S5 Estimation "plus lisse" d'une densité de probabilité (plus lisse que l'histogramme)

Cadre une v.a.r Y et ses réalisations $y \in \mathbb{R}$

indices : 1 2 ---- k ----- N

observations : y_1 y_2 --- y_k ---- y_N

ex $N = 16$

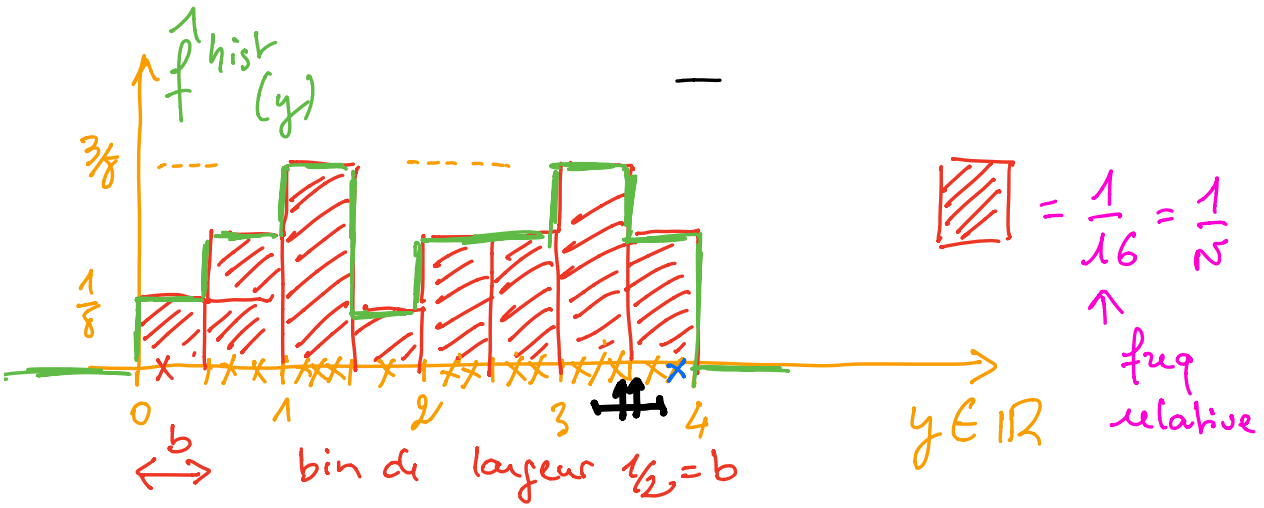
x x_{min} --- x_{max} ---- x

ex $\in [0\,4[$

Réalisation d'un 16-échantillon

Choix des 8 intervalles / "bins" de largeur $b = 1/2$

Classe	$[0;0{,}5[$	$[0{,}5;1[$	$[1;1{,}5[$	$[1{,}5;2[$	----	$[3{,}5\,4[$
Effectif	1	2	3	1	....	2
Freq Relative	1/16	2/16	3/16	1/16		2/16



Une formulation plus "mathématique" du calcul de l'histogramme

$$\hat{f}^{hist}(y) = \text{densité en } y \text{ vraisemblance de } y$$

$$\int_{\mathbb{R}} \hat{f}^{hist}(y)\,dy = \underbrace{b \times \frac{1}{8}}_{\frac{1}{16}} + \underbrace{b \times \frac{2}{8}}_{\frac{2}{16}} + b \times \frac{3}{8} + \dots$$

(1/2 sur le premier terme, 1/2 sur le deuxième terme)

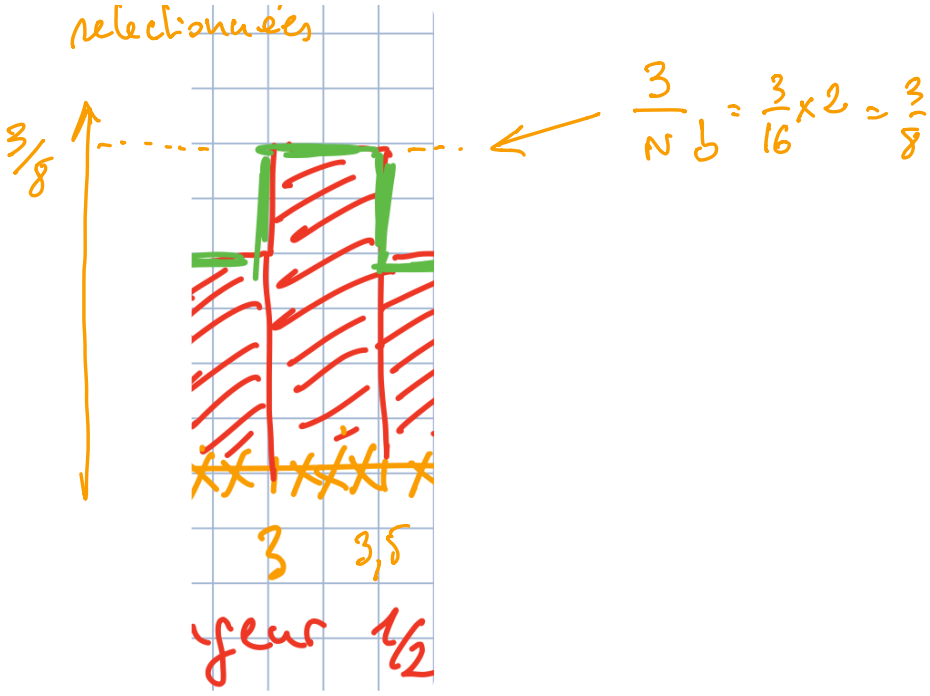
$$= 1$$

Une formule qui traduit le tracé de l'histogramme

"selectionner les réalisations appartenant au bin, les compter, diviser par la taille de l'échantillon (N), diviser par la largeur b du bin"

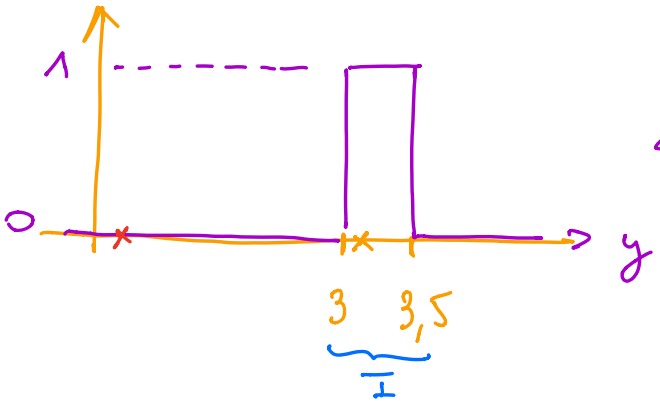
exemple bin $[3 ; 3{,}5[$ (b = 0,5)

3 realisations x x x



Fonction indicatrice d'un intervalle I

d'un bin I



$K(y) = \mathbb{1}_I(y)$

fonction qui "indique" un intervalle I

$$\mathbb{1}_I(y) = \begin{cases} 1 \text{ si } y \in I \\ 0 \text{ sinon} \end{cases}$$

Compter les $\sum_{i=1}^{N} \mathbb{1}_I(y_i) = 3$!

N ← 16 dans notre exemple

réalisation dans I

$\frac{1}{N b} \sum_{i=1}^{N} \mathbb{1}_{I}(y_i) = \frac{3}{16 \cdot \frac{1}{2}} = \frac{3 \times 2}{16} = \frac{3}{8}$

Diviser par N
(Pour trouver la freq relative)
Diviser par b
(Pour rejeter les unités de surface et assurer ainsi $\int \hat{f} = 1$)

$$\hat{f}^{hist}(y) = \frac{1}{N} \frac{\#\text{ de } y_i \text{ dans le } \hat{m} \text{ bin que } y}{\underbrace{\text{largeur du bin}}_{b}}$$

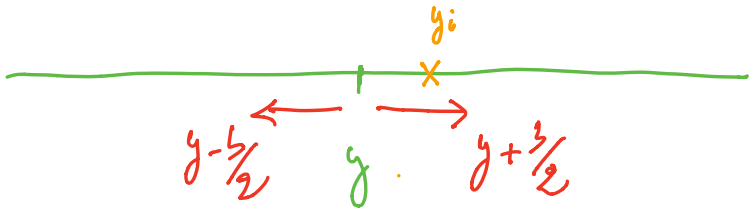
$$= \frac{1}{N} \sum_{i=1}^{N} \frac{1}{b} \mathbb{1}_{Bin(y)}(y_i)$$

Mais alors on peut aussi faire glisser la fenêtre de comptage (au lieu de fixer les Bin)

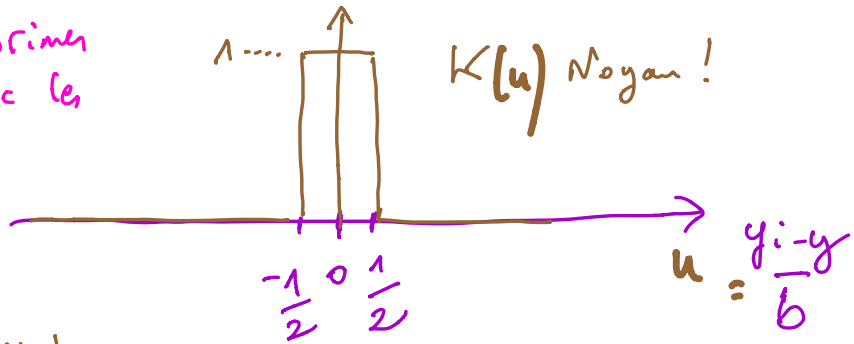
$$\hat{f}^{\text{fenêtre glissante}}(y) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{b} \mathbb{1}_{[y-\frac{b}{2}; y+\frac{b}{2}[}(y_i)$$

inconvénient → c'est l'intervalle qui

dépend de y



Idée : réexprimer la fenêtre avec les écarts



$$K\left(\frac{y_i-y}{b}\right) = 1 \iff -\frac{1}{2} \leq \frac{y_i-y}{b} \leq \frac{1}{2}$$

$$= \mathbb{1}_{[-\frac{1}{2}\,\frac{1}{2}[}\left(\frac{y_i-y}{b}\right) \qquad -\frac{b}{2} \leq y_i - y \leq \frac{b}{2}$$

$$y - \frac{b}{2} \leq y_i \leq y + \frac{b}{2}$$

$$\hat{f}_{\text{fenêtre glissante}}(y) = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{b}\,\mathbb{1}_{[y-\frac{b}{2};\, y+\frac{b}{2}[}(y_i)$$

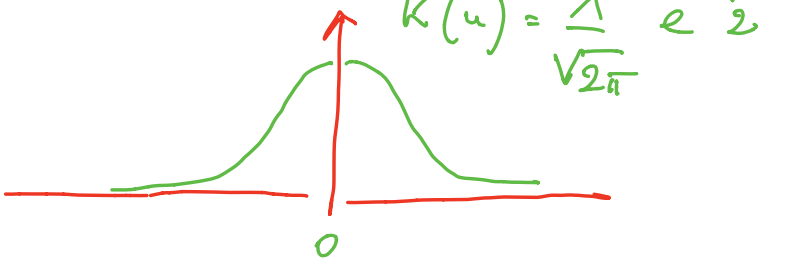
$$= \frac{1}{N}\sum_{i=1}^{N} \frac{1}{b}\,K\left(\frac{y_i-y}{b}\right)$$

Σ de N "noyaux"

Symétriques centrés en les y_i

Mais alors d'autres noyaux K (Kernel) sont possibles

$$\hat{f}^{\text{Noyau}}(y) = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{b} K\left(\frac{y_i - y}{b}\right)$$



Noyau gaussien !

Effet de b !!

bien choisir

Regarder une vidéo sur "Kernel Density Estimation"

Choice of kernel

The *triangular* kernel (or *linear* kernel) is given by

$$f(x) \propto \max(1 - |x|, 0).$$