

Statistiques à deux variables

Prélude - Historique et contexte

À la fin du XIX^e siècle, Francis Galton étudie la taille des enfants en fonction de celle de leurs parents et invente la notion de « régression ». Peu après, Karl Pearson formalise la corrélation. Croiser deux caractères permet de repérer des liens : genre et choix d'orientation, âge et revenus, émissions de gaz à effet de serre et température... Attention cependant : un lien statistique ne prouve pas une relation de cause à effet!

I Deux caractères qualitatifs : tableau croisé

Définition

Un **tableau croisé d'effectifs** (ou tableau à double entrée) donne la répartition d'une population selon deux caractères. On peut aussi le représenter par des diagrammes en barres ou circulaires. On le construit facilement avec un tableur (tableau croisé dynamique).

Exemple

Répartition de 400 élèves de première selon leur genre et leur spécialité préférée :

	SES	SVT	HGGSP	Total
Filles	80	70	60	210
Garçons	90	50	50	190
Total	170	120	110	400

Définition

- **Fréquence marginale** : total d'une ligne (ou d'une colonne) divisé par l'effectif total.
Exemple : fréquence des élèves préférant la SVT : $\frac{120}{400} = 0,3$.
- **Fréquence conditionnelle** : effectif d'une case divisé par le total de sa ligne ou de sa colonne.
Exemple : fréquence des élèves préférant la SVT parmi les filles : $\frac{70}{210} \approx 0,333$; parmi les garçons : $\frac{50}{190} \approx 0,263$.

II Deux caractères quantitatifs : nuage de points

Définition

On étudie deux caractères quantitatifs x et y sur une même population : on obtient une série de couples $(x_i ; y_i)$.

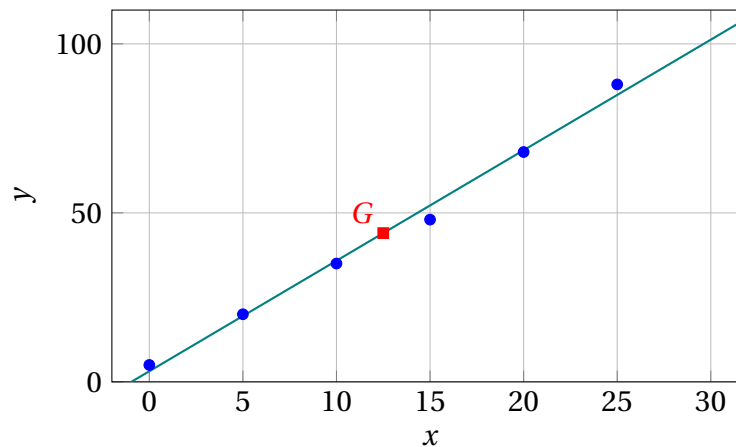
- Le **nuage de points** est l'ensemble des points $M_i(x_i ; y_i)$ placés dans un repère.
- Le **point moyen** du nuage est le point $G(\bar{x} ; \bar{y})$ où $\bar{x}$ et $\bar{y}$ sont les moyennes des x_i et des y_i .

Exemple

Niveau moyen des océans (en mm, par rapport à 1993) :

Rang de l'année x_i (0 = 1995)	0	5	10	15	20	25
Niveau y_i (mm)	5	20	35	48	68	88

$$\bar{x} = \frac{0 + 5 + 10 + 15 + 20 + 25}{6} = 12,5 \quad \text{et} \quad \bar{y} = \frac{5 + 20 + 35 + 48 + 68 + 88}{6} = 44. \text{ Donc } G(12,5 ; 44).$$



III Ajustement affine

Définition

Lorsque les points du nuage sont presque alignés, on peut les approcher par une droite d'équation $y = ax + b$: c'est un **ajustement affine**. Une « bonne » droite d'ajustement passe par le point moyen G (ou très près).

On peut la tracer « au jugé », ou la faire calculer par la calculatrice ou le tableur (droite des moindres carrés, appelée aussi droite de régression).

Exemple

Dans l'exemple précédent, le tableur donne la droite d'ajustement $y = 3,27x + 3,14$ (coefficients arrondis). On vérifie qu'elle passe par G : $3,27 \times 12,5 + 3,14 \approx 44$.

Définition

- **Interpoler**, c'est estimer une valeur **à l'intérieur** de la plage des données.
Exemple : en 2007 ($x = 12$), le niveau est estimé à $3,27 \times 12 + 3,14 \approx 42,4$ mm.
- **Extrapoler**, c'est estimer une valeur **en dehors** de la plage des données, en supposant que la tendance se poursuit.
Exemple : en 2035 ($x = 40$), le modèle prévoit $3,27 \times 40 + 3,14 \approx 134$ mm.

Remarque

Une extrapolation lointaine est risquée : rien ne garantit que le phénomène reste linéaire. Pour le niveau des océans, les scientifiques observent d'ailleurs une accélération.