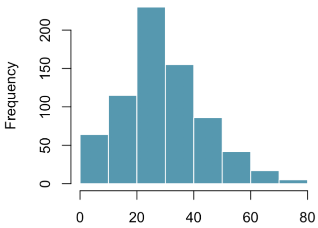
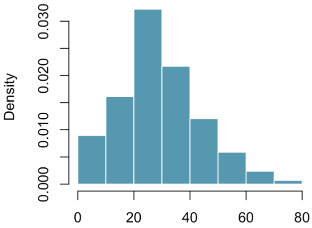
STT 1000 - STATISTIQVES

ARTHVR CHARPENTIER

Histogrammes

Pour faire l'histogramme associé à variable Y pour un échantillon $\{y_1, \cdots, y_n\}$, diviser l'étendue de $\boldsymbol{y}$ ($[\min\{y_i\}, \max\{y_i\}]$) en classes (intervalles), compter le nombre d'observations pour chacune de ces classes, et représenter un rectangle dont la largeur correspond à l'étendue de la classe et dont la hauteur est proportionnelle au nombre (ou au pourcentage) d'observations dans cette classe.





Pour une partition $a_0 < a_1 < \cdots < a_k$,

$$h_j = \sum_{i=1}^{n} \mathbf{1}(y_i \in [a_{j-1}, a_j])$$

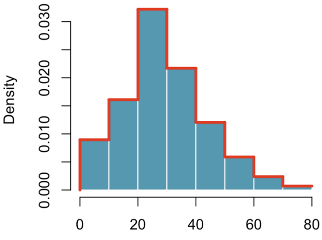
Histogrammes

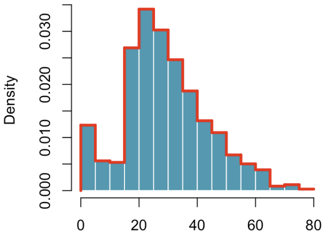
Pour une partition $a_0 < a_1 < \cdots < a_k$,

$$h_j = \sum_{i=1}^{n} \mathbf{1}(y_i \in [a_{j-1}, a_j])$$

Si $a_j - a_{j-1} = h$, et si on pose

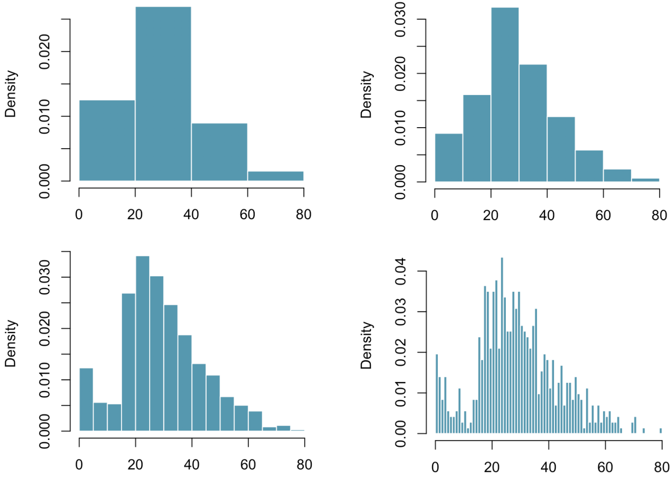
$$\widehat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} \mathbf{1}(y_i \in [a_{j-1}, a_j]) \text{ où } x \in [a_{j-1}, a_j], \int_{-\infty}^{+\infty} f(x)dx = 1$$





Histogrammes

Problème: dépend (fortement) de la partition $(A_i) = ([a_i, a_{i+1}))$



Fonction de répartition empirique

Pour une variable X, $F(x)=\mathbb{P}[X\leq x]$,
Étant donné un échantillon $\{x_1,\cdots,x_n\}$,

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

By the strong law of large numbers, the estimator $\widehat{F}_n(t)$ $\widehat{F}_n(t)$ converges to $F(t)$ as $n\to\infty$ almost surely, for every value of t

$$\widehat{F}_n(t)\xrightarrow{\text{p.s.}}F(t)$$

Glivenko–Cantelli theorem,

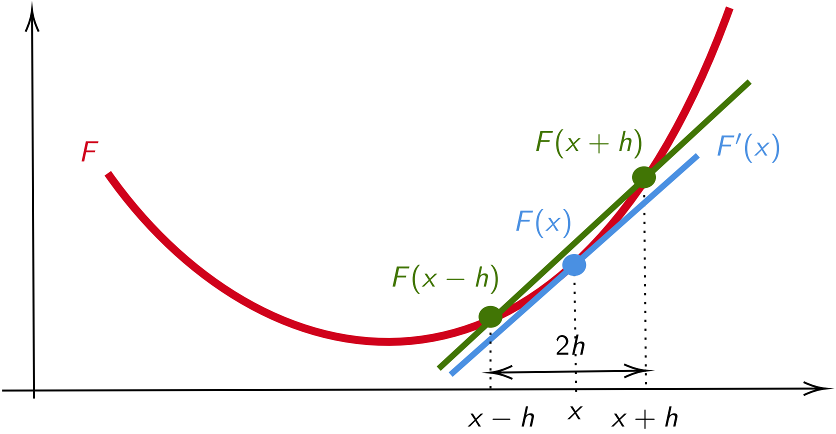
$$\|\widehat{F}_n-F\|_\infty=\sup_{t\in\mathbb{R}}\left|\widehat{F}_n(t)-F(t)\right|\xrightarrow{\text{p.s.}}0.$$

Densité

La fonction $\widehat{F}_n$ n'est pas dérivable mais on peut utiliser

$$\widehat{f}_{n,h}(x) \approx \frac{\widehat{F}_n(x+h) - \widehat{F}_n(x-h)}{2h}$$

pour h suffisement faible.



Densité

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

est l'estimateur empirique de $x \mapsto \mathbb{P}[X \in [x \pm h]]$.
On parle d'histogramme glissant

Moyenne et variance

Étant donné un échantillon $\{x_1, \cdots, x_n\}$, on appelle moyenne

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

et on appelle variance empirique

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2,$$