TP9

STATISTIQUES

But du TP

- Manipuler un tableur (excel ou libreoffice) et calculer des indicateurs statistiques.
- Ouvrir un tableur avec python et la bibliothèque pandas.
- Utiliser une bibliothèque nouvelle à l'aide d'une liste de commandes.
- Comparer des séries statistiques.

Dans ce TP, nous allons étudier les données de deux tableurs :

- « tous-les-pays-du-monde-INED.xls » : estimations 2021 de la population modiale (par continent) et des données de natalités/mortalité.
  Source : Institut national d'études démographiques - www.ined.fr
  World Population Prospects. Nations Unies. 2019
- « Deces-France.xls » : nombre de décès par semaine, par sexe et par tranche d'âge en France de 2000 à 2021.
  Source : Human Mortality Database, Short-Term Mortality Fluctuation data series (STMF), INSERM

I Manipuler un tableur simple

Exercice 1

1. Ouvrir avec **Excel** ou LibreOffice Calc le tableur « tous-les-pays-du-monde-INED.xls ».
2. Remplir les cases vides en utilisant des formules (voir ci-dessous)
   (a) Population totale ;
   (b) Population moyenne par continent ;
   (c) Taux de natalité moyen ;
   (d) Taux de mortalité moyen ;
   (e) Espérance de vie moyenne ;
   (f) Population 65+ totale puis moyenne.

`MOYENNE(...)`	Moyenne
`SOMME(...)`	Somme
`SOMMEPROD(Liste1 ; Liste2)`	Renvoie la somme des produits des éléments des deux listes.

II Python – Bibliothèque pandas

Exercice 2 Pour la suite, nous allons passer par **Jupyter Notebook**. Téléverser les fichiers suivants :

- Python-ECG1-09-statistiques.ipynb
- Deces-France.xls

puis suivez les instructions du notebook.

Nous allons utiliser plusieurs nouvelles bibliothèques dans cet exercices :

```
import pandas as pd             # pandas : gestion de données
import matplotlib.pyplot as plt # matplotlib : graphiques
```

Voici une liste de commandes utiles pour ce TP. Le tableau des données importées sera ici noté `df` (dataframe) et `Colonne` est le *nom* d'une colonne de ce tableau.
Attention : adapter le df en fonction du nom donné à votre tableau au moment de l'import.

`df = pd.read_csv(fichier)`	import d'un fichier csv.
`df = pd.read_excel(fichier)`	import d'un fichier excel
`df`	aperçu du tableau
`df.head(), df.head(n)`	5 premières lignes, n premières lignes
`df.tail() , df.tail(n)`	5 dernières lignes, n dernières lignes
`df.shape`	taille du tableau (lignes, colonnes)
`df.info()`	affiche les colonnes et leur type
`df[Colonne]`	données de la colonne `Colonne`. Attention, si le nom de la colonne est une chaîne de caractères, on écrira son nom entre guillemets.
`df[[Colonne1,Colonne2,...]]`	sélection des colonnes indiquées.
`df[ df[Colonne] == 5 ]`	applique un filtre, ici la valeur de `Colonne` doit être égale à 5.
`df[ (filtre1) & (filtre2) ]`	applique plusieurs filtres (booléens), le `&` signifie 'et'
`df.sort_values(Colonne)`	trie la Colonne par ordre croissant.
`df.describe()`	statistiques de base
`df.mean()`	moyenne
`df.std()`	écart-type
`df.count()`	nombre de valeurs
`df.median()`	médiane
`df.max()`	maximum
`df.min()`	minimum
`df.sum()`	somme des données
`plt.bar(abscisses, ordonnees)`	diagramme en barres
`plt.boxplot(df[Colonne])`	boîte à moustaches
`plt.hist(df[Colonne],range=(a,b),bins=n)`	histogramme de Colonne, sur l'intervalle $[a, b]$ découpé en n sous-intervalles
`df.hist(column = Colonne1, by = Colonne2)`	histogrammes de Colonne1 pour chaque valeur de Colonne2.

Pour la suite de ce TP, ouvrez le Jupyter Notebook du TP9 et suivez les instructions.