

# **MS BGD**

## **MDI 720 : Statistiques**

**Joseph Salmon**

<http://josephsalmon.eu>

Télécom Paristech, Institut Mines-Télécom

# Enseignants

- **Joseph Salmon :**

- Positions : Paris Diderot-Paris 7, Duke University, Télécom ParisTech
- Spécialités : statistiques en grande dimension, agrégation, traitement des images, optimisation
- Email : *joseph.salmon@telecom-paristech.fr*
- Bureau : E410

- **Francois Portier :**

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

# Validation cours MDI 720

- Devoir maison 1 : **20% note finale**
  - Date : sujet mis à disposition le 04/10/2016 et à rendre avant 23h59 le dimanche 9/10/2016
  - Format de rendu : **IPython Notebook**
  - Validation par pairs : à rendre pour le vendredi 14/10/2016
- Devoir maison 2 : **20% note finale**
  - Date : sujet mis à disposition le 12/10/2016 et à rendre avant 23h59 le dimanche 16/10/2016 (zéro passé cette limite)
  - Format de rendu : **IPython Notebook**
  - Validation par pairs : à rendre pour le vendredi 21/10/2016
- Un devoir final : **60% note finale**
  - Date : 9/11/2016
  - Format : partie Quiz (1h30), partie numérique (1h30) qui pourra être rendue jusqu'à 23h59 le 9/11/2016.

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

# Notation des parties numériques

Pour chaque rendu, note totale sur **20**, avec en détails :

- Qualité des réponses aux questions **15** pts
- Qualité de rédaction, de présentation et d'orthographe **2** pts
- Indentation, Style PEP8, commentaires dans le code **2** pts
- absence de bug **1** pt

Retard : 0 pour tout retard, sauf excuse validée par l'administration

Consigne supplémentaire : un fichier unique au format **.ipynb**

**Aucun travail par mail accepté !**

# Bonus

**1 pt** supplémentaire sur **la note finale** pour toute contribution à l'amélioration des cours (présentations, codes, etc.)

## Contraintes :

- ▶ seule la première amélioration reçue est “rémunérée”
- ▶ déposer un fichier **.txt** par proposition dans la partie du site pédagogique intitulée “Propositions d'améliorations”
- ▶ détailler précisément (ligne de code, page des présentations, etc.) l'amélioration proposée, ce qu'elle corrige et/ou améliore
- ▶ Pour les fautes d'orthographe : minimum de 5 fautes trouvées par contribution

# Plan du cours

- Séance 1.** Joseph Salmon (21/09) : Introduction
- Séance 2.** Joseph Salmon (23/09) : Modèle linéaire ( $p < n$ )
- Séance 3.** A. G. / E.N. / F.P. / J. S. (28/09) : Intro. numérique
- Séance 4.** Joseph Salmon (30/09) : Modèle linéaire (suite)
- Séance 5.** A. G. / E.N. / F.P. / J. S. (4/10) : **TP noté #1**
- Séance 6.** Joseph Salmon (5/10) : SVD / IC
- Séance 7.** François Portier (11/10) : IC / Bootstrap
- Séance 8.** A. G. / E. N. / F. P. / J. S. (12/10) : **TP noté #2**
- Séance 9.** Joseph Salmon (19/10) : Tests, Ridge
- Séance 10.** Joseph Salmon (26/10) : Sélection de variables / Lasso
- Séance 11.** Joseph Salmon (02/11) : GLM / Régression logistique
- Séance 12.** F. P. / J. S. (09/11) : **Validation finale**

# Prérequis - à revoir seul

- ▶ Bases de **probabilités** : probabilité, 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)

# Prérequis - à revoir seul

- ▶ Bases de **probabilités** : probabilité, 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)



## Prérequis - à revoir seul

- ▶ Bases de **probabilités** : probabilité, 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

## Prérequis - à revoir seul

- ▶ Bases de **probabilités** : probabilité, 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 : **Conda** / **Anaconda** (tous OS)

Rem: sur ce point je ne peux rien pour vous : entraidez-vous !

Outils conseillés : **Jupyter** / **IPython Notebook** (rendus notamment) **IPython** avec éditeur de texte e.g., **Atom**, **Sublime Text**, **PyCharm**, etc., pour des projets plus importants

- ▶ **Python, Scipy, Numpy** :

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

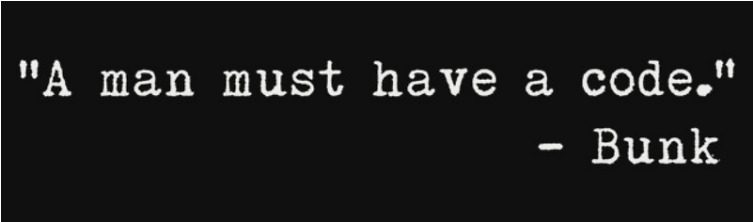
- ▶ **Pandas** : <http://pandas.pydata.org/>

- ▶ **scikit-learn** : <http://scikit-learn.org/stable/>

Rem: en TP, venir éventuellement avec 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, e.g., **AutoPEP8** avec Sublime Text, Atom



"A man must have a code."  
- Bunk

- ▶ Apprenez de bons exemples :  
<https://github.com/scikit-learn/>,  
<http://jakevdp.github.io/>, etc.

# Plan

## Bonus

### Introduction générale

- Modèle statistique

- Biais/Varianace

### Statistiques descriptives

- Résumés basiques d'un jeu de données

- Corrélations/Nuage de points

### 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  $\mathbb{P}_Y$

## Estimation

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

## Prédiction

Souvent : 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** : dépendant de la loi  $\mathbb{P}_Y$ , **inconnue**, et qui génère les observations  
Exemple : l'espérance ou la variance 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  $\mathbb{P}_n := \frac{1}{n} \sum_{i=1}^n \delta_{y_i}$

# Sommaire

## Bonus

### Introduction générale

Modèle statistique

Biais/Variance

### Statistiques descriptives

Résumés basiques d'un jeu de données

Corrélations/Nuage de points

### Rappels de probabilités

Covariances

Les lois gaussiennes



# 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  $\Theta$  pour le “lancé 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 ensemble fini, disons  $p$ , 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, quantiles, 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 (mesurable) 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: classiquement on demande la mesurabilité, mais pour être utilisable il faut aussi que l'estimateur soit calculable efficacement

# Sommaire

## Bonus

### Introduction générale

Modèle statistique

Biais/Variance

### Statistiques descriptives

Résumés basiques d'un jeu de données

Corrélations/Nuage de points

### Rappels de probabilités

Covariances

Les lois gaussiennes

# 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, quel que soit  $\theta \in \Theta$ ,

$$\mathbb{E}(\hat{g}(y_1, \dots, y_n)) = g(\theta)$$

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 à estimer  $g(\theta) = \mathbb{E}(Y)$

## Théorème

La moyenne empirique  $\hat{g}(y_1, \dots, y_n) = \bar{y}_n = \frac{1}{n} \sum_{i=1}^n y_i$  est un estimateur sans biais de l'espérance  $\mathbb{E}(Y)$

Démonstration :

$$\mathbb{E}(\hat{g}(y_1, \dots, y_n)) = \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^n y_i\right) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}(y_i) = \mathbb{E}(Y)$$

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

Rem: l'estimateur  $\hat{g}(y_1, \dots, y_n) = y_1$  est aussi 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 à estimer  $g(\theta) = \text{Var}(Y)$

## 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  $\text{Var}(Y)$

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

---

**Exo**: Vérifier cela par le calcul, par exemple sur des gaussiennes

---



# Propriétés d'un estimateur : la variance

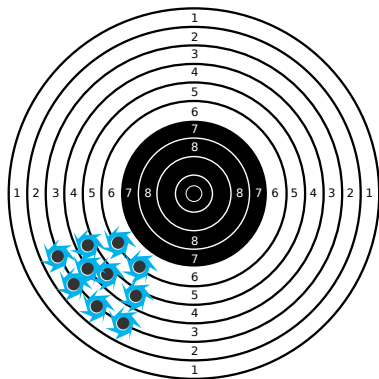
## Définition

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

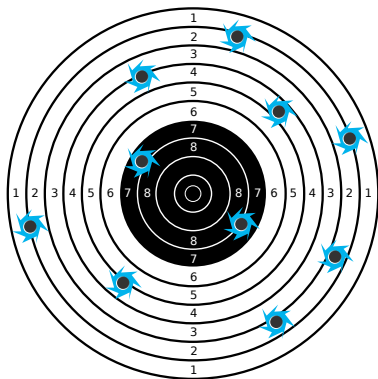
$$\text{Var}(\hat{g}) = \text{Var}(\hat{g}(y_1, \dots, y_n)) = \mathbb{E}_\theta(\hat{g} - \mathbb{E}_\theta(\hat{g}))^2 \quad (\text{dépend de } \theta)$$

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

# Biais ou variance ?

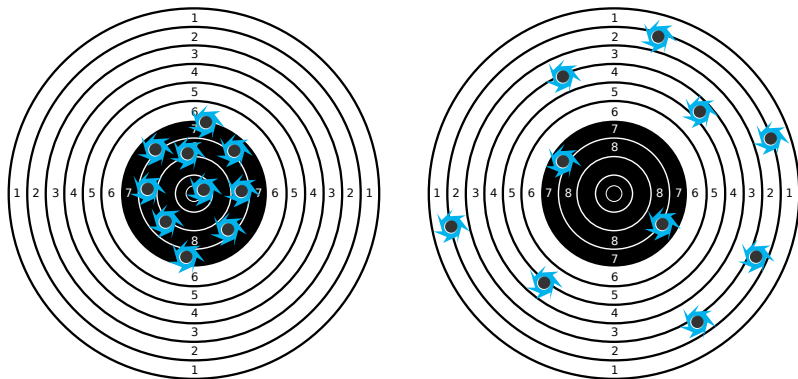


Erreurs systématiques



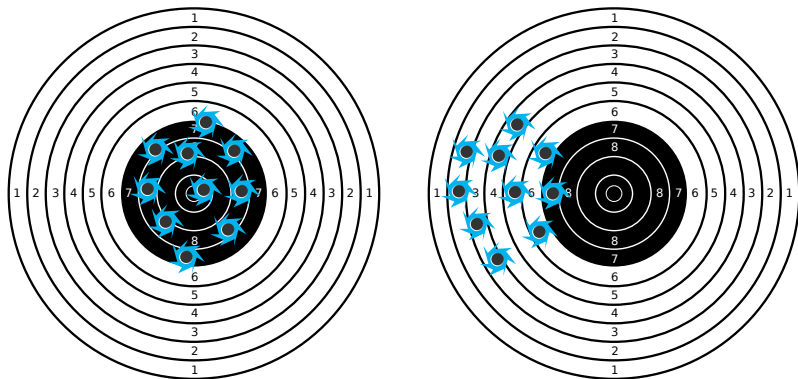
Erreurs stochastiques

## Biais ou variance ?



- 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

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

# Sommaire

## Bonus

## Introduction générale

Modèle statistique

Biais/Variance

## Statistiques descriptives

Résumés basiques d'un jeu de données

Corrélations/Nuage de points

## Rappels de probabilités

Covariances

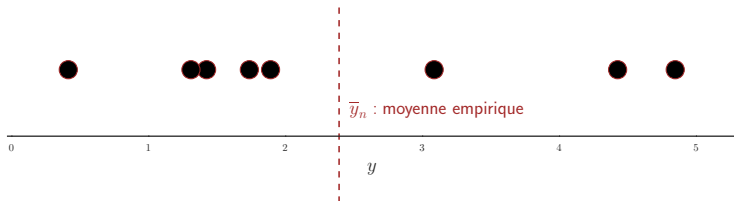
Les lois gaussiennes

# 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  $(y_1, \dots, y_n)$ .

# Moyenne (arithmétique)



## Définition

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

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

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

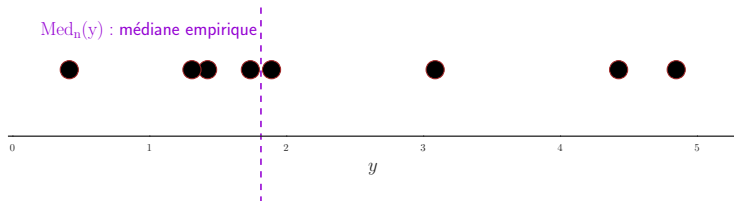
---

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

---



# Médiane



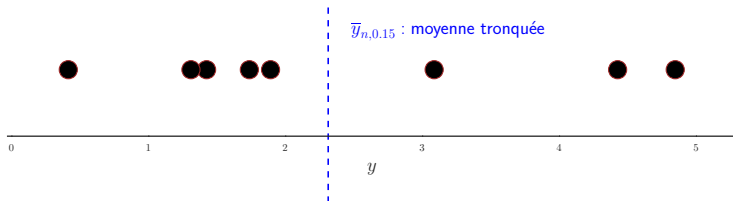
On ordonne les  $y_i$  dans l'ordre croissant :  $y_{(1)} \leq y_{(2)} \leq \dots \leq y_{(n)}$

## 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...

# Moyenne tronquée



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

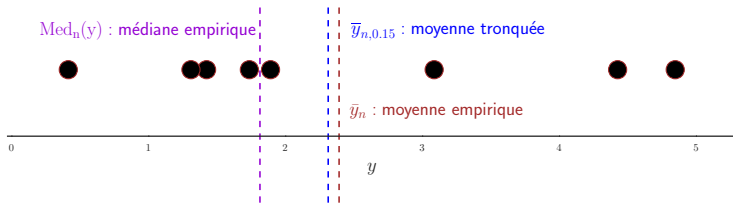
## Définition


**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

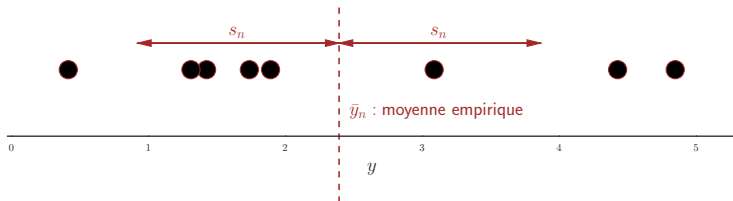
Rem:  $\lfloor u \rfloor$  est le nombre entier tel que  $\lfloor u \rfloor - 1 < u \leq \lfloor u \rfloor$

# Moyenne vs médiane



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

# Dispersion



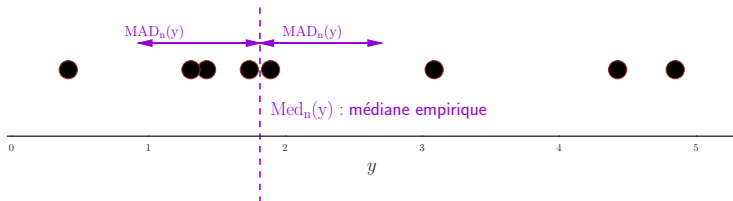
## Définitions

**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

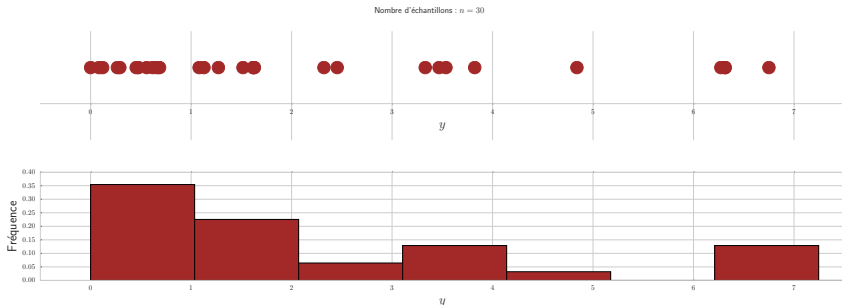


## Définition

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

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

# Estimation de la densité : histogramme

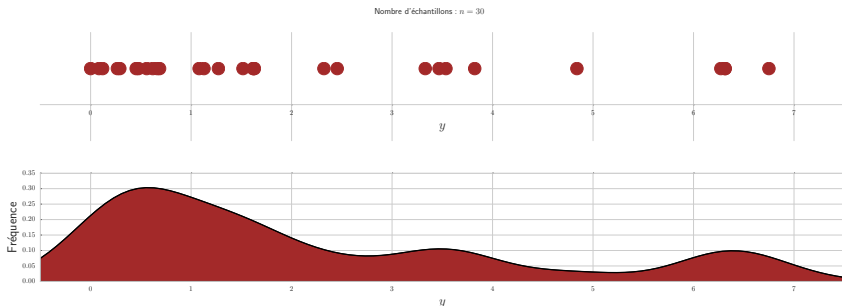


L'**histogramme** est une approximation de la **densité** par une fonction constante par morceaux

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, e.g., avec `normed=False(=True)` dans la fonction `hist`

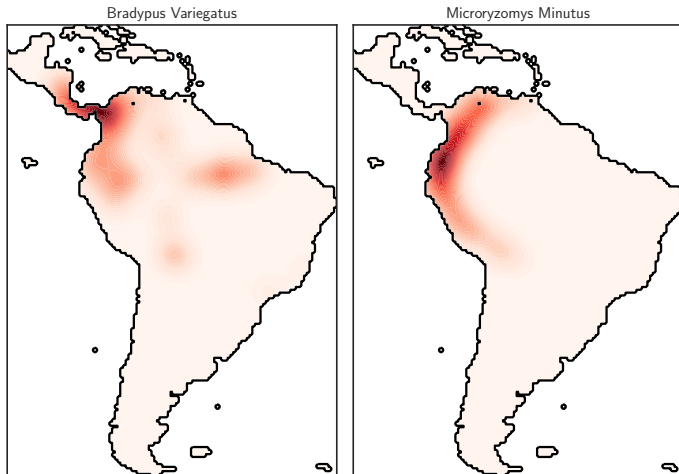
# Estimation de la densité : méthode à noyau



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

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

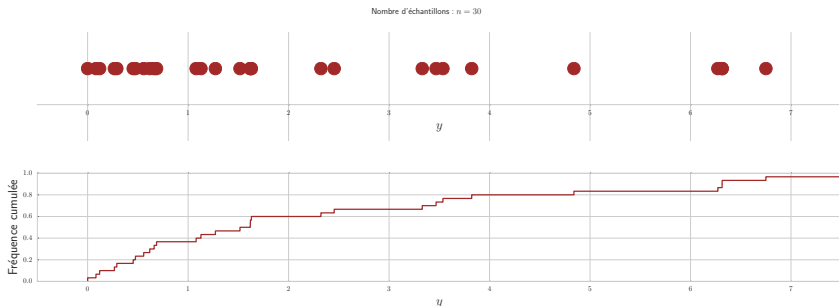
# Densité bi-dimensionnelle



cf. [http://scikit-learn.org/stable/\\_downloads/plot\\_species\\_kde.py](http://scikit-learn.org/stable/_downloads/plot_species_kde.py)



# Fonction de répartition



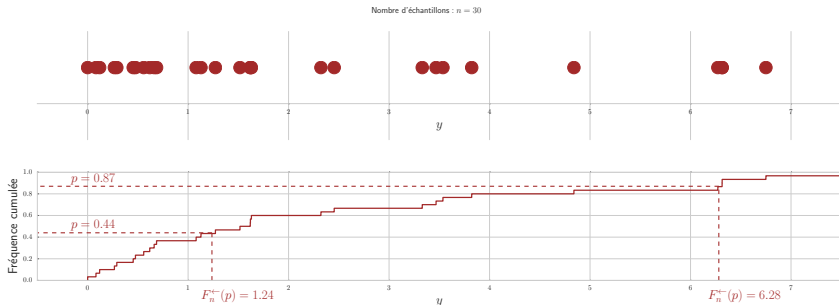
## Définition : fonction de répartition

**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'observation en-dessous 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  $p$ ) :  $F_n^{\leftarrow}(p) = y_{(\lfloor np \rfloor)}$

Rem: c'est l'inverse (généralisée) de la fonction de répartition

# Sommaire

## Bonus

### Introduction générale

Modèle statistique

Biais/Varianace

### Statistiques descriptives

Résumés basiques d'un jeu de données

Corrélations/Nuage de points

### Rappels de probabilités

Covariances

Les lois gaussiennes

# 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  $\text{var}_n(\mathbf{x})$ ,  $\text{var}_n(\mathbf{y})$  :

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

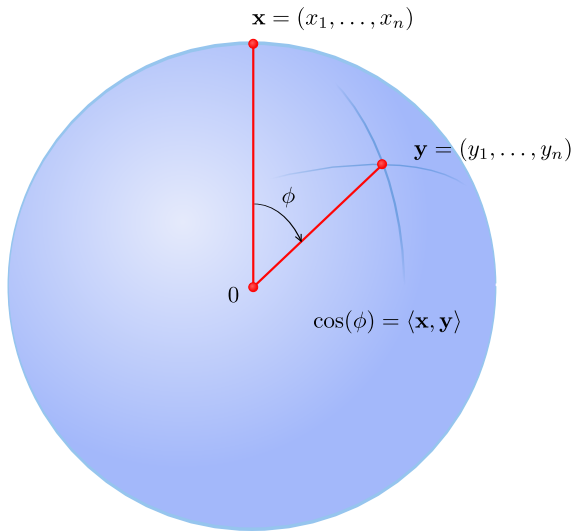
$$\text{cov}_n(\mathbf{x}, \mathbf{y}) = \frac{1}{n} \langle \mathbf{x} - \bar{x}_n \mathbf{1}_n, \mathbf{y} - \bar{y}_n \mathbf{1}_n \rangle$$

## Corrélation empirique

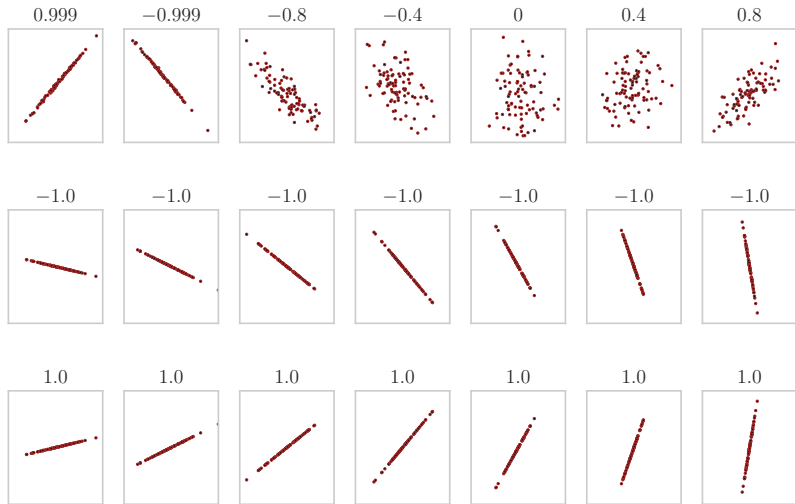
$$\rho = \text{corr}_n(\mathbf{x}, \mathbf{y}) = \frac{\text{cov}_n(\mathbf{x}, \mathbf{y})}{\sqrt{\text{var}_n(\mathbf{x})} \sqrt{\text{var}_n(\mathbf{y})}}, \quad \text{c'est-à-dire}$$

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

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

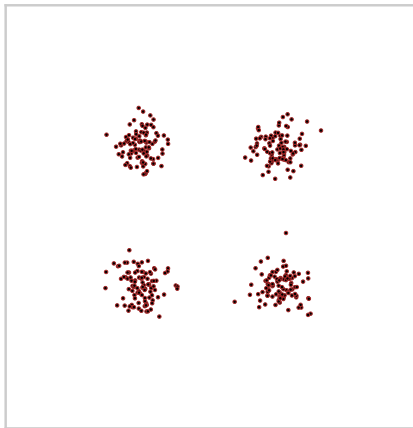


# Exemples de corrélations



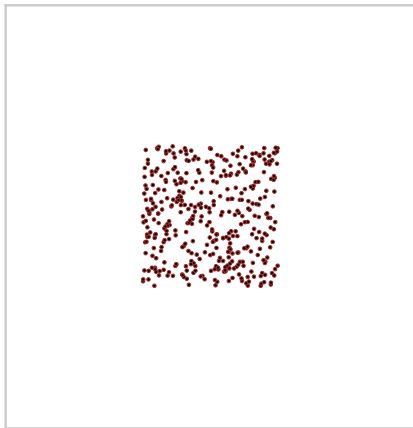
## Exemples de corrélations proches de zéro

Corrélation =  $-0.021$



# Exemples de corrélations proches de zéro

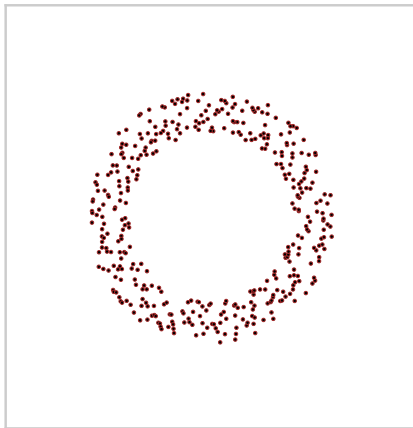
Corrélation = 0.007



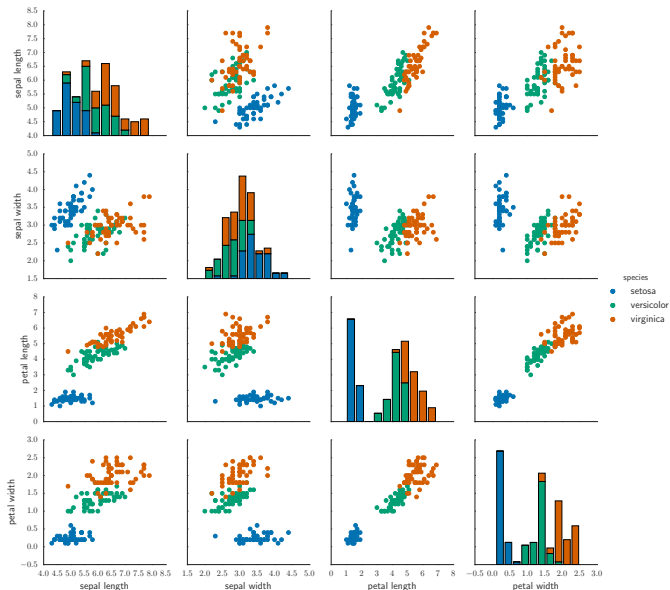


# Exemples de corrélations proches de zéro

Corrélation = 0.011

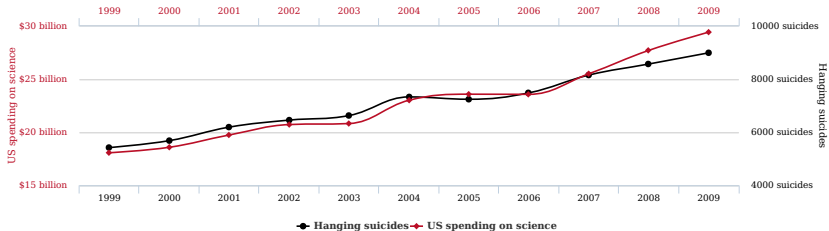


# Nuages de points / Scatter plot / PairGrid



# Covariance $\neq$ causalité

## US spending on science, space, and technology correlates with Suicides by hanging, strangulation and suffocation



tylervigen.com

Corrélation : 0.9979

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

# Sommaire

## Bonus

### Introduction générale

Modèle statistique

Biais/Varianace

### Statistiques descriptives

Résumés basiques d'un jeu de données

Corrélations/Nuage de points

### 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 v.a. 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.

$\forall j \in \llbracket 1, p \rrbracket, \mathbb{E}(X_j^2) < +\infty$  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  $\text{Cov}(X)$ , de taille  $p \times p$ , formée par les  $\sigma_{i,j}$  ( $i^{\text{e}}$  ligne,  $j^{\text{e}}$  colonne). Ainsi,

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

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

---

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

---

## Quelques propriétés de la covariance

- ▶ Une matrice de covariance est symétrique :

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

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

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

Démonstration :

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

---

**Exo:**  $\text{Cov}(AX) = A \text{Cov}(X) A^\top$ , 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 :

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

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

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

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

Vocabulaire :

- les  $\lambda_i \in \mathbb{R}$  sont les **valeurs propres** de  $S$
- les  $\mathbf{u}_i \in \mathbb{R}^n$  sont les **vecteurs propres** de  $S$



## 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

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

Rem: en général  $U$  n'a aucune raison d'avoir une tête sympathique

## 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)

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

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

# Sommaire

## Bonus

### Introduction générale

Modèle statistique

Biais/Variance

### Statistiques descriptives

Résumés basiques d'un jeu de données

Corrélations/Nuage de points

### 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

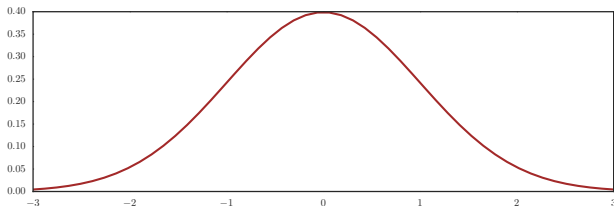
$$\varphi_{0,1}(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

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

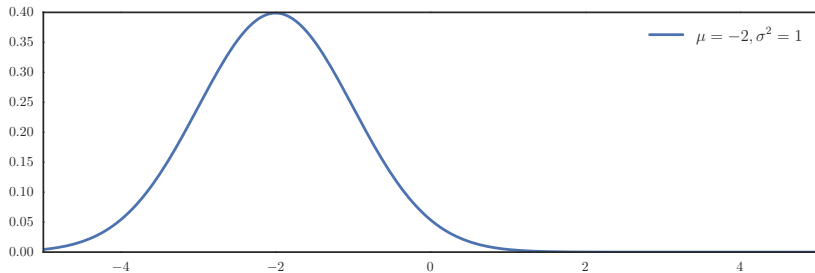
- ▶ Une v.a.  $Y$  suit une loi normale de paramètres  $\mu$  et  $\sigma^2$  si

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

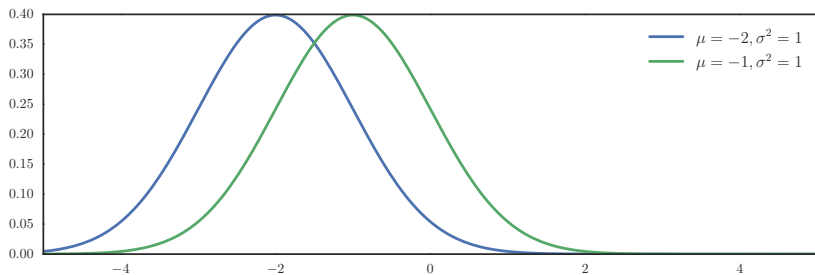
$$\text{Densité : } \varphi_{\mu, \sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



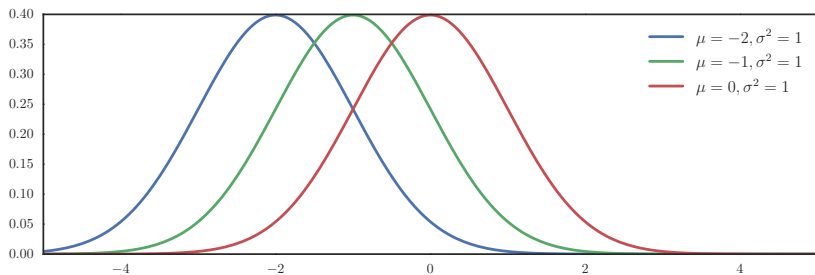
## Exemple : variation sur $\mu$



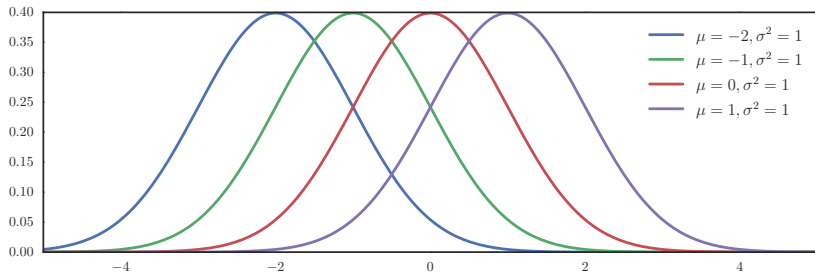
## Exemple : variation sur $\mu$



## Exemple : variation sur $\mu$

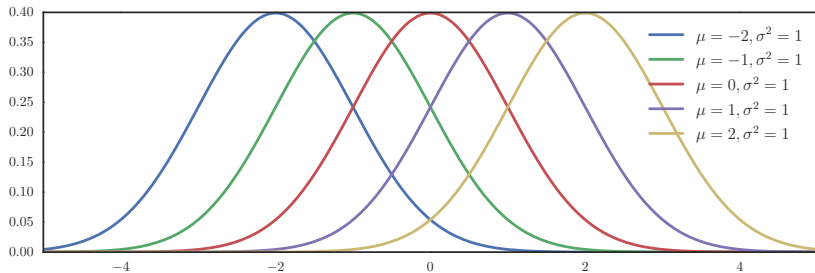


## Exemple : variation sur $\mu$

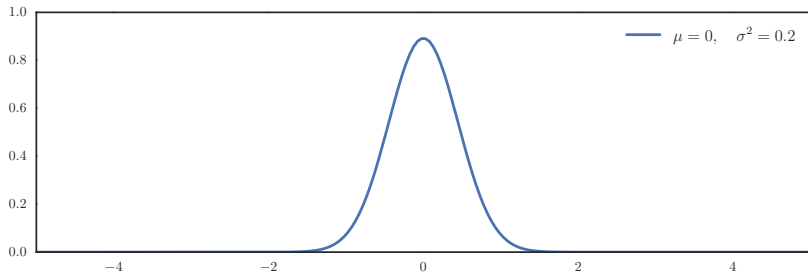




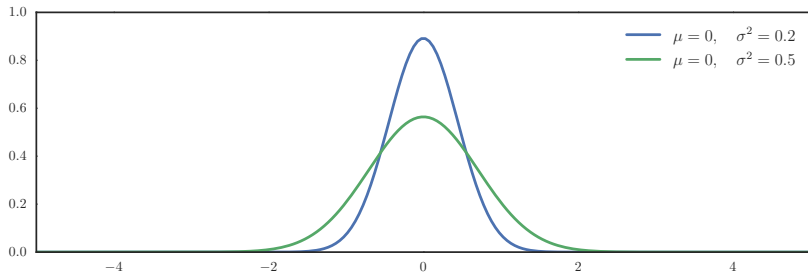
## Exemple : variation sur $\mu$



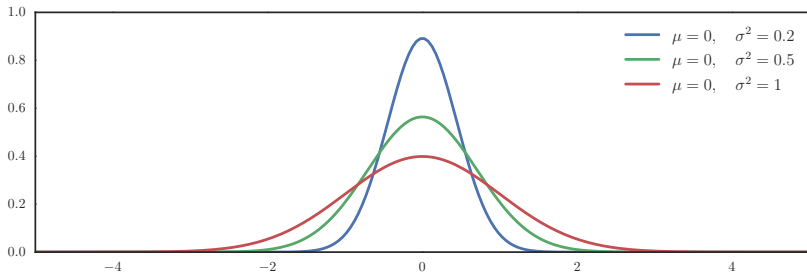
## Exemple : variation sur $\sigma$



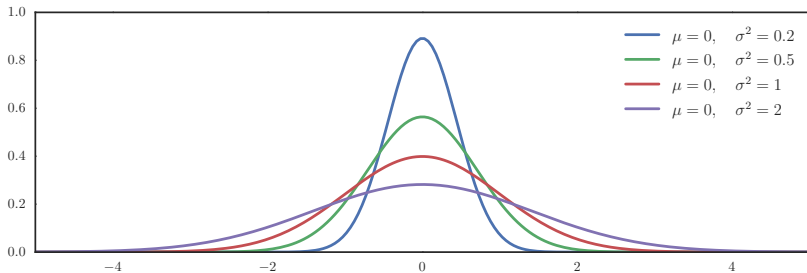
## Exemple : variation sur $\sigma$



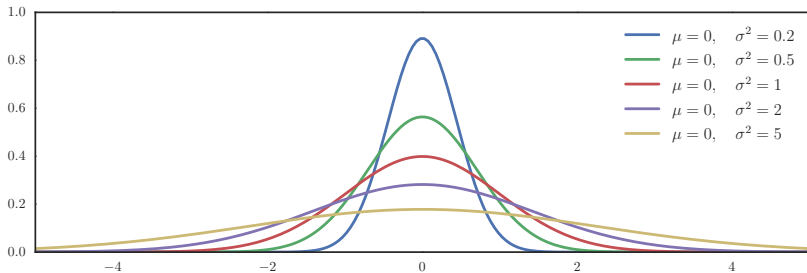
## Exemple : variation sur $\sigma$



## 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 :

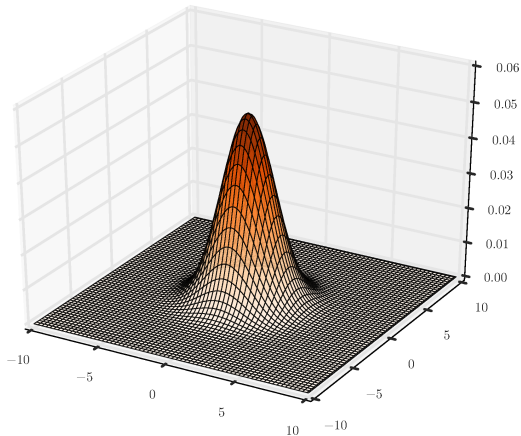
- le vecteur d'espérance  $\mu \in \mathbb{R}^p$
- la matrice de covariance  $\Sigma \in \mathbb{R}^{p \times p}$

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  $\det(\Sigma) = 0$

## Example 2D

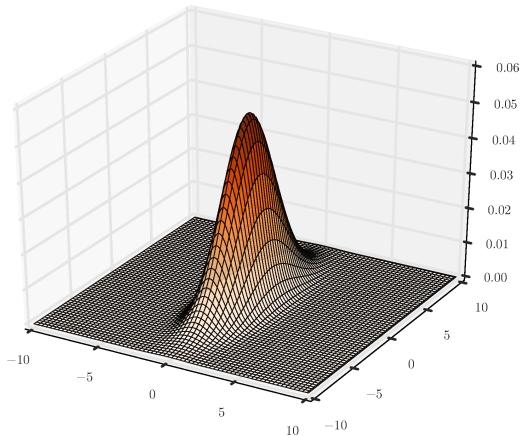
$$\Sigma = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} 3 & \\ & 3 \end{bmatrix} \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix},$$





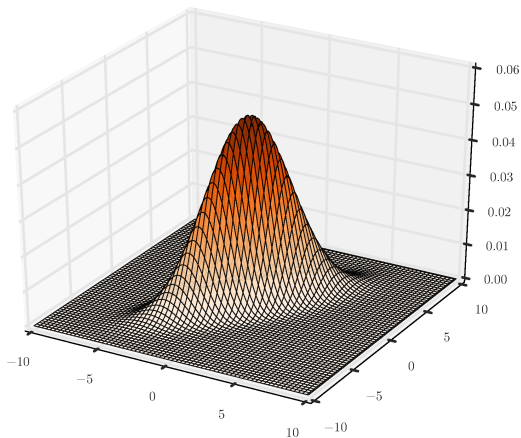
## Example 2D

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



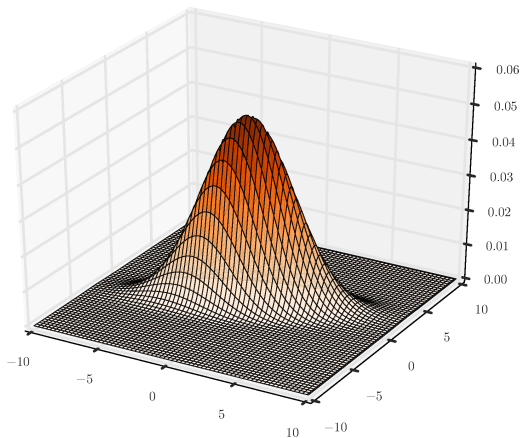
## Example 2D

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



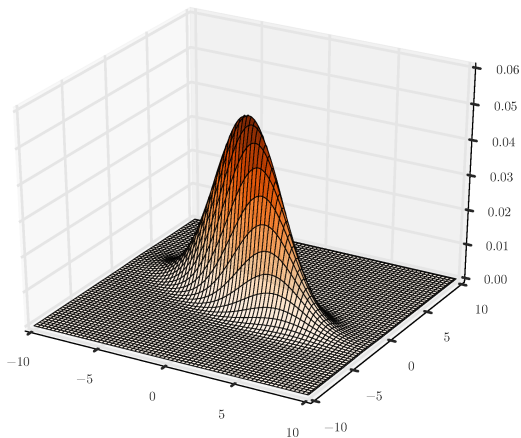
## Example 2D

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



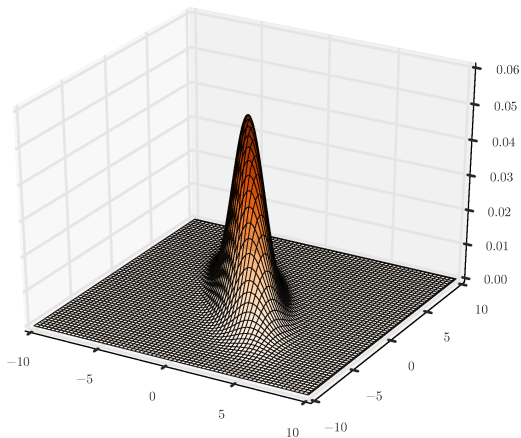
## Example 2D

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



## Example 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  $\mu \in \mathbb{R}^p$ ,  $Y = \mu + LX \sim \mathcal{N}(\mu, \Sigma)$

Calcul :  $\text{Cov}(Y) = \text{Cov}(LX) = L \text{Cov}(X) L^\top = L \text{Id}_p L^\top = \Sigma$   
Trouver  $L$  : peut être obtenu par la factorisation de Cholesky

# Factorisation de Cholesky

## Théorème

Toute matrice symétrique définie positive  $\Sigma \in \mathbb{R}^p$  peut s'écrire :  $\Sigma = LL^\top$  pour une matrice  $L$  triangle inférieure ( $L$  pour “lower”)

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

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

Numériquement on pourra utiliser : `numpy.linalg.cholesky`

# Bibliographie

- ▶ Cours d'algèbre linéaire numérique :  
<http://www.seas.ucla.edu/~vandenbe/103/>
- ▶ Tsybakov (2006) cours de "Statistique appliquée"
- ▶ Hastie *et al.* (2009) : Elements of Statistical Learning
- ▶ McKinney (2012), Raschka (2015), Müller et Guido (2016)  
pour les statistiques/apprentissage avec Python
- ▶ Blog de Jake Vanderplas : <http://jakevdp.github.io/>
- ▶ Livre de Gianluca Bontempi : statistiques pour l'apprentissage automatique  
[http://www.ulb.ac.be/di/map/gbonte/mod\\_stoch/syl.pdf](http://www.ulb.ac.be/di/map/gbonte/mod_stoch/syl.pdf)
- ▶ 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)



# Références I

- ▶ D. P. Bertsekas.  
*Nonlinear programming.*  
Athena Scientific, 1999.
- ▶ S. Boyd and L. Vandenberghe.  
*Convex optimization.*  
Cambridge University Press, Cambridge, 2004.
- ▶ D. Foata and A. Fuchs.  
*Calcul des probabilités : cours et exercices corrigés.*  
Masson, 1996.
- ▶ G. H. Golub and C. F. van Loan.  
*Matrix computations.*  
Johns Hopkins University Press, Baltimore, MD, fourth edition,  
2013.

## Références II

- ▶ R. A. Horn and C. R. Johnson.  
*Topics in matrix analysis.*  
Cambridge University Press, Cambridge, 1994.  
Corrected reprint of the 1991 original.
- ▶ T. Hastie, R. Tibshirani, and J. Friedman.  
*The elements of statistical learning.*  
Springer Series in Statistics. Springer, New York, second edition, 2009.  
<http://www-stat.stanford.edu/~tibs/ElemStatLearn/>.
- ▶ W. McKinney.  
*Python for Data Analysis : Data Wrangling with Pandas, NumPy, and IPython.*  
O'Reilly Media, 2012.

## Références III

- ▶ A. C. Müller and S. Guido.

*Introduction to Machine Learning with Python : A Guide for Data Scientists.*

O'Reilly Media, early access edition, 2016.

- ▶ S. Raschka.

*Python Machine Learning : Unlock deeper insights into Machine Learning with this vital guide to cutting-edge predictive analytics.*

Packt Publishing, 2015.

- ▶ B. W. Silverman.

*Density estimation for statistics and data analysis.*

Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1986.

## Références IV

- A. B. Tsybakov.  
Statistique appliquée, 2006.  
[http://josephsalmon.eu/enseignement/ENSAE/  
StatAppli\\_tsybakov.pdf](http://josephsalmon.eu/enseignement/ENSAE/StatAppli_tsybakov.pdf).