Séries statistiques à deux variables

Table des matières

I Séries statistiques à deux variables

I.1 Nuages de points

Définition

On considère deux variables statistiques x et y observées sur une même population de n individus.
On note x_1, x_2,...,x_n les valeurs relevées pour la variable x et y_1, y_2,...,y_n relevées pour la variable y.
Les couples $(x_1\ ;\ y_1)$, $(x_2\ ;\ y_2)$, ..., $(x_n\ ;\ y_n)$ forment une série statistique à deux variables.
On appelle nuage de points les points de coordonnées $(x_i\ ;\ y_i)$ pour $1 \leqslant i \leqslant n$.

Dans ce chapitre, on va s'intéresser au lien qui peut exister entre ces deux variables.

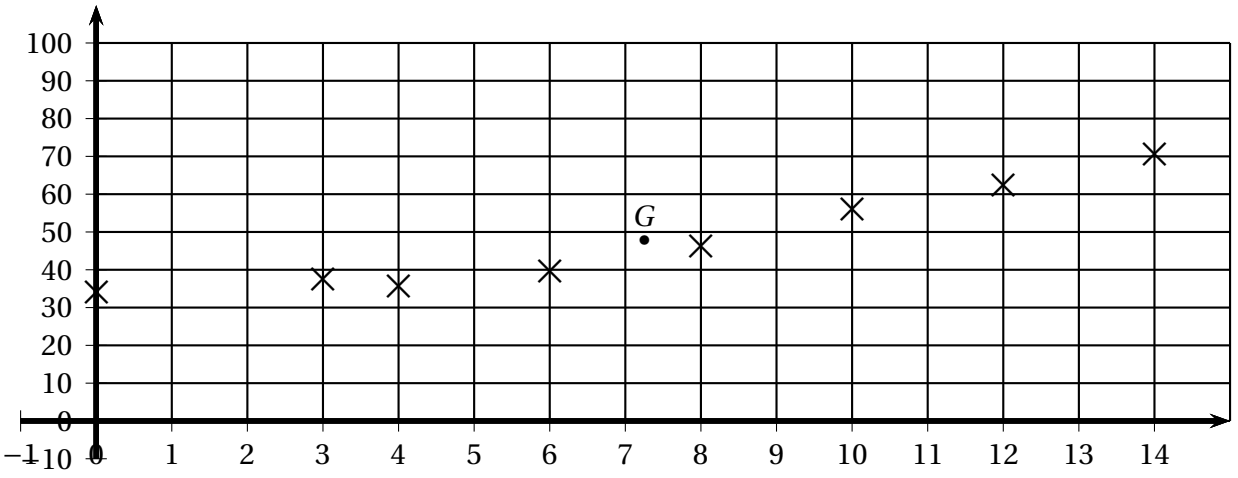
Exemple :

Le tableau ci-dessous indique les pourcentages d'accès au niveau baccalauréat d'une génération d'élèves.

Année x_i	1980	1982	1984	1986	1988	1990	1992	1994
Rang i	0	2	4	6	8	10	12	14
Taux d'accès au niveau baccalauréat y_i	34 %	37,5 %	35,8 %	39,8 %	46,3 %	56,1 %	62,5 %	70,7 %

Source : d'après un document du Ministère de l'éducation nationale

Représentons le nuage de points correspondants.



On constate que ces points sont **à peu près** alignés.

On va essayer de trouver une droite qui passe **au plus près** de ces points.

I.2 Point moyen

Définition

Soient x et y deux séries statistiques prenant chacune n valeurs.
On considère le nuage de points $M_1\left(x_1\,;\,y_1\right), \ldots M_n\left(x_n\,;\,y_n\right)$.
On note $\overline{x}$ la moyenne des abscisses et $\overline{y}$ la moyenne des ordonnées.
On appelle point moyen le point $G\left(\overline{x}\,;\,\overline{y}\right)$.

Par exemple, dans l'exemple précédent : $G(7.25\,;\,47,8375)$

II Ajustement affine

On a vu dans l'exemple précédent que les points sont presque alignés.
On va chercher une droite qui passe près de tous les points.

II.1 Notion d'ajustement affine

Définition

Dans un nuage de points, rechercher une droite qui « approche au mieux » tous les points du nuage d'appelle réaliser un **ajustement affine**.
La droite trouvée s'appelle **droite de régression**.

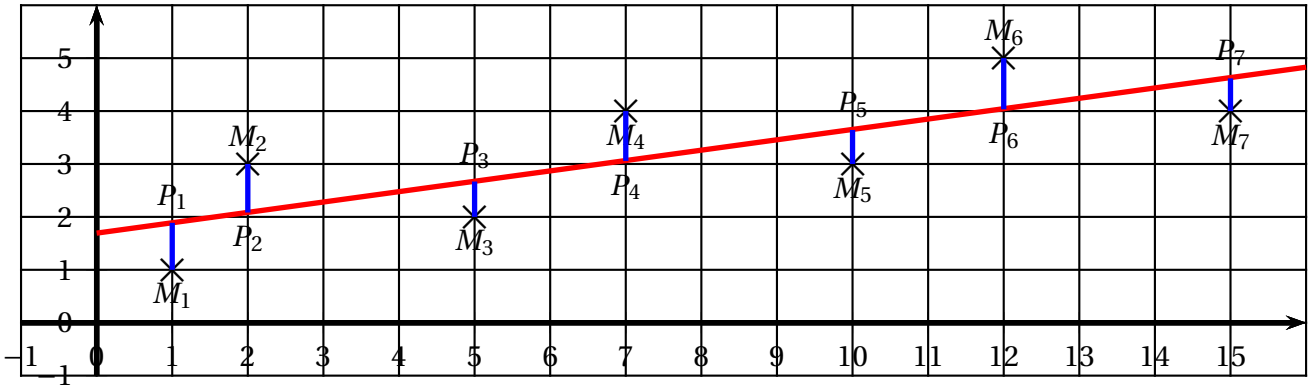
Remarque :

- « approcher au mieux » est très subjectif
- Le point moyen appartient à toute droite de régression (admis).

II.2 Droite d'ajustement obtenu par la méthode des moindres carrés

Propriété (admise)

Soit un nuage de points M_1, M_2, …, M_n.
On considère toutes les droites d qui passent par le point moyen.
On considère les points P_1, P_2, …, P_n de d qui ont les mêmes abscisses que les points M_1, M_2, …, M_n.
Il existe une unique droite d qui minimise $M_1P_1^2 + M_2P_2^2 + \cdots + M_nP_n^2$.
Cette droite d'ajustement linéaire s'appelle droite de régression de y en x obtenue par la méthode des moindres carrés.



Définition et propriété (admise)

On appelle covariance des variables x (prenant les valeurs x_1, x_2, …x_n) et y (prenant les valeurs y_1, y_2, …, y_n) le nombre :

$$\text{cov}(x\,;\,y) = \frac{(x_1-\overline{x})(y_1-\overline{y})+(x_2-\overline{x})(y_2-\overline{y})+\cdots+(x_n-\overline{x})(y_n-\overline{y})}{n}$$

Remarque : on a aussi :

$$\text{cov(x; y)} = \frac{x_1y_1+x_2y_2+\cdots++x_ny_n}{n} - \overline{xy}$$

La droite d de régression de y en x obtenue par la méthode des moindres carrés a pour équation : $y = mx+p$ avec $m = \dfrac{\text{cov}(x\,;\,y)}{\sigma^2(x)}$ et $p = \overline{y} - m\overline{x}$, où $\sigma(x)$ est l'écart-type de la série statistique et $\overline{x}$ et $\overline{y}$ sont les coordonnées du point moyen.

Remarque : ces nombres s'obtiennent facilement avec une calculatrice graphique.

III Coefficient de corrélation linéaire

Quand les points d'un nuage sont plus ou moins alignés, comment décider si un ajustement linéaire est pertinent ou pas ?

Définition

Soit une série statistique à deux variables x et y.
On appelle coefficient de corrélation linéaire le nombre r défini par

$$r = \frac{\text{cov}(x\,;\,y)}{\sigma(x)\sigma(y)}$$

où $\sigma(x)$ et $\sigma(y)$ sont les écart-types des séries x et y.

Propriété

r est un nombre compris entre -1 et 1.

Propriété

Lorsque r est « proche » de -1 ou de 1 (c'est-à-dire $r \leqslant -0,75$ ou $r \leqslant 0,75$), on dit que la corrélation linéaire entre x et y est forte ; on peut alors envisager un ajustement affine.

Remarque :

- Ne pas confondre corrélation et causalité.
- Si r est proche de 0, un ajustement affine n'est pas pertinent ; il se peut qu'un ajustement par un autre type de courbe soit possible (fonction exponentielle, parabole, …)