

# Statistiques non-paramétriques

En Statistique paramétrique

$$X_1, \dots, X_n \text{ iid} \sim P_0$$

$$X_1, \dots, X_T \text{ non iid}$$

$$X_t = \alpha X_{t-1} + \varepsilon_t$$

$$\mathbb{E}[Y|X] = g(X)$$

En Statistique non paramétrique,

$$X_1, \dots, X_n \text{ iid}$$

↳ On veut estimer la densité de  $X_i$

$$\mathbb{E}[Y|X] = g(X)$$

⊕ pas d'hypothèses sur le modèle

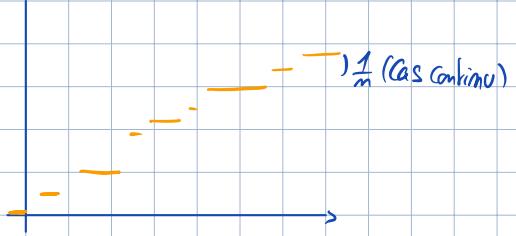
⊖ Il faut davantage d'observations.

# I - Fonction de répartition

## I-1 - Estimateur empirique

Soient  $X_1, \dots, X_m$  iid de fonction de répartition inconnue  $F$ .

L'estimateur empirique est  $\hat{F}_m(x) = \frac{1}{m} \sum_{j=1}^m \mathbb{1}_{X_j \leq x}$



## I-2 - Propriétés de $\hat{F}_m$

- ① Les  $Y_i = \mathbb{1}_{X_i \leq x}, \dots, \mathbb{1}_{X_m \leq x}$  sont iid  $\sim \text{Bernoulli}(p(x))$  avec  $P(X_i \leq x) = F(x)$
- ②  $m \hat{F}_m(x) \sim \mathcal{B}(m, p(x))$
- ③  $\hat{F}_m(x)$  estime  $F(x)$  sans biais
- ④ Risque quadratique  $RQ(\hat{F}_m(x)) = V(\hat{F}_m(x)) = \frac{F(x)(1-F(x))}{m}$

**Remarque :** ⑤  $\hat{F}_m(x) = \begin{cases} 0 & x \leq X(1) \\ \frac{k}{m} & X(k) \leq x \leq X(k+1) \\ 1 & x \geq X(m) \end{cases}$  (cas  $X$  continu)

où  $X(1) < \dots < X(m)$  statistique d'ordre

⑥  $\underbrace{\hat{F}_m(X_i)}_{\text{rang de } X_i} = R_i$

On sait que  $F(x) \sim U(0, 1)$

On peut ainsi transformer  $X_1, \dots, X_m$  iid  $\sim F$

$$\underbrace{F(X_1)}_{Y_1}, \dots, \underbrace{F(X_m)}_{Y_m} \text{ iid } \sim U(0, 1)$$

⑦  $\hat{F}_m(x) \xrightarrow{P} F(x)$

⑧  $\sqrt{m} (\hat{F}_m(x) - F(x)) \xrightarrow{D} N(0, F(x)(1-F(x)))$

⑨ Inégalité DKW (Dvoretzky - Kiefer - Wolfowitz) (Massart)

$$\forall \varepsilon > 0, P\left(\sup_{x \in \mathbb{R}} |\hat{F}_m(x) - F(x)| > \varepsilon\right) \leq 2e^{-2m\varepsilon^2}$$

Consequence:  $P(F(x) \in [\hat{F}_m(x) - \varepsilon, \hat{F}_m(x) + \varepsilon]) = 1 - P(|\hat{F}_m(x) - F(x)| > \varepsilon)$

$$\geq 1 - P(\sup_{x \in \mathbb{R}} |\hat{F}_m(x) - F(x)| > \varepsilon) \\ \geq \frac{1}{2} e^{-2m\varepsilon^2}$$

$\Rightarrow$  on obtient un intervalle de confiance de niveau  $\geq \alpha$

### ⑩ Intervalle de Confiance asymptotique.

en utilisant ⑧ si  $m$  est grand

$$P(a \leq \sqrt{m} \frac{(\hat{F}_m(x) - F(x))}{\sqrt{\hat{F}_m(x)(1 - \hat{F}_m(x))}} \leq b) = 0.95$$

$$\approx P(a \leq N(0,1) \leq b) = 0.95$$

$a = -1.96$        $b = 1.96$

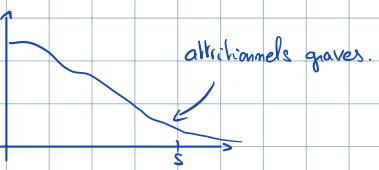
$$= P(\hat{F}_m(x) - Q \leq F(x) \leq \hat{F}_m(x) + Q) = 0.95$$

où  $Q = 1.96 \sqrt{\frac{\hat{F}_m(x)(1 - \hat{F}_m(x))}{m}}$

⑧ Puisque  $\hat{F}_m(x) \xrightarrow{P} F(x)$

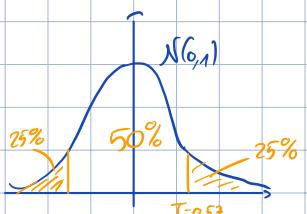
Slutsky,  $\frac{\sqrt{m}(\hat{F}_m(x) - F(x))}{\sqrt{\hat{F}_m(x)(1 - \hat{F}_m(x))}} \xrightarrow{D} N(0,1)$ .

### ⑪ Tester d'une valeur $F(x)$



On peut vouloir tester  $H_0: F(s) = 0.995$

$$\text{Sous } H_0: \sqrt{m} \frac{(\hat{F}_m(s) - 0.995)}{\sqrt{\hat{F}_m(s)(1 - \hat{F}_m(s))}} \xrightarrow{\text{agrand}} N(0,1)$$



### ⑫ Distribution de Kolmogorov

$$D_m = \sup_{x \in \mathbb{R}} |\hat{F}_m(x) - F(x)|$$

$$\text{On a } P(\sqrt{m} D_m > c) \xrightarrow{m \rightarrow \infty} 2 \sum_{i=1}^{\infty} (-1)^{i-1} e^{-2ic^2}$$

$1 - \text{Fonct}^{\circ} \text{de répartit}^{\circ} \text{ de Kolmogorov.}$

Application:  $X_1, \dots, X_m$  iid ~  $F$   
 $H_0: F = F_0 \quad (\forall x)$ .

$$T = \sqrt{m} \sup_{x \in \mathbb{R}} |\hat{F}_m(x) - F(x)|$$

Test asymptotique

cf diapos pour AN.

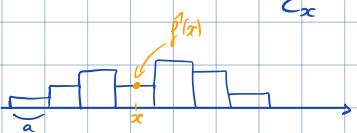
## II - Estimation de la densité

### ① Histogramme.

$X_1, \dots, X_m$  iid ~  $f$  (densité)

$$\hat{f}(x) = \frac{1}{mh} \# \{X_i \in \frac{[x]}{h}\}$$

$h$ : amplitude de la classe



$$\hat{f}(x) = \frac{1}{mh} \sum_{j=1}^m \mathbf{1}_{C_x}(x_j)$$

### ② Estimateur à moyaux

$X_1, \dots, X_m$  iid ~  $f$

$$\hat{f}(x) = \frac{1}{mh} \sum_{i=1}^m \mathbf{1}_{(x-\frac{h}{2} \leq X_i \leq x+\frac{h}{2})}$$

$$= \frac{1}{mh} \sum_{i=1}^m \mathbf{1}_{(-\frac{h}{2} \leq x - X_i \leq \frac{h}{2})}$$

$$= \frac{1}{mh} \sum_{i=1}^m \frac{1}{h} K\left(\frac{x - X_i}{h}\right)$$

*C va être défini*

Autre vision:  $\hat{f}(x) = \frac{\hat{F}_m(x+h) - \hat{F}_m(x-h)}{2h}$

$$= \frac{1}{m} \sum_{j=1}^m \frac{1}{2h} \mathbf{1}_{(X_j \in (x-h, x+h))} = \frac{1}{m} \sum_{j=1}^m \frac{1}{h} K\left(\frac{X_j - x}{h}\right)$$