TP 2 : Statistique univariée et régression linéaire

Dans ce TP, nous utiliserons les bibliothèques suivantes :

```
import numpy as np
import scipy.stats as st
import matplotlib.pyplot as plt
import pandas as pd
```

1 Statistiques unidimensionnelles

Nous allons étudier différents jeux de données avec la librairie `pandas`. Avec `pandas`, une série statistique a pour format `pd.Series`. Dans le cas où l'on a plusieurs données pour chaque individu, les données peuvent être rangées dans un tableau de données `pd.DataFrame`, chaque colonne du tableau étant une `pd.Series`.

`https://pandas.pydata.org/docs/getting_started/intro_tutorials/01_table_oriented.html`

Exercice 1

Un atelier réalise le séchage de boues d'origine industrielle. Il obtient à la fin du processus des déchets (mesurés en kg). On a observé les poids suivants de déchets après le traitement de 100 kg :

$$4,7 \quad 4,3 \quad 4,5 \quad 4,9 \quad 4,2 \quad 4,7 \quad 4,0 \quad 4,2 \quad 5,0 \quad 3,9 \quad 4,6 \quad 4,6$$

$$4,8 \quad 4,4 \quad 4,2 \quad 4,6 \quad 4,3 \quad 4,9 \quad 4,0 \quad 4,5 \quad 4,1 \quad 4,4 \quad 4,3 \quad 4,3$$

Notons x cette série numérique.

1. Définir une `pd.Series` contenant ces données.
2. Créer un tableau de données `pd.DataFrame` contenant les effectifs, les fréquences et les fréquences cumulées. On pourra en particulier utiliser les fonctions `value_counts` et `sort_index`.
3. Tracer le diagramme cumulatif, en utilisant `pandas` : `https://pandas.pydata.org/docs/getting_started/intro_tutorials/04_plotting.html`
4. Rappeler la définition de la moyenne, la variance et l'écart-type. Les calculer pour notre exemple. On pourra pour cela utiliser les fonctions intégrées de `pandas` : `https://pandas.pydata.org/docs/getting_started/intro_tutorials/06_calculate_statistics.html`
5. Définir la médiane et les quartiles, les déterminer.
6. Tracer le boxplot et un histogramme.
7. Supposons que la 9^e valeur soit 50 et non $5,0$ (à cause d'une erreur dans la saisie de la donnée). Que devient alors le boxplot ? Et le résumé numérique ? On pourra mettre les deux séries (les données sans erreur et celles avec erreur) dans un `pd.DataFrame` pour les comparer.

Exercice 2

On lance 100 fois un dé et on note les résultats :

face	1	2	3	4	5	6
nombre d'occurrences	13	16	18	16	13	24

1. Calculer le résumé numérique.
2. Représenter graphiquement ces données de 3 manières différentes.

Exercice 3

Une entreprise d'import-export en textile reçoit ce mois-ci des quantités importantes de chemises provenant d'une usine au Pakistan. Chaque carton contient 100 chemises. Le contrôleur-qualité de l'entreprise ouvre 50 cartons et vérifie les chemises. Il note le nombre de chemises présentant des défauts dans chaque carton pour évaluer la qualité générale de l'arrivage. Voici ses relevés :

2 2 2 2 3 3 3 2 2 1 3 5 4 1 2 6 5 1 4 3 4 4 7 4 3 0 3 3 4 1 6 3 4 3 3 5 6 3 2 0 0 4 6 4 0 3 2 2 5 3

Réalisez une étude statistique descriptive de cet échantillon.

2 Régression linéaire

En faisant juste un copier-coller, récupérer les données disponibles à l'adresse suivante : `https://math.univ-lyon1.fr/~gerber/teaching/2025_Probas_MAT2072L/data.python.txt`.

Exercice 4

Pour chaque jeu de données (X_i, Y_i),

1. tracer le nuage de points grâce à la commande : `plt.plot(X,Y,'o')`.
2. deviner la valeur du coefficient de corrélation ;
3. calculer la valeur du coefficient de corrélation ;
4. ajouter la droite des moindres carrés sur le nuage de points.

Indication : tous les calculs se font sous python grâce à la commande `stats.linregress(X,Y)`.