PR13

Statistiques à deux variables et régression linéaire

Version du 27-01-2023 à 10:34

1. Séries statistiques à deux variables et nuage de points

Définition 1 – Regroupement de données

Soient X et Y deux caractères quantitatifs définis sur une population Ω.

L'étude d'un échantillon de taille n donne n couples de valeurs prises par le couple (X,Y) qui constitue la série statistique du couple (X,Y).

Cette série peut être présentée de deux façons :

Données groupées

Individu	1	...	i	...	n
X	x_1	...	x_i	...	x_n
Y	y_1	...	y_i	...	y_n

où l'on omet parfois la première ligne.

Données groupées

X \ Y	y_1	...	y_i	...	y_s
x_1	$n_{1,1}$	...	$n_{1,j}$	...	$n_{1,s}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$
x_i	$n_{i,1}$	...	$n_{i,j}$	...	$n_{i,s}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$
x_r	$n_{r,1}$	...	$n_{r,i}$	...	$n_{r,s}$

$n_{i,j}$ est l'effectif du couple (x_i,y_j), c'est à dire le nombre d'individus de l'échantillon pour lesquels $X=x_i$ et $Y=y_j$, et on a clairement $\sum_{i=1}^{r}\sum_{j=1}^{s} n_{i,j} = n$.

Ce tableau constitue ainsi un tableau des effectifs.

Utilisation de fréquences

La fréquence du couple (x_i,y_j) est $f_{i,j} = \dfrac{n_{i,j}}{n}$ et on a $\sum_{i=1}^{r}\sum_{j=1}^{s} f_{i,j} = 1$.

Lorsque l'on remplace les effectifs $n_{i,j}$ par les fréquences $f_{i,j}$ dans ce tableau, on a un tableau des fréquences.

Analogie avec les couples de variables aléatoires

- on notera l'analogie avec la loi conjointe d'un couple de variables aléatoires réelles (X, Y), $f_{i,j}$ jouant le rôle de $p_{i,j}$.
- il est équivalent de donner le tableau des effectifs $n_{i,j}$ ou le tableau des fréquences $f_{i,j} = \dfrac{n_{i,j}}{n}$.

□

Définition 2 – Nuage de points

Dans un repère orthogonal, à chaque individu de l'échantillon, on associe le point $M(x, y)$, x et y étant les valeurs prises par X et Y pour cet individu.

Si (X, Y) prend la valeur (x, y) pour m individus, alors on dessine m points autour du point de coordonnées (x, y) ou un disque d'aire proportionnelle à m.

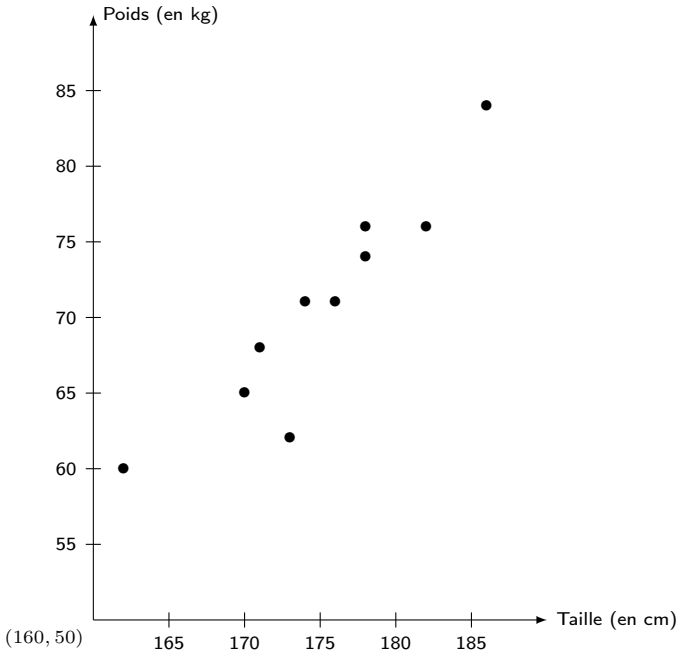
On obtient ainsi le nuage de points représentant la série statistique.

Illustration

Pour un groupe de 10 personnes, on considère la série statistique à deux variables T et P qui, à chaque personne i, associe sa taille T_i (en cm) et son poids P_i (en kg). On récapitule les résultats bruts de ce relevé dans le tableau ci-dessous :

i	1	2	3	4	5	6	7	8	9	10
T_i	174	182	170	176	171	178	173	178	186	162
P_i	71	76	65	71	68	76	62	74	84	60

On représente alors dans un repère les points de coordonnées (T_i, P_i), et on parle alors de nuage de points associé à la série à deux variables.



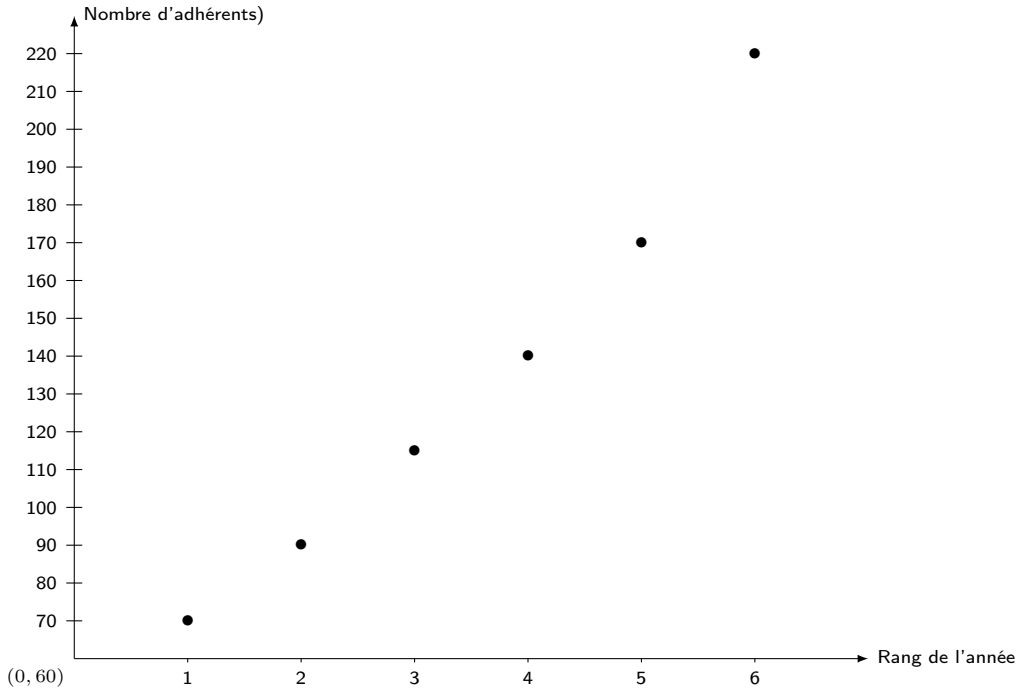
□

Exemple 1 – Exploitation d'un nuage de points

Le tableau suivant donne l'évolution du nombre d'adhérents d'un club de rugby de 2001 à 2006.

Année	2001	2002	2003	2004	2005	2006
Rang x_i	1	2	3	4	5	6
Nombre d'adhérents y_i	70	90	115	140	170	220

Que peut-on prévoir pour l'évolution du nombre d'adhérents pour les années suivantes ?



Le nuage de points associé à une série statistique à deux variables donne donc immédiatement des informations de nature qualitatives.

Le tracé met en évidence la possibilité de « reconnaître » graphiquement une éventuelle relation fonctionnelle entre les deux grandeurs observées (ici rang et nombre d'adhérent).
Le problème de l'établissement d'une telle relation entre les deux séries est le problème de l'ajustement.

□

Définition 3 – Paramètres statistiques

Soient X et Y deux caractères quantitatifs définis sur une population Ω pour lesquels on connaît les distributions marginales :

Individu	1	...	i	...	n
X	x_1	...	x_i	...	x_n
Y	y_1	...	y_i	...	y_n

Moyennes et point moyen

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$
$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Le point $G\left(\overline{X}, \overline{Y}\right)$ est appelé point moyen du nuage.

Écart-type et formule de Huygens

$$\sigma_X^2 = \underbrace{\frac{1}{n}\sum_{i=1}^{n} x_i^2}_{\text{Moyenne de la série statistique } X^2} - \overline{X}^2$$

$$\sigma_Y^2 = \underbrace{\frac{1}{n}\sum_{i=1}^{n} y_i^2}_{\text{Moyenne de la série statistique } Y^2} - \overline{Y}^2$$

Covariance et coefficient de corrélation linéaire

$$\sigma_{XY} = \underbrace{\frac{1}{n}\sum_{i=1}^{n} x_i y_i}_{\substack{\text{Moyenne de la série} \\ \text{statistique } X\times Y}} - \overline{X} \times \overline{Y}$$

$$\begin{aligned}\rho &= \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \\ &= \frac{\displaystyle\sum_{i=1}^{n} x_i y_i - n\overline{XY}}{\sqrt{\left(\displaystyle\sum_{i=1}^{n} x_i^2 - n\overline{X}^2\right)\left(\displaystyle\sum_{i=1}^{n} y_i^2 - n\overline{Y}^2\right)}}\end{aligned}$$

□

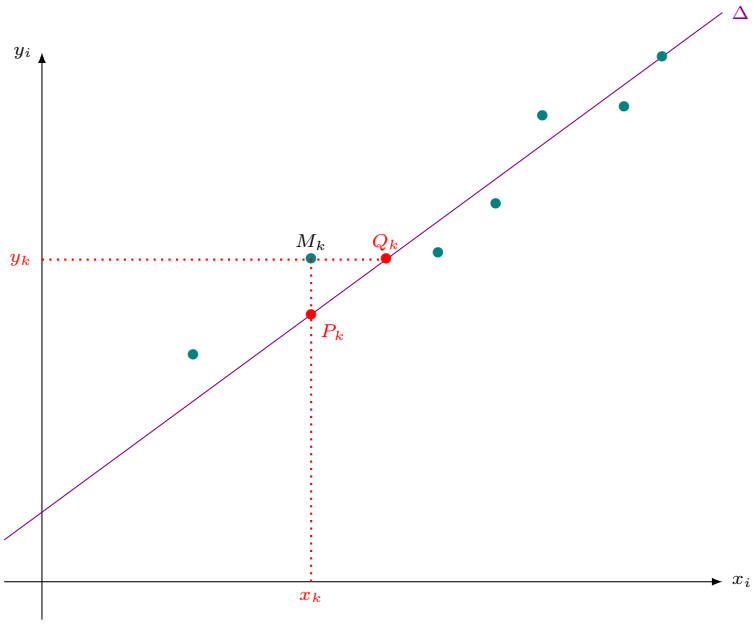
2. Ajustement par la méthode des moindres carrés

Introduction – Objectif

Lorsque le nuage de points représentant la distribution (X, Y) est approximativement rectiligne, on cherche l'équation d'une droite qui « ajuste » le nuage.

On dit que l'on effectue un ajustement linéaire.
On utilise pour cela la méthode des moindres carrés expliquée ci-dessous.

Soit Δ la droite d'équation $y = ax + b$.
À tout point M_k du nuage de points, on associe les points P_k et Q_k projections de M_k sur Δ parallèlement respectivement à l'axe des ordonnées et l'axe des abscisses.



Les points M_k, P_k et Q_k sont donc de coordonnées respectives :

$$M_k\,(x_k, y_k),\ P_k\,(x_k, ax_k + b) \text{ et si } a \neq 0, Q_k\left(\frac{y_k - b}{a}, y_k\right)$$

On cherche alors a et b tels que la somme $S\,(a, b) = \displaystyle\sum_{k=1}^{n} (M_k P_k)^2$ soit minimale.

$$S\,(a, b) = \sum_{k=1}^{n} (y_k - ax_k - b)^2$$

□

Théorème 1 – Droite de régression de y en x

La droite Δ cherchée a pour équation : $y - \overline{Y} = \dfrac{\sigma_{XY}}{\sigma_X^2}\left(x - \overline{X}\right)$

Δ est appelée droite de régression de y en x ou droite des moindres carrés de y en x.

Point moyen et droite de régression

Δ passe par le point moyen du nuage.

Éléments de preuve utilisant une étude de fonctions

Le théorème nous dit que la droite recherchée a donc une équation de la forme $y = \alpha\left(x - \overline{X}\right) + \overline{Y}$. Il s'agit donc de minimiser l'expression $f(\alpha) = \sum_{i=1}^{n}\left(y_i - \overline{Y} - \alpha\left(x_i - \overline{X}\right)\right)^2$ compte-tenu des éléments présentés en introduction.

On étudie donc la fonction $f : \alpha \longmapsto \sum_{i=1}^{n}\left(y_i - \overline{Y} - \alpha\left(x_i - \overline{X}\right)\right)^2$.

Cette dernière est clairement définie, continue et dérivable sur $\mathbb{R}$, et on a :

$$\begin{aligned}
\forall \alpha \in \mathbb{R},\ f'(\alpha) &= \sum_{i=1}^{n} -2\left(x_i - \overline{X}\right)\left(y_i - \overline{Y} - \alpha\left(x_i - \overline{X}\right)\right) \\
&= \sum_{i=1}^{n} -2\left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right) + 2\alpha\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2 \\
&= \underbrace{2\alpha\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2}_{\geq 0} - 2\sum_{i=1}^{n}\left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right)
\end{aligned}$$

où l'on reconnaît ici l'expression d'une fonction affine croissante qui s'annule en $\alpha_0 = \dfrac{\sum_{i=1}^{n}\left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right)}{\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2}$.

On en déduit ainsi le tableau de variation de f sur $\mathbb{R}$:

α	$-\infty$		α_0		$+\infty$
Signe de $f'(\alpha)$		$-$	0	$+$	
Variations de f		↘	$f(\alpha_0)$	↗	

et ainsi f est minimale en α_0.

Par ailleurs, on a : $\alpha_0 = \dfrac{\sum_{i=1}^{n}\left(x_iy_i - x_i\overline{Y} - y_i\overline{X} + \overline{X}\times\overline{Y}\right)}{\sum_{i=1}^{n}\left(x_i^2 - 2x_i\overline{X} + \overline{X}^2\right)}$

$$= \frac{\sum_{i=1}^{n}x_iy_i - \overline{Y}\sum_{i=1}^{n}x_i - \overline{X}\sum_{i=1}^{n}y_i + n\overline{X}\times\overline{Y}}{\sum_{i=1}^{n}x_i^2 - 2\overline{X}\sum_{i=1}^{n}x_i + n\overline{X}^2}$$

$$= \frac{n\left(\frac{1}{n}\sum_{i=1}^{n}x_iy_i - \overline{Y}\times\frac{1}{n}\sum_{i=1}^{n}x_i - \overline{X}\times\frac{1}{n}\sum_{i=1}^{n}y_i + \overline{X}\times\overline{Y}\right)}{n\left(\frac{1}{n}\sum_{i=1}^{n}x_i^2 - 2\overline{X}\times\frac{1}{n}\sum_{i=1}^{n}x_i + \overline{X}^2\right)}$$

$$= \frac{\frac{1}{n}\sum_{i=1}^{n}x_iy_i - \overline{Y}\times\overline{X} - \overline{X}\times\overline{Y} + \overline{X}\times\overline{Y}}{\frac{1}{n}\sum_{i=1}^{n}x_i^2 - 2\overline{X}\times\overline{X} + \overline{X}^2}$$

$$= \frac{\sum_{i=1}^{n}x_iy_i - \overline{X}\times\overline{Y}}{\frac{1}{n}\sum_{i=1}^{n}x_i^2 - \overline{X}^2}$$

$$= \frac{\sigma_{XY}}{\sigma_X^2}$$

Éléments de preuve utilisant le calcul différentiel

Si S présente un minimum global en (a_0, b_0), alors S présente un minimum local en (a_0, b_0).

Cherchons donc si S admet un minimum local. On a : $S(a,b) = \sum_{k=1}^{n}(y_k - ax_k - b)^2$ pour tout $(a,b)\in\mathbb{R}^2$

S est une fonction polynomiale des deux variables a et b, donc S est de classe $\mathcal{C}^2$sur $\mathbb{R}^2$.

$$\frac{\partial S}{\partial a}(a,b) = \sum_{k=1}^{n} -2x_k(y_k - ax_k - b) \quad \text{et} \quad \frac{\partial S}{\partial b}(a,b) = \sum_{k=1}^{n} -2(y_k - ax_k - b)$$

Donc : $\begin{cases}\dfrac{\partial S}{\partial a}(a,b) = 0\\ \dfrac{\partial S}{\partial b}(a,b) = 0\end{cases} \Leftrightarrow \begin{cases}\sum_{k=1}^{n}x_ky_k - a\sum_{k=1}^{n}x_k^2 - b\sum_{k=1}^{n}x_k = 0\\ \sum_{k=1}^{n}y_k - a\sum_{b=1}^{n}x_k - nb = 0\end{cases}$

$$\Leftrightarrow \begin{cases}a\sum_{k=1}^{n}x_k^2 + b\sum_{k=1}^{n}x_k = \sum_{k=1}^{n}x_ky_k\\ a\sum_{k=1}^{n}x_k + nb = \sum_{k=1}^{n}y_k\end{cases}$$

Or $\sum_{k=1}^{n}x_k = n\overline{X}$ et $\sum_{k=1}^{n}y_k = n\overline{Y}$, ainsi que $\sum_{k=1}^{n}x_k^2 = n\left(\sigma_X^2 + \overline{X}^2\right)$ et $\sum_{k=1}^{n}x_ky_k = n\left(\sigma_{XY} + \overline{XY}\right)$.

On en déduit que : $\begin{cases}\dfrac{\partial S}{\partial a}(a,b) = 0\\ \dfrac{\partial S}{\partial b}(a,b) = 0\end{cases} \Leftrightarrow \begin{cases}a\left(\sigma_X^2 + \overline{X}\right) + b\overline{X} = \sigma_{XY} + \overline{XY}\\ a\overline{X} + b = \overline{Y}\end{cases}$

$$\Leftrightarrow \begin{cases}b = \overline{Y} - a\overline{X}\\ a\sigma_X^2 = \sigma_{XY}\end{cases}$$

$$\Leftrightarrow \begin{cases}a = \dfrac{\sigma_{XY}}{\sigma_X^2}\\ b = \overline{Y} - a\overline{X}\end{cases}$$

Et donc S admet un seul point critique $(a_0, b_0) = \left(\dfrac{\sigma_{XY}}{\sigma_X^2}, \overline{Y} - \dfrac{\sigma_{XY}}{\sigma_X^2}\overline{X}\right)$.

Par ailleurs :

$$\frac{\partial^2 S}{\partial a^2}(a,b) = \sum_{k=1}^{n} 2x_k^2 \quad \frac{\partial^2 S}{\partial a \partial b}(a,b) = \sum_{k=1}^{n} 2x_k, \quad \frac{\partial^2 S}{\partial b^2}(a,b) = 2n$$

Donc :

$$r = 2\sum_{k=1}^{n} x_k^2 > 0 \quad \text{et} \quad rt - s^2 = 4n\sum_{k=1}^{n} x_k^2 - 4\left(\sum_{k=1}^{n} x_k\right)^2 = 4n\sigma_X^2 > 0$$

S présente donc un minimum local en (a_0, b_0).
Reste à montrer qu'en (a_0, b_0), S présente un minimum global.
On se place alors dans $\mathbb{R}^n$ et on considère les vecteurs $e = (1, \ldots, 1)$, $x = (x_1, \ldots, x_n)$ et $y = (y_1, \ldots, y_n)$.
La fonction $S(a,b)$ s'interprête donc comme étant le carré de la distance entre les deux vecteurs de $\mathbb{R}^n$ que sont y et $ax + be$.
Ainsi, les valeurs du couple (a,b) qui minimisent $S(a,b)$ sont celles qui correspondant à l'unique vecteur $ax + be$ du sous-espace vectoriel $E = \text{Vect}\ (e,x)$ de $\mathbb{R}^n$ tel que le vecteur $y - (ax + be)$ est orthogonal à E.
On cherche donc (a,b) tel que l'on ait : $\begin{cases} \langle y - (ax+be) \,|x\rangle = 0 \\ \langle y - (ax+be) \,|e\rangle = 0 \end{cases}$
En utilisant la bilinéarité du produit scalaire, et en résolvant le système 2×2 d'inconnue (a,b), on montre alors que :

$$a = \frac{\langle e\,|e\rangle \langle x\,|y\rangle - \langle x\,|e\rangle \langle y\,|e\rangle}{\langle e\,|e\rangle \langle x\,|x\rangle - \langle x\,|e\rangle^2} \text{ et } b = \frac{\langle y\,|e\rangle - a\langle x\,|e\rangle}{\langle e\,|e\rangle}$$

ce qui donnera par la suite les coefficients de l'équation réduite de Δ, puisque :

$$\langle x\,|x\rangle = \sum_{i=1}^{n} x_i^2, \quad \langle x\,|y\rangle = \sum_{i=1}^{n} x_i y_i, \quad \langle y\,|y\rangle = \sum_{i=1}^{n} y_i^2, \quad \langle x\,|e\rangle = \sum_{i=1}^{n} x_i, \quad \langle y\,|e\rangle = \sum_{i=1}^{n} y_i, \quad \langle e\,|e\rangle = n$$

□

Remarque 1 – Droite de régression de x en y

La droite Δ' qui minimise $\displaystyle\sum_{k=1}^{n} (M_k Q_k)^2$ cherchée a pour équation : $x - \overline{X} = \dfrac{\sigma_{XY}}{\sigma_Y^2}(y - \overline{Y})$

Δ est appelée droite de régression de x en y ou droite des moindres carrés de x en y.

Point moyen et droite de régression

Δ' passe par le point moyen du nuage.

□

Définition 4 – Coefficient de corrélation linéaire

Le coefficient de corrélation linéaire d'une série statistique de variables X et Y est le nombre r défini par $r = \dfrac{\sigma_{XY}}{\sigma_X \times \sigma_Y}$.

Interprétation

Ce coefficient sert à mesurer la qualité d'un ajustement affine.

En effet, on peut montrer que le minimum de la quantité $\displaystyle\sum_{i=1}^{n} P_i M_i^2$ vaut $n\sigma_Y^2\left(1 - \left(\dfrac{\sigma_{XY}}{\sigma_X \sigma_X}\right)^2\right)$ et donc plus ce minimum est petit, plus la droite de régression est proche du nuage de points.

Ainsi, plus le coefficient de régression linéaire est proche de 1 en valeur absolue, meilleur est l'ajustement linéaire. Lorsque $r = \pm 1$, la droite de régression passe par tous les points du nuage, qui sont donc alignés.

□