

SÉRIES STATISTIQUES À 2 VARIABLES

I) Rappel : nuage de points. Point moyen

On observe deux caractères (discrets) X et Y pour chaque individu d'une population (ou d'un échantillon).

On obtient une série statistique à 2 variables que l'on représente par un nuage de points qui est l'ensemble des points $M_i(x_i ; y_i)$ où x_i et y_i sont les différentes valeurs des variables X et Y .

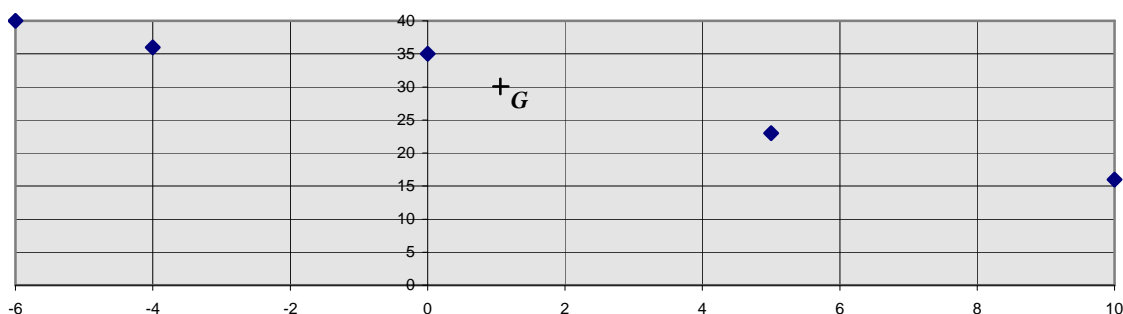
Exemple :

Le tableau ci-dessous donne la consommation quotidienne Y en fuel d'une chaudière (en litres) en fonction des relevés de température extérieure X .

X (en degré C)	$x_1 = -6$	$x_2 = -4$	$x_3 = 0$	$x_4 = 5$	$x_5 = 10$
Y (en litres)	$y_1 = 40$	$y_2 = 36$	$y_3 = 35$	$y_4 = 23$	$y_5 = 16$

On cherche un lien (s'il existe) entre la température extérieure X et la consommation quotidienne de fuel Y .

Nuage de points :



Point moyen du nuage : c'est le point $G(\bar{x} ; \bar{y})$ où $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ et $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$

Dans notre cas, on a $G(1 ; 30)$.

Lorsque les points du nuage sont dans une situation de proche alignement, on peut chercher une droite traduisant cet alignement. Une telle droite s'appelle "droite d'ajustement". On va voir, ci-dessous, qu'on peut déterminer plusieurs droites d'ajustement. (On connaît déjà, par exemple, la droite de Mayer)

II) Les droites de régression

Imaginons que deux élèves aient tracé, à tâtons, des droites d'ajustement. N'y en a-t-il pas une meilleure que l'autre ? Comment "mesurer" la qualité de l'ajustement ?

1) Somme des résidus associée à une droite d'équation $y = ax + b$:

Dans tout ce qui suit, on suppose que le nuage contient au moins deux points d'abscisses différentes. (Si tous les points du nuage ont la même abscisse, ils sont alors alignés sur une droite verticale). On a donc $\sigma_x \neq 0$.

Considérons la somme des carrés des écarts "verticaux" séparant les points $(x_i ; y_i)$ et $(x_i ; ax_i + b)$.

Il s'agit de la quantité $\sum_{i=1}^n [y_i - (ax_i + b)]^2$ qu'on appelle somme des résidus (quadratiques).

Exemple :

Imaginons, que pour l'exemple ci-dessus, on ait tracé la droite d'ajustement d'équation $y = -1,5x + 31,5$.

Calculons la somme des résidus associée à cette droite :

X (en degré C)	$x_1 = -6$	$x_2 = -4$	$x_3 = 0$	$x_4 = 5$	$x_5 = 10$
Y (en litres)	$y_1 = 40$	$y_2 = 36$	$y_3 = 35$	$y_4 = 23$	$y_5 = 16$
$-1,5x_i + 31,5$	40,5	37,5	31,5	24	16,5
$ y_i - (-1,5x_i + 31,5) $	0,5	1,5	3,5	1	0,5
$[y_i - (-1,5x_i + 31,5)]^2$	0,25	2,25	12,25	1	0,25

En additionnant les nombres de la dernière ligne, on obtient la somme des résidus : $\sum_{i=1}^n [y_i - (ax_i + b)]^2 = 16$.

2) Méthode des moindres carrés :

On considère souvent que plus la somme des résidus (qui est une somme de carrés d'écarts) est petite, plus la droite d'ajustement est bonne.

On appelle droite de régression de y en x la droite obtenue lorsque la somme des résidus est MINIMALE.

Voici une méthode pour minimiser la somme des résidus :

Posons $S(a, b) = \sum_{i=1}^n [y_i - ax_i - b]^2$ cette somme des résidus (c'est une fonction des deux variables a et b)

Introduisons une nouvelle variable (afin d'alléger les calculs) : $z = y - ax - b$.

Ainsi, $S(a, b) = \sum_{i=1}^n z_i^2$.

Or, nous savons que : $V(z) = \frac{1}{n} \sum_{i=1}^n (z_i - \bar{z})^2 = \frac{1}{n} \sum_{i=1}^n z_i^2 - \bar{z}^2$ où, par linéarité de la moyenne, $\bar{z} = \bar{y} - a\bar{x} - b$.

Minimiser $S(a, b)$ revient à minimiser $\sum_{i=1}^n z_i^2 = n(V(z) + \bar{z}^2)$.

Or, minimiser la somme $nV(z) + n\bar{z}^2$ revient à minimiser chacun des deux termes (puisque tous deux positifs)

• Minimisation de $nV(z)$:

On a : $z_i - \bar{z} = y_i - ax_i - b - (\bar{y} - a\bar{x} - b) = (y_i - \bar{y}) - a(x_i - \bar{x})$

D'où :

$$nV(Z) = \sum_{i=1}^n (z_i - \bar{z})^2 = \sum_{i=1}^n [(y_i - \bar{y}) - a(x_i - \bar{x})]^2 = \sum_{i=1}^n [(y_i - \bar{y})^2 - 2a(x_i - \bar{x})(y_i - \bar{y}) + a^2(x_i - \bar{x})^2]$$

$$nV(z) = \sum_{i=1}^n (y_i - \bar{y})^2 - 2a \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) + a^2 \sum_{i=1}^n (x_i - \bar{x})^2$$

Posons $\text{cov}(x; y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$. Cette quantité s'appelle la covariance de X et Y encore notée σ_{xy} .

On a finalement (après simplification par $n > 0$) :

$$V(z) = V(x) a^2 - 2\sigma_{xy} a + V(y)$$

On reconnaît un trinôme du second degré en a . Écrivons-le sous forme canonique :

$$V(z) = \left(\sigma_x a - \frac{\sigma_{xy}}{\sigma_x} \right)^2 + V(y) - \left(\frac{\sigma_{xy}}{\sigma_x} \right)^2 = \left(\sigma_x a - \frac{\sigma_{xy}}{\sigma_x} \right)^2 + \frac{V(x)V(y) - \sigma_{xy}^2}{V(x)}$$

Ainsi, $V(z)$ est minimal lorsque $\left(\sigma_x a - \frac{\sigma_{xy}}{\sigma_x} \right)^2 = 0$, c'est-à-dire : $a = \frac{\sigma_{xy}}{V(x)}$.

$$\text{(Et le minimum de } V(z) \text{ est } \frac{V(x)V(y) - \sigma_{xy}^2}{V(x)})$$

- Minimisation de \bar{z}^2 :

On a $\bar{z} = \bar{y} - a\bar{x} - b$. Donc \bar{z} est minimal si $b = \bar{y} - a\bar{x}$. (Et le minimum de \bar{z} est 0)

BILAN :

La droite de régression de y en x a pour équation $y = ax + b$ où $a = \frac{\sigma_{xy}}{V(x)}$ et $b = \bar{y} - a\bar{x}$.

Conséquence : la droite de régression passe par le point moyen du nuage.

En effet, c'est une conséquence immédiate de la relation $\bar{y} = a\bar{x} + b$.

Exemple : déterminons une équation de la droite de régression de y en x pour nos 5 relevés concernant notre chaudière.

$$\text{Calcul de la variance de } X : V(x) = \frac{1}{n} \sum_{i=1}^n z_i^2 - \bar{z}^2 = \frac{1}{5} (36 + 16 + 0 + 25 + 100) - 1 = 34,4.$$

$$\text{Calcul de la covariance de } X \text{ et } Y : \sigma_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x}\bar{y} \text{ (exercice : démontrer cette}$$

dernière égalité)

$$\sigma_{xy} = \frac{1}{5} (-6 \times 40 - 4 \times 36 + 0 \times 35 + 5 \times 23 + 10 \times 16) - 1 \times 30 = -51,8.$$

$$\text{D'où } a = \frac{\sigma_{xy}}{V(x)} = -\frac{51,8}{34,4} \simeq -1,506 \text{ (à } 10^{-3} \text{ près)}$$

$$\text{Enfin, } b = \bar{y} - a\bar{x} = 30 + \frac{51,8}{34,4} \times 1 \simeq 31,506 \text{ (à } 10^{-3} \text{ près)}$$

La droite de régression de y en x a pour équation : $y = -1,506x + 31,506$ (à 10^{-3} près)

Pour finir, calculons la somme des résidus pour la droite de régression : il s'agit de la quantité :

$$n \frac{V(x)V(y) - \sigma_{xy}^2}{V(x)} = 5 \times \frac{34,4 \times 81,2 - 51,8^2}{34,4} = 15,99 \text{ (à } 10^{-2} \text{ près)}$$

Remarque : on peut également s'intéresser à la somme des écarts "horizontaux" (dans le cas où $\sigma_y \neq 0$).

Considérons la droite d'équation $x = a'y + b'$; la somme des résidus est :

$$\sum_{i=1}^n [x_i - (a'y_i + b')]^2$$

La droite pour laquelle cette somme est minimale s'appelle droite de régression de x en y .

On démontre, comme ci-dessus, qu'il suffit de choisir : $a' = \frac{\sigma_{xy}}{V(y)}$ et $b' = \bar{x} - a'\bar{y}$

En général, les deux droites de régressions sont différentes.

III) Coefficient de corrélation linéaire

On suppose dans ce paragraphe que les points du nuage ne sont pas tous alignés sur une même droite verticale, ni sur une même droite horizontale. On a donc $\sigma(x) \neq 0$ et $\sigma(y) \neq 0$.

Lorsqu'on a recherché la droite de régression de y en x , on a démontré que la somme des résidus minimale était :

$$n \frac{V(x)V(y) - \sigma_{xy}^2}{V(x)}$$

Comme la somme des résidus est positive, on a : $V(x)V(y) \geq \sigma_{xy}^2$

Comme $\sigma(x) \neq 0$ et $\sigma(y) \neq 0$, on a donc :

$$-1 \leq \frac{\sigma_{xy}}{\sigma_x \sigma_y} \leq 1$$

Le réel $r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$ s'appelle coefficient de corrélation linéaire entre les deux variables X et Y .

Si $r^2 = 1$ alors la somme des résidus est nulle et les points du nuage sont alignés. (Les deux droites de régressions sont alors confondues)

Si r est proche de 1 ou -1 alors la somme des résidus est proche de 0. On dit alors que la corrélation entre X et Y est forte.

Dans la pratique, on considère qu'une corrélation est forte lorsque $r \geq 0,95$. Dans le cas contraire un ajustement affine n'est pas conseillé. (Attention, d'autres critères comme $r^2 \geq \frac{3}{4}$ existent mais ne sont pas tous pertinents)

Exemple : dans l'exemple précédent, calculer le coefficient de corrélation linéaire :

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{51,8}{\sqrt{34,4} \sqrt{81,2}} \approx 0,98 \text{ (à } 10^{-2} \text{ près)}$$

L'ajustement affine est donc amplement justifié dans ce cas.

Formule donnant la somme des résidus quadratiques S de la droite de régression de y en x :

On a vu que dans ce cas :

$$S = n \frac{V(x)V(y) - \sigma_{xy}^2}{V(x)}$$

D'où

$$S = nV(y) \left(1 - \frac{\sigma_{xy}^2}{V(x)V(y)} \right) = nV(y)(1 - r^2)$$

On démontre, de même, que pour la droite de régression de x en y , on a : $S = nV(x)(1 - r^2)$

SÉRIES STATISTIQUES À 2 VARIABLES : RÉSUMÉ

Données : deux caractères X et Y dont on dispose n relevés généralement présentés sous forme de tableau :

X	x_1	x_2	x_3	\dots	x_n
Y	y_1	y_2	y_3	\dots	y_n

Ce que l'on peut vous demander de faire (ou de calculer) :

1) **Nuage de points** : dans un repère, c'est l'ensemble des points $M_i(x_i; y_i)$ pour $1 \leq i \leq n$.

2) **Point moyen d'un nuage** : c'est le point $G(\bar{x}; \bar{y})$ où $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ et $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$.

3) **Calcul du coefficient de corrélation linéaire** : $r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$ où :

$$* \sigma_{xy} \text{ désigne la covariance de } X \text{ et } Y : \sigma_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x}\bar{y}$$

$$* \sigma_x \text{ est l'écart type de } X : \sigma_x = \sqrt{V(x)} \text{ avec } V(x) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2$$

$$* \sigma_y \text{ est l'écart type de } Y : \sigma_y = \sqrt{V(y)} \text{ avec } V(y) = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 = \frac{1}{n} \sum_{i=1}^n y_i^2 - \bar{y}^2$$

Si $r \geq 0,95$ alors la corrélation linéaire entre X et Y est **forte**. Un ajustement affine est alors **justifié**.

(Les points du nuage sont dans une situation de proche alignement).

Si $r^2 = 1$ (c'est-à-dire $r = 1$ ou $r = -1$) alors les points du nuage sont **alignés**.

4) **Déterminer un ajustement affine** (uniquement dans le cas où $r \geq 0,95$) :

* Droite de régression de y en x par la méthode des moindres carrés :

$$\text{c'est la droite d'équation } y = ax + b \text{ avec : } a = \frac{\sigma_{xy}}{V(x)} \text{ et } b = \bar{y} - a\bar{x}$$

* Droite de régression de x en y (peu utilisée) par la méthode des moindres carrés :

$$\text{c'est la droite d'équation } x = a'y + b' \text{ avec : } a' = \frac{\sigma_{xy}}{V(y)} \text{ et } b' = \bar{x} - a'\bar{y}$$

Les droites de régression passent toujours par le point moyen du nuage.

Instruction calculatrice : LINREG

5) **Déterminer un autre type d'ajustement** (quadratique, exponentielle ou autre ...)

Dans ce cas, l'énoncé vous guide à l'aide d'un changement de variable ($z_i = \sqrt{y_i}$ ou $z_i = \ln y_i$ ou autre ...)

afin de se ramener à un ajustement affine entre X et Z .

6) **Calculer la somme des résidus quadratiques d'une droite d'ajustement** :

Soit D une droite d'ajustement d'équation $y = ax + b$. Sa somme des résidus quadratiques de y en x est :

$$S = \sum_{i=1}^n [y_i - (ax_i + b)]^2 \text{ (somme des carrés des "écarts verticaux" entre les points du nuage et la droite } D)$$

Dans le cas de la droite de régression de y en x : $S = nV(y)(1 - r^2)$.