

# 2

## 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  $1 \leq i \leq n$  alors on peut représenter la série par le tableau ci dessous.

Caractère $x$	$x_1$	$x_2$	$\dots$	$x_n$
Caractère $y$	$y_1$	$y_2$	$\dots$	$y_n$

**Exemple** On note les poids et tailles de 4 individus.

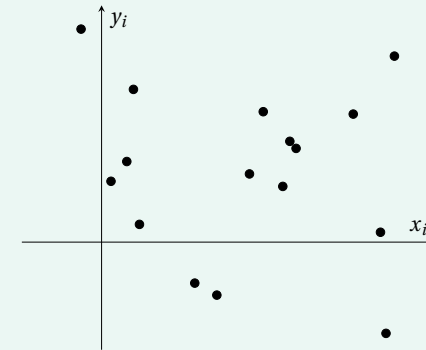
Taille (en m)	1,71	1,64	1,82	1,77
Poids (en kg)	64	76	89	59

#### Définition

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

$x_i$	$x_1$	$x_2$	$\dots$	$x_n$
$y_i$	$y_1$	$y_2$	$\dots$	$y_n$

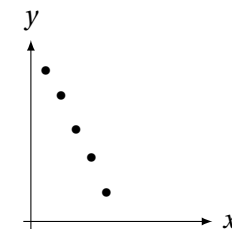
- Dans un repère du plan, on appelle **nuage de points** l'ensemble des points  $M$  de coordonnées  $(x_i, y_i)$ .



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

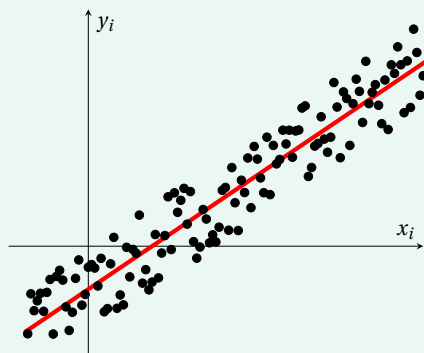
En effet,  $\bar{x} = \frac{1+2+3+4+5}{5} = 3$  et  $\bar{y} = \frac{9,9+8,2+6+4,1+1,8}{5} = 6$ .

## 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.

*Démonstration.* Admise. □

### Définition | Covariance

Soient deux séries statistiques  $x$  et  $y$  d'effectif total  $n$  et de moyennes respectives  $\bar{x}$  et  $\bar{y}$ .

On appelle **covariance** de  $x$  et  $y$  le nombre :

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

### Définition

Soit  $(x, y)$  une série statistiques à deux variables.

La droite  $\mathcal{D}$  de régression par les moindres carrés (de  $y$  en  $x$ ) est la droite d'équation  $\mathcal{D} : y = ax + b$  avec :

$$\blacktriangleright a = \frac{\text{Cov}(x, y)}{\text{Var}(x)}$$

$$\blacktriangleright b = \bar{y} - a\bar{x}$$

### Théorème

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

*Démonstration.* Admise. □

**Exemple** Reprenons la série de l'exemple précédent. On a  $\bar{x} = 3$  et  $\bar{y} = 6$ .

Ensuite, calculons  $\text{Var}(x)$  et  $\text{Cov}(x, y)$  puis  $a = \frac{\text{Cov}(x, y)}{\text{Var}(x)}$  et  $b = \bar{y} - a\bar{x}$ .

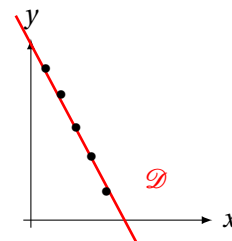
$$\text{Var}(x) = \frac{1}{5} \sum_{i=1}^5 (x_i - \bar{x})^2 = \frac{(1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2}{5} = 2.$$

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

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

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

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



### 3 Corrélation

#### Définition

On appelle **coefficient de corrélation linéaire** de deux statistiques  $x$  et  $y$ , d'écarts-type  $\sigma(x)$  et  $\sigma(y)$ , le nombre réel  $r$  défini par :

$$r = \frac{\text{Cov}(x, y)}{\sigma(x)\sigma(y)}.$$

#### Propriété

Le coefficient de corrélation linéaire  $r$  est compris entre  $-1$  et  $1$  inclus.

$$-1 \leq r \leq 1$$

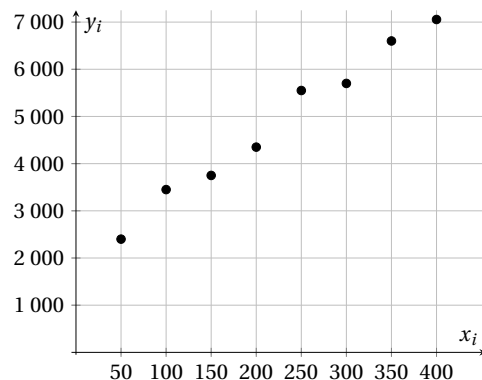
#### Définition

Lorsque  $|r|$  est très proche de  $1$ , on dit que la corrélation est forte entre  $x$  et  $y$ .

**Exemple** Considérons le tableau suivant exprimant le chiffre d'affaires  $y$  en euro par rapport à la quantité d'engrais (en L) utilisée sur des parcelles de maïs de  $10\,000\text{ m}^2$ .

Quantité $x_i$	50	100	150	200	250	300	350	400
CA $y_i$	2400	3450	3750	4350	5550	5700	6600	7055

Le nuage de points associé est le suivant.



L'ajustement affine semble convenir puisque les points sont "plutôt" alignés. Réalisons-le par les moindres carrés.

Calculons les différents indicateurs nécessaires.

►  $\bar{x} = 225$

►  $\text{Var}(x) = 15\,000$  donc  $\sigma(x) = \sqrt{\text{Var}(x)} = \sqrt{15\,000} \approx 122,474$

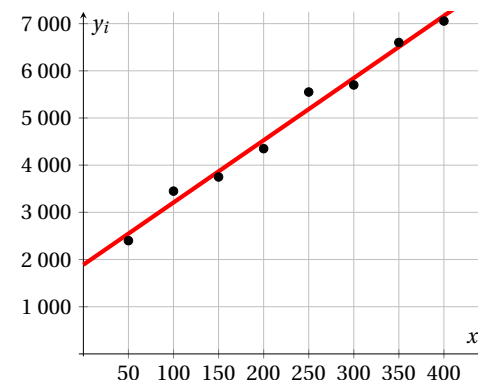
►  $\bar{y} = 4856,875$

►  $\text{Var}(y) \approx 2\,651\,306,696$  donc  $\sigma(y) = \sqrt{\text{Var}(y)} \approx \sqrt{2\,651\,306,696} \approx 1\,628,283$

►  $\text{Cov}(x, y) = 173\,078,125$

La droite des moindres carrés a pour équation :  $y = \frac{\text{Cov}(x, y)}{\text{Var}(x)}x + \bar{y} - \frac{\text{Cov}(x, y)}{\text{Var}(x)}\bar{x}$  ce qui donne pour équation approchée :

$$y \approx 13,187x + 1\,889,821.$$



Vérifions la corrélation grâce au coefficient de corrélation linéaire  $r$ .

$$r = \frac{\text{Cov}(x, y)}{\sigma(x)\sigma(y)} \approx \frac{173\,078,125}{122,474 \times 1\,628,283} \approx 0,868$$

Comme  $|r| \geq 0,75$ , nous pouvons dire que la corrélation linéaire entre  $x$  et  $y$  est très forte. L'ajustement affine était justifié.