

Chapitre 8

Statistiques bivariées.

Sommaire

8.1 Statistique descriptive bivariée	158
8.1.1 Définition	158
8.1.2 Rappels	158
8.1.3 Covariance et corrélation	159
8.1.4 Représentation graphique - Nuage de points	160
8.2 Ajustement affine par la méthode des moindres carrés	161
8.2.1 Méthode des moindres carrés	161
8.2.2 Coefficient de corrélation	167
8.2.3 Remarque fondamentale : corrélérer n'est pas expliquer	170

Ce chapitre est à mettre en étroite relation avec le TP3 d'informatique : Statistiques descriptives bivariées.

On étudiera des séries statistiques à deux variables, pour cela on tracera le nuage de points associé à la série étudiée. On parlera de point moyen du nuage. On calculera la covariance et le coefficient de corrélation linéaire empirique d'une série statistique double.

Si le nuage de points associé à une série statistique double possède une forme étirée, on peut avoir l'idée de chercher quelle droite approcherait au mieux les points de ce nuage. Le problème consiste donc à identifier une droite $y = ax + b$ qui ajuste bien le nuage de points, ce qui conduira à un ajustement des moindres carrés et aux droites de régression.

Compétences attendues. ⇨ Représenter un nuage de points associé à une série statistique double.

⇨ Représenter la droite des moindres carrés.

⇨ Calculer le coefficient de corrélation linéaire et interpréter sa valeur.

Aussi loin que l'on remonte dans le temps et dans l'espace (en Chine et en Égypte, par exemple), les États ont toujours senti le besoin de disposer d'informations sur leurs sujets ou sur les biens qu'ils possèdent et produisent. Mais les recensements de population et de ressources, les statistiques (du latin status : état) sont restées purement descriptives jusqu'au 17ème siècle.

Puis s'est développé le calcul des probabilités et des méthodes statistiques sont apparues en Allemagne, en Angleterre et en France. Beaucoup de scientifiques de tout ordre ont apporté leur contribution au développement de cette science :

Pascal, Gauss, Laplace, Quetelet, De Moivre, Galton, Tchebychev, Okun ...

La loi d'Okun, en économie, a été proposée par Arthur Okun en 1962. Elle décrit une relation linéaire empirique entre le taux de croissance (du PIB) et la variation du taux de chômage. En dessous d'un certain seuil de croissance, le chômage augmente; au-dessus de ce seuil, il diminue, à élasticité constante. La loi d'Okun constitue « un lien de toute première importance entre les marchés des produits et celui du travail » et permet aux économistes de prévoir les effets d'un taux de croissance donné sur le chômage.

Cette loi s'est révélée être une des relations les plus durables de la macroéconomie moderne. Elle séduit également par sa durabilité et sa simplicité, deux qualités rares dans cette discipline qu'est la science économique.

Toutefois, cette loi soulève quelques critiques : la variabilité du coefficient d'Okun en fonction du temps ou son inexistence dans certaines économies en sont des exemples.

Quelques citations :

Il y a trois sortes de mensonges : les mensonges, les gros mensonges et les statistiques (Mark Twain).
Je ne crois aux statistiques que lorsque je les ai moi-même falsifiées (Winston Churchill).

Dans tout le chapitre n désigne un entier naturel non nul.

8.1 Statistique descriptive bivariable

8.1.1 Définition

On s'intéresse à l'étude, sur une population donnée, du lien qui peut exister entre deux caractères.

Valeurs du 1 ^{er} caractère x_i	x_1	x_2	x_3	$\dots$	x_k
Valeurs du 2 ^{ème} caractère y_i	y_1	y_2	y_3	$\dots$	y_k

Définition 8.1 (Série statistique double).

Soient une population Ω et deux variables quantitatives discrètes X et Y relatives à cette population.

On note x la série statistique $(x_i)_{i \in [1, n]}$, des modalités de X observées sur un échantillon de taille n de la population $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$, on note y la série statistique $(y_i)_{i \in [1, n]}$, des modalités de Y observées sur le même échantillon. On appelle série statistique double la donnée de la liste : $[(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)]$, chaque couple $(x_i, y_i)_{i \in [1, n]}$ étant associé à un seul individu ω_i de la population.

Remarque. C'est le recueil simultané des modalités de deux variables X et Y chez les mêmes sujets. L'intérêt se porte le plus souvent sur la relation entre les deux variables : la recherche de corrélation.

Exemple 8.1. L'étude du coût de maintenance annuel d'une installation de chauffage dans un immeuble de bureaux, en fonction de l'âge de l'installation, a donné les résultats suivants :

Age x_i (années)	1	2	3	4	5	6
Coût y_i (k€)	7,55	9,24	10,74	12,84	15,66	18,45

Y'a-t-il un lien crédible entre l'âge de l'installation et le coût de maintenance? Si oui, peut-on le quantifier, et peut-on, par exemple, prévoir le coût de maintenance d'une installation de 7 ans? 8 ans? 10 ans?

8.1.2 Rappels

Propriétés 8.1 (Moyenne).

On rappelle que $\bar{X}$ désigne la moyenne de X , à savoir : $\bar{X} = \frac{1}{n} \sum_{i=1}^n x_i$.

$\bar{Y}$ désigne la moyenne de Y , à savoir : $\bar{Y} = \frac{1}{n} \sum_{i=1}^n y_i$.

Propriétés 8.2 (Variance et formule de Koenig-Huygen).

On rappelle que les variances de X et Y notées s_X^2 et s_Y^2 sont définies par :

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

$$s_Y^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{Y})^2 = \frac{1}{n} \sum_{i=1}^n y_i^2 - \bar{Y}^2.$$

8.1.3 Covariance et corrélation

Définition 8.2 (Covariance empirique).

On appelle covariance empirique de la série statistique double $(x_i, y_i)_{i \in [1, n]}$, le réel :

$$s_{X,Y} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y}).$$

Propriété 8.3 (formule de Koenig-Huygen). On a $s_{X,Y} = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{X}\bar{Y}$.

☛ Python : liste des commandes exigibles

C_1 : `np . mean ( x )` renvoie la moyenne arithmétique $\bar{x}$ de la série $(x_i)_{i \in [1, n]}$

C_2 : `np . var ( x )` renvoie la variance empirique de la série $(x_i)_{i \in [1, n]}$.

C_3 : `np . std ( x )` renvoie l'écart-type σ_x de la série $(x_i)_{i \in [1, n]}$

C_4 : `np . mean ( x*y ) - np . mean ( x ) * np . mean ( y )` renvoie la covariance $s_{x,y}$ de la série (x_i, y_i) .

☛ Liste des commandes non exigibles mais commode d'utilisation

C_2^* : `np . cov ( x , y , bias=True )` renvoie une matrice carrée d'ordre 2 symétrique, avec comme coefficients diagonaux les variances de x et y et les deux autres coefficients sont égaux à la covariance de x et y .

Définition 8.3 (Coefficient de corrélation linéaire).

Le coefficient de corrélation linéaire de la série statistique double $(x_i, y_i)_{i \in [1, n]}$ est le réel :

$$r_{X,Y} = \frac{s_{X,Y}}{s_X * s_Y}.$$

Remarques.

1. On note parfois $\sigma_X = \sqrt{s_X^2} = s_X$, où s_X^2 représente la variance de X .
On a $|s_{X,Y}| \leq s_X s_Y$ donc $|r_{X,Y}| \leq 1$.

2. Le coefficient de corrélation mesure la dépendance linéaire entre deux variables.
- S'il est proche de 1 ou -1, alors X et Y sont fortement corrélés.
 - S'il est proche de 0, alors X et Y sont faiblement corrélés (voire ne le sont pas).

☛ Liste des commandes exigibles

C_5 : `np . mean (x*y)-np . mean (x)*np . mean(y)/(np . std (x)*np . std(y))` renvoie le coefficient de corrélation $r_{x,y}$ de la série $(x_i, y_i)_{i \in \llbracket 1, n \rrbracket}$.

☛ Python : Liste des commandes non exigibles mais commode d'utilisation

C_1^* : `np . corrcoef (x, y)` renvoie la matrice $\begin{pmatrix} 1 & r_{x,y} \\ r_{x,y} & 1 \end{pmatrix}$.

C_2^* : `np . corrcoef (x, y)[0,1]` renvoie donc le coefficient de corrélation linéaire de la série (x_i, y_i) .

C_3^* : si l'on suppose déclarés p vecteurs $x_1, x_2, \dots, x_p$ contenant p séries statistiques notées $x_1, \dots, x_p$ alors `np . corrcoef ([x1, x2, ..., xp])` renvoie la matrice dont le coefficient de la i -ème ligne et la j -ième colonne est r_{x_i, x_j} .

EXERCICE 8.1.

Le tableau ci-dessous donne le nombre de retraités en France métropolitaine entre 1975 et 2005 :

Année	Rang de l'année X	Nombre de retraités Y (en millions)
1975	0	4,1
1980	1	5
1985	2	5,9
1990	3	7,4
1995	4	8,3
2000	5	9,7
2005	6	10,7

1. Calculer les moyennes $\bar{X}$ et $\bar{Y}$ ainsi que leurs variances.
2. Calculer le coefficient de corrélation linéaire de cette série.

• Point méthode

☛ On calcule $\frac{1}{n} \sum_{i=1}^n x_i y_i$ puis $\frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{X} \bar{Y}$.

SOLUTION.

Calcul des moyennes :

$$\bar{X} = \frac{0+1+2+3+4+5+6}{7} = 3 \text{ et } \bar{Y} = \frac{4,1+5+5,9+7,4+8,3+9,7+10,7}{7} = 7,3.$$

Calcul des variances et des écarts types :

$$s_X^2 = \frac{0+1+4+9+16+25+36}{7} - 3^2 = 4 \text{ donc } s_X = 2.$$

$$s_Y^2 = \frac{4,1^2+25+5,9^2+7,4^2+8,3^2+9,7^2+10,7^2}{7} - (7,3)^2 = 5,12 \text{ donc } s_Y = \sqrt{5,12} \approx 2,26.$$

Calcul de la covariance de la série :

$$s_{X,Y} = \frac{1 \times 5 + 2 \times 5,9 + 3 \times 7,4 + 4 \times 8,3 + 5 \times 9,7 + 6 \times 10,6}{7} - 3 \times 7,3 = 4,47$$

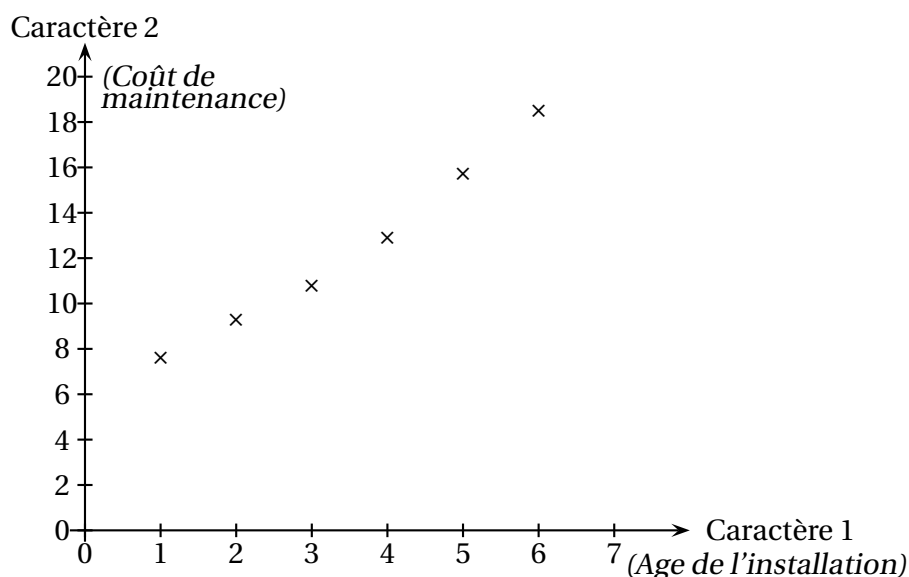
$$\text{Le coefficient de corrélation est : } r_{X,Y} = \frac{s_{X,Y}}{s_X s_Y} \approx \frac{4,47}{2 \times \sqrt{5,12}} \approx 0,989.$$

8.1.4 Représentation graphique - Nuage de points

Définition 8.4 (Nuage de points).

On appelle **nuage de points**, l'ensemble des points M_i de coordonnées (x_i, y_i) .

Exemple 8.2. en lien avec exemple 8.1



Définition 8.5 (Point moyen).

Le **point moyen** du nuage de points est le point de coordonnées $(\bar{X}; \bar{Y})$.

Exemple 8.3. Dans l'exemple 8.1, le point moyen G a pour coordonnées $(3,5; 12,41)$.

Python : liste des commandes non exigibles mais commode d'utilisation

On importera les librairies numpy et matplotlib.pyplot :

```
import numpy as np
```

```
import matplotlib.pyplot as plt
```

C_4^* : `plt.plot(x, y, '+')` permet de tracer le nuage de point de coordonnées (x_i, y_i) , les points étant matérialisés par des «+».

Si l'on veut des points on remplacera '+' par '.'

8.2 Ajustement affine par la méthode des moindres carrés

8.2.1 Méthode des moindres carrés

Les points de l'exemple précédents ne sont pas alignés. Néanmoins, ces points semblent se distribuer approximativement autour d'une droite.

Si le nuage de points associé à une série statistique double possède une forme étirée, on peut avoir l'idée de chercher quelle droite approcherait au mieux les points de ce nuage.

Le problème consiste donc à identifier une droite $y = ax + b$ qui ajuste bien le nuage de points.

L'erreur que l'on commet en utilisant la droite de régression pour prédire y_i à partir de x_i est $y_i - (ax_i + b)$.

Pour déterminer la valeur des coefficients a et b , on utilise le principe des moindres carrés qui

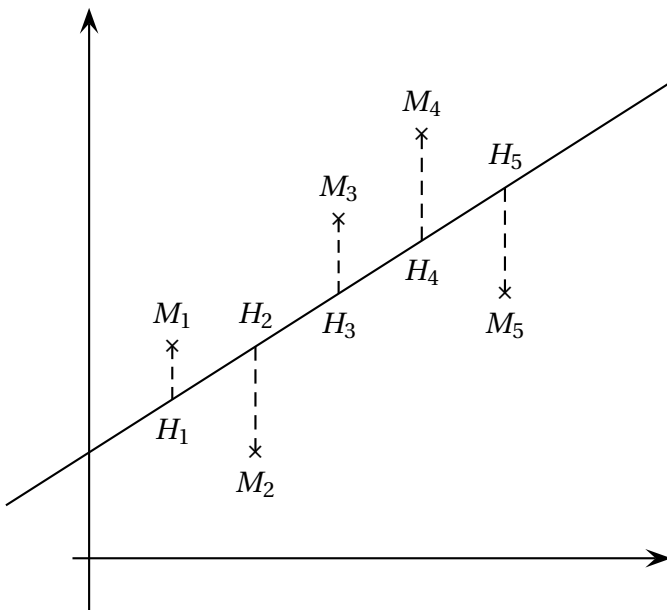
consiste à chercher la droite qui minimise la somme des carrés de ces erreurs : $\sum_{i=1}^n (y_i - ax_i - b)^2$.

La méthode des moindres carrés permet de déterminer l'équation de la « meilleure » droite passant dans le nuage de points, ainsi que de quantifier la « qualité de l'alignement des points » du nuage.

Exemple 8.4. On considère un nuage de points $M_k(x_k; y_k)$.

Pour une droite quelconque, on peut définir la "distance" de la droite au nuage de points par la somme des distances $M_k H_k$.

Ainsi, la "meilleure" droite passant dans le nuage de points est celle dont la distance au nuage de points est la plus petite.



Il existe une unique droite telle que la somme des distances

$$d = M_1 H_1^2 + M_2 H_2^2 + \dots + M_n H_n^2 = \sum_{k=1}^n M_k H_k^2$$

soit minimale. Cette droite est appelée **droite de régression de Y en X**, ou encore **droite des moindres carrés**.

Proposition 8.4 (Droite de régression linéaire).

L'unique droite rendant minimale $\sum_{i=1}^n (y_i - ax_i - b)^2$ est la droite d'équation :

$$y = \frac{s_{X,Y}}{s_X^2} (x - \bar{X}) + \bar{Y}.$$

Cette droite est appelée droite de régression linéaire de Y en X.

On dit que X est la variable explicative et Y la variable expliquée.

EXERCICE 8.2.

Faire la preuve de cette proposition 8.4. On considère (x_i) et (y_i) deux séries de variances non nulles et on pose pour tout (a, b) dans $\mathbb{R}^2$: $f(a, b) = \sum_{i=1}^n (y_i - ax_i - b)^2$.

1. (a) On suppose que $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i = 0$.

Montrer alors que pour tout (a, b) dans $\mathbb{R}^2$, $f(a, b) \geq a^2 \sum_{i=1}^n x_i^2 - 2a \sum_{i=1}^n x_i y_i + \sum_{i=1}^n y_i^2$.

- (b) Montrer qu'il y a égalité si et seulement si $b = 0$.

- (c) En déduire que $f(a, b)$ est minimale quand $a = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$ et $b = 0$.

2. Dans le cas général, pour quelle valeur de a et b , la quantité $f(a, b)$ est-elle minimale? Indication : on posera $x'_i = x_i - \bar{X}$, $y'_i = y_i - \bar{Y}$. On montrera que ces deux nouvelles séries sont de moyennes nulles, on exprimera $f(a, b)$ par rapport à x'_i et y'_i et on se ramènera au cas précédent en posant $a' = a$, $b' = b - \bar{Y} + a\bar{X}$.

SOLUTION.

On considère (x_i) et (y_i) deux séries de variances non nulles et on pose pour tout (a, b) dans $\mathbb{R}^2$:

$$f(a, b) = \sum_{i=1}^n (y_i - ax_i - b)^2.$$

Cas particulier : on suppose que $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i = 0$.

Alors en développant l'expression $(y_i - ax_i - b)^2$, on trouve : $f(a, b) = \sum_{i=1}^n y_i^2 - 2a \sum_{i=1}^n x_i y_i + a^2 \sum_{i=1}^n x_i^2 + \sum_{i=1}^n b^2 -$

$$2b \sum_{i=1}^n y_i + 2ab \sum_{i=1}^n x_i = \sum_{i=1}^n y_i^2 - 2a \sum_{i=1}^n x_i y_i + a^2 \sum_{i=1}^n x_i^2 + nb^2 \text{ et cette dernière quantité est bien supérieure à } a^2 \sum_{i=1}^n x_i^2 - 2a \sum_{i=1}^n x_i y_i + \sum_{i=1}^n y_i^2.$$

Donc on a bien pour tout (a, b) dans $\mathbb{R}^2$, $f(a, b) \geq a^2 \sum_{i=1}^n x_i^2 - 2a \sum_{i=1}^n x_i y_i + \sum_{i=1}^n y_i^2 (*)$.

On a égalité si et seulement si : $nb^2 = 0$ soit pour $b = 0 (**)$.

De plus $a^2 \sum_{i=1}^n x_i^2 - 2a \sum_{i=1}^n x_i y_i + \sum_{i=1}^n y_i^2$ est un polynôme du second degré en a de la forme $P(a) = \alpha a^2 + \beta a + \gamma$

et $P(a)$ minimale ssi $a = -\frac{\beta}{2\alpha}$ soit pour $a = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$.

D'où d'après (*) et (**), $f(a, b)$ est minimale quand $a = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$ et $b = 0$.

Cas général :

On pose $x'_i = x_i - \bar{X}$, $y'_i = y_i - \bar{Y}$ alors $\overline{x'} = \sum_{i=1}^n x'_i = \frac{1}{n} \sum_{i=1}^n x_i - \bar{X} = \bar{X} - \bar{X} = 0$.

De même avec $\overline{y'} = \sum_{i=1}^n y'_i = \bar{Y} - \bar{Y} = 0$.

$$\text{Puis } f(a, b) = \sum_{i=1}^n (y'_i + \bar{Y} - a(x'_i + \bar{X}) - b)^2 = \sum_{i=1}^n (y'_i - ax'_i - (b - \bar{Y} + a\bar{X}))^2.$$

Alors, en posant $a = a'$ et $b' = b - \bar{Y} + a\bar{X}$, on a $f(a', b') = \sum_{i=1}^n (y'_i - a'x'_i - b')^2$, avec $\sum_{i=1}^n x'_i = \sum_{i=1}^n y'_i = 0$.

On est ramené donc au cas particulier.

On sait alors que pour que $f(a', b')$ soit minimale on a $b' = 0$ donc pour $b = \bar{Y} - a\bar{X}$ et $a' = a = \frac{\sum_{i=1}^n x'_i y'_i}{\sum_{i=1}^n x'^2_i} =$

$$\frac{\sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{X})^2} = \frac{s_{X,Y}}{s_X^2}.$$

Il existe une autre façon de démontrer la proposition qui fait référence aux fonctions à deux variables.

Preuve. On considère la fonction $F(a, b) = \sum_{i=1}^n (y_i - ax_i - b)^2$.

Le minimum de cette fonction se trouve parmi les points critiques de cette fonction.

Recherche des points critiques :

On cherche les points où les dérivées partielles s'annulent; pour cela on calcule les dérivées partielles.

$$\partial_1(F)(a, b) = -2 \sum_{i=1}^n x_i (y_i - ax_i - b) = -2 \sum_{i=1}^n x_i y_i - 2a \sum_{i=1}^n x_i^2 - 2b \sum_{i=1}^n x_i = -2n \left(s_{X,Y} + \overline{XY} \right) - a(s_X^2 + \overline{X^2}) - b\overline{X}.$$

$$\partial_2(F)(a, b) = -2 \sum_{i=1}^n (y_i - ax_i - b) = -2n \left(\overline{Y} - a\overline{X} - b \right).$$

$$\text{On résout } \nabla_F(a, b) = 0 \Leftrightarrow \begin{cases} -2n \left(s_{X,Y} + \overline{XY} \right) - a(s_X^2 + \overline{X^2}) - b\overline{X} = 0 \\ -2n \left(\overline{Y} - a\overline{X} - b \right) = 0 \end{cases}$$

$$\Leftrightarrow \begin{cases} -2n \left(s_{X,Y} + \overline{XY} \right) - a(s_X^2 + \overline{X^2}) - \overline{Y} + a\overline{X} = 0 \\ \overline{Y} - a\overline{X} = b \end{cases} \Leftrightarrow \begin{cases} s_{X,Y} - as_X^2 = 0 \\ \overline{Y} - a\overline{X} = b \end{cases} \Leftrightarrow \begin{cases} a = \frac{s_{X,Y}}{s_X^2} \\ b = \overline{Y} - a\overline{X} \end{cases}$$

On a donc trouvé un unique point critique (a^*, b^*) tel que :

$$\begin{cases} a^* = \frac{s_{X,Y}}{s_X^2} \\ b^* = \overline{Y} - a^*\overline{X} \end{cases}$$

On va arranger l'expression de F sous deux formes :

On développe l'expression de F :

$$\begin{aligned} F(a, b) &= \sum_{i=1}^n (y_i - ax_i - b)^2 = \sum_{i=1}^n (y_i^2 + a^2 x_i^2 + b^2 - 2ax_i y_i - 2y_i b + 2abx_i) \\ &= ns_Y^2 + n\overline{Y^2} + na^2 s_X^2 + na^2 \overline{X^2} + nb^2 - 2ans_{X,Y} + 2n\overline{XY} - 2bn\overline{Y} + 2abn\overline{X} \\ &= n \left[(\overline{Y} - a\overline{X} - b)^2 + a^2 s_X^2 - 2as_{X,Y} + s_Y^2 \right] \quad (1) \end{aligned}$$

Puis une deuxième forme, qui sera la forme canonique en partant de (1) :

$$F(a, b) = \left[(\overline{Y} - a\overline{X} - b)^2 + \left(as_X - \frac{s_{X,Y}}{s_X} \right)^2 - \frac{s_{X,Y}^2}{s_X^2} + s_Y^2 \right] \quad (2).$$

Calculons maintenant $F(a^*, b^*)$ en partant de (1) et en remplaçant par l'expression de b^* :

$$F(a^*, b^*) = n \left(a^{*2} s_X^2 - 2a^* s_{X,Y} + s_Y^2 \right).$$

Puis par l'expression de a^* :

$$F(a^*, b^*) = n \left(s_Y^2 - \frac{s_{X,Y}^2}{s_X^2} \right) \quad (3) = ns_Y^2 \left(1 - r_{X,Y}^2 \right) \quad (4).$$

Comparons enfin $F(a, b)$ à $F(a^*, b^*)$ en utilisant (2) et (3).

$$\begin{aligned} F(a, b) - F(a^*, b^*) &= \left[(\overline{Y} - a\overline{X} - b)^2 + \left(as_X - \frac{s_{X,Y}}{s_X} \right)^2 - \frac{s_{X,Y}^2}{s_X^2} + s_Y^2 \right] - n \left(s_Y^2 - \frac{s_{X,Y}^2}{s_X^2} \right) \\ &= \left[(\overline{Y} - a\overline{X} - b)^2 + \left(as_X - \frac{s_{X,Y}}{s_X} \right)^2 \right] \end{aligned}$$

Or cette dernière quantité est toujours positive donc (a^*, b^*) est l'unique minimum de F .

Par conséquent l'équation cherchée est bien :

$$y = \frac{s_{X,Y}}{s_X^2} (x - \overline{X}) + \overline{Y}.$$

◇

Propriété 8.5.

Les droites de régression linéaire passent toujours par le point moyen du nuage : $G(\overline{X}; \overline{Y})$.

Remarque. On peut aussi demander la droite de régression linéaire de X en Y (il suffit d'inverser

les rôles de x et y) : $x = \frac{s_{X,Y}}{s_Y^2}(y - \bar{Y}) + \bar{X}$.

C'est l'unique droite rendant minimale $\sum_{i=1}^n (x_i - ay_i - b)^2$.

☛ Python : liste des commandes exigibles

On importera les librairies numpy et matplotlib.pyplot :

import numpy as np

import matplotlib.pyplot as plt

C_5 :

a=(np . mean (x*y)-np . mean(x)*np . mean(y))/np . var(x),

b=np . mean(y)-a*np . mean(x) renvoient a et b les réels a et b tels que : $y = ax + b$ soit la droite de régression linéaire de y en x .

C_6 : **plt . plot (x,a*x+b)** permet de tracer la droite de régression linéaire de y en x , si l'on connaît a et b .

☛ Python : liste des commandes non exigibles mais commode d'utilisation

C_5^* : **a, b=np.polyfit (x, y, 1)** affecte aux variables a et b les réels a et b tels que : $y = ax + b$ soit la droite de régression linéaire de y en x .

EXERCICE 8.3.

La droite de régression de l'exemple 8.1 de y en x a pour équation $y = 2,17x + 4,82$. Vérifier-le. Retrouver cette équation à l'aide de Python. Estimer à partir de ce modèle le coût de maintenance pour une installation de 7 ans, de 8 ans, puis de 10 ans.

SOLUTION.

On a $\bar{X} = 3,5$, $\bar{Y} = 12,41$.

Calcul de la variance de X : $s_X^2 = \frac{1+4+9+16+25+36}{6} - (3,5)^2 \simeq 15,17 - 12,25 \simeq 2,92$.

Calcul de la covariance de la série :

$$s_{X,Y} = \frac{1 \times 7,55 + 2 \times 9,24 + 3 \times 10,74 + 4 \times 12,84 + 5 \times 15,66 + 6 \times 18,45}{6} - 3,5 \times 12,41 \simeq 6,33$$

D'où l'équation de la droite de régression de y en x est :

$$y = \frac{s_{X,Y}}{s_X^2}(x - \bar{X}) + \bar{Y} = 2,17(x - 3,5) + 12,41 = 2,17x + 4,82.$$

Avec Python :

```
import numpy as np
import matplotlib.pyplot as plt
x=np.arange(1,7)
y=np.array[7.55,9.24,10.74,12.84,15.66,18.45]
a=(np . mean ( x*y )-np . mean( x )*np . mean( y ))/np . var( x )
b=np . mean( y )-a*np . mean( x )
plt . plot ( x,a*x+b )
plt.show()
```

Estimation des coûts de maintenance :

Pour $x = 7$ on trouve environ un coût de maintenance de : $y = 2,17 \times 7 + 4,82 = 20$.

Pour $x = 8$ on trouve environ un coût de maintenance de : $y = 22,2$.

Pour $x = 10$ on trouve environ un coût de maintenance de : $y = 26,5$.

8.2.2 Coefficient de corrélation

La droite de régression est la droite la plus proche de tous les points du nuage. Néanmoins, l'idée d'approcher tous les points du nuage par une droite peut-être plus ou moins pertinent.

Le coefficient de corrélation est un nombre qui quantifie justement ce degré de pertinence.

Proposition 8.6.

- Le coefficient de corrélation $r_{X,Y}$ prend des valeurs entre -1 et 1 : $-1 \leq r_{X,Y} \leq 1$
- $r_{X,Y}$ a le même signe que le coefficient directeur de la droite de régression.
- 1. Si $r_{X,Y} > 0$, alors la droite est de pente positive : X et Y varient dans le même sens.
- 2. Si $r_{X,Y} < 0$, alors la droite est de pente négative : X et Y varient dans des sens opposés.
- La corrélation est d'autant meilleure que $|r_{X,Y}|$ est proche de 1 (si $r_{X,Y}=1$ ou $r_{X,Y} = -1$, les points sont alignés et la corrélation est parfaite).

Preuve. On va démontrer le dernier point de la proposition.

On a prouvé dans la preuve 8.2.1, que $F(a^*, b^*)$ vérifiait la relation suivante

$$F(a^*, b^*) = ns_Y^2 (1 - r_{X,Y}^2) \quad (4).$$

Ainsi, plus $|r_{X,Y}|$ est proche de 1, plus $F(a^*, b^*)$ est petit et donc plus le nuage de points est proche de la droite de régression. $\diamond$

EXERCICE 8.4.

Temps de chargement et fréquentation d'un site web

Le temps de chargement d'une page sur internet dépend de nombreux paramètres, entre autre le nombre d'utilisateurs qui y sont connectés simultanément.

Par ailleurs, le temps de chargement influe en retour sur le nombre de visiteurs : plus le temps de chargement est long, plus les utilisateurs sont susceptibles de se diriger vers d'autres ressources.

Le responsable d'un site a relevé le nombre d'internautes sur son site en fonction de sa durée de chargement :

Nombre d'internautes connectés (en millier), x_i	0,5	1	2,5	3	4	5	6
Durée de chargement (en secondes), y_i	0,3	0,4	0,6	0,9	1,3	2	2,8

1. Représenter le nuage de points de coordonnées $(x_i; y_i)$ associés à cette série statistique. (Axes orthogonaux; unités : 2 cm pour 1000 internautes et 1cm pour 0,2 seconde).
2. Est-il pertinent de réaliser un ajustement linéaire?
3. Déterminer l'équation $y = ax + b$ de la droite d'ajustement $\mathcal{D}$ obtenue par la méthode des moindres carrés (Arrondir les coefficients au millième).

• **Point méthode : comment faire un ajustement linéaire de Y en X ?**

Il faut calculer $\bar{X}, \bar{Y}, s_X^2, s_Y^2, s_{X,Y}$ et $r_{X,Y}$.

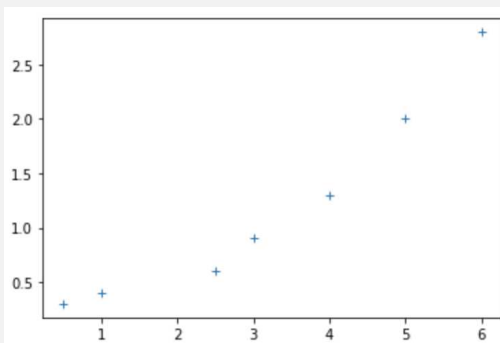
Si $r_{X,Y}$ est proche de 1 ou de -1, l'ajustement est pertinent et on donne l'équation cherchée par la formule :

$$y = \frac{s_{X,Y}}{s_X^2} (x - \bar{X}) + \bar{Y}.$$

4. Pour la suite, on prendra $y = 0,44x - 0,19$ pour équation de la droite $\mathcal{D}$.
- Tracer la droite $\mathcal{D}$.
 - Avec ce modèle, estimer la durée de chargement pour 8000 personnes connectées.
 - Une étude indépendante a montré que 60% des internautes cesse de charger une page pour se diriger vers un autre site dès que le temps de chargement dépasse 3,5 secondes. Avec le modèle précédent, estimer le nombre de visiteurs sur ce site lorsque la durée de chargement est de 3,5 secondes.
Combien de visiteurs perdrait-il alors?

SOLUTION. 1. Nuage de points avec Python :

```
import numpy as np
import matplotlib.pyplot as plt
x=[0.5,1,2.5,3,4,5,6]
y=[0.3,0.4,0.6,0.9,1.3,2,2.8]
plt.plot(x,y,'+')
plt.plot(np.mean(x),np.mean(y),'k.')#point moyen du nuage pour info
plt.show()
Résultat :
```



2. Le nuage de points a une forme alignée donc graphiquement il semble pertinent de faire un ajustement linéaire.
Par le calcul, on va chercher le coefficient de corrélation associé à cette série statistique.
Avec Python :

```
import numpy as np
import matplotlib.pyplot as plt
x=[0.5,1,2.5,3,4,5,6]
y=[0.3,0.4,0.6,0.9,1.3,2,2.8]
def r(x,y):
    A=np.cov(x,y,bias=True)
    return A[0,1]/np.sqrt(A[0,0]*A[1,1])
print(r(x,y))
#ou
print(np.corrcoef(x,y)[0,1])
```

Résultat : 0.9560241549849803

Sans Python, à la main :

Calcul des moyennes :

$$\bar{X} = \frac{0.5 + 1 + 2.5 + 3 + 4 + 5 + 6}{7} = 22/7 \approx 3.14 \text{ et } \bar{Y} = \frac{0.3 + 0.4 + 0.6 + 0.9 + 1.3 + 2 + 2.8}{7} = 8.3/7 \approx 1.18.$$

Calcul des variances et des écarts types :

$$s_X^2 = \frac{0.5^2 + 1^2 + 2.5^2 + 3^2 + 4^2 + 5^2 + 6^2}{7} - (3.14)^2 = 3.48 \text{ donc } s_X = 1.86.$$

$$s_Y^2 = \frac{0.3^2 + 0.4^2 + 0.6^2 + 0.9^2 + 1.3^2 + 2^2 + 2.8^2}{7} - (1.18)^2 \approx 0.73 \text{ donc } s_Y \approx \sqrt{0.73} \approx 0.85.$$

Calcul de la covariance de la série :

$$s_{X,Y} = \frac{0.5 \times 0.3 + 1 \times 0.4 + 2.5 \times 0.6 + 3 \times 0.9 + 4 \times 1.3 + 5 \times 2 + 6 \times 2.8}{7} - 3.14 \times 1.18 = 1.52$$

Le coefficient de corrélation est : $r_{X,Y} = \frac{s_{X,Y}}{s_X s_Y} \approx \frac{1.52}{1.86 \times \sqrt{0.85}} \approx 0.96$. Il est proche de 1 donc il est pertinent de faire un ajustement linéaire.

Avec Python :

```
import numpy as np
import matplotlib.pyplot as plt
x=[0.5,1,2.5,3,4,5,6]
y=[0.3,0.4,0.6,0.9,1.3,2,2.8]
print(np.mean(x))#moyenne des x
print(np.mean(y))#moyenne des y
print(np.var(x))#variance de x
print(np.var(y))#variance de y
print(np.std(x))#écart-type de x
print(np.std(y))#écart-type de y
print(np.mean(x*y)-np.mean(x)*np.mean(y))#covariance de x,y
print(np.cov(x,y,bias=True))#ou variances, covariance
```

3 Obtention de la droite de régression de Y en X par la méthode des moindres carrés : $y = \frac{s_{X,Y}}{s_X^2}(x - \bar{X}) + \bar{Y}$:

$$y = 1.52/3.48(x - 3.14) + 1.18 \text{ soit } y = 0.44x - 0.19.$$

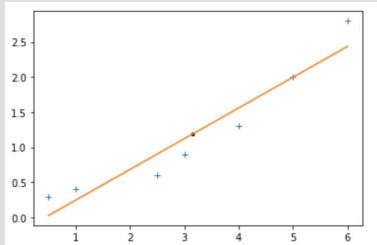
Avec python on rajoute au programme précédent :

```
a=(np.mean(x*y)-np.mean(x)*np.mean(y))/np.var(x)
b=np.mean(y)-a*np.mean(x)
print(a,b)
#ou
a,b=np.polyfit(x,y,1)
print(a,b)
```

on trouve : 0.4378299120234604 -0.19032258064516117

- 4 (a) on trace la droite à la main ou avec Python :

```
plt.plot(x,a*x+b)
```



- (b) on peut estimer la durée de chargement pour 8000 personnes connectées en faisant : $y = 0,44 * 8 - 0,19 \approx 3,33$ secondes.
- (c) Si la durée de chargement est de 3,5 secondes alors on a :
 $3,5 = 0,44 * x - 0,19$ d'où $x = (3,5 + 0,19) / 0,44 \approx 8400$ visiteurs sur ce site.
 Le site perdrait donc 5040 visiteurs (60% de 8400).

8.2.3 Remarque fondamentale : corrélérer n'est pas expliquer

Une erreur (malheureusement) assez répandue consiste à confondre corrélation avec causalité.

Observer que deux variables sont corrélées entre elles ne signifie pas que l'une soit la cause ou la conséquence de l'autre, c'est-à-dire qu'il y ait un lien de cause à effet.

Par exemple en France au 20ème siècle, le nombre de mariages a augmenté ainsi que le nombre de suicides. Ces deux variables sont sûrement corrélées, ce qui ne montre en aucun cas l'existence d'un lien de cause à effet d'un phénomène à l'autre (en fait ces deux augmentations peuvent être directement reliées à une variable commune, ici cachée : l'augmentation de la démographie).

EXERCICE 8.5.

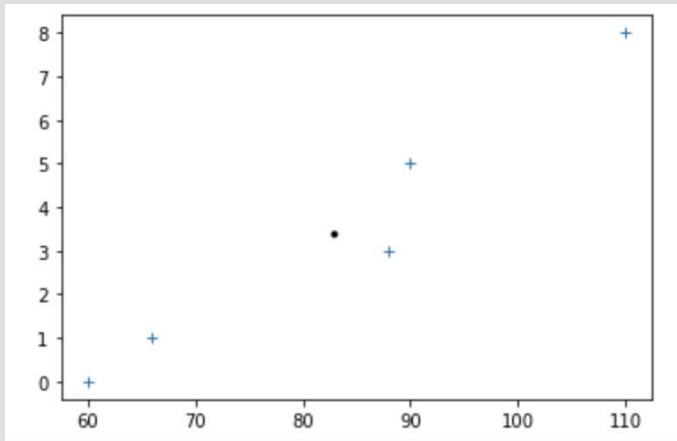
Le tableau suivant donne les évolutions, de Mai à Septembre, du nombre de climatiseurs vendus et de noyade par accident dans un secteur littoral.

	Nombre de climatiseurs x_i	Nombre de noyades y_i
Mai	66	1
Juin	88	3
Juillet	90	5
Août	110	8
Septembre	60	0

1. Représenter graphique le nuage de points correspondant.
2. Un ajustement affine semble-t'il pertinent? Donner l'équation de la droite d'ajustement par moindres carrés.
3. Prévoir le nombre de noyades si en Avril de l'année d'après, 88 climatiseurs sont vendus?
4. Commenter la relation de causalité entre les variables étudiées.

SOLUTION. 1. Nuage de points correspondant avec Python :

```
import numpy as np
import matplotlib.pyplot as plt
x=np.array([66,88,90,110,60])
y=np.array([1,3,5,8,0])
plt.plot(np.mean(x),np.mean(y),'k.')
plt.plot(x,y,'+')
```



2. Un ajustement affine semble-t-il pertinent? Les points du nuage sont assez alignés pour envisager un ajustement affine. On calcule le coefficient de corrélation, sans Python, à la main :

Calcul des moyennes :

$$\overline{X} = \frac{66 + 88 + 90 + 110 + 60}{5} = 414/5 = 82,8 \text{ et } \overline{Y} = \frac{1 + 3 + 5 + 8 + 0}{5} = 3,4.$$

Calcul des variances et des écarts types :

$$s_X^2 = \frac{66^2 + 88^2 + 90^2 + 110^2 + 60^2}{5} - (82,8)^2 \approx 324,168 \text{ donc } s_X \approx 18.$$

$$s_Y^2 = \frac{1 + 9 + 25 + 64}{5} - (3,4)^2 = 8,24 \text{ donc } s_Y \approx \sqrt{8,24} \approx 2,87.$$

Calcul de la covariance de la série :

$$s_{X,Y} = \frac{66 + 88 \times 3 + 90 \times 5 + 110 \times 8}{5} - 82,8 \times 3,4 = 50,48$$

Le coefficient de corrélation est : $r_{X,Y} = \frac{s_{X,Y}}{s_X s_Y} \approx \frac{50,48}{18 \times 2,87} \approx 0,97$. Il est proche de 1 donc il est pertinent de faire un ajustement linéaire.

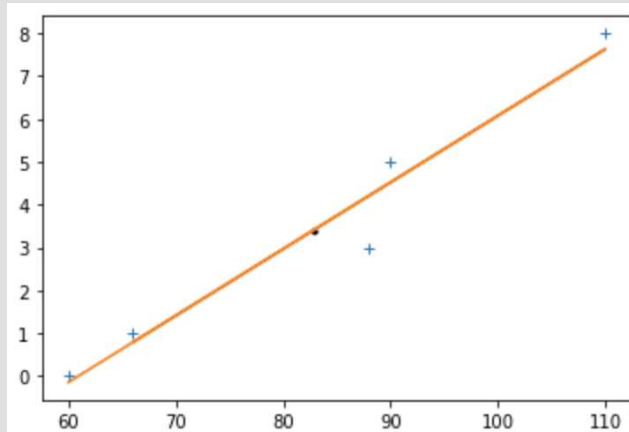
On calcule le coefficient de corrélation : avec Python, on rajoute

```
print(np.mean(x))#moyenne des x
print(np.mean(y))#moyenne des y
print(np.var(x))#variance de x
print(np.var(y))#variance de y
print(np.std(x))#écart-type de x
print(np.std(y))#écart-type de y
print(np.mean(x*y)-np.mean(x)*np.mean(y))#covariance de x,y
print(np.cov(x,y,bias=True))#variances, covariance
print((np.mean(x*y)-np.mean(x)*np.mean(y))/(np.std(x)*np.std(y)))#coeff
correlation
#ou
print(np.corrcoef(x,y)[0,1])
```

résultat : 0.9767333766271783 Il est très proche de 1, il est pertinent de faire un ajustement affine.

Donnons l'équation de la droite d'ajustement par moindres carrés. Avec Python :

```
a=(np.mean(x*y)-np.mean(x)*np.mean(y))/np.var(x)
print(a)
#0.1557255676209279
print(np.mean(y)-a*np.mean(x))
#-9.494076999012831
#ou
a,b=np.polyfit(x,y,1)
print(a,b)
plt.plot(x,a*x+b)
```



- 3 Prévoir le nombre de noyades si en Avril de l'année d'après, 88 climatiseurs sont vendus? Si l'on utilise l'équation de la droite de régression : $y = 0,16x - 9,49$, on a alors : $0,16 \times 88 - 9,49 = 4,6$ soit entre 4 et 5 noyades possibles.
- 4 Commenter la relation de causalité entre les variables étudiées. Il n'y a aucune relation de causalité entre les variables étudiées, les deux variables sont corrélées mais cela ne signifie que l'une soit la cause de l'autre! Pas de lien de cause à effet.

EXERCICE 8.6.

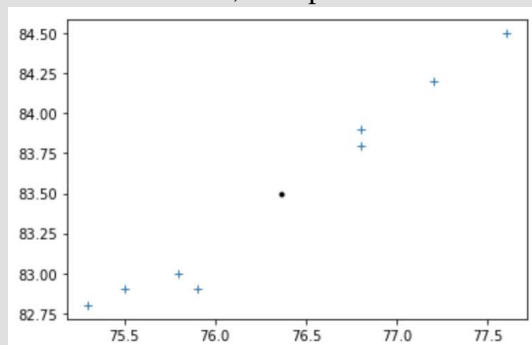
Le tableau suivant nous renseigne sur l'espérance de vie à la naissance pour les femmes et les hommes en France, entre 2000 et 2007 :

Année	Hommes	Femmes
2000	75,3	82,8
2001	75,5	82,9
2002	75,8	83
2003	75,9	82,9
2004	76,8	83,9
2005	76,8	83,8
2006	77,2	84,2
2007	77,6	84,5

1. Étudier la pertinence d'un ajustement linéaire entre les variables mesurant les espérances de vie pour les femmes et les hommes.
2. Déterminer éventuellement la droite de régression linéaire de Y (espérance de vie des femmes) en X (espérance de vie des hommes).

SOLUTION.

Pour fixer les idées, on représente tout d'abord le nuage de points associés à cette série à l'aide de Python.



La forme du nuage peut s'apparenter à une droite donc on peut conjecturer qu'un ajustement linéaire semble être pertinent, on va le prouver en calculant le coefficient de corrélation.

On note X l'espérance de vie des hommes, Y l'espérance de vie des femmes.

Calcul de la moyenne :

$$\bar{X} \approx 76,36, \bar{Y} \approx 83,5.$$

Pour la commodité des calculs, on va noter $X' = X - 76$, $Y' = Y - 83$, $(x'_i, y'_i)_{i \in \llbracket 1,8 \rrbracket}$ la série associée.

Les paramètres de cette série seront plus simples à calculer que ceux de la série initiale.

On a de plus : $\bar{X} = \bar{X}' + 76$, $\bar{Y} = \bar{Y}' + 83$, $s_X^2 = s_{X'}^2$, $s_Y^2 = s_{Y'}^2$, $s_{X,Y} = s_{X',Y'}$, $r_{X,Y} = r_{X',Y'}$.

On dispose les calculs comme ceci :

$x_i'^2$	x'_i	$x'_i y'_i$	y'_i	$y_i'^2$
0,49	-0,7	0,14	-0,2	0,04
0,25	-0,5	0,05	-0,1	0,01
0,04	-0,2	0	0	0
0,01	-0,1	0,01	-0,1	0,01
0,64	0,8	0,72	0,9	0,81
0,64	0,8	0,64	0,8	0,64
1,44	1,2	1,44	1,2	1,44
2,56	1,6	2,4	1,5	2,25
$\sum x_i'^2 = 6,07$	$\sum x'_i = 2,9$	$\sum x'_i y'_i = 5,4$	$\sum y'_i = 4$	$\sum y_i'^2 = 5,2$

Calcul de s_X^2 : $s_X^2 = s_{X'}^2 = 6,07/8 - (2,9/8)^2 \approx 0,63$.

Calcul de s_Y^2 : $s_Y^2 = s_{Y'}^2 = 5,2/8 - (4/8)^2 \approx 0,4$.

Calcul de $s_{X,Y}$: $s_{X,Y} = s_{X',Y'} = 5,4/8 - (2,9/8) \times (4/8)^2 = 0,4$.

Calcul de $r_{X,Y}$: $r_{X,Y} = \frac{s_{X,Y}}{s_X s_Y} = \frac{0,49}{\sqrt{0,63} \sqrt{0,4}} \approx 0,976$.

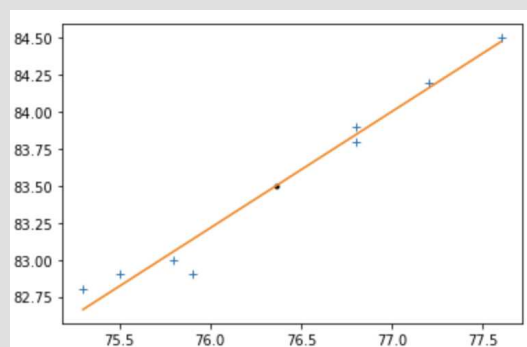
$r_{X,Y}$ étant proche de 1, un ajustement linéaire est pertinent.

Équation de la droite de régression de Y en X :

$$y = \frac{0,49}{0,63}(x - 76,36) + 83,5 \text{ soit en arrondissant au centième :}$$

$$y = 0,78x + 24,11.$$

NB : on peut vérifier avec Python et tracer cette droite.



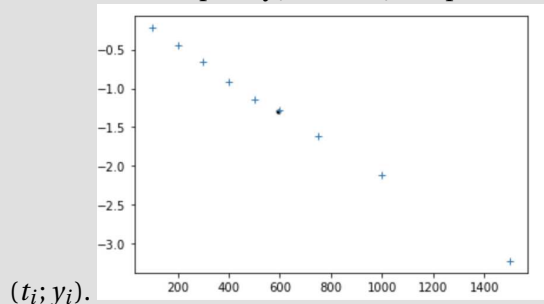
EXERCICE 8.7 (Durée de vie et maintenance d'équipements.).

Les pourcentages $R(t_i)$ des appareils mécaniques encore en service après un nombre t_i d'heures de fonctionnement ont été relevés et notés dans le tableau suivant :

t_i	100	200	300	400	500	600	750	1000	1500
$R(t_i)$	0,80	0,64	0,52	0,40	0,32	0,28	0,20	0,12	0,04

1. On pose $y_i = \ln R(t_i)$. Représenter graphiquement le nuage de points M_i de coordonnées $(t_i; y_i)$.
2. Peut-on envisager un ajustement affine de ce nuage de points?
Donner l'équation de la droite de régression de y en t .
En déduire une expression de la forme $R(t) = ke^{-\lambda t}$, avec k et λ des constantes.
3. Déterminer à l'aide du modèle précédent, le nombre d'équipements encore en service au bout de 900 heures de fonctionnement.

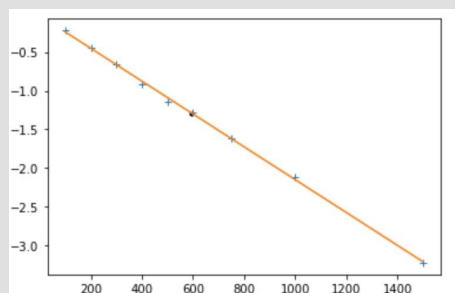
SOLUTION. 1. On pose $y_i = \ln R(t_i)$. Représentation graphique du nuage de points M_i de coordonnées



2. Peut-on envisager un ajustement affine de ce nuage de points? Oui car le nuage de points donne une forme de points alignés. On peut le vérifier avec Python

```
import numpy as np
import matplotlib.pyplot as plt
x=np.array([100,200,300,400,500,600,750,1000,1500])
y=np.array([np.log(0.8),np.log(0.64),np.log(0.52),np.log(0.4),np.log(0.32),
np.log(0.28),np.log(0.2),np.log(0.12),np.log(0.04)])
plt.plot(x,y,'+')

```



On calcule aussi le coefficient de corrélation : Avec Python, on rajoute

```
print((np.mean(x*y)-np.mean(x)*np.mean(y))/(np.std(x)*np.std(y)))
#ou
print(np.corrcoef(x,y)[0,1])
```

On trouve : -0.9995401483304258 Il est très très proche de -1, donc oui un ajustement affine est très pertinent. Donnons l'équation de la droite de régression de y en t .

Avec python :

```
a=(np.mean(x*y)-np.mean(x)*np.mean(y))/np.var(x)
print(a)
print(np.mean(y)-a*np.mean(x))
a,b=np.polyfit(x,y,1)
print(a,b)
```

Ce qui donne : $y = -0,002t - 0,03$.

On a $\ln R(t) = -0,002t - 0,03 \Leftrightarrow R(t) = e^{0,03} \cdot e^{-0,002t} = 1,03e^{-0,002t}$.

3 A l'aide du modèle précédent, le nombre d'équipements encore en service au bout de 900 heures de fonctionnement est donné par :

$R(900) = 1,03 \cdot e^{-0,002 \times 900} = 0,17$ soit 17% des appareils mécaniques sont en service après 900 heures de fonctionnement.

Feuille d'exercices n°8

STATISTIQUES BIVARIÉES

Exercice 1.

Le tableau ci-dessous donne le nombre de retraités en France métropolitaine entre 1975 et 2005 :

Année	Rang de l'année X	Nombre de retraités Y (en millions)
1975	0	4,1
1980	1	5
1985	2	5,9
1990	3	7,4
1995	4	8,3
2000	5	9,7
2005	6	10,7

1. Calculer les moyennes $\bar{X}$ et $\bar{Y}$ ainsi que leurs variances.
2. Calculer le coefficient de corrélation linéaire de cette série.

• Point méthode

On calcule $\frac{1}{n} \sum_{i=1}^n x_i y_i$ puis $\frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{X}\bar{Y}$.

Exercice 2. Temps de chargement et fréquentation d'un site web

Le temps de chargement d'une page sur internet dépend de nombreux paramètres, entre autre le nombre d'utilisateurs qui y sont connectés simultanément.

Par ailleurs, le temps de chargement influe en retour sur le nombre de visiteurs : plus le temps de chargement est long, plus les utilisateurs sont susceptibles de se diriger vers d'autres ressources.

Le responsable d'un site a relevé le nombre d'internautes sur son site en fonction de sa durée de chargement :

Nombre d'internautes connectés (en millier), x_i	0,5	1	2,5	3	4	5	6
Durée de chargement (en secondes), y_i	0,3	0,4	0,6	0,9	1,3	2	2,8

1. Représenter le nuage de points de coordonnées $(x_i; y_i)$ associés à cette série statistique. (Axes orthogonaux; unités : 2 cm pour 1000 internautes et 1cm pour 0,2 seconde).
2. Est-il pertinent de réaliser un ajustement linéaire?

3. Déterminer l'équation $y = ax + b$ de la droite d'ajustement $\mathcal{D}$ obtenue par la méthode des moindres carrés (Arrondir les coefficients au millièmes).

• **Point méthode : comment faire un ajustement linéaire de Y en X ?**

Il faut calculer $\bar{X}$, $\bar{Y}$, s_X^2 , s_Y^2 , $s_{X,Y}$ et $r_{X,Y}$.

☛ Si $r_{X,Y}$ est proche de 1 ou de -1, l'ajustement est pertinent et on donne l'équation cherchée par la formule :

$$y = \frac{s_{X,Y}}{s_X^2}(x - \bar{X}) + \bar{Y}.$$

4. Pour la suite, on prendra $y = 0,44x - 0,19$ pour équation de la droite $\mathcal{D}$.
- Tracer la droite $\mathcal{D}$.
 - Avec ce modèle, estimer la durée de chargement pour 8000 personnes connectées.
 - Une étude indépendante a montré que 60% des internautes cesse de charger une page pour se diriger vers un autre site dès que le temps de chargement dépasse 3,5 secondes. Avec le modèle précédent, estimer le nombre de visiteurs sur ce site lorsque la durée de chargement est de 3,5 secondes.
Combien de visiteurs perdrait-il alors ?

Exercice 3.

Le tableau suivant donne les évolutions, de Mai à Septembre, du nombre de climatiseurs vendus et de noyade par accident dans un secteur littoral.

	Nombre de climatiseurs x_i	Nombre de noyades y_i
Mai	66	1
Juin	88	3
Juillet	90	5
Août	110	8
Septembre	60	0

- Représenter graphique le nuage de points correspondant.
- Un ajustement affine semble-t-il pertinent ? Donner l'équation de la droite d'ajustement par moindres carrés.
- Prévoir le nombre de noyades si en Avril de l'année d'après, 88 climatiseurs sont vendus ?
- Commenter la relation de causalité entre les variables étudiées.

Exercice 4.

Le tableau suivant nous renseigne sur l'espérance de vie à la naissance pour les femmes et les hommes en France, entre 2000 et 2007 :

Année	Hommes	Femmes
2000	75,3	82,8
2001	75,5	82,9
2002	75,8	83
2003	75,9	82,9
2004	76,8	83,9
2005	76,8	83,8
2006	77,2	84,2
2007	77,6	84,5

1. Étudier la pertinence d'un ajustement linéaire entre les variables mesurant les espérances de vie pour les femmes et les hommes.
2. Déterminer éventuellement la droite de régression linéaire de Y en X .

Exercice 5. Équilibrer offre et demande

Une étude statistique effectuée sur un produit a permis de quantifier l'offre et la demande de ce produit, pour différentes valeurs de son prix unitaire.

Prix unitaire (en euros) x_i	Demande (en milliers d'unités) y_i	Offre (en milliers d'unités) z_i
1,2	8,4	0,75
2,5	6	1,25
3,5	5	1,75
4,5	4,2	2,25
5	3,5	2,5
7	2,1	3,5
8,5	1,2	4,25

1. Proposer un programme Python qui permet de représenter graphiquement, sur un même graphique, les nuages de points des séries à deux variables :
 - (a) la demande en fonction du prix unitaire (série à deux variables x et y).
 - (b) l'offre en fonction du prix unitaire (série à deux variables x et z).
2. Un ajustement affine semble-t-il pertinent? Justifiez graphiquement et par le calcul votre réponse.
Compléter le programme afin d'obtenir une droite d'ajustement pour chacun des deux nuages.
3. Estimer graphiquement,
 - (a) la demande et l'offre pour un prix unitaire de 6 euros;
 - (b) le prix unitaire pour une demande de 7 milliers;
 - (c) le prix unitaire pour une offre de 1 millier.
4. Le programme donne en sortie :

```
-0.9384133611691025 8.659558604235016
0.4866910229645094 0.0826498657918288
```

A quoi correspondent ces valeurs, en déduire les deux équations des droites d'ajustement par moindres carrés. Vérifier les estimations de la question précédente.

5. Graphiquement pour quel prix unitaire y a-t-il équilibre entre l'offre et la demande? Vérifier le par le calcul.

Exercice 6.

Le tableau suivant donne le nombre de clients annuel, en millier, d'une nouvelle chaîne de magasins.

Année	2006	2007	2008	2009	2010	2011	2012	2013
Rang de l'année	0	1	2	3	4	5	6	7
Nombre de clients	11,2	20,6	29,7	37,0	39,6	41,7	44,5	48,0

Représenter le nuage de points $(x_i; y_i)$ du nombre de clients en fonction du rang de l'année avec Python.

Partie A. Ajustement affine.

1. Déterminer avec Python une équation de la droite de régression de y en x obtenue par la méthode des moindres carrés.
Pour la suite, on utilisera l'ajustement affine donné par la droite D d'équation $y = 4,9x + 16,7$.
2. Faire tracer la droite D sur le nuage de points précédent.
3. Prévoir à l'aide de ce modèle le nombre de clients en 2015 et 2016.

Partie B. Ajustement par une fonction logistique.

Un autre ajustement est obtenu à l'aide de la fonction f définie par $f(x) = \frac{52}{1 + 3e^{-0,6x}}$.

1. Compléter le tableau de valeurs avec Python :

Rang de l'année x	0	1	2	3	4	5	6	7
$f(x)$		19,6	27,3					

2. Faire tracer l'allure de la courbe $\mathcal{C}$ représentative de la fonction f sur le nuage de points précédent.
3. Pourquoi l'ajustement logistique est meilleure que celui donnée par la droite de régression?
4. Donner à l'aide de ce modèle le nombre de clients estimé en 2015 et 2016.

Exercice 7. Durée de vie et maintenance d'équipements.

Les pourcentages $R(t_i)$ des appareils mécaniques encore en service après un nombre t_i d'heures de fonctionnement ont été relevés et notés dans le tableau suivant :

t_i	100	200	300	400	500	600	750	1000	1500
$R(t_i)$	0,80	0,64	0,52	0,40	0,32	0,28	0,20	0,12	0,04

1. On pose $y_i = \ln R(t_i)$. Représenter graphiquement le nuage de points M_i de coordonnées $(t_i; y_i)$.
2. Peut-on envisager un ajustement affine de ce nuage de points?
Donner l'équation de la droite de régression de y en t .
En déduire une expression de la forme $R(t) = ke^{-\lambda t}$, avec k et λ des constantes.
3. Déterminer à l'aide du modèle précédent, le nombre d'équipements encore en service au bout de 900 heures de fonctionnement.

Exercice 8.

Le tableau suivant donne l'évolution du chiffre d'affaire (en millions d'euros) d'une entreprise depuis sa création en 2002 :

Année X	2002	2003	2004	2005	2006	2007	2008	2009
Chiffre d'affaire Y	0,7	1,6	2	2,4	2,5	2,8	3	3

On s'intéresse à la série statistique (X, Y) .

1. Étudier la pertinence d'un ajustement linéaire.
2. Calculer le coefficient de corrélation de X et $\exp(Y)$. Le comparer à celui de X et Y .
3. En déduire un ajustement logarithmique de Y en X .

Exercice 9. Droite de Mayer

Le tableau suivant donne la durée moyenne d'intervention, en minutes, sur les postes de télévision en panne dans un atelier de dépannage, de 1992 à 2000.

Rang x_i	1	2	3	4	5	6	7	8	9
Année	1992	1993	1994	1995	1996	1997	1998	1999	2000
Durée moyenne d_i	83	82	80	75	73	74	71	71	70

Partie A. Ajustement à l'aide de la droite de régression

1. Calculer le coefficient de corrélation linéaire entre x et d (à 10^{-3} près).
Semble-t-il y avoir une dépendance affine entre l'année et la durée moyenne des interventions?
2. Donner la droite de régression des moindres carrés de d en x (valeurs arrondies à 10^{-2} près).
3. En supposant que l'évolution se poursuit ainsi pendant les 5 années futures, estimer la durée moyenne d'intervention dans cet atelier en 2002.

Partie B. Ajustement à l'aide de la droite de Mayer

La méthode de Mayer consiste à partager la série en 2. Soit S_1 la série correspondant aux années 1992-1996, et S_2 la série correspondant aux années 1997-2000.

1. Calculer les coordonnées du point moyen G_1 de la série S_1 , et du point moyen G_2 de la série S_2 .
2. Déterminer l'équation de la droite ($G_1 G_2$) appelée droite de Mayer.
3. Estimer la durée moyenne d'intervention dans cet atelier en 2002 avec la droite de Mayer et comparer avec la droite de régression.

Exercice 10. Ajustement exponentiel

Après un accident nucléaire, on procède à intervalles de temps réguliers à des mesures de radioactivité sur un site donné. Le tableau suivant donne les résultats de ces mesures.

Rang x_i de la mesure	1	2	3	4	5	6
Valeur y_i mesurée	100	61	37	22	14	7

1. Tracer le nuage de points correspondant. Utiliser une droite de régression linéaire semble-t-il pertinent?
2. Pour chaque mesure on pose $z_i = \ln y_i$ et on étudie alors la série statistique $(x_i; z_i)$.
Compléter le tableau :

Rang x_i de la mesure	1	2	3	4	5	6
$z_i = \ln y_i$						

3. Calculer le coefficient de corrélation de cette série à 0,001 près. Commenter le résultat.
4. Donner une équation de la droite D de régression de z en x (arrondir les coefficients à 0,01 près).
5. En déduire une relation entre x et y du type $y = \alpha e^{\beta x}$, où α et β sont deux constantes à déterminer.
6. En supposant que le modèle reste valable, en déduire pour la prochaine mesure ($x_i = 7$) une estimation de y .

7. En supposant toujours que le modèle reste valable, déterminer à partir de quelle mesure la valeur y mesurée sera inférieure à 0,01.

Exercice 11. Ajustement carré

Au cours d'une séance d'essai, un pilote automobile doit, quand il reçoit un signal sonore dans son casque, arrêter le plus rapidement possible son véhicule. Au moment du top sonore, on mesure la vitesse de l'automobile puis la distance nécessaire pour arrêter le véhicule.

Pour six expériences, on a obtenu les résultats suivants :

v_i (km/h)	27	43	62	80	98	115
distance y_i d'arrêt (m)	6,8	20,5	35,9	67,8	101,2	135,8

On pose $x_i = v_i^2$ et on considère la série $(x_i; y_i)$.

- Compléter le tableau

x_i						
y_i	6,8	20,5	35,9	67,8	101,2	135,8
- Représenter le nuage de points associé à cette nouvelle série.
- Déterminer, l'équation de la droite de régression de y en x . Tracer cette droite dans le repère précédent.
 - A l'aide de cette équation, déterminer la valeur estimée de x correspondant à une distance d'arrêt de 180 m, puis la vitesse correspondante du véhicule.
 - Quelle est la vitesse d'arrêt estimée correspondant à une vitesse de 150 km/h.
 - Le manuel du code de la route donne, pour calculer la distance d'arrêt, en mètres, la méthode suivante : "Prendre le carré de la vitesse exprimé en dizaines de kilomètres par heure."
Comparer le résultat obtenu au c. à celui que l'on obtiendrait par cette méthode.