

Cours 2 – Décrire, comparer, visualiser des données OWID

Moyenne, médiane, dispersion et premiers pas avec pandas

Etienne Dagorn

Université de Lille – L2 Économie

2026–2027

Objectif de la séance

But

Passer de petits objets Python à une première analyse descriptive : lire des données, calculer quelques indicateurs, visualiser et interpréter avec prudence.

- ▶ Python reste l'objectif principal.
- ▶ Les statistiques servent de langage pour résumer les données.
- ▶ Les exemples suivent la logique des TD OWID.

Où va le cours ?

Séance	Python	Statistique support
TD1	variables, listes	moyenne, min, max, pourcentage
TD2	boucles, conditions	proportions, seuils
TD3	fonctions	comparaison de groupes
TD4	graphiques, simulation	distribution, stabilisation d'une moyenne
TD5	pandas	mini-projet descriptif OWID

Fil rouge

Une bonne analyse décrit une distribution, pas seulement un chiffre isolé.

Donnée, observation, variable

Mot	Sens simple	Exemple OWID
Observation	une ligne du tableau	France en 2021
Variable	une colonne du tableau	PIB par habitant
Valeur	une case du tableau	46000
Unité	sens de la valeur	dollars par habitant

Réflexe

Avant de calculer : quelle est l'observation ? quelle est la variable ? quelle est l'unité ?

Même information, trois formes

```
valeurs = [46, 54, 34]
```

```
gdp = {  
    "France": 46,  
    "Germany": 54,  
    "Spain": 34  
}
```

```
donnees = [  
    {"pays": "France", "gdp":  
     46},  
    {"pays": "Germany", "gdp":  
     54},  
    {"pays": "Spain", "gdp":  
     34}  
]
```

Idée

Plus les données deviennent riches, plus une table devient pratique. C'est le rôle d'un DataFrame pandas.

Langage, bibliothèque, environnement

Terme	Rôle	Exemple
Langage	règles pour écrire des programmes	Python
Interpréteur	programme qui lit le langage	python3
Bibliothèque	outils ajoutés au langage	pandas, matplotlib
Environnement	ensemble installé sur la machine	Python + packages

Erreur fréquente

Si `import pandas` échoue, Python fonctionne peut-être très bien : c'est la bibliothèque ou l'environnement qui manque.

Avant de calculer avec un tableau

Rituel de lecture

Avec un vrai fichier de données, on ne calcule pas avant d'avoir regardé la structure minimale du tableau.

Question	Commande pandas	Ce que l'on vérifie
Quelles lignes ?	<code>df.head()</code>	aperçu des premières observations
Quelle taille ?	<code>df.shape</code>	nombre de lignes et colonnes
Quels noms ?	<code>df.columns</code>	noms exacts à utiliser dans le code

Pseudo-code : décrire une liste

```
# Objectif : resumer une liste de valeurs économiques
# Entree : une liste de nombres
# 1. Compter les observations
# 2. Calculer la moyenne
# 3. Trier pour trouver la mediane
# 4. Reperer minimum et maximum
# 5. Observer la dispersion
# Sortie : chiffres + interpretation courte
```

Programmation

Le pseudo-code dit quoi faire. Python impose ensuite une syntaxe précise pour le faire.

Moyenne et médiane : ce qu'elles disent

Indicateur	Question	Limite
Moyenne	niveau général	sensible aux valeurs extrêmes
Médiane	valeur centrale	ignore l'ampleur des écarts
Minimum	plus petite valeur	ne résume pas le groupe
Maximum	plus grande valeur	peut être atypique

À retenir

Moyenne et médiane résument le centre. Elles ne suffisent pas à décrire la forme de la distribution.

Même moyenne, même médiane, distributions différentes

```
groupe_a = [8, 9, 10, 11, 12]  
groupe_b = [0, 5, 10, 15, 20]
```

Groupe	Moyenne	Médiane	Lecture
A	10	10	valeurs concentrées
B	10	10	valeurs très dispersées

Question

Si ces valeurs sont des revenus, les deux groupes racontent-ils la même situation économique ?

Distribution : regarder toutes les valeurs

- ▶ Une distribution décrit la façon dont les valeurs se répartissent.
- ▶ On peut regarder le centre : moyenne, médiane.
- ▶ On peut regarder les extrêmes : min, max, étendue.
- ▶ On peut regarder la dispersion : écarts à la moyenne, variance, écart-type.
- ▶ On peut visualiser : histogramme, barres, points.

Idée économique

Deux pays ou deux groupes peuvent avoir le même niveau moyen mais des inégalités ou dispersions très différentes.

Écarts à la moyenne

Pour `groupe_b = [0, 5, 10, 15, 20]`, la moyenne vaut 10.

```
ecart = valeur - moyenne
```

Valeur	0	5	10	15	20
Écart	-10	-5	0	5	10

Problème

Si on additionne les écarts, les valeurs négatives et positives se compensent. Il faut donc mesurer la taille des écarts autrement.

Variance et écart-type : intuition

Variance

La variance est la moyenne des carrés des écarts à la moyenne.

$$\text{variance} = \frac{1}{n} \sum (x_i - \bar{x})^2$$

- ▶ On met les écarts au carré pour éviter les compensations.
- ▶ Les grands écarts comptent davantage.
- ▶ L'écart-type est la racine carrée de la variance.
- ▶ L'écart-type revient dans l'unité de départ, donc il est plus lisible.

Pseudo-code : variance sans panique

```
# Objectif : mesurer la dispersion d'une liste
# Entree : liste de valeurs
# 1. Calculer la moyenne
# 2. Pour chaque valeur :
#     calculer l'ecart a la moyenne
#     mettre cet ecart au carre
# 3. Calculer la moyenne des ecarts au carre
# 4. Prendre la racine pour obtenir l'ecart-type
# Sortie : variance et ecart-type
```

Position dans le cours

À connaître intuitivement. En obligatoire, on insiste surtout sur moyenne, médiane, min, max et graphiques.

Livecoding 1 : résumer deux groupes

Départ en direct :

```
groupe_a = [8, 9, 10, 11, 12]
groupe_b = [0, 5, 10, 15, 20]

def moyenne(valeurs):
    return sum(valeurs) / len(valeurs)

print(moyenne(groupe_a))
print(moyenne(groupe_b))
```

À coder ensemble : min, max, étendue, puis une phrase d'interprétation.

Livecoding 2 : médiane et dispersion

Objectif : transformer une idée statistique en fonction Python.

```
def mediane_impair(valeurs):  
    valeurs_triees = sorted(valeurs)  
    milieu = len(valeurs_triees) // 2  
    return valeurs_triees[milieu]  
  
def etendue(valeurs):  
    return max(valeurs) - min(valeurs)
```

Question en direct : quel groupe a la plus grande étendue ? Est-ce visible dans la moyenne ?

Livecoding 3 : variance pas à pas

Version volontairement explicite, pour comprendre la logique.

```
def variance(valeurs):  
    m = moyenne(valeurs)  
    carres = []  
    for x in valeurs:  
        ecart = x - m  
        carres.append(ecart ** 2)  
    return moyenne(carres)
```

Extension en direct si le groupe suit :

```
import math  
print(math.sqrt(variance(groupe_b)))
```

Graphiques : voir ce que les chiffres cachent

Graphique	Sert à voir	TD
Barres	comparer des catégories	TD4, TD5
Histogramme	forme d'une distribution	TD4
Courbe	évolution dans le temps	extension TD5
Nuage de points	relation entre deux variables	plus tard

Règle

Un graphique doit répondre à une question précise. Il n'est pas une décoration.

Livecoding 4 : visualiser une distribution

```
import matplotlib.pyplot as plt

valeurs = [46, 54, 34, 38, 22, 61]
pays = ["France", "Germany", "Spain", "Italy", "Poland", "Sweden"]

plt.bar(pays, valeurs)
plt.xticks(rotation=45, ha="right")
plt.ylabel("PIB par habitant, milliers de dollars")
plt.tight_layout()
plt.show()
```

À demander : quel pays tire la moyenne vers le haut ?

Lire un fichier Our World in Data

Colonne OWID	Sens	Exemple
Entity	pays ou région	France
Code	code pays	FRA
Year	année	2022
Dernière colonne	indicateur	PIB par habitant

Astuce

Les fichiers OWID ont souvent une structure régulière : identifier la colonne de valeur est la première étape.

Pseudo-code : pipeline pandas OWID

```
# Objectif : deccrire un indicateur OWID
# 1. Lire le CSV depuis une URL
# 2. Afficher les premieres lignes
# 3. Renommer Entity, Code, Year, valeur
# 4. Filtrer une annee recente
# 5. Selectionner quelques pays
# 6. Calculer moyenne, mediane, min, max
# 7. Tracer un graphique
# 8. Rediger une interpretation prudente
```

Livecoding 5 : lire un CSV OWID

```
import pandas as pd

url = "https://ourworldindata.org/grapher/gdp-per-capita-worldbank.csv"
df = pd.read_csv(url)

print(df.head())
print(df.shape)
print(df.columns)
```

Question : quelle colonne contient l'indicateur économique ?

Débogage pandas : trois cas courants

Message probable	Hypothèse utile
ModuleNotFoundError	la bibliothèque n'est pas installée dans l'environnement actif
FileNotFoundError	le chemin ou le dossier courant n'est pas le bon
KeyError	le nom de colonne utilisé n'existe pas exactement

Réflexe

Avant d'utiliser une colonne, afficher `df.columns` et recopier le nom exact.

Livecoding 6 : nettoyer et filtrer

```
value_col = df.columns[-1]
df = df.rename(columns={
    "Entity": "country",
    "Code": "code",
    "Year": "year",
    value_col: "gdp_pc"
})

latest_year = df["year"].max()
latest = df.loc[df["year"] == latest_year].dropna(subset=["gdp_pc"])
```

À demander : pourquoi retirer les valeurs manquantes avant de calculer ?

Livecoding 7 : calculer et comparer

```
pays = ["France", "Germany", "Spain", "Poland", "United States", "India"]
selection = latest.loc[latest["country"].isin(pays)].copy()

print(selection[["country", "gdp_pc"]])
print(selection["gdp_pc"].mean())
print(selection["gdp_pc"].median())
print(selection["gdp_pc"].min())
print(selection["gdp_pc"].max())
```

Interprétation : la France est-elle au-dessus ou au-dessous de la moyenne de cette sélection ?

Décrire sans surinterpréter

Phrase acceptable

Dans cette sélection de pays et pour cette année, la France est au-dessus ou au-dessous de la moyenne observée du PIB par habitant.

Phrase trop forte

La France est plus riche parce que sa politique économique est meilleure.

- ▶ On décrit des données.
- ▶ On ne prouve pas une cause.
- ▶ On précise toujours le périmètre : pays, année, indicateur, source.

Ce que les TD rendent obligatoire

TD	Obligatoire
TD3	fonctions, moyenne, médiane, min, max, comparaison
TD4	graphique, simulation simple, distribution
TD5	CSV OWID, pandas, graphique, interprétation

Priorité pédagogique

Écrire du Python clair avant d'empiler les notions statistiques.

Résumé

- ▶ Moyenne et médiane résument le centre d'une distribution.
- ▶ Elles ne capturent pas toute la dispersion.
- ▶ Variance et écart-type mesurent les écarts à la moyenne, mais restent secondaires dans ce cours.
- ▶ Les graphiques aident à voir ce qu'un chiffre unique cache.
- ▶ pandas permet de passer d'une petite liste à un vrai tableau OWID.
- ▶ Toute interprétation doit rester descriptive et prudente.