# PROJET - TRAITEMENT DES DONNÉES EN SHS

licence MIASHS - S3

Mathieu Dehouck - Thibault Liétard
mathieu.dehouck@inria.fr - thibault.lietard@univ-lille3.fr

18 septembre 2017

équipe MAGNET, INRIA Lille



# PRÉSENTATION DU PROJET

# Présentation du projet

## Traitement de données...

- nettoyage et analyse des données
- extraction d'information

## en SHS

- étude de données spécifiques
- mise en évidence de phénomènes
- conclusion

## URL du cours

`http://lietard.fr/tdshs`

## Cadre

- 2 intervenants
- 10 séances
- 2 langages (Python et LaTeX)
- en groupes de 2

## Notation

- 1 compte-rendu intermédiaire (1 à 2 pages)
- 1 oral à la fin des 10 séances (10 minutes + 5 minutes de questions)
- 1 rapport écrit (4 à 8 pages)

# STATISTIQUES DESCRIPTIVES

## Variables aléatoires

- quantitatives (discrètes ou continues) : âge, taille,...
- qualitatives : couleur des yeux,...

## Échantillon

Un échantillon est un ensemble de réalisations d'une variable aléatoire.
$X = [x_1, x_2, x_3, ..., x_N]$

# DESCRIPTEURS 1

## Moyenne

$\bar{X} = \frac{1}{N} \sum_X x_i$

## Médiane

Valeur centrale de X : la moitié des $x_i$ y est inférieure, l'autre moitié supérieure

## Mode

Valeur la plus souvent observée pour une variable discrète

### Variance

mesure la dispersion d'un échantillon : $V = \frac{1}{N} \sum_X (x_i - \bar{X})^2$

### Écart type

$\sigma_X = \sqrt{(V)}$

### Covariance

mesure la variation simultanée de 2 variables (nulle si indépendance) : $Cov(X, Y) = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{X})(y_i - \bar{Y})$

### Quantiles

Le quantile d'ordre $\alpha$ est la valeur de l'échantillon telle que $\alpha\%$ des valeurs de l'échantillon y sont inférieures

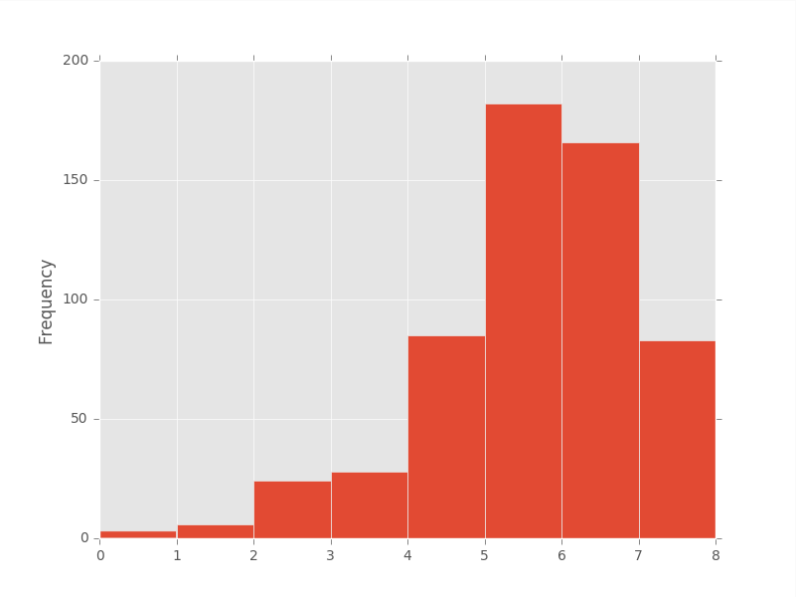
### Coefficient de corrélation

Mesure la corrélation entre deux variables.
coefficient de corrélation de Pearson : $\frac{Cov(X,Y)}{\sigma_X.\sigma_Y}$

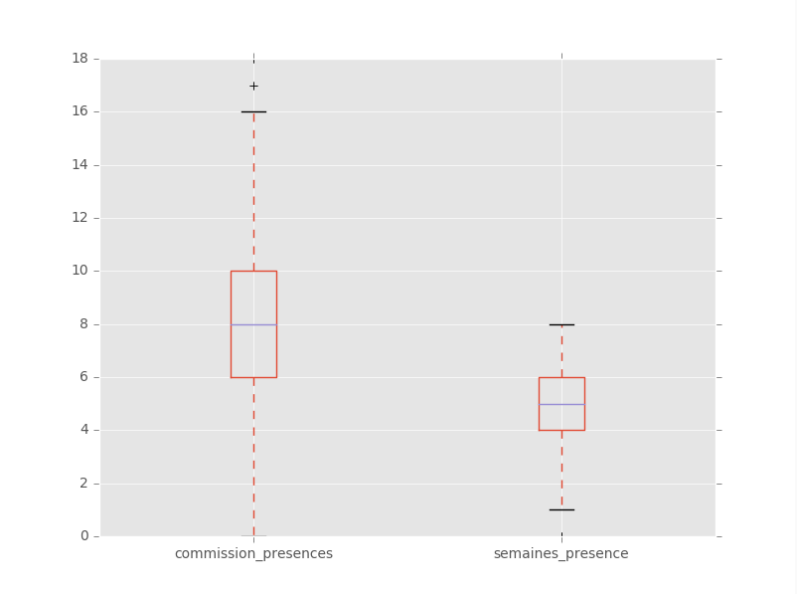
# DÉCRIRE UNE VARIABLE QUANTITATIVE

- Moyenne, médiane, mode (si discrète)
- Variance, étendue ($X_{max} - X_{min}$)
- intervalle inter-quartile : intervalle entre les quantiles d'ordre 0.25 et 0.75
- graphiques :

Histogramme



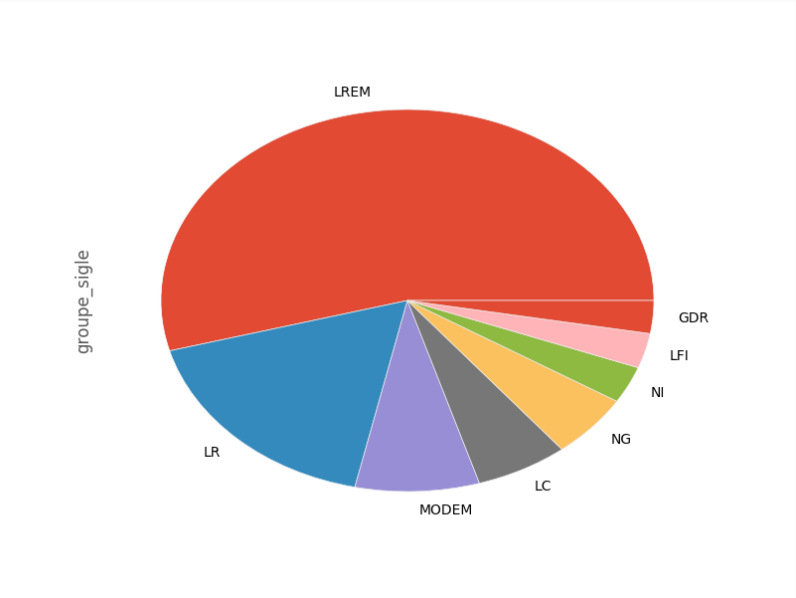

Boîte à moustaches



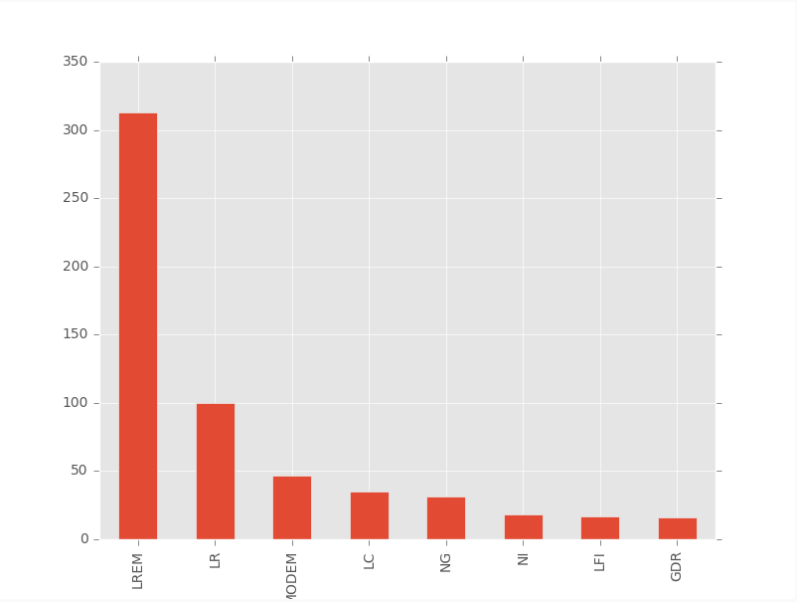
# DÉCRIRE UNE VARIABLE QUALITATIVE

- nombre de réalisations des modalités
- fréquence des modalités
- graphiques :

## Camembert




## Barres



# BIEN DÉCRIRE UNE VARIABLE

## Manipulation des variables continues

résumé statistique d'une variable
histogramme
boîte à moustaches

## Manipulation des variables discrète

modalités d'une variable discrète
nombre de réalisations des modalités
camemberts
diagramme en barres

# PYTHON - INTRODUCTION

## Utilisations (sous Linux)

- Console
- Fichiers .py

## Depuis le terminal

Ouvrir python : *python*
Exécuter un fichier .py : *python ./monfichier.py*

## Ressources

`https://openclassrooms.com/courses/apprenez-a-programmer-en-python`
`http://pandas.pydata.org/pandas-docs/stable/tutorials.html`

## Importer les librairies

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import sys
```

## Calculs simples

```
2 + 3
np.sqrt(4)
x = 8
x %% 3
```

## Types de données

- int : 1, 2, -20,...
- float : 1.2, 3.14159, -3.36
- String : ”bonjour”, ”12”, ”c”
- char : ’c’, ’0’

## Listes

```
list1 = [1, 2, 3]
list2 = [4, 5, 6]
list3 = [tab1, tab2]
list1[0] + list3[2][1]
```

## Tableaux

```
tab1 = np.array([1,2,3])
tab2 = np.array(list2)
mat1 = np.array([tab1, tab2])
tab1[0] + mat1[2][1]
```

## Fonction

```
def ma_fonction(param1, param2, ...) :
    instr 1
    instr 2
    ...
```

## Boucles for

```
for i in range(0,12) :
    intst1
    instr2
```

## Boucles while

```
while x > 2 :
    intst1
    instr2
```

Test if

```
if x > 2 :
    print >> sys.stdout, "x plus grand que 2"
else :
    print >> sys.stdout, "x plus petit que 2"
```

## création

```
df = pd.DataFrame({
    'noms' : ['Marc', 'Jean', 'Luc', 'Matthieu'],
    'taille' : [157, 175, 182, 168]
})
```

## chargement

```
df = pd.read_table("./deputes.data", sep=';', index_col=0)
```

## manipulation des variables

| | |
|---|---|
| lister les variables | `df.columns` |
| afficher une variable (quali.) | `df['nom']` |
| afficher une variable (quanti.) | `df['semaines_presence']` |
| afficher deux variables | `df[['nom', 'semaines_presence']]` |

## Ajouter/supprimer des variables

| | |
|---|---|
| ajouter | `df['mandats1000'] = 1000 * df['nb_mandats']` |
| supprimer | `del df('mandats1000')` |

compter les modalités d'une variable discrète

```
df['groupe_sigle'].value_counts()
```

Créer un sous dataframe

```
df_majorite = df[df.groupe_sigle == 'LREM']
```

# FONCTIONS STATISTIQUES

| | |
|---|---|
| max | `df['semaines_presence'].max()` |
| min | `df['semaines_presence'].min()` |
| moyenne | `df['semaines_presence'].mean()` |
| médiane | `df['semaines_presence'].median()` |
| mode | `df['groupe_sigle'].mode()` |
| quantile | `df['semaines_presence'].quantile(0.2)` |
| variance | `df['semaines_presence'].var()` |
| écart type | `df['semaines_presence'].std()` |
| covariance | `df['semaines_presence'].cov(df['nb_mandats'])` |

## Histogramme

```
df['semaines_presence'].plot.hist(bins=[0,1,2,3,4,5,6,7,8],
    range=(1,12))
plt.show()
```

## Boxplot

```
df[['commission_presences', 'semaines_presence']].plot.box()
plt.show()
```

## Barplot

```
df['groupe_sigle'].value_counts().plot.bar(); plt.show()
```

## Piechart

```
df['groupe_sigle'].value_counts().plot.pie(); plt.show()
```