Statistiques à deux variables

Prélude - Historique et contexte

À la fin du XIX^{e} siècle, Francis Galton étudie le lien entre la taille des parents et celle des enfants et invente la notion de « régression ». Karl Pearson, puis d'autres statisticiens, développent les outils qui permettent de résumer un nuage de points par une droite. On s'en sert pour étudier le lien entre deux grandeurs (dépenses publicitaires et ventes, âge et taille) et pour faire des prévisions.

I Nuage de points et point moyen

Définition

On étudie deux caractères quantitatifs x et y sur une même population : on obtient une **série statistique à deux variables** $(x_i ; y_i)$.
Dans un repère, l'ensemble des points $M_i(x_i ; y_i)$ est le **nuage de points** associé à la série.

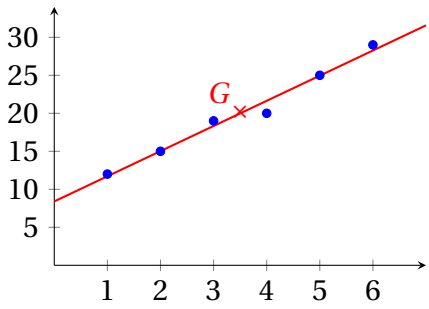
Définition - Point moyen

Le **point moyen** du nuage est le point $G(\overline{x};\overline{y})$, où $\overline{x}$ est la moyenne des x_i et $\overline{y}$ la moyenne des y_i.

Exemple

Dépenses publicitaires x_i (en milliers d'euros) et chiffre d'affaires y_i (en milliers d'euros) d'une entreprise sur 6 mois :

x_i	1	2	3	4	5	6
y_i	12	15	19	20	25	29



$\overline{x} = \dfrac{21}{6} = 3,5$ et $\overline{y} = \dfrac{120}{6} = 20$: $G(3,5;20)$.

II Ajustement affine

Définition

Lorsque les points du nuage sont presque alignés, on peut effectuer un **ajustement affine** : on remplace le nuage par une droite d'équation $y = ax + b$ qui passe « au plus près » des points.

Méthode - Droite passant par deux points

On peut choisir une droite passant par le point moyen G et par un autre point bien placé, puis déterminer son équation.

Exemple : droite passant par $G(3,5;20)$ et $A(1;12)$: $a = \dfrac{20-12}{3,5-1} = 3,2$ et $b = 12 - 3,2 \times 1 = 8,8$: $y = 3,2x + 8,8$.

Propriété - Droite des moindres carrés

La calculatrice ou le tableur donnent la droite d'ajustement de y en x par la **méthode des moindres carrés**. Elle passe toujours par le point moyen G.
Ici : $y \approx 3,31x + 8,4$ (coefficients arrondis au centième).

Méthode - Faire une prévision

On utilise l'équation de la droite d'ajustement :

- **interpolation** ou **extrapolation** : pour $x = 8$, $y \approx 3,31 \times 8 + 8,4 \approx 34,9$: un budget de 8 000 € de publicité laisse prévoir un chiffre d'affaires d'environ 34 900 € ;
- pour atteindre $y = 40$, on résout $3,31x + 8,4 = 40$: $x \approx 9,5$.

⚠ Une prévision n'a de sens que si l'on suppose que la tendance se maintient ; elle est d'autant moins fiable qu'on s'éloigne des données.