

I - Rappels : statistiques descriptives univariées

On considère une population Ω formée de n individus sur laquelle on étudie un caractère quantitatif (données réelles) $X = \{x_1, \dots, x_n\}$. On parle alors de la série statistique $(x_i)_{i \in \llbracket 1; n \rrbracket}$.

Définition 1

- Paramètres de position
 - On appelle minimum (resp. maximum) de la série statistique $(x_i)_{i \in \llbracket 1; n \rrbracket}$, notée $\bar{x}$ sa plus petite (resp. plus grande) valeur.
 - On appelle moyenne empirique de la série statistique $(x_i)_{i \in \llbracket 1; n \rrbracket}$, notée $\bar{x}$, la quantité : $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$.
 - On appelle médiane de la série statistique $(x_i)_{i \in \llbracket 1; n \rrbracket}$ toute valeur m qui la partage en deux de sorte que 50% des données sont inférieures ou égales à m et 50% sont supérieures ou égales à m .
- Paramètres de dispersion
 - On appelle variance empirique de la série statistique $(x_i)_{i \in \llbracket 1; n \rrbracket}$, notée σ_x^2 , la quantité : $\sigma_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$.
 - On appelle écart-type empirique de la série statistique $(x_i)_{i \in \llbracket 1; n \rrbracket}$, notée σ_x , la quantité : $\sigma_x = \sqrt{\sigma_x^2}$.

Exercice 1

- Écrire une fonction moyemp(X) qui prend en paramètre une série statistique (sous forme de liste) et qui renvoie sa moyenne empirique (sans utiliser la fonction np.mean).
 - Écrire une fonction varemp(X) qui prend en paramètre une série statistique et qui calcule et renvoie sa variance empirique (en utilisant la fonction moyemp(X) mais sans utiliser la fonction np.mean).
 - Écrire une fonction med(X) qui prend en paramètre une série statistique rangée par ordre croissant et qui calcule et renvoie sa médiane. On distinguera selon la parité de la taille de la série.
La commande Python n%2 renvoie 0 si n est pair et 1 si n est impair.

Python

Si X est une liste numpy (ou une série pandas) :

- La commande "X.min()" (resp. "X.max()") (ou "np.min(X)" (resp. "np.max(X)")) renvoie le minimum (resp. maximum) de la série statistique contenue dans X .
- La commande "X.mean()" (ou "np.mean(X)") renvoie la moyenne empirique de la série statistique contenue dans X .
- La commande "X.median()" (ou "np.median(X)") renvoie la médiane de la série statistique contenue dans X .
- La commande "X.var()" (ou "np.var(X)") renvoie la variance de la série statistique contenue dans X .
- La commande "X.std()" (ou "np.std(X)") renvoie l'écart-type de la série statistique contenue dans X .

II - La bibliothèque pandas de Python

Dans ce TP, on utilise la bibliothèque pandas de Python qui permet la lecture de fichiers .csv (Comma Separated Values) et la manipulation de tables de données.

Python

- La commande "import pandas as pd" permet de charger la bibliothèque.
- La commande "Donnees = pd.read_csv('C:\Dossier\Fichier.csv', sep = ',') " permet de récupérer les données du fichier "Fichier.csv" qui sont alors stockées dans la variable "Donnees".
- On travaille désormais avec la variable "Donnees".
 - La commande "Donnees.head()" permet de visualiser les premières lignes du fichier.
 - La commande "Donnees.columns" permet d'afficher la liste des colonnes.
 - La commande "Donnees['Nom de colonne']" permet d'accéder aux données d'une colonne en renvoyant une liste de ces données.

Exercice 2

Le fichier "Notes.csv" contient des données sur les notes aux concours d'une classe du sud ouest de la France. On exécute les instructions suivantes.

```
1 >>> import pandas as pd # On charge la bibliothèque pandas
2 >>> Notes = pd.read_csv ("Notes.csv", sep = ',')
3 >>> Notes.columns
4 ['Etudiant', 'DCG EDHEC/ESSEC', 'DCG EML/HEC', 'ESH ESCP',
5 'ESH ESSEC/HEC', 'LVA ELVi', 'LVB ELVi', 'Maths EML', 'Maths Edhec',
6 'Maths ESSEC II', 'Maths HEC', 'Synthèse ESCP']
```

- Pour une épreuve de votre choix, écrire des instructions permettant d'afficher la meilleure note ainsi que la moyenne et l'écart-type de cette épreuve.
- Écrire une fonction Python d'entête "def NbrNotes(M,eps,Matiere) :" qui renvoie le nombre de notes contenues dans la liste "Matiere" et qui sont entre "M-eps" et "M+eps".
- Compléter le script ci-dessous pour qu'il affiche un graphe représentant les notes obtenues à cette épreuve. On utilisera en abscisses une liste d'entiers numérotant les élèves.

```
1 import matplotlib.pyplot as plt
2 Matiere = .....
3 plt.plot(range(.....), ..... )
4 plt.show()
```

III - Statistiques descriptives bivariées

On considère une population Ω formée de n individus. On souhaite étudier deux caractères quantitatifs d'un même individu dans cette population et éventuellement conclure à une relation entre ces deux caractères.

Définition 2

On considère deux caractères quantitatifs $X = \{x_1, \dots, x_n\}$ et $Y = \{y_1, \dots, y_n\}$ d'une population Ω .

- On appelle série statistique double la donnée de n couples $(x_i, y_i)_{i \in \llbracket 1; n \rrbracket}$, correspondant aux valeurs prises respectivement par les caractères X et Y d'un même individu.
- On appelle nuage de points de la série statistique double $(x_i, y_i)_{i \in \llbracket 1; n \rrbracket}$ l'ensemble des points de coordonnées (x_i, y_i) .
- On appelle point moyen du nuage de points $(x_i, y_i)_{i \in \llbracket 1; n \rrbracket}$ le point $(\bar{x}, \bar{y})$ où $\bar{x}$ et $\bar{y}$ représentent les moyennes empiriques de $(x_i)_{i \in \llbracket 1; n \rrbracket}$ et de $(y_i)_{i \in \llbracket 1; n \rrbracket}$.
- On appelle covariance empirique de la série statistique double $(x_i, y_i)_{i \in \llbracket 1; n \rrbracket}$, notée $\text{cov}(x, y)$, la quantité :

$$\text{cov}(x, y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

- On appelle coefficient de corrélation linéaire empirique de la série statistique double $(x_i, y_i)_{i \in \llbracket 1; n \rrbracket}$, noté $\rho_{x,y}$, la quantité :

$$\rho_{x,y} = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y}$$

- Le coefficient de corrélation linéaire est un réel compris entre -1 et 1 : $-1 \leq \rho_{x,y} \leq 1$.
- Il est égal à 1 (resp. -1) si l'une des variables est fonction affine croissante (resp. décroissante) de l'autre variable.
- Plus le coefficient est proche des valeurs extrêmes -1 et 1 , plus la corrélation entre les variables est forte.

Exercice 3

- Démontrer la formule suivante : $\text{cov}(x, y) = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y}$.
 - En déduire une fonction Python d'entête "def Covariance(x,y) :" qui calcule la covariance empirique de la série double (x, y) stockée dans deux tableaux numpy donnés en paramètre. On pourra utiliser la fonction mean.
 - Utiliser la fonction précédente pour calculer la variance empirique de la série statistique $(x_i)_{i \in \llbracket 1; n \rrbracket}$.
 - Écrire une fonction Python d'entête "def CoeffCor(x,y) :" qui calcule le coefficient de corrélation linéaire empirique de la série double (x, y) stockée dans deux tableaux numpy donnés en paramètre.

Exercice 4

On considère la série statistique double contenant les notes aux épreuves de Maths EDHEC et de Maths EML des élèves de notre classe du sud ouest de la France.
Compléter le programme suivant pour qu'il trace le nuage de points ainsi que le point moyen associés à cette série double.

```
1 Notes = pd.read_csv ("Notes.csv",sep = ',')
2 X = .....
3 Y = .....
4 plt.plot(....., ..... )
5 plt.plot(....., ..... )
6 plt.show()
```

IV - Droite de régression linéaire

Lorsque le nuage de points associé à une série statistique double présente une forme allongée, on peut chercher à « approcher » ce nuage de point par une droite, appelée droite de régression linéaire.

Théorème 1

- La droite de régression de y en x , est la droite d'équation :

$$y = ax + b \quad \text{avec} \quad a = \frac{\text{cov}(x,y)}{\sigma_x^2} \quad \text{et} \quad b = \bar{y} - a \cdot \bar{x}$$

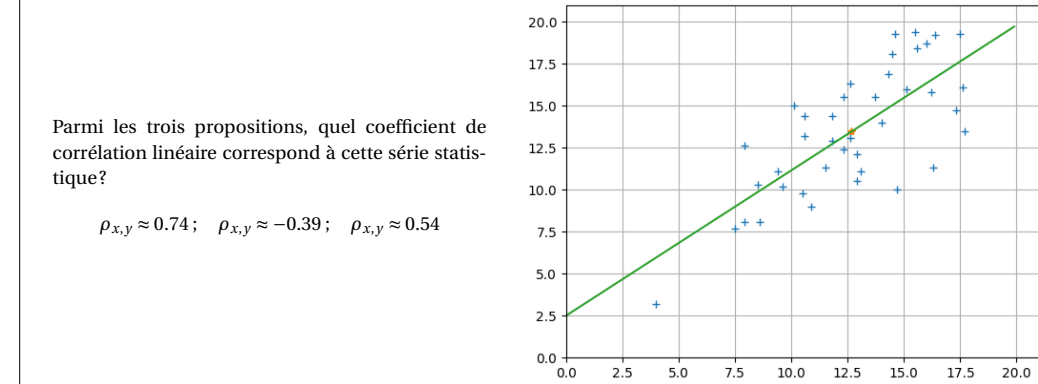
- Cette droite de régression de y en x passe par le point moyen $(\bar{x}, \bar{y})$.
- On dit que x est la variable explicative et que y est la variable à expliquer.

Exercice 5

1. On reprend le programme de l'exercice 4 et on ajoute les lignes ci-dessous.
Compléter ce programme pour qu'il détermine les coefficients a et b de la droite de régression linéaire de Y en X puis expliquer ce qu'il fait.

```
1 cov = Covariance(X,Y) # Fonction écrite plus haut
2 a = .....
3 b = .....
4 abs = np.arange(0, 20 ,0.1)
5 plt.plot(abs ,a*abs+b )
6 plt.show()
7 print(CoeffCorr(X,Y)) # Fonction écrite plus haut
```

2. On obtient le graphique suivant.

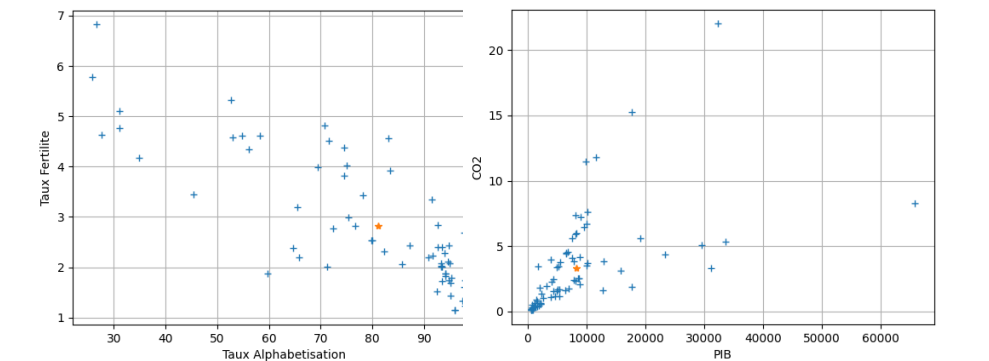


Exercice 6

Le fichier "Pays.csv" contient les données suivantes sur certains pays du monde : Émissions de CO2, taux de fertilité des femmes, PIB, taux d'alphabétisation des femmes.
On exécute les instructions suivantes.

```
1 >>> import pandas as pd # On charge la bibliothèque pandas
2 >>> Pays = pd.read_csv ("Pays.csv",sep = ',')
3 >>> Pays.columns
4 ['Pays', 'CO2', 'Fertilite', 'PIB', 'Alphabetisation']
```

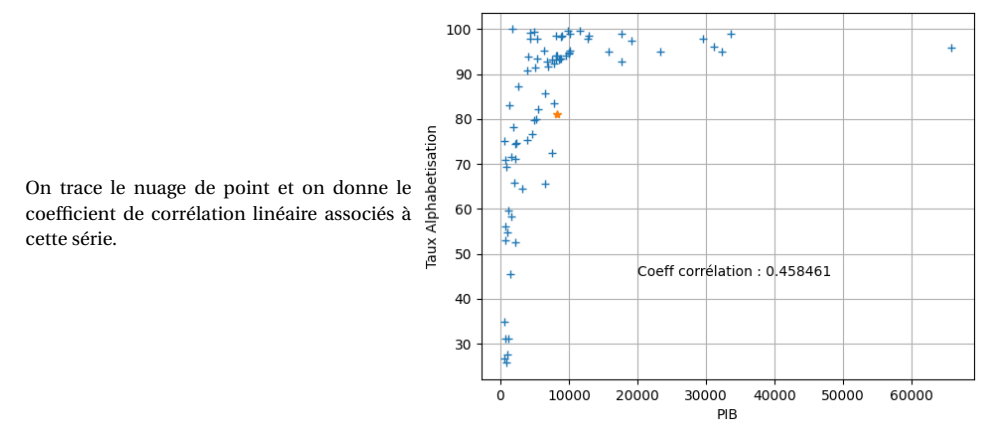
1. On trace les nuages de points suivants.



Associer à chaque graphique le coefficient de corrélation linéaire correspond et commenter.

$\rho_{x,y} \approx 0.56; \quad \rho_{x,y} \approx -0.79; \quad \rho_{x,y} \approx 0.85 \quad \rho_{x,y} \approx -0.25$

2. On considère à présent la série statistique double (PIB, Taux d'alphabétisation).



- (a) Une approximation par régression linéaire est-elle pertinente pour modéliser le taux d'alphabétiser par rapport au PIB?
(b) On considère les instructions suivantes.

```
1 >>> X, Y = Pays['PIB'], Pays['Alphabetisation']
2 >>> Z = np.log(X)
3 >>> print(CoeffCorr(Z,Y))
4 0.7786212347927588
```

Une approximation par régression linéaire entre Z et Y est-elle plus pertinente?
Quelle relation peut-on supposer exister entre X et Y ?