MATHÉMATIQUES

Statistique à deux variables : corrélation et causalité

Terminale maths complémentaires

Nuage de points • Point moyen G • Ajustement affine • Droite des moindres carrés • Coefficient de corrélation • Changement de variable • Interpolation • Extrapolation • Corrélation et causalité

Repères du programme

Ce cours suit le **programme de l'enseignement optionnel de mathématiques complémentaires de terminale générale** (arrêté du 19 juillet 2019, contenus repris par l'arrêté du 26 février 2026 pour la rentrée 2027), partie « Statistique à deux variables quantitatives », et le thème d'étude « Corrélation et causalité ». Il couvre :

- nuage de points, point moyen ;
- ajustement affine, droite des moindres carrés, coefficient de corrélation ;
- ajustement se ramenant à un ajustement affine par changement de variable ;
- interpolations et extrapolations, regard critique ;
- **démonstration** : droite des moindres carrés (minimum d'un trinôme).

La calculatrice ou le tableur donnent les coefficients ; le cours explique ce qu'ils signifient.

Objectif : représenter une série statistique à deux variables, déterminer une droite de régression (directement ou après un changement de variable), l'utiliser pour interpoler ou extrapoler, et distinguer corrélation et causalité.

Document pédagogique – mathlvl.fr

mathlvl.fr — Tous droits réservés. Toute reproduction, diffusion, revente ou exploitation commerciale, même partielle, est strictement interdite. Tout contrevenant s'expose à des poursuites en vertu du Code de la propriété intellectuelle (art. L.335-2 et suivants).

Table des matières

1 Nuage de points et point moyen **3**

2 La droite des moindres carrés **3**

3 Changement de variable **4**

4 Interpoler, extrapoler, et rester critique **5**

5 Erreurs classiques **6**

6 Fiche de synthèse **6**

7 Exercices d'application **6**

8 Corrigés des exercices **7**

mathlvl.fr — Tous droits réservés. Toute reproduction, diffusion, revente ou exploitation commerciale, même partielle, est strictement interdite.
Tout contrevenant s'expose à des poursuites en vertu du Code de la propriété intellectuelle (art. L.335-2 et suivants).

1. Nuage de points et point moyen

Série à deux variables

On étudie deux caractères quantitatifs x et y sur n individus : couples $(x_1\,;y_1),\dots,(x_n\,;y_n)$. Dans un repère, les points $M_i(x_i\,;y_i)$ forment le **nuage de points**. Le **point moyen** est

$$G\,(\overline{x}\,;\overline{y}) \quad \text{avec} \quad \overline{x}=\frac{x_1+\cdots+x_n}{n} \quad \text{et} \quad \overline{y}=\frac{y_1+\cdots+y_n}{n}.$$

Heures de révision et note

Heures de révision x_i	1	2	3	4	5	6
Note obtenue y_i	7	9	10	12	13	15

$\overline{x}=\frac{21}{6}=3{,}5$ et $\overline{y}=\frac{66}{6}=11$: $G(3{,}5\,;11)$. Le nuage est presque aligné : un ajustement affine est pertinent.

2. La droite des moindres carrés

Ajustement affine

Ajuster le nuage par une droite $y=ax+b$, c'est choisir a et b pour que la droite passe « au plus près » des points. Pour chaque point, l'écart vertical est $y_i-(ax_i+b)$ (le **résidu**). La **droite des moindres carrés** (ou de régression de y en x) est celle qui rend minimale la somme des carrés des résidus :

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

Coefficients de la droite des moindres carrés

La droite des moindres carrés **passe par le point moyen** G et a pour coefficient directeur

$$\boxed{a=\frac{\operatorname{cov}(x,y)}{V(x)}} \quad \text{où} \quad \operatorname{cov}(x,y)=\frac{1}{n}\sum(x_i-\overline{x})(y_i-\overline{y}), \quad V(x)=\frac{1}{n}\sum(x_i-\overline{x})^2,$$

et $\boxed{b=\overline{y}-a\,\overline{x}}$.

Pourquoi la droite passe par G. Fixons a et regardons S comme une fonction de b. En notant $z_i=y_i-ax_i$, de moyenne $\overline{z}=\overline{y}-a\overline{x}$:

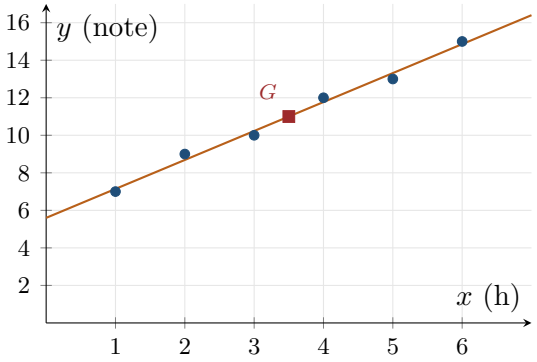
$$S(b)=\sum(z_i-b)^2=nb^2-2b\sum z_i+\sum z_i^2=nb^2-2nb\,\overline{z}+\sum z_i^2.$$

C'est un trinôme en b, de coefficient $n>0$: il est minimal pour $b=\dfrac{2n\overline{z}}{2n}=\overline{z}=\overline{y}-a\overline{x}$, c'est-à-dire quand la droite passe par G. On montre de même, en remplaçant b par $\overline{y}-a\overline{x}$, que S est un trinôme en a, minimal pour $a=\frac{\operatorname{cov}(x,y)}{V(x)}$.

mathlvl.fr — Tous droits réservés. Toute reproduction, diffusion, revente ou exploitation commerciale, même partielle, est strictement interdite. Tout contrevenant s'expose à des poursuites en vertu du Code de la propriété intellectuelle (art. L.335-2 et suivants).

Révisions (suite)

La calculatrice donne $a \approx 1{,}543$ et $b = 11 - 1{,}543 \times 3{,}5 \approx 5{,}6$: $y = 1{,}543x + 5{,}6$. Prévision pour 4,5 heures (interpolation) : $y \approx 12{,}5$. Chaque heure de révision supplémentaire est associée, en moyenne, à 1,5 point de plus.



Nuage, point moyen $G(3{,}5\,;11)$ et droite des moindres carrés $y = 1{,}543x + 5{,}6$.

Coefficient de corrélation linéaire

$$r = \frac{\operatorname{cov}(x,y)}{\sigma(x)\,\sigma(y)}, \qquad -1 \leq r \leq 1.$$

r a le signe de a. Plus $|r|$ est proche de 1, plus les points sont proches de la droite ; $|r| = 1$ si et seulement si ils sont alignés. On considère souvent l'ajustement affine comme justifié quand $|r| \geq 0{,}9$ environ. Pour les révisions, $r \approx 0{,}996$.

Loi d'Ohm

On mesure la tension U (en V) aux bornes d'un conducteur pour différentes intensités I (en A) :

I	0,1	0,2	0,3	0,4	0,5	0,6
U	2,1	4,0	6,2	7,9	10,1	12,0

La régression donne $U \approx 19{,}86\,I + 0{,}10$ avec $r \approx 0{,}9996$. L'ordonnée à l'origine est négligeable (erreurs de mesure) : on retrouve $U = RI$ avec une résistance $R \approx 20\ \Omega$.

3. Changement de variable

Se ramener à un ajustement affine

Si le nuage a une forme courbe, on cherche une transformation qui le redresse :

- croissance exponentielle $y = k\,\mathrm{e}^{ax}$: on pose $z = \ln y$, alors $z = ax + \ln k$ est affine en x ;
- autres modèles : $z = \sqrt{y}$, $z = \frac{1}{y}$, ou $X = x^2$, selon la forme du nuage.

On ajuste z en x, puis on revient à y.

mathlvl.fr — Tous droits réservés. Toute reproduction, diffusion, revente ou exploitation commerciale, même partielle, est strictement interdite. Tout contrevenant s'expose à des poursuites en vertu du Code de la propriété intellectuelle (art. L.335-2 et suivants).

Croissance d'une culture de bactéries

t (h)	0	1	2	3	4	5
N	100	150	220	340	500	760
$z = \ln N$	4,605	5,011	5,394	5,829	6,215	6,633

Le nuage $(t\,;z)$ est très bien aligné ($r \approx 0{,}9999$) : $z \approx 0{,}405t + 4{,}601$. Donc $N = \mathrm{e}^z \approx \mathrm{e}^{4{,}601} \times \mathrm{e}^{0{,}405t} \approx 99{,}6 \times 1{,}50^t$: la population est multipliée par 1,5 chaque heure. Extrapolation à $t = 7$ h : $N \approx 1\,700$.

La loi de Moore

En 1965, Gordon Moore remarque que le nombre de transistors par puce double environ tous les deux ans : en échelle logarithmique, les données de l'industrie s'alignent sur une droite pendant cinquante ans. Le même procédé sert pour la décroissance radioactive ou l'élimination d'un médicament.

4. Interpoler, extrapoler, et rester critique

Interpolation, extrapolation

Utiliser l'ajustement **à l'intérieur** de l'intervalle des données, c'est **interpoler** ; **au-delà**, c'est **extrapoler**. L'extrapolation suppose que la tendance continue : elle est d'autant plus risquée qu'on s'éloigne des données.

Une extrapolation absurde

Avec les révisions, $x = 10$ heures donnerait $y \approx 21$ sur 20 ! Le modèle affine n'est valable que dans le domaine observé. De même, les bactéries ne peuvent croître indéfiniment (espace, nourriture) : à long terme, un modèle logistique est plus réaliste. Quand la variable explicative est le temps (climat, démographie, économie), l'extrapolation est l'enjeu principal et doit être accompagnée d'hypothèses explicites.

Corrélation n'est pas causalité

Un coefficient $|r|$ proche de 1 indique que deux grandeurs varient **ensemble**, pas que l'une **cause** l'autre. Trois situations fréquentes :

- un **facteur commun** : les ventes de glaces et les noyades sont corrélées, car toutes deux augmentent l'été ; la pointure et le niveau de lecture des enfants sont corrélés, car tous deux dépendent de l'âge ;
- une causalité **inversée** : dans les villes, plus il y a de policiers, plus il y a de délits… parce qu'on affecte des policiers là où il y a des délits ;
- une **coïncidence** : deux séries qui croissent avec le temps sont toujours corrélées.

Établir une causalité demande une expérience contrôlée (essai clinique avec groupe témoin tiré au hasard) ou un mécanisme explicatif (loi physique, comme la loi d'Ohm).

mathlvl.fr — Tous droits réservés. Toute reproduction, diffusion, revente ou exploitation commerciale, même partielle, est strictement interdite. Tout contrevenant s'expose à des poursuites en vertu du Code de la propriété intellectuelle (art. L.335-2 et suivants).

5. Erreurs classiques

À éviter

- Oublier que la droite des moindres carrés passe par G : c'est un contrôle immédiat.
- Conclure à une relation affine avec $|r|$ faible, ou à l'absence de lien : $r \approx 0$ n'exclut pas une relation non affine (une parabole).
- Revenir à y en écrivant $y = \mathrm{e}^{ax} + \mathrm{e}^{b}$: c'est $y = \mathrm{e}^{b}\mathrm{e}^{ax}$.
- Extrapoler loin sans précaution.
- Confondre corrélation et causalité.

6. Fiche de synthèse

Point moyen	$G(\overline{x}\,;\overline{y})$
Moindres carrés	minimise $\sum(y_i - ax_i - b)^2$; $a = \frac{\mathrm{cov}(x,y)}{V(x)}$; $b = \overline{y} - a\overline{x}$ (passe par G)
Corrélation	$r = \frac{\mathrm{cov}(x,y)}{\sigma(x)\sigma(y)} \in [-1\,;1]$; ajustement pertinent si $\lvert r\rvert$ proche de 1
Changement de variable	$y = k\mathrm{e}^{ax}$: $z = \ln y$ affine en x, puis $y = \mathrm{e}^{z}$
Prévisions	interpoler (dans les données) ; extrapoler (au-delà) avec prudence
Causalité	corrélation $\neq$ causalité : facteur commun, causalité inversée, coïncidence

7. Exercices d'application

Exercice 1 – Émissions d'une usine

Les émissions de CO_2 d'une usine (en milliers de tonnes) sont :

Rang de l'année x	0	1	2	3	4	5
Émissions y	52	50	49	46	45	43

1. Calculer les coordonnées du point moyen G.
2. La calculatrice donne la droite $y = -1{,}8x + 52$ et $r \approx -0{,}993$. Vérifier qu'elle passe par G et interpréter le coefficient $-1{,}8$.
3. Estimer les émissions pour le rang 12. Quelle réserve formuler ?

Exercice 2 – Changement de variable

Un ajustement de $z = \ln y$ en x donne $z = -0{,}35x + 2{,}3$. Exprimer y en fonction de x sous la forme $y = k\,\mathrm{e}^{ax}$, puis estimer y pour $x = 4$.

Exercice 3 – Esprit critique

Une étude constate une forte corrélation entre le nombre de cafés bus par jour et le risque de maladie cardiaque. Peut-on conclure que le café est dangereux pour le cœur ? Proposer une explication alternative.

mathlvl.fr — Tous droits réservés. Toute reproduction, diffusion, revente ou exploitation commerciale, même partielle, est strictement interdite. Tout contrevenant s'expose à des poursuites en vertu du Code de la propriété intellectuelle (art. L.335-2 et suivants).

8. Corrigés des exercices

Corrigé de l'exercice 1

Corrigé. 1. $\overline{x} = 2{,}5$ et $\overline{y} = \frac{285}{6} = 47{,}5$: $G(2{,}5\,;47{,}5)$. 2. $-1{,}8 \times 2{,}5 + 52 = 47{,}5$: la droite passe par G. Les émissions baissent en moyenne de 1 800 tonnes par an. 3. $y = -1{,}8 \times 12 + 52 = 30{,}4$ milliers de tonnes, en supposant que la baisse se poursuit au même rythme, ce qui est une extrapolation lointaine (une baisse linéaire finirait par donner des émissions négatives).

Corrigé de l'exercice 2

Corrigé. $y = \mathrm{e}^{z} = \mathrm{e}^{2{,}3}\mathrm{e}^{-0{,}35x} \approx 9{,}97\,\mathrm{e}^{-0{,}35x}$. Pour $x = 4$: $y \approx 9{,}97\,\mathrm{e}^{-1{,}4} \approx 2{,}46$.

Corrigé de l'exercice 3

Corrigé. Non : une corrélation ne prouve pas une causalité. Un facteur commun peut l'expliquer : les grands buveurs de café fument peut-être davantage, dorment moins ou sont plus stressés, ce qui augmente le risque cardiaque. Seule une étude contrôlée (groupes comparables tirés au hasard) permettrait de conclure.

mathlvl.fr — Tous droits réservés. Toute reproduction, diffusion, revente ou exploitation commerciale, même partielle, est strictement interdite. Tout contrevenant s'expose à des poursuites en vertu du Code de la propriété intellectuelle (art. L.335-2 et suivants).