7

STATISTIQUES À DEUX VARIABLES

Résumé

En statistiques, la corrélation entre deux variables montre le lien de dépendance qu'il existe entre ces variables. Pour autant, on ne peut pas affirmer que ce lien soit toujours un lien de cause à effet et c'est l'un des principaux moyens de désinformation.

1 Série statistique à deux variables

Définition

Une **série statistiques à deux variables** est une série statistique dont la population possède deux caractéristiques quantitatives distinctes ; si, pour un effectif total n, la première caractéristique est notée x_i et la seconde y_i avec $1 \leqslant i \leqslant n$ alors on peut représenter la série par le tableau ci dessous.

Caractère x	x_1	x_2	…	x_n
Caractère y	y_1	y_2	…	y_n

Exemple On note les poids et tailles de 4 individus.

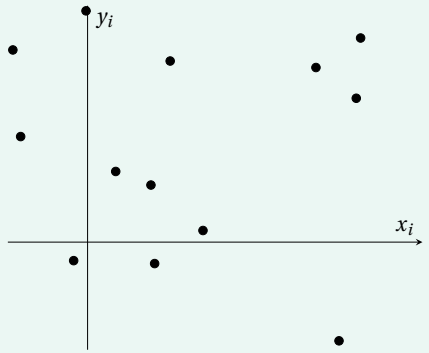
Taille (en m)	1,71	1,64	1,82	1,77
Poids (en kg)	64	76	89	59

Définition

Soit une série statistique à deux variables définies par le tableau ci-dessous.

x_i	x_1	x_2	…	x_n
y_i	y_1	y_2	…	y_n

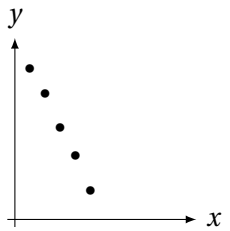
- Dans un repère du plan, on appelle **nuage de points** l'ensemble des points M de coordonnées (x_i, y_i).



- On appelle **point moyen** le point $G\left(\overline{x}, \overline{y}\right)$ où $\overline{x}$ est la moyenne des x_i et $\overline{y}$ celle des y_i.

Exemple Considérons la série statistique à deux variables définies par le tableau ci-contre.

x_i	1	2	3	4	5
y_i	9,9	8,2	6	4,1	1,8

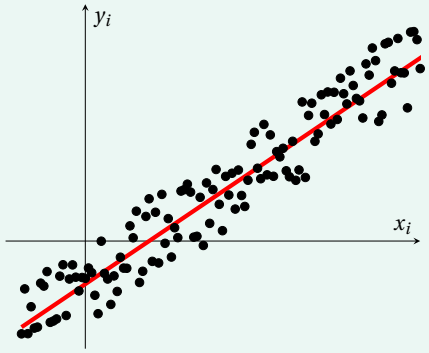


Le point moyen G ici a pour coordonnées $G(3;6)$. Il s'agit d'un point de la série mais ce n'est pas toujours le cas.
En effet, $\overline{x} = \frac{1+2+3+4+5}{5} = 3$ et $\overline{y} = \frac{9{,}9+8{,}2+6+4{,}1+1{,}8}{5} = 6$.

2 Ajustement affine

Définition

Lorsque les points du nuage d'une série statistique à deux variables sont « presque alignés » on peut construire une droite qui « passe le plus près possible de ces points ».
On dit que cette droite réalise un **ajustement affine** du nuage de points et s'appelle **droite de régression**.



Propriété

Dans un ajustement affine *pertinent*, le point moyen appartient à la droite de régression.

Définition | Droite des points extrêmes

Lorsque le nuage de points est approximativement aligné, on peut tracer la droite passant par les deux points extrêmes du nuage, c'est-à-dire le point d'abscisse minimale et le point d'abscisse maximale.
Cette droite fournit un **ajustement affine par les points extrêmes**.

Exemple Soit la série :

x_i	1	2	3	4	5
y_i	9,9	8,2	6	4,1	1,8

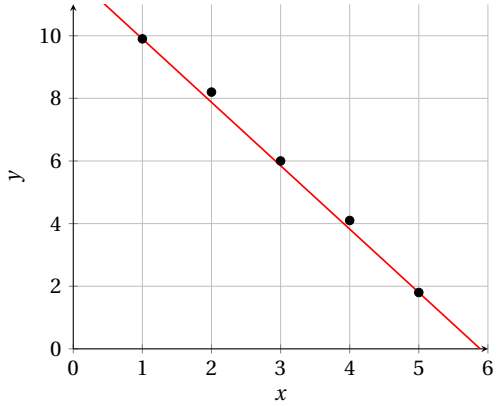
Les points extrêmes sont $A(1;9,9)$ et $B(5;1,8)$.

$$a = \frac{1,8-9,9}{5-1} = \frac{-8,1}{4} = -2,025$$

On peut déterminer b en considérant $y = -2,025x + b$.
Sachant que A est sur cette droite, on a :
$9,9 = -2,025 \times 1 + b$ donc $b = 9,9 + 2,025 = 11,925$.

La droite d'ajustement par les points extrêmes est :

$$\boxed{y = -2,025x + 11,925}.$$



Définition | Droite de Mayer

On ordonne les points suivant les abscisses croissantes, puis on partage la série en deux groupes de même effectif (ou presque égal si l'effectif est impair).
On calcule ensuite les points moyens G_1 et G_2 de chaque groupe.
La droite passant par G_1 et G_2 s'appelle la **droite de Mayer**.

Exemple On reprend la série :

x_i	1	2	3	4	5
y_i	9,9	8,2	6	4,1	1,8

On partage en deux groupes :

$$\{(1;9.9),(2;8.2),(3;6)\} \qquad \{(4;4.1),(5;1.8)\}$$

Points moyens :

$$G_1\left(\frac{1+2+3}{3};\frac{9.9+8.2+6}{3}\right) = (2;8,03)$$

$$G_2\left(\frac{4+5}{2};\frac{4,1+1,8}{2}\right) = (4,5;2,95)$$

Coefficient directeur :

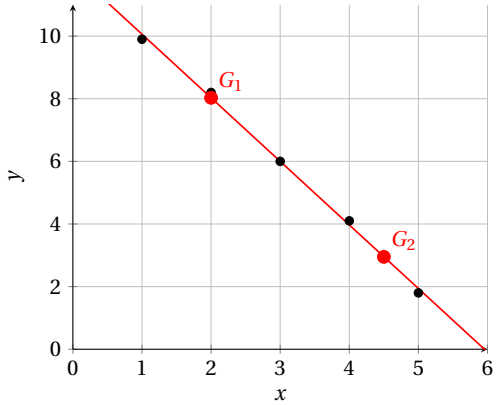
$$a = \frac{2,95 - 8,03}{4,5 - 2} = \frac{-5,08}{2,5} = -2,03$$

Ordonnée à l'origine :

$$b = 8,03 - (-2,03) \times 2 = 12,09$$

La droite de Mayer est donc :

$$\boxed{y = -2,03x + 12,09}$$



3 Moindres carrés

Définition | Covariance

Soient deux séries statistiques x et y d'effectif total n et de moyennes respectives $\overline{x}$ et $\overline{y}$.
On appelle **covariance** de x et y le nombre :

$$\text{Cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \frac{(x_1 - \overline{x})(y_1 - \overline{y}) + \cdots + (x_n - \overline{x})(y_n - \overline{y})}{n}.$$

Définition

Soit (x, y) une série statistiques à deux variables.
La droite $\mathscr{D}$ de régression par les moindres carrés (de y en x) est la droite d'équation $\mathscr{D} : y = ax + b$ avec :

- $a = \dfrac{\text{Cov}(x,y)}{\text{Var}(x)}$
- $b = \overline{y} - a\overline{x}$

Théorème

La droite $\mathscr{D}$ des moindres carrés est une droite de régression.

Exemple Reprenons la série de l'exemple précédent. On a $\overline{x} = 3$ et $\overline{y} = 6$.
Ensuite, calculons $\text{Var}(x)$ et $\text{Cov}(x,y)$ puis $a = \dfrac{\text{Cov}(x,y)}{\text{Var}(x)}$ et $b = \overline{y} - a\overline{x}$.

$$\text{Var}(x) = \tfrac{1}{5}\sum_{i=1}^{5}(x_i - \overline{x})^2 = \frac{(1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2}{5} = 2.$$

$$\begin{aligned}\text{Cov}(x,y) &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \\ &= \frac{(1-3)(9,9-6) + (2-3)(8,2-6) + (3-3)(6-6) + (4-3)(4,1-6) + (5-3)(1,8-5)}{5} \\ &= -3,66\end{aligned}$$

$$a = \frac{\text{Cov}(x,y)}{\text{Var}(x)} = \frac{-3,66}{2} = -1,88$$

$$b = \overline{y} - a\overline{x} = 6 - (-1,88) \times 3 = 11,64$$

Finalement, la droite $\mathscr{D}$ des moindres carrés admet $a = \dfrac{\text{Cov}(x,y)}{\text{Var}(x)}$ comme coefficient directeur et $b = \overline{y} - a\overline{x}$ pour ordonnée à l'origine.