

PARTIE 2: INTRODUCTION À LA STATISTIQUE NON PARAMÉTRIQUE

Irene Balelli

Centre Inria d'Université Côte d'Azur

irene.balelli@inria.fr

OBJECTIFS ET PRINCIPAUX THÈMES ABORDÉS

1. Introduction : qu'est-ce que la statistique non paramétrique ?
2. Rappels sur la fonction de densité
3. Estimateur de la densité par Histogramme
4. Estimateur de la densité à noyau
5. La régression non paramétrique
6. Estimateur par régressogramme de la fonction de régression
7. Régression par la méthode du noyau
8. Prévision non paramétrique
9. Quelques tests non paramétriques

I. INTRODUCTION :

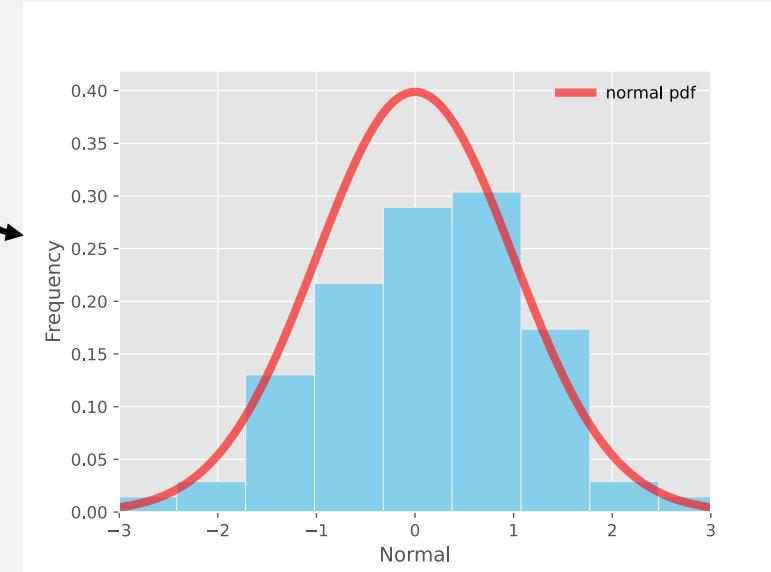
QU'EST-CE QUE LA STATISTIQUE NON PARAMÉTRIQUE ?

INTRODUCTION : QU'EST-CE QUE LA STATISTIQUE NON PARAMÉTRIQUE ?

- **Statistique paramétrique:**
 - La loi de l'échantillon qu'on souhaite analyser est supposée être connue, e.g. Gaussienne

INTRODUCTION : QU'EST-CE QUE LA STATISTIQUE NON PARAMÉTRIQUE ?

- **Statistique paramétrique:**
 - La loi de l'échantillon qu'on souhaite analyser est supposée être connue, e.g. Gaussienne



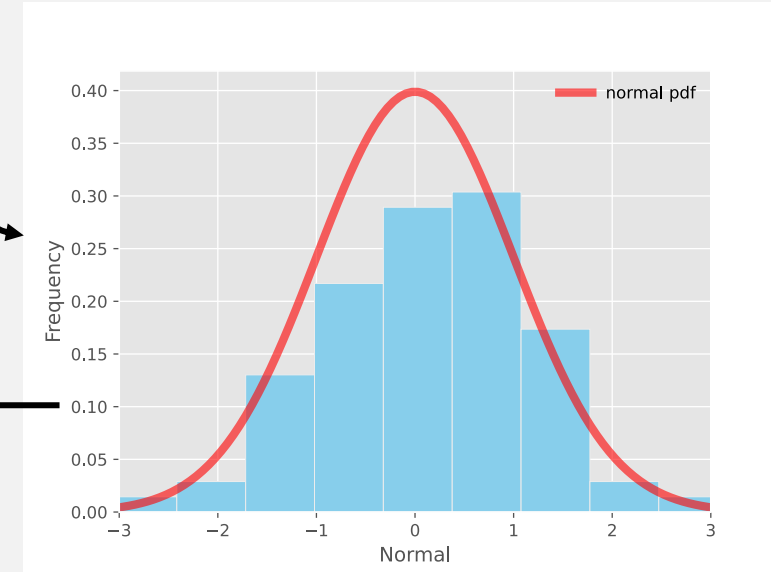
INTRODUCTION : QU'EST-CE QUE LA STATISTIQUE NON PARAMÉTRIQUE ?

- **Statistique paramétrique:**

- La loi de l'échantillon qu'on souhaite analyser est supposée être connue, e.g. Gaussienne
- D'où, le nombre de paramètre à estimer = nombre de paramètres inconnus du modèle choisi, e.g. la moyenne et la variance pour le modèle gaussien

$$x \sim \mathcal{N}(\mu, \sigma^2) \Rightarrow f_{\mu, \sigma}(x) := \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{x - \mu}{\sigma} \right)^2}$$

↑
2 paramètres inconnus



- **Partie I :** plusieurs méthodes pour estimer les paramètres inconnus, étant donné modèle + échantillon (e.g. méthode des moments, maximum de vraisemblance)

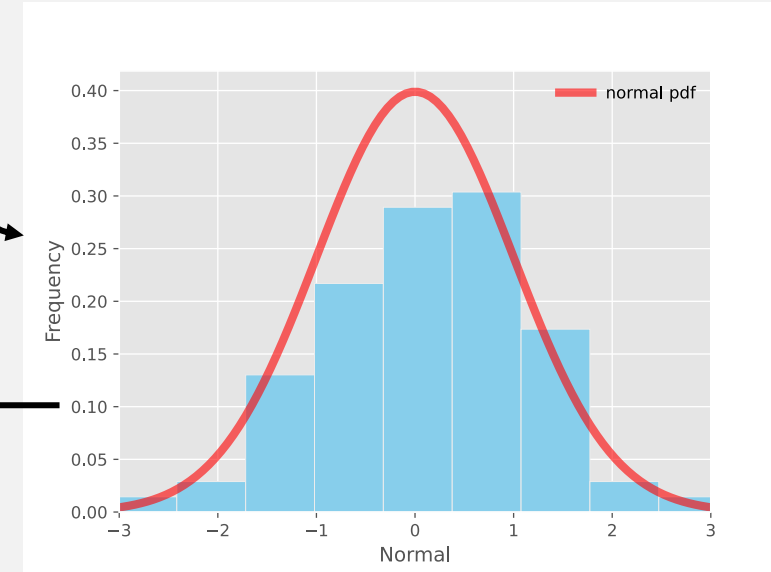
INTRODUCTION : QU'EST-CE QUE LA STATISTIQUE NON PARAMÉTRIQUE ?

- **Statistique paramétrique:**

- La loi de l'échantillon qu'on souhaite analyser est supposée être connue, e.g. Gaussienne
- D'où, le nombre de paramètre à estimer = nombre de paramètres inconnus du modèle choisi, e.g. la moyenne et la variance pour le modèle gaussien

$$x \sim \mathcal{N}(\mu, \sigma^2) \Rightarrow f_{\mu, \sigma}(x) := \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{x - \mu}{\sigma} \right)^2}$$

↑
2 paramètres inconnus



- **Partie I :** plusieurs méthodes pour estimer les paramètres inconnus, étant donné modèle + échantillon (e.g. méthode des moments, maximum de vraisemblance)

- **Statistique non paramétrique:**

- Pas d'informations concernant la distribution de la population observée → pas de modèle a priori

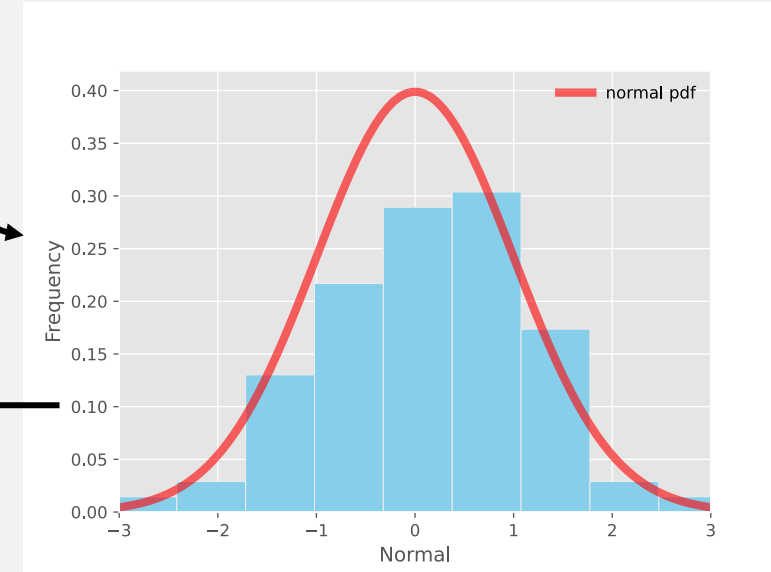
INTRODUCTION : QU'EST-CE QUE LA STATISTIQUE NON PARAMÉTRIQUE ?

- **Statistique paramétrique:**

- La loi de l'échantillon qu'on souhaite analyser est supposée être connue, e.g. Gaussienne
- D'où, le nombre de paramètre à estimer = nombre de paramètres inconnus du modèle choisi, e.g. la moyenne et la variance pour le modèle gaussien

$$x \sim \mathcal{N}(\mu, \sigma^2) \Rightarrow f_{\mu, \sigma}(x) := \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{x - \mu}{\sigma} \right)^2}$$

↑
2 paramètres inconnus



- **Partie I :** plusieurs méthodes pour estimer les paramètres inconnus, étant donné modèle + échantillon (e.g. méthode des moments, maximum de vraisemblance)

- **Statistique non paramétrique:**

- Pas d'informations concernant la distribution de la population observée → pas de modèle a priori
- La distribution doit être entièrement apprise à partir de données

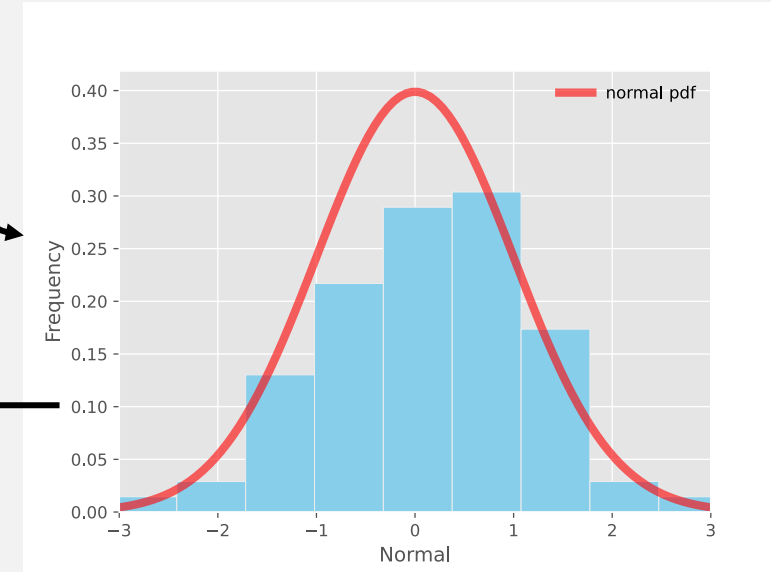
INTRODUCTION : QU'EST-CE QUE LA STATISTIQUE NON PARAMÉTRIQUE ?

- **Statistique paramétrique:**

- La loi de l'échantillon qu'on souhaite analyser est supposée être connue, e.g. Gaussienne
- D'où, le nombre de paramètre à estimer = nombre de paramètres inconnus du modèle choisi, e.g. la moyenne et la variance pour le modèle gaussien

$$x \sim \mathcal{N}(\mu, \sigma^2) \Rightarrow f_{\mu, \sigma}(x) := \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{x - \mu}{\sigma} \right)^2}$$

↑
2 paramètres inconnus



- **Partie I :** plusieurs méthodes pour estimer les paramètres inconnus, étant donné modèle + échantillon (e.g. méthode des moments, maximum de vraisemblance)

- **Statistique non paramétrique:**

- Pas d'informations concernant la distribution de la population observée → pas de modèle a priori
- La distribution doit être entièrement apprise à partir de données
- Nous pouvons faire de hypothèse sur la famille des distributions possibles : forme, nature, type, support, dérivabilité etc.

INTRODUCTION : QU'EST-CE QUE LA STATISTIQUE NON PARAMÉTRIQUE ?

- **Estimation non paramétrique : inconvénients**

- I. Le poids des données : plus sensible au bruit et à l'échantillon d'entraînement.

INTRODUCTION : QU'EST-CE QUE LA STATISTIQUE NON PARAMÉTRIQUE ?

- **Estimation non paramétrique : inconvénients**
 1. Le poids des données : plus sensible au bruit et à l'échantillon d'entraînement.
 2. Généralement, un grand nombre d'exemples est nécessaire pour assurer une bonne couverture de l'espace → une bonne estimation de la densité d'où la population a été tirée, surtout si haute dimension.

INTRODUCTION : QU'EST-CE QUE LA STATISTIQUE NON PARAMÉTRIQUE ?

- **Estimation non paramétrique : inconvénients**

1. Le poids des données : plus sensible au bruit et à l'échantillon d'entraînement.
2. Généralement, un grand nombre d'exemples est nécessaire pour assurer une bonne couverture de l'espace → une bonne estimation de la densité d'où la population a été tirée, surtout si haute dimension.
3. Un petit nombre de paramètres est suffisant pour décrire l'échantillon dans un contexte paramétrique (compression de l'information) : l'échantillon d'entraînement doit être stocké dans le cas non paramétrique pour la prédiction → mémoire.

INTRODUCTION : QU'EST-CE QUE LA STATISTIQUE NON PARAMÉTRIQUE ?

- **Estimation non paramétrique : inconvénients**

1. Le poids des données : plus sensible au bruit et à l'échantillon d'entraînement.
2. Généralement, un grand nombre d'exemples est nécessaire pour assurer une bonne couverture de l'espace → une bonne estimation de la densité d'où la population a été tirée, surtout si haute dimension.
3. Un petit nombre de paramètres est suffisant pour décrire l'échantillon dans un contexte paramétrique (compression de l'information) : l'échantillon d'entraînement doit être stocké dans le cas non paramétrique pour la prédiction → mémoire.
4. Plus lent en phase de déploiement.

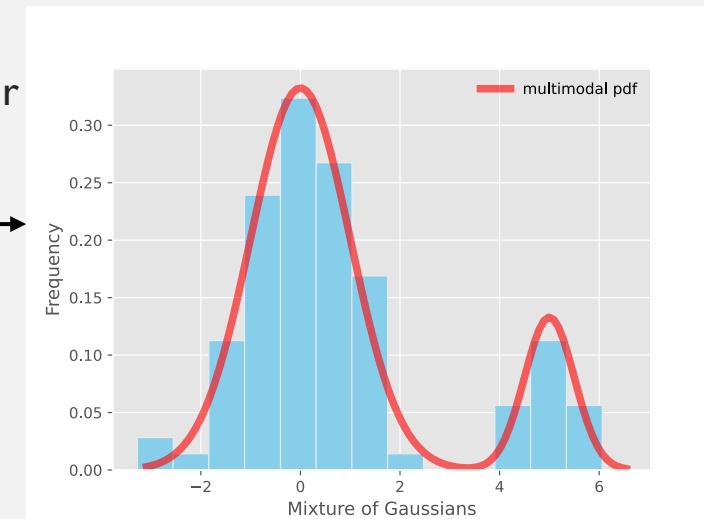
INTRODUCTION : QU'EST-CE QUE LA STATISTIQUE NON PARAMÉTRIQUE ?

- **Estimation non paramétrique : inconvénients**

1. Le poids des données : plus sensible au bruit et à l'échantillon d'entraînement.
2. Généralement, un grand nombre d'exemples est nécessaire pour assurer une bonne couverture de l'espace → une bonne estimation de la densité d'où la population a été tirée, surtout si haute dimension.
3. Un petit nombre de paramètres est suffisant pour décrire l'échantillon dans un contexte paramétrique (compression de l'information) : l'échantillon d'entraînement doit être stocké dans le cas non paramétrique pour la prédiction → mémoire.
4. Plus lent en phase de déploiement.

- **Estimation non paramétrique : avantages**

1. Généralement, nous ne connaissons pas la « vrai » densité de probabilité. Avoir un a priori sur le modèle peut induire en erreur, si cet a priori n'est pas correct, e.g. une hypothèse Gaussienne unimodale alors que la distribution est en réalité multimodale



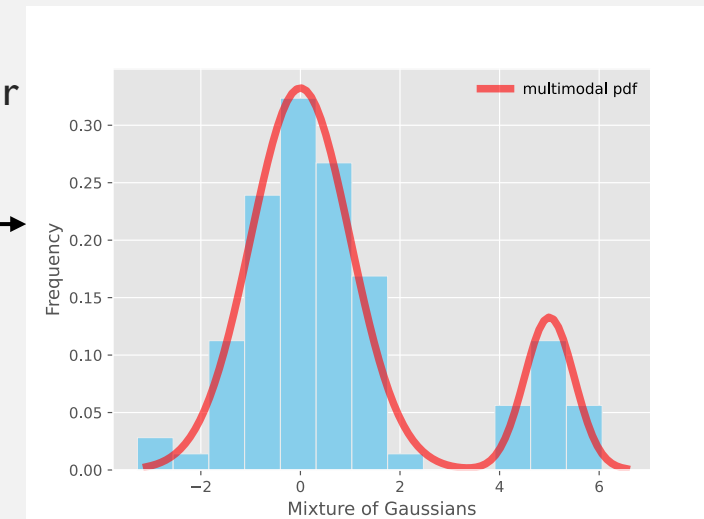
INTRODUCTION : QU'EST-CE QUE LA STATISTIQUE NON PARAMÉTRIQUE ?

- **Estimation non paramétrique : inconvénients**

1. Le poids des données : plus sensible au bruit et à l'échantillon d'entraînement.
2. Généralement, un grand nombre d'exemples est nécessaire pour assurer une bonne couverture de l'espace → une bonne estimation de la densité d'où la population a été tirée, surtout si haute dimension.
3. Un petit nombre de paramètres est suffisant pour décrire l'échantillon dans un contexte paramétrique (compression de l'information) : l'échantillon d'entraînement doit être stocké dans le cas non paramétrique pour la prédiction → mémoire.
4. Plus lent en phase de déploiement.

- **Estimation non paramétrique : avantages**

1. Généralement, nous ne connaissons pas la « vrai » densité de probabilité. Avoir un a priori sur le modèle peut induire en erreur, si cet a priori n'est pas correct, e.g. une hypothèse Gaussienne unimodale alors que la distribution est en réalité multimodale
2. Si nous n'arrivons pas à ajuster les observations à aucune distribution paramétrique ou nous ne savons pas e.g. le nombre de composantes à mettre dans un mélange



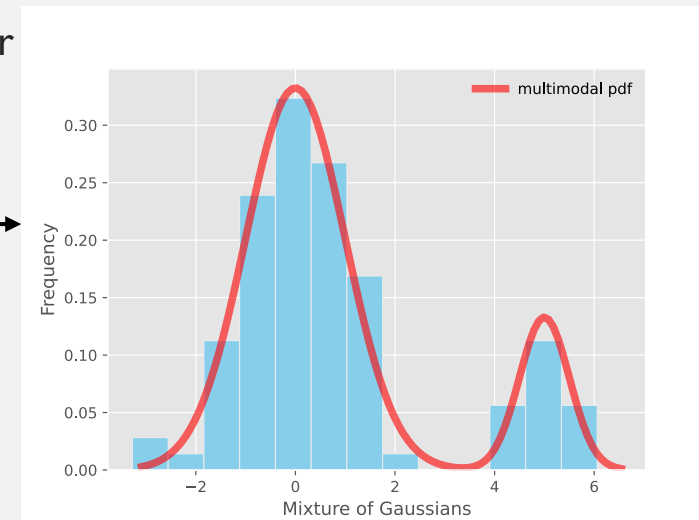
INTRODUCTION : QU'EST-CE QUE LA STATISTIQUE NON PARAMÉTRIQUE ?

- **Estimation non paramétrique : inconvénients**

1. Le poids des données : plus sensible au bruit et à l'échantillon d'entraînement.
2. Généralement, un grand nombre d'exemples est nécessaire pour assurer une bonne couverture de l'espace → une bonne estimation de la densité d'où la population a été tirée, surtout si haute dimension.
3. Un petit nombre de paramètres est suffisant pour décrire l'échantillon dans un contexte paramétrique (compression de l'information) : l'échantillon d'entraînement doit être stocké dans le cas non paramétrique pour la prédiction → mémoire.
4. Plus lent en phase de déploiement.

- **Estimation non paramétrique : avantages**

1. Généralement, nous ne connaissons pas la « vrai » densité de probabilité. Avoir un a priori sur le modèle peut induire en erreur, si cet a priori n'est pas correct, e.g. une hypothèse Gaussienne unimodale alors que la distribution est en réalité multimodale
2. Si nous n'arrivons pas à ajuster les observations à aucune distribution paramétrique ou nous ne savons pas e.g. le nombre de composantes à mettre dans un mélange
3. En cas de haute dimension, un modèle paramétrique peut être difficile à estimer en raison du nombre élevé de paramètres à estimer → problèmes d'identifiabilité



INTRODUCTION : QU'EST-CE QUE LA STATISTIQUE NON PARAMÉTRIQUE ?

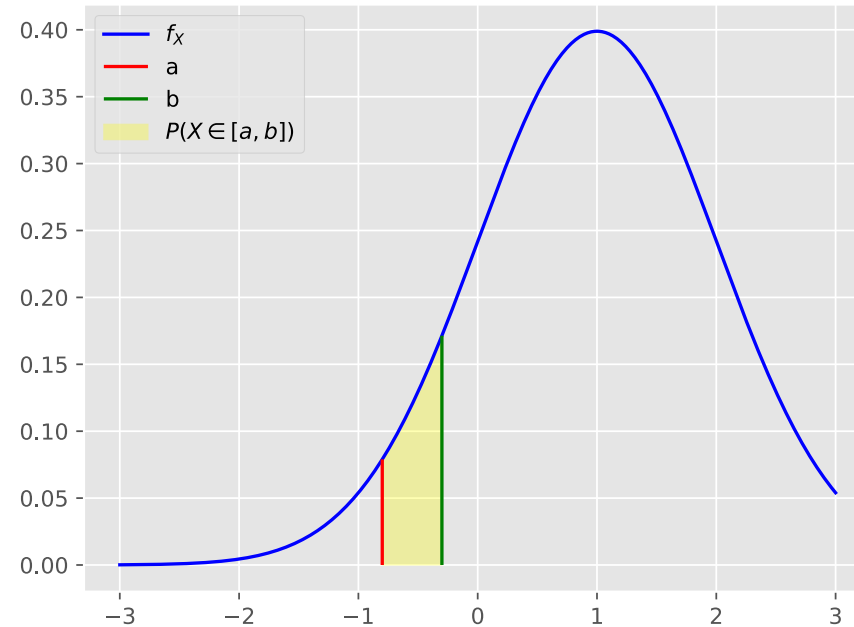
Paramétrique	Non paramétrique
La population es bien connue	Aucune information sur la population n'est disponible
Hypothèses sur la population et sa « vrai » distribution	Aucune hypothèse n'est faite sur la population/distribution
Echantillon des données basé sur la distribution	Echantillons de données arbitraires

2. RAPPELS SUR LA FONCTION DE DENSITÉ

RAPPELS SUR LA FONCTION DE DENSITÉ

- Soit X une variable aléatoire continue, i.e. définie sur \mathbb{R} (support infini) ou un intervalle borné $C \subset \mathbb{R}$ (support fini).
- Dans ce cas, pour tout $(a, b) \in \mathbb{R}^2, a < b$, la probabilité que la variable X se trouve dans l'intervalle $[a, b]$ est déterminée à travers sa fonction de densité, f_X , à l'aide d'un calcul intégrale :

$$P(X \in [a, b]) := \int_a^b f_X(x) dx$$



RAPPELS SUR LA FONCTION DE DENSITÉ

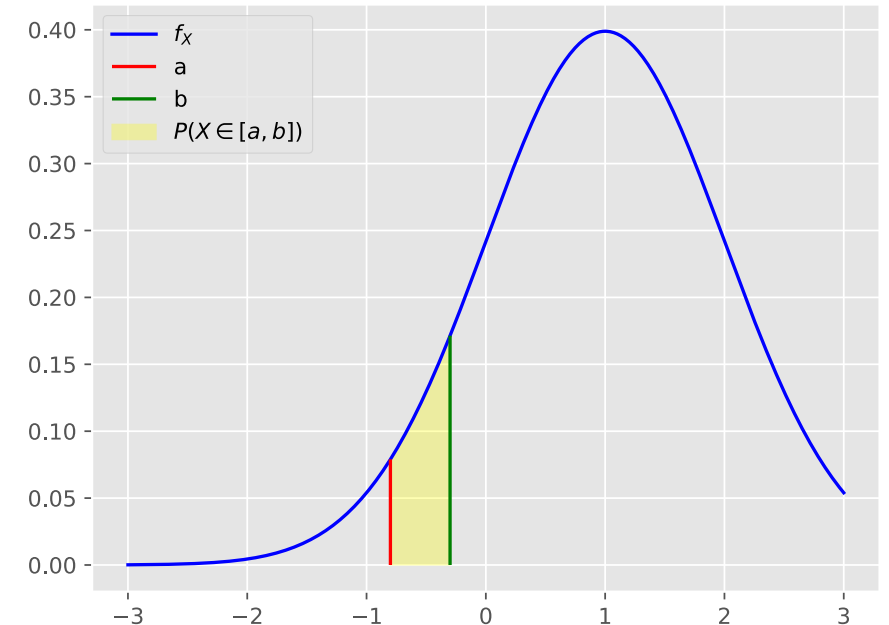
- Soit X un variable aléatoire continue, i.e. définie sur \mathbb{R} (support infini) ou un intervalle borné $C \subset \mathbb{R}$ (support fini).
- Dans ce cas, pour tout $(a, b) \in \mathbb{R}^2, a < b$, la probabilité que la variable X se trouve dans l'intervalle $[a, b]$ est déterminée à travers sa fonction de densité, f_X , à l'aide d'un calcul intégrale :

$$P(X \in [a, b]) := \int_a^b f_X(x) dx$$

En autre termes, pour $\epsilon \rightarrow 0$, $P(X \in [a, a + \epsilon]) \approx \epsilon f_X(a)$

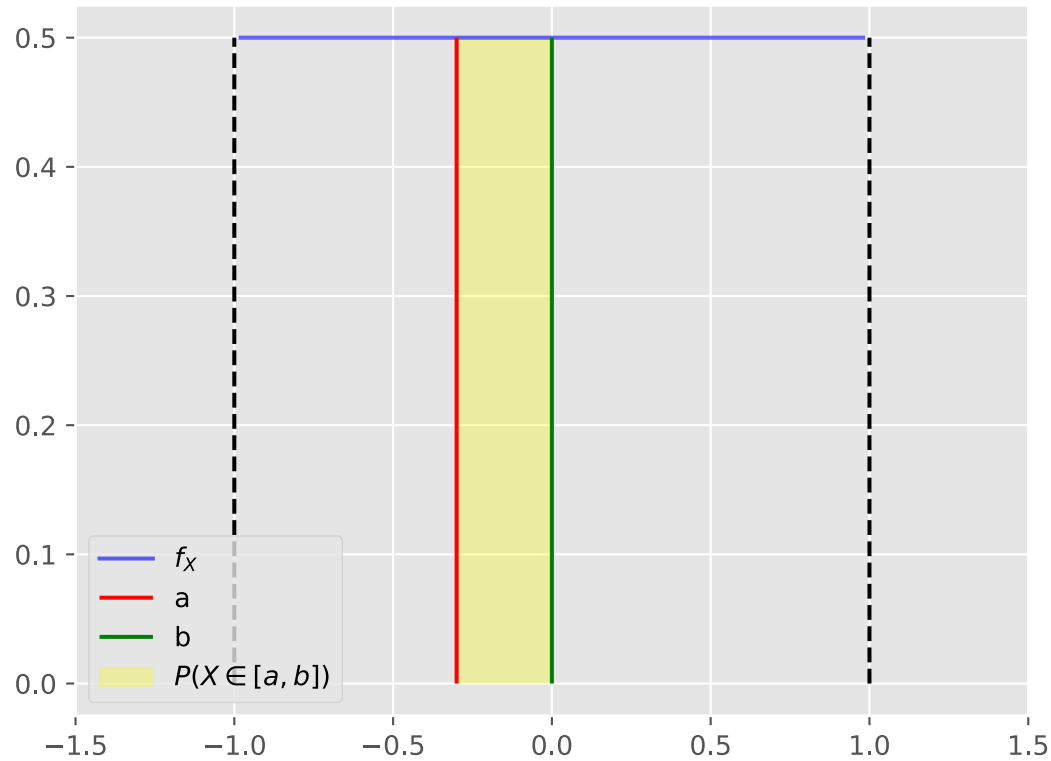
Quelques exemples de variables aléatoires à densité :

- Loi gaussienne (support = \mathbb{R})
- Loi exponentielle (support = \mathbb{R}^+)
- Loi uniforme, loi triangulaire (support borné)
- Extension naturelle à des variables continues vectorielles (\mathbb{R}^d)

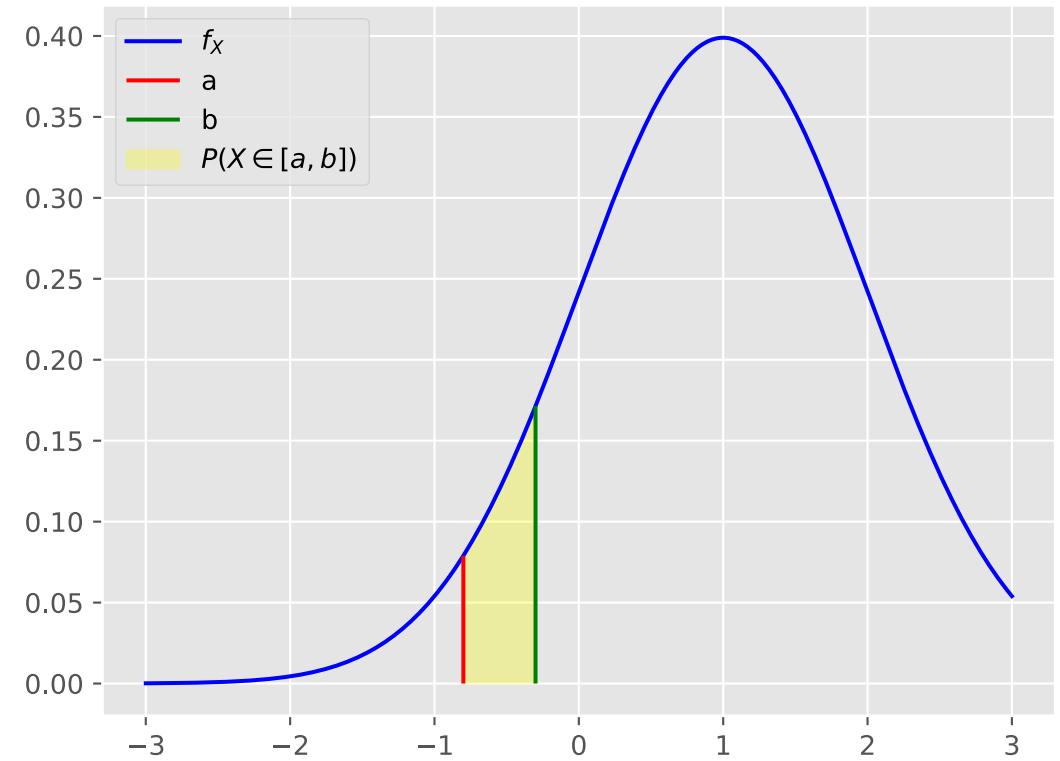


RAPPELS SUR LA FONCTION DE DENSITÉ

Uniforme : $X \sim \mathcal{U}([-1,1])$, $\text{supp}_X = [-1,1]$



Gaussian : $X \sim \mathcal{N}(0,1)$, $\text{supp}_X = \mathbb{R}$



Propriétés de la fonction de densité f_X d'une variable X :

- $f_X(x) \geq 0$ pour tout x dans son support, et elle est intégrable

Propriétés de la fonction de densité f_X d'une variable X :

- $f_X(x) \geq 0$ pour tout x dans son support, et elle est intégrable

- L'intégrale de f_X sur son support est égale à 1 :
$$\int_{\text{supp}_X} f_X(x) dx = 1$$

Propriétés de la fonction de densité f_X d'une variable X :

- $f_X(x) \geq 0$ pour tout x dans son support, et elle est intégrable

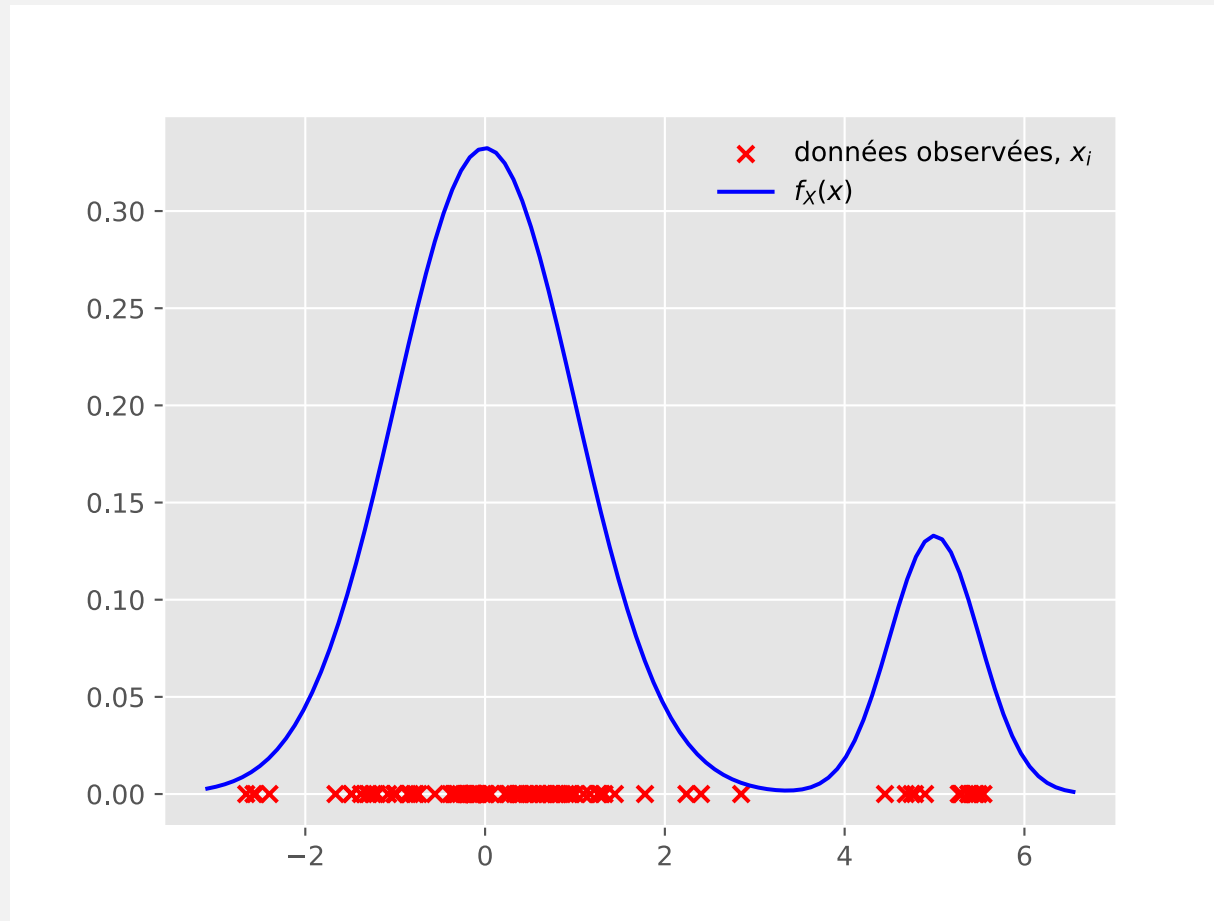
- L'intégrale de f_X sur son support est égale à 1 :
$$\int_{\text{supp}_X} f_X(x) dx = 1$$

- Lien avec le moment d'ordre k de X (sous condition que l'intégrale correspondant existe et soit fini) :

$$\mathbb{E}[X^k] := \int_{\text{supp}_X} x^k f_X(x) dx$$

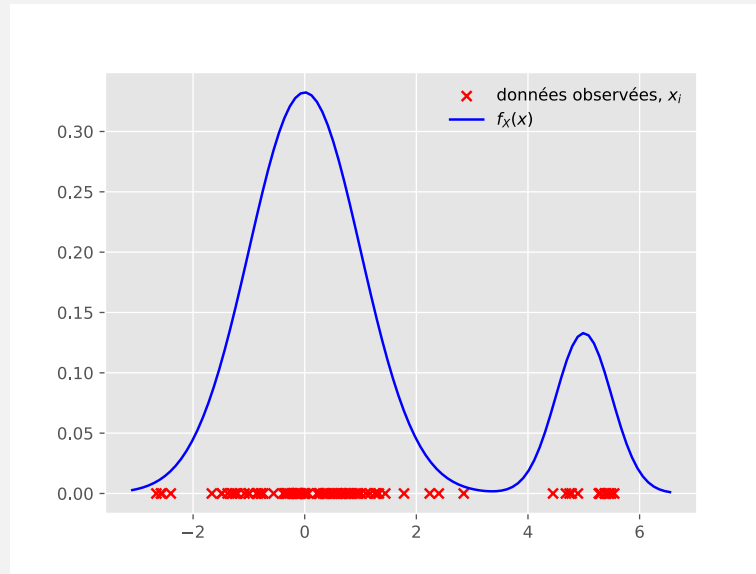
PROBLÈME : ESTIMATION DE LA DENSITÉ

Comment estimer une densité inconnue, à partir uniquement d'un échantillon observé ?



PROBLÈME : ESTIMATION DE LA DENSITÉ

Comment estimer une densité inconnue, à partir uniquement d'un échantillon observé ?



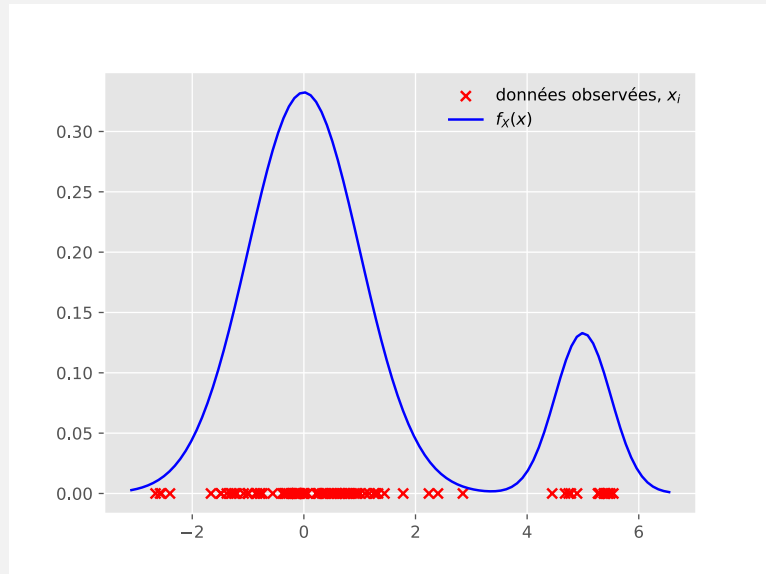
Soit $\mathcal{D}_N := \{x_1, \dots, x_N\} \subset \mathbb{R}$ un échantillon aléatoire composé de N observations : de quelle densité f est-il issu ?

Répondre à cette question peut nous apporter plusieurs informations :

- Quelles sont les régions à densité élevée ?
- Unimodale ou multimodale ?
- Quel est le support de la distribution ?
- Y a-t-il des outliers dans mon échantillon ?
- Quoi s'attendre d'un modèle décisionnel construit sur la base de cette densité ?

PROBLÈME : ESTIMATION DE LA DENSITÉ

Comment estimer une densité inconnue, à partir uniquement d'un échantillon observé ?



Approches non paramétriques:

1. Estimation par histogramme
2. Estimation par noyaux
3. Estimation par k-plus-proches-voisins

Soit $\mathcal{D}_N := \{x_1, \dots, x_N\} \subset \mathbb{R}^d$ un échantillon aléatoire composé de N observations : de quelle densité f est-il issu ?

Répondre à cette question peut nous apporter plusieurs informations :

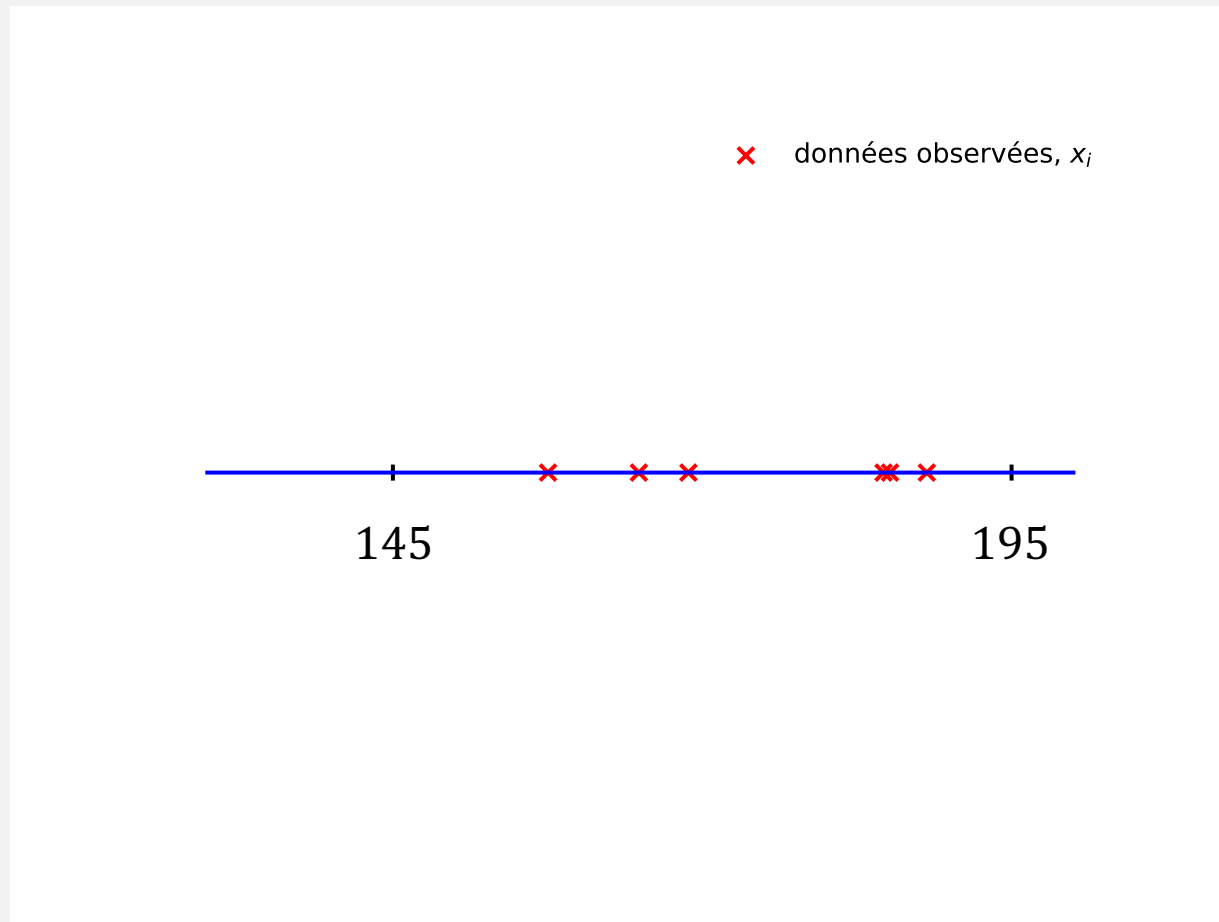
- Quelles sont les régions à densité élevée ?
- Unimodale ou multimodale ?
- Quel est le support de la distribution ?
- Y a-t-il des outliers dans mon échantillon ?
- Quoi s'attendre d'un modèle décisionnel construit sur la base de cette densité ?

3. ESTIMATEUR DE LA DENSITÉ PAR HISTOGRAMME

ESTIMATEUR PAR HISTOGRAMME

La méthode plus intuitive et élémentaire d'estimation de la densité est l'histogramme. Pour le construire on doit suivre 3 simples étapes.

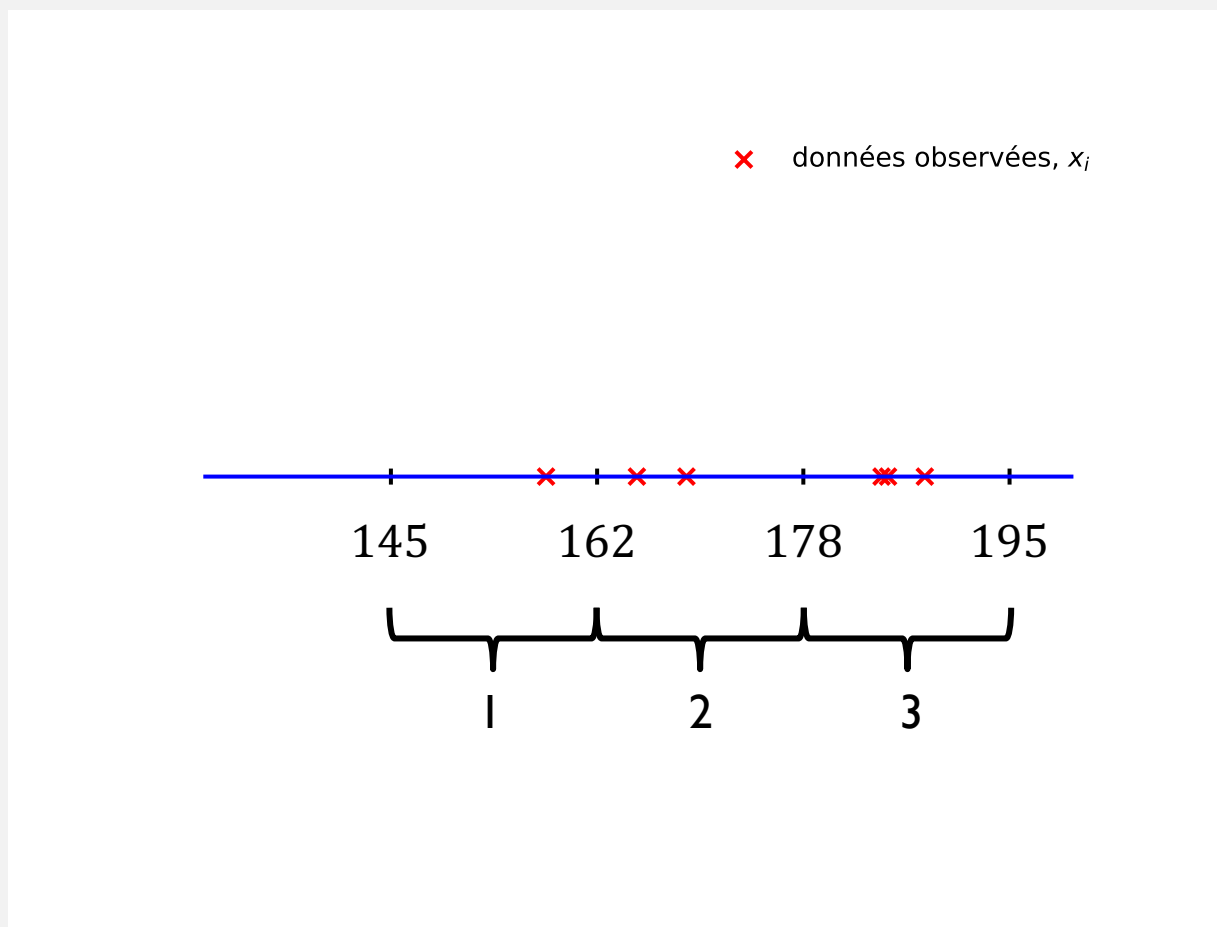
Supposons, pour simplicité, d'avoir un échantillon composé de 6 observations (la taille, en cm, de 6 personnes adultes – hommes et femmes) :



ESTIMATEUR PAR HISTOGRAMME

La méthode plus intuitive et élémentaire d'estimation de la densité est l'histogramme. Pour le construire on doit suivre 3 simples étapes.

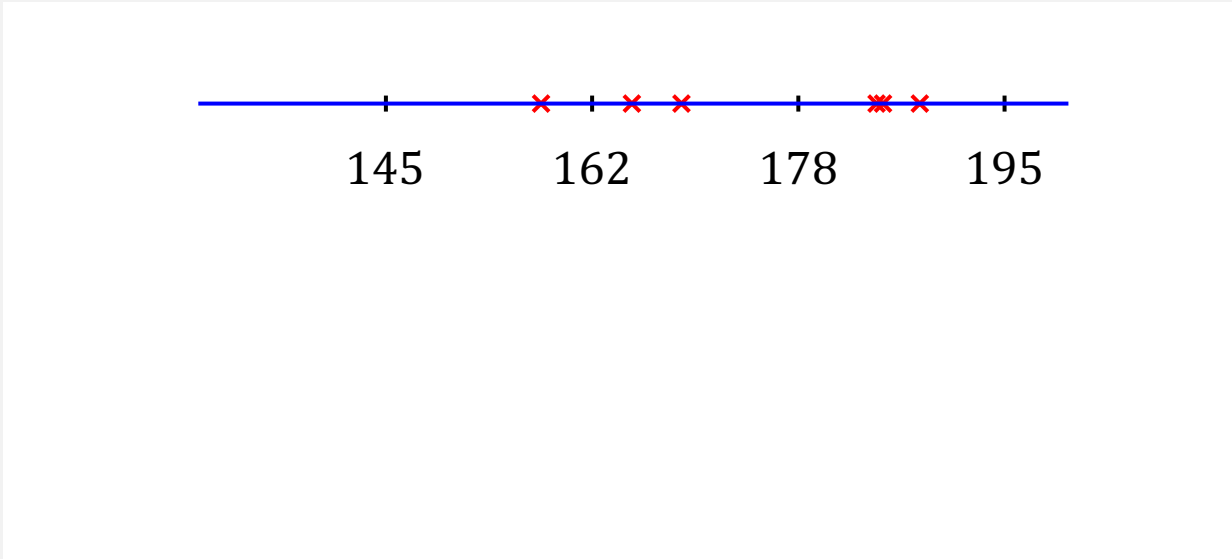
I. Nous allons diviser l'espace des observations en intervalles (ou « boîtes » dans le cas multidimensionnel), de même taille :



ESTIMATEUR PAR HISTOGRAMME

La méthode plus intuitive et élémentaire d'estimation de la densité est l'histogramme. Pour le construire on doit suivre 3 simples étapes.

2. Nous approximations la densité de chaque intervalle avec la fréquence des observations qu'il contient (probabilité empirique) :

$$\forall i = 1, \dots, \text{nombre intervalles}, f_i := \frac{N_i}{N}$$


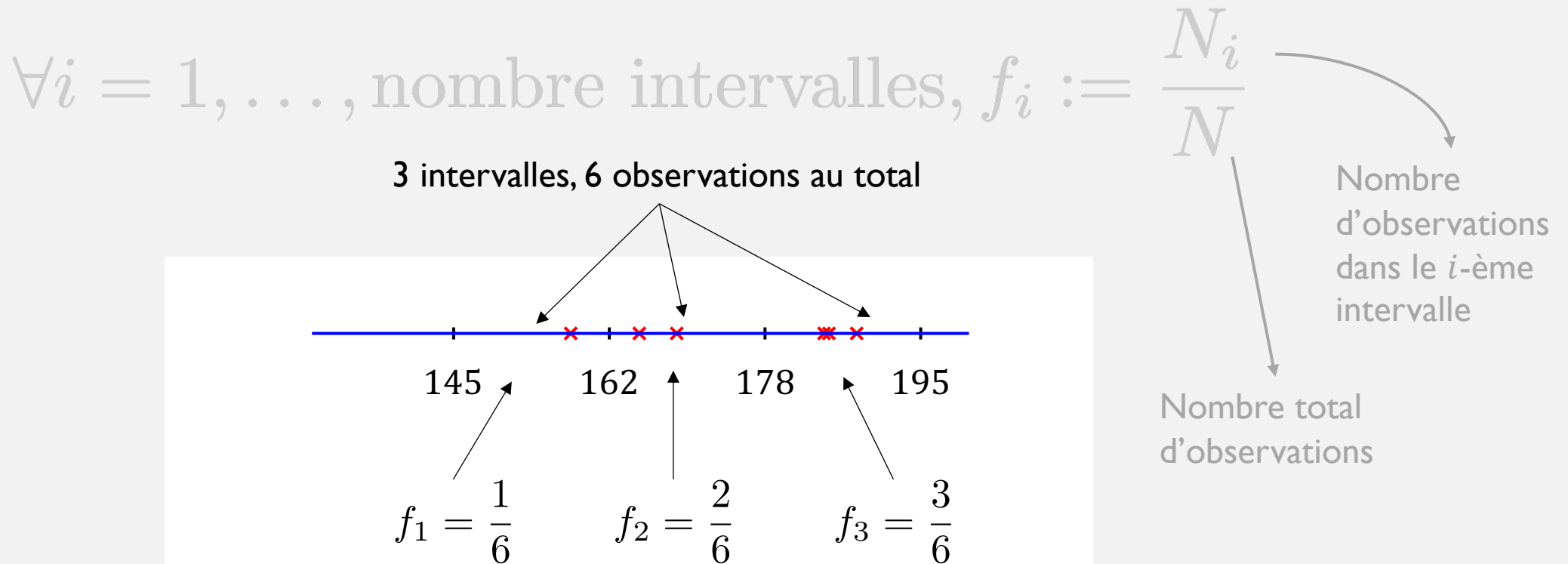
Nombre d'observations dans le i -ème intervalle

Nombre total d'observations

ESTIMATEUR PAR HISTOGRAMME

La méthode plus intuitive et élémentaire d'estimation de la densité est l'histogramme. Pour le construire on doit suivre 3 simples étapes.

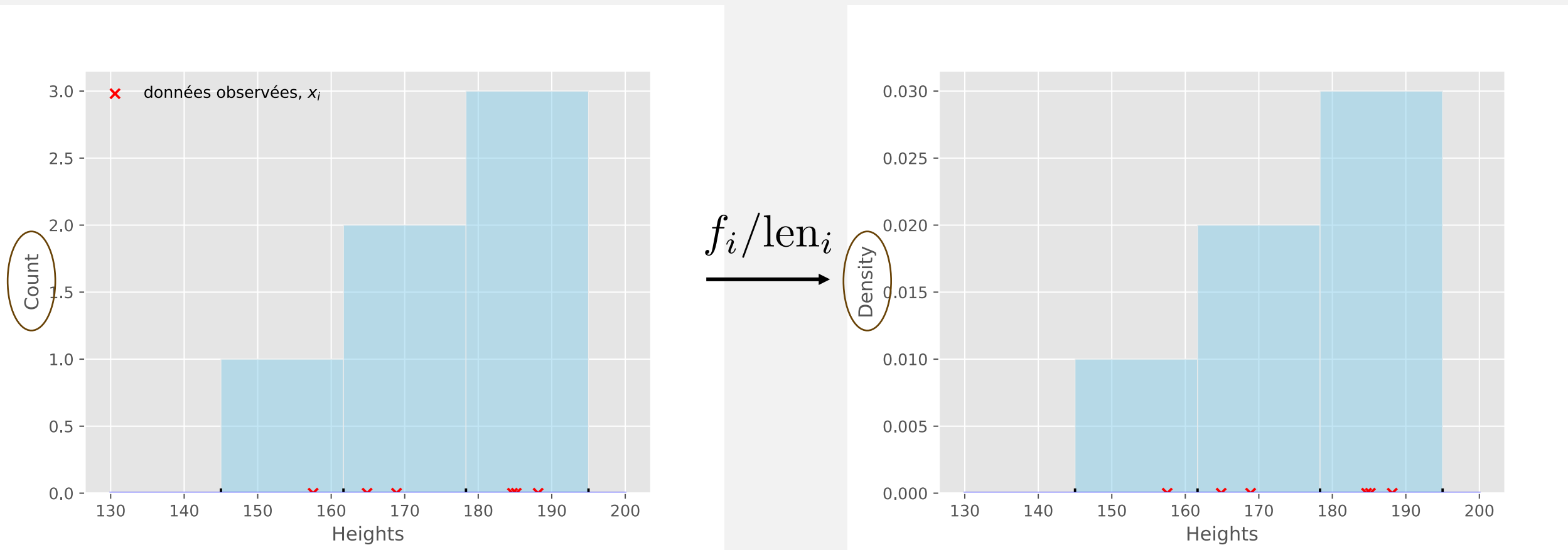
2. Nous approximations la densité de chaque intervalle avec la fréquence des observations qu'il contient (probabilité empirique) :



ESTIMATEUR PAR HISTOGRAMME

La méthode plus intuitive et élémentaire d'estimation de la densité est l'histogramme. Pour le construire on doit suivre 3 simples étapes.

3. Nous divisons ensuite par la longueur (ou volume) de l'intervalle (boîte), supposée ici constante, pour normaliser et ainsi obtenir une densité de probabilité « valide » :



ESTIMATEUR PAR HISTOGRAMME

Un peu de notation :

$\mathcal{D}_N := \{x_1, \dots, x_N\} \in [m, m + l]$, où l est la longueur de l'intervalle considéré, en supposant qu'en dehors de $[m, m + l]$ la densité peut être assimilée à 0 (hypothèse de support)

Dans notre exemple:

- $N = 6$
- $m = 145 \text{ cm}$
- $l = 50 \text{ cm}$

Un peu de notation :

$\mathcal{D}_N := \{x_1, \dots, x_N\} \in [m, m + l]$, où l est la longueur de l'intervalle considéré, en supposant qu'en dehors de $[m, m + l]$ la densité peut être assimilée à 0 (hypothèse de support)

- Nous avons divisé $[m, m + l]$ en un total de b intervalles. La taille de chaque intervalle est donc $\nu := l/b$

Dans notre exemple:

- $N = 6$
- $m = 145 \text{ cm}$
- $l = 50 \text{ cm}$
- $b = 3 \Rightarrow \nu = \frac{50 \text{ cm}}{3} \approx 17 \text{ cm}$

ESTIMATEUR PAR HISTOGRAMME

Un peu de notation :

$\mathcal{D}_N := \{x_1, \dots, x_N\} \in [m, m + l]$, où l est la longueur de l'intervalle considéré, en supposant qu'en dehors de $[m, m + l]$ la densité peut être assimilée à 0 (hypothèse de support)

- Nous avons divisé $[m, m + l]$ en un total de b intervalles. La taille de chaque intervalle est donc $\nu := l/b$
- Pour chaque intervalle, nous avons procédé au comptage du nombre d'observations contenues :

$$\forall i = 1, \dots, b, C_i := \sum_{j=1}^N \mathbb{I}_{[m+(i-1)\nu, m+i\nu]}(x_j)$$

$$\mathbb{I}_A(x) := \begin{cases} 1 & \text{si } x \in A \\ 0 & \text{autrement} \end{cases}$$

Dans notre exemple:

$$C_1 = 1, C_2 = 2, C_3 = 3$$

ESTIMATEUR PAR HISTOGRAMME

Un peu de notation :

$\mathcal{D}_N := \{x_1, \dots, x_N\} \in [m, m + l]$, où l est la longueur de l'intervalle considéré, en supposant qu'en dehors de $[m, m + l]$ la densité peut être assimilée à 0 (hypothèse de support)

- Nous avons divisé $[m, m + l]$ en un total de b intervalles $\nu := l/b$. Nous appelons ν la fenêtre.
- Pour chaque intervalle, nous avons procédé au comptage du nombre d'observations contenues :

$$\forall i = 1, \dots, b, C_i := \sum_{j=1}^N \mathbb{I}_{[m+(i-1)\nu, m+i\nu]}(x_j)$$

$$\mathbb{I}_A(x) := \begin{cases} 1 & \text{si } x \in A \\ 0 & \text{autrement} \end{cases}$$

- Pour chaque $x \in [m + (i - 1)\nu, m + i\nu]$ et pour chaque $i = 1, \dots, b$, la densité estimée de x est donnée par :

$$\hat{f}_b^{\text{Hist}}(x) := \frac{C_i}{N\nu}$$

\hat{f}_b^{Hist} est-elle une densité de probabilité ?

EXERCICE. Démontrer que :

1 $\forall x, \hat{f}_b^{\text{Hist}} \geq 0$

2 \hat{f}_b^{Hist} est intégrable

3 $\int_{\text{support}} \hat{f}_b^{\text{Hist}}(x) dx = 1$

SOLUTION 3.

$$\begin{aligned} \int_{\text{support}} \hat{f}_b^{\text{Hist}}(x) dx &= \int_m^{m+l} \hat{f}_b^{\text{Hist}}(x) dx \\ &= \sum_{i=1}^b \int_{m+(i-1)\nu}^{m+i\nu} \frac{C_i}{N\nu} dx \\ &= \frac{1}{N\nu} \sum_{i=1}^b C_i \nu = 1 \end{aligned}$$

ESTIMATEUR PAR HISTOGRAMME

Quels paramètre avons-nous du fixer ?

- La valeur (ou coordonnée) m
- Le nombre total d'intervalles (ou boîtes) b
- La longueur (ou volume) de chaque intervalle/boîte v

Nous avons donc du fixer le support, et la granularité de recouvrement de l'espace.

ESTIMATEUR PAR HISTOGRAMME

Quels paramètre avons-nous du fixer ?

- La valeur (ou coordonnée) m
- Le nombre total d'intervalles (ou boîtes) b
- La longueur (ou volume) de chaque intervalle/boîte v

Nous avons donc du fixer le support, et la granularité de recouvrement de l'espace.

Nous allons reprendre l'exemple vu précédemment, et voir empiriquement l'effet de ces choix. Nous allons aussi observer combien la taille de l'échantillon va jouer sur notre estimation.

- Télécharger le fichier `TPI_Histogramme.ipynb` : ibalelli.github.io → Teaching → Modélisation statistique avancée
- Ouvrir un terminal, aller dans le dossier où vous avez enregistré le fichier → jupyter notebook

ESTIMATEUR PAR HISTOGRAMME : RISQUE

Nous avons vu que une modification des paramètres b , le nombre de partitions, et $\nu = \frac{l}{b}$, la taille de chaque partition, peut avoir un effet important sur la qualité de notre estimation (tout comme la taille de l'échantillon, que par contre nous ne pouvons pas modifier a priori). Pouvons-nous optimiser ces deux paramètres ?

ESTIMATEUR PAR HISTOGRAMME : RISQUE

Nous avons vu que une modification des paramètres b , le nombre de partitions, et $\nu = \frac{l}{b}$, la taille de chaque partition, peut avoir un effet important sur la qualité de notre estimation (tout comme la taille de l'échantillon, que par contre nous ne pouvons pas modifier a priori). Pouvons-nous optimiser ces deux paramètres ?

Nous allons tout d'abord évaluer la « qualité » de l'estimateur, et calculer son risque :

- Il faut définir une distance qui permette de caractériser l'écart entre \hat{f}_b^{Hist} et f , e.g. la distance \mathbb{L}^2
- Il faut définir une fonction de perte, e.g. la fonction de perte quadratique ($\omega: u \mapsto u^2$)

Pour tout x^* dans la j -ème boîte, nous pouvons estimer la dépendance par rapport à la fenêtre ν de l'estimateur au point x^* comme la moyenne de l'erreur quadratique (ou **MSE**, de l'anglais *Mean Squared Error*) :

$$\text{MSE}_f(x^*, \nu) = \mathbb{E}_f[(\hat{f}_\nu^{\text{Hist}}(x^*) - f(x^*))^2]$$

Il est possible de décomposer le MSE en biais et variance :

$$\text{MSE}_f(x^*, \nu) = \underbrace{(\mathbb{E}_f[\hat{f}_\nu^{\text{Hist}}(x^*)] - f(x^*))^2}_{\text{Biais}} + \underbrace{\text{Var}[\hat{f}_\nu^{\text{Hist}}(x^*)]}_{\text{Variance}}$$

ESTIMATEUR PAR HISTOGRAMME : RISQUE

Pour tout x^* dans la j -ème boîte, nous pouvons estimer la dépendance par rapport à la fenêtre ν de l'estimateur au point x^* comme la moyenne de l'erreur quadratique (ou **MSE**, de l'anglais *Mean Squared Error*) :

$$\text{MSE}_f(x^*, \nu) = \mathbb{E}_f[(\hat{f}_\nu^{\text{Hist}}(x^*) - f(x^*))^2]$$

Il est possible de décomposer le MSE en biais et variance :

$$\text{MSE}_f(x^*, \nu) = \underbrace{(\mathbb{E}_f[\hat{f}_\nu^{\text{Hist}}(x^*)] - f(x^*))^2}_{\text{Biais}} + \underbrace{\text{Var}[\hat{f}_\nu^{\text{Hist}}(x^*)]}_{\text{Variance}}$$

Nous pouvons ensuite évaluer ces deux termes séparément. En particulier, soit p_j la probabilité d'être dans la j -ème boîte après 1 tirage (dont l'estimation empirique est $\frac{c_j}{N}$), nous avons (astuce : se ramener à une loi de Bernoulli) :

ESTIMATEUR PAR HISTOGRAMME : RISQUE

Pour tout x^* dans la j -ème boîte, nous pouvons estimer la dépendance par rapport à la fenêtre ν de l'estimateur au point x^* comme la moyenne de l'erreur quadratique (ou **MSE**, de l'anglais *Mean Squared Error*) :

$$\text{MSE}_f(x^*, \nu) = \mathbb{E}_f[(\hat{f}_\nu^{\text{Hist}}(x^*) - f(x^*))^2]$$

Il est possible de décomposer le MSE en biais et variance :

$$\text{MSE}_f(x^*, \nu) = \underbrace{(\mathbb{E}_f[\hat{f}_\nu^{\text{Hist}}(x^*)] - f(x^*))^2}_{\text{Biais}} + \underbrace{\text{Var}[\hat{f}_\nu^{\text{Hist}}(x^*)]}_{\text{Variance}}$$

Nous pouvons ensuite évaluer ces deux termes séparément. En particulier, soit p_j la probabilité d'être dans la j -ème boîte après 1 tirage (dont l'estimation empirique est $\frac{c_j}{N}$), nous avons (astuce : se ramener à une loi de Bernoulli) :

$$\mathbb{E}_f[\hat{f}_\nu^{\text{Hist}}(x^*)] = \frac{p_j}{\nu} \quad \text{et} \quad \text{Var}_f[\hat{f}_\nu^{\text{Hist}}(x^*)] = \frac{p_j(1 - p_j)}{N\nu^2}$$

Nous souhaitons maintenant avoir une estimation globale du risque de \hat{f}_b^{Hist} , c'est pourquoi nous allons considérer l'intégrale du MSE sur l'intégralité du support \rightarrow risque quadratique intégré (ou MISE, *Mean Integrated Squared Error*) :

$$\text{MISE}_f(\nu) = \int_{\text{support}} \text{MSE}_f(x, \nu) dx$$

EXERCICE. En utilisant la dérivation vue précédemment, plus le fait que $\sum_j p_j = \int_{\text{support}} f(x) dx = 1$, dériver l'expression du risque quadratique intégré (en fonction de f, p_j, N, ν)

SOLUTION

Nous avons d'une part :

I

$$\int_{\text{support}} \text{Var}_f[f_\nu^{Hist}(x)] dx = \frac{1}{N\nu} \left(1 - \sum_{j=1}^b p_j^2 \right)$$

SOLUTION

Et d'autre part :

$$2 \int_{\text{support}} \left(\mathbb{E}_f[f_{\nu}^{Hist}(x)] - f(x) \right)^2 dx = \int_{\text{support}} f^2(x) dx - \frac{1}{\nu} \sum_{j=1}^b p_j^2$$

SOLUTION

Ce qui nous amène enfin au résultat suivant :

$$\text{MISE}_f(\nu) = \int_{\text{support}} f^2(x) dx + \frac{1}{N\nu} - \frac{N+1}{N\nu} \sum_{j=1}^b p_j^2$$

SOLUTION

Ce qui nous amène enfin au résultat suivant :

$$\text{MISE}_f(\nu) = \underbrace{\int_{\text{support}} f^2(x) dx}_{\|f\|_2^2} + \frac{1}{N\nu} - \frac{N+1}{N\nu} \sum_{j=1}^b p_j^2$$

$$\|f\|_2^2$$

ESTIMATEUR PAR HISTOGRAMME : RISQUE

Sous de conditions de régularité de la densité f , et en supposant de définir ν en fonction de la taille de l'échantillon N de sorte que $\nu_N \mapsto 0$ quand $N \mapsto +\infty$, nous pouvons démontrer le résultat asymptotique suivant :

$$\text{MISE}_f(\nu) = \underbrace{\frac{\nu^2}{12} \int_{\text{support}} f'(x) dx + \frac{1}{N\nu}}_{\text{Terme principale du risque}} + \underbrace{\mathcal{O}(\nu^3) + \mathcal{O}\left(\frac{1}{N}\right)}_{\text{Terme résiduel}} \text{ lorsque } N \rightarrow +\infty$$

Le terme principale du risque est minimisé pour :

$$\nu_N^{\text{opt}} = \left(\frac{N}{6} \int_{\text{support}} f'(x) dx \right)^{-1/3}$$

ESTIMATEUR PAR HISTOGRAMME : RISQUE

Sous de conditions de régularité de la densité f , et en supposant de définir ν en fonction de la taille de l'échantillon N de sorte que $\nu_N \mapsto 0$ quand $N \mapsto +\infty$, nous pouvons démontrer le résultat asymptotique suivant :

$$\text{MISE}_f(\nu) = \underbrace{\frac{\nu^2}{12} \int_{\text{support}} f'(x) dx + \frac{1}{N\nu}}_{\text{Terme principale du risque}} + \underbrace{\mathcal{O}(\nu^3) + \mathcal{O}\left(\frac{1}{N}\right)}_{\text{Terme résiduel}} \text{ lorsque } N \rightarrow +\infty$$

Le terme principale du risque est minimisé pour :

$$\nu_N^{\text{opt}} = \left(\frac{N}{6} \int_{\text{support}} f'(x) dx \right)^{-1/3}$$

Même si cette fenêtre optimale ne peut pas être déterminée précisément (car f est inconnue), ce résultat nous permet de conclure que lorsque N est grand, la fenêtre optimale ν_N doit être de l'ordre de $N^{-1/3}$

Nous cherchons maintenant à estimer la fenêtre optimale de notre histogramme, en estimant le risque uniquement à partir des observations.

Nous cherchons à définir un estimateur \hat{J} de $MISE_f(\nu) - ||f||_2^2$ qu'il soit sans biais. Pour toute densité f et tout ν . Soit :

$$\begin{aligned} J(\nu, x_1, \dots, x_N) &= MISE_f(\nu) - ||f||_2^2 \\ &= \frac{1}{N\nu} - \frac{N+1}{N\nu} \sum_{j=1}^b p_j^2 \end{aligned}$$

Afin que \hat{J} estimateur de J soit sans biais, il suffit d'avoir un estimateur sans biais de p_j^2 pour tout j , en se rappelant que p_j correspond à la fréquence théorique des observations se situant dans le j -ème intervalle.

ESTIMATEUR PAR HISTOGRAMME : FENÊTRE OPTIMALE PAR VALIDATION CROISÉE

Un approche naïf consiste à estimer p_j^2 par \hat{p}_j^2 , où \hat{p}_j est la fréquence empirique, c'est-à-dire $\hat{p}_j := C_j/N$, en suivant les notations vu plus tôt dans ce cours.

EXERCICE. Calculer le biais de \hat{p}_j^2

Rappel : $C_j \sim \text{Binom}(N, p_j)$

SOLUTION

Etant donné que $C_j \sim \text{Binom}(N, p_j)$, il en découle :

$$\mathbb{E}[\hat{p}_j] = p_j; \text{Var}[\hat{p}_j] = \frac{p_j(1 - p_j)}{N}$$

Et par conséquent :

$$\mathbb{E}[\hat{p}_j^2] = \text{Var}[\hat{p}_j] + (\mathbb{E}[\hat{p}_j])^2 = p_j^2 \left(1 - \frac{1}{N}\right) + \frac{p_j}{N}$$

QUESTION. L'estimateur \hat{p}_j^2 est-il un estimateur sans biais de p_j^2 ?

ESTIMATEUR PAR HISTOGRAMME : FENÊTRE OPTIMALE PAR VALIDATION CROISÉE

Nous pouvons en déduire les observations suivantes :

- \hat{p}_j^2 est un estimateur biaisé de p_j^2
- $\hat{p}_j^2 - \hat{p}_j/N$ est un estimateur sans biais de $p_j^2(1 - 1/N)$
- D'où :

$$\tilde{p}_j^2 := \frac{\hat{p}_j^2 - \hat{p}_j/N}{1 - 1/N} = \frac{N}{N-1} \hat{p}_j^2 - \frac{1}{N-1} \hat{p}_j$$

est un estimateur asymptotiquement ($N \mapsto \infty$) sans biais de p_j^2 .

Nous pouvons en déduire les observations suivantes :

- \hat{p}_j^2 est un estimateur biaisé de p_j^2
- $\hat{p}_j^2 - \hat{p}_j/N$ est un estimateur sans biais de $p_j^2(1 - 1/N)$
- D'où :

$$\tilde{p}_j^2 := \frac{\hat{p}_j^2 - \hat{p}_j/N}{1 - 1/N} = \frac{N}{N-1} \hat{p}_j^2 - \frac{1}{N-1} \hat{p}_j$$

est un estimateur asymptotiquement ($N \mapsto \infty$) sans biais de p_j^2 .

Ceci nous permet enfin de proposer l'estimateur suivant (rappel : $\sum_{j=1}^b \hat{p}_j = 1$) :

$$\hat{J}(\nu, x_1, \dots, x_N) = \frac{2}{(N-1)\nu} - \frac{N+1}{(N-1)\nu} \sum_{j=1}^b \tilde{p}_j^2, \quad \hat{p}_j = C_j/N$$

ESTIMATEUR PAR HISTOGRAMME : FENÊTRE OPTIMALE PAR VALIDATION CROISÉE

L'estimateur \hat{f} peut être utilisé pour déterminer automatiquement la fenêtre optimale ν par la méthode de validation croisée :

1. Poser : $m = \min_i x_i; l = \max_i x_i - m, i = 1, \dots, N$
2. Initialiser : $\nu_{CV} = 1, b_{CV} = 1, \hat{f}_{CV} = -1$
3. Tant que $b < N$:
 - Calculer $J := \hat{f}(l/b, x_1, \dots, x_N)$
 - Si $J < \hat{f}$:
 - $b_{CV} = b$
 - $\hat{f}_{CV} = J$
 - $m += 1$
4. Et enfin : $\nu_{CV} = l/b_{CV}$

→ Reprendre le TP, **Partie II.**

Quels avantages/inconvénients de l'estimateur par histogramme ?

Quels avantages/inconvénients de l'estimateur par histogramme ?

- Très naturel et intuitif → Simple à réaliser et à analyser
- Le résultat dépend fortement de l'échantillon ainsi que de certains paramètres qui peuvent être choisis a priori, comme en particulier la partition considéré du support (ou plus précisément, de la partie du support où l'on possède des données)

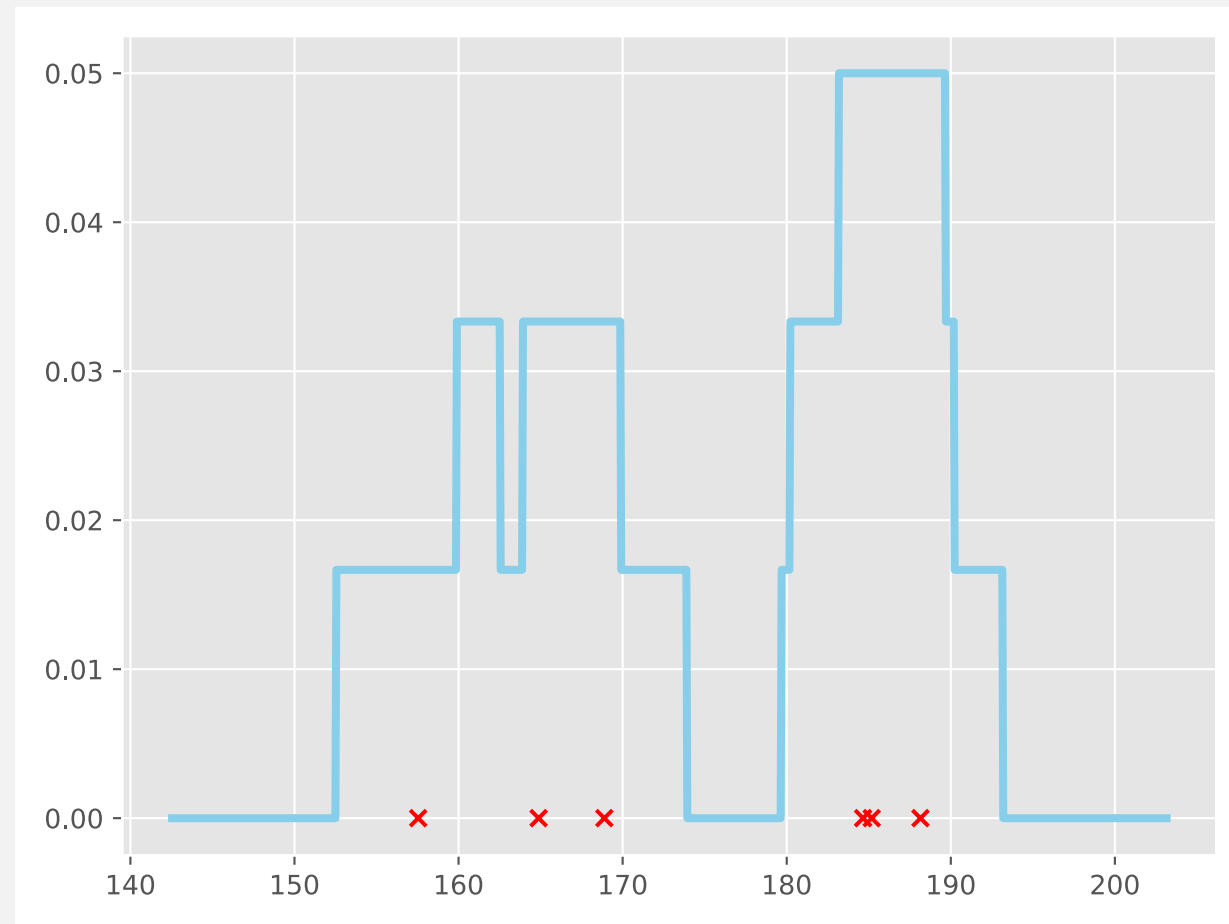
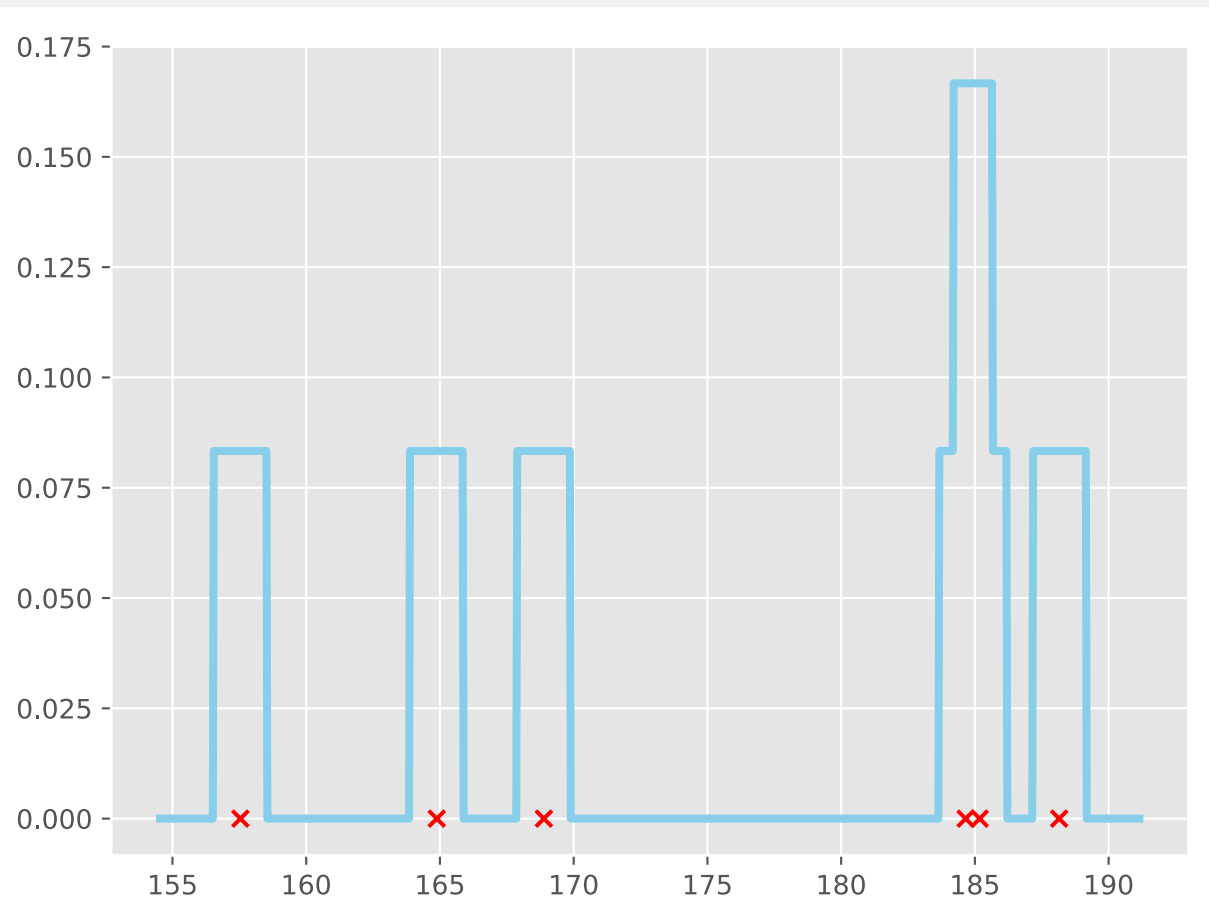
Une première idée pour résoudre un des limites de cette méthodes (le choix du support et de sa partition), et celle de supposer que chaque x_i dans notre échantillon est le centre d'une classe de l'histogramme de longueur ν .

Cela nous amène à l'expression suivante :

$$\hat{f}_\nu^{\mathcal{U}} := \frac{1}{N\nu} \sum_{i=1}^N \mathbb{I}(|x_i - x| \leq \nu/2) = \frac{1}{N\nu} \sum_{i=1}^N \mathbb{I}\left(\frac{|x_i - x|}{\nu} \leq \frac{1}{2}\right)$$

Cela revient, en pratique, à « construire » autour de chaque observation un « bloc » dont l'aire est égale à $1/N$, ce qui conduit à un estimateur toujours constant par morceaux, mais avec des plateaux de longueur variable.

ESTIMATEUR PAR HISTOGRAMME : CONCLUSION



Pouvons-nous rendre cette estimation plus « lisse » ?

Pouvons-nous rendre cette estimation plus « lisse » ?



Estimateur par noyaux