Data Science for Actuaries (ACT6100)

Arthur Charpentier

\# 05 - Densités, histogrammes et fonctions de répartition

été 2022

Fonction de répartition I

Pour une variable aléatoire X, on note $F(x) = \mathbb{P}[X \leq x]$ sa fonction de répartition. F est croissante, et à valeurs dans $[0, 1]$.

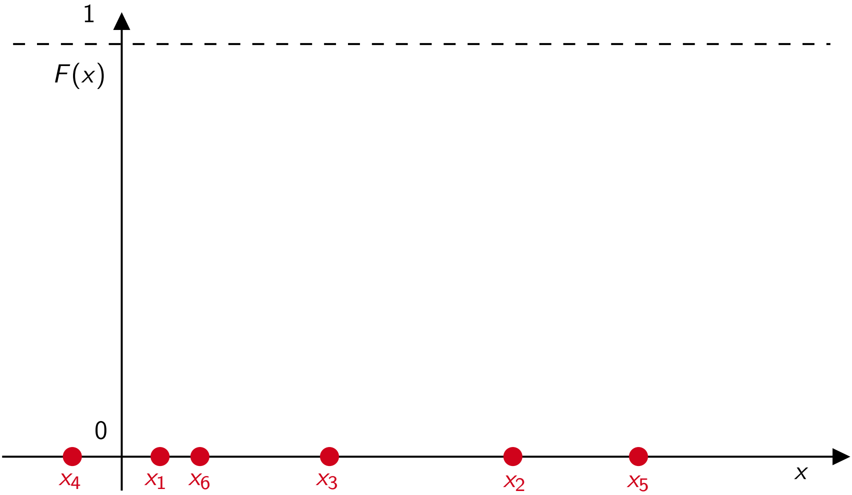
Fonction de répartition empirique $\widehat{F}$

Consider a sample $\boldsymbol{x} = \{x_1, x_2, \cdots, x_n\}$, a natural estimator de F is the empirical cumulative distribution function $\widehat{F}$

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

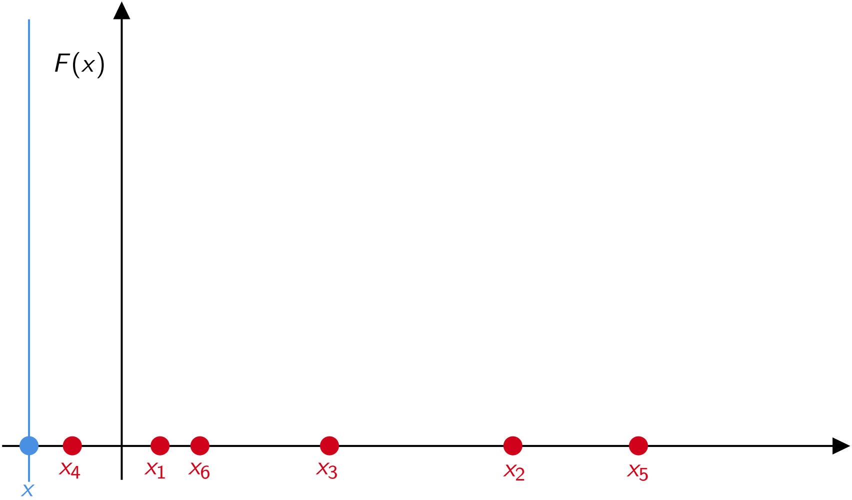
Fonction de répartition II

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



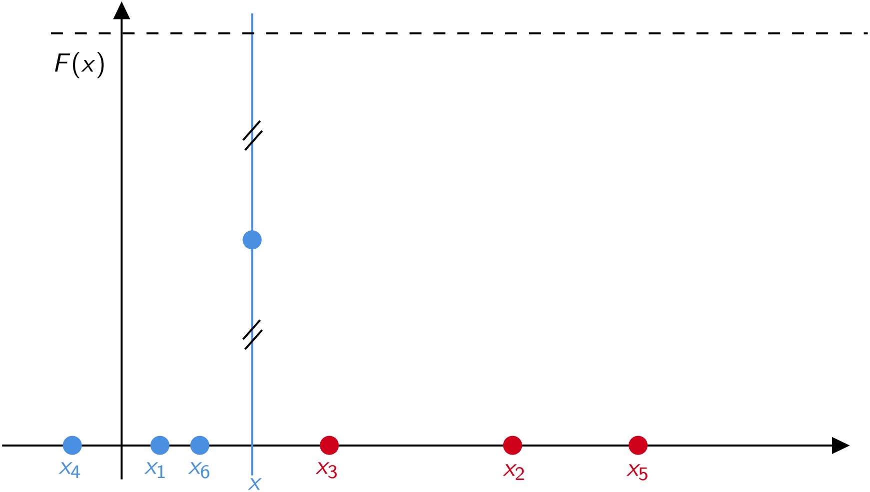
Fonction de répartition III

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



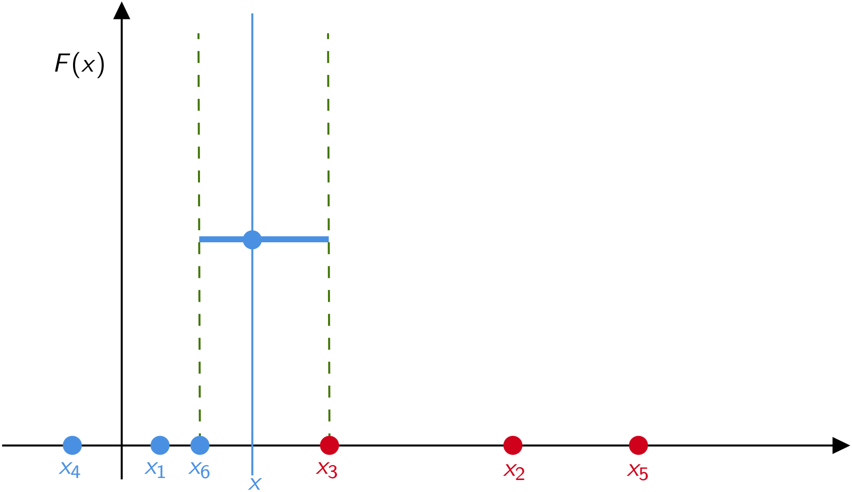
Fonction de répartition IV

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



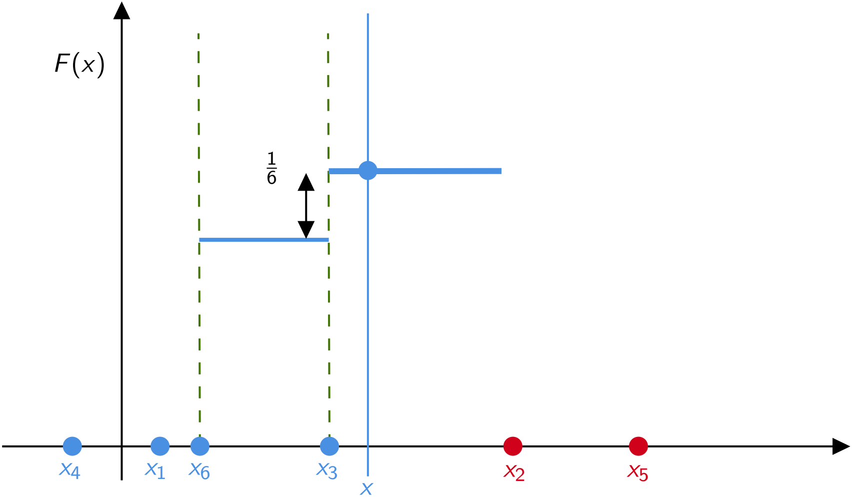
Fonction de répartition V

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



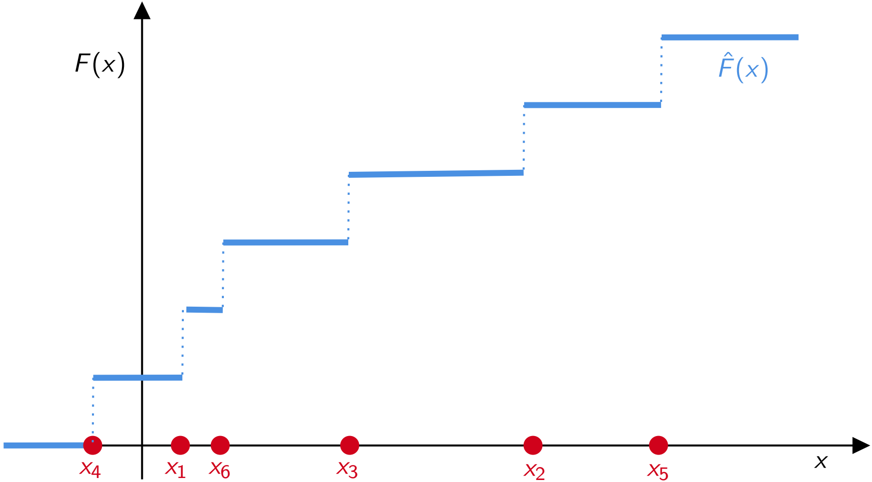
Fonction de répartition VI

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



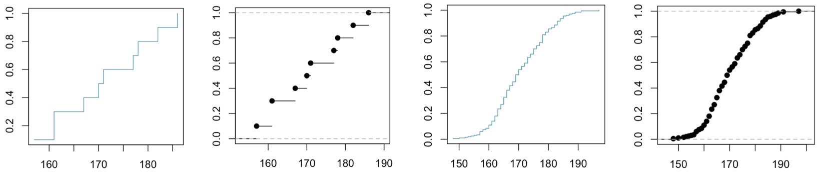
Fonction de répartition VII

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



Cumulative Distribution Function

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



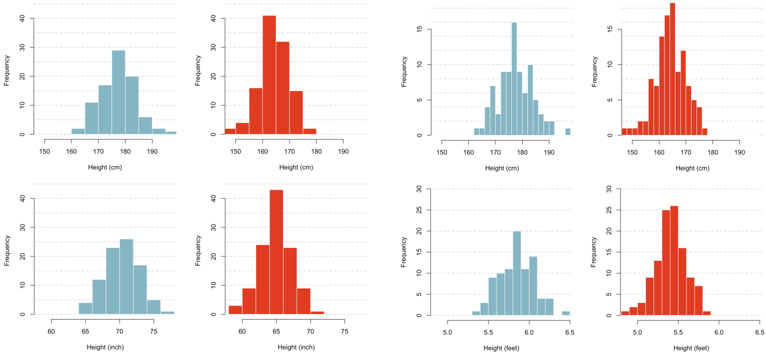
La fonction $x \mapsto \widehat{F}(x)$ est une fonction en escalier, qui fait un saut de $1/n$ dès qu'elle croise une observation x_i.

Densité et Histogramme I

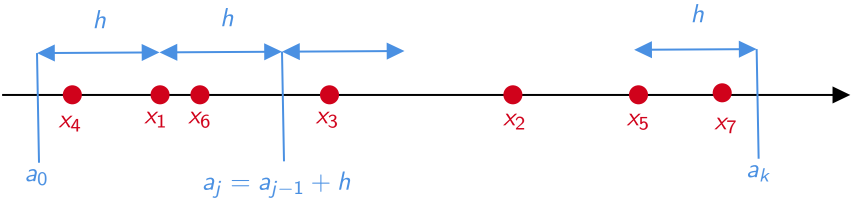
Given a random variable X, f is such that $F(x) = \int_{-\infty}^{x} f(t)dt$

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

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



Densité et Histogramme II



Histogramme et estimation de densité $\widehat{f}$

Étant donné un échantillon $\boldsymbol{x} = \{x_1, x_2, \cdots, x_n\}$, soient

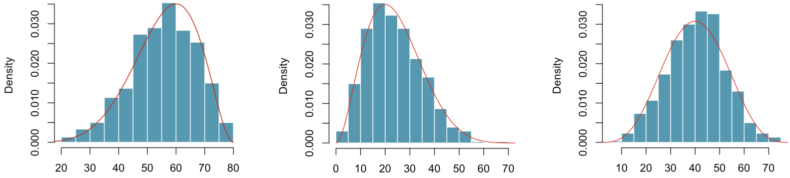
$$\begin{cases} a_0 < \min\{x_1, x_2, \cdots, x_n\}, \;\; a_k > \max\{x_1, x_2, \cdots, x_n\} \\ a_{j+1} = a_j + h = a_0 + (j+1)h \end{cases}$$

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

Densité et Histogramme III

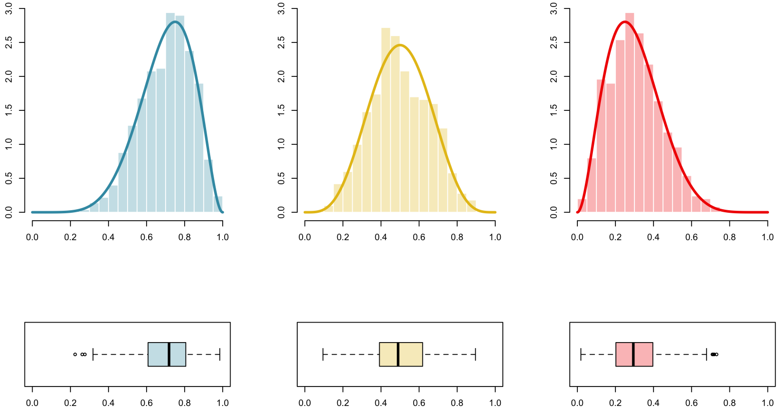
```
> hist(x)
> hist(x, probability=TRUE)
```

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.



Densité et Histogramme IV

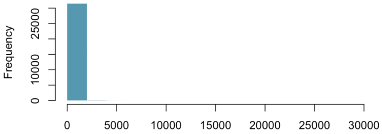
```
> hist(x)
> boxplot(x)
```

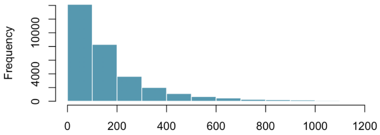


Histogramme I

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])
```

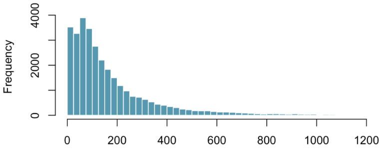


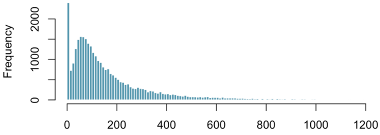


Histogramme II

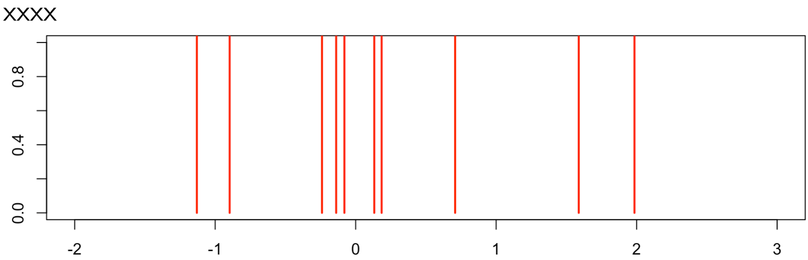
On peut se restreindre aux 31,247 appels de moins de 20 minutes

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





Densité ? I



Soit k une fonction symétrique positive centrée sur zéro et intégrant à 1, c'est à dire une densité d'une variable de moyenne nulle), et K la fonction de répartition associée.

Densité ? II

Noyau k

Un noyau k est une fonction de densité centrée sur 0, i.e. $\displaystyle\int_{-\infty}^{+\infty} xk(x)dx = 0$.

Définissons $k_h(x)$ la fonction obtenue après changement d'échelle,

$$k_h(x) = \frac{1}{h} k\left(\frac{x}{h}\right)$$

Plus h est grand, plus cette fonction sera étendue, et plus h est petit, moins elle le sera.
En fait, si k est la densité de la loi $\mathcal{N}(0,1)$, k_h est la densité de la loi $\mathcal{N}(0,h^2)$. Et $x \mapsto k_h(x-x_i)$ est la densité de la loi $\mathcal{N}(x_i,h^2)$.
On définit $\displaystyle\widehat{f}_h(x) = \frac{1}{n}\sum_{i=1} k_h(x-x_i)$

Densité ? III

Densité estimée par noyau k

Étant donné un échantillon $\boldsymbol{x} = \{x_1, x_2, \cdots, x_n\}$, un estimateur de la densité f est $\widehat{f}_h$, pour $k > 0$,

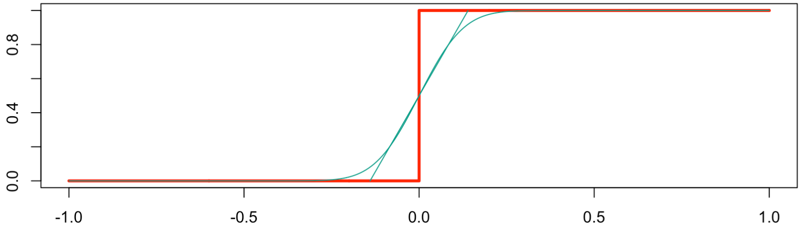
$$\widehat{f}_h(x) = \frac{1}{n} \sum_{i=1} k_h(x - x_i), \text{ où } k_h(\cdot) = \frac{1}{h} k\left(\frac{\cdot}{h}\right)$$

Densité ? IV

Rappelons que $\widehat{F}(x) = \frac{1}{n}\sum_{i=1} \mathbf{1}_{[x_i,\infty)}(x) = \frac{1}{n}\sum_{i=1} \mathbf{1}_{[0,\infty)}(x - x_i)$

La fonction de répartition associée à $\widehat{f}_h$ est

$$\widehat{F}_h(x) = \frac{1}{n}\sum_{i=1} K_h(x - x_i)$$



Densité ? V

Une autre approche est basée sur le fait que

$$f(x) = F'(x) = \lim_{h\to 0} \underbrace{\frac{F(x+h) - F(x-h)}{2h}}_{=f_h(x)}$$

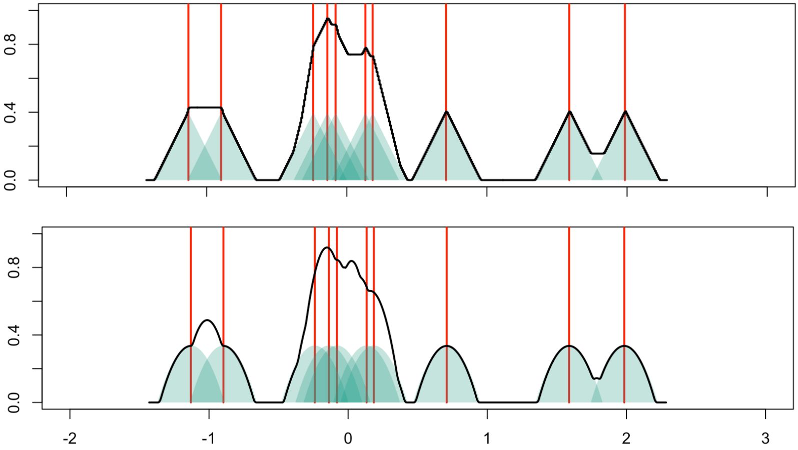
On peut estimer f_ζ par

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

qui est l'expression précédente si $k_h(x) = \dfrac{1}{2h}\mathbf{1}_{[-h,+h]}(x)$, ou $k(x) = \dfrac{1}{2}\mathbf{1}_{[-1,+1]}(x)$

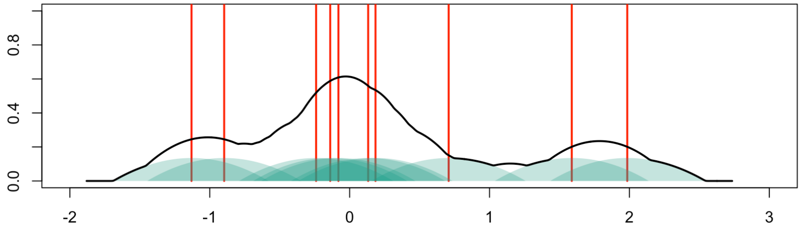
Densité ? VI

$$\widehat{f}_h(x) = \frac{1}{n}\sum_{i=1} k_h(x - x_i)$$



Densité ? VII

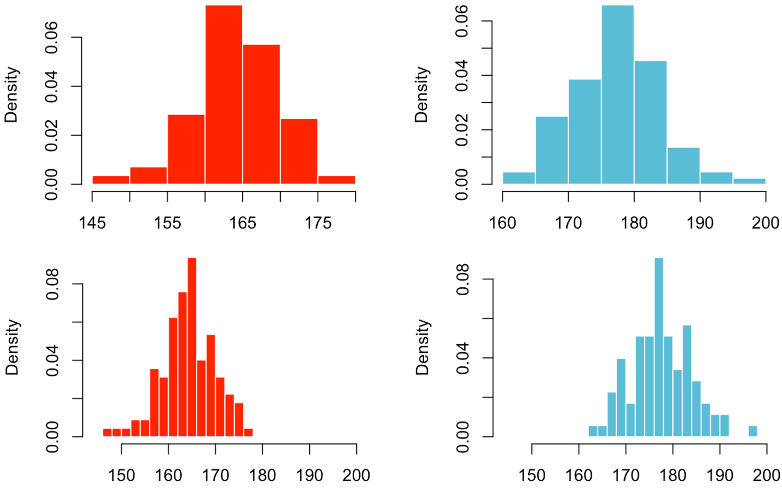
$$\widehat{f}_h(x) = \frac{1}{n}\sum_{i=1} k_h(x - x_i)$$




```
> loc = "http://socserv.socsci.mcmaster.ca/jfox/Books/
   Applied-Regression-2E/datasets/Davis.txt"
> Davis = read.table(loc)
> Davis[12,c(2,3)] = Davis[12,c(3,2)]
> x = Davis$height[Davis$sex == "F"]
> y = Davis$height[Davis$sex == "M"]
```

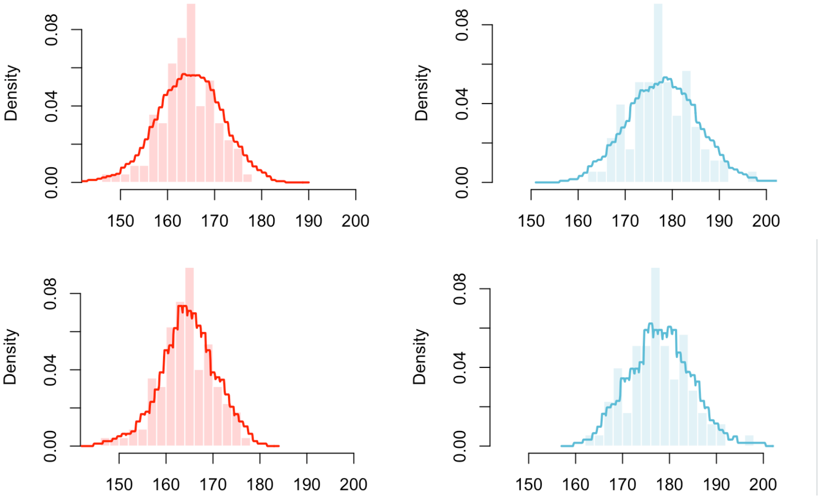
La taille des élèves I

```
> hist(x,probability=TRUE)
> hist(y,probability=TRUE)
```



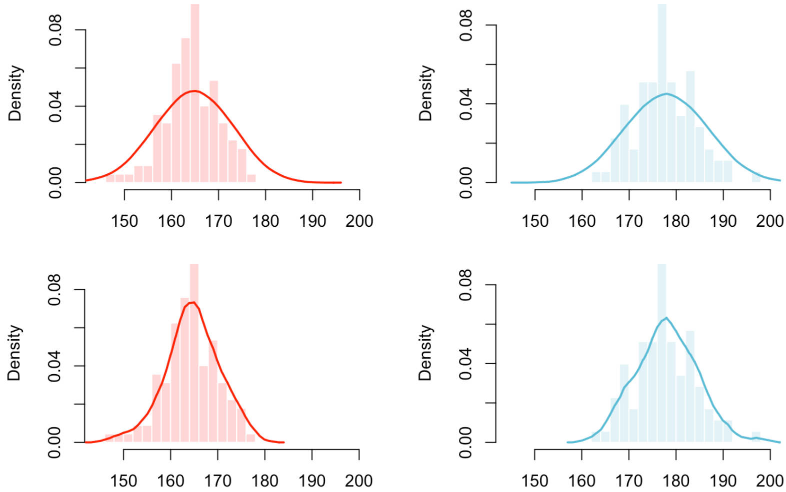
La taille des élèves II

```
> lines(density(x, kernel = "rectangular", bw=2))
> lines(density(y, kernel = "rectangular", bw=2))
```



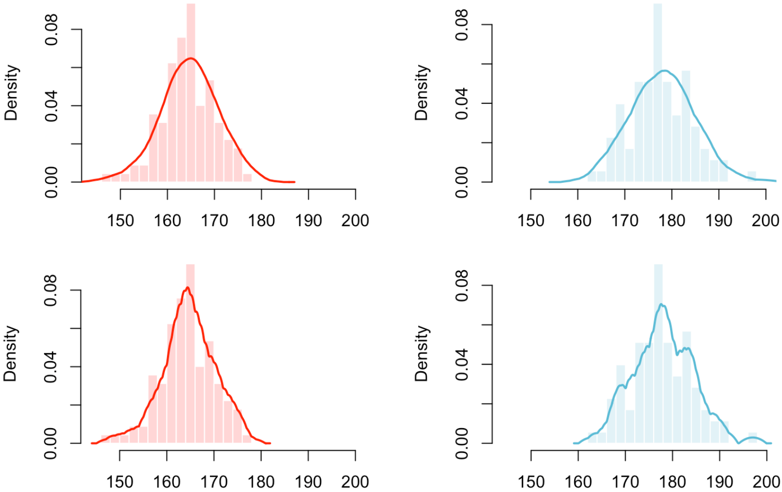
La taille des élèves III

```
> lines(density(x, kernel = "triangular", bw=2))
> lines(density(y, kernel = "triangular", bw=2))
```



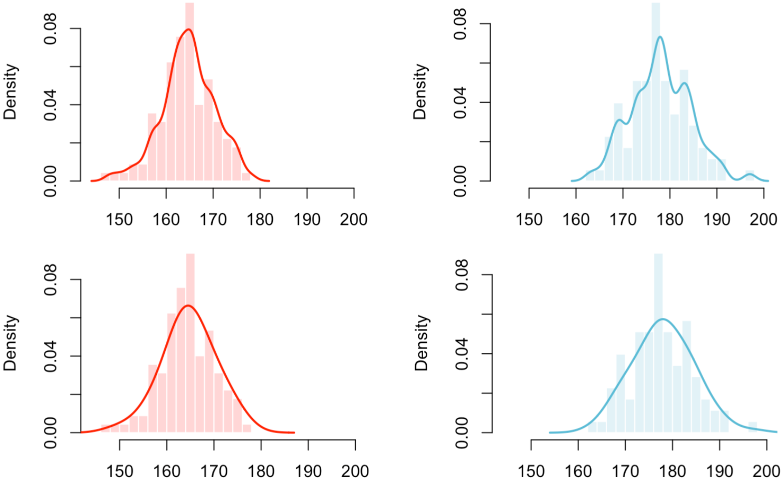
La taille des élèves IV

```
1 > lines(density(x, kernel = "epanechnikov", bw=2))
```



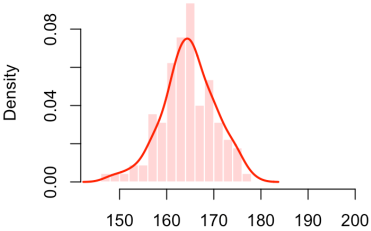
La taille des élèves V

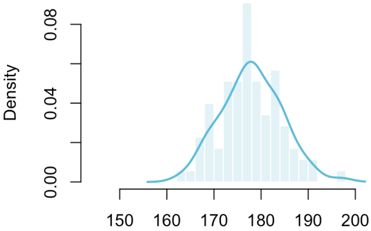
```
> lines(density(x, kernel = "gaussian", bw=2))
> lines(density(y, kernel = "gaussian", bw=2))
```



La taille des élèves VI

```
> lines(density(x, kernel = "gaussian", bw=2))
> lines(density(y, kernel = "gaussian", bw=2))
```



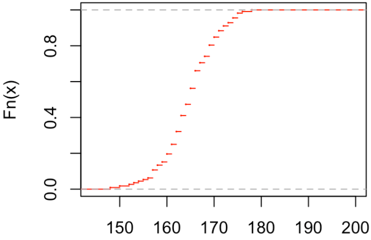


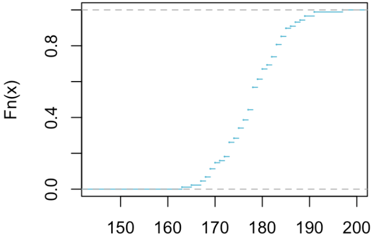

```
> density(x)$bw
[1] 1.8951
> density(y)$bw
[1] 2.367441
```

La taille des élèves VII

On peut aussi regarder la fonction de répartition, avec la fonction de répartition empirique, $\widehat{F}(x) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(x_i \leq x)$

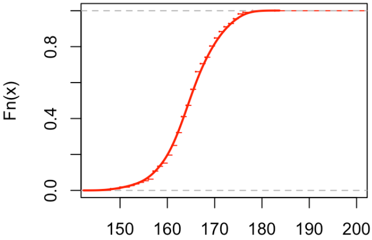
```
> plot(ecdf(x), cex=.1)
> plot(ecdf(y), cex=.1)
```

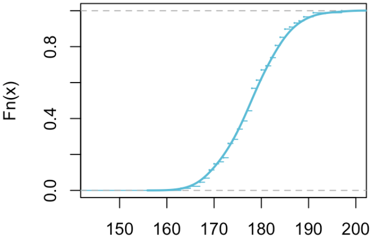




La taille des élèves VIII

```
> Dx = density(x)
> pasx = diff(Dx$x)[1]
> cumDx = cumsum(Dx$y*pasx)
> lines(Dx$x, cumDx)
```





On peut aussi cumuler la densité lissée, $\tilde{F}(x) = \int_{-\infty}^{x} \widehat{f}_h(t)dt$