Apprentissage non supervisé : Bases statistiques

Pavlo Mozharovskyi

LTCI, Telecom Paris, Institut Polytechnique de Paris

Parcours Data Science BPCE

Paris, le 5 juin 2023

Enseignants

- **Pavlo Mozharovskyi** :
  - Précédemment : Institut Polytechnique de Kiev (bachelor et master), Université de Cologne (PhD, post-doc), Agrocampus Rennes (post-doc), École Nationale de la Statistique et de l'Analyse de l'Information (Assistant Professor).
  - Spécialités : machine learning statistique, intelligence artificielle explicable, statistiques robustes, statistique computationnelle, profondeur des données.
    Applications : classification (supervisée et non-supervisée), détection d'anomalies, analyse d'efficacité.
  - Email : *pavlo.mozharovskyi@telecom-paris.fr*
  - Bureau : 5C30.

- **Thomas Belhalfaoui** :
  - Études : Télécom Paris et ENS Paris-Saclay (master MVA).
  - Précédemment : Ministère des Armées, JobTeaser, Shift Technology.
  - Spécialités : machine learning, réseaux de neurones, systèmes de recommandation, metric learning.
  - E-mail : *belhalfaoui@gmail.com*

Plan du cours

- **Séance 1.** 05/06 matin : Cours *Bases statistiques*.
- **Séance 2.** 05/06 après-midi : TP (`R` + `Python`).
- **Séance 3.** 06/06 matin : Cours *Clustering*.
- **Séance 4.** 06/06 après-midi : TP (`Python`).
- **Séance 5.** 13/06 matin : Cours *Détection d'anomalies*.
- **Séance 6.** 13/06 après-midi : TP (`R` + `Python`).

Littérature supplémentaire - à revoir seul

- Bases de **probabilités** : probabilité, densité, espérance, loi des grands nombres, lois gaussiennes, théorème central limite
  Lecture : Foata et Fuchs (1996)
- Bases de l'**optimisation** : fonctions convexes, condition du premier ordre, descente de gradient, méthode de Newton
  Lecture : Boyd et Vandenberghe (2004), Bertsekas (1999)
- Bases de l'**algèbre (bi-)linéaire** : espaces vectoriels, normes, produit scalaire, matrices, déterminants, diagonalisation
  Lecture : Horn et Johnson (1994)
- Bases de l'**algèbre linéaire numérique** : résolution de système, factorisation de matrices, conditionnement, etc.
  Lecture : Golub et VanLoan (2013)

Cadre statistique standard

On notera $\mathbb{P}, \mathbb{E}$ pour probabilité et l'espérance

- 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 $\mathbb{P}_Y$

Rem : on note souvent Y une variable aléatoire et y une réalisation

Estimation

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

Prédiction

On se prépare à observer y_{n+1} : comment approcher y_{n+1}, quantifier une incertitude sur cette grandeur, etc. ?

Vocabulaire

- Observations $y = y_{1:n} = (y_1, \dots, y_n)$: **échantillon** de **taille** n
- Grandeurs **théoriques** : dépendent de la loi $\mathbb{P}_Y$ (**inconnue**) et contrôlent la génération des observations
  Example : l'espérance $\mathbb{E}(Y)$ ou la variance $\mathbb{V}\mathrm{ar}(Y)$ de Y
- Grandeurs **empiriques** : calculées à partir des observations y_i
  Example : la moyenne empirique $\bar{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i$
- Objectif général : apprendre les caractéristiques théoriques de $\mathbb{P}_Y$ à partir de résumés empiriques.

Rem : les grandeurs théoriques dépendent de $\mathbb{P}_Y$ alors que les grandeurs empiriques dépendent de $\mathbb{P}_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{y_i}$ (ici δ_{y_i} est la mesure de Dirac au point y_i)

Modèle statistique : contexte

Rappel

- 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 $\mathbb{P}_Y$

- Selon la situation, la loi $\mathbb{P}_Y$ a certaines caractéristiques. Example : "Pile ou face" : on sait que $\mathbb{P}_Y = \text{Bernoulli}(\theta)$ pour un certain $\theta \in [0, 1]$ inconnu

- Reformulation : on dispose d'une **famille de lois candidates**, (parfois naturelle) pour $\mathbb{P}_Y$
  Example : la famille des lois de Bernoulli

Exo : Quel est un modèle naturel pour "un lancer de dé" ?

Modèle statistique

- La loi cible $\mathbb{P}_Y$ est indexée par un **paramètre** $\theta \in \Theta$: $\mathbb{P}_Y = \mathbb{P}_\theta$ pour un θ inconnu, et Θ est l'ensemble d'indexation

  Example : "Pile ou face", $\theta \in \Theta = [0, 1]$ et $\mathbb{P}_\theta = \text{Bernoulli}(\theta)$

Définition

Un **modèle statistique** est une famille de lois

$$\mathcal{M} = \{\mathbb{P}_\theta : \theta \in \Theta\}$$

indexée par un ensemble de paramètres Θ.

Exo : Proposer un modèle $\mathcal{M}$ pour le "lancer de dé".

Modèle statistique paramétrique

Définition

Un **modèle paramétrique** est une famille de lois $\mathcal{M} = \{\mathbb{P}_\theta : \theta \in \Theta\}$ indexée par un nombre fini p de paramètres : $\Theta \subset \mathbb{R}^p$. On note aussi $\mathbb{E}_\theta$ l'espérance associée.

Rem : le modèle est indexé par un nombre ou un vecteur réel ; p est la dimension du modèle

Example :

- Modèle de Bernoulli (ou "Pile ou face") : $\Theta = [0, 1]$.
- Modèle gaussien : $\theta = (\mu, \sigma^2)$, $\Theta = \mathbb{R} \times \mathbb{R}_+^*$.

Rem : le modèle est dit **non-paramétrique** s'il n'est pas indexable par un paramètre de dimension finie, *e.g.* $\{f : \int f = 1, \text{ et } f \geq 0\}$

Rem : dans le cadre **fréquentiste**, on suppose qu'il existe un vrai paramètre inconnu, tel que $\mathbb{P}_Y = \mathbb{P}_\theta$

Estimateur

- *Objectif :* estimer une quantité $g = g(\theta)$ qui ne dépend que de la loi $\mathbb{P}_\theta$ des observations. g est une constante inconnue **déterministe** *i.e.* non aléatoire.

  Example : espérance, quantile, variance, écart-type, etc.

- *Intuition :* un **estimateur** $\widehat{g}$ est calculé à partir de l'échantillon $(y_1, \ldots, y_n)$, dans le but d'approcher $g(\theta)$.

Définition

Un **estimateur** $\widehat{g}$ de g est une fonction des observations :

$$\widehat{g} : (y_1, \ldots, y_n) \mapsto \widehat{g}(y_1, \ldots, y_n)$$

Rem : un estimateur est parfois aussi appelé une **statistique**

Rem : en pratique l'estimateur doit être calculable efficacement

Propriétés d'un estimateur : le biais

Définition

Le **biais** d'un estimateur $\widehat{g}$ est l'espérance de son écart au paramètre :

$$\text{Biais}(\widehat{g}, g) = \mathbb{E}_\theta(\widehat{g}(Y_1, \ldots, Y_n)) - g(\theta) \qquad \text{(dépend de } \theta\text{)}$$

Définition

Un estimateur $\widehat{g}$ de g est dit **non biaisé** (ou **sans biais**) si :

$$\forall \theta \in \Theta, \quad \mathbb{E}_\theta(\widehat{g}(Y_1, \ldots, Y_n)) = g(\theta)$$

Rem : le biais mesure l'erreur systématique d'un estimateur

Estimateur sans biais de l'espérance

- L'espérance 'théorique' dépend de la loi $\mathbb{P}_\theta$
- On cherche ici à estimer $g(\theta) = \mathbb{E}_\theta(Y)$

Théorème

Sous l'hypothèse que l'échantillon est *i.i.d.*, la moyenne empirique $\overline{y}_n = \frac{1}{n}\sum_{i=1}^n y_i$ est un estimateur sans biais de l'espérance $\mathbb{E}(Y)$

Démonstration :

$$\mathbb{E}_\theta\left(\frac{1}{n}\sum_{i=1}^n Y_i\right) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}(Y_i) = \mathbb{E}(Y)$$

car $\mathbb{E}(Y_i) = \mathbb{E}(Y)$ (caractère *i.i.d.* des Y_i)

Rem : $\widehat{g}(y_1, \ldots, y_n) = y_1$ est un estimateur sans biais de l'espérance

Estimateur sans biais de la variance

- La variance 'théorique' dépend de la loi $\mathbb{P}_\theta$
- On cherche ici à estimer $g(\theta) = \mathbb{V}\mathrm{ar}_\theta(Y)$

Théorème

L'estimateur $\widehat{g}(y_1, \ldots, y_n) = \dfrac{1}{n-1} \displaystyle\sum_{i=1}^{n} (y_i - \overline{y}_n)^2$ est un estimateur sans biais de la variance $\mathbb{V}\mathrm{ar}_\theta(Y)$

Rem : l'estimateur $\dfrac{1}{n} \displaystyle\sum_{i=1}^{n} (y_i - \overline{y}_n)^2$ est lui biaisé

Exo : Vérifier cette propriété par le calcul

Propriétés d'un estimateur : la variance

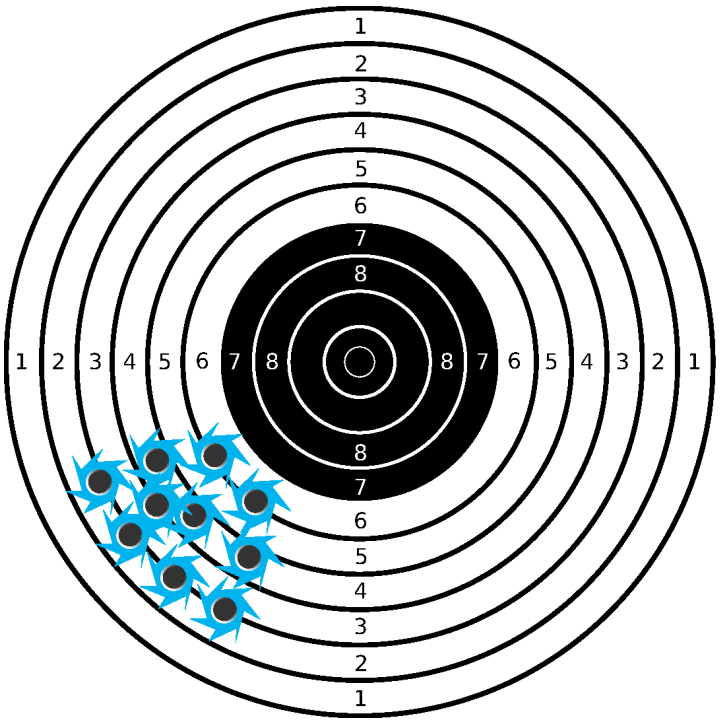
Définition

La **variance** d'un estimateur $\widehat{g}$ est sa variance théorique :

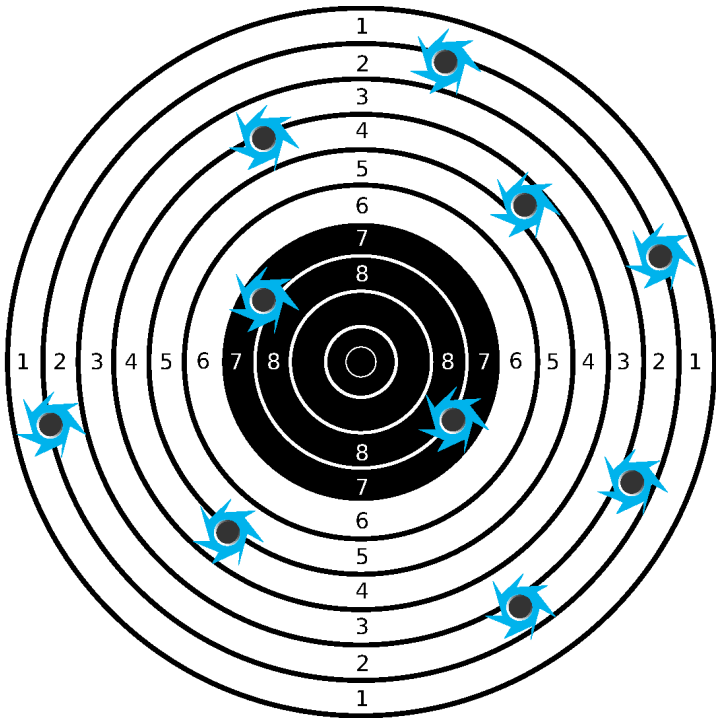
$$\mathbb{V}\mathrm{ar}_\theta(\widehat{g}) = \mathbb{V}\mathrm{ar}_\theta(\widehat{g}(Y_1, \ldots, Y_n)) = \mathbb{E}_\theta(\widehat{g} - \mathbb{E}_\theta(\widehat{g}))^2 \quad (\text{dépend de } \theta)$$

Rem : la variance mesure la dispersion autour de l'espérance

Biais ou variance ?

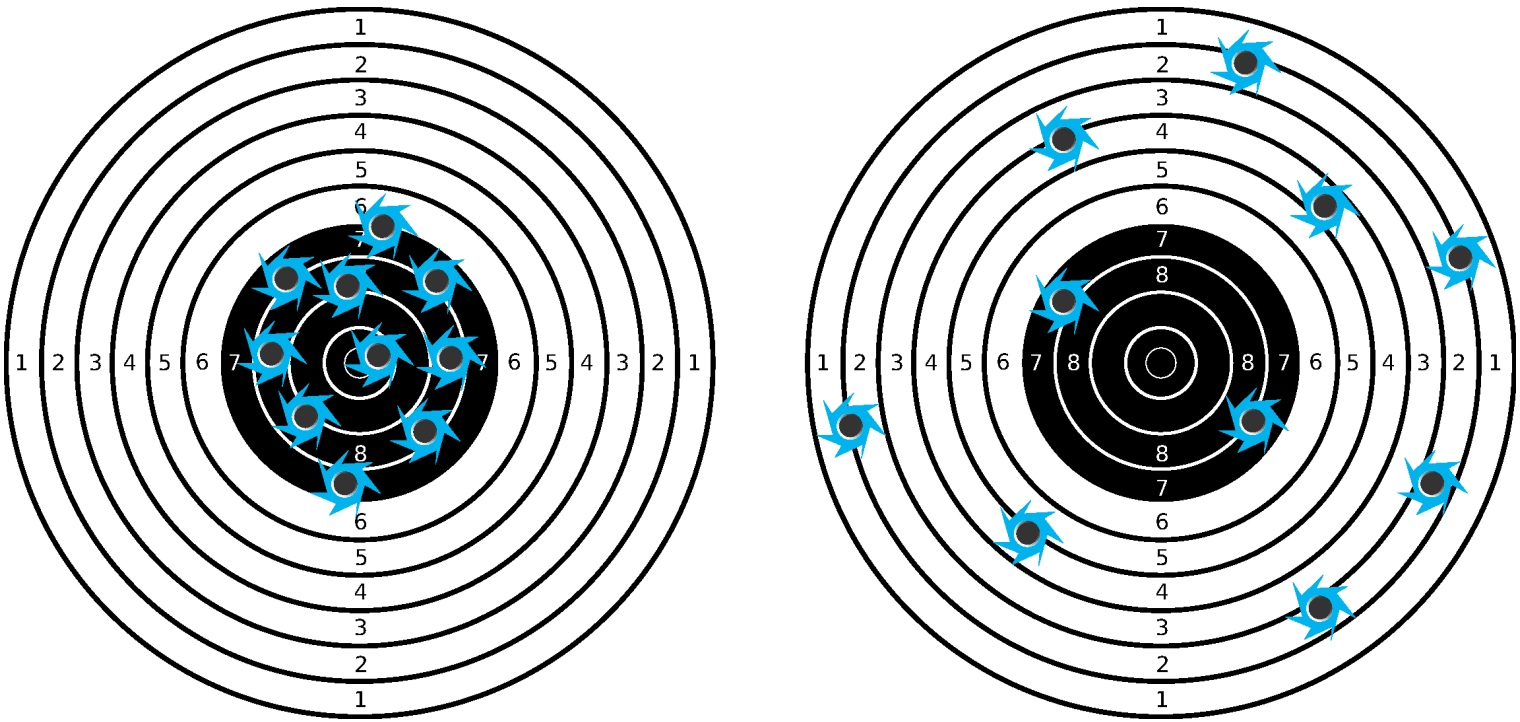


Erreurs systématiques



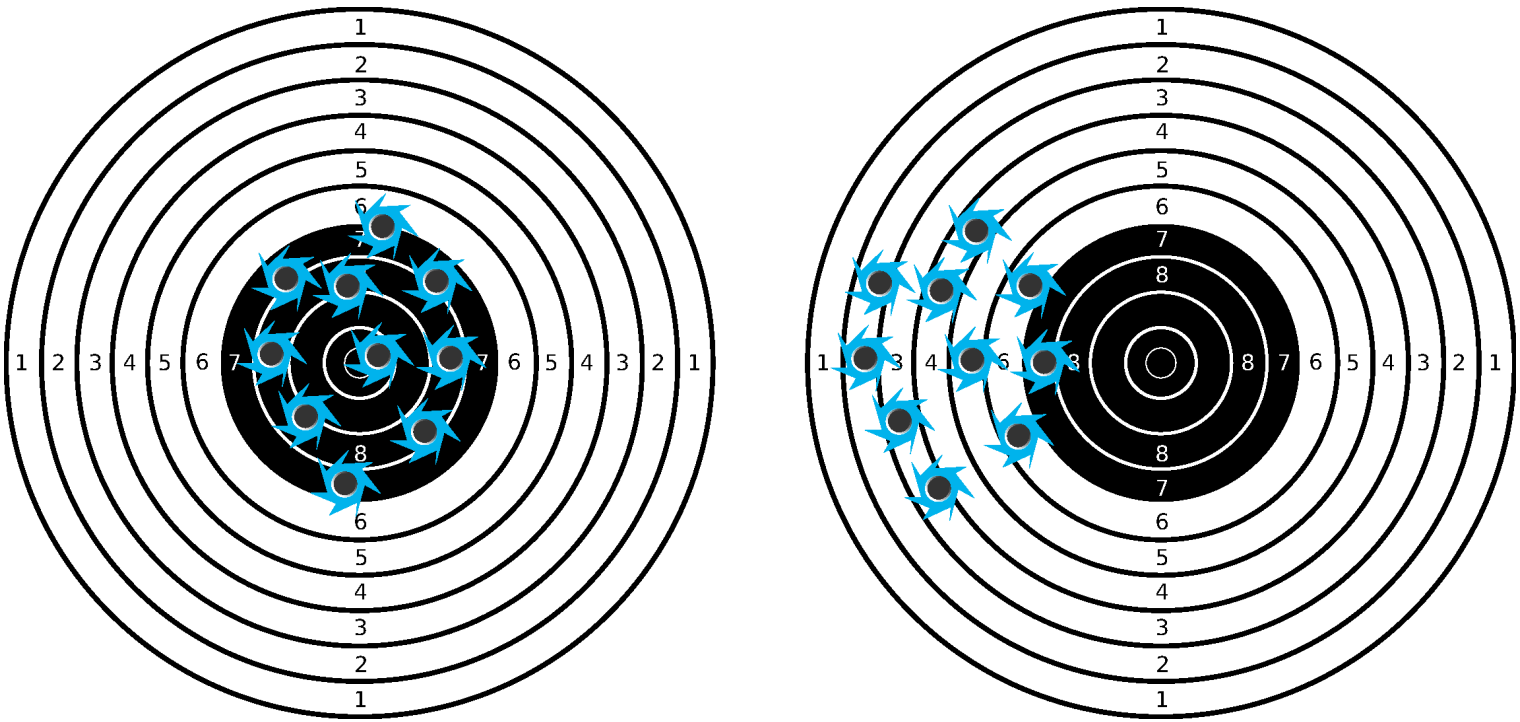
Erreurs stochastiques

Biais ou variance ?



- Si $\widehat{g_0}$ et $\widehat{g_1}$ sont sans biais, on préfère avoir une faible variance

Biais ou variance ?



- Si $\widehat{g_0}$ et $\widehat{g_1}$ ont la même variance, on préfère un biais faible

Risque quadratique / compromis biais-variance

Définition

Le **risque quadratique** d'un estimateur $\widehat{g}$ est l'espérance de son erreur au carré :

$$R(\widehat{g}) = \mathbb{E}\left[(\widehat{g} - g)^2\right]$$

Règle de choix : prendre l'estimateur dont le risque est le plus petit

Théorème : décomposition biais / variance

$$\text{Risque}(\widehat{g}) = \text{Variance}(\widehat{g}) + \big(\text{Biais}(\widehat{g})\big)^2$$

Démonstration : faire apparaître le biais $B = \mathbb{E}(\widehat{g}) - g$; développer

$$\begin{aligned} R(\widehat{g}) &= \mathbb{E}\left[(\widehat{g} - \mathbb{E}(\widehat{g}) + B)^2\right] \\ &= \mathbb{E}\left[(\widehat{g} - \mathbb{E}(\widehat{g}))^2 + B^2 + 2B(\widehat{g} - \mathbb{E}(\widehat{g}))\right] \\ &= \mathbb{V}\text{ar}(\widehat{g}) + B^2 + 2B \underbrace{\mathbb{E}\left[\widehat{g} - \mathbb{E}(\widehat{g})\right]}_{=0} = \mathbb{V}\text{ar}(\widehat{g}) + B^2 \end{aligned}$$

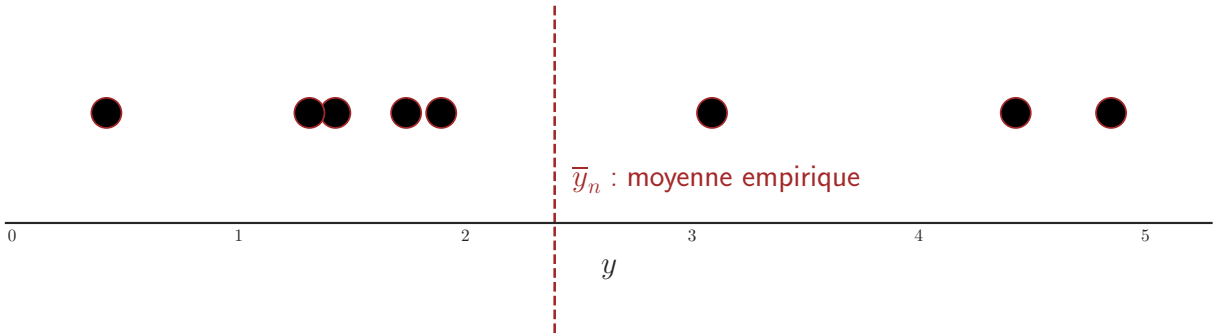
Statistique exploratoire et descriptive

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

Rappel : **statistique** = **estimateur**, c'est une fonction (mesurable) des observations $(y_1, \ldots, y_n)$ (et qu'on espère être une fonction calculable des observations $(y_1, \ldots, y_n)$!)

Rem : les enjeux computationnels seront a prendre en compte dans la plupart de vos applications pratiques

Moyenne (arithmétique)



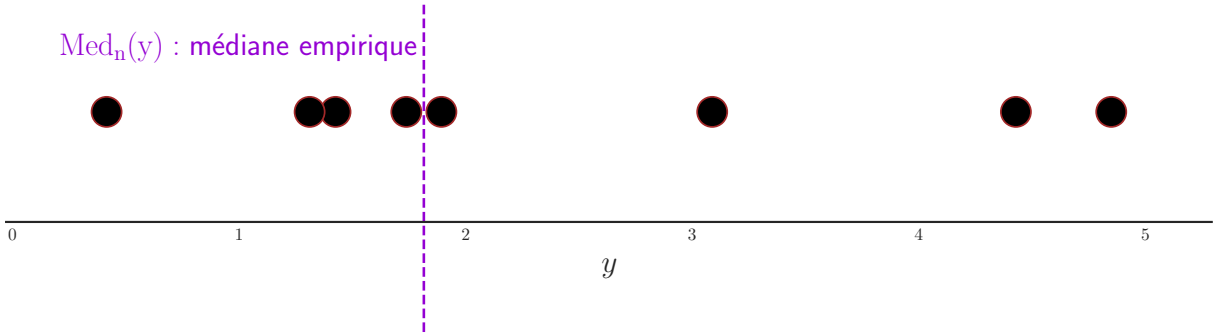
Définition

Moyenne (arithmétique) : $\overline{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i$

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

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

Médiane



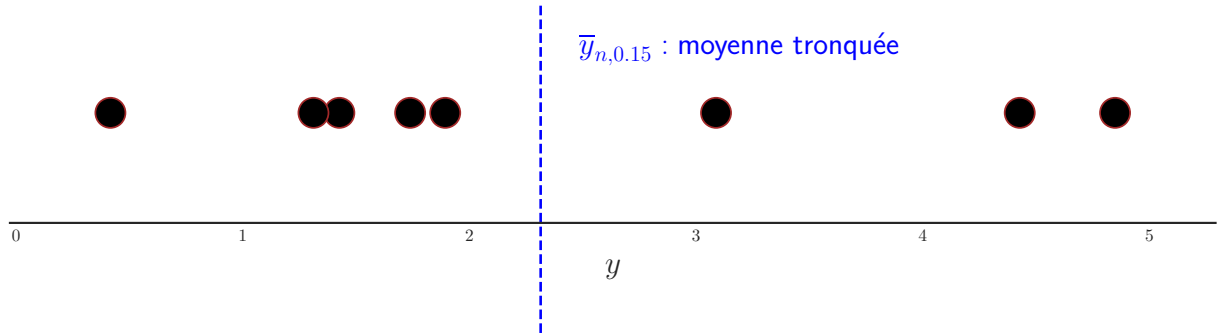
On ordonne les y_i dans l'ordre croissant : $y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(n)}$

Définition

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

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

Moyenne tronquée



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

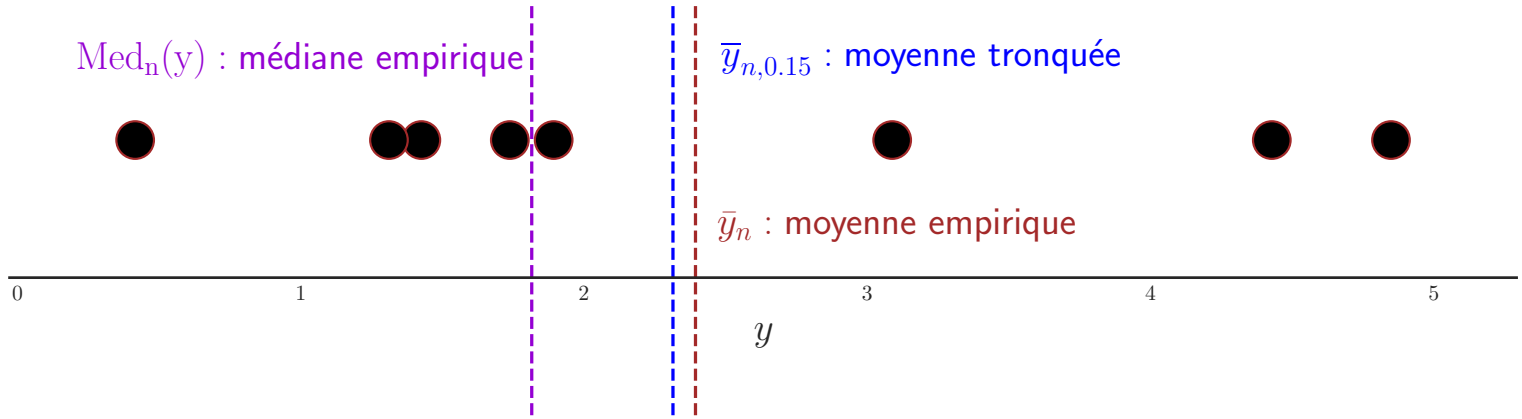
Définition

Moyenne tronquée (à l'ordre α) : $\overline{y}_{n,\alpha} = \overline{z}_n$

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

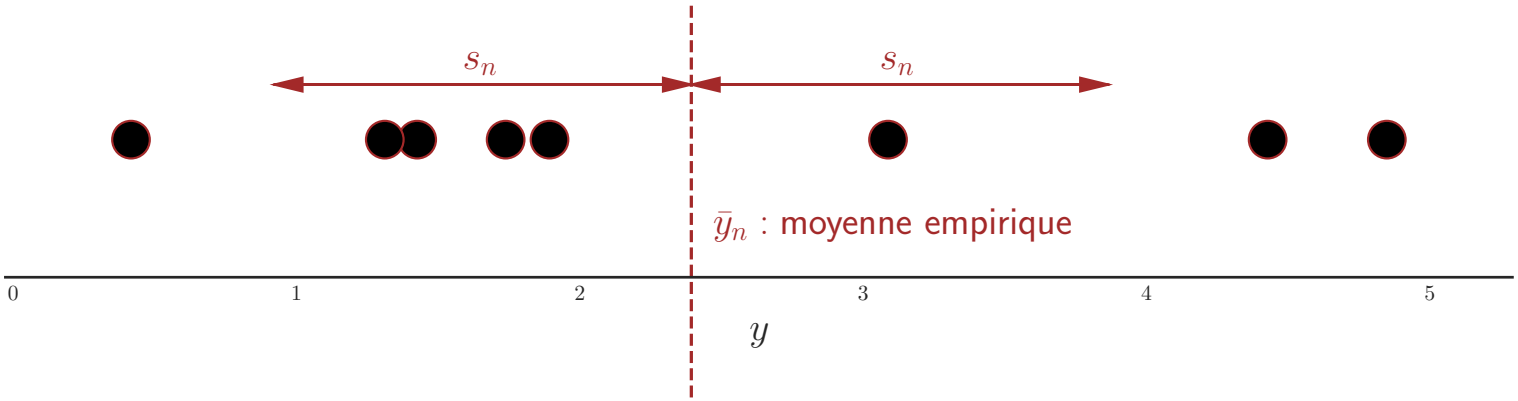
Rem : $\lfloor u \rfloor$ est le nombre entier tel que $\lfloor u \rfloor - 1 < u \leq \lfloor u \rfloor$

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



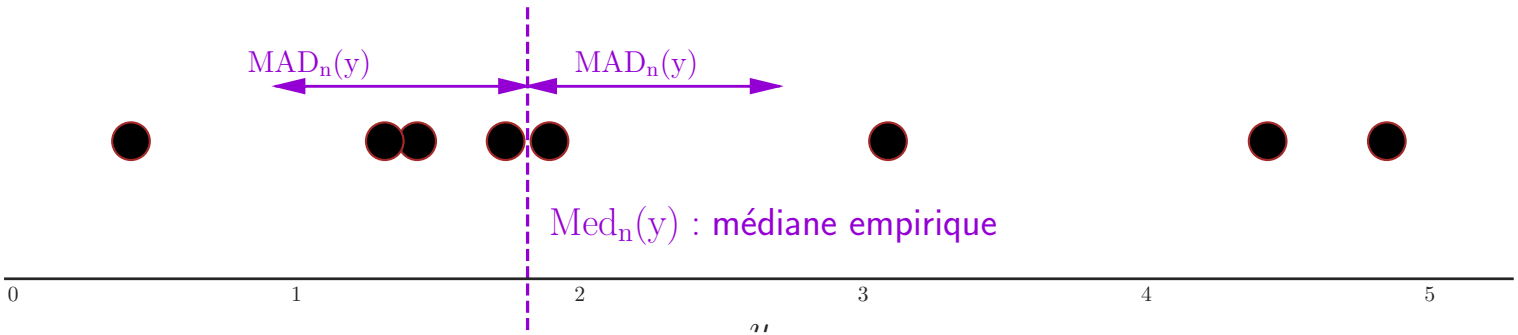
Définitions

Variance : $\mathrm{var}_n(\mathsf{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_n)^2 = \frac{1}{n}\|\mathsf{y} - \bar{y}_n \mathbb{1}_n\|^2$

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

Exo : Quels sont les vecteurs $\mathsf{y} \in \mathbb{R}^n$ tels que $\mathrm{var}_n(\mathsf{y}) = 0$?

Dispersion : MAD

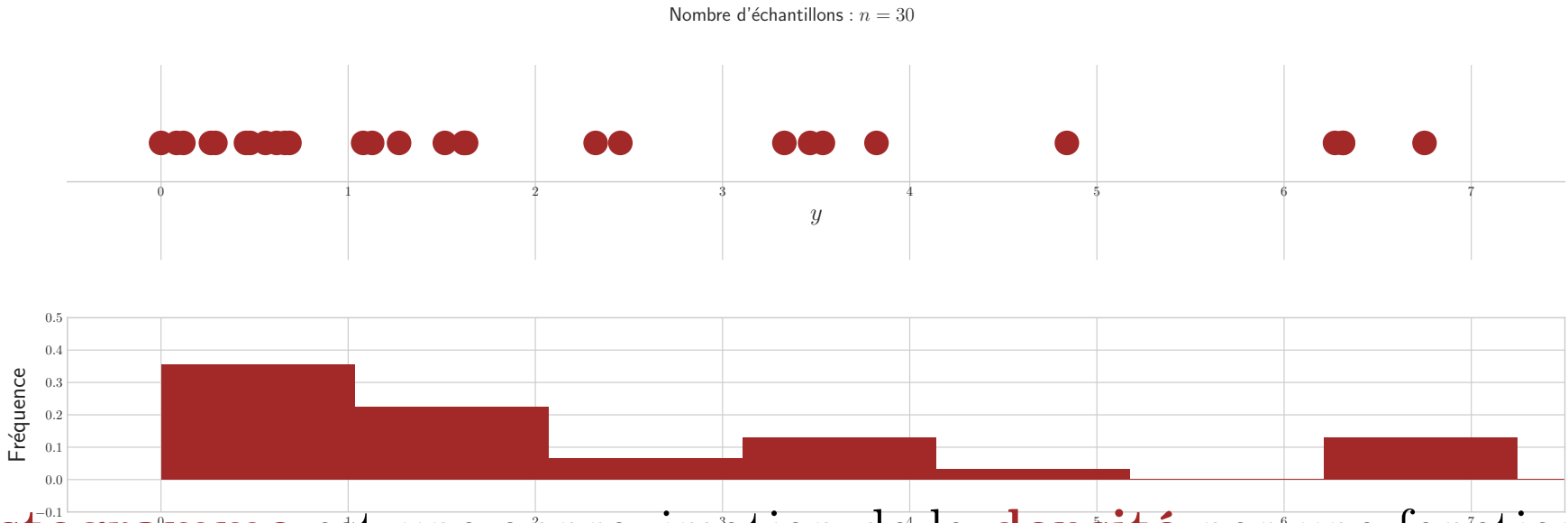


Définition

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

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

Estimation de la densité : histogramme

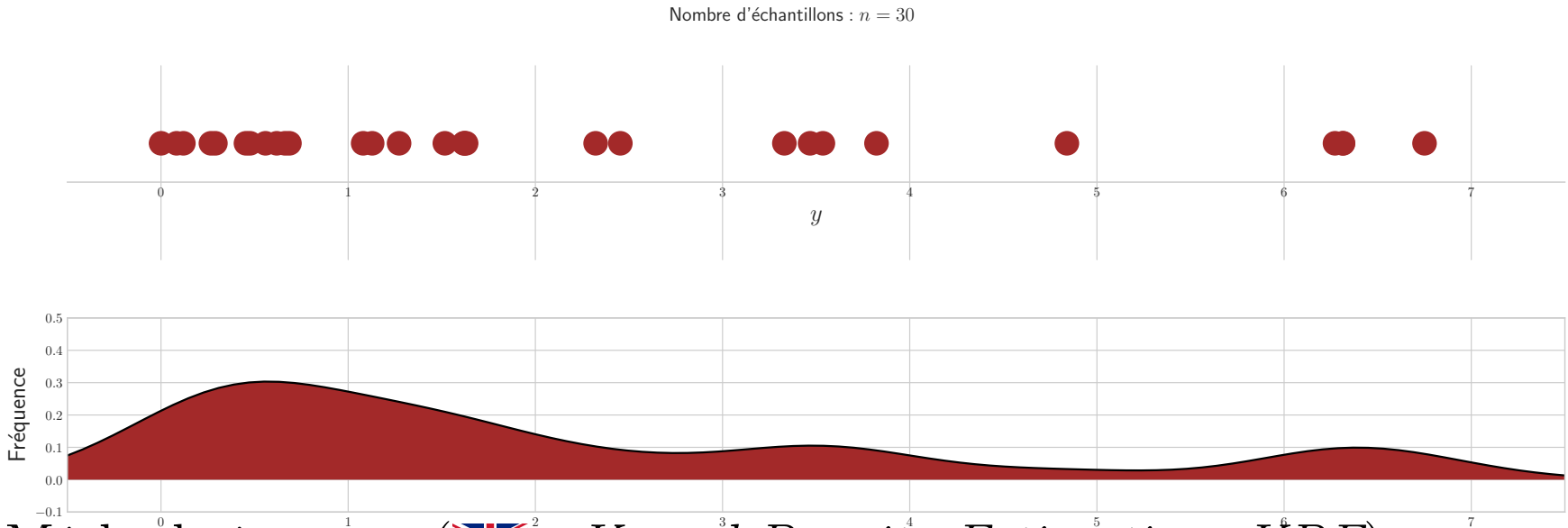


L'**histogramme** est une approximation de la **densité** par une fonction constante par morceaux

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

Rem : en Python, on compte le nombre ou la proportion de données par case, par exemple avec `normed=False(True)` dans la fonction `hist`

Estimation de la densité : méthode à noyau

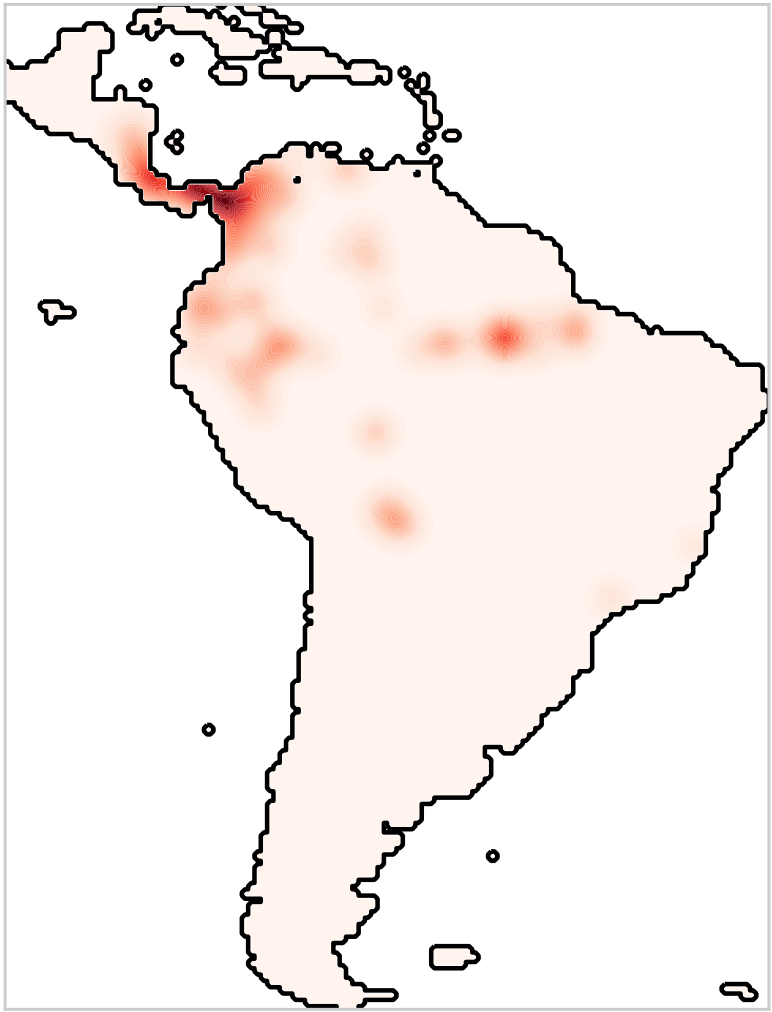


- Méthode à noyau ( : *Kernel Density Estimation, KDE*) : approche non-paramétrique estimant la densité par une fonction continue – généralisation de l'histogramme

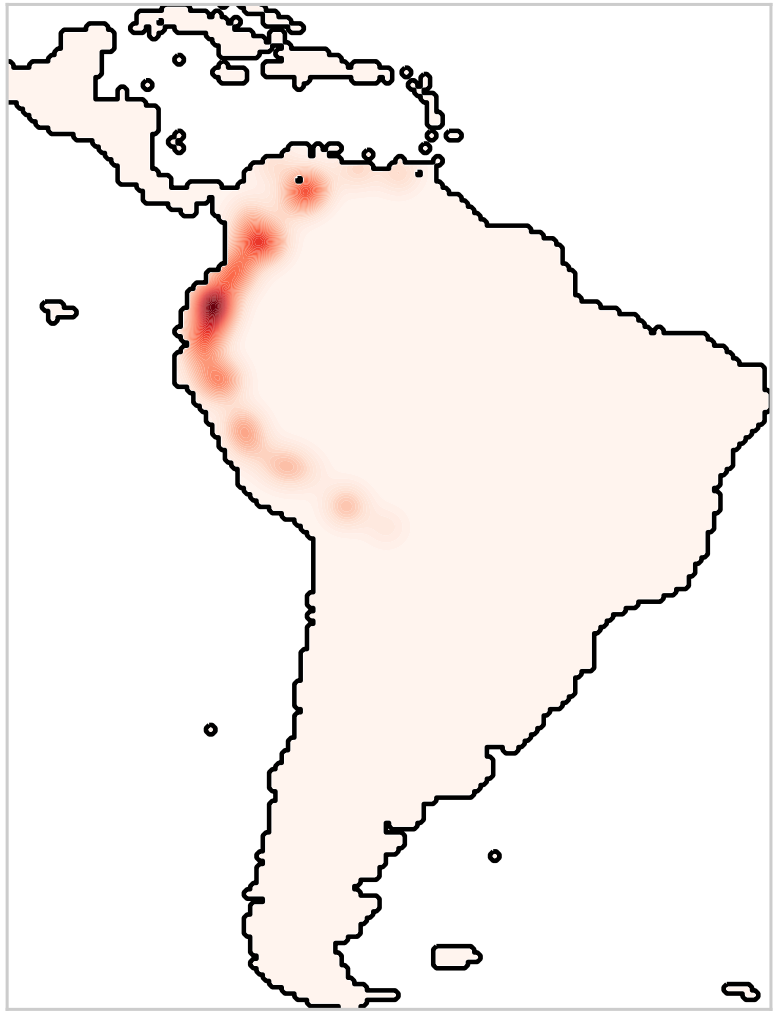
Pour plus de détails voir le livre Silverman (1986)

Densité bi-dimensionnelle (spatiale)

Bradypus Variegatus

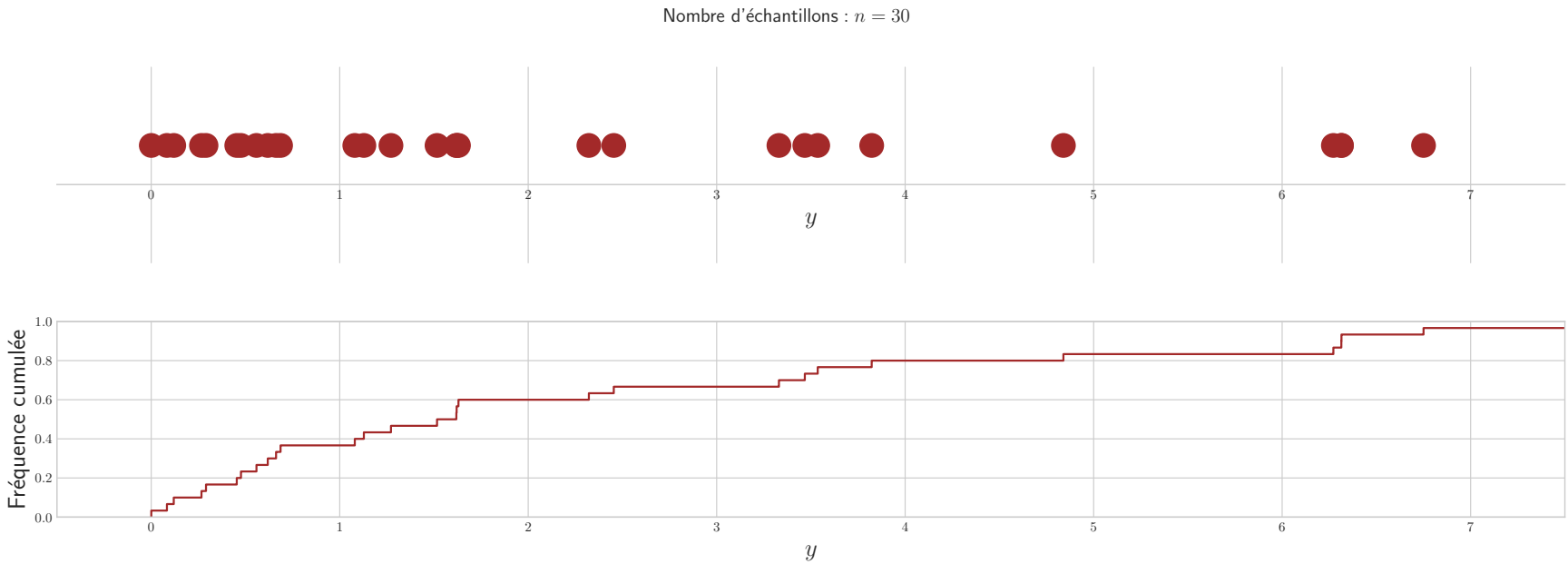


Microryzomys Minutus



`http://scikit-learn.org/stable/_downloads/plot_species_kde.py` c.f.

Fonction de répartition



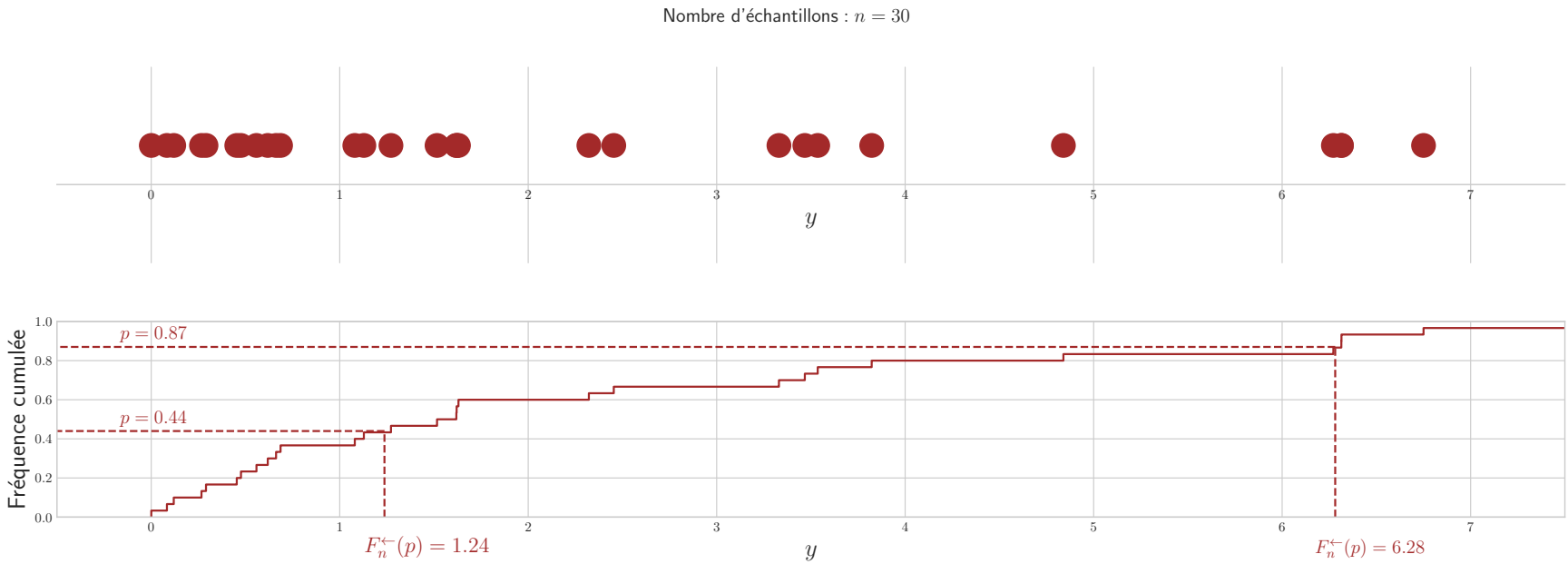
Définition : fonction de répartition

Théorique : $$F(u) = \mathbb{P}(Y \leq u) = \int_{-\infty}^{u} f_Y(x)dx$$

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

Interprétation : proportion d'observations sous un certain niveau

Fonction quantile



Définition

Pour $p \in]0, 1]$,

Quantile théorique (d'ordre p) : $F^{\leftarrow}(p) = \inf\{u \in \mathbb{R} : F(u) \geq p\}$

Quantile empirique (d'ordre p) : $F_n^{\leftarrow}(p) = y_{(\lfloor (n-1)p \rfloor + 1)}$

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

Covariances et corrélations empiriques

Covariance empirique

Pour deux échantillons x et y de moyennes et variances empiriques $\overline{x}_n$, $\overline{y}_n$ et $\mathrm{var}_n(\mathsf{x})$, $\mathrm{var}_n(\mathsf{y})$:

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

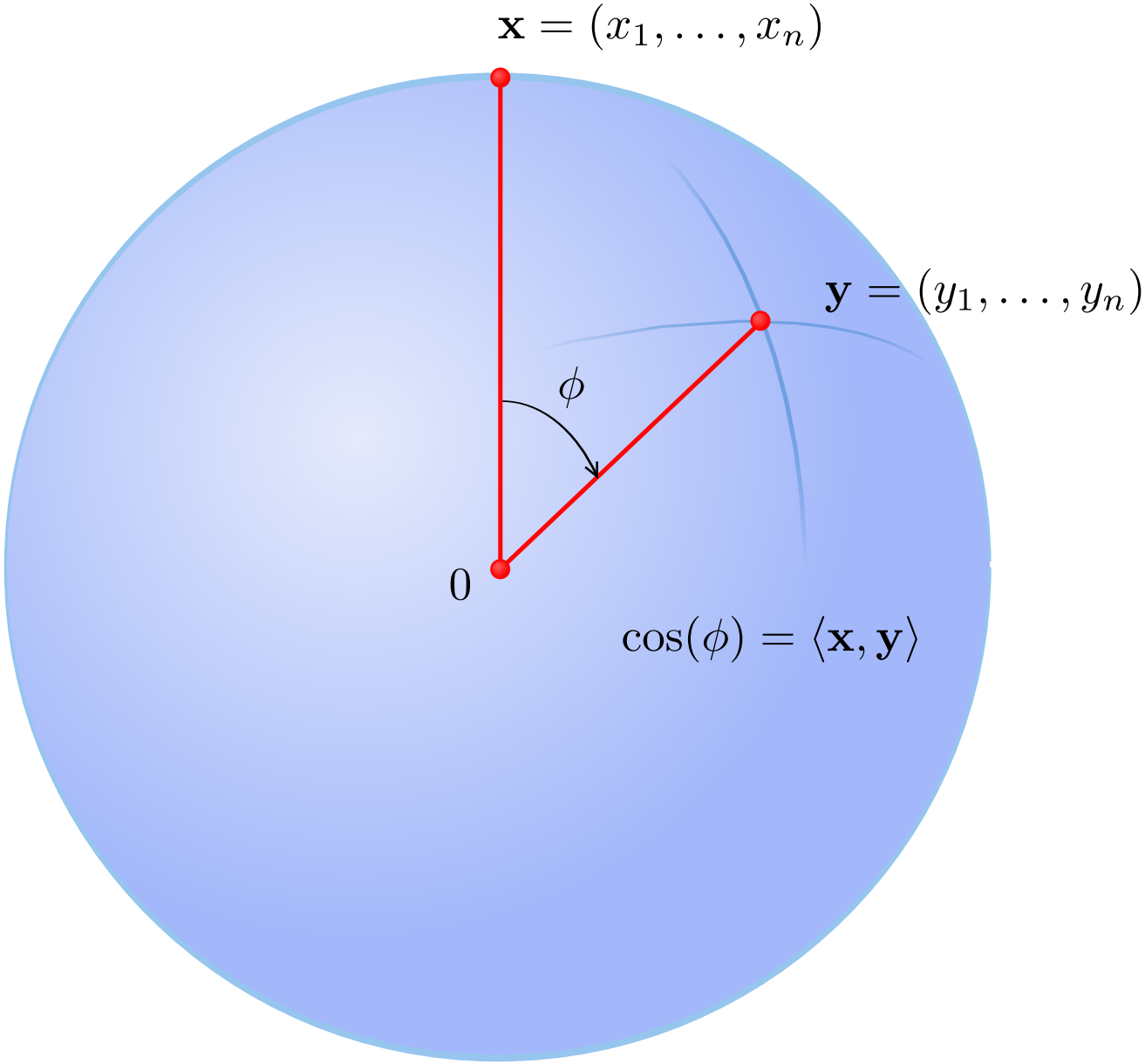
$$\mathrm{cov}_n(\mathsf{x},\mathsf{y}) = \frac{1}{n}\langle \mathsf{x} - \overline{x}_n \mathbb{1}_n, \mathsf{y} - \overline{y}_n \mathbb{1}_n \rangle$$

Corrélation empirique

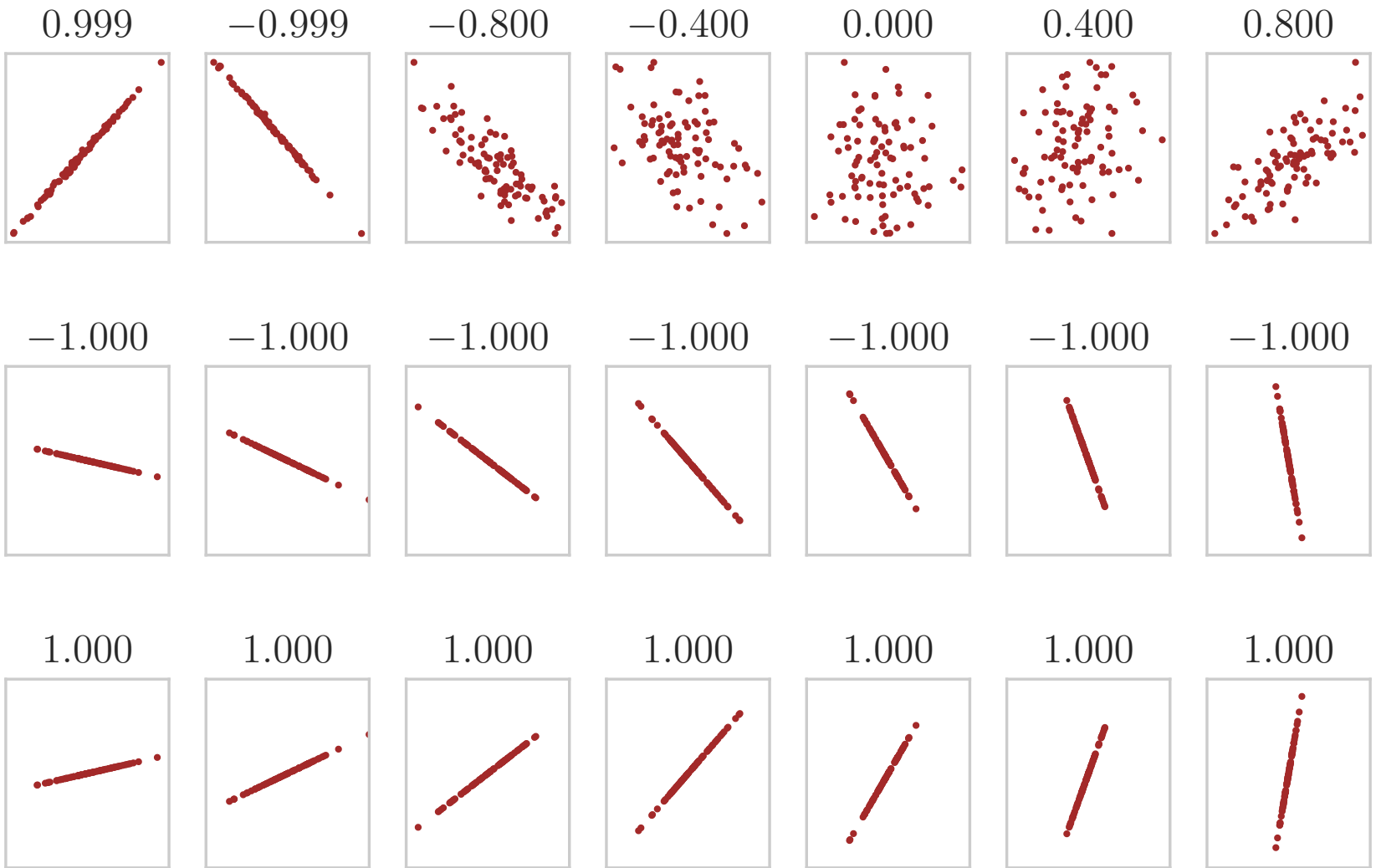
$$\rho = \mathsf{corr}_n(\mathsf{x},\mathsf{y}) = \frac{\mathrm{cov}_n(\mathsf{x},\mathsf{y})}{\sqrt{\mathrm{var}_n(\mathsf{x})}\sqrt{\mathrm{var}_n(\mathsf{y})}}, \quad \text{c'est-à-dire}$$

$$\rho = \frac{\langle \mathsf{x} - \overline{x}_n \mathbb{1}_n, \mathsf{y} - \overline{y}_n \mathbb{1}_n \rangle}{\|\mathsf{x} - \overline{x}_n \mathbb{1}_n\| \, \|\mathsf{y} - \overline{y}_n \mathbb{1}_n\|} = \cos(\mathsf{x} - \overline{x}_n \mathbb{1}_n, \mathsf{y} - \overline{y}_n \mathbb{1}_n)$$

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

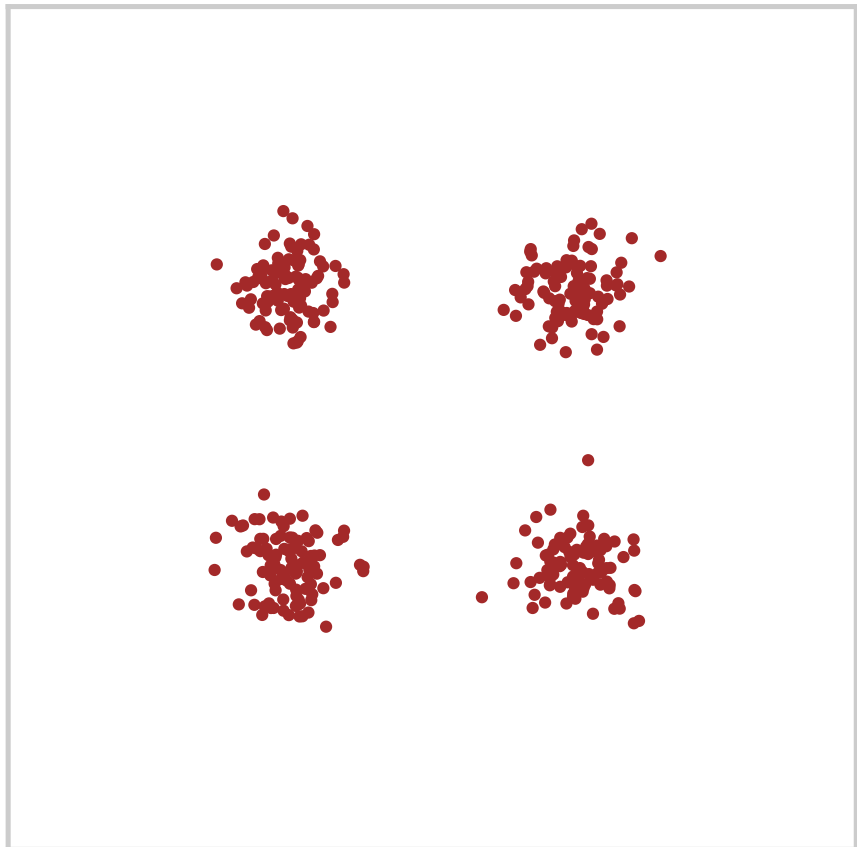


Exemples de corrélations



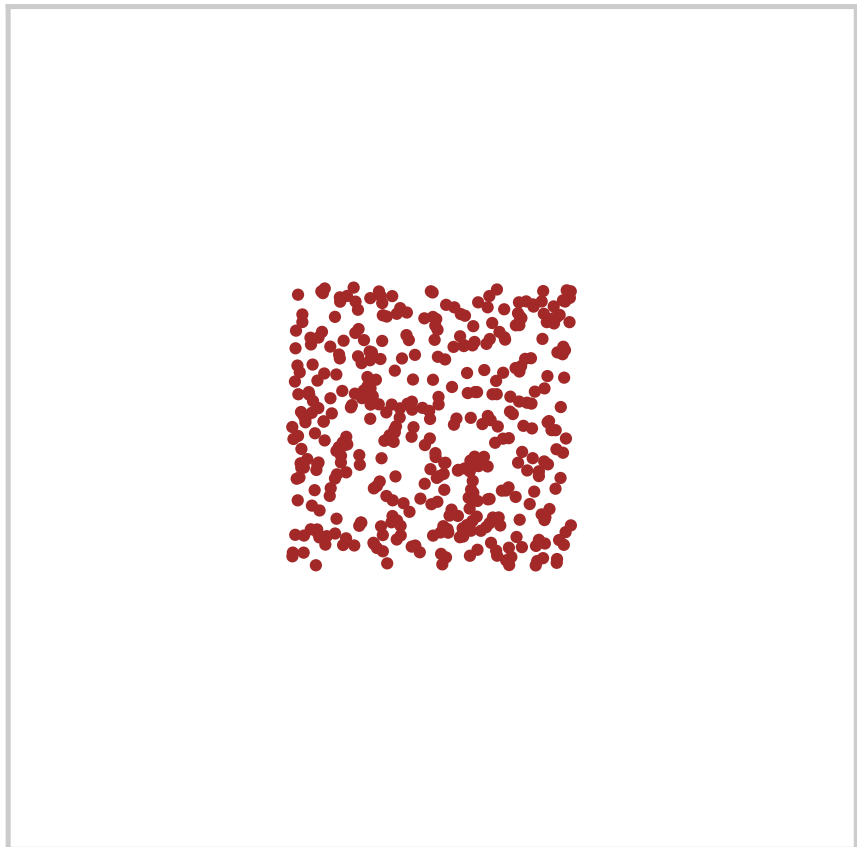
Exemples de corrélations proches de zéro

Corrélation $= -0.021$



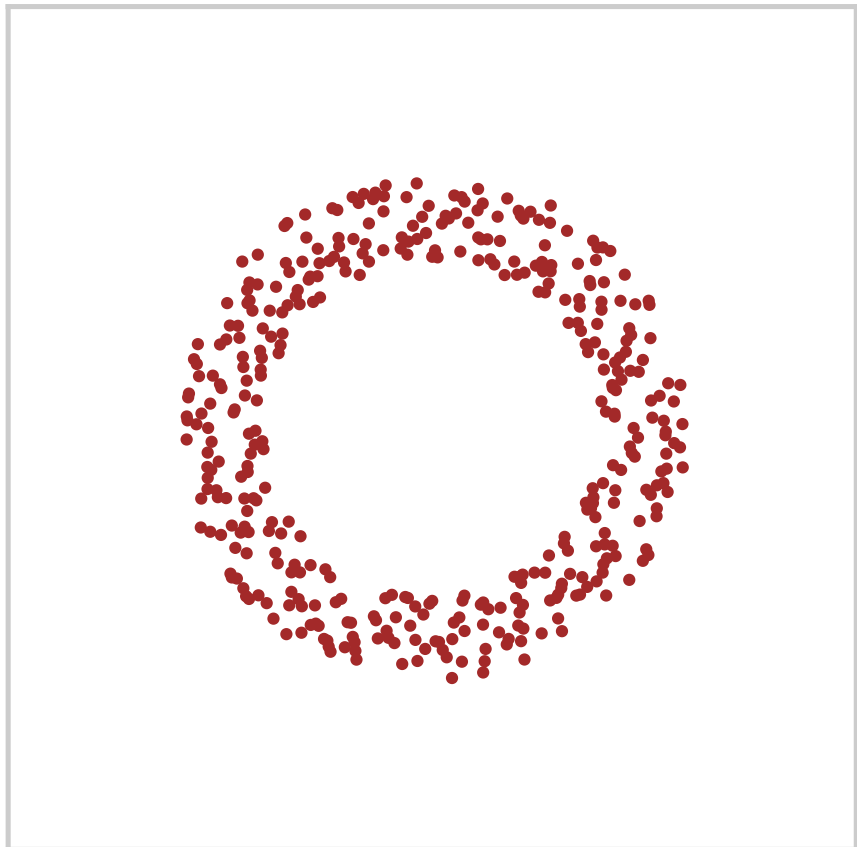
Exemples de corrélations proches de zéro

Corrélation $= 0.007$

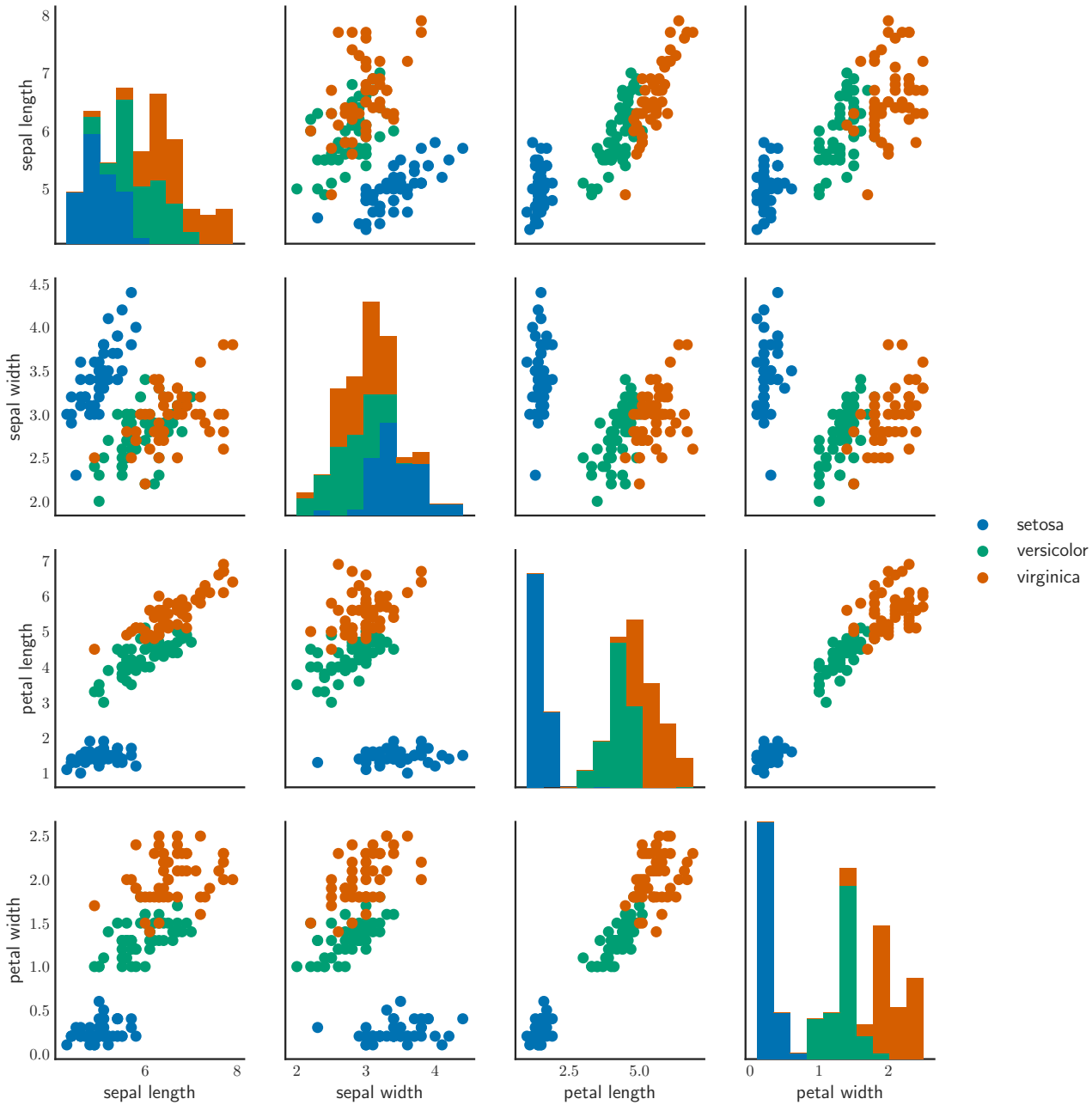


Exemples de corrélations proches de zéro

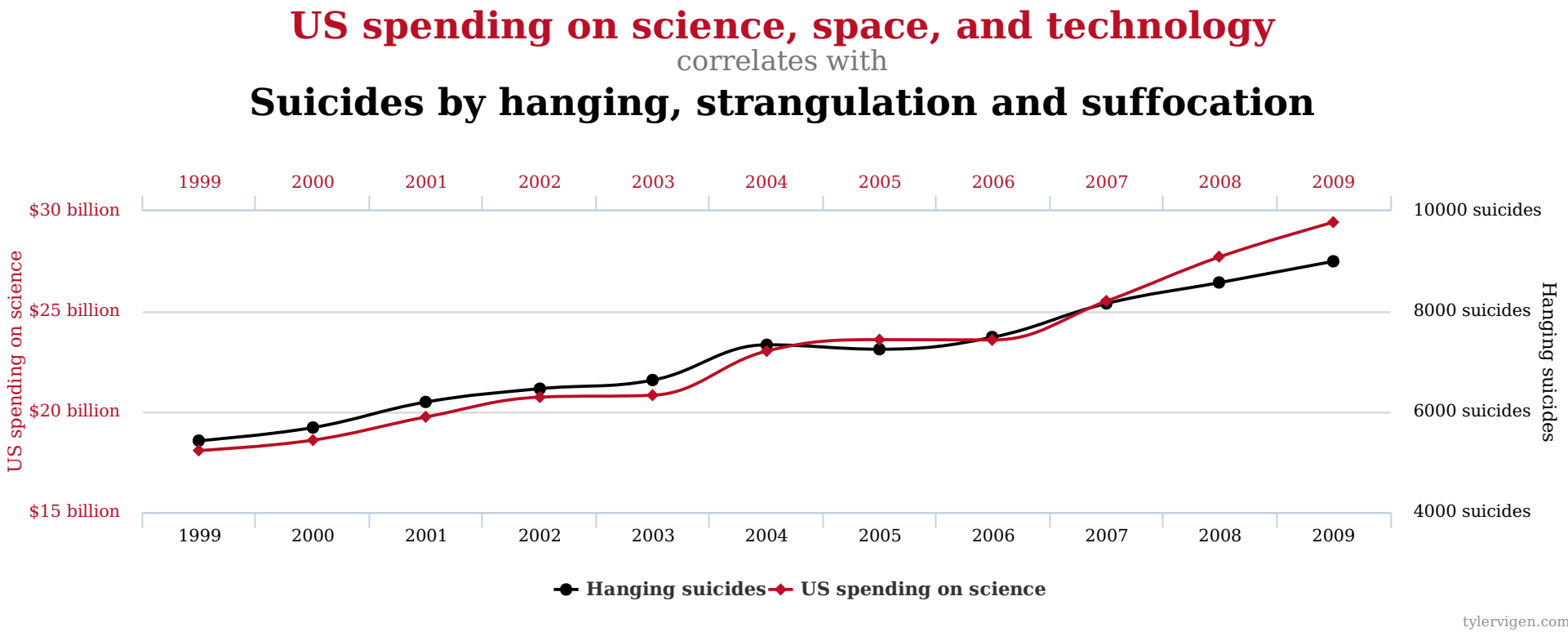
Corrélation $= 0.011$



Nuages de points / Scatter plot / PairGrid



Covariance ≠ causalité



Corrélation : 0.9979

c.f. `http://www.tylervigen.com/spurious-correlations`

Covariance d'un couple de V.A.

Soient X et Y des variables aléatoires réelles de carré intégrable.

Définition

La **covariance** de X et Y est la moyenne des fluctuations jointes :

$$\mathbb{C}\mathrm{ov}(X, Y) = \mathbb{E}\left[\left(X - \mathbb{E}(X)\right)\left(Y - \mathbb{E}(Y)\right)\right]$$

Propriété : la covariance est bilinéaire, pour tous $\alpha, \beta \in \mathbb{R}$ et toutes variables aléatoires réelles X_1, X_2, Y_1, Y_2 on a

$$\begin{aligned}\mathbb{C}\mathrm{ov}(\alpha X_1 + \beta X_2, Y_1) &= \alpha\mathbb{C}\mathrm{ov}(X_1, Y_1) + \beta\mathbb{C}\mathrm{ov}(X_2, Y_1)\\ \mathbb{C}\mathrm{ov}(X_1, \alpha Y_1 + \beta Y_2) &= \alpha\mathbb{C}\mathrm{ov}(X_1, Y_1) + \beta\mathbb{C}\mathrm{ov}(X_1, Y_2)\end{aligned}$$

Rappel : inégalité de Cauchy–Schwarz dans ce cadre

$$|\mathbb{C}\mathrm{ov}(X, Y)| \leq \sqrt{\mathbb{V}\mathrm{ar}(X)\,\mathbb{V}\mathrm{ar}(Y)}$$

Matrice de covariance d'un vecteur aléatoire

Notation : $X = (X_1, \ldots, X_p)^\top$ est vecteur aléatoire t.q.
$\forall j \in \{1, \ldots, p\}, \mathbb{E}(X_j^2) < +\infty$ et $\sigma_{i,j} = \mathrm{cov}(X_i, X_j)$ $(\sigma_{i,i} = \mathrm{var}(X_i))$

Définition

La **matrice de covariance** du vecteur X est la matrice $\mathbb{C}\mathrm{ov}(X)$, de taille $p \times p$, formée par les $\sigma_{i,j}$ (i^{e} ligne, j^{e} colonne). Ainsi,

$$\mathbb{C}\mathrm{ov}(X) = \begin{pmatrix} \mathrm{var}(X_1) & \mathrm{cov}(X_1,X_2) & \ldots & \mathrm{cov}(X_1,X_p) \\ \mathrm{cov}(X_2,X_1) & \mathrm{var}(X_2) & & \vdots \\ \vdots & & \ddots & \vdots \\ \mathrm{cov}(X_p,X_1) & \ldots & & \mathrm{var}(X_p) \end{pmatrix} \in \mathbb{R}^{p \times p}$$

Version condensée : $\mathbb{C}\mathrm{ov}(X) = \mathbb{E}\Big[\big(X - \mathbb{E}(X)\big)\big(X - \mathbb{E}(X)\big)^\top\Big]$

Exo : Montrer que pour μ déterministe $\mathbb{C}\mathrm{ov}(X + \mu) = \mathbb{C}\mathrm{ov}(X)$

Quelques propriétés de la covariance

- Une matrice de covariance est symétrique :

$$\mathbb{C}\mathrm{ov}(X) = \mathbb{C}\mathrm{ov}(X)^\top \Leftrightarrow \forall (i,j) \in \{1,\dots,p\}^2,\ \mathbb{C}\mathrm{ov}(X_i, X_j) = \mathbb{C}\mathrm{ov}(X_j, X_i)$$

- Une matrice de covariance est (semi-définie) positive :

$$\forall u \in \mathbb{R}^p,\ u^\top \mathbb{C}\mathrm{ov}(X) u \geq 0$$

Démonstration :

$$u^\top \mathbb{C}\mathrm{ov}(X) u = \sum_{i=1}^{p}\sum_{j=1}^{p} u_i u_j \mathbb{C}\mathrm{ov}(X_i, X_j) = \underbrace{\mathbb{C}\mathrm{ov}\left(\sum_{i=1}^{p} u_i X_i, \sum_{j=1}^{p} u_j X_j\right)}_{=\mathbb{V}\mathrm{ar}\left(\sum_{j=1}^{p} u_j X_j\right) \geq 0}$$

Exo : $\mathbb{C}\mathrm{ov}(AX) = A\mathbb{C}\mathrm{ov}(X)A^\top$, pour toute matrice $A \in \mathbb{R}^{m \times p}$

La décomposition spectrale

Théorème spectral

Une matrice symétrique $S \in \mathbb{R}^{n\times n}$ est diagonalisable en base orthonormée, *i.e.* il existe $\lambda_1 \geq \cdots \geq \lambda_n$ et une matrice orthogonale $U \in \mathbb{R}^{n\times n}$ telle que :

$$S = U\mathrm{diag}(\lambda_1, \ldots, \lambda_n)U^\top \text{ ou } SU = U\mathrm{diag}(\lambda_1, \ldots, \lambda_n)$$

Rappel : une matrice orthogonale $U \in \mathbb{R}^n$ est une matrice telle que $U^\top U = UU^\top = \mathrm{Id}_n$ ou $\forall (i,j) \in \{1, \ldots, n\}, \mathrm{u}_i^\top \mathrm{u}_j = \langle \mathrm{u}_i, \mathrm{u}_j \rangle = \delta_{i,j}$

Rem : si l'on écrit $U = [\mathrm{u}_1, \ldots, \mathrm{u}_n]$ cela signifie que :

$$S = \sum_{i=1}^{n} \lambda_i \mathrm{u}_i \mathrm{u}_i^\top \quad \text{et} \quad \forall i \in \{1, \ldots, n\},\ S\mathrm{u}_i = \lambda_i \mathrm{u}_i$$

Vocabulaire :

- les λ_i sont les **valeurs propres** de S (🇬🇧 : *eigenvalues*)
- les u_i sont les **vecteurs propres** de S (🇬🇧 : *eigenvectors*)

La décomposition spectrale : exemple

$$A = \begin{pmatrix} 1 & 2 & 0 & 2 \\ 2 & 1 & 2 & 0 \\ 0 & 2 & 1 & 2 \\ 2 & 0 & 2 & 1 \end{pmatrix} = UDU^\top$$

avec

$$D = \begin{pmatrix} 5 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -3 \end{pmatrix} \quad \text{et} \quad U = \begin{pmatrix} \frac{1}{2} & 0 & -\frac{\sqrt{2}}{2} & -\frac{1}{2} \\ \frac{1}{2} & -\frac{\sqrt{2}}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{\sqrt{2}}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{\sqrt{2}}{2} & 0 & \frac{1}{2} \end{pmatrix}$$

La décomposition spectrale : numérique

```
import numpy as np
from scipy.linalg import toeplitz
from numpy.linalg import eigh

A = toeplitz([1, 2, 0, 2])
[Dint, Uint] = eigh(A)
# use eigh not eig for symmetric matrices

idx = Dint.argsort()[::-1]
D = Dint[idx]
U = Uint[:, idx]

print(np.allclose(U.dot(np.diag(D)).dot(U.T), A))
```

Loi normale unidimensionnelle

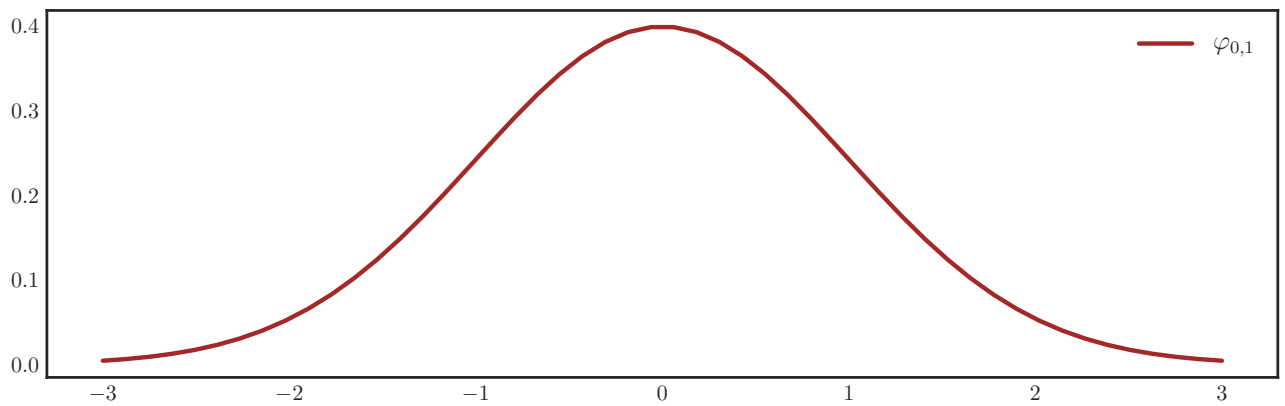
- Une v.a. réelle X suit une « **loi normale** standard » (ou « **loi gaussienne** » ou « loi de Laplace-Gauss ») si sa densité vaut

$$\varphi_{0,1}(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

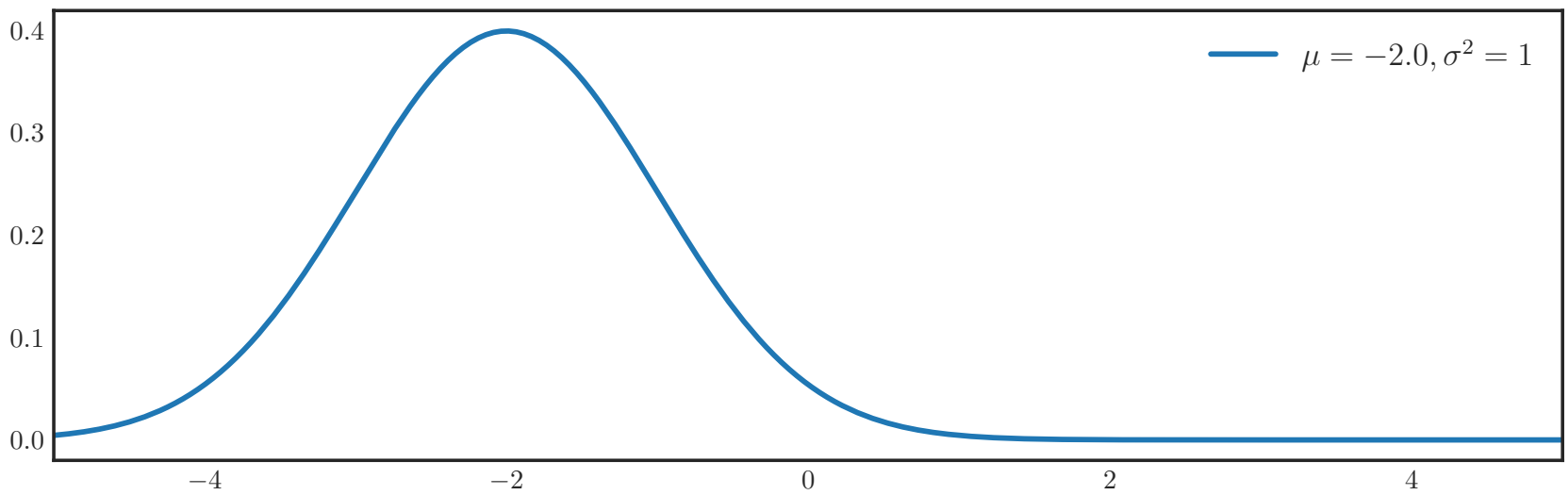
  On note alors $X \sim \mathcal{N}(0,1)$.
- Une v.a. Y suit une loi normale de paramètres μ et σ^2 si

$$Y = \mu + \sqrt{\sigma^2} X\,, \text{ où } X \sim \mathcal{N}(0,1), \text{ et on note } Y \sim \mathcal{N}(\mu, \sigma^2)$$

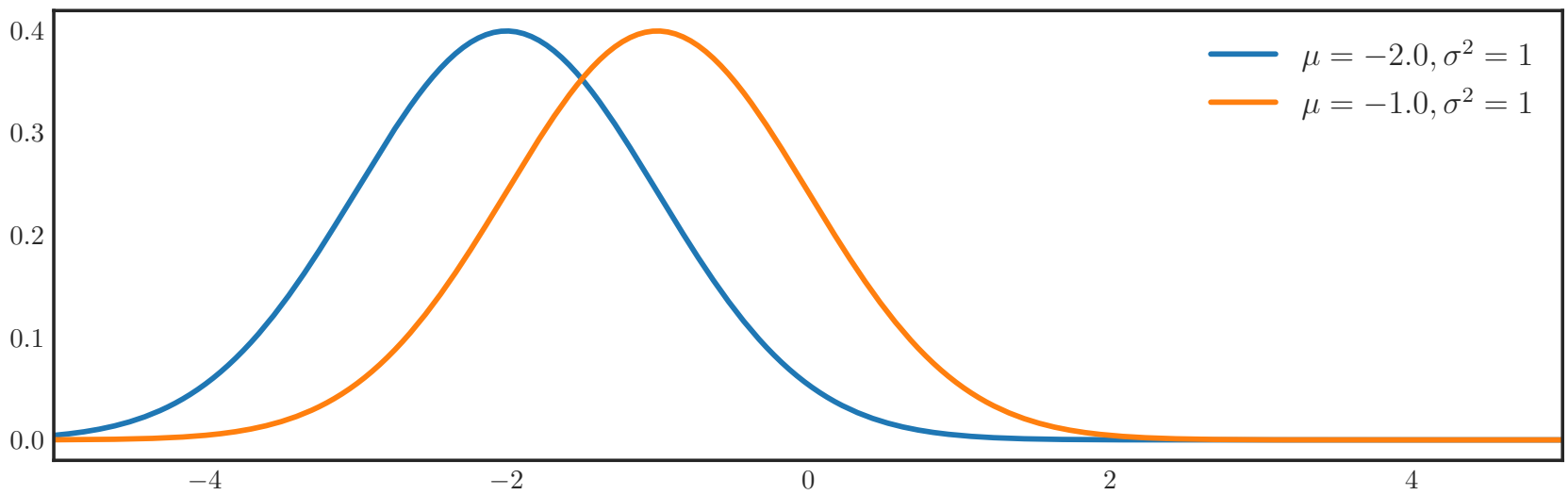
Densité : $\boxed{\varphi_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}}$



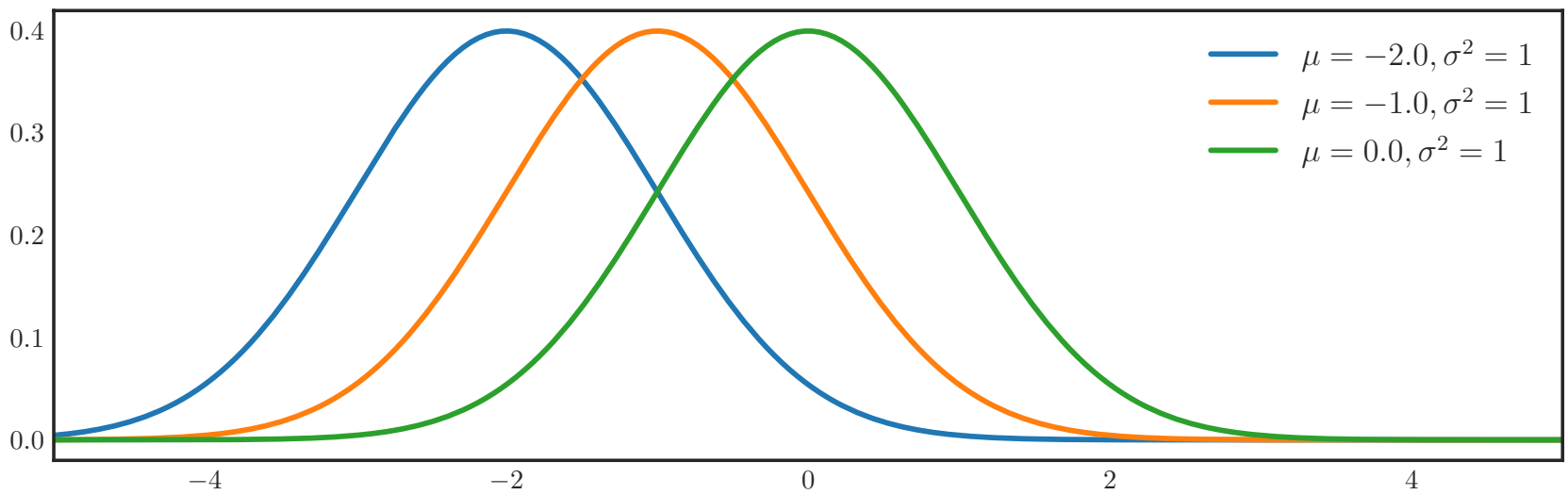
Exemple : variation sur μ



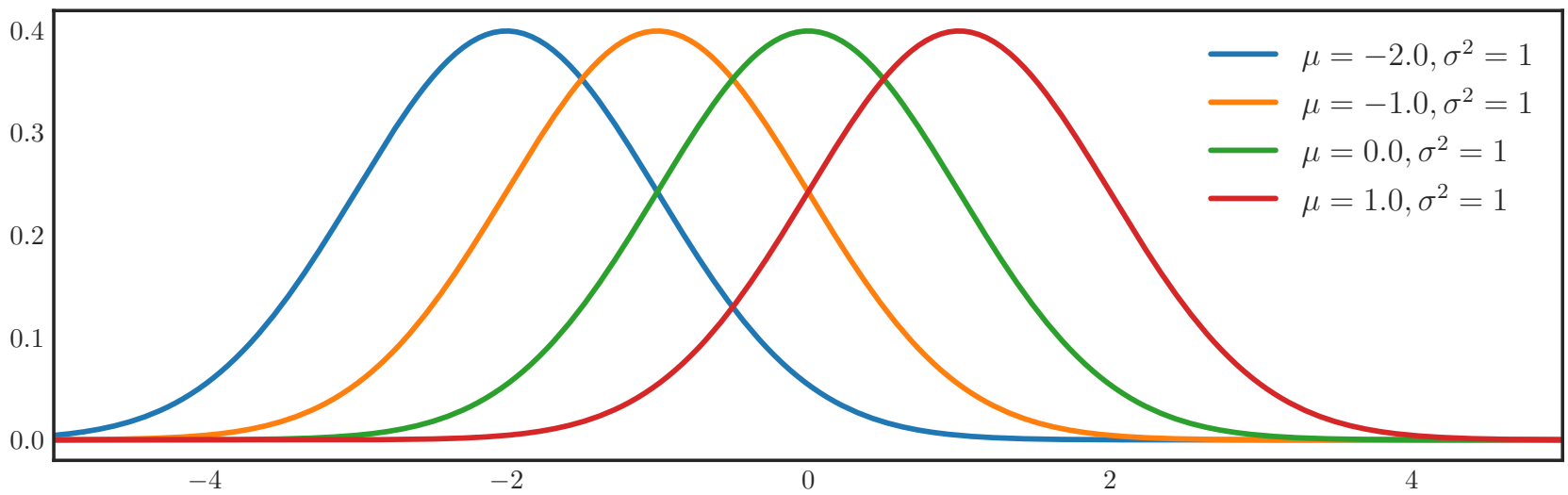
Exemple : variation sur μ



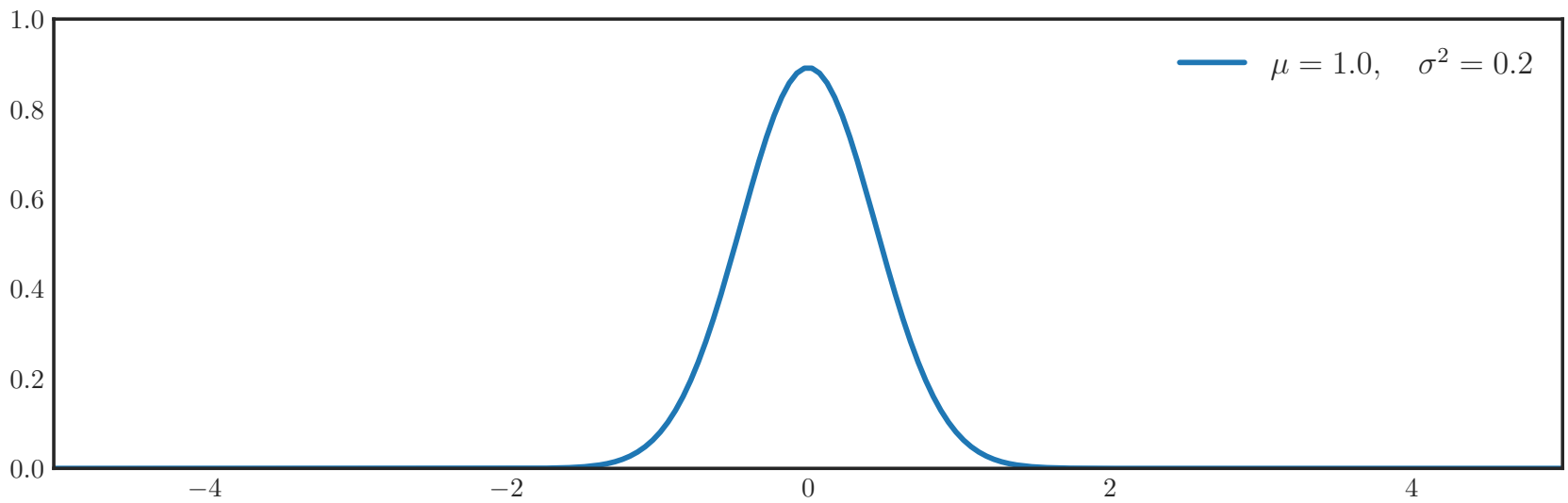
Exemple : variation sur μ



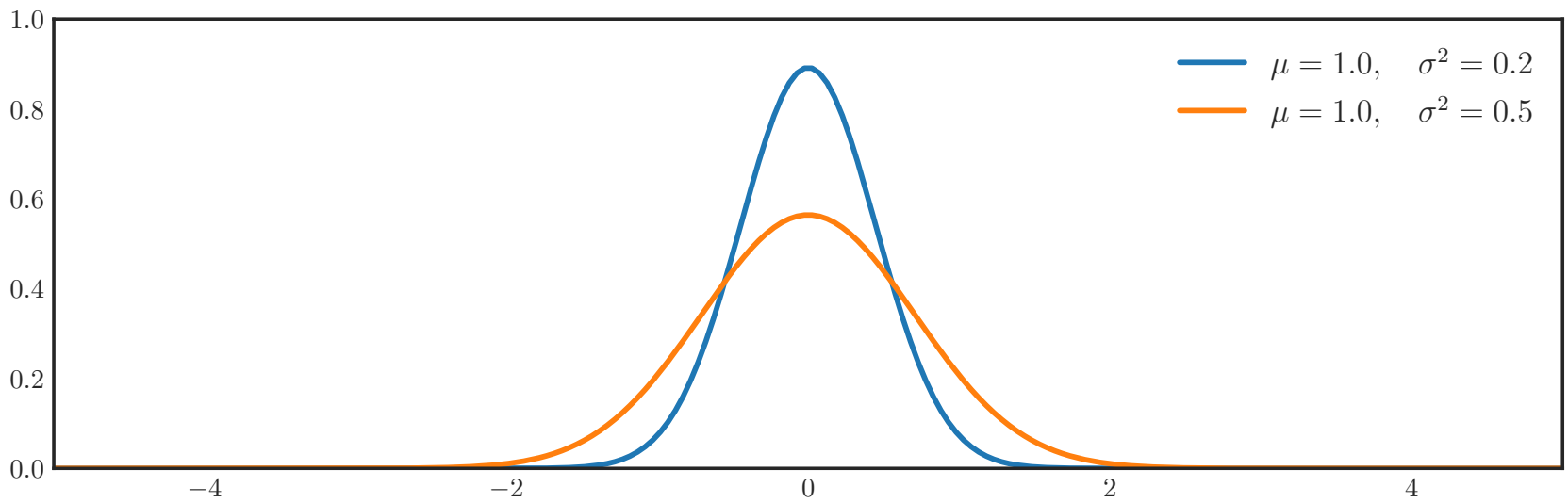
Exemple : variation sur μ



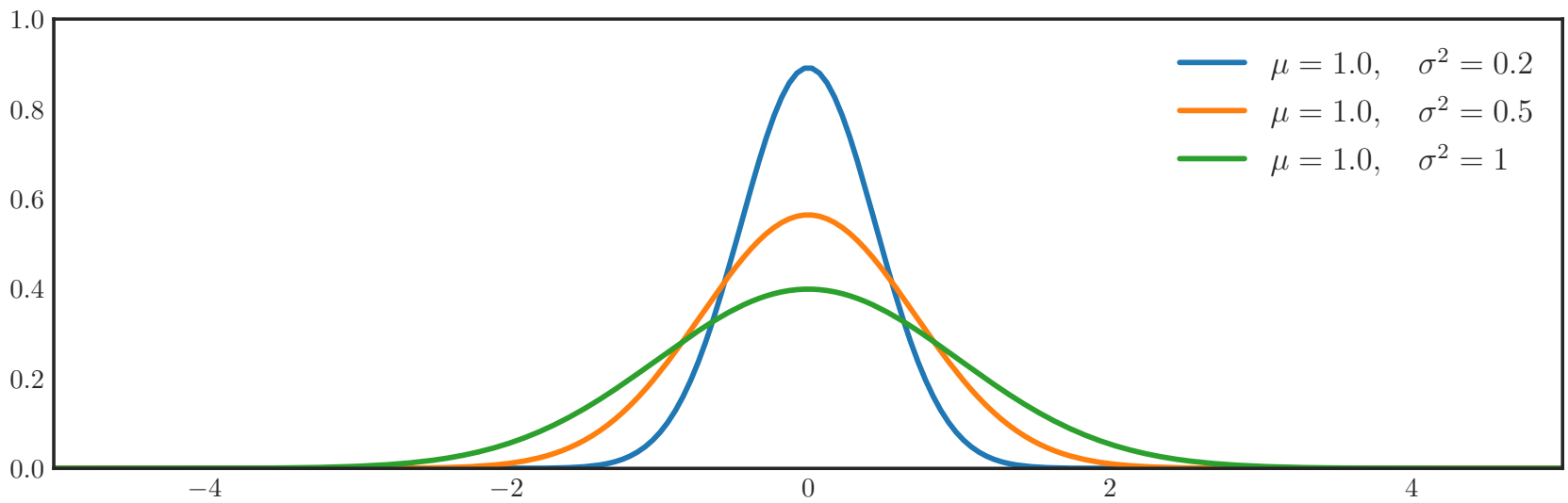
Exemple : variation sur σ



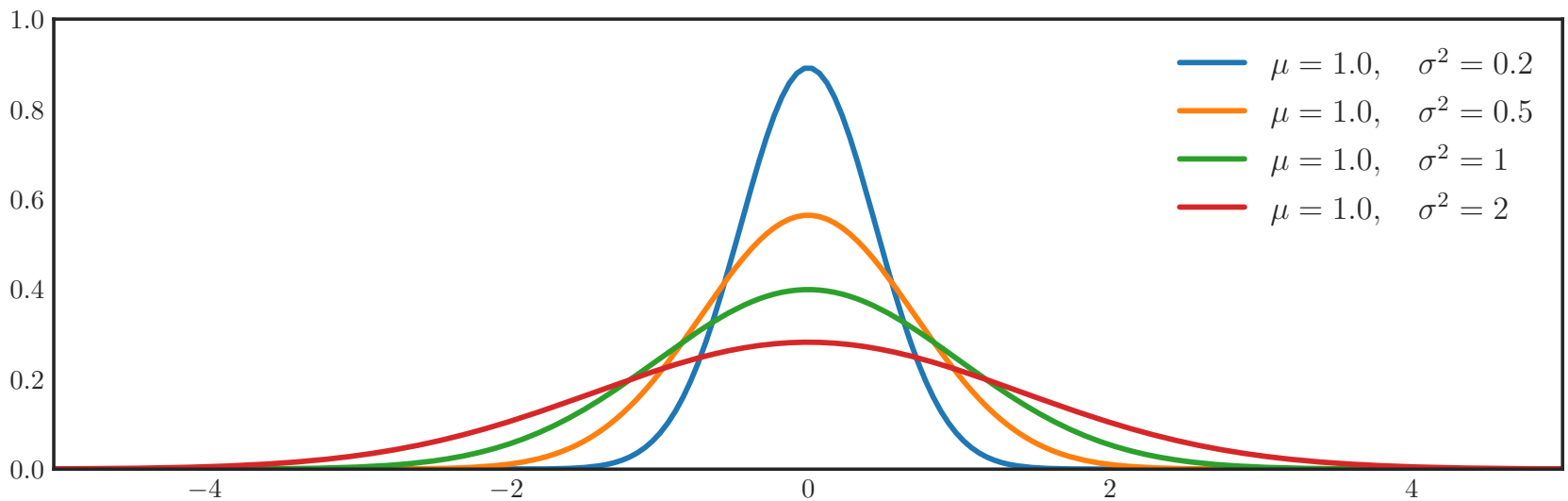
Exemple : variation sur σ



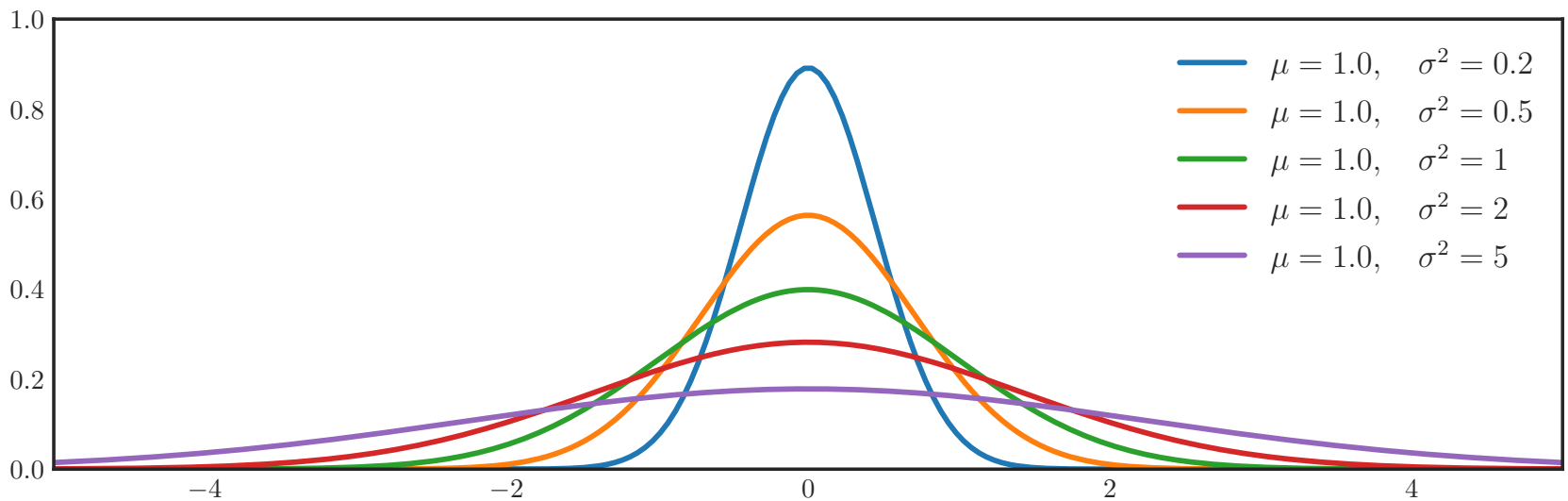
Exemple : variation sur σ



Exemple : variation sur σ



Exemple : variation sur σ



Vecteurs gaussiens

En dimension p, les lois gaussiennes ont des densités de la forme :

$$\boxed{\varphi_{\mu,\Sigma}(x) = \frac{1}{(2\pi)^{p/2}\sqrt{|\Sigma|}} \exp\left\{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right\}}.$$

La fonction $\varphi_{\mu,\Sigma}$ est gouvernée par deux paramètres :

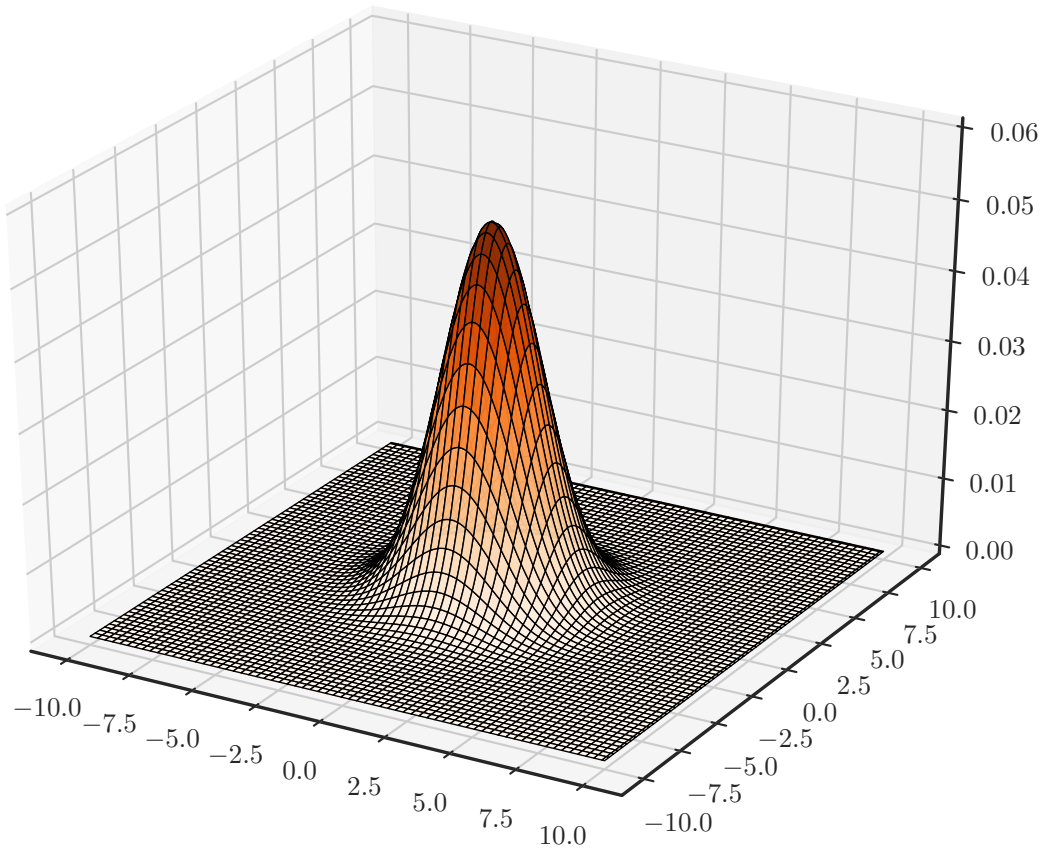
- le vecteur d'espérance $\mu \in \mathbb{R}^p$
- la matrice de covariance $\Sigma \in \mathbb{R}^{p\times p}$

Notation : lorsque le vecteur aléatoire X suit une loi normale d'espérance μ et de covariance Σ, on note $X \sim \mathcal{N}(\mu, \Sigma)$ qu'on suppose définie positive

Rem : $|\Sigma| = \det(\Sigma)$ est le produit des valeurs propres de Σ. On parle de cas dégénéré quand $\det(\Sigma) = 0$

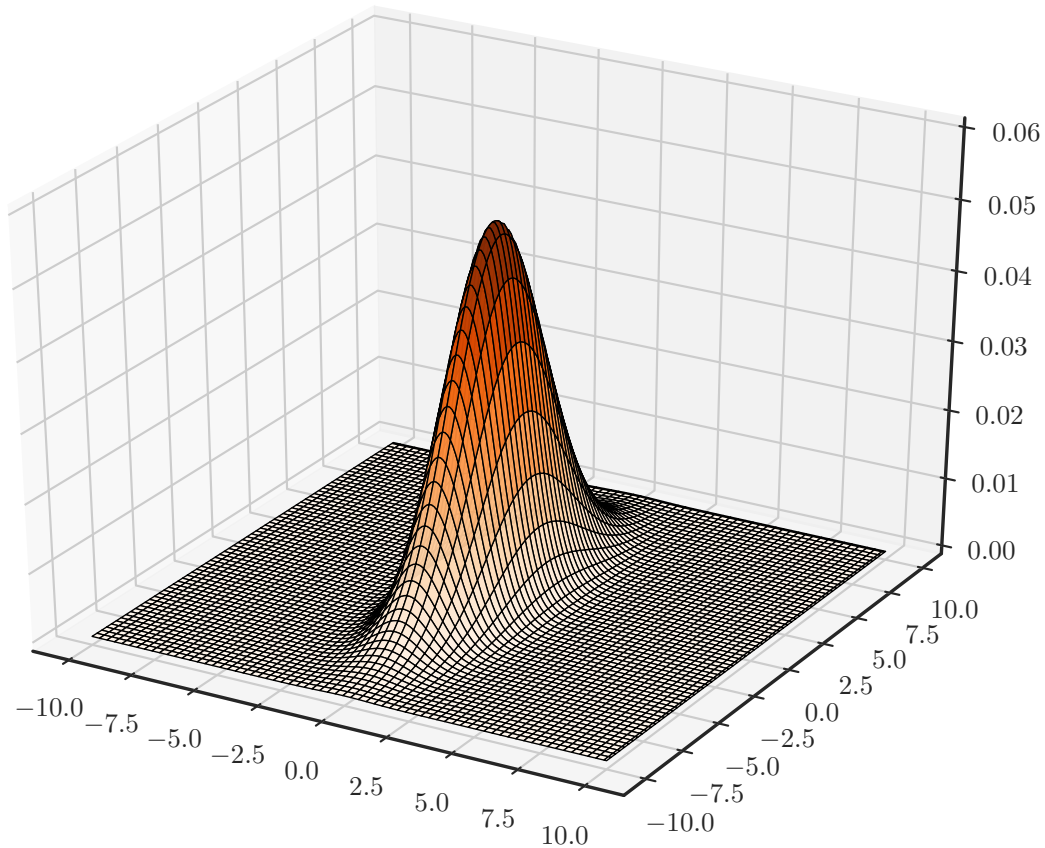
Exemple 2D

$$\Sigma = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} 3 & \\ & 3 \end{bmatrix} \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix},$$



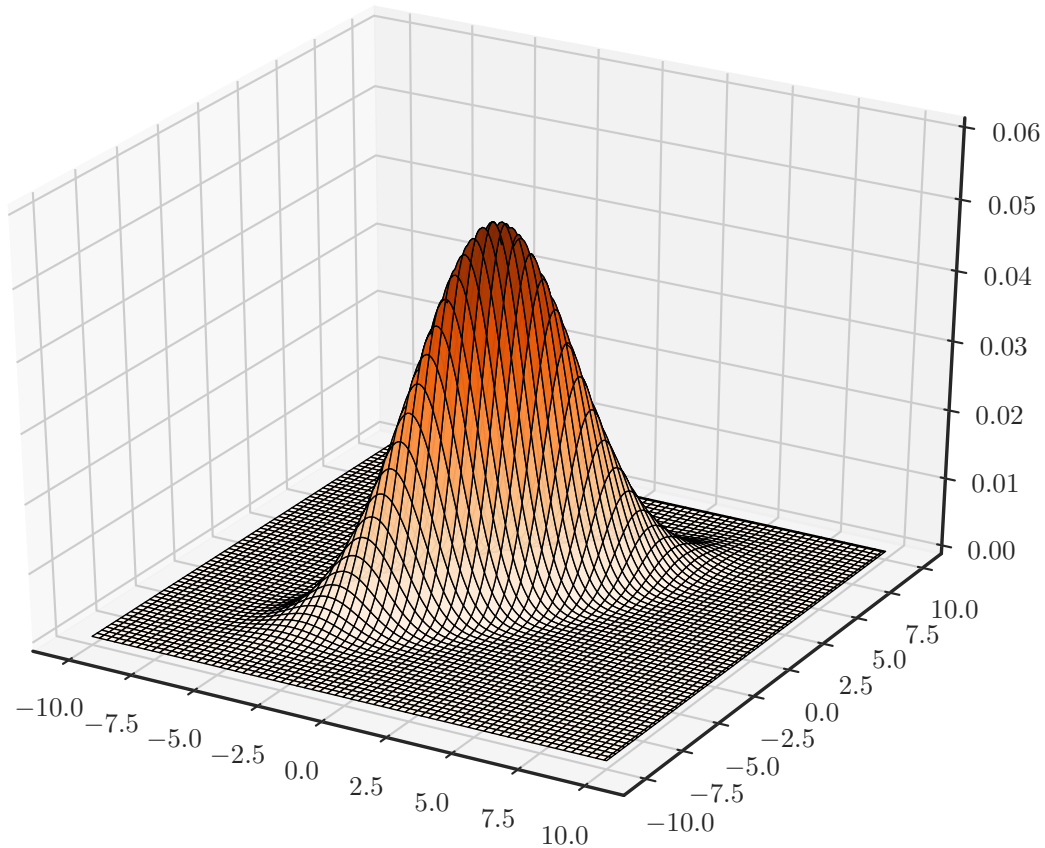
Exemple 2D

$$\Sigma = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} 1 & \\ & 9 \end{bmatrix} \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix}, \theta = \qquad 0$$



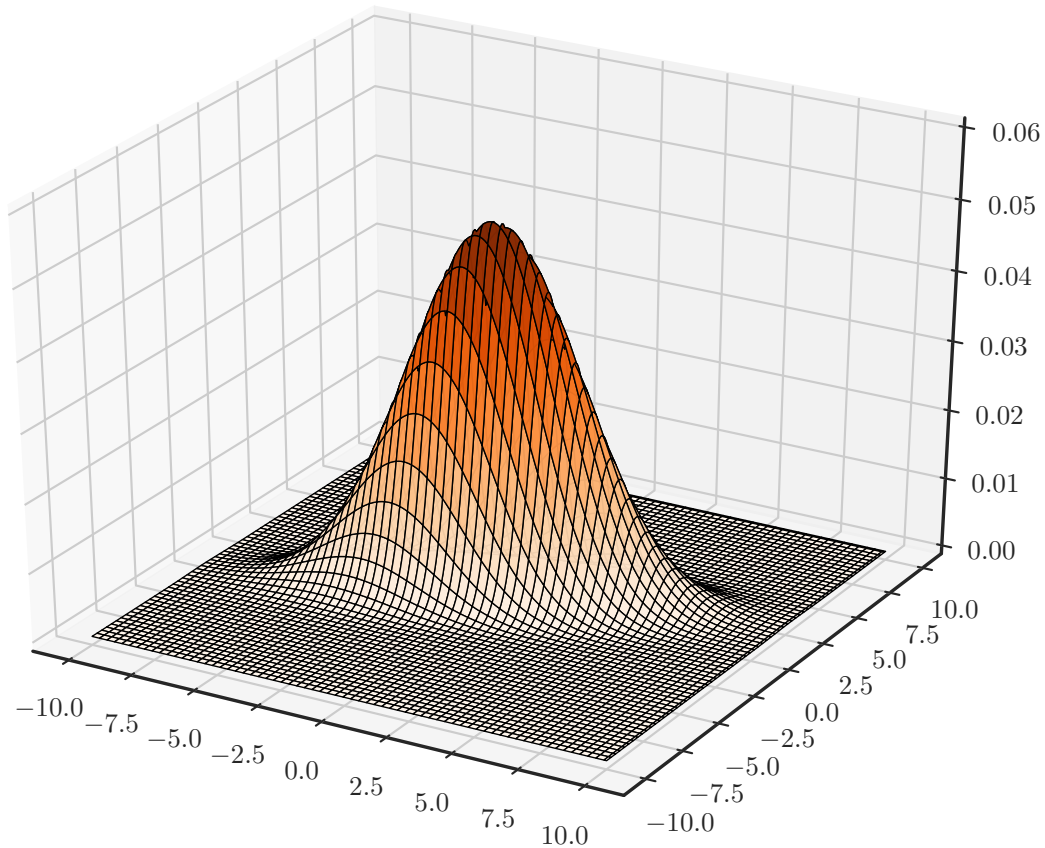
Exemple 2D

$$\Sigma = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} 1 & \\ & 9 \end{bmatrix} \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix}, \theta = 1 \cdot \pi/5$$



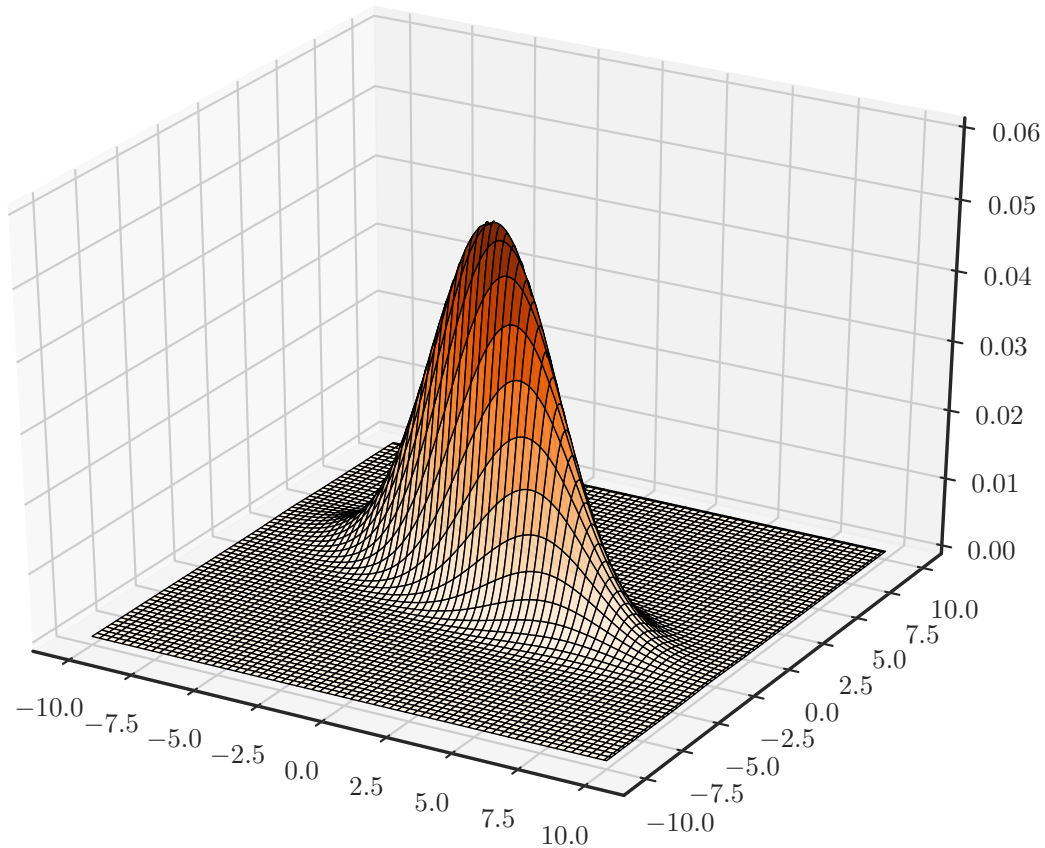
Exemple 2D

$$\Sigma = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} 1 & \\ & 9 \end{bmatrix} \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix}, \theta = 2 \cdot \pi/5$$



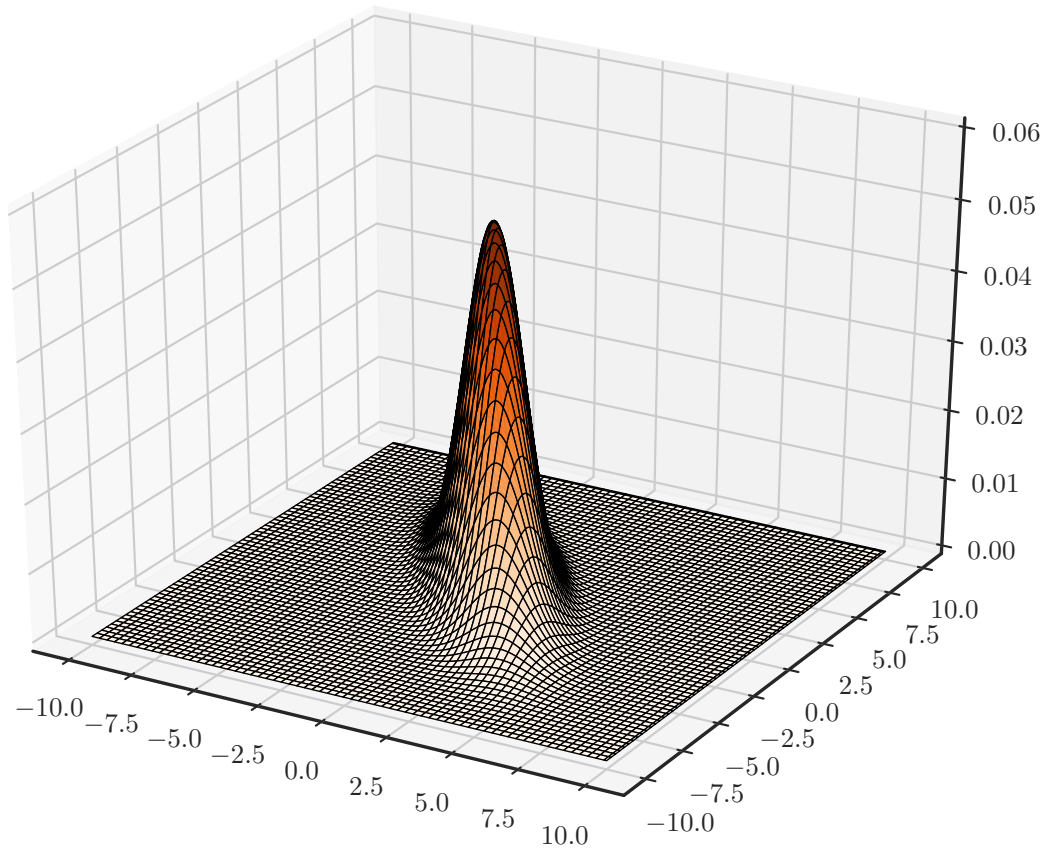
Exemple 2D

$$\Sigma = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} 1 & \\ & 9 \end{bmatrix} \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix}, \theta = 3 \cdot \pi/5$$



Exemple 2D

$$\Sigma = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} 1 & \\ & 9 \end{bmatrix} \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix}, \theta = 4 \cdot \pi/5$$



Propriétés des vecteurs gaussiens

Proposition

Si X est un vecteur gaussien de $\mathbb{R}^p$, et si A est une matrice de $\mathbb{R}^{m\times p}$ et que b est un vecteur de $\mathbb{R}^m$ alors $Y = AX + b$ est un vecteur gaussien de $\mathbb{R}^m$

Construction

Soit $X \in \mathbb{R}^p$ un vecteur gaussien centré-réduit $X \sim \mathcal{N}(0, \mathrm{Id}_p)$.
Supposons que l'on connaisse $L \in \mathbb{R}^{p\times p}$ telle que $LL^\top = \Sigma$, alors pour tout $\mu \in \mathbb{R}^p$, $Y = \mu + LX \sim \mathcal{N}(\mu, \Sigma)$

Démonstration : $\mathbb{C}\mathrm{ov}(Y) = \mathbb{C}\mathrm{ov}(LX) = L\mathbb{C}\mathrm{ov}(X)L^\top = L\,\mathrm{Id}_p\,L^\top = \Sigma$

Rem : L peut être obtenue par la factorisation de Cholesky

Factorisation de Cholesky

Théorème

Toute matrice symétrique définie positive $\Sigma \in \mathbb{R}^{p \times p}$ peut s'écrire $\Sigma = LL^\top$ pour une matrice L triangulaire inférieure

$$L = \begin{bmatrix} L_{11} & 0 & \cdots & 0 \\ L_{21} & L_{22} & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ L_{p1} & L_{p2} & \cdots & L_{pp} \end{bmatrix}$$

Rem : on peut imposer que les éléments diagonaux de la matrice L soient tous positifs ; la factorisation correspondante est alors unique

Rem : numériquement L est obtenue par la méthode du pivot de Gauss, *e.g.* avec `numpy.linalg.cholesky`

Bibliographie

DataScience :

- Blog + videos de Jake Vanderplas : `http://jakevdp.github.io/`, `http://jakevdp.github.io/blog/2017/03/03/reproducible-data-analysis-in-jupyter/`
- VanderPlas (2016), Müller et Guido (2016) : statistiques/apprentissage avec `Python`

Math :

- Hastie *et al.*(2009) : *Elements of Statistical Learning*
- James *et al.*(2013) : *An introduction to statistical learning* (version simplifiée du précédent)
- Tsybakov (2006) cours de "Statistique appliquée"

Références I

Bertsekas, D. P. (1999). *Nonlinear programming*. Athena Scientific.

Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press, Cambridge.

Foata, D. and Fuchs, A. (1996). *Calcul des probabilités : cours et exercices corrigés*. Masson.

Golub, G. H. and van Loan, C. F. (2013). *Matrix computations*. Johns Hopkins University Press, Baltimore, MD, fourth edition.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning*. Springer Series in Statistics. Springer, New York, second edition.
`http://www-stat.stanford.edu/~tibs/ElemStatLearn/`.

Horn, R. A. and Johnson, C. R. (1994). *Topics in matrix analysis*. Cambridge University Press, Cambridge. Corrected reprint of the 1991 original.

Références II

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 6. Springer.

Müller, A. C. and Guido, S. (2016). *Introduction to Machine Learning with Python : A Guide for Data Scientists*. O'Reilly Media, early access edition.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. Monographs on Statistics and Applied Probability. Chapman & Hall, London.

Tsybakov, A. B. (2006). Statistique appliquée.

VanderPlas, J. (2016). *Python Data Science Handbook*. O'Reilly Media.