Modèles Linéaires Appliqués

Arthur Charpentier

Automne 2020

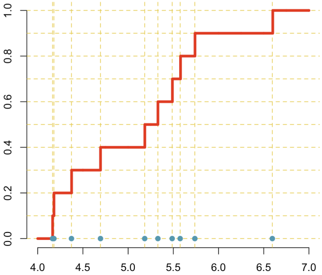
Rappels #3.3 (estimate F and f)

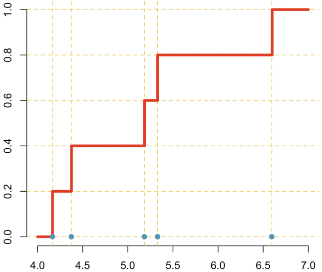
Cumulative Distribution Function

Given a random variable X, $F(x)$, i.e. $x \mapsto \mathbb{P}[X \leq x]$ is an increasing function, taking values in $[0,1]$.
Consider a sample $\mathbf{x} = \{x_1, x_2, \cdots, x_n\}$, a natural estimator is

$$\widehat{F}(x) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(x_i \leq x)$$

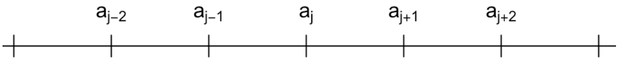
```
> sample_x = sort(sample_x)
> n = length(sample_x)
> y = (1:n)/n
> plot(ecdf(sample_x))
> Fhat = function(x)
    mean(sample_x <= x)
```





Density & Histogram

Given a random variable X, f is such that $F(x) = \int_{-\infty}^{x} f(t)dt$
or conversely, $f(x) = F'(x)$.
But we cannot define $\widehat{f}(x) = \widehat{F}'(x)$
For an histogram, consider $a_0 \leq a_1 \leq \cdots \leq a_{k-1} \leq a_k$ so that
$\forall i,\ x_i \in (a_0, a_k)$, and $\forall j,\ a_{j+1} - a_j = h$ (constant).



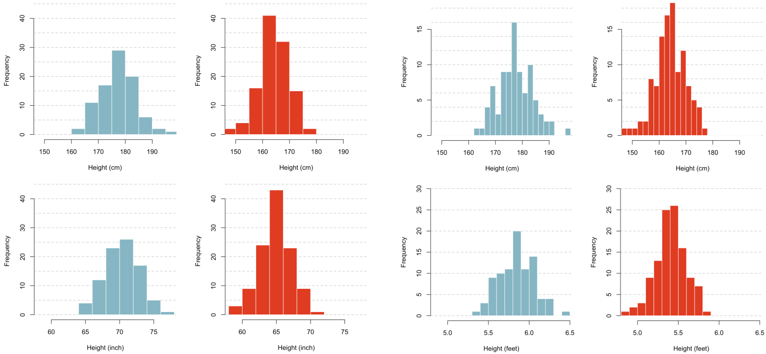
$$\text{if } x \in [a_j, a_{j+1}),\ \widehat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} \mathbf{1}_{[a_j, a_{j+1})}(x)$$

Problem: very sensitive to a_0 and h...

Density & Histogram

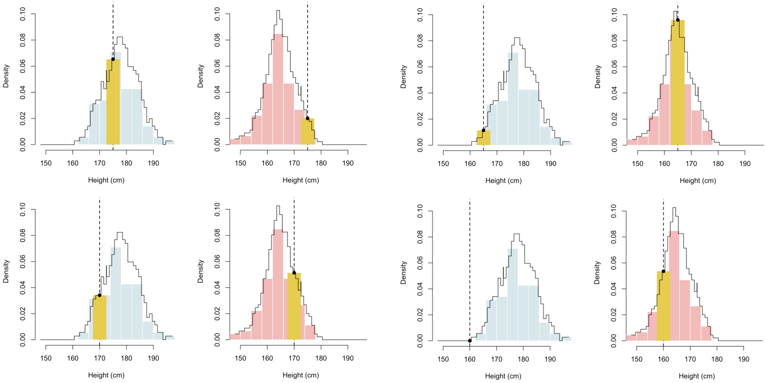
$$\text{if } x \in [a_j, a_{j+1}),\ \widehat{f}(x) = \underbrace{\frac{1}{nh} \sum_{i=1}^{n} \mathbf{1}_{[a_j, a_{j+1})}(x_i)}_{\text{histogram}}$$

Here, $\displaystyle\int_{a_0}^{a_k} \widehat{f}(x)dx = \int_{\mathbb{R}} \widehat{f}(x)dx = 1.$



Moving Histogram

$$\widehat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} \mathbf{1}(|x_i - x| \leq h/2)$$



$\widehat{F}$ cannot be differentiated, but we can consider

$$f_h(x) = \frac{1}{h}\Big[\underbrace{F(x+h/2) - F(x-h/2)}_{\mathbb{P}(X\in[x\pm h/2])}\Big]$$

i.e.

$$\widehat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} \mathbf{1}\left(x_i \in [x-h/2, x+h/2]\right)$$

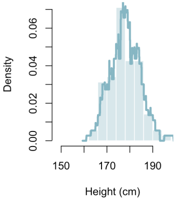
$$\widehat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} \mathbf{1}(|x_i - x| \leq h/2)$$

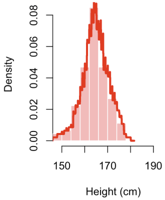
One can prove that $\mathbb{E}(\widehat{f}_h(x)) = f_h(x) \sim f(x) + \dfrac{h^2}{24} f''(x)$

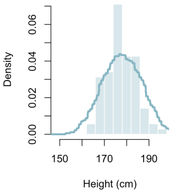
i.e. $\text{bias}(\widehat{f}_h(x)) \sim \dfrac{h^2}{24} f''(x)$, while $\text{Var}(\widehat{f}_h(x)) \sim \dfrac{1}{nh} \cdot f_h(x)$

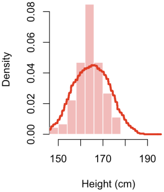
Moving Histogram

$$\widehat{f_h}(x) = \frac{1}{nh}\sum_{i=1}^{n} \mathbf{1}(|x_i - x| \le h/2)$$







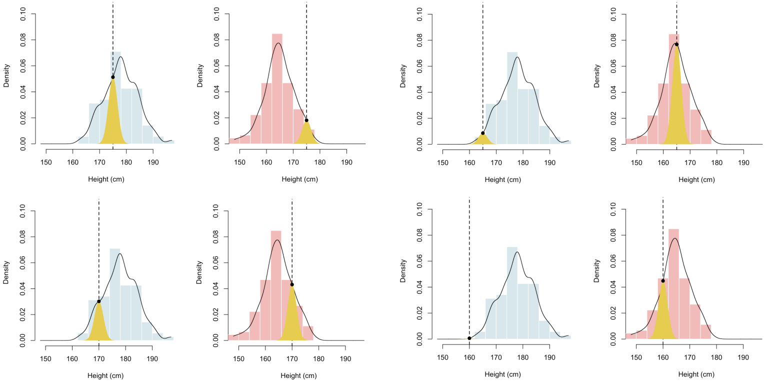


small h
bias $\text{bias}(\widehat{f_h}(x))$ small
variance $\text{Var}(\widehat{f_h}(x))$ large

large h
bias $\text{bias}(\widehat{f_h}(x))$ large
variance $\text{Var}(\widehat{f_h}(x))$ small

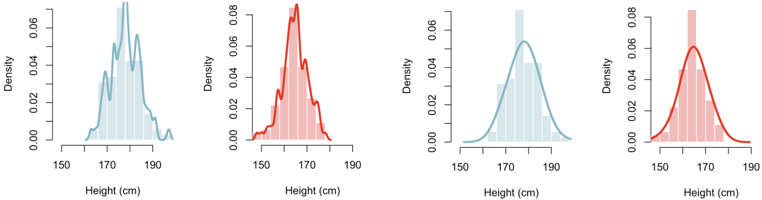
Kernel Density

$$\widehat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x_i - x}{h}\right)$$



Kernel Density

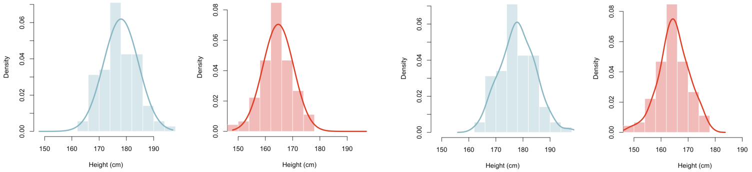
$$\widehat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x_i - x}{h}\right)$$



small h
bias $\text{bias}(\widehat{f}_h(x))$ small
variance $\text{Var}(\widehat{f}_h(x))$ large

large h
bias $\text{bias}(\widehat{f}_h(x))$ large
variance $\text{Var}(\widehat{f}_h(x))$ small

Histogram & Density



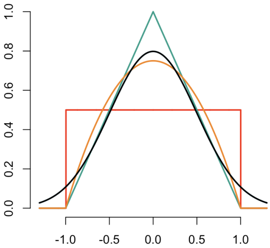

```
> hist(x, probability=TRUE)
> plot(density(x))
> plot(density(x), kernel="gaussian", bw=1)
```

Rectangle: $k(u) = \dfrac{1}{2}\mathbf{1}_{[-1,+1]}(u)$

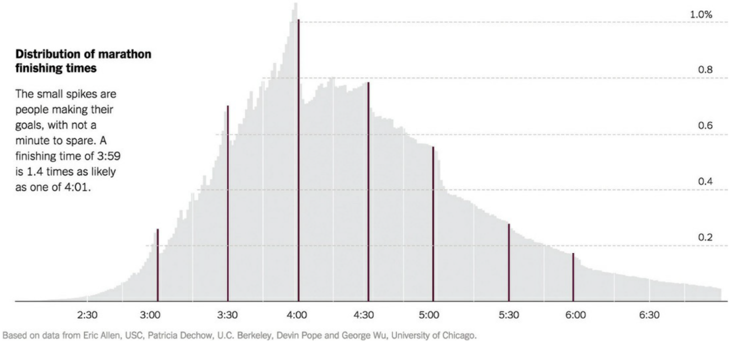
Triangle: $k(u) = (1-|u|)_+$

Epanechnikov: $k(u) = \dfrac{3}{4}(1-u^2)_+$

Gaussian: $k(u) = \dfrac{1}{\sqrt{2\pi}}\exp\left(-\dfrac{u^2}{2}\right)$



Histogram & Density



via Reference-Dependent Preferences: Evidence from Marathon Runners