

LEÇON N° 8 :

Séries statistiques à deux variables numériques. Nuage de points associé. Ajustement affine par la méthode des moindres carrés. Droites de régression. Applications. L'exposé pourra être illustré par un ou des exemples faisant appel à l'utilisation d'une calculatrice.

Pré-requis :

- Résultats sur les séries statistiques à une variable ;
- Trinôme du second degré (forme canonique, minimum) ;
- Equation d'une droite dans \mathcal{P} .

On se place dans un plan affine euclidien \mathcal{P} , rapporté à un repère orthogonal¹ (O, \vec{i}, \vec{j}) , de direction \vec{P} .

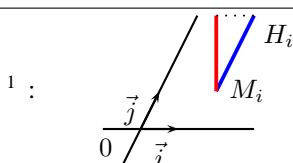
8.1 Séries statistiques à deux variables

Définition 1 : Soit $\Omega = \{\omega_1, \dots, \omega_n\}$ une population de taille $n \in \mathbb{N}^*$. On dit que deux variables X et Y définissent sur Ω une série statistique double $(x_i, y_i)_{1 \leq i \leq n}$, avec $X(\omega_i) = x_i$ et $Y(\omega_i) = y_i$, lorsque :

- $x_1 \leq \dots \leq x_n$;
- $X(\Omega)$ et $Y(\Omega)$ ne sont pas des singletons.

Conséquence - notation : Nous avons donc les résultats suivants (aussi valables en remplaçant X par Y et x_i par y_i), qui introduisent des notations utilisées dans la suite :

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n x_i \quad \text{et} \quad V(X) = \sigma_X^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{X})^2 = \overline{X^2} - \bar{X}^2.$$



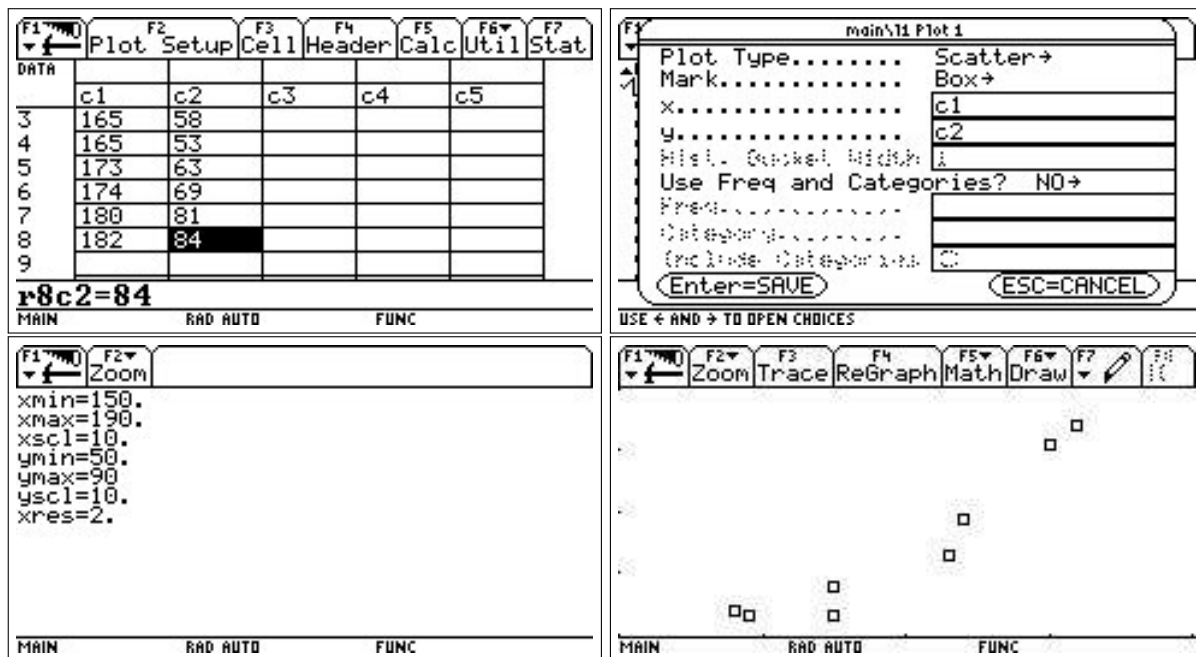
car $y_i - ax_i - b \neq M_i H_i$ si le repère n'est pas orthogonal. En effet, pour le calcul, $M_i H_i$ correspond à la longueur du segment bleu, et $y_i - ax_i - b$ correspond à la longueur du segment rouge. Cette notion d'orthogonalité doit être présente, et amène la variante de démonstration du théorème 1 présente en fin de leçon.

Séries statistiques à deux variables numériques

Exemple : On demande à 8 élèves de terminale leur taille (T) et leur poids (M) (ou plutôt... « masse » pour être physiquement exact !). Voici les résultats :

| Taille (cm) | 158 | 159 | 165 | 165 | 172 | 174 | 180 | 182 |
|-------------|-----|-----|-----|-----|-----|-----|-----|-----|
| Poids (kg) | 54 | 53 | 58 | 53 | 63 | 69 | 81 | 84 |

On entre (x_i) et (y_i) dans les deux premières colonnes de l'éditeur de listes de la calculatrice, et on fait tracer le nuage de points associé à cette série statistique double (attention à bien configurer la fenêtre graphique !!!) :



Cet exemple sera utilisé dans toute la suite de cette leçon.

Définition 2 : Dans (O, \vec{i}, \vec{j}) , on appelle *nuage de points* associé à la série statistique double $(x_i, y_i)_{1 \leq i \leq n}$ l'ensemble des points $M_i \in \mathcal{P}$ de coordonnées (x_i, y_i) . Le point de coordonnées (\bar{X}, \bar{Y}) est appelé *point moyen* et est noté \bar{G} .

Remarque 1 : \bar{G} est l'isobarycentre du système de points $\{M_i\}_{1 \leq i \leq n}$.

Définition 3 : Soit $(x_i, y_i)_{1 \leq i \leq n}$ une série statistique double. On appelle *covariance du couple* (X, Y) le réel noté $\text{Cov}(X, Y)$ ou $\sigma_{X,Y}$, égal à

$$\text{Cov}(X, Y) = \sigma_{X,Y} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y}).$$

Exemple : $\bar{G} = \left(\frac{1355}{8}, \frac{129}{2} \right) = (169, 375; 64, 5)$ et $\sigma_{T,M} = \frac{1503}{16} = 93, 9375$.

Proposition 1 :

(i) $\sigma_{X,Y} = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{X} \bar{Y}$;

(ii) Pour tous réels a, b, c, d , $\sigma_{aX+b, cY+d} = ac \sigma_{X,Y}$;

(iii) $|\sigma_{X,Y}| \leq \sigma_X \sigma_Y$, avec égalité si et seulement si les M_i sont tous alignés.

démonstration :

(i) Il suffit de faire quelques calculs pour démontrer cette égalité :

$$\begin{aligned} \sigma_{X,Y} &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y}) = \frac{1}{n} \sum_{i=1}^n (x_i y_i - \bar{Y} x_i - \bar{X} y_i + \bar{X} \bar{Y}) \\ &= \frac{1}{n} \sum_{i=1}^n x_i y_i - 2\bar{X} \bar{Y} + \bar{X} \bar{Y} = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{X} \bar{Y}. \end{aligned}$$

(ii) Il suffit à nouveau de faire des calculs, en utilisant le résultat précédent :

$$\begin{aligned} \sigma_{aX+b, cY+d} &= \frac{1}{n} \sum_{i=1}^n (ax_i + b)(cy_i + d) - \overline{aX + b} \overline{cY + d} \\ &= \frac{1}{n} \sum_{i=1}^n (acx_i y_i + adx_i + bcy_i + bd) - (a\bar{X} + b)(c\bar{Y} + d) \\ &= ac \frac{1}{n} \sum_{i=1}^n x_i y_i + ad\bar{X} + bc\bar{Y} + bd - ac\bar{X}\bar{Y} - ad\bar{X} - bc\bar{Y} - bd \\ &= ac \left(\frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{X} \bar{Y} \right) = ac \sigma_{X,Y}. \end{aligned}$$

(iii) Pour tout $\lambda \in \mathbb{R}$, on a $\sigma_{\lambda X+Y}^2 \geq 0$. Or $\sigma_{\lambda X+Y}^2 = \dots = \lambda^2 \sigma_X^2 + 2\lambda \sigma_{X,Y} + \sigma_Y^2$. Notons que $\sigma_X \neq 0$ car $X(\Omega)$ n'est par définition pas un singleton. Nous sommes donc en présence d'un trinôme du second degré qui est positif, son discriminant Δ est donc négatif, c'est-à-dire $\sigma_{X,Y}^2 - \sigma_X^2 \sigma_Y^2 \leq 0$, soit $|\sigma_{X,Y}| \leq \sigma_X \sigma_Y$.

De plus, $\sigma_{X,Y}^2 = \sigma_X^2 \sigma_Y^2 \Leftrightarrow \Delta = 0 \Leftrightarrow \exists \lambda_0 \in \mathbb{R} \mid \sigma_{\lambda_0 X+Y}^2 = 0$. Or $\sigma_{\lambda_0 X+Y}^2 = 0 \Leftrightarrow \dots \Leftrightarrow \forall i \in \{1, \dots, n\}, \lambda_0(x_i - \bar{X}) + (y_i - \bar{Y}) = 0 \Rightarrow \forall i, M_i(x_i, y_i) \in d$, où d est la droite d'équation $\lambda_0(x - \bar{X}) + (y - \bar{Y}) = 0$. Réciproquement, s'il existe une droite d'équation $y = ax + b$ telle que pour tout $i, y_i = ax_i + b$, alors $\bar{Y} = a\bar{X} + b$, et le calcul donne $\sigma_{X,Y}^2 = a^2 \sigma_X^2 = \sigma_X \sigma_Y$. ■

Remarque 2 : L'inégalité de (iii) porte généralement le nom d'*inégalité de Schwarz*.

8.2 Ajustement affine

On cherche une droite d'équation $y = ax + b$ qui approche au mieux tous les points du nuage d'une série statistique double. Soit $(x_i, y_i)_{1 \leq i \leq n}$ une telle série. Il existe alors plusieurs méthodes :

- manuelle : on trace une telle droite selon le « bon sens » sur le graphique, et l'on en déduit a et b .
- moyenne : il s'agit de calculer pour chaque sous-nuage les coordonnées du point moyen. On obtient donc un nouveau nuage de points : $\overline{G}_1, \overline{G}_2, \dots$ et l'on recommence avec ce nuage.
- des moindres carrés : c'est celle que l'on va développer ci-dessous.

On cherche a et b tels que $\varphi(a, b) = \sum_{i=1}^n (y_i - ax_i - b)^2$ soit minimale. Dans (O, \vec{i}, \vec{j}) , si l'on se donne la droite \mathcal{D} d'équation $y = ax + b$ et H_i le projeté de M_i parallèlement à l'axe (Oy) pour tout i entre 1 et n , alors on a

$$\varphi(a, b) = \sum_{i=1}^n (M_i H_i)^2.$$

Définition 4 : Si a et b minimisent φ , alors $\mathcal{D} : y = ax + b$ est la droite réalisant un ajustement affine du nuage de points selon la méthode des moindres carrés. On dit que \mathcal{D} est la droite de régression de Y en X .

Théorème 1 : Il existe une unique droite \mathcal{D} réalisant un ajustement affine du nuage de points selon la méthode des moindres carrés. Son coefficient directeur est $a = \sigma_{X,Y} / \sigma_X^2$ et elle passe par le point moyen. On a donc :

$$\mathcal{D} : y = \frac{\sigma_{X,Y}}{\sigma_X^2} x + (\overline{Y} - a\overline{X}).$$

démonstration : On a :

$$\begin{aligned} \sum_{i=1}^n (M_i H_i)^2 &= \sum_{i=1}^n (y_i - ax_i - b)^2 = \sum_{i=1}^n [(y_i - ax_i)^2 - 2b(y_i - ax_i) + b^2] \\ &= \sum_{i=1}^n y_i^2 - 2a \sum_{i=1}^n x_i y_i + a^2 \sum_{i=1}^n x_i^2 + n(b - (\overline{Y} - a\overline{X}))^2 - n(\overline{Y} - a\overline{X})^2 \\ &= \left(\sum_{i=1}^n y_i^2 - n\overline{Y}^2 \right) - 2a \left(\sum_{i=1}^n x_i y_i - n\overline{X}\overline{Y} \right) + a^2 \left(\sum_{i=1}^n x_i^2 - n\overline{X}^2 \right) + n(b - (\overline{Y} - a\overline{X}))^2 \\ &= n \left[(b - \overline{Y} + a\overline{X})^2 + \left(a\sigma_X - \frac{\sigma_{X,Y}}{\sigma_X} \right)^2 + \frac{1}{\sigma_X^2} (\sigma_Y^2 \sigma_X^2 - \sigma_{X,Y}^2) \right]. \end{aligned}$$

Or $\sigma_X^{-2} (\sigma_Y^2 \sigma_X^2 - \sigma_{X,Y}^2)$ est un nombre positif indépendant de a et b , donc

$$\sum_{i=1}^n (M_i H_i)^2 \geq \frac{1}{\sigma_X^2} (\sigma_Y^2 \sigma_X^2 - \sigma_{X,Y}^2),$$

avec égalité si et seulement si

$$\begin{cases} b - \overline{Y} + a\overline{X} = 0 \\ a\sigma_X - \frac{\sigma_{X,Y}}{\sigma_X} = 0 \end{cases} \Leftrightarrow \begin{cases} a = \frac{\sigma_{X,Y}}{\sigma_X^2} \\ b = \overline{Y} - a\overline{X}. \end{cases}$$

En fin de leçon est proposée une variante à cette démonstration. ■

Remarques :

1. D'après cette démonstration, $\sum_{i=1}^n M_i H_i = 0 \Leftrightarrow \sigma_X^{-2}(\sigma_X^2 \sigma_Y^2 - \sigma_{X,Y}^2) = 0 \Leftrightarrow |\sigma_{X,Y}| = \sigma_X \sigma_Y$, et l'on retrouve un résultat précédent ;
2. On peut aussi déterminer la droite \mathcal{D}' de régression de X en Y . Si l'on note $\mathcal{D}' : x = a'y + b'$, alors (en inversant les rôles de X et Y dans le théorème précédent), on a

$$a' = \frac{\sigma_{X,Y}}{\sigma_Y^2} \quad \text{et} \quad b' = \bar{X} - a'\bar{Y}.$$

Donc :

- * $\mathcal{D}' \ni \bar{G}$ car $\bar{X} = a'\bar{Y} + b'$.
- * Si $\sigma_{X,Y} = 0$, alors $a = a' = 0$, donc $\mathcal{D} \parallel (Ox)$ et $\mathcal{D}' \parallel (Oy)$.
- * Si $\sigma_{X,Y} \neq 0$, alors $a' \neq 0$ et donc

$$\mathcal{D} : y = \frac{1}{a'}x - \frac{b'}{a'}.$$

Alors $\mathcal{D} = \mathcal{D}' \Leftrightarrow a = 1/a' \Leftrightarrow \sigma_{X,Y}^2 = \sigma_X^2 \sigma_Y^2 \Leftrightarrow M_i$ alignés. Il est à noter que la condition $b = -b'/a'$ n'est pas utile puisque les deux droites passent par le point moyen.

- * a et a' ont même signe, celui de $\sigma_{X,Y}$, donc

$$\sigma_{X,Y}^2 \leq \sigma_X^2 \sigma_Y^2 \Rightarrow |a| \leq \left| \frac{1}{a'} \right|.$$

Définition 5 : On appelle *coefficient de corrélation linéaire entre X et Y* le réel noté R égal à

$$R = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}.$$

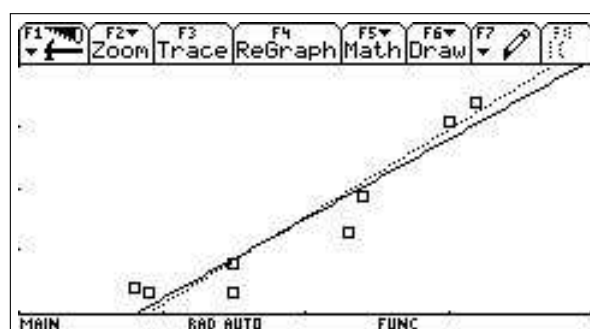
Remarque 4 :

- $-1 \leq R \leq 1$ (car $|\sigma_{X,Y}| \leq \sigma_X \sigma_Y$);
- Plus les points du nuage sont « alignés », plus $|R|$ sera proche de 1.

Exemple : On détermine dans notre exemple que

$$\mathcal{D} : y = 1,305x - 156,53 \quad \text{et} \quad \mathcal{D}' : y = 1,424x - 176,66,$$

ainsi que $R = 0,957$, d'où une bonne « corrélation » entre P et T . Voici la capture d'écran obtenue à la calculatrice (dans l'éditeur de liste, la possibilité de calculer l'équation d'une droite de régression et de la mémoriser dans une variable se fait *via* le menu F5) :



8.3 Applications

8.3.1 Ajustement par une fonction exponentielle

Si l'on a l'impression à la calculatrice que le nuage de points pourrait être approché par une fonction exponentielle, on détermine d'abord une droite de régression $y = mx + p = \ln(a)x + \ln(\lambda)$ (a et λ existent dans \mathbb{R}_+^* car $\ln : \mathbb{R}_+^* \mapsto \mathbb{R}$ est une bijection) du nuage de points associé à la série double $(x_i, \ln y_i)$. Alors le nuage de points initial est ajusté par $y = \exp(ax + b) = \lambda a^x$.

8.3.2 Ajustement par une fonction puissance

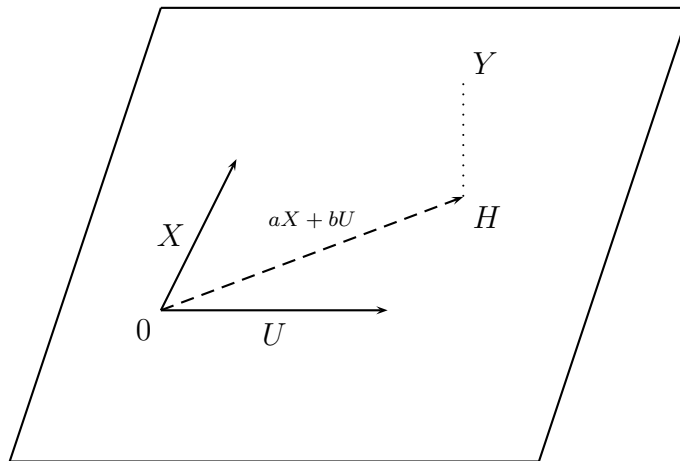
Si les points $M_i(x_i, y_i)$ sont proche de la courbe d'équation $y = \lambda x^a$, alors les points $(\ln x_i, \ln y_i)$ sont proches de la droite d'équation $y = ax + \ln \lambda$, et réciproquement.

8.3.3 Autres

Evolution (linéaire, exponentielle, ...) d'une statistique simple (par exemple une population, le tarif d'un produit, ...) en fonction du temps.

Variante de la démonstration du théorème 1

On pose $Y = (y_1, \dots, y_n)$, $X = (x_1, \dots, x_n)$, $U = (1, \dots, 1)$ et $\varphi(a, b) = \|Y - aX - bU\|_2^2$:



On cherche donc a et b tels que $aX + bU = \overrightarrow{OH}$. Sachant que $\overrightarrow{OY} - \overrightarrow{OH} = \overrightarrow{HY}$, on a

$$\begin{cases} (Y - aX - bU) \cdot X = 0 & (1) \\ (Y - aX - bU) \cdot U = 0 & (2) \end{cases}$$

(a, b) est unique par unicité du projeté orthogonal H de Y (si X et U sont non colinéaires, ce qui est exclu par le fait que $X(\Omega)$ n'est pas un singleton). Alors (2) $\Leftrightarrow n\bar{Y} - an\bar{X} - bn = 0 \Leftrightarrow b = \bar{Y} - a\bar{X}$ (ou encore

$\bar{Y} = a\bar{X} + b$, donc \bar{G} est sur la droite). On conclut ensuite avec l'équation (1) :

$$\begin{aligned}(1) &\Leftrightarrow \sum_{i=1}^n (y_i - ax_i - b) x_i = 0 \Leftrightarrow \sum_{i=1}^n x_i y_i - a \sum_{i=1}^n x_i^2 - b \sum_{i=1}^n x_i = 0 \\ &\Leftrightarrow n\bar{X}\bar{Y} - an\bar{X}^2 - (\bar{Y} - a\bar{X})n\bar{X} = 0 \\ &\Leftrightarrow a\bar{X}^2 - a\bar{X}^2 = \bar{X}\bar{Y} - \bar{X}\bar{Y} \\ &\Leftrightarrow a = \frac{\sigma_{X,Y}}{\sigma_X^2}.\end{aligned}$$