Chapitre 12

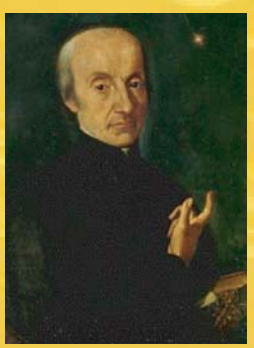
SÉRIES STATISTIQUES A DEUX CARACTÈRES

I . Activités de rappel

II . Nuage de points–Cœfficient de corrélation linéaire

III . Ajustements affines

IV . Exemples d'ajustements non affines



Giuseppe Piazzi

Le 1er janvier 1801, l'astronome G. Piazzi a décelé dans le ciel un objet inconnu. Après l'avoir étudié pendant quelques temps, il fit part de sa découverte. Mais lorsque ses collègues voulurent observer cet objet, ils ne le trouvèrent plus.
Le mathématicien Carl Friedrich Gauss appliqua alors la méthode des moindres carrés qu'il avait mise au point quelques années plus tôt et réussit à prédire, à partir des observations de Piazzi, où serait précisément cet objet céleste le dernier jour de l'année.

I. Activités de rappel

Activité 1 :

Voici les notes du dernier devoir de contrôle de mathématiques de la classe de 4ème sciences de l'informatique :

14 16 12 10 12 14 15 15 16 16 15 12 12 16 10 18 16 14 9 12 10 18 .

Calculer la moyenne $\overline{X}$, la variance V et l'écart type σ de cette série.

➢ La moyenne arithmétique d'une série statistique (x_i, n_i) est donnée par : $\overline{X} = \dfrac{\sum_{i=1}^{n} n_i x_i}{\sum_{i=1}^{n} n_i}$

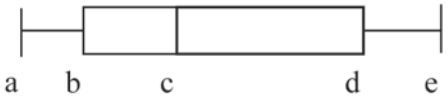
➢ La variance d'une série statistique (x_i, n_i) est le réel V défini par :

$$V = \frac{\sum_{i=1}^{n} n_i (x_i - \overline{X})^2}{\sum_{i=1}^{n} n_i} = \frac{\sum_{i=1}^{n} n_i x_i^{\,2}}{\sum_{i=1}^{n} n_i} - \left(\overline{X}\right)^2 \ ;$$

➢ L'écart type d'une série statistique (x_i, n_i) est le réel σ défini par : $\sigma = \sqrt{V}$

Activité 2 :

Le digramme en boîte ci-contre résume une série statistique. Identifier les paramètres indiqués sur le diagramme.



Activité 3 :

Un directeur a noté ses fonctionnaires sur 40 ; la moyenne est $\overline{X}$ =16.

1) Sur les dossiers, ce directeur a écrit les notes sur 20 ; quelle est la moyenne $\overline{x}$ de ces notes ?

2) Trouvant la moyenne faible par rapport à la production, le directeur décide de rajouter 1 point sur 20 à tous les fonctionnaires ; quelle est la moyenne m ?

3) Exprimer m en fonction de $\overline{x}$.

Activité 4 :

La série suivante donne les espérances de vie (en année) en 1995 de 21 pays d'Afrique de plus de 10 millions d'habitants

55	69	47	47	55	47	46	58	49	52	39
53	69	40	41	52	42	51	53	49	40	

1) Calculer une valeur médiane et les deux quartiles (Q_1 et Q_2) de cette série

2) Construire le diagramme en boîte de cette série.

3) L'espérance de vie la plus basse des pays d'Amérique (Haïti) était, la même année de 49 ans

et celle de l'Inde de 61 ans. Placer ces valeurs sur le diagramme précédent.

➢ La médiane partage la série ordonnée (dans l'ordre croissant) en deux groupes de même effectif.

➢ Les quartiles partagent la série ordonnée en quatre groupes de même effectif.

➢ L'intervalle interquartile $[Q_1;Q_3]$ contient 50% des observations

➢ L'écart interquartile $I = Q_3 - Q_1$ mesure la dispersion de la série par rapport à la médiane.

Activité 5 :

Pour une année donnée, cette série donne le nombre d'interventions chirurgicales quotidiennes dans une clinique.

Interventions quotidiennes	0	1	2	3	4	5	6
Effectif (en jours)	84	105	72	59	28	15	2

1) Dresser le tableau des fréquences.

2) Utiliser les fréquences pour calculer la moyenne et l'écart-type. On arrondit les valeurs au dixième.

3) Calculer l'arrondi au dixième du pourcentage des valeurs appartenant à l'intervalle $\left[\overline{X} - \sigma\ ;\ \overline{X} + \sigma\right]$

Activité 6 :

La série suivante donne le chiffre d'affaire en milliers de dinars des 500 magasins d'une chaîne de distribution

Chiffre d'affaires	effectif	Effectif cumulé croissant
[100 ; 150[	40	
[150 ; 200[	50	
[200 ; 250[	60	
[250 ; 300[	70	
[300 ; 350[	100	
[350 ; 400[	80	
[400 ; 450[	60	
[450 ; 500[	40	

1) a) Reproduire et compléter le tableau ci-contre.

b) Vérifier que le premier quartile Q_1 se trouve dans la classe [200 ; 250[.

c) dans quel intervalle se trouve la médiane *Me* ? le troisième quartile Q_3 ?

2) Pour calculer une valeur approchée du premier quartile Q_1, on fait l'hypothèse que les valeurs de la série sont répartie uniformément dans la classe [200 ; 250[.

a) Expliquer la relation : $\dfrac{Q_1 - 200}{125 - 90} = \dfrac{250 - 200}{150 - 90}$

Chiffre d'affaires	200	Q_1	250
Effectif cumulé croissant	90	125	150

b) Calculer une valeur approchée arrondie au dixième du premier quartile Q_1.

c) Calculer de la même façon *Me* et Q_3.

Activité 7 :

La touche « Random » d'une calculatrice a donné 100 nombres au hasard (entre 0 et 9). Ils peuvent être considérés comme un échantillon issu de la population des nombres aléatoires que cette calculatrice peut fournir.

1	7	5	5	9	4	2	8	3	2	0	0	7	2	8	0	0	5	7	5	1	7	8	4	9
8	8	1	4	8	8	5	4	0	9	5	8	6	7	8	6	5	0	4	7	7	2	5	6	5
6	1	6	2	4	0	7	5	8	8	1	0	6	4	6	1	1	2	8	0	8	2	8	9	6
0	5	8	9	0	4	0	7	6	3	4	8	0	5	2	6	7	5	8	7	3	6	4	3	9

1) a) A quelle moyenne approximative peut-on s'attendre ?
   b) Vérifier cette prévision avec cette série S_1.
   c) Calculer l'arrondi au dixième de l'écart-type de cette série.
2) On regroupe ces 100 nombres par 4 (les 4 premiers, les 4 suivants,…).
a) Calculer la moyenne des 4 nombres dans chacun des 25 groupes obtenus.
b) On obtient ainsi une série S_2 de 25 moyennes. Calculer la moyenne de cette série. Quel résultat obtient-on ? Est-ce prévisible ?
c) Calculer l'écart-type de la série S_2.
d) Comparer avec l'écart-type de la série S_1. Est-ce prévisible ?
3) On regroupe ces 100 nombres de départ par 10. On obtient ainsi une série S_3 de 10 moyennes.
a) Calculer la moyenne et l'ecart-type de la série S_3.
b) Représenter les séries S_2 et S_3 par leur diagramme en boîte.
c) Quelle conclusion peut-on faire?

II. Nuage de point– cœfficient de corrélation linéaire

Activités de découverte

Activité 1:

Ce tableau donne pour les personnes de plus de 15 ans la consommation de tabac en g par jour et l'évolution du prix du tabac (base 100 en 1970).

Année	**Rang de l'année**	**Consommation** x_i	**Indice** y_i **du prix**
1995	1	5 ,1	121
1996	2	5	126
1997	3	4,8	136
1998	4	4,7	138
1999	5	4,7	143
2000	6	4,6	147
2001	7	4,6	152
2002	8	4,4	162
2003	9	3,9	180
2004	10	3,3	220

Dans ce tableau, la population est un ensemble d'années qu'on peut noter $A = \{a_1, a_2, ..., a_{10}\}$. A chaque individu a_i de A on associe un couple de réels (x_i, y_i) représentant respectivement la consommation et l'indice du prix de cet individu.

1) a) Ecrire les 10 couples (x_i, y_i) où $1 \leq i \leq 10$.

L'ensemble des 10 couples (x_i, y_i) où $1 \leq i \leq 10$, est **une série statistique à deux caractères**.

b) Calculer la moyenne arithmétique $\overline{X}$ de la série statistique des consommations $(x_i)_{1 \leq i \leq j}$.

c) Calculer la moyenne arithmétique $\overline{Y}$ de la série statistique des indices du prix $(y_i)_{1 \leq i \leq 10}$.

2) a) Dans un repère orthogonal du plan, placer les dix points $M_i(x_i; y_i)$ (unités : 1cm pour un gramme en abscisse et 1cm pour 10 points d'indice du prix en ordonnée en commençant la graduation à l'indice du prix 120). Placer en rouge sur ce graphique le point G de coordonnées $(\overline{Y} ; \overline{X})$.

➢ L'ensemble des 10 points $M_i(x_i; y_i)$ où $1 \leq i \leq 10$ est **le nuage de points** de la série statistique à deux caractères et le point G $(\overline{X}; \overline{Y})$ est **le point moyen** de ce nuage.

b) Tracer une droite qui est la plus proche possible de tous les points M_i .

Comparer avec les droites tracées par d'autres élèves.

c) Afin que tous les élèves aient la même droite, on décide de choisir la droite (GM_{10}) .

Tracer (GM_{10}) et en donner une équation en arrondissant les coefficients au centième.

3) On suppose que dans les années à venir, la consommation de tabac va continuer à évoluer de la même façon. On peut alors utiliser la droite (GM_{10}) pour estimer la consommation de tabac lorsque l'indice sera 250. Estimer cette consommation par lecture graphique, puis par le calcul en utilisant l'équation de la droite (GM_{10}) .

Activité 2:

On considère la série double (X,Y) ci-dessous :

1) Calculer les moyennes arithmétiques $\overline{X}$ et $\overline{Y}$ ainsi que $\sigma(x)$ et $\sigma(y)$.

x_i	10	11	13	15	17	18
y_i	105	107	110	111	112	115

2) Etant donnée une série statistique double $(x_i; y_i)_{1 \leq i \leq n}$, définie en données individuelles,

➢ On appelle **covariance** du couple $(X ; Y)$ le réel noté $cov(X ; Y)$ et défini par :

$$\text{cov}(X;Y) = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right)$$

Calculer la covariance de cette série.

3) Montrer que, dans le cas général, on a :

- $\text{cov}(X;Y) = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{X}\,\overline{Y}$
- $\text{cov}(aX + b;Y) = a\,\text{cov}(X;Y)$
- $\text{cov}(X;X) = V(X)$

➢ Ainsi un changement d'unité ou d'origine du repère affecte la covariance .

4) On démontre et nous admettons que : $\text{cov}^2(X;Y) \leq V(X)V(Y)$.

Déduire alors que : $\left|\text{cov}(X;Y)\right| \leq \sigma(X)\sigma(Y)$.

5) Quand est ce que $V(X)$ et $V(Y)$ sont nuls ?

➢ Dans le cas contraire, on considère le réel $\frac{\text{cov}(X;Y)}{\sigma(X)\sigma(Y)}$ appelé **le cœfficient de corrélation linéaire** du couple $(X;Y)$. Ce réel garde la même valeur quelque soit l'unité choisie .

Calculer ce cœfficient pour la série double définie au début

Activité 3:

a) Pour chacune des séries doubles suivantes, construire son nuage de points, dans un repère orthogonal et calculer son cœfficient de corrélation

z_i	10	11	13	15	17	18
t_i	105	107	110	111	112	115

x_i	30	60	114	122
y_i	9	14	14	51

b) Quelle conjecture peut-on faire sur la relation entre la forme du nuage de points de la série et son cœfficient de corrélation ?

A retenir

Nuage de points

Définitions

X et Y désignent deux variables statistiques numériques observées sur n individus d'une même population. Pour $1 \leq i \leq n, x_i$ et y_i désignent les mesures relevées respectives de X et Y .

- Les n couples $(x_i ; y_i)$ forment **une série statistique à deux caractères.**
- Dans un repère orthogonal, l'ensemble des n points $M_i(x_i; y_i)$ constitue l**e nuage de points** associé à cette série statistique double.

- Le **point moyen** G du nuage de points $M_i(x_i; y_i)$, $1 \leq i \leq n$, dans un repère, est le point $G(\ \overline{X}\ ;\ \overline{Y}\ )$

Covariance

Définition

Soit une série statistique double $(x_i; y_i)_{1 \leq i \leq n}$ définie en données individuelles. On désigne par $x_1, x_2, ..., x_n$ les valeurs du 1er caractère quantitatif X et par

$y_1, y_2, ..., y_n$ celles du 2ème caractère quantitatif Y. On appelle covariance du couple $(X ; Y)$ le réel noté cov$(X ; Y)$ et défini par :

$$\text{cov}(X;Y) = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right) \quad \text{où} \quad \overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{et} \quad \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Propriétés

Avec les notations de la définition, on a :

- $\text{cov}(X ; Y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{X}\,\overline{Y}$
- cov$(X ; Y)$ = cov(Y ;X)
- cov$(a\,X + b\ ; Y)$ = a cov$(X ; Y)$ pour tous réels a et b .
- cov$(X;X) = V(X)$.
- $\text{cov}^2(X;Y) \leq V(X)V(Y)$

Cœfficient de corrélation linéaire

Définition

Soient X et Y deux variables quantitatives, non constantes et observées dans une même population.

On appelle cœfficient de corrélation linéaire du couple $(X ; Y)$ le réel r défini par

$r = \dfrac{\text{cov}(X;Y)}{\sigma(X)\sigma(Y)}$, $\sigma(X)$ et $\sigma(Y)$ étant les écart-types respectifs de X et Y.

Propriétés

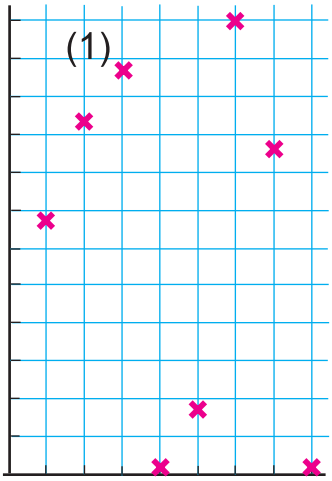
- $-1 \leq r \leq 1$.
- r est invariant par changement d'unité ou d'origine.
- Si $|r| \geq 0{,}75$ alors la corrélation linéaire entre X et Y est forte .
- Si $|r| < 0{,}75$ alors la corrélation linéaire entre X et Y est faible.

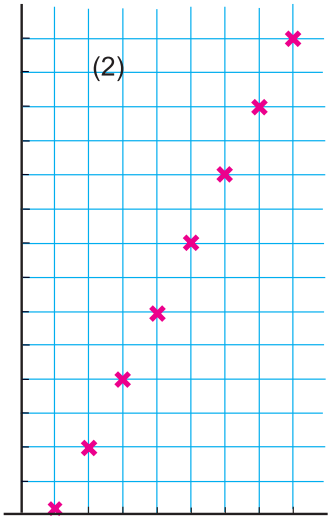
Applications

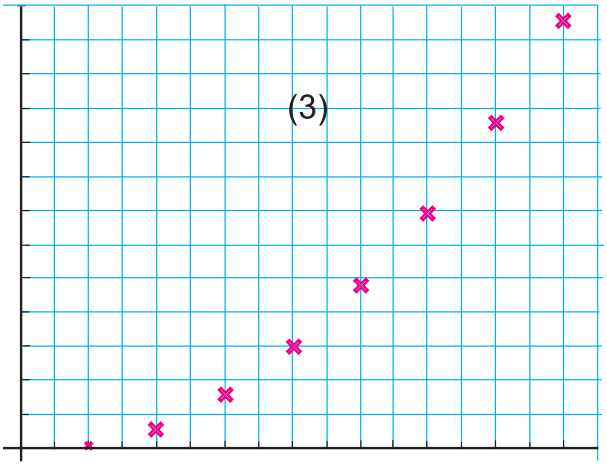
1 **Vrai ou faux**. Corriger les réponses fausses.

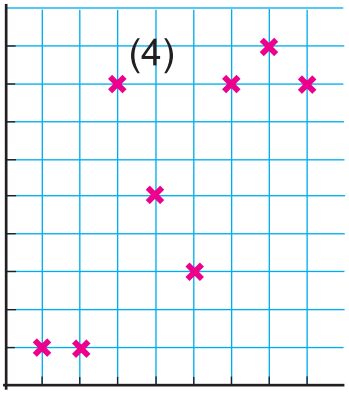
On considère les cinq séries du tableau et quatre nuages de points ci-dessous

x_i	1	2	3	4	5	6	7	8
y_i	1	1	8	5	3	8	9	8
z_i	0	3	8	15	24	35	48	63
t_i	0	10	20	30	40	50	60	70
w_i	6,71	9,29	10,71	0,23	1,68	11,96	8,64	0,23









a) Le nuage de points $M_i(x_i;w_i)$ est le nuage (4).

b) Le nuage de points $N_i(x_i;z_i)$) est le nuage (2).

c) Le point moyen du nuage (1) est un des points de ce nuage.

d) Les points du nuage (2) sont alignés.

e) le point moyen du nuage (2) est aligné avec les autres points de nuage.

f) Le point du nuage (3) sont ceux d'une parabole.

g) Le point moyen de la série $(x_i;y_i)$ est $G_1\ (4{,}5\ ;\ 4)$.

h) Le point moyen de la série $(x_i;t_i)$ est $G_2\ (4{,}5\ ;\ 35)$

2 Le tableau suivant donne la part des salariés parmi les actifs

Année : x_i	1975	1980	1985	1990	1995	2000	2001
Part en % y_i:	82,3	84	85,3	87	89,5	91,1	91,3

1) a) Représenter le nuage de points associé à cette série dans un repère orthogonal d'origine O(1975 ; 80).

b) Vérifier que le point moyen du nuage *a* pour coordonnées (1989,4 ; 87,2).

c) Calculer la covariance de cette série et vérifier que le cœfficient de corrélation linéaire du couple (X,Y) est 0,99.

2) Pour cette série on pose $z_i = x_i - 1975$ et $t_i = y_i - 80$. On définit ainsi deux séries Z et T.

a) Recopier et compléter le tableau ci-dessous.

$z_i = x_i - 1975$	0						
Part en % *yi* :	2.3						

b) Déterminer les coordonnées du point moyen de cette série.

c) Calculer la covariance de cette série et vérifier que le cœfficient de corrélation linéaire du couple $(Z\ ,\ T)$ est 0,99.

3) a) On pose $u_i = \frac{y_i}{100}$. Recopier et compléter le tableau ci-dessous

$z_i = x_i - 1975$	0	5					
$u_i = \frac{y_i}{100}$	0.823	0.84					

b) Déterminer les coordonnées du point moyen de cette série.
c) Calculer la covariance de cette série et vérifier que le cœfficient de corrélation linéaire du couple (Z , U) est 0,99.
4) Qu'est ce qu'on peut en déduire ? Expliquer la réponse.

Activité 1:
Le tableau suivant donne la consommation tunisienne en tonnes d'une matière première M pour la période de 1996 à 2003.

Année	1996	1997	1998	1999	2000	2001	2002	2003
Consommation en Tonne	7740	7800	7880	7900	7920	8000	8020	8060

III. Ajustements affines

Activités de découverte

On appelle x_i le rang de l'année exprimé à partir de 1996 et y_i la consommation tunisienne en tonnes de la matière M.
1) Représenter le nuage de points de coordonnées $M_i(x_i;y_i)$ (on choisira un repère orthogonal avec 2cm pour 1 unité en abscisses et 1cm pour 20 unités en ordonnées en graduant cet axe à partir de 7720).
2) On constate que le nuage de points a une forme allongée. On se propose de déterminer une droite qui indique la tendance de l'évolution de la consommation.
On partage le nuage en deux parties : le premier sous-nuage formé des points M_1, M_2, M_3, M_4 et le deuxième sous-nuage formé par les autres points.
a) Calculer les cordonnées des points moyens G_1 et G_2 de chacun de ces sous-nuages.
b) Déterminer une équation de la droite $(G_1 G_2)$. Tracer cette droite sur le graphique.
c) On suppose que la fonction affine associée à cette droite modélise la croissance de consommation entre 1996 et 2007. Utiliser cette équation pour estimer la consommation de la matière M en 2007.

3) a) On aurait pu partager le nuage de points en deux parties autre que les précédentes. Choisir deux autres sous-nuages de 4 points. Déterminer leurs points moyens G'_1 et G'_2, puis une équation de la droite $(G'_1G'_2)$. En déduire une estimation de la consommation en 2007.

b) Peut-on savoir si une estimation est meilleure que l'autre ?

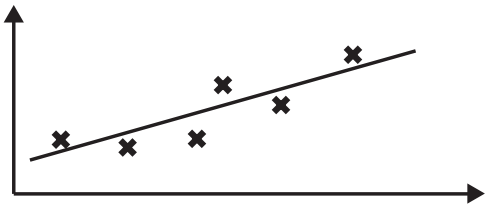
➢ La droite $(G_1\ G_2)$ est appelée **la droite de MAYER** (1814-1878)

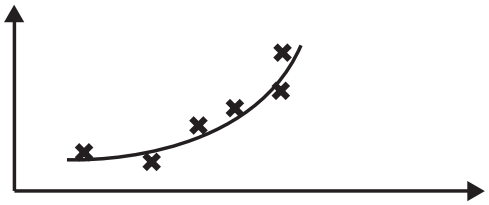
Commentaire

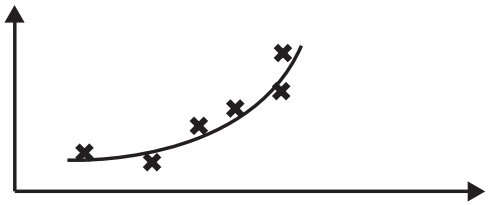
Lorsque deux caractères X et Y sont liés l'un à l'autre, l'étude de la forme du nuage de points permet de **modéliser** cette liaison. La forme d'un nuage de points associé à une série à deux variables X et Y invite à retenir, pour l'ajuster, des modèles de fonctions familières : soit le modèle affine $y = ax + b$, soit le modèle exponentiel $y = a.b^x$ ou autres.

Effectuer un **ajustement de Y en X** d'un nuage de points consiste à trouver une fonction f telle que la courbe d'équation $y = f(x)$ passe « le ➢ plus près possible » des points du nuage.

Pour les statisticiens, une formule telle que $y = f(x)$ est appelé un **modèle** ; x est la variable **explicative** et y la variable **expliquée**.







Activité 2 :

Ci-contre une série statistique à deux variables (X, Y).

x_i	1	2	3
y_i	3	4	6

D est une droite variable qui passe par le point moyen G du nuage ; on note a son cœfficient directeur.

1) a) Calculer les coordonnées de G.

b) Écrire une équation de D.

c) Calculer le cœfficient de corrélation du couple (X, Y).

d) Construire dans un repère orthogonal le nuage de points de cette série.

2) Soient les points P_1, P_2 et P_3 de D d'abscisses respectives 1, 2 et 3.

a) Exprimer, en fonction de a , la somme $\sum_{i=1}^{3}(M_iP_i)^2$.

b) Déterminer a pour que cette somme soit minimale.

➢ La droite ainsi obtenue est appelée **droite de régression** de Y en X.

3) Construire, dans le même repère, la droite D.

Activités 3 :

Le tableau ci-dessous donne pour une grande entreprise industrielle la relation entre sa charge mensuelle en milliers d'heures de travail et sa production mensuelle en milliers de produits.

Production x_i	20	50	80	90	100	120	160	180
Charge y_i	60	85	90	105	115	125	140	160

1) Représenter le nuage de points de coordonnées $(x_i;y_i)_{1\le i\le 8}$ dans un repère orthogonal.

2) Calculer les coordonnées $(\overline{X};\overline{Y})$ du point moyen G du nuage, puis vérifier que le cœfficient de corrélation du couple (X, Y) est 0,98.

3) Soit D une droite d'équation $y = ax + b$ et passant par G $(\overline{X};\overline{Y})$ .Montrer que $b = \overline{y} - \overline{ax}$

4) On considère les points $M_i(x_i,y_i)$ et $P_i(x_i,ax_i+b)$ et on se propose de déterminer D pour que la somme des écarts $f(a) = \sum_{i=1}^{8}(M_iP_i)^2$ soit minimale:

a) Montrer que $8f(a) = a^2V(X) - 2aCov(X,Y) + V(Y)$

b) Montrer que $a = \dfrac{\text{cov}(X,Y)}{V(X)}$.

5) Calculer a puis tracer la droite D dans le même repère.

➢ $\sum_{i=1}^{8}(M_iP_i)^2$ est minimale pour la valeur de a déterminée au b). La droite D tracée passe donc «le plus proche possible» des points du nuage. C'est-à-dire que la somme des carrés des écarts entre les points M_i du nuage et les points P_i de la droite D de même abscisse, est la plus petite possible .

On dit qu'on a effectué un ajustement linéaire par la méthode **des moindres carrés**.

6) a) Vérifier que l'on trouve que D a pour équation $y = 0{,}6\,x + 50{,}4$.

b) Pour une production de 300 unités estimer la charge nécessaire à l'aide de la droite D.

➢ Pour une valeur donnée x_i de la variable X, la fonction $f(x) = ax + b$ permet de prévoir approximativement la valeur correspondante de Y ; pour cela on calcule $f(x_i)$.

• Si x_i appartient à l'intervalle d'observation des valeurs de X, on dit qu'on fait une **interpolation.**

• Si x_i n'appartient pas à cet intervalle, on parle d'**extrapolation**, mais dans ce cas il faut faire l'hypothèse que le modèle reste plausible à l'extérieur de cet intervalle.

A retenir

Soient X et Y deux variables statistiques quantitatives, non constantes et observées dans une population donnée.Lorsque le cœfficient de corrélation r vérifie $|r| \geq 0,75$ ou lorsque le nuage de points a une forme allongée, alors il est possible d'effectuer un ajustement affine du nuage de points $M(x_i, y_i)$.

Théorème et définition

Lors d'un ajustement affine de Y en X par la méthode des moindres carrés, la droite obtenue passe par le point moyen du nuage et a pour équation :

$$y = a(x - \overline{X}) + \overline{Y} \quad \text{où} \quad a = \frac{\text{cov}(X,Y)}{V(X)}.$$

Cette droite s'appelle la **droite de régression de Y en X**.

La droite de régression de X en Y obtenue par la méthode des moindres carrés, lors d'un ajustement affine, passe par le point moyen du nuage et a pour équation :

$$x = a'(y - \overline{Y}) + \overline{X} \quad \text{où} \quad a' = \frac{\text{cov}(X,Y)}{V(Y)}.$$

Remarques

On suppose ici que les points du nuage ne sont pas tous alignés sur une même droite verticale, ni sur une même droite horizontale. On a donc

et

$\sigma(X) \neq 0 \qquad \sigma(Y) \neq 0$

- Les deux droites de régression : $\begin{cases} D : y = ax + b \\ D' : x = a'y + b' \end{cases}$ passent par le point moyen G $(\overline{X}; \overline{Y})$.
- Les deux coefficients a et a' sont de même signe et le cœfficient corrélation r vérifie $r^2 = aa'$

Méthode de Mayer

Cette méthode d'ajustement consiste à partager les données en deux groupes de mêmes effectifs (à un près) après un tri en fonction des valeurs de la première variable. On calcule ensuite les coordonnées des points moyens G_1 et G_2 de chaque groupe. On construit alors la droite (G_1G_2).

Applications

1 Dans le tableau ci-dessous, on donne la taille moyenne (en cm) des nouveaux nés en fonction du nombre de l'âge gestationnel (en semaines).

Age gestationnel (semaines)	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45
Taille (cm)	47.5	48,5	49	49,7	50	50, 5	50, 8	51, 2	51, 5	51, 8	52, 2	52, 5	52, 8	53	53, 5	53, 7

1) Représenter le nuage de points dans un repère orthogonal en prenant comme unités :
- En abscisse : 1 cm pour 1 semaine (commencer la graduation à 20 semaines)
- En ordonnée : 2 cm par unité (commencer la graduation à 45 cm)

2) on propose d'ajuster ce nuage de points par application de la méthode de Mayer.
- Calculer les coordonnées des points moyens G_1 et G_2.
- Tracer la droite d'ajustement passant par les points G_1 et G_2.

3) Déterminer l'équation de cette droite d'ajustement.

2 Une équation de la droite de régression par la méthodes des moindres carrés, de X en Y est : $x = -0,43y + 12,3$. Les valeurs du caractéres X sont 1 ; 2 ; 5 ; 7 ; 11 ; 13.

3 Calculer les coordonnées du point moyen.

Ce tableau donne, l'espérance de vie (en années) des hommes à la naissance pour certaines années.

Année x_i	1980	1985	1990	1995	1998	1999	2000	2001	2002	2003	2004
Espérance y_i	70,02	71,3	72,8	73,9	74,6	74,9	75,3	75,5	75,8	75,9	76,7

1) Dans un repére, représenter le nuage de points associé à cette série.
Un ajustement affine est-t-il justifié ?

2) On donne les calculs suivants : $\overline{X} = \dfrac{21957}{11} \approx 1996,09$; $\overline{Y} = \dfrac{816,9}{11} \approx 74,26$;

$\sum_{i=1}^{n}(x_i - \overline{X})^2 = 616,91$; $\sum_{i=1}^{n}\left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right) = 159,84$

Déterminer alors une équation de la droite d'ajustement D de y en x par la méthode des moindres carrés et tracer cette droite sur le graphique précédent.
3) Estimer l'espérance de vie à la naissance des hommes en 2007 en supposant que ce modèle reste plausible.
4) Peut-on, à l'aide de cet ajustement, estimer l'espérance de vie à la naissance des hommes en 2500 ? En l'an 3000 ?

4 1) **Q.C.M**. Donner toutes les bonnes réponses.
On considère la série statistique double $(x_i ; y_i)$ suivante :

x_i	1	2	3	5	6
y_i	4	7	10	14	16

a) La droite D_1 d'équation $y = 2x + 4$ passe par trois points du nuage, donc est un bon ajustement.
b) La droite D_2 d'équation $y = 3x + 1$ passe par trois points du nuage, donc D_2 est le meilleur ajustement.
c) La droite de régression Δ de y en x a pour équation $y = 2{,}36x + 2{,}17$, les coefficients étant arrondis à 10^{-2}
2) **VRAI OU FAUX**. Corriger les réponses fausses.
a) Pour déterminer le meilleur ajustement du nuage selon la méthode des moindres carrés, pour chacune des droites Δ ,D_1 et D_2 , on calcule la somme : $S = \sum\left(y_i - (ax_i + b)\right)^2$
b) Pour la droite D_1 , les trois derniers termes de la somme S sont nuls.
c) Pour la droite Δ , tous les termes de la somme S sont nuls.
d) La somme S prend la même valeur pour les droites D_1 et D_2 .
e) La somme S est minimale pour la droite Δ .
f) La somme S vaut 5 pour la droite D_1.

5 Une équation de la droite de régression de y en x est : $y = -0{,}43x + 12{,}3$.
La moyenne de X est $\overline{X} = 5{,}7$ et le cœfficient de corrélation est $r = 0{,}85$. Donner une équation de la droite de régression de x en y.

IV. Exemples d'ajustements non affines Ajustement par une fonction puissance

L'offre et la demande d'un produit ont fait l'objet d'une étude statistique. x désigne le prix unitaire en dinars, y désigne la demande en milliers d'unités et z désigne l'offre en milliers.

x	1,5	2,5	3,5	4,5	5	7	8,5
y	8,4	5,3	5,3	3,1	2,8	2,1	1,7
z	0,75	1,23	1,75	2,25	2,5	3,5	4,25

1) Vérifier que la quantité offerte est proportionnelle au prix unitaire. En déduire la fonction g, telle que $z = g(x)$ pour $x \in [0;10]$.
2) Donner une équation de la droite de régression de y en x par la méthode des moindres carrés. A l'aide de cet ajustement, déterminer le prix d'équilibre (prix pour lequel la quantité offerte est égale à la quantité demandée).
3) On se propose de rechercher un autre ajustement de la fonction de demande par une fonction puissance.
a) Etablir le tableau des valeurs $X_i = \ln(x_i)$ et $Y_i = \ln(y_i)$, arrondies à 10^{-3} .
b) Calculer le cœfficient de corrélation linéaire de (X,Y). Interpréter ce résultat.
c) Déterminer un ajustement affine par moindres carrés de Y en X.
d) En déduire la nouvelle relation de la demande sous la forme $y = k\,x^{\alpha}$.

Ajustement exponentiel

On a mesuré entre 1995 et 2000, l'effet de la pollution sur la population piscicole d'une rivière. Les résultats présentés dans le tableau suivant donnent une estimation du nombre y_i de poissons, exprimé en milliers, correspondant à l'année dont le rang est x_i .

année	1995	1996	1997	1998	1999	2000
x_i	1	2	3	4	5	6
y_i	951,3	106,7	96,5	63,2	21	9,4

1) Représenter le nuage de points $M_i(x_i;\ln(y_i))$ dans un repère orthogonal. Expliquer pourquoi un ajustement exponentiel semble justifié.
2) On pose $z = \ln(y)$.
a) calculer le cœfficient de corrélation linéaire de (x, z) . Interpréter ce résultat.
b) Ecrire une équation de la droite d'ajustement de z en x par la méthode des moindres carrés sous la forme $z = ax + b$ en arrondissant a et b au centième.
c) En déduire un ajustement exponentiel de y en x sous la forme $y = Ae^{Bx}$.
3) On suppose que l'évolution de cette population se poursuit sur le même modéle.
a) A partir de quelle année cette population sera-t-elle inférieure à 1000 ?
b) Donner une estimation de la population de cette rivière en l'an 2008 ?

Ajustement parabolique

Les 11 élèves d'une classe travaillant sur la proportionnalité doivent chacun tracer un disque sur une feuille de papier quadrillé, puis évaluer l'aire de ce disque. Ce tableau donne les résultats de cette expérience : x_i est le rayon d'un disque en cm et A_i l'aire en cm² du disque correspondant.

x_i	2	2,5	3	3,5	4	4,5	5	5,5	6	6,5	7
A_i	12	20	28	38	50	50	78	95	113	113	154

1) Les deux séries sont-elles proportionnelles ? Justifier la réponse.
2) On pose $y = \sqrt{A}$.
a) Présenter dans un tableau la série statistique $(x_i;y_i)$,chaque y_i sera arrondi au dixiéme.
b) Représenter dans un repère orthogonal, le nuage de points associé à cette série (unité graphique : 2cm sur chaque axe).
c) Déterminer une équation de la droite d'ajustement de y en x en arrondissant les cœfficients au centième.
d) Tracer cette droite dans le repére précédent en faisant apparaître le point moyen G de cette série statistique.
3) L'aire A d'un disque de rayon X étant $A = \pi X^2$, $Y = \sqrt{A}$ est alors proportionnel à X. Quel est le cœfficient de proportionnalité ? En déduire une valeur approchée du nombre π en utilisant le résultat trouvé à la question 2.c).

Situation 1

Un chef d'entreprise reçoit de la part de ses collaborateurs la demande d'obtenir des véhicules de fonction plus confortables et plus puissants. Il sollicite alors son comptable afin que celui-ci examine la demande et sa faisabilité.

Le comptable utilise le tableau ci-dessous, donnant le prix de revient kilométrique (PRK) des véhicules d'une puissance fiscale de 4 à 8 CV et en fait une projection sur les véhicules plus puissants.

Puissance fiscale des véhicules (CV)	4	5	6	7	8
Prix de revient kilométrique (D)	0,424	0,471	0,492	0,513	0,555

1)Représenter cette série statistique par un nuage de points dans un repère orthogonal.

2) Calculer les coordonnées du point moyen G.

3) On admet que la droite d'ajustement de cette série a pour équation : $y = 0{,}03x + 0{,}311$

a) Montrer que le point G appartient à cette droite.

b) Tracer cette droite dans le repère précédent.

4) En utilisant la droite d'ajustement, quel est le prix de revient d'une voiture de 10 CV ?
Laisser apparents les traits nécessaires à la lecture.

5) Le comptable fixe le prix de revient kilométrique maximum à 0,650 D.
Calculer la puissance maximale du véhicule qui correspond à cette exigence.

Vers une solution :

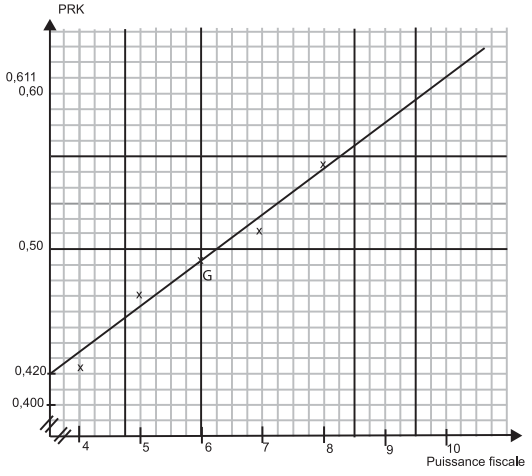
G(6 ; 0,491)

5) Résolution de l'inéquation :

$$0{,}03x + 0{,}311 < 0{,}65$$

donc $x < 11{,}3$

La puissance fiscale maximale autorisée par le comptable est de 11 CV



Situation 2 :

Sur un parcours donné, la consommation y d'une voiture est donnée en fonction de sa vitesse moyenne x par le tableau suivant :

X (en km/h)	80	90	100	110	120
Y (en litres/100km)	4	4,8	6,3	8	10

1) La consommation est-elle proportionnelle à la vitesse moyenne ? Justifier la réponse.

2) a) Représenter le nuage de points correspondant à la série statistique $(x_i ; y_i)$ dans un repère orthogonal du plan (on prendra 2cm pour 10km/h sur l'axe des abscisses et 1 cm pour 1 litre sur l'axe des ordonnées).

b) Déterminer les coordonnées du point moyen G du nuage et le placer sur le graphique.

c) Donner une équation de la droite d'ajustement affine de y en x par la méthode des moindres carrés et tracer cette droite.

d) En utilisant cet ajustement, estimer la consommation aux 100km (arrondie au dixiéme) de la voiture pour une vitesse de 130km/h.

3) La forme du nuage permet d'envisager un ajustement exponentiel. On pose : $z = \ln(y)$ et on admet que la droite d'ajustement obtenue pour les cinq points $(x ; z)$ du nuage par la méthode des moindres carrés a pour équation : $z = 0{,}0234\,x - 0{,}508$.

a) Ecrire y sous la forme $y = Ae^{Bx}$ (donner A et B arrondis à 10^{-4}).

b) Tracer, sur le même graphique, la courbe d'équation $y = A\,e^{Bx}$ pour $x \in [80 ; 120]$.

c) En utilisant cet ajustement, estimer la consommation aux 100km (arrondie au dixiéme) de la voiture, pour une vitesse de 130km/h.

4) Des valeurs obtenues dans les questions 2.d) et 3.c), laquelle vous semble la plus proche de la consommation réelle ? Expliquer votre choix.

Vers une solution :

1) $\frac{4}{80} \neq \frac{10}{120}$; conclure.

2) c) $y = 0{,}152\,x - 8{,}58$. Pour $x = 130$, on trouve $y = 11{,}18$, soit une consommation d'environ 11,2 litres aux 100km.

3) a) $y = e^{0{,}0234x - 0{,}508}$; $A = e^{-0{,}508} \approx 0{,}6017$

c) Pour $x = 130$, on obtient $y = 0{,}6017e^{0{,}0234\times 130}$, soit une consommation d'environ 12,6 litres aux 100km .

4) La seconde valeur (12,6) est la plus vraisemblable.

Les augmentations de consommation, par tranche de 10km/h, sont de plus en plus grandes. Entre 110 et 120, la consommation augmente de 2 litres aux 100km. On peut penser qu'elle augmentera encore plus entre 120 et 130, ce qui donnera plutôt 12,6 que 11,2.

Situation 3

Monsieur Math est un papa heureux. Son fils bénéficie d'une excellente santé. Il a noté son poids (en kg) à chacun de ses anniversaires. Soucieux de l'avenir, Monsieur Math souhaiterait avoir une idée de l'évolution du poids de son héritier.

1) Représenter cette série par un nuage de points (1 cm pour un an en abscisse et 1 cm pour 4 kg en ordonnée).

2) Posons $z_i = \sqrt{y_i}$ et compléter le tableau précédent avec les z_i arrondis à 10^{-2} près.

Age (en années) x_i	7	8	9	10	11	12
Poids y_i	22	24	28	34	42	51

3) a) Sur un autre graphique, représenter les points de coordonnées (x_i , z_i). Calculer le cœfficient de corrélation linéaire entre x et z . Calculer les coordonnées du point moyen G.

b) Donner une équation réduite de la droite de régression de z en x par moindres carrés.

4) En utilisant cette droite, calculer quel pourrait être le poids de l'héritier à 20 ans et à 25 ans. Que faut-il penser de tels calculs ? Monsieur Math doit-il réellement se faire du souci ?

MEILLEUR AJUSTEMENT ET TABLEUR

On compare les taux de chômage, exprimés en %, au Japon et en France entre 1985 et 2003

année	1985	1990	1995	2000	2001	2002	2003
rang x_i	0	5	10	15	16	17	18
Japon y_i	2,6	2,1	3,1	4,7	5	5,4	5,3
France z_i	10,1	8,6	11,3	9,3	8,5	8,8	9,4

Travail sur papier

1) Calculer le pourcentage d'évolution du taux de chômage au Japon et en France entre 1985 et 2003.

2) On admet que le pourcentage d'évolution annuel moyen est de 12,6% pour le Japon et de -1,19% pour la France. En supposant que l'évolution reste la même pour les deux pays, estimer le taux de chômage en France et au Japon en 2004.

Détermination de la tendance à l'aide d'un tableur

Pour déterminer les droites de régressions des nuages de points de coordonnées $(x_i;y_i)$ et $(x_i;z_i)$, construire le tableau ci-dessous.

	A	B	C	D
1	Taux de chomage			
2	année	rang xi		France zi
3	1985	0	2,6	10,1
4	1990	5	2,1	8,6
5	1995	10	3,1	11,3
6	2000	15	4,7	9,3
7	2001	16	5	8,5
8	2002	17	5,4	8,8
9	2003	18	5,3	9,4

Représenter, dans un même repére, les nuages de points de coordonnées : $(x_i;y_i)$ et $(x_i;z_i)$.

Sélectionner la plage B2 :D9, puis cliquer sur :

Clic-droit sur un point de la série $(x_i;y_i)$ pour entrer dans le format **série de données**, puis cliquer sur :

Ajouter la droite de régression pour le nuage de points $(x_i;z_i)$.

Remarque :

pour faire apparaître l'équation de la courbe de tendance, clic-droit sur la droite, puis sur :

Format de la courbe de tendance | Option | Afficher l'équation sur le graphique | Ok

Question : Donner les équations des droites de régressions ainsi obtenues. Les ajustements affines semblent-ils appropriés ?

Pour aller plus loin

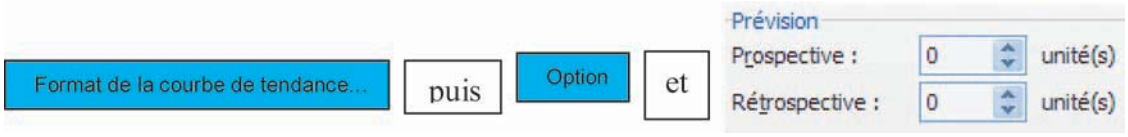
On cherche un autre ajustement pour la série $(x_i ; y_i)$.

a) Clic-droit sur un point de la série $(x_i ; y_i)$ et cliquer sur

Choisir un ordre 2, 3, 4, …, correspondant au degré du polynôme.

b) Déterminer, de même, la courbe de tendance la mieux adaptée au nuage de points $(x_i ; z_i)$.

c) Visualiser la prévision sur le graphique : pour cela, clic-droit sur la courbe de tendance, puis sur :



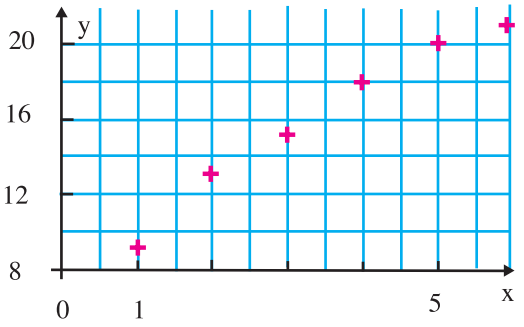
En déduire une estimation du taux de chômage, en 2004 puis 2007, au Japon.

d) Reproduire la prévision pour le taux de chômage, en 2004, en France.

e) Comparer ces résultats à ceux trouvés au début, et au taux de chômage réel de 9% en France.

VRAI-FAUX

1 Une série statistique est représentée par le nuage de points $M_i(x_i ; y_i)$ ci-dessous.



On note G le point moyen de la série.

1) Le nuage de points est celui de la série :

a) $\{9;\ 13;\ 15;\ 18;\ 20;\ 21\}$

b) $\{(1;9),\ (2;13),\ (3;15),\ (4;18),\ (5;20),\ (6;21)\}$

c) $\{(1991;19),(1992;13),(1993;15),$

$(1994;18),(1995;20),(1996;21)\}$

2) Pour déterminer l'ordonnée de G, on calcule :

a) la médiane des y_i **b)** la moyenne des y_i **c)** la demi somme y_1 et y_6 .

3) Le point moyen est :

a) G (3,5 ; 16) **b)** G(3,5 ; 16,5) **c)** G (3,5 ; 15).

4) D'aprés la forme du nuage, on peut envisager un ajustement par une fonction :

a) de la forme $ax^2 + bx + c$, où a est positif.

b) de la forme $ax^2 + bx + c$, où a est négatif.

c) affine de la forme $ax + b$, où a est positif.

d) affine de la forme $ax + b$, où a est négatif.

5) La meilleure droite d'ajustement du nuage est :

a) la droite passant par le premier et le dernier point du nuage.

b) la droite de MAYER.

c) la droite de régression de y en x.

6) La droite de régression de y en x par la méthode des moindres carrés est la droite obtenue en :

a) minimisant $S = \sum (y_i - (ax_i + b))$

b) minimisant $S = \sum (y_i - (ax_i + b))^2$

c) minimisant $S = \sum |y_i - (ax_i + b)|$

7) L'équation de la droite de régression de y en x du nuage est :
a) $y = 2,4x + 7,6$ b) $y = 7,6x + 2,4$ c) $y = -2,4x + 7,6$.

8) Un ajustement affine d'une série statistique à deux variables est :
a) toujours possible **b)** toujours pertinent **c)** le meilleur ajustement du nuage

2 Le tableau ci-dessous donne les effectifs d'une série statistique double

x_i	14	20	28	30	36	45	50
y_i	8	10	17	23	29	32	40

1) Appliquer la méthode de Mayer pour déterminer un ajustement affine de cette série statistique double.
2) Tracer cette droite de Mayer.
3) Calculer les coordonnées de G point moyen du nuage.
4) Vérifier que G appartient à la droite (G_1G_2) .

3 Pour chacune des séries doubles suivantes, calculer le cœfficient de corrélation linéaire et dire si un ajustement affine est justifié.

a)

x_i	1	2	3	4
y_i	3	5	4	5

b)

x_i	1	2	3	4
y_i	6,5	4	2,5	1

c)

x_i	8,4	8,5	8,6	8,7	8,8
y_i	8,1	8,3	8,6	8,9	9,2

d)

x_i	1	2	3	4	5
y_i	54	48	70	60	130

4 On donne les moyennes trimestrielles des 22 éléves de 4^{éme} sciences de l'informatique.

Math	6	10	10	12	2	11	7	8	9	11	11
Info.	10	7	10	11	9	13	10	10	6	7	12
Philo	11	11	13	12	11	14	11	9	12	11	13

Math	8	10	1	10	6	11	8	9	15	11	7
Info.	10	11	9	7	12	10	11	13	10	10	6
Philo	12	15	15	13	13	10	11	11	11	13	11

a) Calculer le cœfficient de corrélation linéaire entre les notes de mathématiques et celles de l'informatique.
b) Même question pour les notes de l'informatique et celles de philosophie. Conclusion.

5 Le tableau suivant donne la consommation mondiale de sucre (en millions de tonnes) entre 1900 et 2004 pour certaines années.

1900	8,1
1910	12,3
1920	13
1930	24,7
1940	26,7
1950	29,4
1960	49,3
1970	70,5

1980	88,6
1995	116,6
2000	129
2001	130,7
2002	135,7
2003	139,2
2004	143,3

a) Représenter cette série chronologique par un nuage de points.
b) Déterminer le point moyen du nuage en arrondissant au dixiéme.
c) La croissance de la consommation de sucre de 1900 à 2004 peut-elle être modélisée par une fonction affine ?

6 Un professeur de terminale a constaté une certaine relation entre la note en mathématiques et le temps de travail hebdomadaire dans cette matière. Voici selon lui la note sur 20 que peut espérer un «éléve en fonction de son temps de travail » en minutes.

Temps	5	60	180	300
Note	3	8	14	18

Trouver une courbe qui ajuste au mieux ces données et qui soit compatible avec la réalité.

VRAI - FAUX

7 Préciser s'il est possible que les droites D et D', dont une équation est donnée ci-dessous, soient les deux droites de régression d'une série statistique double. Sinon dire pourquoi.

a) $y = 12{,}5x + 7{,}2$ et $x = -10^{-2}y + 3$

b) $y = 3x + 5$ et $x = \frac{1}{3}y - 5$

c) $y = 7x - 14$ et $x = \frac{13}{8}y$

d) $y = 0{,}5x$ et $y = 3x$

e) $y = -7$ et $x = 5$

8 On se propose d'étudier l'influence de la température sur la durée d'incubation des oeufs de grenouilles. On choisit 6 échantillons de 200 oeufs chacun. Le nombre x d'éclosion au 22éme jour est le suivant :

Température t_i d'incubation en d°C	6	6,4	6,8	7,2	7,6	8
Nombre x_i d'éclosions à la température t_i	131	144	157	170	190	189

1) Dessiner le nuage des données et tracer « à l'oeil » une droite D qui a l'air de bien approcher ce nuage.
2) Calculer le cœfficient de corrélation et écrire l'équation de la droite de régression de x en t . Etudier la qualité de l'ajustement.
3) Calculer le nombre d'éclosions prédit pour un échantillon de 200 oeufs au 22éme jour pour une température de 7,5 °C.

9 On étudie un échantillon de taille n = 100 sur lequel ont été mesurés deux caractéres x et y, on a observé les résultats suivants :

$$\sum_{i=1}^{100} x_i = 800 \quad ; \quad \sum_{i=1}^{100} y_i = 1200 \quad ; \quad \sum_{i=1}^{100} x_i^2 = 7200 \quad ; \quad \sum_{i=1}^{100} y_i^2 = 16000 \quad ; \quad \sum_{i=1}^{100} x_i y_i = 10200$$

1) Déterminer l'équation de la droite de régression de y en x par la méthode des moindres carrés.
2) Déterminer l'équation de la droite de régression de x en y par la méthode des moindres carrés.

10 Pour vérifier les relations d'allométrie entre insectes, on a retenu les deux variables :
x = logarithme de la longueur de l'élytre
y = logarithme de la largeur de la tête
les mesures sur 50 insectes, notées $(x_i ; y_i)$ ont fourni les résultats suivants :

$\sum_{i=1}^{50} x_i = 155$; $\sum_{i=1}^{50} y_i = 125$; $\sum_{i=1}^{50} x_i y_i = 391{,}1$; $\sum_{i=1}^{50} x_i^2 = 482{,}5$; $\sum_{i=1}^{50} y_i^2 = 320{,}5$;

$\sum_{i=1}^{50} x_i^2 y_i^2 = 3468{,}7$

1) Calculer

a) La moyenne et l'écart-type du caractère x sur l'échantillon observé.

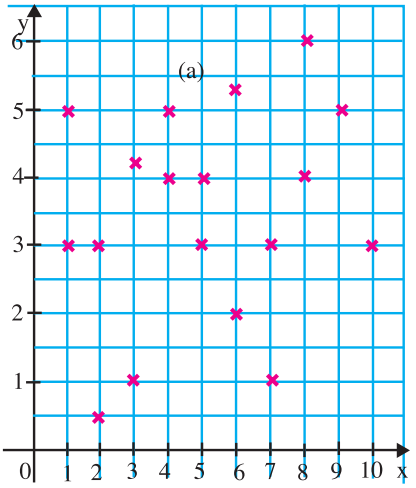
b) La moyenne et l'écart-type du caractère y sur l'échantillon observé.

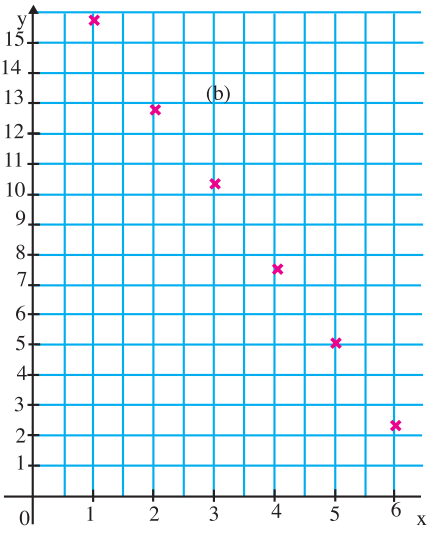
c) La covariance et le cœfficient de corrélation des variables x et y.

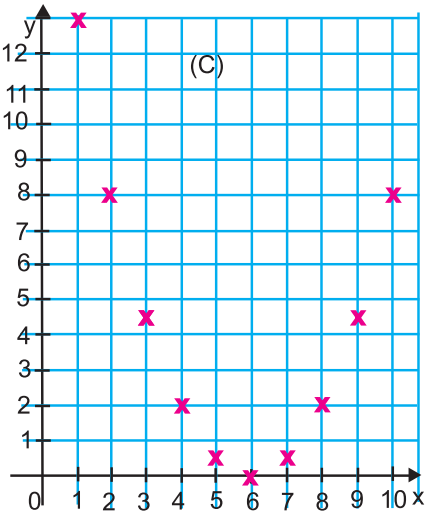
d) L'équation de la droite de régression de y en x obtenue par la méthode des moindres carrés.

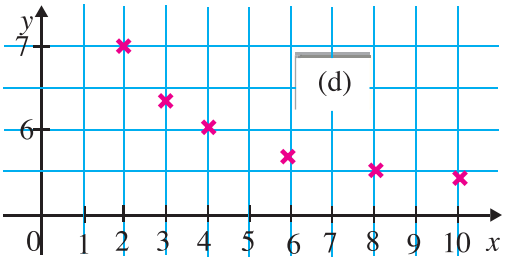
2) En déduire la loi d'allométrie exprimant la largeur de la tête en fonction de la longueur de l'élytre

11









1) Pour chacun des nuages , un ajustement affine serait-il un bon ajustement ?
2) Donner le cœfficient directeur d'une droite ajustant le nuage (b).
3) Quel nuage pourrait être ajusté par une fonction associée à la fonction inverse ?
4) Quel nuage peut être ajusté par une parabole? Donner l'équation de cette parabole .

12 Le tableau suivant donne représente l'évolution du chiffre d'affaire en milliers de dinars d'une entreprise pendant dix années.

Année	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
Rang de l'année x_i	0	1	2	3	4	5	6	7	8	9
Chiffre d'affaires y_i	110	130	154	180	190	210	240	245	270	295

1) Représenter le nuage de points $M_i\left(x_i;y_i\right)$
On choisira un repère orthogonal ayant pour unités 2cm en abscisse et 1cm pour 20 milliers de dinars en ordonnée.
2) Quel est, en pourcentage, l'augmentation du chiffre d'affaires entre les années 1995 et 2004 ?
3) Soit G le point moyen du nuage. Calculer les coordonnées de G et placer G sur le dessin.
4) Justifier qu'il est judicieux de procéder pour cette série à un ajustement affine.
Donner l'équation de la droite d'ajustement D obtenue par la méthode des moindres carrés.
Vérifier que G appartient à la droite D et tracer D sur le dessin.
5) En admettant que l'évolution continue au même rythme et en utilisant l'ajustement affine, quel chiffre d'affaires peut-on atteindre pour l'année 2010 ?
6) On suppose qu'à partir de l'année 2004, le chiffre d'affaires progresse de 8% par an.
Quel est alors le chiffre d'affaires prévisible en 2010 ?

13 On a mesuré les variables x et y sur 10 individus et obtenu les résultats suivants :

Individu n°i	1	2	3	4	5	6	7	8	9	10
x_i	18	20	19	16	19	16	19	21	15	17
y_i	43	110	70	17	91	29	80	134	15	34

On cherche une relation logarithmique entre x et y du type : $y = a\ln x + b$
Pour cela on pose : $z = \ln x$
1) Représenter graphiquement le nuage de points $(z_i;y_i)$ et déterminer la droite de régression linéaire de y en z .
2) Quelle valeur de y peut-on prédire pour un individu présentant la valeur $x = 22$?

14 Une étude de marché sur un produit de grande nécessité a conduit à la série statistique $(x_i ; y_i)$ où x_i est le prix en dinars au kg de ce produit et y_i la quantité demandée en centaines de tonnes.

Prix x_i	10	11,5	12	13	13,7	15	16,5	18,8	20
Quantité y_i	4,7	4,1	4	3,7	3,5	3,2	2,9	2,6	2,4

1) a) Représenter le nuage de points $M_i(x_i ; y_i)$ dans un repère orthogonal : unités 1cm pour 1D en abscisse et 2cm pour 100 tonnes en ordonnée. Un ajustement affine est-il justifié ?
b) Donner l'équation réduite de la droite de régression Δ de y en x. On arrondira les coefficients à 10^{-2} .
Tracer cette droite sur le graphique.
Calculer la quantité demandée pour un prix de 24, 5D par kg.

2) On pose $z = \dfrac{100}{y}$ et on se propose d'établir un autre ajustement.

a) Calculer les valeurs z_i , arrondies à 0,1 près.
b) Déterminer l'équation réduite de la droite de régression de z en x , sous la forme $z = ax + b$ où a et b sont arrondis à l'unité.
c) En déduire la fonction f qui au prix associe la quantité demandée y suivant cet ajustement et vérifier que $f(24,5) = 2$.
3) Pour un prix de 24,5 D par kg, on sait que la demande est de 210 tonnes.

Quel est l'ajustement le plus judicieux ?

15 Une maison d'édition *a* ouvert le 1er janvier 2002, sur Internet, un site de vente par correspondance.
Le tableau suivant donne l'évolution du nombre de livres vendus par mois en milliers.

Mois	Janvier 2002	Janvier 2003	Juillet 2003	janvier 2004	Juillet 2004
Rang du mois x_i	1	13	19	25	28
Nombre de livres y_i	1,2	2,5	3,5	5,1	6

1) Représenter le nuage de points (x_i, y_i) dans un repère (unités graphiques : 1cm représente deux mois en abscisse et 1cm représente 500 livres en ordonnée).
2) L'allure du nuage permet d'envisager un ajustement exponentiel plutôt qu'un ajustement affine. Pour cela, on pose $z_i = \ln(y_i)$.
Aprés l'avoir recopié, compléter le tableau suivant où z_i est arrondi 10^{-3}

Rang du mois x_i	1	13	19	25	28
$z_i = \ln(y_i)$.			1,253		

3) Ecrire une équation de la droite d'ajustement affine D de z en x par la méthode des moindres carrés.
4) Déduire une relation entre y et x de la forme : $y = \alpha\, e^{kx}$.
5) En supposant que l'évolution se poursuive de cette façon :
a) Donner une estimation à l'unité prés du nombre de livres qui seront vendus en janvier 2205.
b) A partir de quel mois peut-on prévoir que le nombre de livres vendus dépasse 13 000 ?
6) On admet que le nombre moyen m de livres vendus chaque mois entre janvier 2002 et avril 2004 est donné par la formule : $\frac{1}{28}\int_0^{28} 1{,}14e^{0{,}06}dx$
Calculer m. On donnera la valeur exacte de m, puis une valeur approchée à l'unité près

16 Le tableau suivant donne les indices des prix à la consommation pour les années 1990 à 1997.

Année	90	91	92	93	94	95	96	97
Rang de l'année x_i	0	1	2	3	4	5	6	7
Indice y_i	100	103,2	105,7	107,9	109,7	111,6	113,6	115,2

1) Représenter le nuage de points associé à la série statistique (x_i, y_i) dans un repère orthogonal. Calculer les coordonnées du point moyen et placer ce point.
2) Donner une équation de la droite d'ajustement affine D par la méthode des moindres carrés. Représenter la droite D dans le repére précédent.

3) On envisage l'ajustement du nuage par une branche de parabole d'équation $y = ax^2 + bx + c$ et l'on cherche les trois nombres a, b, c . Pour cela on pose $z_i = \sqrt{1198 - 10y_i}$ une équation de la droite d'ajustement affine de z en x par la méthode des moindres carrés est alors : $z = -x + 14$.

a) Vérifier que $y = -0{,}1x^2 + 2{,}8x + 100{,}2$.

b) Dans le repére précédent, et sans étudier la fonction correspondante, tracer la branche de la parabole d'équation $y = -0{,}1x^2 + 2{,}8x + 100{,}2$ pour x appartenant à l'intervalle $[0;7]$.

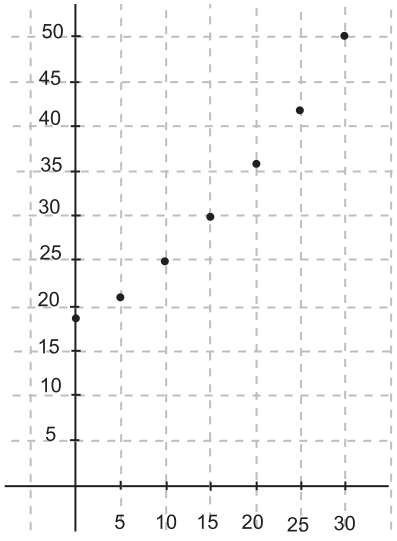
c) En choisissant ce dernier ajustement, quelle prévision de l'indice des prix à la consommation pouvait-on faire fin 1997 pour 1998 ?

d) On sait aujourd'hui que l'indice des prix à la consommation en 1998 était de 116. Calculer le pourcentage de l'erreur commise en utilisant la prévision trouvée en 3. c).

17 Le tableau suivant donne la population d'une nouvelle ville entre les années 1970 et 2000.

Année	1970	1975	1980	1985	1990	1995	2000
Rang de l'année x	0	5	10	15	20	25	30
Population en milliers d'habitants y	18	21	25	30	36	42	50

Le nuage de points associé à ce tableau est représenté graphiquement ci-dessous ; le rang x de l'année est en abscisse et la population y en ordonnée.



1) a) Déterminer une équation de la droite d'ajustement affine de y en x par la méthode des moindres carrés (les coefficients seront arrondis au centième). Tracer cette droite sur le graphique ci-contre.

b) Déduire de cet ajustement une estimation de la population en 2003, à un millier près.

2) a) L'allure du nuage incite à chercher un ajustement par une fonction définie sur $[0;+\infty[$ par $f(x) = ae^{bx}$ où a et b sont des réels. Déterminer a et b tels que $f(0) = 18$ et

On donnera une valeur arrondie de b au millième.

b) Déduire de cet ajustement une estimation de la population en 2003, à un millier près.

c) Tracer la courbe représentative de f sur le graphique .

d) La population en 2003 était de 55 milliers. Lequel des deux ajustements vous semble le plus pertinent ? Justifier votre choix.

3) On considère maintenant que, pour une année, la population est donnée en fonction du rang x par $f(x) = 18e^{0,034x}$

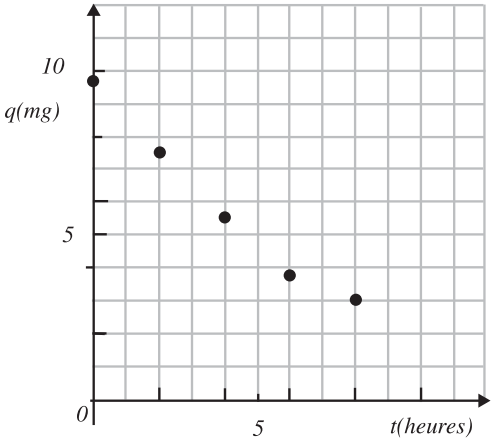
a) Calculer la valeur moyenne de la fonction f sur $[0;30]$; on donnera le résultat arrondi au dixiéme.

b) A l'aide d'une lecture graphique, déterminer l'année au cours de laquelle la population atteint cette valeur moyenne.

18 Un médicament est injecté par voie intraveineuse. Dans les heures qui suivent, la substance est éliminée par les reins. La quantité présente dans le sang (en milligrammes) à l'instant (en heures) a été mesurée par des prises de sang toutes les deux heures : q_i t_i

Le nuage de points associé à la série statistique $(t_i;q_i)$ est représenté dans un repère orthogonal ci-contre.

t_i (heures)	0	2	4	6	8
q_i (mg)	9,9	7,5	5,5	3,9	3



1) a) Déterminer une équation de la droite D d'ajustement affine de q en t par la méthode des moindres carrés (coefficients arrondis à 10^{-1}).

b) En supposant que ce modèle reste valable pendant 12 heures, quelle estimation obtient-on de la quantité de médicament présente dans le sang au bout de 12 heures ? Qu'en pensez-vous ?.

2) a) On pose $y_i = \ln q_i$. Recopier et compléter le tableau ci-dessous (valeurs arrondies au centième).

t_i					
y_i					

b) Déterminer une équation de la droite d'ajustement affine de y en t par la méthode des moindres carrés (coefficients arrondis au centième).

c) Montrer que l'expression de q en fonction de t obtenue à partir de cet ajustement est de la forme $q = ae^{-bt}$.

d) Etudier le sens de variation de la fonction f définie sur $[0;15]$ par : $f(t) = 10e^{-0,15t}$.

Tracer sa courbe représentative C .

e) On suppose que ce nouveau modèle reste valable pendant 12 heures. Calculer à 10^{-1} près la quantité de médicament présente dans le sang au bout de 12 heures.

19 Un hypermarché dispose de 20 caisses. Le tableau suivant donne le temps moyen d'attente à une caisse en fonction du nombre de caisses ouvertes :

Nombre de caisses ouvertes X	3	4	5	6	8	10	12
Temps moyen d'attente (en minutes) Y	16	12	9,6	7,9	6	4,7	4

1) Construire le nuage de points $M_i(x_i;y_i)$ correspondant à cette série statistique (unités graphiques : 1cm pour une caisse ouverte, 1cm pour une minute d'attente).

2) Calculer les coordonnées du point moyen G du nuage et le placer sur le graphique.

3) Un ajustement affine.

a) Calculer le cœfficient de corrélation linéaire r .

b) Déterminer l'équation de la droite de régression linéaire D de y en x par la méthode des moindres carrés. Tracer la droite D sur le graphique.

c) Estimer à l'aide d'un calcul utilisant l'équation de la droite D :

- Le nombre de caisses à ouvrir pour que le temps moyen d'attente à une caisse soit 5 minutes.
- Le temps moyen d'attente à la caisse lorsque 15 caisses sont ouvertes.
- Pensez-vous que, dans le cas du dernier résultat, l'ajustement affine soit fiable ?

4) Un ajustement non affine.

On considère la fonction f définie sur $]0;+\infty[$ par $f(x)=\dfrac{\lambda}{x}$.

a) Déterminer λ de façon à avoir : $f(3)=16$.

b) Tracer alors la représentation graphique C de f dans le repère utilisé pour le nuage.

c) Estimer à l'aide d'un calcul utilisant la fonction f :

- Le nombre de caisses à ouvrir pour que le temps moyen d'attente à une caisse soit de 5 minutes.
- Le temps moyen d'attente à la caisse lorsque 15 caisses sont ouvertes.

Une brève histoire des outils statistiques

Comment interpréter «l'avalanche de chiffres» de la réalité statistique sans outils théoriques? L'humanité a mis fort longtemps avant de découvrir des procédés de calcul efficaces et des représentations pertinentes. Depuis, ces outils ont envahi tous les domaines de la connaissance.

L'astronomie semble bien être mère de toutes les sciences ; les statistiques ne font pas exception.

Si on cherche l'origine du besoin d'ordonner des observations, de les représenter par des tableaux et des graphiques, de rechercher des valeurs typiques, de construire des outils spécifiques, on constate que c'est en astronomie que sont apparus ces concepts.

Il y a près de 2500ans, les Babyloniens ont établi des procédés destinés à mesurer les mouvements des planètes et du soleil sur des bases de relevés à intervalles réguliers.

La loi statistique ne s'impose pas à l'esprit humain avec le même caractère de nécessité que les lois naturelles.

Emile Borel

Ptolémée, astronome grec du IIe siécle, a développé son système en se basant surlesdécouvertes d'Aristarque de Samos et d'Hipparque, nés avant lui et qui avaient obtenu des résultats importants sur la position des étoiles et la périodicité de leur retour. Le fait de posséder plusieurs valeurs observées a conduit ces astronomes à proposer des valeurs uniques accompagnées de mesure de variation.

Les valeurs typiques d'une série univariée

Le choix de « valeur typique d'une série » est donc un problème très ancien. Il semble que les premiers paramètres de position qui aient été utilisés soient le mode valeur apparaissant le plus fréquemment, et le « milieu de l'intervalle défini par les valeurs extrêmes».

La «moyenne arithmétique» apparaît clairement dans l'oeuvre de Tycho Brahé (1546-1601) qui, en constituant un ensemble de données sur le mouvement des planètes, permit à Képler de formuler ses lois.

En 1722, Roger Cotes, qui dispose d'observations qui ne sont pas toutes aussi fiables, propose d'utiliser une moyenne pondérée dont les coefficients sont inversement proportionnels à la dispersion des erreurs d'observations.

On peut noter que la médiane voit naître son intérét à la même époque, en 1757 et que la moyenne géométrique et la moyenne harmonique ont été introduites en Angleterre en 1874.

Si l'idée de la variabilité n'est pas récente puisque des astronomes grecs l'ont utilisée, c'est Galileo Galilée qui, en 1632, s'intéressant à la détermination de la distance entre la terre et une étoile nouvelle, dit clairement pour la première fois : «que ces observateurs ont tous fait des erreurs…que ces erreurs doivent être corrigées… nous nous consacrerons à appliquer les modifications minimales et les corrections les plus petites possibles, juste pour sortir les observations de l'impossible et les remettre dans le possible…».

La variance naît au XIXe siècle avec les moindres carrés ; Gauss lui préfère l'écart-type.

Ajustement, corrélation et régression

Le problème de l'ajustement d'un ensemble de points représentés dans un système d'axes par une droite, ou plus généralement par une courbe, est essentiel dans le développement de la statistique.

Au XVIIIe siècle, Leonhard Euler et Johan Tobias Maver, développant, indépendamment l'un de l'autre, la méthode des moyennes permettant d'ajuster des points par une droite.
Le premier texte paru faisant mention de la méthode des moindres carrés est dû à Adrien Marie Legendre dans un article sur ses « nouvelles méthodes pour la détermination des orbites des Comètes » publié en 1805. Un an plus tard Gauss fait aussi allusion à cette méthode.
C'est avec l'apparition de la «loi normale» que cette mëthode va trouver sa justification et va devenir pour longtemps La Méthode d'ajustement.
La paternité de la corrélation a donné lieu a une abondante littérature. Signalons simplement que Galton exprime le désir de construire un cœfficient de réversion qui se mutera en régression et qu'en 1888 il utilise les termes de Partial co-relation annonçant déjà à la corrélation multiple.
En 1896, Karl Pearson reprend les concepts de Galton pour leur donner leur forme actuelle.
Au XX^{é} siècle d'autres mesures d'association allaient naître comme, en 1904 le cœfficient de corrélation de rang avec Spearman et la même année la statistique «classique» du chi-deux par Karl Pearson (encore lui !).

Tangente J073-n° 77-Octobre -Novembre 2000.