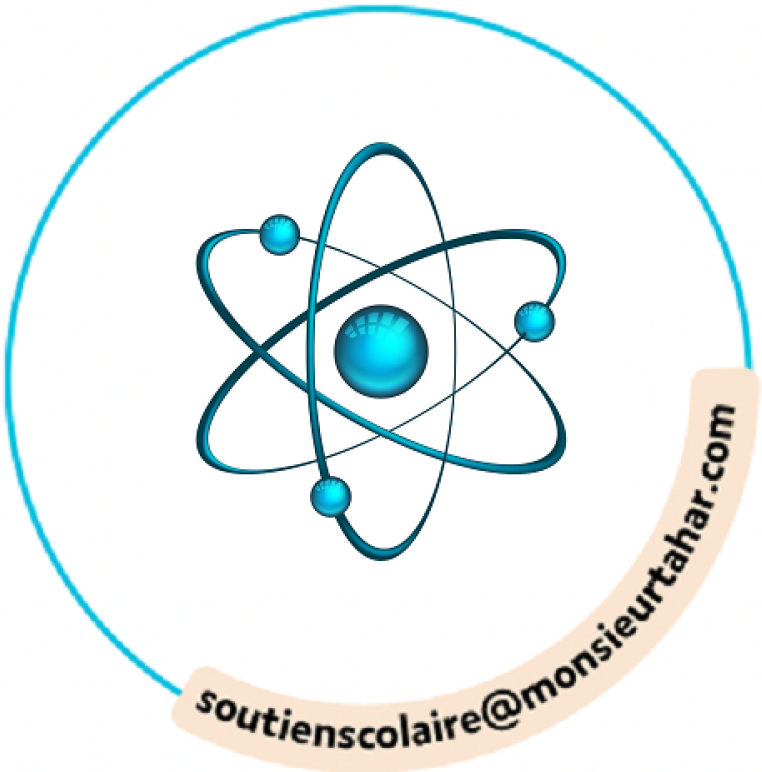
MATHEMATIQUES



CHAPITRE 9

Séries statistiques à deux variables

1. Séries statistiques doubles

1. Tableaux de données

Dans certaines situations d'étude de populations, il semble exister un lien entre deux caractères de la population. Par exemple, le poids et la taille d'un nouveau-né, la résistance thermique d'un mur et son épaisseur ou le montant des salaires d'une entreprise et son chiffre d'affaires sont sûrement liés. On ne peut cependant pas affirmer que ce lien est un lien de cause à effet (causalité).
Pour étudier simultanément deux caractères quantitatifs d'une même population, on introduit deux variables x et y prenant respectivement les valeurs du premier et du deuxième caractère.

Définition

Sur une population d'effectif n, on étudie **simultanément deux variables statistiques** x et y. Pour chaque individu i, avec $1 \leqslant i \leqslant n$, on mesure la valeur x_i de la variable x et la valeur y_i de la variable y.
Les couples $(x_1 ; y_1), (x_2 ; y_2), \ldots, (x_n ; y_n)$ forment une **série statistique double** ou **série statistique à deux variables x et y.**
Généralement, on présente cette série sous forme d'un tableau.

variable x	x_1	x_2	…	x_n
variable y	y_1	y_2	…	y_n

Exemple

Le tableau ci-dessous donne, pour huit villes des États-Unis, le nombre moyen de jours d'ensoleillement dans l'année, ainsi que la température mensuelle moyenne en °C.

Villes	Phenix	Miami	San Diego	Sacramento	Las Vegas	Denver	San Francisco	Oklahoma City
Ensoleillement x_i	161	131	127	150	159	129	158	128
Températures y_i	21,5	24,3	17,2	15,6	18,8	10,1	14,1	15,7

Population : les 8 villes, donc $n = 8$; variable x : nombre de jours d'ensoleillement ; variable y : température annuelle moyenne.

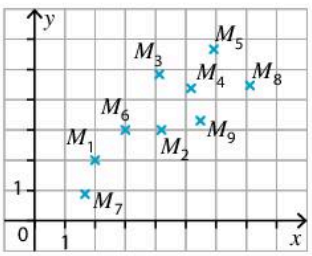
Remarque

L'un des deux caractères étudiés peut être des dates ou des années.

2. Nuage de points

Définition

On représente une série statistique à deux variables x et y par un **nuage de points** dans un repère $(O ; I, J)$, constitué des points $M_i(x_i ; y_i)$, x_i et y_i étant respectivement les valeurs de x et y pour $i = 1 ; \ldots ; n$.



3. Point moyen

Définition

Soient x et y deux variables statistiques prenant chacune n valeurs.
On appelle **point moyen** du nuage de points $M_1(x_1 ; y_1)$, $M_2(x_2 ; y_2)$, …, $M_n(x_n ; y_n)$ le point $G(\overline{x} ; \overline{y})$, où $\overline{x}$ est la moyenne arithmétique des valeurs de la variable x et $\overline{y}$ la moyenne arithmétique des valeurs de la variable y.

On a donc $G\left(\dfrac{x_1 + x_2 + \ldots + x_n}{n} ; \dfrac{y_1 + y_2 + \ldots + y_n}{n}\right)$.

Méthode 1 Modéliser par une série statistique à deux variables

Parmi les situations suivantes, dire si l'étude de la série statistique à deux variables semble pertinente ou non.

1. Le tableau ci-dessous donne le prix de différentes croisières d'une semaine en bateau et l'âge des capitaines des bateaux.

Nom du bateaux	Sea week	La traversée	Sun boat	L'espoir	Mer Azur	Trans océan	Medit-sun	Nord Mer
Prix x_i en €	768	892	1 230	750	899	1 599	1 250	799
Age y_i	53	57	48	59	44	45	52	50

2. Le tableau ci-dessous donne le poids moyen d'un nouveau-né, chaque mois lors de la première année de sa vie.

Nombre de mois x_i après la naissance	1	2	3	6	8	10	12
Poids y_i en kg.	3,6	3,8	4,3	4,9	5,4	6,2	6,8

Solution commentée

1. Cette étude ne semble pas pertinente car le prix d'une croisière n'a a priori rien à voir avec l'âge du capitaine d'un bateau.

2. Cette étude semble pertinente pour donner une référence aux parents de nouveaux-nés au sujet de la croissance de leur enfant lors de sa première année.

Méthode 2 Tracer un nuage de points

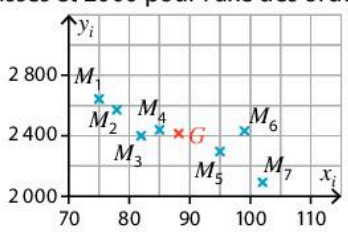
En prévision du lancement d'un nouveau produit, une société a effectué une enquête auprès de ses clients potentiels pour déterminer le futur prix de vente du produit. Les résultats sont donnés dans le tableau ci-dessous.

Prix x_i en €	75	78	82	85	95	99	102
Nombres y_i d'acheteurs potentiels	2 650	2 580	2 400	2 450	2 300	2 440	2 099

1. Tracer le nuage de points de cette série statistique à deux variables x et y.
2. Déterminer le point moyen du nuage de points et le placer sur le graphique.

Solution commentée

1. On représente les points de coordonnées $M_i(x_i\ ;\ y_i)$ pour $i = 1, \ldots, 7$ dans un repère. Pour faciliter le graphique, on commence les graduations sur chaque axe à une valeur proche des valeurs minimales de chaque série (70 pour l'axe des abscisses et 2000 pour l'axe des ordonnées).



2. $\overline{x} = \dfrac{75+78+82+85+95+99+102}{7} = 88$

$\overline{y} = \dfrac{2\,650+2\,580+2\,400+2\,450+2\,300+2\,440+2\,099}{7} = 2\,417$

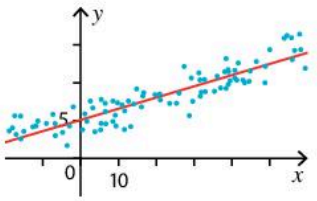
2. Ajustement affine

1. Notion d'ajustement affine

Définition

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

La droite trouvée s'appelle un ajustement affine du nuage de points ou une **droite de régression**.



Remarques

- Dire que la droite « approche au mieux » le nuage de points est très subjectif.
- Si le nuage de points possède des points qui paraissent globalement alignés, trouver un ajustement affine semble assez naturel.

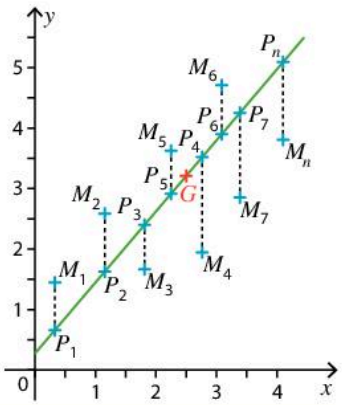
Propriété (admise)

Dans tout ajustement affine, le point moyen G appartient à la droite de régression.

2. Droite des moindres carrés

Propriété (admise)

Soit un nuage de points $M_1, M_2, \ldots, M_n$. On considère toutes les droites d passant par le point moyen G, et on note $P_1, P_2, \ldots, P_n$ les points de d qui ont les mêmes abscisses que $M_1, M_2, \ldots, M_n$.
Il existe une unique droite d qui minimise $M_1P_1^2 + M_2P_2^2 + \ldots + M_nP_n^2$.
Cette droite est un ajustement affine qui s'appelle la **droite des moindres carrés** ou la **droite de régression de y en x**.



Définition et propriété (admise)

On appelle **covariance des variables x** (prenant les valeurs $(x_1\ ;\ x_2\ ;\ \ldots\ ;\ x_n)$ et y (prenant les valeurs $(y_1\ ;\ y_2\ ;\ \ldots\ ;\ y_n)$, le nombre :

$$\text{cov}(x;y) = \frac{(x_1-\overline{x})(y_1-\overline{y}) + (x_2-\overline{x})(y_2-\overline{y}) + \ldots + (x_n-\overline{x})(y_n-\overline{y})}{n} = \frac{x_1y_1 + x_2y_2 + \ldots + x_ny_n}{n} - \overline{x}\,\overline{y}$$

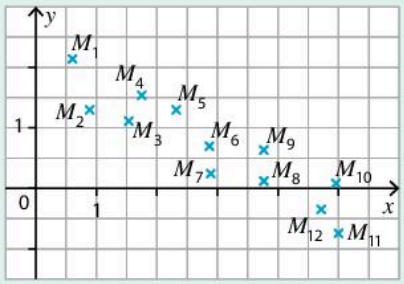
La droite d des moindres carrés (appelée aussi droite de régression de y en x) a pour équation $y = mx + p$ avec $m = \dfrac{\text{cov}(x\ ;\ y)}{\sigma^2(x)}$ et $p = \overline{y} - m\overline{x}$, où $\sigma(x)$ est l'écart type de la série statistique x et $(\overline{x}\ ;\overline{y})$ sont les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

Remarque

En pratique, on obtient la droite des moindres carrés en utilisant une calculatrice ou un ordinateur.

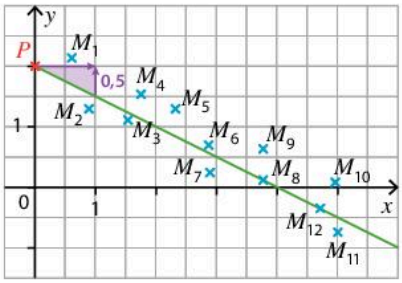
Méthode 1 Réaliser un ajustement affine « au jugé »

On considère le nuage de points ci-contre.



1. Tracer un ajustement affine qui semble approcher au mieux le nuage de points.
2. Lire graphiquement l'équation de la droite de régression ainsi tracée.

Solution commentée



1. On trace une droite en essayant d'approximer au mieux le nuage de point.
2. On lit $m = -\frac{1}{2}$ et $p = 2$.

La droite de régression ainsi tracée a pour équation $y = -\frac{1}{2}x + 2$.

Remarque : Lorsqu'on cherche un ajustement affine « au jugé », plusieurs droites peuvent convenir.

Méthode 2 Déterminer l'équation de la droite des moindres carrés

On considère la série statistique à deux variables x et y suivantes.

x_i	2	2,5	3	4	5,5	7	8,5	9	10	10,5
y_i	5	8	10	15	18	25	32	45	50	62

• Déterminer par le calcul l'équation de la droite des moindres carrés. Arrondir les calculs à 10^{-3}.

Solution commentée

On calcule les coordonnées du point moyen G.

$\overline{x} = \frac{2 + 2{,}5 + 3 + 4 + 5{,}5 + 7 + 8{,}5 + 9 + 10 + 10{,}5}{10} = 6{,}2$;

$\overline{y} = \frac{5 + 8 + 10 + 15 + 18 + 25 + 32 + 45 + 50 + 62}{10} = 27$.

On a donc $G(6{,}2\ ;\ 27)$.

On calcule $\text{cov}(x\ ;\ y)$. Pour cela, on réalise un tableau.

x_i	2	2,5	3	4	5,5	7	8,5	9	10	10,5
$(x_i - \overline{x})$	−4,2	−3,7	−3,2	−2,2	−0,7	0,8	2,3	2,8	3,8	4,3
y_i	5	8	10	15	18	25	32	45	50	62
$(y_i - \overline{y})$	−22	−19	−17	−12	−9	−2	5	18	23	35

$$\text{cov}(x\ ;y) = \frac{-4{,}2-22+(-3{,}7)(-19)+(-3{,}2)(-17)+(-2{,}2)(-12)+(-0{,}7)(-9)+0{,}8(-2)+2{,}3\times 5+2{,}8\times 18+3{,}8\times 23+4{,}3\times 35}{10} = 54{,}8$$

On calcule $\sigma^2(x)$ avec une calculatrice : $\sigma^2(x) \approx 9{,}36$.

On calcule le coefficient directeur m de la droite des moindres carrés :

$m = \frac{\text{cov}(x\ ;\ y)}{\sigma^2(x)} \approx \frac{54{,}8}{9{,}36} \approx 5{,}855$ et $p = 27 - 5{,}855 \times 6{,}2 \approx -9{,}301$.

Remarque : la calculatrice donne : $m \approx 5{,}855$ et $p \approx -9{,}299$, les coefficients étant arrondis au millième.

3. Corrélation et ajustements

1. Coefficient de corrélation linéaire

Définition

On appelle **coefficient de corrélation linéaire** d'une série statistique à deux variables x et y, le nombre r défini par $r = \dfrac{\text{cov}(x\,;y)}{\sigma(x)\sigma(y)}$, où $\sigma(x)$ est l'écart type de la série x et $\sigma(y)$ est l'écart type de la série y.

Propriété (admise)

r est un nombre réel tel que $-1 \leqslant r \leqslant 1$.

Définition

Lorsque r est très proche de 1 ou de -1 ($r \geqslant 0{,}75$ ou $r \leqslant -0{,}75$), on dit que la **corrélation linéaire entre les séries x et y est forte.**

Remarques

- **Il ne faut pas confondre corrélation et causalité.** Une forte corrélation entre deux variables ne signifie pas qu'il y a un lien de cause à effet entre les valeurs des deux variables ni que l'une est la cause de l'autre.
- Lorsque le coefficient de corrélation linéaire est proche de 0, cela signifie que le nuage de points ne peut pas être « ajusté au mieux » par une droite. Il se peut qu'un autre type de courbes puisse l'ajuster au mieux.

2. Ajustement se ramenant à un ajustement affine

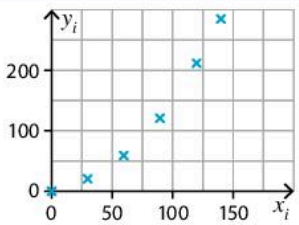
Lorsqu'un nuage de points est constitué de points qui ne paraissent pas globalement alignés, on peut être amené à déterminer d'autres types d'ajustements que l'ajustement affine.

Exemple

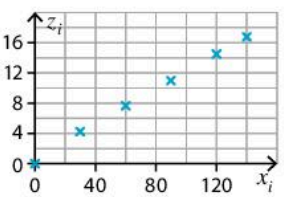
Le tableau ci-contre donne les vitesses x_i d'un véhicule (en km · h^{-1}) et la distance de freinage y_i (en m) correspondante pour chacune des vitesses.

Vitesse x_i	0	30	60	90	120	140
Distance y_i	0	18	58	120	212	285

Le nuage de points obtenu avec les valeurs x et y ne permet pas d'envisager un ajustement affine.
On procède à un changement de variable pour se ramener à une méthode d'ajustement affine connue.
Les valeurs sont arrondies au dixième.



x_i	0	30	60	90	120	140
$z_i = \sqrt{y_i}$	0	4,2	7,6	11	14,6	16,9



Le nouveau nuage de points $(x_i\,;z_i)$ peut être ajusté par une droite des moindres carrés, où z est exprimé en fonction de x.
On trouve avec une calculatrice et en arrondissant au centième, l'équation de la droite $z = 0{,}12x + 0{,}31$.
On en déduit donc l'expression de y en fonction de x :

$$z = \sqrt{y}, \text{ donc } y = z^2 = (0{,}12x + 0{,}31)^2.$$

On a ainsi réalisé un ajustement (appelé ici ajustement quadratique) du nuage de points $(x_i\,;y_i)$.

Remarque

- Un ajustement s'appelle aussi une **interpolation**. Lorsqu'on l'utilise, on dit qu'on fait une **extrapolation**.

Méthode 1 Déterminer un coefficient de corrélation linéaire par le calcul

Les séries statistiques ci-dessous montrent le nombre de véhicules en circulation dans cinq départements et le nombre d'accidents correspondant recensés sur une année.

Départements	Dép. 1	Dép. 2	Dép. 3	Dép. 4	Dép. 5
Nombre de véhicules x_i	132 980	112 232	109 530	107 070	102 354
Nombre d'accidents y_i	3 345	2 234	1 764	1 669	2 133

• Calculer et interpréter le coefficient de corrélation linéaire r de cette série à deux variables x et y.

Solution commentée

On calcule la covariance de x et de y. On construit donc le tableau suivant.

x_i	132 980	112 232	109 530	107 070	102 354
$(x_i - \overline{x})$	20 146,8	−601,2	−3 303,2	−5 763,2	−10 479,2
y_i	3 345	2 234	1 764	1 669	2 133
$(y_i - \overline{y})$	1 116	5	−465	−560	−96

$$\text{cov}(x\,;\,y) = \frac{1}{5}(20\,146{,}8 \times 1116 + (-601{,}2) \times 5 + (-3\,303{,}2)(-465) + (-5\,763{,}2)(-560) + (-10\,479{,}2) \times (-96))$$
$$= 5\,650\,041{,}2$$

On a également, avec une calculatrice, $\sigma(x) \approx 10584{,}84$ et $\sigma(y) \approx 597{,}41$.

On a donc $r = \dfrac{\text{cov}(x\,;\,y)}{\sigma(x)\sigma(y)} \approx 0{,}89$ arrondi au centième. Cette valeur est proche de 1, donc on peut dire qu'il y a une forte corrélation entre les deux séries x et y. On ne peut pas en déduire pour autant que le nombre d'accidents est seulement dû au nombre de véhicules en circulation.

Méthode 2 Déterminer un ajustement par changement de variable

Une entreprise vend des perles pour la fabrication de bijoux fantaisie. Le tableau ci-dessous indique la quantité vendue y_i (en tonne), pour un prix au kilogramme fixé x_i (en euro).

Prix x_i en €	3	4	5,5	7	8,5	9,3	10	10,8
Quantité y_i en tonne	4,926	3,773	2,773	2,197	1,820	1,669	1,554	1,430

1. Tracer le nuage de points. Peut-on envisager un ajustement affine ?
2. On pose $z = \dfrac{100}{y}$. Déterminer les valeurs de la série statistiques z. Arrondir à 10^{-3}.
3. Déterminer, avec une calculatrice, l'équation de la droite des moindres carrés des séries x et z. En déduire l'expression de y en fonction de x.
4. En extrapolant avec ce modèle, calculer la quantité de perles que vendrait l'entreprise si le prix montait à 24 € le kg.

Solution commentée

1. Le nuage de points des séries x et y semble indiquer qu'un ajustement affine n'est pas approprié car les points ne semblent pas globalement alignés.

2. La série z est la suivante.

z_i	20,300	26,504	36,062	45,517	54,945	59,916	64,350	69,930

3. On trouve, en arrondissant au millième, $z = 6{,}330x + 1{,}219$. On a $z = \dfrac{100}{y}$, donc $y = \dfrac{100}{z} = \dfrac{100}{6{,}330x + 1{,}219}$.

4. On prend $x = 24$. On a alors $y = \dfrac{100}{6{,}330 \times 24 + 1{,}219} \approx 0{,}653$.

Si le prix monte à 24 €, la quantité de perles vendues selon ce modèle, serait d'environ 0,653 tonne.