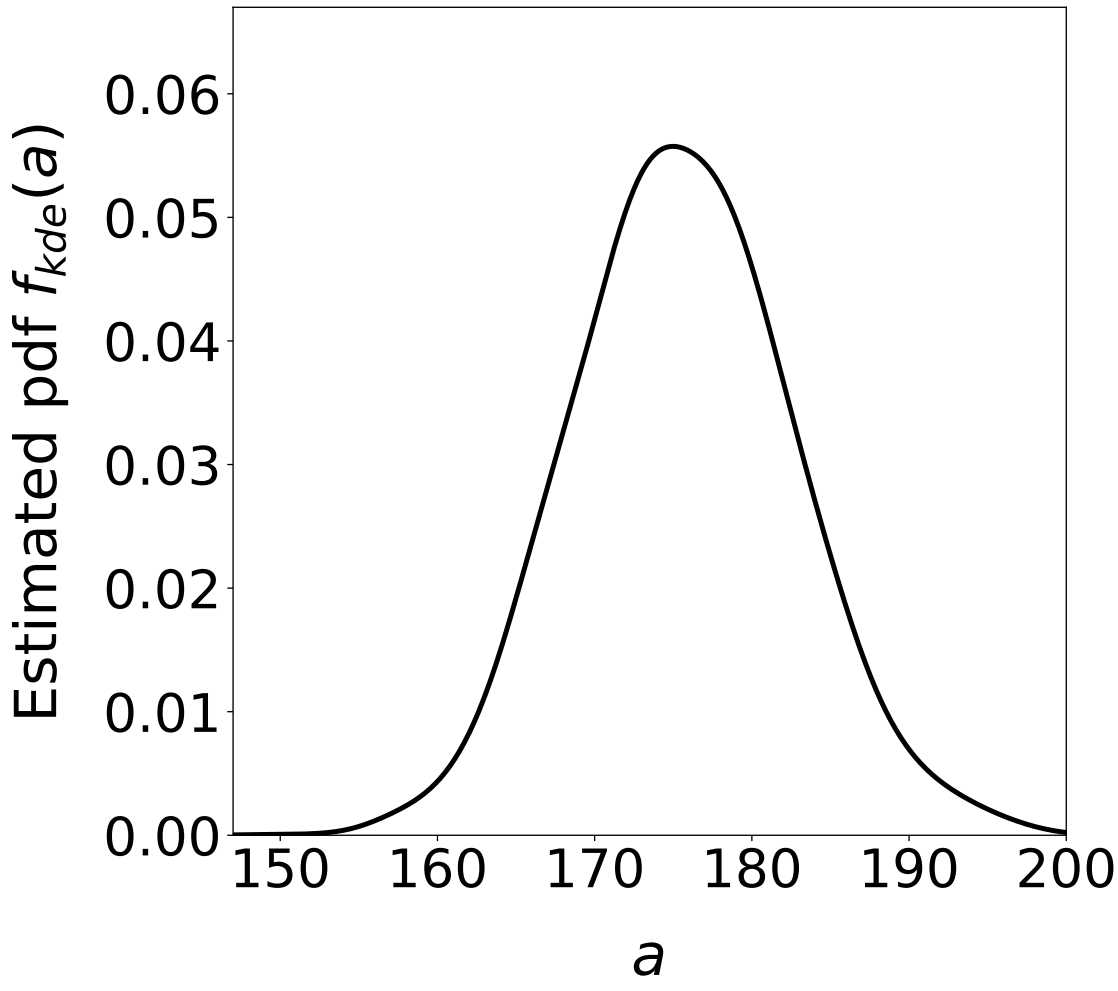
Estimate is composed of copies of the kernel centered at each data point

Is estimate a valid pdf?

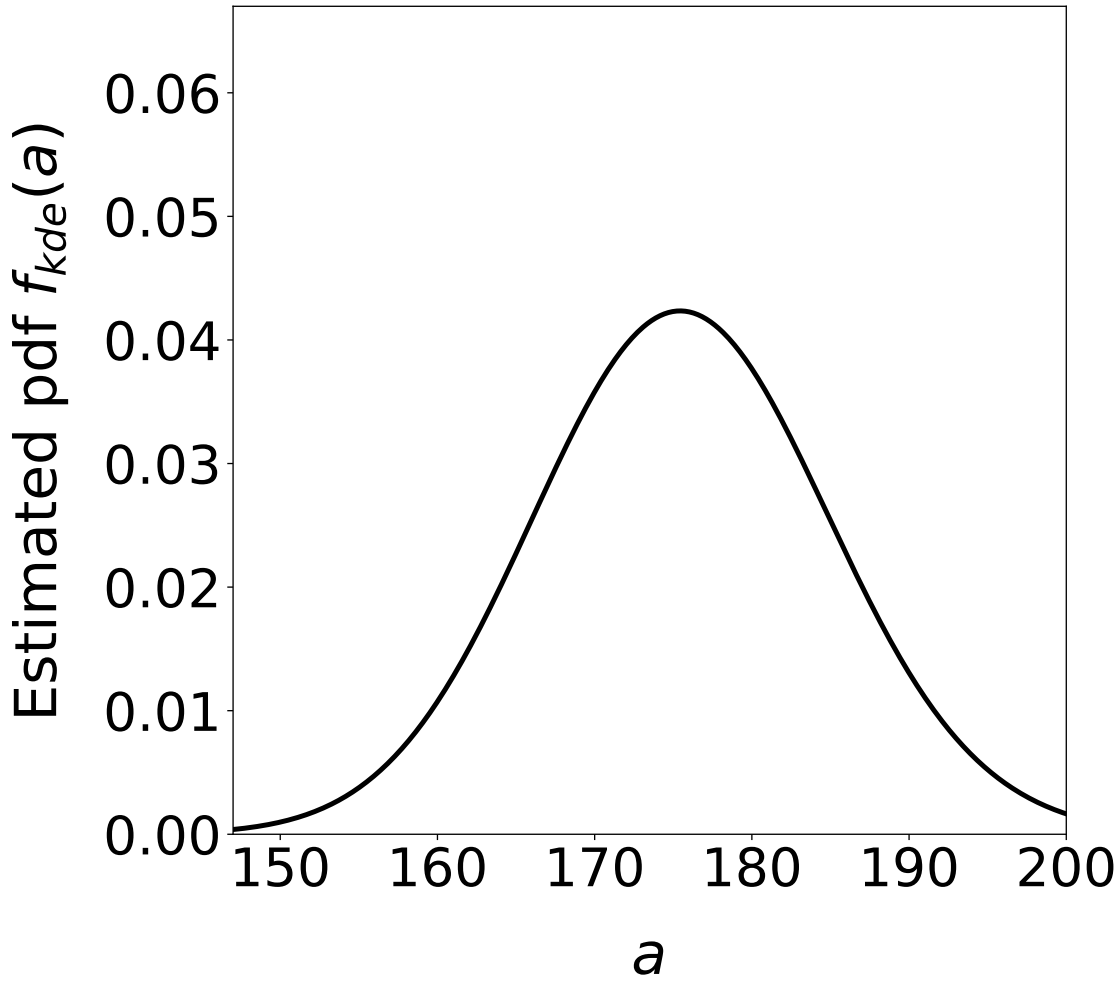
Height in US army: $h = 0.25$



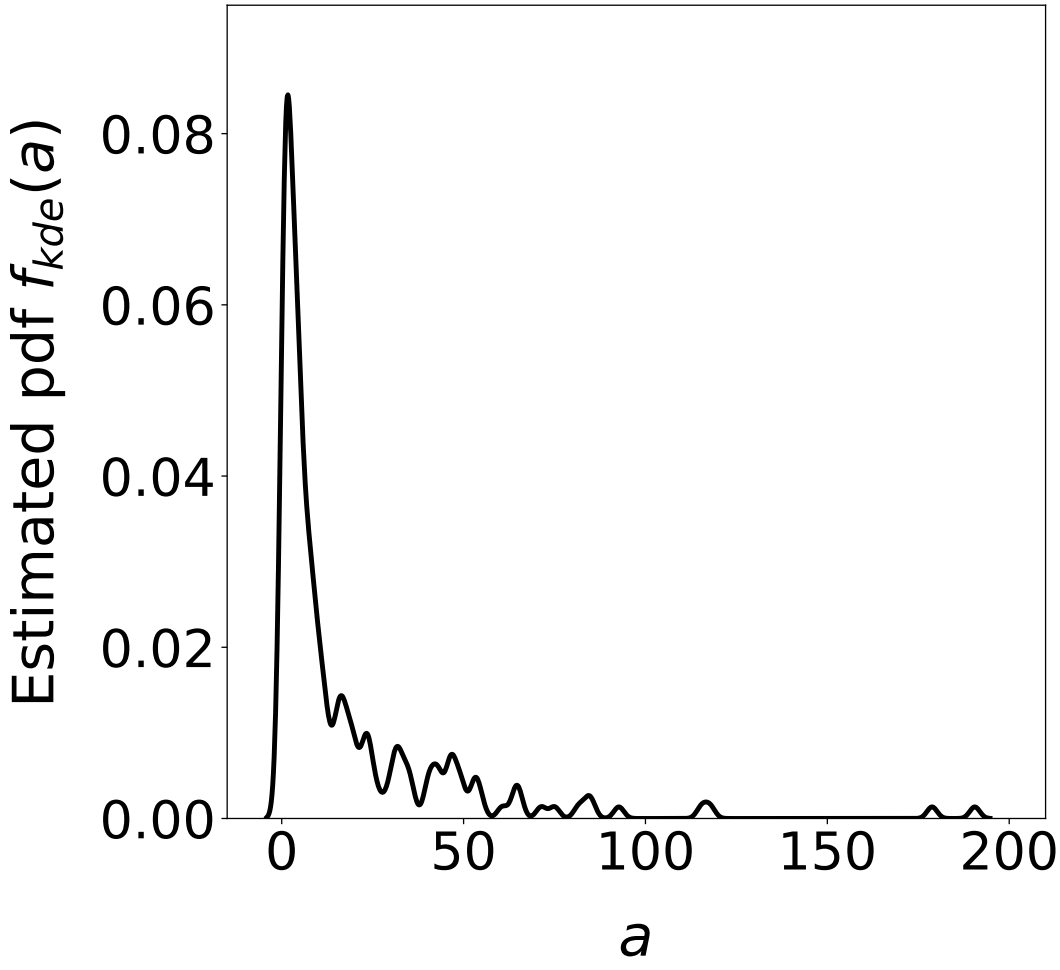
Height in US army: $h = 1.5$



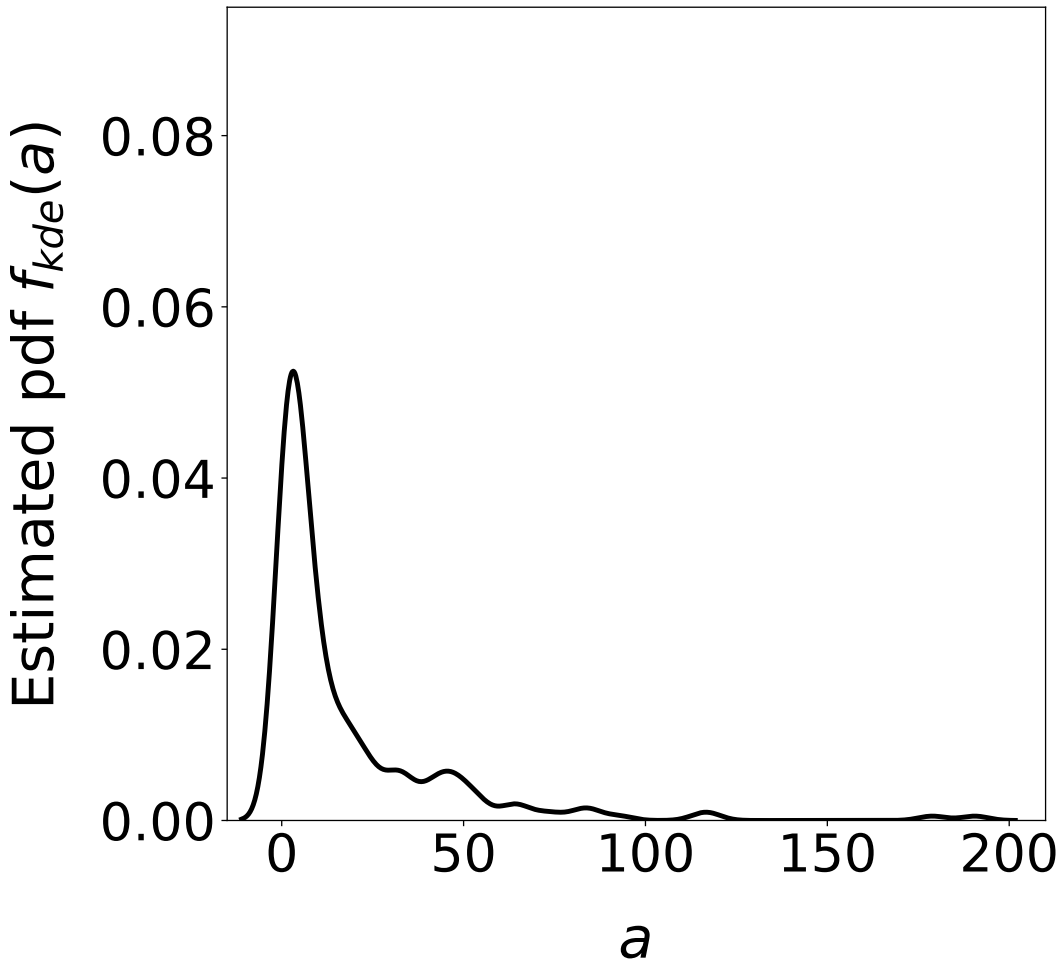
Height in US army: $h = 5$



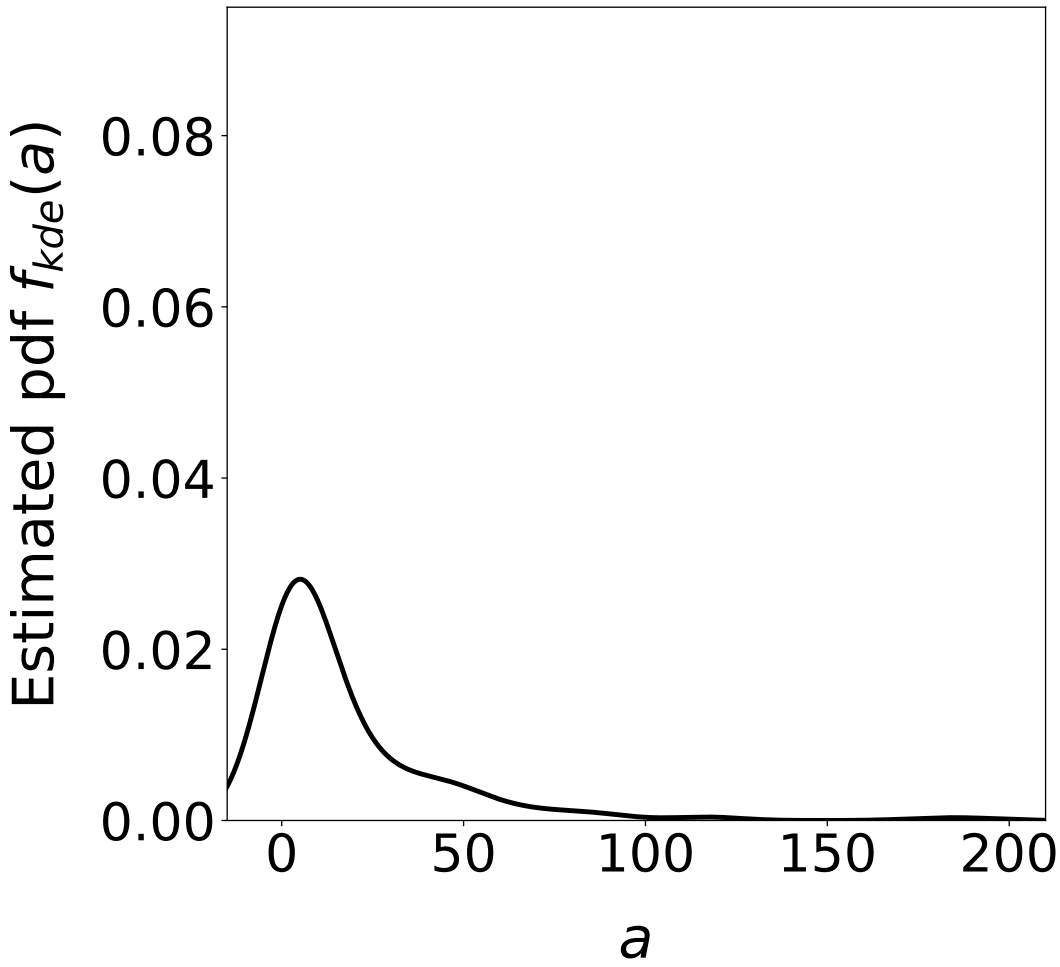
GDP per capita: $h = 1.5$



GDP per capita: $h = 3$



GDP per capita: $h = 6$



What have we learned?

How to estimate probability densities from data using the histogram and kernel density estimation