Statistiques à deux variables

Définition

Quand on étudie deux caractères statistiques sur une même population, on obtient une série statistique à deux variables.

Si les valeurs prises par le premier caractère sont $x_1, x_2, ..., x_n$ et celles prises par le second caractères sont $y_1, y_2, ..., y_n$, alors cette série est définie par le tableau ci-dessous :

x_i	x_1	x_2	...	x_n
y_i	y_1	y_2	...	y_n

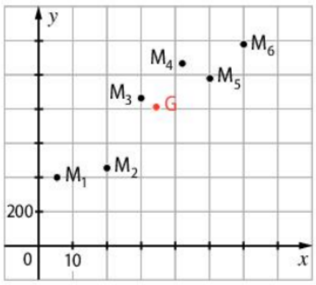
Exemple :
Une grande surface intéresse au lien entre ses dépenses publicitaires et son chiffre d'affaires. Les résultats sont donnés dans le tableau ci-dessous (en milliers d'euros).
Les deux caractères étudiés sont le montant des dépenses publicitaires (série des x_i) et le chiffre d'affaire (série des y_i).

Dépenses publicitaires x_i	5	20	30	42	50	60
Chiffre d'affaires y_i	400	460	870	1 070	980	1 170

Définition

Dans un repère du plan, l'ensemble des points $M_1(x_1; y_1), M_2(x_2; y_2), ..., M_n(x_n; y_n)$ constitue le nuage de points associé à cette série statistique.
Le point moyen du nuage de points de cette série statistique est le point $G(\overline{x}; \overline{y})$ où $\overline{x}$ est la moyenne de la série $x_1, x_2, ..., x_n$ et $\overline{y}$ est la moyenne de la série $y_1, y_2, ..., y_n$.

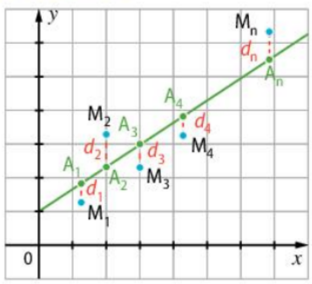
Exemple :
Avec les valeurs de l'exemple précédent, le nuage de points de la série statistique à deux variables est constitué des six points : $M_1(5; 40), M_2(20; 460), ..., M_6(60; 1170)$.
Le point moyen G du nuage a pour coordonnées $(34,5; 825)$.



Définition

Lorsque les points du nuage d'une série statistique à deux variables sont sensiblement alignés, on peut construire une droite passant au plus près de ces points.
On dit que cette droite réalise un ajustement affine du nuage de points.

Déterminer un ajustement affine par la méthode des moindres carrés consiste à déterminer la droite telle que la somme des carrés des écarts soit minimale.
Sur le graphique ci-contre, cette droite minimise la somme $A_1M_1^2 + A_2M_2^2 + ... + A_nMn^2$.



Propriété

La droite obtenue par la méthode des moindres carrés est la droite d'équation $y = ax + b$ avec $a = \dfrac{\text{cov}_{xy}}{\sigma_x^2}$.
La quantité notée cov_{xy} s'appelle la covariance de la série statistique à deux variables et a pour valeur $\text{cov}_{xy} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{i=n}(x_i - \overline{x})(y_i - \overline{y})$.

La quantité σ_x est l'écart type de la série $x_1, x_2, ..., x_n$ c'est à dire $\sigma_x = \sqrt{\dfrac{1}{n}\displaystyle\sum_{i=1}^{i=n}(x_i - \overline{x})}$.

Cette propriété se démontre et je me tiens à votre disposition pour les plus curieux.

> Dans la pratique la droite des moindres carrés sera souvent obtenue à l'aide de la calculatrice.
> Des fiches calculatrices sont disponibles dans votre livre aux pages 288 (CASIO) 290 (TI) et 292 (Num Works)

Dans notre exemple, la calculatrice indique : $\overline{x} = 34,5$ $\overline{y} = 825$ $a \approx 14,98$ et $b \approx 308$.

Définition

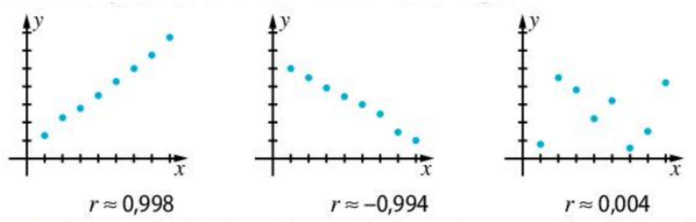
On appelle coefficient de corrélation linéaire de la série statistique à deux variables x et y le nombre r défini par $r = \dfrac{\text{cov}_{xy}}{\sigma_x \sigma_y}$.

Dans la pratique la calculatrice ou le tableur permet de calculer cette valeur sans connaître la formule.

Propriété

$-1 \leq r \leq 1$.
Il permet de mesurer la qualité de l'ajustement affine : plus il proche des valeurs extrêmes (-1 ou 1) plus la relation affine entre les variables est forte.
Plus sa valeur est proche de 0 plus la relation affine entre les variables est faible.

Si ce coefficient est proche de 1, les variables évoluent dans le même sens (la droite a un coefficient directeur positif).
Si ce coefficient est proche de -1, les variables évoluent en sens opposé (la droite a un coefficient directeur négatif).



Une corrélation forte n'est pas forcément la preuve d'une liaison de cause à effet entre les caractères étudiés.
Vous trouverez des exemples de corrélation fortes sans pour autant de liaison de cause à effet dans votre livre p.259

Exemple :
La calculatrice indique $r \approx 0,94$. La relation affine entre ces deux variables est donc forte et l'ajustement affine semble justifié.