# TP6 Python (statistiques)

En statistiques, on utilise le module `numpy` pour calculer les indicateurs et le module `matplotlib.pyplot` pour les graphiques.

Soient $x = (x_1, ..., x_n)$ et $y = (y_1, ..., y_n)$ des séries statistiques quantitatives, écrites en Python sous forme de listes ou de tableaux numpy (matrices).

Formules à connaître au concours :

- $\overline{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$ = moyenne de $x$ `np.mean(x)`
- $s_x^2 = \dfrac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2$ = variance de $x$ `np.var(x)`
- $s_x = \sqrt{s_x^2}$ = écart-type [1] de $x$ `np.std(x)`
- $s_{xy} = \dfrac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$ = covariance de $(x, y)$
- $\rho_{xy} = \dfrac{s_{xy}}{s_x s_y}$ = coefficient de corrélation linéaire de $(x, y)$.

Propriétés à connaître au concours :

- $-1 \leq \rho_{xy} \leq 1$.
- $\rho_{xy} \approx 1$ ou $\rho_{xy} \approx -1 \iff y$ est une fonction presque affine de $x$.
- $s_x^2 = \dfrac{1}{n}\sum_{i=1}^{n} x_i^2 - (\overline{x})^2$ (formule de Koënig).

Propriété hors programme à savoir démontrer :

$s_{xy} = \dfrac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x}\,\overline{y}$ (formule de Huygens)

démonstration :

$$\begin{aligned} s_{xy} &= \frac{1}{n}\sum_{i=1}^{n} (x_i y_i - \overline{x} y_i - x_i \overline{y} + \overline{x}\,\overline{y}) \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x}\frac{1}{n}\sum_{i=1}^{n} y_i - \overline{y}\frac{1}{n}\sum_{i=1}^{n} x_i + \frac{1}{n}\overline{x}\,\overline{y}\sum_{i=1}^{n} 1 \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x}\,\overline{y} - \overline{x}\,\overline{y} + \overline{x}\,\overline{y} \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x}\,\overline{y}. \end{aligned}$$

1. en anglais = standard deviation

Exercice 1

Soient $x$ et $y$ les listes données par $x = [1, 2, 3]$ et $y = [3, 5, 10]$.

1)a)En revenant à la définition, calculer $\overline{x}$ et $\overline{y}$.

b)Montrer que $s_x^2 = \frac{2}{3}$ et $s_y^2 = \frac{26}{3}$, puis vérifier l'exactitude des résultats trouvés en utilisant les commandes `mean` et `var`.

2)a)En revenant à la définition, montrer que $s_{xy} = \frac{7}{3}$.

b)Compléter la fonction `covariance` ci-dessous prenant comme arguments des listes $x$ et $y$ de taille $n$ et renvoyant $s_{xy}$.

```
def covariance(x,y,n):
    liste=[]
    mx=np.mean(x)
    my=np.mean(y)
    for i in range(......):
        liste.append((x[i]-mx)*(y[i]-my))
    return(.................)
```

c)A l'aide de cette fonction, vérifier l'exactitude du résultat trouvé en 2)a).

3)Calculer $\rho_{xy}$ et vérifier que $-1 \leq \rho_{xy} \leq 1$.

Exercice 2

1)Construire une série statistique de 3 valeurs dont la médiane fait 10 et la moyenne fait 100.

2)Idem avec 4 valeurs.

3)En France, en 2019, le salaire moyen était de 2424 euros, alors que le salaire médian était de 1940 euros. Comment l'expliquer ?

Exercice 3

On rappelle que la commande rd.randint(a,b) du module numpy.random renvoie un entier aléatoire entre a et b-1.

1)Ecrire un programme Python effectuant les tâches suivantes :

- création et affichage d'une liste de 100 entiers aléatoires entre 1 et 1000,
- affichage de la moyenne, médiane et écart-type de la liste,
- affichage de la boîte à moustache.

2)Quelles informations donne la boîte à moustache ?

Exercice 4

Une entreprise de 360 employés est constituée de cadres qui gagnent 4000 euros par mois et d'ouvriers qui gagnent 2400 euros par mois.
Le salaire moyen d'un employé est de 2800 euros par mois.

1)Combien l'entreprise possède t-elle de cadres et d'ouvriers ?

2)L'entreprise décide d'augmenter le salaire des ouvriers de 5%.
Quel est alors le salaire moyen d'un employé ?

Exercice 5

L'allométrie (du grec allos=autres) étudie comment les caractéristiques d'un individu (taille du cerveau, poids, ...) changent avec sa taille.

Le tableau ci-dessous donne la taille $x$ en centimètres et le poids $y$ en grammes de 15 bars prelevés lors d'une pêche.

| $x$ | 55 | 74 | 80 | 91 | 30 | 88 | 17 | 96 | 83 | 30 | 21 | 41 | 61 | 25 | 98 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $y$ | 1750 | 3200 | 4000 | 7500 | 250 | 6640 | 50 | 9360 | 5000 | 270 | 90 | 720 | 2200 | 150 | 10500 |

1)Ecrire un programme qui crèe et affiche $x$ et $y$ sous forme de tableaux numpy.

2)a)A l'aide de la commande `mean`, préciser la taille moyenne et le poids moyen des poissons pêchés.

2)b)Déterminer la taille médiane et le poids médian des poissons pêchés, puis vérifier le résultat trouvé à l'aide de la commande `median`.

3)On étudie maintenant s'il existe une corrélation entre $x$ et $y$.

a)Visualiser le nuage de points à l'aide de la commande `plt.scatter(x,y)`. Existe t-il une corrélation affine entre $y$ et $x$ ?

b)L'opération * entre deux tableaux Numpy de même format permet de multiplier ces tableaux coefficient par coefficient.
A l'aide de la formule de Huygens, compléter la fonction ci-dessous prenant en argument deux matrices lignes numpy x et y de même taille et renvoyant la covariance de (x,y).

```
import numpy as np
def covariance(x,y):
    prod=x*y
    mx=np.mean(x)
    my=np.mean(y)
    return ......................
```

c)Ecrire un programme qui affiche le coefficient de corrélation linéaire de la série statistique $(\ln x, \ln y)$ ainsi que son nuage de points.
Vérifier la bonne corrélation affine entre $\ln y$ et $\ln x$.

d)En déduire qu'il existe des constantes réelles $\alpha$ et $\beta$ telles que $y \approx \beta x^{\alpha}$.

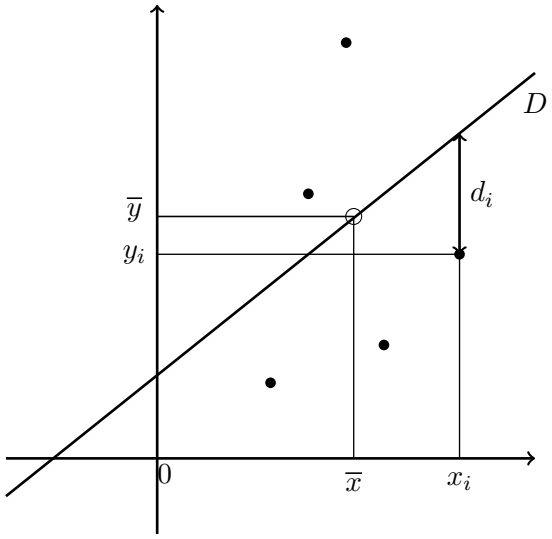
> Remarque
> $\alpha$ s'appelle le coefficient d'allométrie. Chez le bar, $\alpha \approx 3$ et $\beta \approx 1/100$.

Exercice 6 (méthode des moindres carrés)

Soit $(x_1, y_1)$, ..., $(x_n, y_n)$ un nuage de points.

Le point $(\overline{x}, \overline{y})$ est appelé *point moyen* du nuage.

On cherche à construire une droite $D$ passant par le point moyen et telle que la somme des distances au carré des points du nuage à $D$ soit minimale, c'est-à-dire telle que $\sum_{i=1}^{n} d_i^2$ soit minimale.




1)Montrer que l'équation de la droite $D$ est de la forme : $y = \overline{y} + a(x - \overline{x})$ .

2)En déduire que $\forall i \in [\![1, n]\!],\ di^2 = \left(y_i - \overline{y} - a(x_i - \overline{x})\right)^2$.

3)Soit $f : a \mapsto \sum_{i=1}^{n} d_i^2$.

a)Montrer que $\forall a \in \mathbf{R},\ f'(a) = 2nas_x^2 - 2ns_{xy}$.

b)En déduire que $f$ admet un minimum en $\dfrac{s_{xy}}{s_x^2}$.

c)En déduire l'équation [2] de $D$.

2. D est appelée droite de regression linéaire de $y$ en $x$