Logiciels scientifiques - HLMA310
Partie 1 : `Python`

Statistiques descriptives et `pandas`

Joseph Salmon

http://josephsalmon.eu

Université de Montpellier

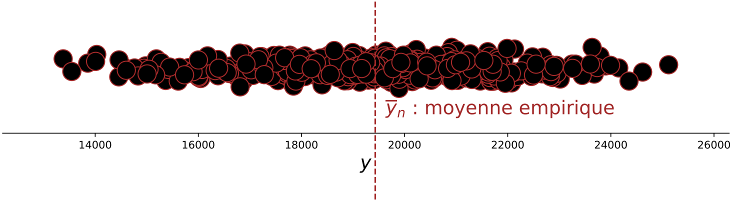
Introduction

Base de données : `belgianmunicipalities.csv`, données des revenus des foyers belges, par région et par villes

```
from download import download
url = "http://josephsalmon.eu/enseignement/datasets/belgianmunicipalities.csv"
path_target = "./belgianmunicipalities.csv"
download(url, path_target, replace=False)
```

On s'intéressera ici qu'à la variable `medianincome` dans la suite.

Moyenne (arithmétique)



Définition : Moyenne (arithmétique)

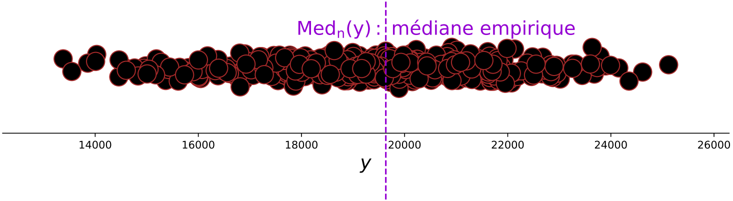
$$\overline{y}_n = \frac{1}{n} \sum_{i=1}^{n} y_i$$

Si $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{n} x_i y_i$ (produit scalaire) et $\mathbf{1}_n = (1, \ldots, 1)^\top \in \mathbb{R}^n$:

$$\overline{y}_n = \left\langle \mathbf{y}, \frac{\mathbf{1}_n}{n} \right\rangle$$

Exo: Le vecteur $\overline{y}_n \mathbf{1}_n$ est la projection de $\mathbf{y}$ sur l'espace $\mathrm{vect}(\mathbf{1}_n)$

Médiane



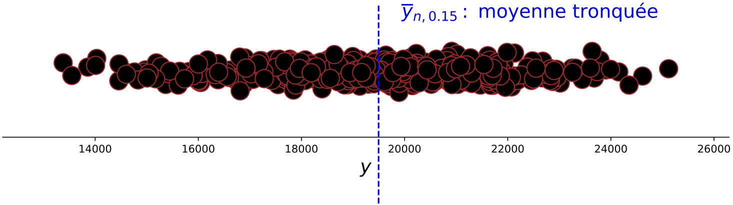
On ordonne les y_i dans l'ordre croissant : $y_{(1)} \leqslant y_{(2)} \leqslant \ldots \leqslant y_{(n)}$

Définition : Médiane

$$\mathrm{Med}_n(\mathbf{y}) = \begin{cases} \frac{y_{(\frac{n}{2})} + y_{(\frac{n}{2}+1)}}{2}, & \text{si } n \text{ est pair} \\ y_{(\frac{n+1}{2})}, & \text{si } n \text{ est impair} \end{cases}$$

Rem: la définition d'une médiane est non-unique, et peut être parfois ambigu...

Moyenne tronquée



Pour un paramètre α (*e.g.*, $\alpha = 15\%$), on calcule la moyenne en enlevant les $\alpha\%$ plus grandes et plus petites valeurs

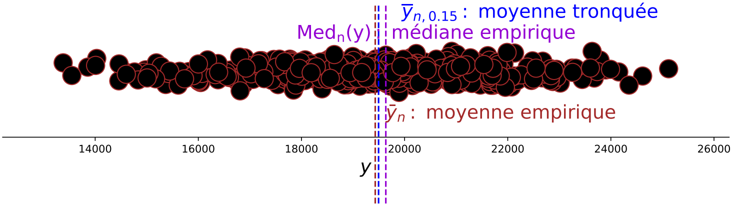
Définition : Moyenne tronquée (à l'ordre α)

$$\textbf{Moyenne tronquée} : \quad \overline{y}_{n,\alpha} = \overline{z}_n$$

où $\mathbf{z} = (y_{(\lfloor \alpha n \rfloor)}, \ldots, y_{(\lfloor (1-\alpha) n \rfloor)})$ est l'échantillon α-tronqué

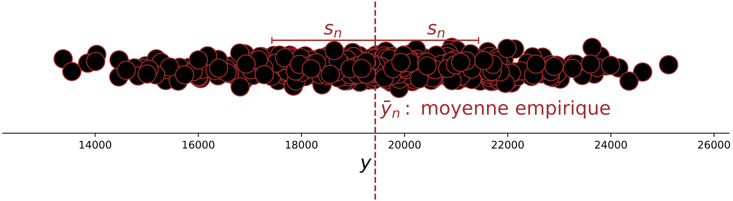
Rem: $\lfloor u \rfloor$ est le nombre entier tel que $\lfloor u \rfloor - 1 < u \leqslant \lfloor u \rfloor$

Moyenne *vs* médiane



- Les trois statistiques ne coïncident pas
- Moyennes tronquées et médianes sont robustes aux points atypiques (🇬🇧 : *outliers*), la moyenne non !

Dispersion : variance / écart-type

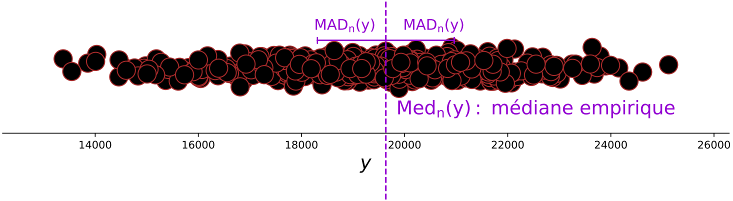


Définitions

Variance : $$\mathrm{var}_n(\mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y}_n)^2 = \frac{1}{n}\left\|\mathbf{y} - \overline{y}_n \mathbf{1}_n\right\|^2$$

Écart-type : $$s_n(\mathbf{y}) = \sqrt{\mathrm{var}_n(\mathbf{y})} \quad (\text{où } \|\mathbf{z}\|^2 = \sum_{i=1}^{n} z_i^2)$$

Dispersion : MAD

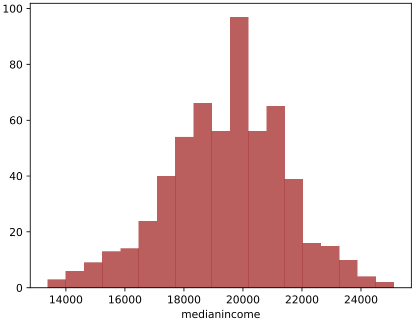


Définition

Déviation médiane absolue (🇬🇧 : *Mean Absolute Deviation*) :

$$\mathrm{MAD}_n(\mathbf{y}) = \mathrm{Med}_n\left(|\mathrm{Med}_n(\mathbf{y}) - \mathbf{y}|\right)$$

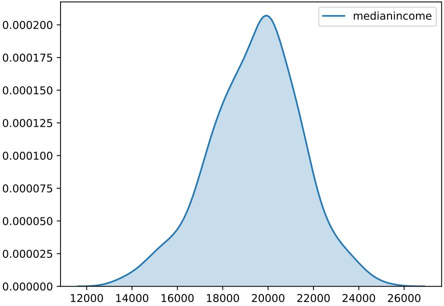
Estimation de la densité : histogramme



L'**histogramme** est une approximation de la **densité** par une fonction constante par morceaux

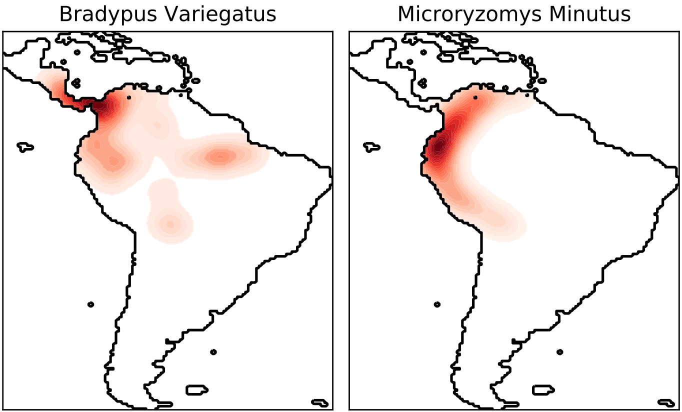
Rem: les « cases » (🇬🇧 : *bins*) ont une aire proportionnelle au nombre de données qu'elles contiennent

Estimation de la densité : méthode à noyau



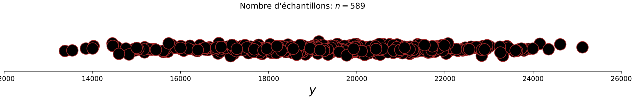
- Méthode à noyau (🇬🇧 : *Kernel Density Estimation, KDE*) : approche non-paramétrique estimant la densité par une fonction continue – généralisation de l'histogramme

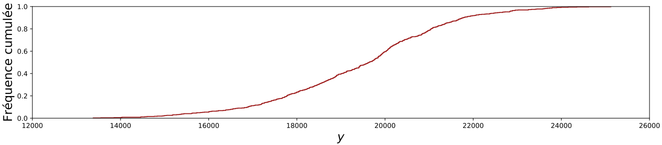
Densité bi-dimensionnelle (spatiale)



cf. http://scikit-learn.org/stable/_downloads/plot_species_kde.py

Fonction de répartition



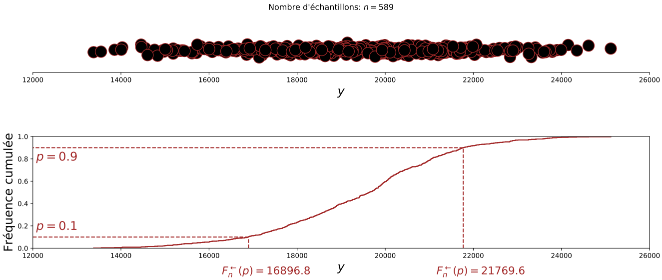


Définition : fonction de répartition

Empirique : $$F_n(u) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{y_i \leqslant u\}}$$

Interprétation : proportion d'observations sous un certain niveau

Fonction quantile



Définition : fonction quantile

$$\text{Pour } p \in]0,1], \quad F_n^{\leftarrow}(p) = \inf\{u \in \mathbb{R} : F_n(u) \geqslant p\}$$

Rem: c'est l'inverse (généralisée) de la fonction de répartition ; sa définition admet plusieurs conventions, *cf.* percentile in `numpy`