STT 1000 - STATISTIQUES

ARTHUR CHARPENTIER

Cumulative Distribution Function

Given a random variable X, $F(x) = \mathbb{P}[X \leq x]$
F is an increasing function, taking values in $[0, 1]$.

Fonction de répartition empirique $\widehat{F}$

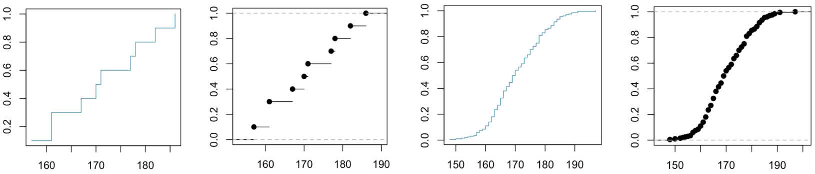
Consider a sample $\boldsymbol{x} = \{x_1, y_2, \cdots, x_n\}$, a natural estimator is the empirical cumulative distribution function $\widehat{F}$

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

On peut prouver que $\mathbb{E}[\widehat{F}(x)] = F(x)$.

Cumulative Distribution Function

```
> x = sort(x)
> n = length(x)
> y = (1:n)/n
> plot(x,y,type="s")
> plot(ecdf(x))
```



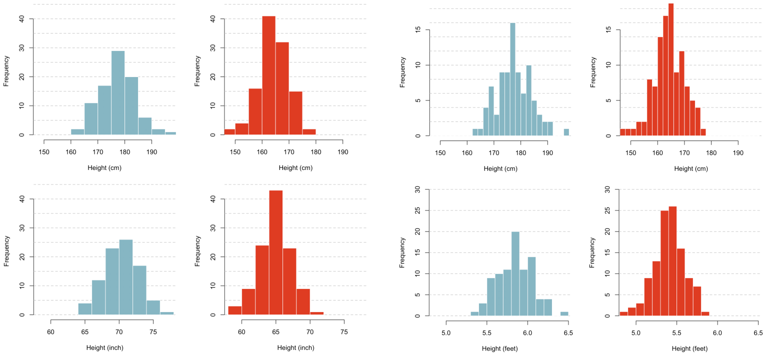
La fonction $x \mapsto \widehat{F}(x)$ est une fonction en escalier, qui fait un saut de $1/n$ dès qu'elle croise une observation x_i.

Density & Histogram

Given a random variable X, f is such that $F(x) = \int_{-\infty}^{x} f(t)dt$

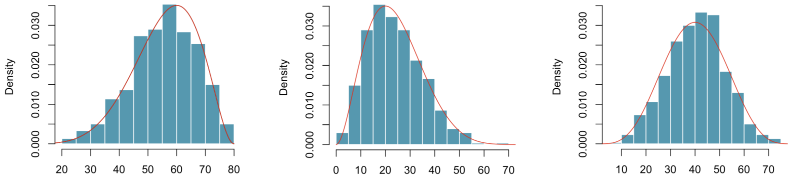
or conversely, $f(x) = F'(x)$.

Thus, $\mathbb{P}(X \in [a,b]) = \int_a^b f(t)dt$



Density & Histogram

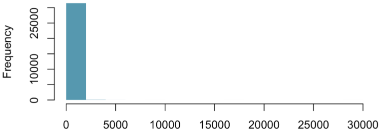
On dit qu'une distribution est unimodale si elle ne possède qu'un pic majeur.
Quand une distribution n'est pas symétrique, elle est dite asymétrique ; on dit qu'une distribution est asymétrique à droite si l'aile (queue) droite de la distribution est plus longue que l'aile gauche.

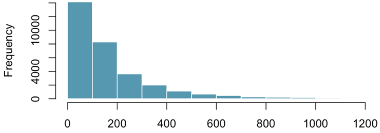


Histogram

Données de Larry Brown et Haipeng Shen, durée des appels au service à la clientèle d'une banque pendant un mois : 31,492 appels

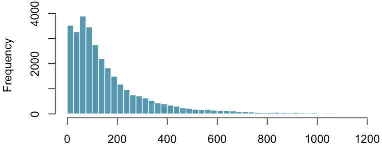
```
> hist(bankcall$Time)
> hist(bankcall$Time[bankcall$Time<1200])
```

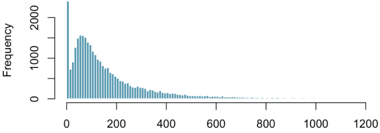




On peut se restreindre aux 31,247 appels de moins de 20 minutes

```
> hist(bankcall$Time[bankcall$Time<1200], breaks=seq
    (0,1200,by=10))
```

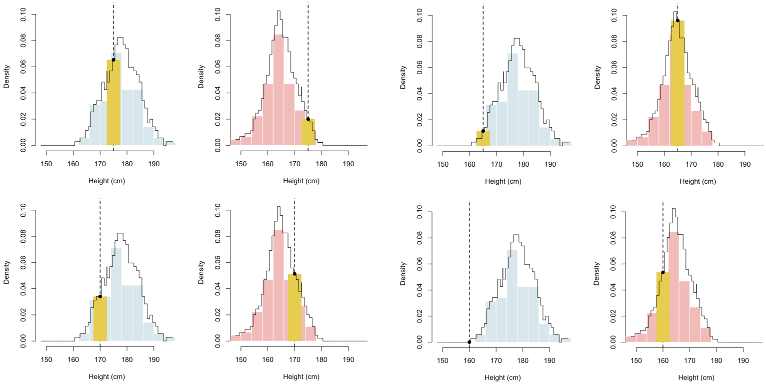




Moving Histogram

Histogramme glissant $\widehat{f}$

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



Moving Histogram

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

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

i.e.

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

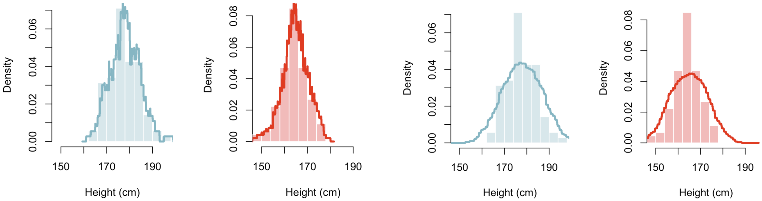
$$\widehat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} \mathbf{1}(|x_i - x| \leq h/2)$$

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

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

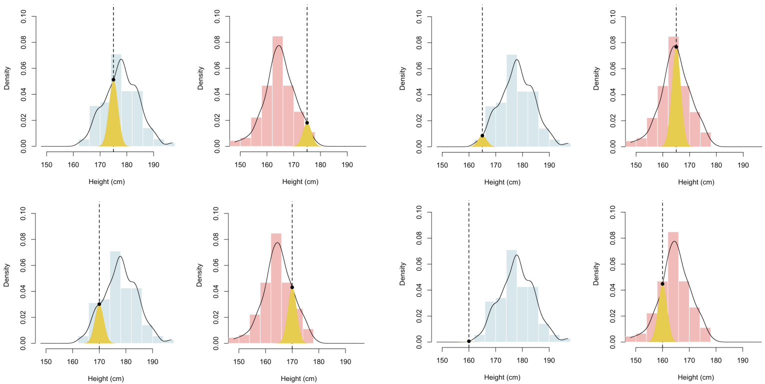


small h
bias bias$(\widehat{f}_h(x))$ small
variance Var$(\widehat{f}_h(x))$ large

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

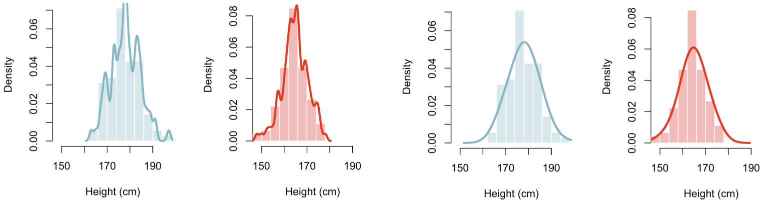
Kernel Density

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



Kernel Density

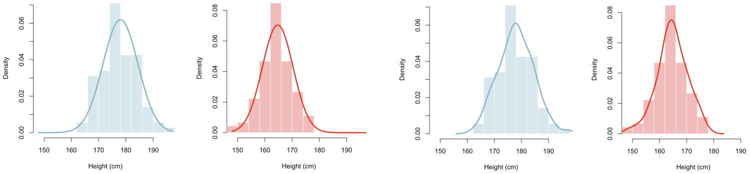
$$\widehat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x_i - x}{h}\right)$$



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

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

Histogram & Density




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