

MS BGD: MDI720

Statistiques, machine learning, modèle linéaire

François Portier, Anne Sabourin  
Telecom ParisTech

Septembre 2019

- **Anne Sabourin :**

- Précédemment : Université Lyon 1, LSCE-CEA Saclay, Télécom ParisTech
- Spécialités : statistiques des valeurs extrêmes, réduction de la dimension. Applications : détection d'anomalie, risques liés aux événements rares.
- Email : *anne.sabourin@telecom-paristech.fr*
- Bureau : E307

- **François Portier :**

- Précédemment : Université de Rennes 1, Université catholique de Louvain, Télécom ParisTech
- Spécialités : régression parcimonieuse, bootstrap, estimation semi-paramétrique, méthodes à noyaux
- Email : *fportier@enst.fr*
- Bureau : E 302

## 1. Aspects pratiques du cours

## 2. Introduction générale

Modèle statistique

Biais/Variance

## 3. Statistiques descriptives

Résumés basiques d'un jeu de données

Corrélations/Nuage de points

## 4. Rappels de probabilités

Covariances

Les lois gaussiennes

# Calendrier de validation

- Devoir maison 1 : **25% note finale**
  - Sujet disponible le 02/10 à rendre avant 23h59 le vendredi 05/10
  - Correction par les pairs : à rendre pour le 12/10
- Devoir maison 2 : **25% note finale**
  - Sujet disponible 31/10 et à rendre avant 23h59 le 4/11
  - Correction par les pairs : à rendre pour le 12/11

**ATTENTION : le travail rendu doit être personnel!!!**

- Un examen final : **50% note finale**
  - Date : 24/10
  - Format : 3h, sur des thèmes voisin du quizz disponible dès maintenant sur le site pédagogique.

# Plan du cours

**Séance 1.** Anne Sabourin (9/09) : Introduction

**Séance 2.** Anne Sabourin (16/09) : Modèle linéaire ( $p < n$ )

**Séance 3.** (16/09) (Après midi)

**TP non noté. Intro numérique (python)**

**Séance 4.** François Portier (23/09 + 30/09) : Modèle linéaire (suite : propriétés)

**Séance 5.** 30/09 après-midi : **TP noté #1**

**Séance 6.** Anne Sabourin (7/10 + 14/10) : Intervalles de confiance et tests

**Séance 7.** François Portier (14/10) : Bootstrap

**Séance 8.** François Portier (21/10) : Ridge, PCA, SVD

**Séance 9.** François Portier (4/11) : Sélection de variables / Lasso / Cross-validation

**Séance 10.** François Portier (13/11) : Révision

**Séance 11.** 13/11 après-midi : **TP noté #2**

**Séance 12.** (18/11) : **Examen final**

## Prérequis - 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 \(2013\)](#), [Applied Numerical Computing](#) par L. Vandenberghe

# Aspects algorithmiques : quelques conseils

Installation Python (3) : **Conda** / **Anaconda** (tous OS)

Rem: sur ce point entraidez-vous !

Outils :

- Rendus **Jupyter** / **IPython Notebook**
- Projets plus importants : **IPython** + éditeur de texte avancé ; *e.g.* **Atom**, **Sublime Text**, **PyCharm**, etc.

- **Python, Scipy, Numpy** : Reproducible Data Analysis in Jupyter (Tutos de Jake Vanderplas) :

[http://perso.telecom-paristech.fr/~gramfort/liesse\\_python/](http://perso.telecom-paristech.fr/~gramfort/liesse_python/)

- **Pandas** : <https://github.com/jorisvandenbossche/pandas-tutorial>
- **scikit-learn** : <http://scikit-learn.org/stable/tutorial/index.html>

Rem: en TP, prenez vos portables si vous préférez garder votre environnement (packages, versions, etc.)

# Conseils généraux pour l'année

- Utilisez un système de versionnement de fichiers pour vos travaux en groupe : **Git** (*e.g.* **Bitbucket**, **Github**, etc.)
- Adoptez des règles d'écriture de code et tenez-y vous! Exemple : **PEP8** pour Python (utiliser **AutoPEP8**, <https://github.com/kenko000/jupyter-autopep8>)
- Utilisez **Markdown** (.md) (markdown-preview-plus avec **Atom**), *e.g.* pour les parties rédigées / comptes-rendus
- Apprenez de bons exemples (ouvrez les codes sources!) : <https://github.com/scikit-learn/>, <http://jakevdp.github.io/>, etc.



## 1. Aspects pratiques du cours

## 2. Introduction générale

Modèle statistique

Biais/Variance

## 3. Statistiques descriptives

Résumés basiques d'un jeu de données

Corrélations/Nuage de points

## 4. Rappels de probabilités

Covariances

Les lois gaussiennes

# Cadre statistique standard

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

- On observe des réalisations  $(y_1, \dots, y_n)$  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  $(y_1, \dots, y_n)$  ?

## Prédiction

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

# Vocabulaire

- Observations  $\mathbf{y} = y_{1:n} = (y_1, \dots, y_n)$  : **échantillon** de **taille**  $n$
- Grandeurs **théoriques** : dépendent de la loi  $\mathbb{P}_Y$  (**inconnue**) et contrôlent la génération des observations

Exemple : l'espérance  $\mathbb{E}(Y)$  ou la variance  $\text{Var}(Y)$  de  $Y$

- Grandeurs **empiriques** : calculées à partir des observations  $y_i$

Exemple : 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  $\delta_{y_i}$  est la mesure de Dirac au point  $y_i$ )

1. Aspects pratiques du cours

2. Introduction générale

Modèle statistique

Biais/Variance

3. Statistiques descriptives

4. Rappels de probabilités

# Modèle statistique : contexte

## Rappel

- On observe des réalisations  $(y_1, \dots, y_n)$  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. Exemple : “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  $\mathbb{P}_Y$   
Exemple : 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  $\theta$  inconnu, et  $\Theta$  est l'ensemble d'indexation

Exemple : “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  $\Theta$ .

---

**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

Exemple :

- 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.

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

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

## Définition

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

$$\hat{g} : (y_1, \dots, y_n) \mapsto \hat{g}(y_1, \dots, y_n)$$

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

Rem: en pratique l'estimateur doit être calculable efficacement



1. Aspects pratiques du cours

2. Introduction générale

Modèle statistique

Biais/Variance

3. Statistiques descriptives

4. Rappels de probabilités

# Propriétés d'un estimateur : le biais

## Définition

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

$$\text{Biais}(\hat{g}, g) = \mathbb{E}_{\theta}(\hat{g}(Y_1, \dots, Y_n)) - g(\theta) \quad (\text{dépend de } \theta)$$

## Définition

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

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

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  $\bar{y}_n = \frac{1}{n} \sum_{i=1}^n y_i$  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:  $\hat{g}(y_1, \dots, y_n) = y_1$  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  $g(\theta) = \text{Var}_\theta(Y)$

### Théorème

L'estimateur  $\hat{g}(y_1, \dots, y_n) = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y}_n)^2$  est un estimateur sans biais de la variance  $\text{Var}_\theta(Y)$

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

---

**Exo**: Vérifier cette propriété par le calcul

---

# Propriétés d'un estimateur : la variance

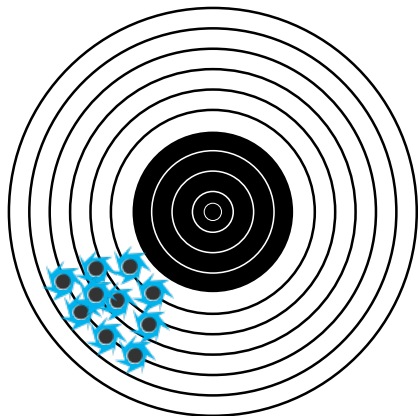
## Définition

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

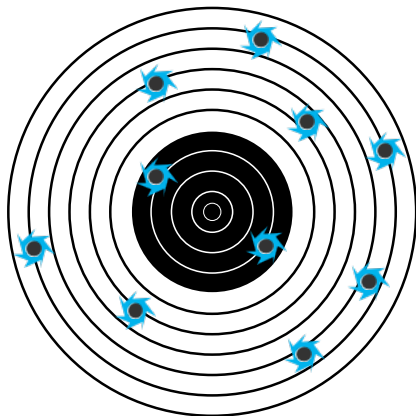
$$\text{Var}_\theta(\hat{g}) = \text{Var}_\theta(\hat{g}(Y_1, \dots, Y_n)) = \mathbb{E}_\theta(\hat{g} - \mathbb{E}_\theta(\hat{g}))^2 \quad (\text{dépend de } \theta)$$

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

## Biais ou variance ?

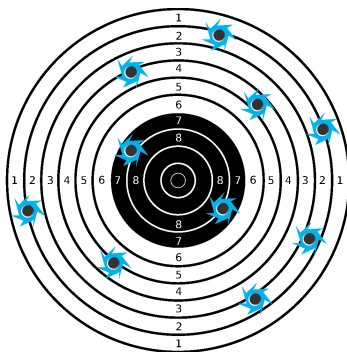
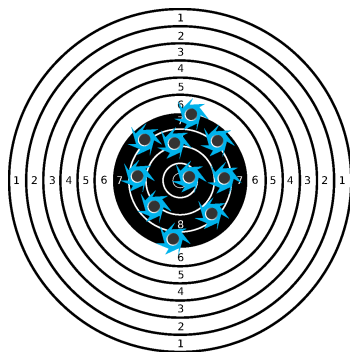


Erreurs systématiques



Erreurs stochastiques

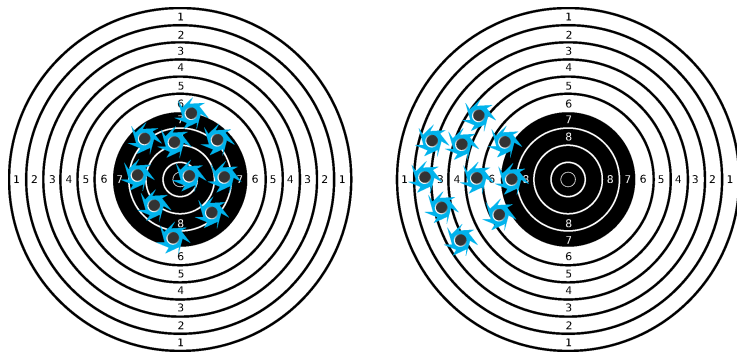
# Biais ou variance ?



1

- Si  $\hat{g}_0$  et  $\hat{g}_1$  sont sans biais, on préfère avoir une faible variance

## Biais ou variance ?



- Si  $\hat{g}_0$  et  $\hat{g}_1$  ont la même variance, on préfère un biais faible



# Risque quadratique / compromis biais-variance

## Définition

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

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

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

## Théorème : décomposition biais / variance

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

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

$$\begin{aligned} R(\hat{g}) &= \mathbb{E} [(\hat{g} - \mathbb{E}(\hat{g}) + B)^2] \\ &= \mathbb{E} [(\hat{g} - \mathbb{E}(\hat{g}))^2 + B^2 + 2B(\hat{g} - \mathbb{E}(\hat{g}))] \\ &= \text{Var}(\hat{g}) + B^2 + 2B \underbrace{\mathbb{E} [\hat{g} - \mathbb{E}(\hat{g})]}_{=0} = \text{Var}(\hat{g}) + B^2 \end{aligned}$$

## 1. Aspects pratiques du cours

## 2. Introduction générale

Modèle statistique

Biais/Variance

## 3. Statistiques descriptives

Résumés basiques d'un jeu de données

Corrélations/Nuage de points

## 4. Rappels de probabilités

Covariances

Les lois gaussiennes

1. Aspects pratiques du cours

2. Introduction générale

3. Statistiques descriptives

Résumés basiques d'un jeu de données

Corrélations/Nuage de points

4. Rappels de probabilités

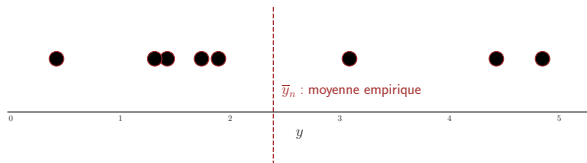
# Statistique exploratoire et descriptive

- 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  $(y_1, \dots, y_n)$  (et qu'on espère être une fonction calculable des observations  $(y_1, \dots, y_n)$ !)

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

# Moyenne (arithmétique)



## Définition

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

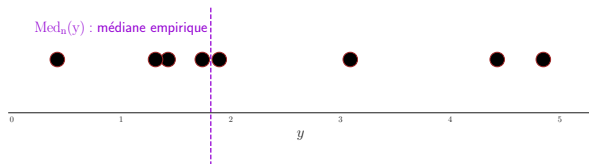
Si  $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^n x_i y_i$  (produit scalaire) et  $\mathbb{1}_n = (1, \dots, 1)^\top \in \mathbb{R}^n$  :

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

---

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

# Médiane



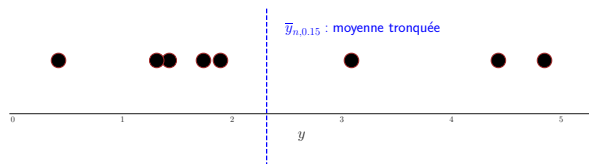
On ordonne les  $y_i$  dans l'ordre croissant :  $y_{(1)} \leq y_{(2)} \leq \dots \leq y_{(n)}$

## Définition

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

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

# Moyenne tronquée



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

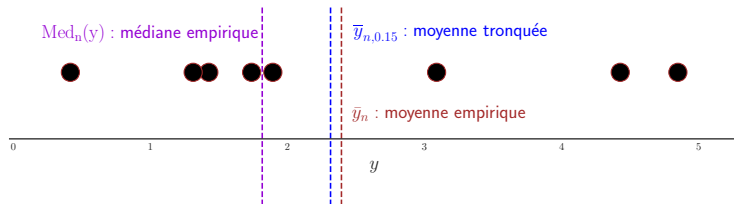
## Définition

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

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

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

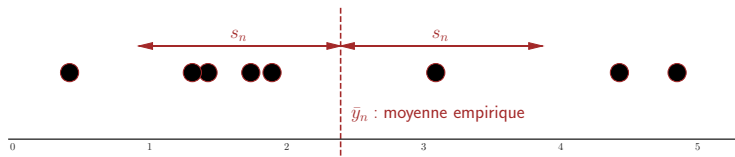
# Moyenne vs médiane



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



## Dispersion : variance / écart-type



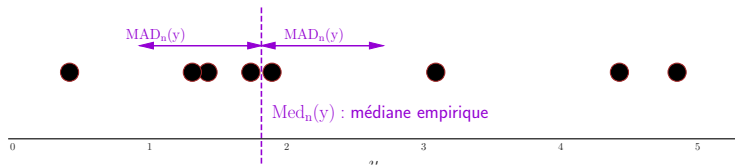
### Définitions

**Variance :**  $\text{var}_n(\mathbf{y}) = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y}_n)^2 = \frac{1}{n} \|\mathbf{y} - \bar{y}_n \mathbf{1}_n\|^2$

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

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

# Dispersion : MAD

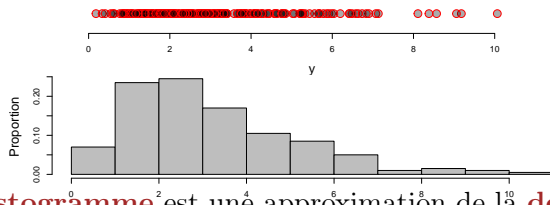


## Définition

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

$$\text{MAD}_n(\mathbf{y}) = \text{Med}_n(|\text{Med}_n(\mathbf{y}) - \mathbf{y}|)$$

# Estimation de la densité : histogramme

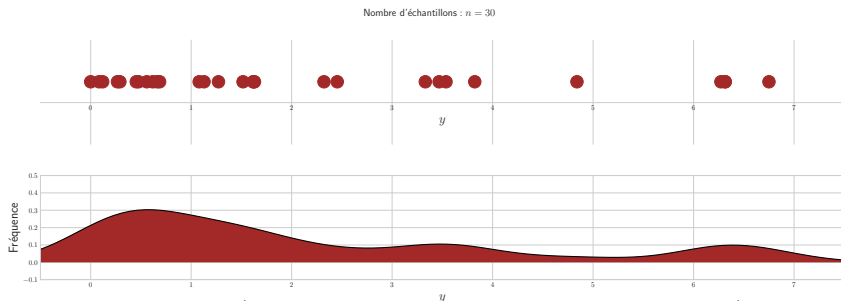


L'**histogramme** est une approximation de la **densité** par une fonction constante par morceaux

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`

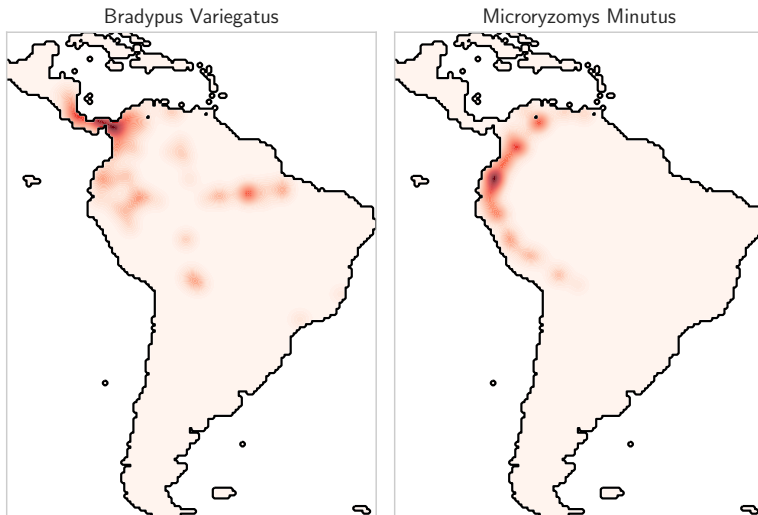
# Estimation de la densité : méthode à noyau



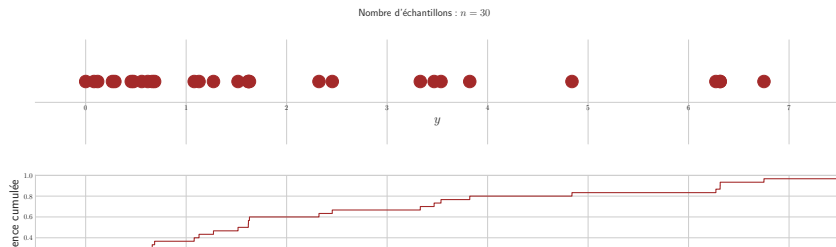
- Méthode à noyau (🇬🇧 : *Kernel Density Estimation, KDE*) :  
approche non-paramétrique estimant la densité par une fonction  
continue – généralisation de l'histogramme

Pour plus de détails voir le livre [Silverman \(1986\)](#)

## Densité bi-dimensionnelle (spatiale)



# Fonction de répartition



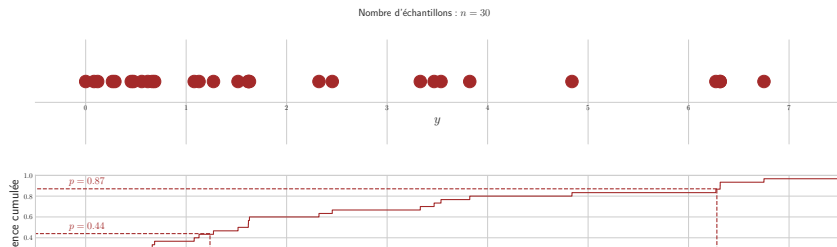
## Définition : fonction de répartition

**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

1. Aspects pratiques du cours

2. Introduction générale

3. Statistiques descriptives

Résumés basiques d'un jeu de données

Corrélations/Nuage de points

4. Rappels de probabilités



# Covariances et corrélations empiriques

## Covariance empirique

Pour deux échantillons  $\mathbf{x}$  et  $\mathbf{y}$  de moyennes et variances empiriques  $\bar{x}_n$ ,  $\bar{y}_n$  et  $\text{var}_n(\mathbf{x})$ ,  $\text{var}_n(\mathbf{y})$  :

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

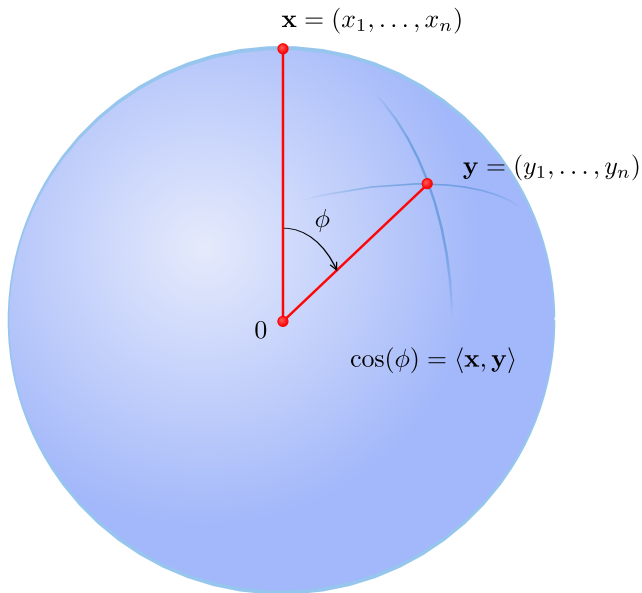
$$\text{cov}_n(\mathbf{x}, \mathbf{y}) = \frac{1}{n} \langle \mathbf{x} - \bar{x}_n \mathbb{1}_n, \mathbf{y} - \bar{y}_n \mathbb{1}_n \rangle$$

## Corrélation empirique

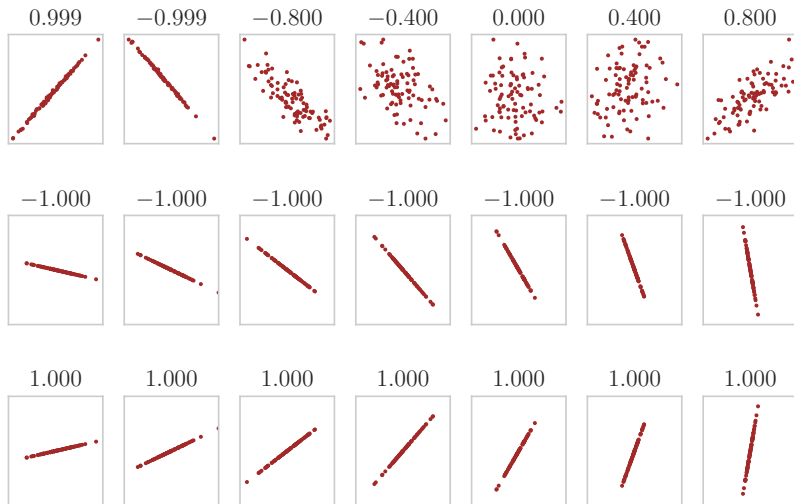
$$\rho = \text{corr}_n(\mathbf{x}, \mathbf{y}) = \frac{\text{cov}_n(\mathbf{x}, \mathbf{y})}{\sqrt{\text{var}_n(\mathbf{x})} \sqrt{\text{var}_n(\mathbf{y})}}, \quad \text{c'est-à-dire}$$

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

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

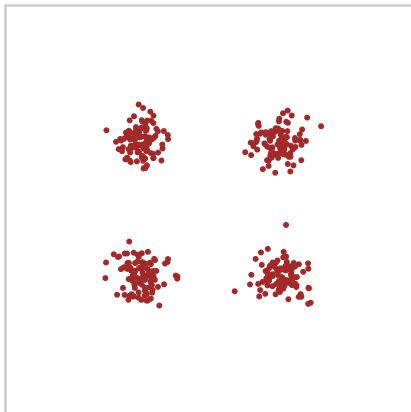


# Exemples de corrélations



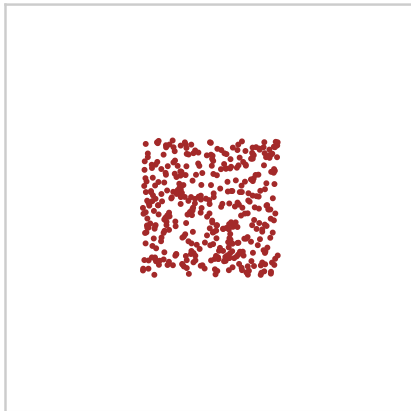
## Exemples de corrélations proches de zéro

Corrélation =  $-0.021$



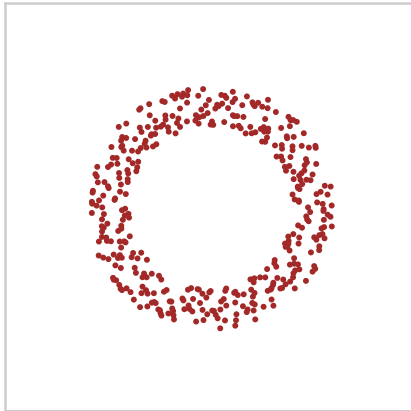
## Exemples de corrélations proches de zéro

Corrélation = 0.007

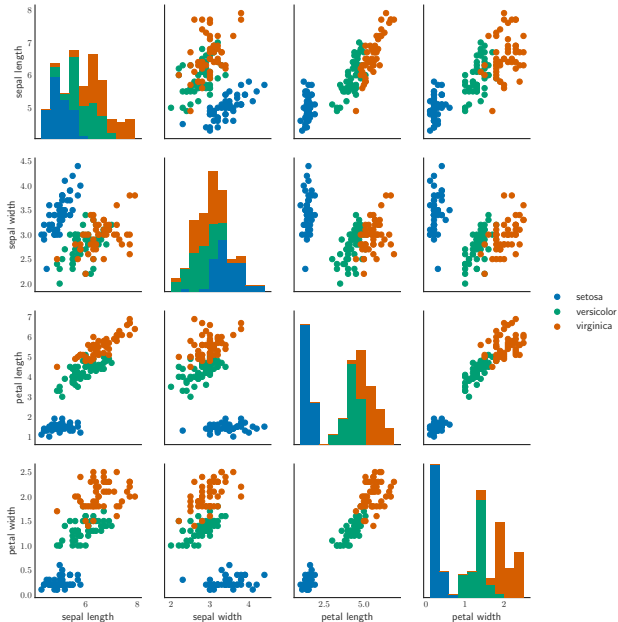


## Exemples de corrélations proches de zéro

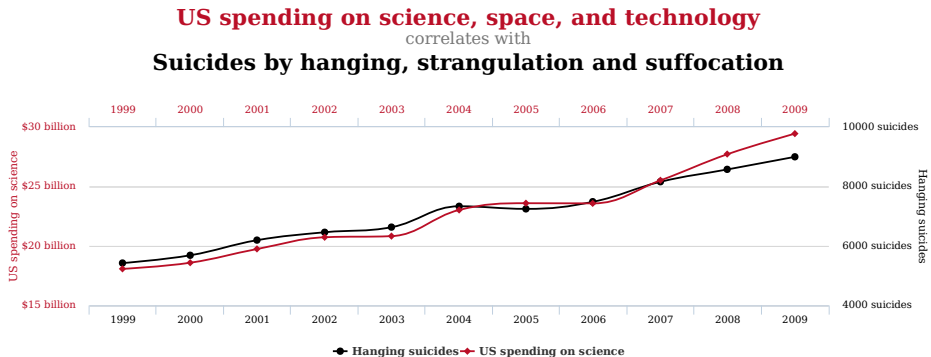
Corrélation = 0.011



# Nuages de points / Scatter plot / PairGrid



Covariance  $\neq$  causalité



tylervigen.com

Corrélation : 0.9979

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



## 1. Aspects pratiques du cours

## 2. Introduction générale

Modèle statistique

Biais/Variance

## 3. Statistiques descriptives

Résumés basiques d'un jeu de données

Corrélations/Nuage de points

## 4. Rappels de probabilités

Covariances

Les lois gaussiennes

1. Aspects pratiques du cours

2. Introduction générale

3. Statistiques descriptives

4. Rappels de probabilités

Covariances

Les lois gaussiennes

## 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 :

$$\text{Cov}(X, Y) = \mathbb{E} [(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))]$$

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

$$\text{Cov}(\alpha X_1 + \beta X_2, Y_1) = \alpha \text{Cov}(X_1, Y_1) + \beta \text{Cov}(X_2, Y_1)$$

$$\text{Cov}(X_1, \alpha Y_1 + \beta Y_2) = \alpha \text{Cov}(X_1, Y_1) + \beta \text{Cov}(X_1, Y_2)$$

Rappel : inégalité de Cauchy–Schwarz dans ce cadre

$$|\text{Cov}(X, Y)| \leq \sqrt{\text{Var}(X) \text{Var}(Y)}$$

## Matrice de covariance d'un vecteur aléatoire

Notation :  $X = (X_1, \dots, X_p)^\top$  est vecteur aléatoire t.q.

$\forall j \in \{1, \dots, p\}, \mathbb{E}(X_j^2) < +\infty$  et  $\sigma_{i,j} = \text{cov}(X_i, X_j)$  ( $\sigma_{i,i} = \text{var}(X_i)$ )

### Définition

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

$$\mathbb{Cov}(X) = \begin{pmatrix} \text{var}(X_1) & \text{cov}(X_1, X_2) & \dots & \text{cov}(X_1, X_p) \\ \text{cov}(X_2, X_1) & \text{var}(X_2) & & \vdots \\ \vdots & & \ddots & \vdots \\ \text{cov}(X_p, X_1) & \dots & & \text{var}(X_p) \end{pmatrix} \in \mathbb{R}^{p \times p}$$

Version condensée :  $\mathbb{Cov}(X) = \mathbb{E}[(X - \mathbb{E}(X))(X - \mathbb{E}(X))^\top]$

---

**Exo:** Montrer que pour  $\mu$  déterministe  $\mathbb{Cov}(X + \mu) = \mathbb{Cov}(X)$

---

## Quelques propriétés de la covariance

- Une matrice de covariance est symétrique :

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

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

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

Démonstration :

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

---

**Exo:**  $\mathbb{Cov}(AX) = A\mathbb{Cov}(X)A^\top$ , 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  $\lambda_1 \geq \dots \geq \lambda_n$  et une matrice orthogonale  $U \in \mathbb{R}^{n \times n}$  telle que :

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

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

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

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

Vocabulaire :

- les  $\lambda_i$  sont les **valeurs propres** de  $S$  ( : *eigenvalues*)
- les  $\mathbf{u}_i$  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))
```



1. Aspects pratiques du cours

2. Introduction générale

3. Statistiques descriptives

4. Rappels de probabilités

Covariances

Les lois gaussiennes

# 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

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

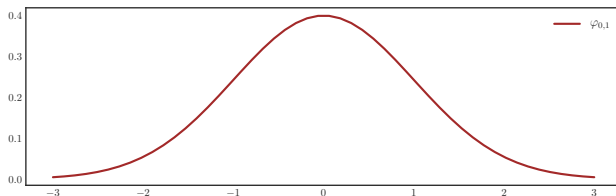
On note alors  $X \sim \mathcal{N}(0, 1)$ .

- Une v.a.  $Y$  suit une loi normale de paramètres  $\mu$  et  $\sigma^2$  si

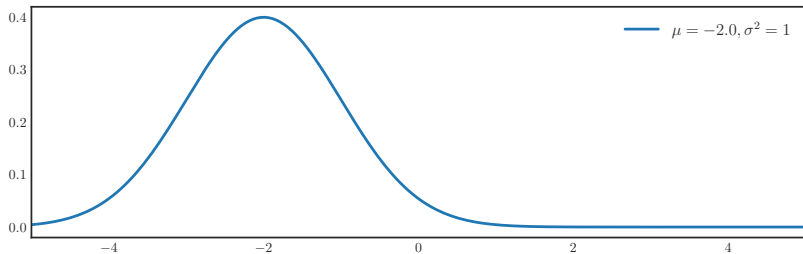
$$Y = \mu + \sqrt{\sigma^2}X, \text{ où } X \sim \mathcal{N}(0, 1), \text{ et on note } Y \sim \mathcal{N}(\mu, \sigma^2)$$

Densité :

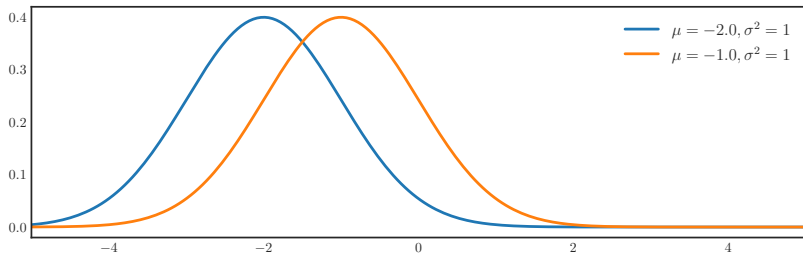
$$\varphi_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



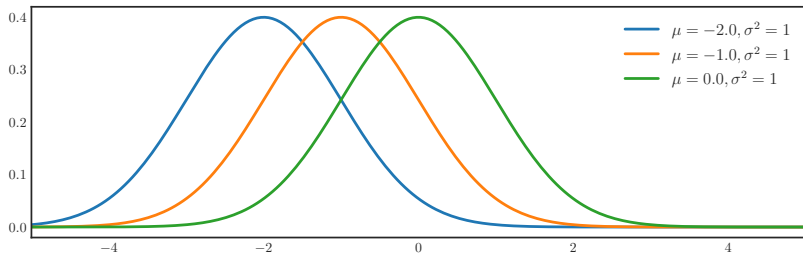
## Exemple : variation sur $\mu$



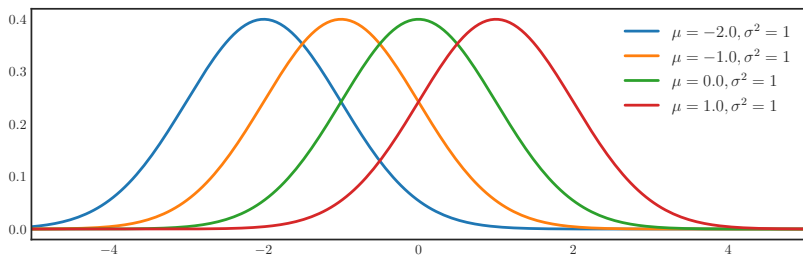
## Exemple : variation sur $\mu$



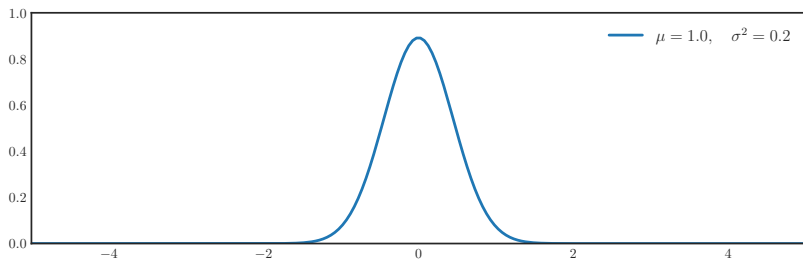
## Exemple : variation sur $\mu$



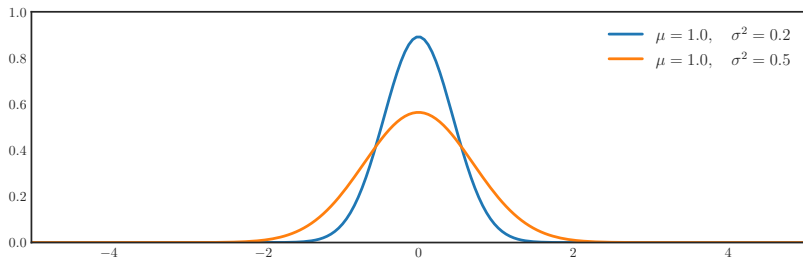
## Exemple : variation sur $\mu$



## Exemple : variation sur $\sigma$

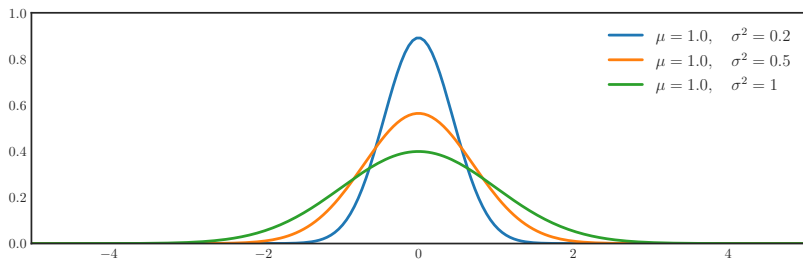


## Exemple : variation sur $\sigma$

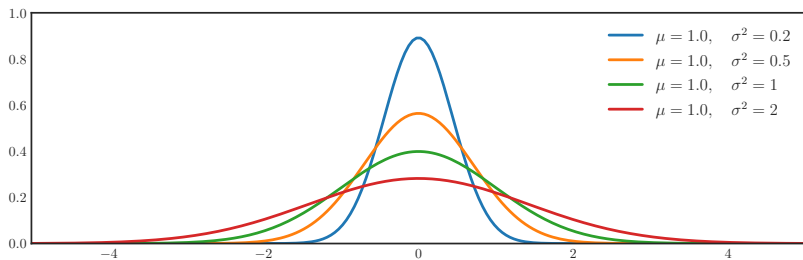




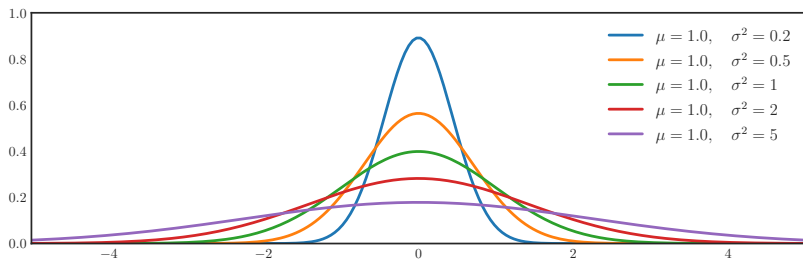
## Exemple : variation sur $\sigma$



## Exemple : variation sur $\sigma$



## Exemple : variation sur $\sigma$



## Vecteurs gaussiens

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

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

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

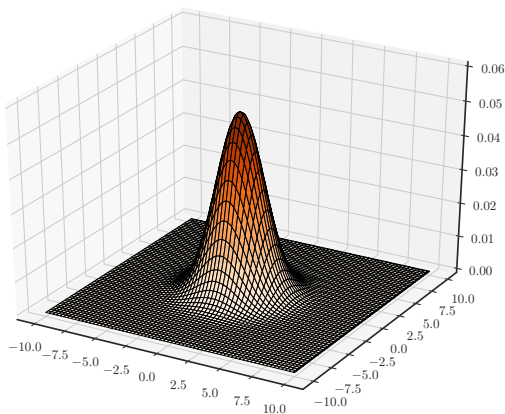
- le vecteur d'espérance  $\mu \in \mathbb{R}^p$
- la matrice de covariance  $\Sigma \in \mathbb{R}^{p \times p}$

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

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

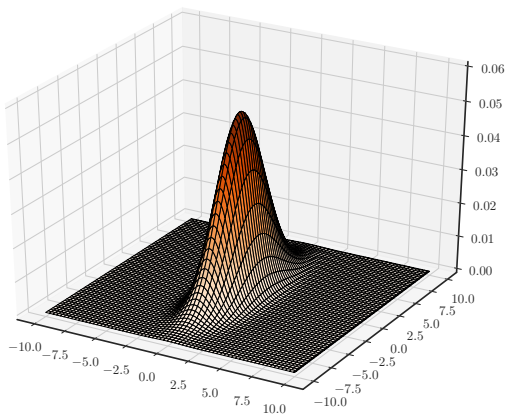
## Exemple 2D

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



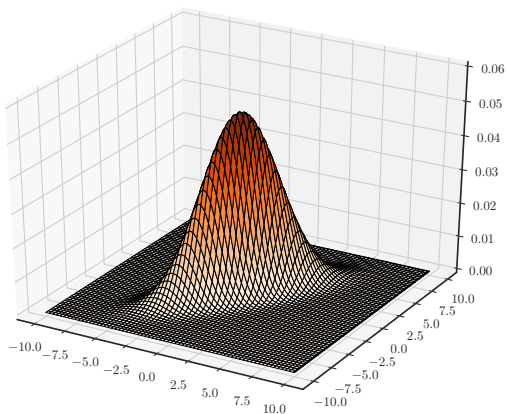
## Exemple 2D

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



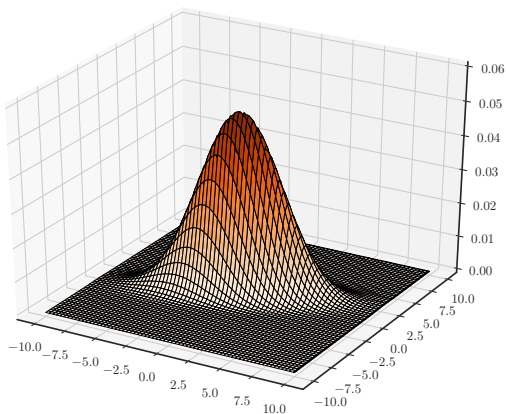
## Exemple 2D

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



## Exemple 2D

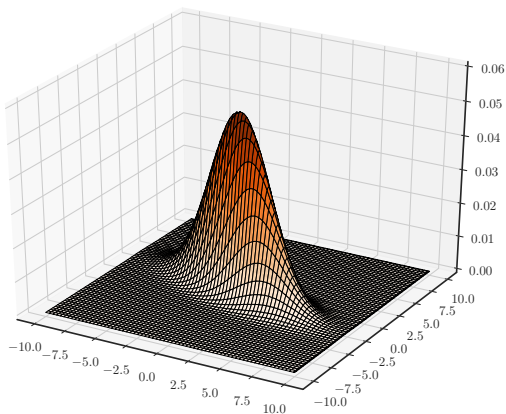
$$\Sigma = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} 1 & \\ & 9 \end{bmatrix} \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix}, \theta = 2 \cdot \pi/5$$





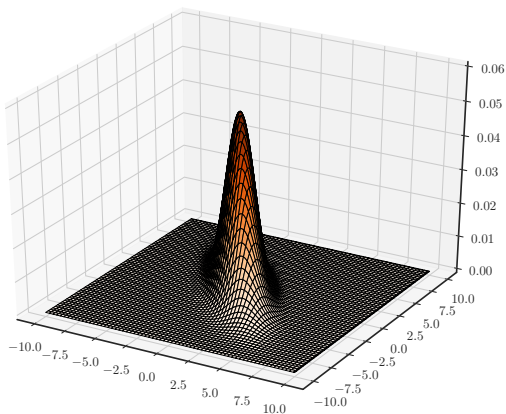
## Exemple 2D

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



## 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  $X \sim \mathcal{N}(0, \text{Id}_p)$ .  
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 :  $\text{Cov}(Y) = \text{Cov}(LX) = L\text{Cov}(X)L^\top = L\text{Id}_p L^\top = \Sigma$

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
- Exemples d'application de `scikit-learn` : [http://www.baglom.com/b/10-scikit-learn-case-studies-examples-tutorials-cm572/?utm\\_content=bufferbde5d&utm\\_medium=social&utm\\_source=twitter.com&utm\\_campaign=buffer](http://www.baglom.com/b/10-scikit-learn-case-studies-examples-tutorials-cm572/?utm_content=bufferbde5d&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer)

## 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”
- Delyon (2015) cours de Régression

# Références I

- Bertsekas, D. P. (1999). *Nonlinear programming*. Athena Scientific.
- Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press, Cambridge.
- Delyon, B. (2015). Régression. <https://perso.univ-rennes1.fr/bernard.delyon/regression.pdf>.
- 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/>.

## Références II

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