Statistique : Statistiques descriptives

Joseph Salmon

Septembre 2014

Statistique

- On observe des réalisations $(y_1, \ldots, y_n)$ de variables aléatoires inconnues (éventuellement vectorielles)
- On suppose ici que les variables sont indépendantes et identiquement distribuées (*i.i.d.*) selon une loi $\mathbb{P}_Y$

But de l'estimation

Comment apprendre certaines caractéristiques de $\mathbb{P}_Y$ à partir de $(y_1, \ldots, y_n)$?

Souvent : on se prépare à observer y_{n+1}.

Cas de la prédiction

Que peut-on attendre de y_{n+1} ? (en moyenne, ou avec une certaine probabilité ?)

Vocabulaire

- Observations $\mathbf{y} = y_{1:n} = (y_1, \ldots, y_n)$: échantillon de taille n.
- Grandeurs théoriques : dépendant de la loi $\mathbb{P}_Y$ **inconnue**
  Exemple: l'espérance de la variable y sous la loi $\mathbb{P}_Y$.
- Grandeurs empiriques : calculées à partir des observations y_i.
  Exemple: $\bar{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i$ est la moyenne empirique
- Objectif général : apprendre les caractéristiques théoriques de $\mathbb{P}_Y$ à partir de résumés empiriques.

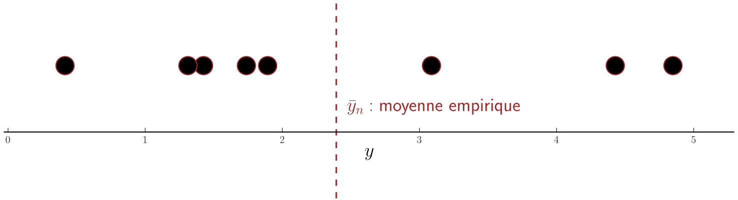
Statistique exploratoire et descriptive

- Première analyse sans hypothèse sur la loi $\mathbb{P}_Y$.
- Analyse qualitative du jeu de données /échantillon

Définition : Statistique

Une statistique est une fonction des observations $(y_1, \ldots, y_n)$.

Moyenne



Défintion : Moyenne

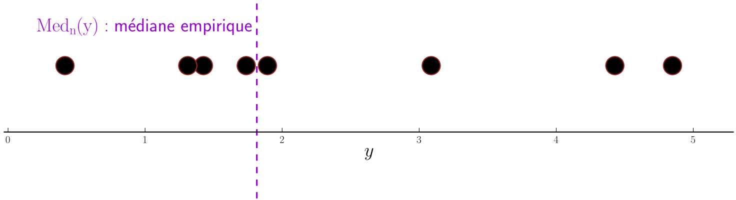
$$\overline{y}_n = \frac{1}{n} \sum_{i=1}^{n} y_i$$

Notons $\mathbf{1}_n$ le vecteur $(1, \ldots, 1) \in \mathbb{R}^n$. La moyenne est (à facteur $1/n$ près) un produit scalaire dans $\mathbb{R}^n$:

$$\overline{y}_n = \langle \mathbf{y}, \mathbf{1}_n / n \rangle$$

cf. McKinney (2012) pour les statistiques avec `python`

Médiane empirique

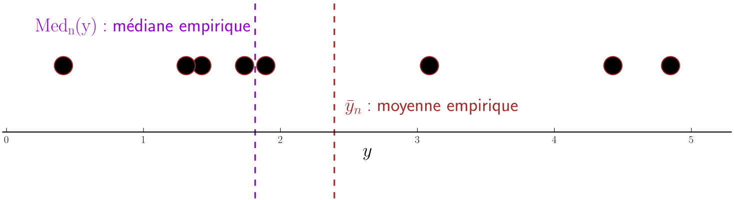


On ordonne les y_i : $y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(n)}$

Défintion : Médiane (NON-UNIQUE !)

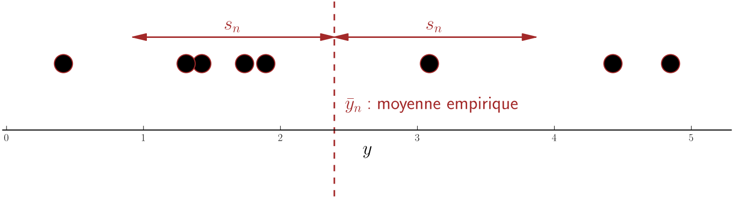
$$\mathrm{Med}_n(\mathbf{y}) = \begin{cases} \frac{y_{(\lfloor \frac{n}{2} \rfloor)} + y_{(\lfloor \frac{n}{2} \rfloor + 1)}}{2} & \text{Si } n \text{ est pair} \\ y_{(\frac{n+1}{2})} & \text{Si } n \text{ est impair} \end{cases}$$

Moyenne *vs* médiane



- Les deux statistiques ne coïncident pas
- Une médiane est plus robuste aux points atypiques (en anglais : *outliers*)

Dispersion



Variance empirique

Moyenne des écarts quadratiques à la moyenne (empirique)

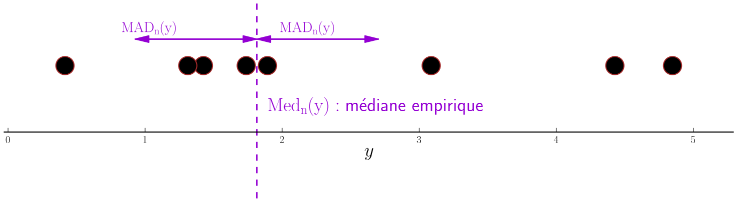
$$\mathrm{var}_n(\mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y}_n)^2 = \frac{1}{n}\left\|\mathbf{y} - \overline{y}_n\mathbf{1}_n\right\|^2$$

($\|\cdot\|$: norme euclidienne dans $\mathbb{R}^n$)

Écart-type empirique

$$s_n(\mathbf{y}) = \sqrt{\mathrm{var}_n}(\mathbf{y}) \qquad \left(= \frac{1}{\sqrt{n}}\left\|\mathbf{y} - \overline{y}_n\mathbf{1}_n\right\| \right)$$

Dispersion

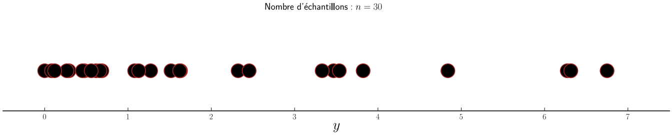


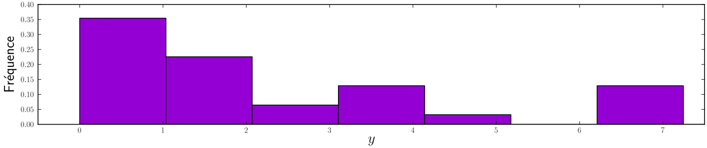
Mean Absolute deviation

Déviation médiane absolue :

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

Histogramme

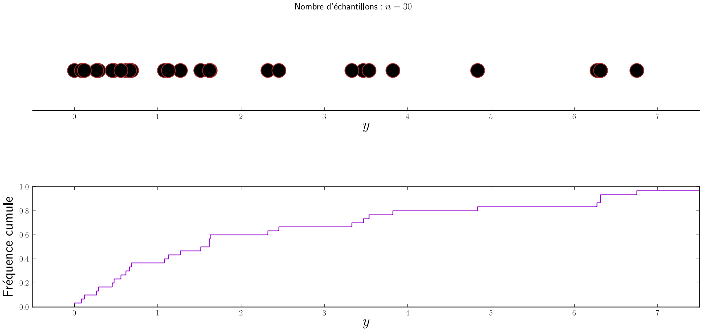




Répartition des données dans des « cases »
L'aire de chaque case est proportionnelle à la fraction des données qui « tombent » dans la case.

L'histogramme est une approximation de la **densité** de y

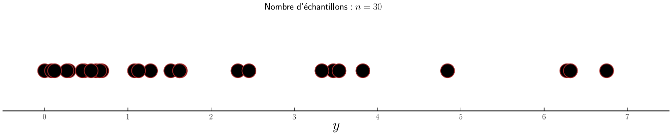
Fonction de répartition empirique

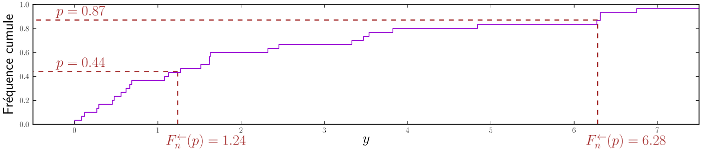


- *Rappel :* Fonction de répartition : $F(u) = \mathbb{P}_Y(-\infty, u]$
- Version empirique : proportion des données en-dessous de u

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

Quantiles empiriques





- Inverse de la fonction de répartition empirique.
- Soit $\lceil u \rceil$ le nombre entier tel que $\lceil u \rceil - 1 < u \leq \lceil u \rceil$.

Quantiles empiriques

$$\begin{aligned} \text{quantile d'ordre } p &= y_{(\lceil np \rceil)} \quad (p \in [0,1]) \\ &= F_n^{\leftarrow}(p) \end{aligned}$$

Covariance et corrélation empirique

Covariance empirique

Pour deux échantillons $x_{1:n}$ et $y_{1:n}$ de moyennes et variances empiriques $\mathbf{x} = \overline{x}_n$, $\mathbf{y} = \overline{y}_n$ et $\mathrm{var}_n(\mathbf{x})$, $\mathrm{var}_n(\mathbf{y})$:

$$\mathrm{cov}_n(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n)(y_i - \overline{y}_n) \quad \text{c'est-à-dire}$$

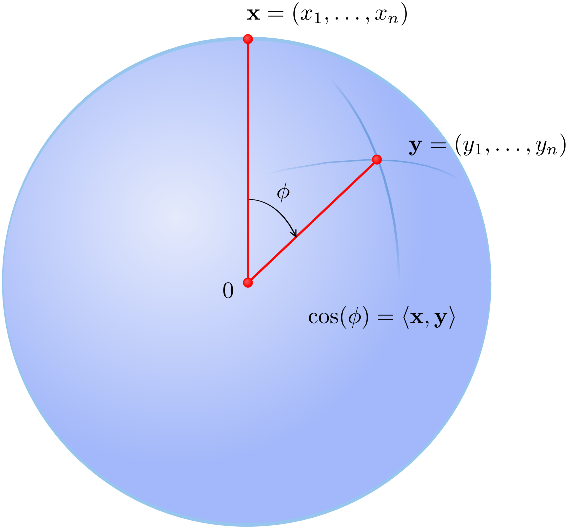
$$\mathrm{cov}_n(x, y) = \frac{1}{n}\langle x_{1:n} - \overline{x}_n \mathbf{1}_n, y_{1:n} - \overline{y}_n \mathbf{1}_n \rangle$$

Corrélation empirique

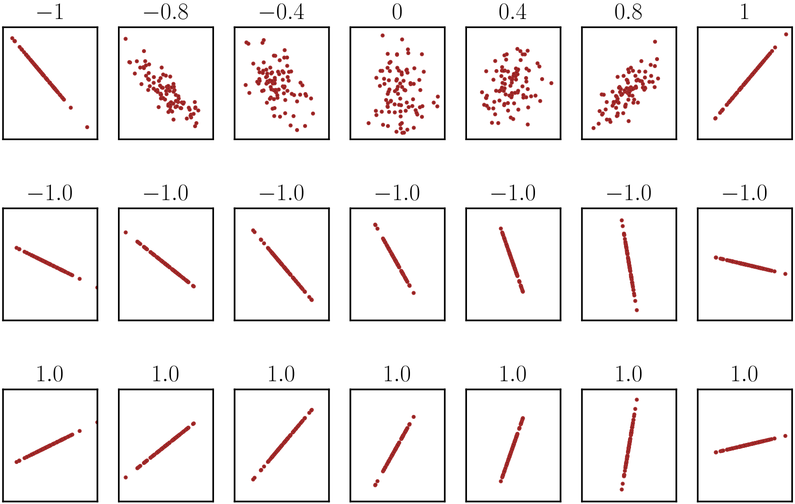
$$\rho = \mathrm{corr}_n(x, y) = \frac{\mathrm{cov}_n(x, y)}{\sqrt{\mathrm{var}_n(\mathbf{x})}\sqrt{\mathrm{var}_n(\mathbf{y})}}, \quad \text{c'est-à-dire}$$

$$\rho = \frac{\langle x_{1:n} - \overline{x}_n \mathbf{1}_n, y_{1:n} - \overline{y}_n \mathbf{1}_n \rangle}{\|x - \overline{x}_n\| \, \|y - \overline{y}_n\|} = \cos(x_{1:n} - \overline{x}_n \mathbf{1}_n, y_{1:n} - \overline{y}_n \mathbf{1}_n)$$

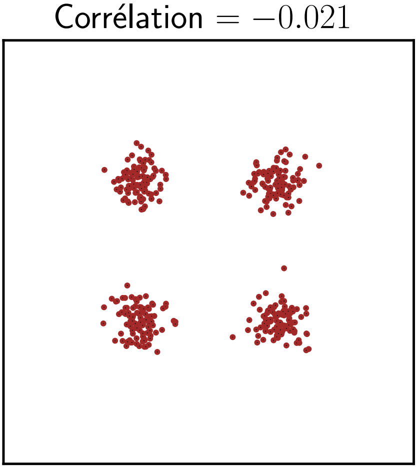
Interprétation pour $n = 3$ et $\|\mathbf{x}\| = \|\mathbf{y}\| = 1$



Exemples de corrélations

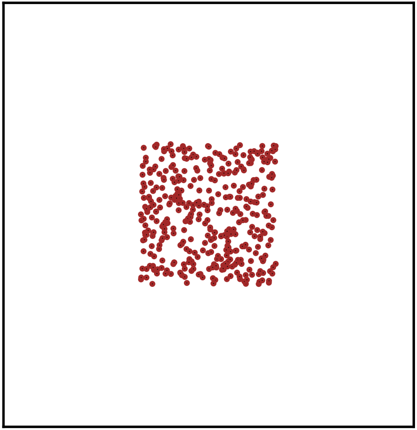


Exemples de corrélations proches de zéros



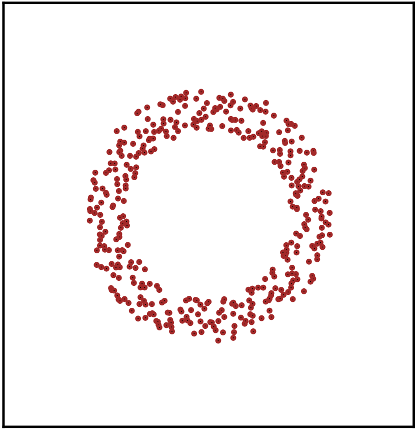
Exemples de corrélations proches de zéros

Corrélation $= 0.007$

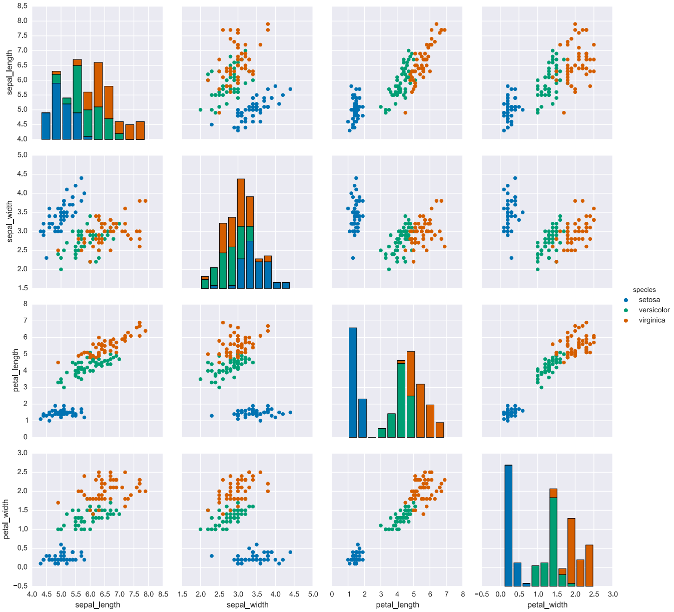


Exemples de corrélations proches de zéros

Corrélation $= 0.011$



Exemples de visualisation



Références I

- W. McKinney.

  Python for Data Analysis : Data Wrangling with Pandas, NumPy, and IPython.

  O'Reilly Media, 2012.