

Statistiques à deux variables

Prélude - Historique et contexte

En 1801, l'astronome Piazzi découvre l'astéroïde Cérès puis le perd de vue. Le jeune Gauss, à partir de quelques observations et de sa méthode des moindres carrés, prédit où le retrouver : Cérès réapparaît exactement à l'endroit annoncé. Legendre publie la méthode en 1805. Aujourd'hui, les ingénieurs l'utilisent pour étalonner un capteur, modéliser une consommation ou prévoir l'usure d'une pièce; c'est aussi la base de nombreux algorithmes d'apprentissage automatique.

I Nuage de points

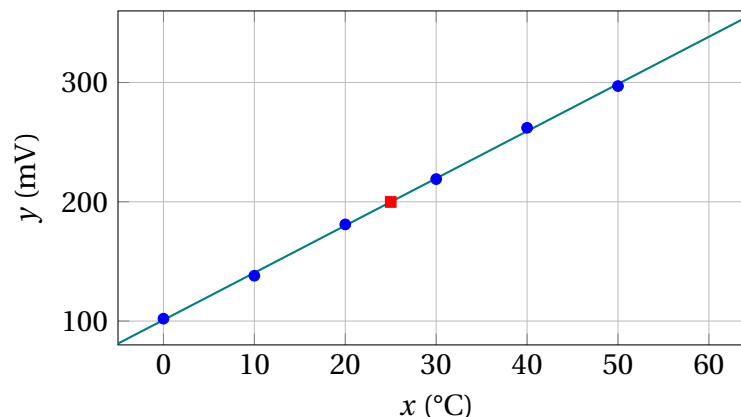
Définition

Une série statistique à deux variables est une liste de couples $(x_i ; y_i)$, $1 \leq i \leq n$. Leur représentation dans un repère est le **nuage de points**. Le **point moyen** est $G(\bar{x} ; \bar{y})$.

Exemple - Étalonnage d'un capteur de température

Température x_i (°C)	0	10	20	30	40	50
Tension y_i (mV)	102	138	181	219	262	297

$$\bar{x} = 25; \bar{y} = \frac{1199}{6} \approx 199,8.$$



II Ajustement affine

Définition - Méthode des moindres carrés

Pour une droite $y = ax + b$, on appelle **résidu** l'écart vertical $y_i - (ax_i + b)$. La **droite des moindres carrés** est la droite qui rend minimale la somme des carrés des résidus :

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

Elle passe par le point moyen G . Ses coefficients sont donnés par la calculatrice ou le tableur (fonctions DROITEREG, PENTE, ORDONNEE.OIGINE).

Exemple

Pour le capteur : $y \approx 3,96x + 100,9$. Estimation (interpolation) pour 25°C : $3,96 \times 25 + 100,9 \approx 199,9$ mV. Extrapolation pour 60°C : $\approx 338,5$ mV. Inversement, une tension de 240 mV correspond à $x = \frac{240-100,9}{3,96} \approx 35,1^\circ\text{C}$.

Remarque

Un ajustement peut aussi être fait « au jugé » (à la règle) en faisant passer une droite par G au plus près des points. L'extrapolation loin des données est risquée : le modèle peut ne plus être valable.

III Changement de variable

Méthode

Si le nuage n'a pas une forme rectiligne, on pose une nouvelle variable (par exemple $z = \ln y$, $z = \frac{1}{x}$, $z = \sqrt{x}$) pour obtenir un nuage $(x_i ; z_i)$ ou $(z_i ; y_i)$ allongé. On ajuste alors affinement, puis on revient aux variables initiales.

Exemple - Décharge d'un condensateur

t_i (ms)	0	2	4	6	8
U_i (V)	12	7,3	4,4	2,7	1,6
$z_i = \ln U_i$	2,48	1,99	1,48	0,99	0,47

Le nuage $(t_i ; z_i)$ est quasiment aligné. L'ajustement donne $z \approx -0,25t + 2,49$.

Donc $U = e^z \approx e^{2,49} \times e^{-0,25t} \approx 12e^{-0,25t}$: on retrouve une décroissance exponentielle.

Méthode - Automatiser les moindres carrés en Python

```
x = [0, 10, 20, 30, 40, 50]
y = [102, 138, 181, 219, 262, 297]
n = len(x)
xm, ym = sum(x)/n, sum(y)/n
a = sum((x[i]-xm)*(y[i]-ym) for i in range(n)) / sum((x[i]-xm)**2 for i in range(n))
b = ym - a*xm
print(a, b)    # environ 3.96 et 100.9
```

Remarque - Corrélation et causalité

Deux grandeurs peuvent être fortement corrélées sans que l'une soit la cause de l'autre : un troisième facteur peut les influencer toutes les deux.