

Corrélation et causalité

Séries à deux variables, ajustement affine, changement de variable

TERMINALE · MATHS COMPLÉMENTAIRES

1 Statistiques à deux variables

Nuage de points et point moyen

Une série statistique à deux variables est une liste de couples $(x_i; y_i)$, $1 \leq i \leq n$.

- Le **nuage de points** est l'ensemble des points $M_i(x_i; y_i)$ dans un repère.
- Le **point moyen** est $G(\bar{x}; \bar{y})$, avec $\bar{x} = \frac{1}{n} \sum x_i$ et $\bar{y} = \frac{1}{n} \sum y_i$.

Variance, covariance

$$V(x) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \quad \text{cov}(x, y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

La covariance est positive quand x et y ont tendance à varier dans le même sens.

2 Ajustement affine

Droite des moindres carrés

Lorsque le nuage a une forme allongée « rectiligne », on l'ajuste par la droite $y = ax + b$ qui rend minimale la somme $\sum (y_i - ax_i - b)^2$ des carrés des écarts verticaux. Elle est donnée par

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

Cette droite **passé par le point moyen** G . En pratique, on obtient a et b à la calculatrice ou au tableur.

Coefficient de corrélation linéaire

$$r = \frac{\text{cov}(x, y)}{\sigma(x)\sigma(y)} \in [-1; 1]$$

Plus $|r|$ est proche de 1, plus les points sont proches de la droite d'ajustement (en pratique, l'ajustement est jugé pertinent si $|r| \geq 0,9$ environ). Le signe de r est celui de la pente a .

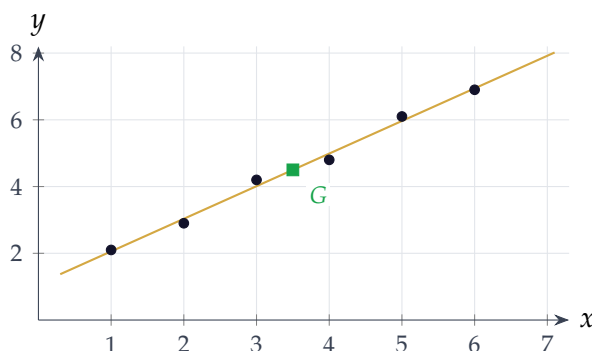
Méthode — ajustement affine et prévision

Exemple : Chiffre d'affaires y (en M€) d'une entreprise selon le rang x de l'année :

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

- Point moyen : $\bar{x} = 3,5$, $\bar{y} = \frac{27}{6} = 4,5$, donc $G(3,5; 4,5)$.

2. Calculatrice : $\text{cov}(x, y) = 2,85$ et $V(x) = \frac{35}{12} \approx 2,917$, d'où $a \approx 0,977$ et $b = 4,5 - 0,977 \times 3,5 \approx 1,08$: $y = 0,98x + 1,08$.
3. $r \approx 0,997$: l'ajustement affine est très pertinent.
4. Prévision pour l'année 8 : $y \approx 0,98 \times 8 + 1,08 \approx 8,9 \text{ M€}$ (**extrapolation** : on suppose que la tendance se poursuit).



Corrélation n'est pas causalité

Un coefficient $|r|$ proche de 1 indique une **corrélacion** (les deux variables évoluent ensemble), pas une **causalité** (l'une provoque l'autre). La corrélation peut venir d'un **facteur commun caché** : les ventes de glaces et le nombre de noyades sont fortement corrélées... parce que les deux augmentent avec la chaleur estivale. Seule une étude expérimentale (ou un mécanisme explicatif) permet d'établir un lien de cause à effet.

3 Ajustement et changement de variable

Modèle exponentiel

Si le nuage $(x_i; y_i)$ a une allure exponentielle (croissance « en pourcentage »), on pose $z_i = \ln y_i$. Si le nuage $(x_i; z_i)$ est rectiligne, on l'ajuste par $z = ax + b$, et alors

$$y = e^z = e^{ax+b} = k e^{ax} \quad \text{avec } k = e^b$$

Méthode — ajustement exponentiel

Exemple : Nombre N (en milliers) de bactéries dans une culture au bout de t heures :

t_i	0	1	2	3	4	5
N_i	2,3	3,1	4,6	6,8	9,9	14,6
$z_i = \ln N_i$	0,833	1,131	1,526	1,917	2,293	2,681

1. Le nuage $(t_i; N_i)$ est courbé, mais le nuage $(t_i; z_i)$ est quasiment aligné : $r \approx 0,999$.
2. Droite des moindres carrés (calculatrice) : $z \approx 0,375t + 0,793$.
3. Retour à N : $N = e^{0,375t+0,793} = e^{0,793} e^{0,375t} \approx 2,21 e^{0,375t}$.
4. Prévision à $t = 6$ h : $N \approx 2,21 e^{2,25} \approx 20,9$ milliers de bactéries.

Interprétation : $e^{0,375} \approx 1,45$, la population augmente d'environ 45 % par heure.

À retenir

- Point moyen $G(\bar{x}; \bar{y})$: la droite des moindres carrés passe par G .
- $a = \frac{\text{cov}(x, y)}{V(x)}$, $b = \bar{y} - a\bar{x}$; coefficient $r \in [-1; 1]$, pertinent si $|r|$ proche de 1.
- Allure exponentielle : $z = \ln y$, ajustement $z = ax + b$, puis $y = e^b e^{ax}$.
- **Corrélation** $\neq$ **causalité** : méfiance envers les facteurs cachés.
- **Piège** : une prévision par extrapolation loin des données est hasardeuse, le modèle peut cesser d'être valable.