TP 6 – Statistique bivariée

1 Séries statistiques

On appelle *variable statistique* une caractéristique qui peut prendre différentes valeurs (par exemple la taille), et *série statistique* une liste finie de mesures d'une variable statistique donnée, dans une population identifiée.

Si $x = (x_1, \dots, x_n)$ est une série statistique associée à la variable statistique X, on appelle

- *moyenne* de la série :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

- *variance* de la série :

$$V(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

On a par ailleurs la formule de Koenig-Huygens :

$$V(x) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2.$$

Les commandes Python pour les calculer sont :

- `np.mean(x)` pour la moyenne,
- `np.var(x)` pour la variance.

On considère $y = (y_1, \dots, y_n)$ la série statistique associée à une autre variable statistique Y, pour la même population que x (c'est-à-dire que x_i et y_i sont les mesures respectives des caractères X et Y pour le i-ème individu de la population étudiée).

On dit alors que (x, y) est une série statistique double, ou bivariée. On appelle alors *covariance* de la série bivariée (x, y) le réel

$$\begin{aligned}\mathrm{Cov}(x,y) &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}.\end{aligned}$$

Cette dernière écriture porte encore le nom de formule de Koenig-Huygens.

Exemple. 10 enfants de 6 ans sont mesurés et pesés. On note X la variable désignant la taille de l'enfant (en centimètres) et Y celle désignant le poids de l'enfant (en kilogrammes). On obtient la série statistique double suivante :

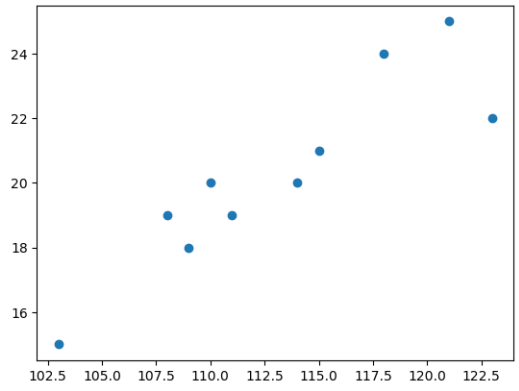
	1	2	3	4	5	6	7	8	9	10
x	121	123	108	118	111	109	114	103	110	115
y	25	22	19	24	19	18	20	15	20	21

Exercice 1. Ecrire une fonction `cov(x,y)` qui prend en entrée une série bivariée (x, y), et calcule la covariance de x et y.

2 Représentation graphique

On représente une série statistique double au moyen d'un nuage de points. On utilisera la syntaxe suivante pour afficher le nuage de points.

```
import matplotlib.pyplot as plt
plt.scatter(x,y)
plt.show()
```



Exercice 2. Représenter le nuage de points associé à l'exemple ci-dessus.

3 Régression linéaire

On cherche la droite d'équation $y = ax + b$ qui passe "au plus près" du nuage de points au sens des moindres carrés, c'est-à-dire qu'on cherche des réels a et b tels que la quantité

$$\sum_{i=1}^{n}(y_i - (ax_i + b))^2$$

soit minimale. On appelle cette droite la droite de régression linéaire associée à la série double.

Le problème revient à chercher $U = \binom{a}{b} \in \mathcal{M}_{2,1}(\mathbb{R})$ solution du problème de minimisation

$$\min_{U\in\mathcal{M}_{2,1}(\mathbb{R})} \|AU - Y\|^2,$$

où $A = \begin{pmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix}$, et $Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$.

Il y a une unique solution U à ce problème, caractérisée par

$${}^tAAU \;=\; {}^tAY.$$

Comme A est de rang 2, si les x_i ne sont pas tous identiques, la matrice ${}^tAA$ est inversible dans ce cas, et on a alors

$$U \;=\; ({}^tAA)^{-1}\,{}^tAY,$$

ce qui fournit les valeurs de a et b :

$$a = \frac{\mathrm{Cov}(x,y)}{V(x)}, \quad \text{et} \quad b = \bar{y} - a\bar{x},$$

ce qui exprime que la droite de régression passe par le point $(\bar{x}, \bar{y})$ dit point moyen. Plus précisément, l'équation de la droite est donnée par

$$y \;=\; \frac{\mathrm{Cov}(x,y)}{V(x)}\,(x - \bar{x}) + \bar{y}.$$

Exercice 3.

1. Déterminer les valeurs de a et b pour la série statistique double de l'exemple.
2. Représenter la droite de régression linéaire sur le nuage de points.
3. Représenter le point moyen $(\bar{x}, \bar{y})$.
4. Estimer le poids d'un enfant de 6 ans qui mesure 1,20m.

Exercice 4.

1. Créer deux tableaux `x` et `y` de taille 1000 contenant des simulations de la loi $\mathcal{U}([0,1])$.
2. Représenter le nuage des points (x_i, y_i), et la droite de régression linéaire.
3. Calculer le coefficient de corrélation.
4. Créer un tableau `u` de taille 1000 contenant des simulations de la loi $\mathcal{U}([-1,1])$. Créer ensuite le tableau `v` donné par `v = u**2`.
5. Représenter le nuage de points (u_i, v_i), et la droite de régression linéaire. Que constate-t-on ?
6. Calculer le coefficient de corrélation. Que constate-t-on ?