Chapitre 7

Statistiques

Huygens (1669) : Espérance de vie.

"Par les observations faites à Londres avec beaucoup d'exactitude.
De 100 personnes conçues, il en meurt [...].
Donc, de 100 personnes, ceux qui atteignent l'âge de 6 ans sont 64, de 16 ans sont 40, de 26 ans sont 25, de 36 ans sont 16, de 46 ans sont 10, de 56 ans sont 6, de 66 ans sont 3, de 76 ans est 1 et de 86 ans est 0.
Qui gagerait qu'un enfant conçu virait jusqu'à 6 ans, peut mettre 64 contre 36 ou 16 contre 9. Et qui gagerait [...].
De 100 enfants conçus, il en meurt 36 avant l'age de 6 ans, lesquels on peut dire ont vécu l'un portant l'autre 3 ans.
Des 64 restants, il en meurt 24 avant l'âge de 16 ans [...]."
Une correspondance de Huygens sur la statistique démographique.
Huygens aboutit au total 1822 en multipliant 36 par 3, 24 par 11, jusqu'à 1 par 81 et en ajoutant tous les produits ainsi obtenus puis calcul le quotient de 1822 par 100 et déclare :" Et le quotient qui est ici 18 ans et environ 2 mois et demi, ce n'est pas à dire qu'il soit apparent qu'il vivra si longtemps, car il est beaucoup plus apparent qu'il mourra avant ces termes."

(J Dhombres et al,
Mathématiques au fil des âges,
1987).

Chapitre 7

Statistiques

Il arrive que l'on soit amené à effectuer deux séries de mesure X et Y sur un même échantillon composé de n individus et que l'on s'interroge sur les relations possibles entre ces mesures.

On dit alors que l'on a une série statistique double (X, Y).

I. Distributions marginales

Activité 1

On a relevé dans le tableau ci-dessous, l'intensité de travail X (en kilojoules par minute) et la fréquence cardiaque Y de 100 personnes.

Y \ X	9.6	12.8	18.4	31.2	36.8	47.2	49.6	56.8	Total
70	4	2							6
86	3	5	6	4	1			2	21
90	2	5	12	3	4	1	1		28
104		1	12	14	8	5	3	2	45
Total	9	13	30	21	13	6	4	4	100

1. a. Quelle est la signification du nombre 12 encadré dans le tableau ?
   b. Quelle est le nombre d'individus dont la fréquence cardiaque est supérieure à 100 ?
   c. Quelle est le nombre d'individus qui ont fourni un travail d'intensité supérieure à 49 ?
   d. Quelle est le nombre d'individus qui ont fourni un travail d'intensité supérieure à 49 et ayant une fréquence cardiaque supérieure à 100 ?
2. Déterminer la distribution marginale de X, puis calculer la moyenne $\overline{X}$ et l'écart- type σ_X.
3. Déterminer la distribution marginale de Y, puis calculer la moyenne $\overline{Y}$ et l'écart- type σ_Y.

Activité 2

On a recueilli dans le tableau ci-contre la distance parcourue avant la première grande panne et la puissance en chevaux de 20 voitures.

1. a. Quelle sont les valeurs prises par Y ?

   b. Calculer la moyenne $\overline{Y}$ et l'écart type σ_Y.

2. Compléter le tableau suivant donnant la répartition des 20 voitures suivant la distance parcourue.

X (distance parcourue en mille km)	Effectifs
Moins de 50	
$[50,60[$	
$[60,70[$	
70 et plus	

3. Déterminer le pourcentage des voitures ayant parcouru une distance inférieure à 60000 km et qui ont une puissance supérieure ou égale à 6 chevaux.

(X) Distance parcourue en mille km	(Y) Puissance en chevaux
42	4
55	5
57	6
81	6
64	4
70	7
75	6
58	5
61	4
65	5
48	4
58	4
65	4
72	6
75	4
80	7
65	7
73	5
43	6
61	5

Définitions

Soit (X,Y) une série statistique double sur un échantillon de taille n et soit $(x_i,y_i)_{1\le i\le n}$ les valeurs numériques prises respectivement par les variables X et Y.

La distribution marginale de la variable X est la distribution des valeurs $(x_i)_{1\le i\le n}$ prises par la variable X.

La distribution marginale de la variable Y est la distribution des valeurs $(y_i)_{1\le i\le n}$ prises par la variable Y.

Soit X une série statistique sur un échantillon de taille n.

Si $\overline{X}$, $V(X)$ et σ_X désignent respectivement la moyenne, la variance et l'écart-type de la série, alors $\overline{X}=\frac{1}{n}\sum_{i=1}^{p} n_i x_i$, $V(X)=\frac{1}{n}\sum_{i=1}^{p} n_i\left(x_i-\overline{X}\right)^2$, $\sigma_X=\sqrt{V(X)}$, où les valeurs $x_1, x_2, ..., x_p$ désignent les valeurs distinctes prise par la variable X si elle est discrète, ou les centres des classes si la variable X est continue. L'entier n_i désigne l'effectif de la valeur x_i.

Activité 3

Dans le tableau suivant, on a reproduit les effectifs d'individus d'un échantillon selon leur poids X (en kg)et leur taille Y(en cm).

Y \ X	[40,45[	[45,50[	[50,55[	[55,60[	Effectif selon la taille	Fréquence selon la taille
[120,155[	20	9	1	0	30	0.30
[155,160[	2	18	4	1	25	0.25
[160,165[	0	5	12	6	23	0.23
[165,170[	0	1	7	14	22	0.22
Effectif selon le poids	22	33	24	21	100	1
Fréquence selon le poids	0.22	0.23	0.24	0.21	1	

1. Déterminer la distribution marginale de X et celle de Y.
2. Calculer la moyenne $\overline{X}$ et l'écart- type σ_X de la variable X.
3. Calculer la moyenne $\overline{Y}$ et l'écart- type σ_Y de la variable Y.

II. Covariance d'une série statistique double

II. 1 Cas d'un échantillon simple

Activité 1

Dans le tableau ci-dessous, on a relevé les exportations (en million de dinars) et les importations (en million de dinars) mensuelles de la Tunisie pour l'année 2006.

Mois	exportations (X)	importations (Y)
Janvier	1081.1	1312.1
Février	1225.6	1367.6
Mars	1378.6	1641.6
Avril	1193.7	1613.1
Mai	1205.8	1827.3
Juin	1374.6	1705.8
Juillet	1283.8	1713.4
Août	1157.8	1494.1
Septembre	1349.4	1859.8
Octobre	1230.1	1668.1
Novembre	1488.5	1902.6
Décembre	1347.3	1660.6

1. Déterminer la taille de l'échantillon étudié.
2. a. Calculer la moyenne $\overline{X}$ et l'écart- type σ_X de la variable X.

   b. Calculer la moyenne $\overline{Y}$ et l'écart- type σ_Y de la variable Y.
3. Calculer $\frac{1}{12}\sum_{i=1}^{12} x_i y_i - \overline{X}\overline{Y}$.

Définition

Soit (X,Y) une série statistique double sur un échantillon de taille n.

On appelle covariance de (X,Y) le réel, noté $\text{cov}(X,Y)$ défini par

$\text{cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{XY}$, où (x_i, y_i) est la valeur observée pour l'individu i si X et Y sont discrètes, ou le centre de la classe si l'une des variables est continue.

Il découle de la définition que $\text{cov}(X,Y) = \text{cov}(Y,X)$.

Interprétation de la covariance

La covariance mesure la tendance qu'ont les variables X et Y à varier ensemble.
La covariance est positive si X et Y ont tendance à varier dans le même sens.
La covariance est négative si X et Y ont tendance à varier en sens contraire.

Activité 2

On a relevé dans le tableau suivant le nombre de logements (en milliers) et le nombre de logements modernes (villa, appartement) durant quelque années.

Année	1984	1989	1994	1999	2004
X : Nombre de logements (en milliers)	1313.1	1512.3	1870.1	2204.3	2501.0
Y : Nombre de logements modernes(en milliers)	265.2	343.3	630.2	848.7	1128.0

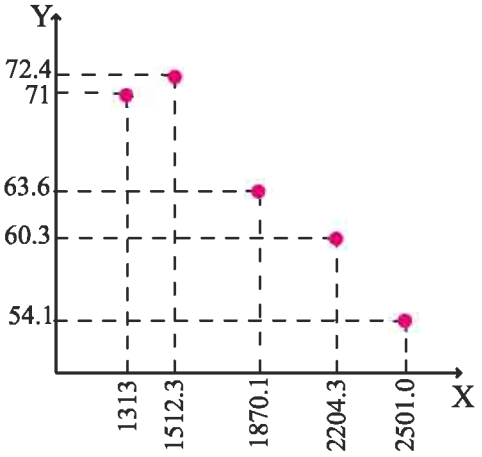
1. Représenter le nuage de points de la série (X,Y).
2. a. Calculer $\overline{X}$ et $\overline{Y}$.

   b. Calculer $\text{cov}(X,Y)$. Interpréter le résultat.

Activité 3

Dans le graphique ci-contre, on a représenté les points $M(X,Y)$, où X désigne le nombre de logements (en milliers) et Y le pourcentage des logements traditionnels pour la même année.

Quel est le signe de $\text{cov}(X,Y)$?



II. 2 Cas d'un échantillon groupé

Définition

Soit (X,Y) une série statistique double de taille n.

Soit n_{ij} le nombre de fois qu'apparaît le couple (x_i,y_j).

Alors $\text{cov}(X,Y)=\frac{1}{n}\sum_{j=1}^{q}\sum_{i=1}^{p}n_{ij}x_iy_j-\overline{X}\,\overline{Y}$.

Exercice résolu

Le tableau ci-dessous donne le poids Y (en kg) de 63 nouveaux-nés ainsi que le poids maternel X.

Y \ X	]40,50]	]50,60]	]60,70]	]70,80]	Total
]1.5,2.5]	1	0	1	0	2
]2.5,3.5]	11	17	13	2	43
]3.5,4.5]	4	4	8	2	18
Total	16	21	22	4	63

1. Calculer $\overline{X}$ et σ_X, ainsi que $\overline{Y}$ et σ_Y.
2. Déterminer la covariance de X et Y. Interpréter.

Solution

1. • Etude de la variable X

x_i (centres des classes de la variable X)	45	55	65	75	
n_i	16	21	22	4	$\sum_{i=1}^{4} n_i = 63$
x_i^2	2025	3025	4225	5625	
$n_i x_i$	720	1155	1430	300	$\sum_{i=1}^{4} n_i x_i = 3605$
$n_i x_i^2$	32400	63525	92950	22500	$\sum_{i=1}^{4} n_i x_i^2 = 211375$

Le calcul donne

$$\overline{X} = \frac{1}{63}\sum_{i=1}^{4} n_i x_i \simeq 57.2222,\ V(X) = \frac{1}{63}\sum_{i=1}^{4} n_i x_i^2 - \left(\overline{X}\right)^2 \simeq 80.776,$$

$$\sigma_X = \sqrt{V(X)} \simeq 8.9875.$$

• Etude de la variable Y

y_j	2	3	4	
n_j	2	43	18	$\sum_{j=1}^{3} n_j = 63$
y_j^2	4	9	16	
$n_j y_j$	4	129	72	$\sum_{j=1}^{3} n_j y_j = 205$
$n_j y_j^2$	8	387	288	$\sum_{j=1}^{3} n_j y_j^2 = 683$

Le calcul donne

$$\overline{Y} = \frac{1}{63}\sum_{j=1}^{3} n_j y_j \simeq 3.2539\ ,\ V(Y) = \frac{1}{63}\sum_{j=1}^{3} n_j y_j^2 - \left(\overline{Y}\right)^2 \simeq 0.2529\ ,\ \sigma_Y = \sqrt{V(Y)} \simeq 0.5029.$$

2. Dressons les couples distincts des valeurs observées et leurs effectifs.

Couples (x_i, y_j)	(45,2)	(45,3)	(45,4)	(55,3)	(55,4)	(65,2)	(65,3)	(65,4)	(75,3)	(75,4)
Effectifs n_{ij}	1	11	4	17	4	1	13	8	2	2
$n_{ij}.x_i.y_j$	90	1485	720	2805	880	130	2535	2080	450	600

Le calcul donne $\sum_{j=1}^{3}\sum_{i=1}^{4} n_{ij}x_i y_j = 11775$.

D'où $\text{cov}(X,Y) = \frac{1}{63} \times 11775 - \overline{X}\,\overline{Y} \simeq 0.7$.

Interprétation

La covariance est positive donc X et Y ont tendance à varier dans le même sens.

Utilisation d'une calculatrice

Les calculatrices et les ordinateurs actuels permettent de retrouver les résultats précédents. A titre d'exemple, on donne le mode d'emploi d'une calculatrice.

- Pour choisir le mode de fonctionnement en statistique appuyer sur [MODE] [1].
- Appuyer sur [1] pour sélectionner le sous mode statistique à deux variables.
- Pour entrer les données taper [x_i] [STO] [y_j] [STO] [n_{ij}] [M+]

Par exemple pour le couple $(55,3)$ taper [55] [STO] [3] [STO] [17] [M+].

- On appuie sur [RCL] [n] la calculatrice affiche 63.
- On appuie sur [RCL] [$\sum x$] la calculatrice affiche 3605 (la valeur de $\sum_{i=1}^{4} n_i x_i$).
- On appuie sur [RCL] [$\sum x^2$] la calculatrice affiche 211375 (la valeur de $\sum_{i=1}^{4} n_i x_i^2$).
- On appuie sur [RCL] [$\overline{X}$] la calculatrice affiche 57.22222222.
- On appuie sur [RCL] [σ_X] la calculatrice affiche 8.987547725.
- On appuie sur [RCL] [σ_X] [x^2] la calculatrice affiche 80.77601411 (la valeur de V(X)).
- On appuie sur [RCL] [$\sum xy$] la calculatrice affiche 11775 (la valeur de $\sum_{j=1}^{3}\sum_{i=1}^{4} n_{ij}x_i y_j$).
- On appuie sur [RCL] [$\sum xy$] [÷] [63] [−] ([RCL] [$\overline{X}$] [×] [RCL] [$\overline{Y}$]) la calculatrice affiche 0.705467372 (la valeur de $\text{cov}(X,Y)$).

Activité 4

Dans une population de 100 ménages, on a considéré le nombre d'enfants X et le revenu du chef de famille Y (en DT).

Y \ X	0	1	2	3	4	5	Total
Moins de 200	6	4	1	0	0	0	11
$[200, 400[$	3	11	10	5	1	0	30
$[400, 600[$	1	3	16	13	4	1	38
$[600, 800[$	0	1	3	5	8	4	21
Total	10	19	30	23	13	5	100

1. a. Déterminer le nombre de ménages qui ont 4 enfants et dont le revenu est supérieur à 600 dinars.
   b. Déterminer le nombre de ménages qui n'ont pas d'enfants et ayant un revenu inférieur à 200 dinars.
   c. Déterminer le nombre de ménages qui ont moins de 4 enfants et dont le revenu est compris entre 400 et 600 dinars.
2. a. Calculer la moyenne $\overline{X}$ et l'écart- type σ_X de la variable X.
   b. Calculer la moyenne $\overline{Y}$ et l'écart- type σ_Y de la variable Y.
3. a. Peut-on prévoir le signe de la covariance de X et Y ?
   b. Calculer la covariance de X et Y.

III. Ajustement d'une série statistique double

Lorsque un statisticien étudie une série statistique double. L'une des questions qu'il se pose est : peut-on prévoir la valeur de Y lorsqu'on connaît la variable X ?
Pour répondre à une telle question, le statisticien essaiera de trouver une fonction f qui modélise le phénomène étudié, grâce à la relation $Y = f(X)$.

Dans ce cas , on dit que X est la variable explicative et Y est la variable expliquée.
La fonction f cherchée dépendra de l'allure du nuage de points.
Si le nuage de point a l'allure d'une droite, le statisticien essaiera de trouver une fonction affine f qui sera la plus proche des points du nuage. On dit que le statisticien effectue un ajustement affine.
Par conséquent faire un ajustement affine consiste à déterminer deux réels a et b tels que $Y = aX + b$ soit un modèle acceptable du phénomène étudié.
La droite d'équation $y = ax + b$ sera appelée droite d'ajustement affine de Y en X.

III. 1 Méthode de Mayer

Activité 1

On a relevé dans le tableau ci-contre, le montant total (en million de dinars) du commerce extérieur en Tunisie (importations et exportations) depuis l'année 1990 jusqu'à l'année 2004.

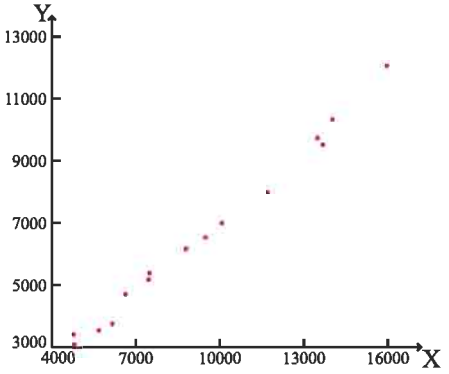
Année	Importations (X)	Exportations (Y)
1990	4826.4	3087.4
1991	4788.9	3417.1
1992	5688.8	3549.7
1993	6172.1	3760
1994	6647.3	4696.6
1995	7464.3	5172.5
1996	7498.8	5372
1997	8793.5	6147.9
1998	9489.5	6518.3
1999	10070.5	6966.9
2000	11738	8004.8
2001	13697.3	9536.2
2002	13510.9	9748.6
2003	14038.9	10342.6
2004	15960.3	12054.9

1. a. Calculer la moyenne $\overline{X}$ et l'écart-type σ_X de la variable X.

b. Calculer la moyenne $\overline{Y}$ et l'écart- type σ_Y de la variable Y.

2. On a représenté ci-contre, dans un même repère, le nuage de points de la série double (X,Y).

Placer le point $\overline{G}\ (\overline{X},\overline{Y})$.



3. On scinde l'ensemble des 15 points du nuage en deux parties. La première partie (I) correspond aux valeurs observées entre 1990 et 1997 et la deuxième partie (II) correspond aux valeurs observées entre 1998 et 2004.

On désigne par G_1 et G_2 les points moyens respectifs de la partie I et de la partie (II).

a. Déterminer les coordonnées de G_1 et G_2.

Vérifier que G, G_1 et G_2 sont alignés et tracer la droite (G_1G_2).

b. Comment semblent se répartir les points du nuage autour de la droite (G_1G_2).

c. Donner alors un ajustement affine de la série double (X,Y).

d. Donner une estimation du montant des exportations si le montant de l'importation est égal à 17000 millions de dinars.

Définition

Soit (X,Y) une série statistique double de valeurs $(x_i,y_i)_{1\le i\le n}$.

L'ensemble des points M_i de coordonnées (x_i,y_i) dans un repère orthogonal est appelé nuage de points représentant la série statistique.

Le point moyen du nuage est le point dont les coordonnées sont les moyennes $\overline{X}$ et $\overline{Y}$.

Principe de la méthode de Mayer

Soit un nuage de points représentant une série statistique double (X,Y) et G son point moyen.

On scinde le nuage de points de (X,Y) en deux parties contenant à peu prés le même nombre de points.

On considère alors les points moyens G_1 et G_2 des deux nuages obtenus.

La droite (G_1G_2) définit un ajustement affine du nuage de points représentant la série statistique double (X,Y).

La droite (G_1G_2) est appelée droite de Mayer et passe par le point moyen G du nuage global.

Activité 2

Le mur d'une habitation est constitué par une couche de béton et une couche de polystyrène d'épaisseur variable X (en cm). On a mesuré la résistance thermique R (en $m^2.°C/W$) de ce mur pour divers valeurs de X et on a obtenu les résultats ci-dessous.

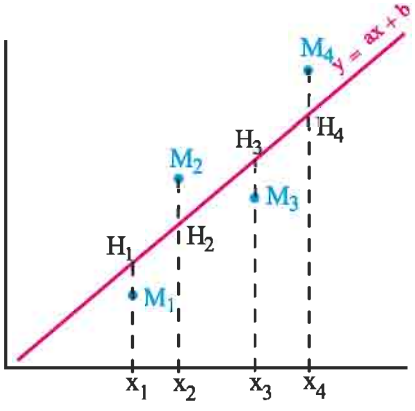
X	2	4	6	8	10	12	14	16	18
R	0.83	1.34	1.63	2.3	2.44	2.93	3.44	3.85	4.28

1. Tracer le nuage de la série (X,R).
2. Déterminer un ajustement affine de R en X par la méthode de Mayer.
3. Quelle résistance thermique peut-on espérer obtenir avec une épaisseur de polystyrène de 25 cm ?

III. 2 Méthode d'ajustement par les moindres carrés

Nous avons représenté ci-contre le nuage de points $M_i(x_i,y_i)$, $1\le i\le n$ d'une série statistique double, ainsi qu'une droite D d'équation $y=ax+b$.

Pour tout entier $1\le i\le n$, on note $H_i(x_i,z_i)$ le point de la droite D de même abscisse que M_i.



Le principe de la méthode d'ajustement par la méthode des moindres carrés consiste à déterminer les réels a et b tels que la somme $\sum_{i=1}^{n} M_i H_i^2$ soit minimale.

Dans ce cas, le statisticien pourra faire des prévisions en remplaçant la valeur observée y_i par la valeur théorique $z_i = ax_i + b$.

Activité 3

Le tableau ci-dessous donne le chiffre d'affaire annuel en mille DT d'une société pendant huit années consécutives.

Rang de l'année (X)	1	2	3	4	5	6	7	8
Chiffre d'affaires en mille DT	13.6	15	15.8	17	18	20	19	20

1. a. Représenter le nuage de points de la série (X,Y).
   b. Un ajustement affine de cette série est-il justifié ?
2. Déterminer les coordonnées du point moyen G de ce nuage.
3. a. Tracer dans le même repère la droite D d'équation $y = 1.6x + 10.1$.
   b. Calculer la somme $S_D = \sum_{i=1}^{8} \left[y_i - (1.6x_i + 10.1) \right]^2$.
4. On considère une droite Δ d'ajustement de Y par rapport à X obtenue par la méthode de Mayer.
   a. Déterminer l'équation de Δ sous la forme $y = ax + b$. (on donnera a et b à 10^{-1} prés)
   b. Calculer la somme $S_\Delta = \sum_{i=1}^{8} \left[y_i - (ax_i + b) \right]^2$.
5. Comparer S_D et S_Δ.
6. Estimer le chiffre d'affaires de cette société à sa dixième année.

Théorème (admis)

Soit (X,Y) une série statistique double sur un échantillon de taille n et telle que $\sigma_X \neq 0$.

Soit $(x_i, y_i)_{1 \le i \le n}$ les valeurs observées de la série. Alors la somme $\sum_{i=1}^{n} (ax_i + b - y_i)^2$ est minimale pour le couple (a_0, b_0) tel que $a_0 = \dfrac{\text{cov}(X,Y)}{\sigma_X^2}$ et $b_0 = \left(\overline{Y} - \dfrac{\text{cov}(X,Y)}{\sigma_X^2} \overline{X} \right)$.

Définition

Soit (X,Y) une série statistique double sur un échantillon de taille n.

La droite d'équation $y=\frac{cov(X,Y)}{\sigma_X^2}(x-\overline{X})+\overline{Y}$ est appelée droite des moindres carrés de Y en X, ou droite de régression de Y en X.

La droite d'équation $x=\frac{cov(X,Y)}{\sigma_Y^2}(y-\overline{Y})+\overline{X}$ est appelée droite des moindres carrés de X en Y, ou droite de régression de X en Y.

Conséquence

Les droites des moindres carrés de Y en X et de X en Y passent par le point moyen G du nuage associé à la série (X,Y).

Activité 1

Dans le tableau ci-dessous, on a relevé le poids (en Kg) et la surface corporelle (en m^2) correspondante de 15 sujets.

	Masse (X)	Surface corporelle (Y)
1	15	0.58
2	20	0.7
3	25	0.8
4	29	0.9
5	34	1
6	39	1.1
7	44	1.2
8	50	1.3
9	55	1.4
10	60	1.5
11	65	1.56
12	70	1.64
13	74	1.74
14	80	1.8
15	85	1.9

1. a. Calculer la moyenne $\overline{X}$ et l'écart- type σ_X de la variable X.

b. Calculer la moyenne $\overline{Y}$ et l'écart- type σ_Y de la variable Y.

2. Déterminer la covariance la série (X,Y).

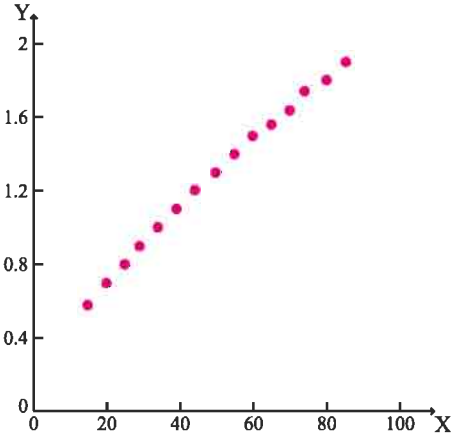
3. On a représenté ci-contre le nuage de la série (X, Y).

a. Placer le point G $(\overline{X},\overline{Y})$.

b. Comment semblent se répartir les points du nuage ?

c. Donner alors un ajustement affine par les moindres carrés de la série double (X,Y).

4. Donner une estimation de la surface corporelle d'un sujet qui pèse 62 KG.



III. 3 Coefficient de corrélation linéaire

On peut toujours au vu des formules précédentes construire une droite de régression.
Mais parfois cette dernière n'est d'aucune efficacité, dans la mesure où les prédictions que l'on fait à partir de cette droite ne sont pas raisonnables.
Pour savoir s'il est pertinent d'ajuster un nuage de point par les moindres carrés, on calcule un réel appelé coefficient de corrélation linéaire.

Définition

Soit (X,Y) une série statistique double. On appelle coefficient de corrélation linéaire le réel noté ρ_{XY} défini par $\rho_{XY} = \dfrac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$.

Propriétés

Soit (X,Y) une série statistique double. Alors $-1 \leq \rho_{XY} \leq 1$.
Le coefficient de corrélation linéaire est invariant par changement d'unité ou d'origine.

Interprétation du coefficient de corrélation linéaire

Les statisticiens conviennent que lorsque $|\rho_{XY}| > \dfrac{\sqrt{3}}{2}$, l'ajustement affine est justifié et les prédictions faites au moyen de cet ajustement sont raisonnables.

Activité 4

Le tableau suivant donne l'effectif de la population scolaire de la $3^{ème}$ année de l'enseignement secondaire du mois d'octobre 1997 au mois d'octobre 2002.

Année (X)	1997	1998	1999	2000	2001	2002
Population scolaire en $3^{ème}$ année (Y)	67755	74581	79266	76138	80123	90087

1. Calculer le coefficient de corrélation linéaire.
2. Déterminer un ajustement par les moindres carrés de la série double (X,Y) puis donner une estimation de la population scolaire en $3^{ème}$ année secondaire au mois d'octobre 2010.

Activité 5

On donne la série double suivante, relative aux voitures selon leur puissance Y et la durée des pneumatiques X (en millier de kilomètres).

Y \ X	2	3	4	
20	0	8	30	38
25	5	20	7	32
30	25	3	2	30
	30	31	39	100

1. Calculer le coefficient de corrélation linéaire.
2. Un ajustement par les moindres carrés est-il justifié ?

III. 4 Exemple d'ajustement non affine

Exercice résolu

Le tableau ci-contre indique l'évolution du personnel paramédical tunisien dans le secteur public (techniciens supérieurs, infirmiers, auxiliaires de santé) de 1990 à 2005.

Année	Paramédicaux
1990	23743
1991	24555
1992	25070
1993	25291
1994	25466
1995	25874
1996	26130
1997	26369
19989	26676
1999	27050
2000	27392
2001	30292
2002	28629
2003	29976
2004	29584
2005	29607

1. En numérotant les années de 0 à 15, déterminer les valeurs de la série double $(X,\ln Y)$, où X est le rang de l'année et Y est le nombre de paramédicaux de l'année correspondante.
2. On pose $Z=\ln Y$.
   a. Calculer le coefficient de corrélation et justifier que l'on peut procéder à un ajustement affine par les moindres carrés de la série (X,Z).
   b. Donner la droite de régression de Z en X.
3. Quel sera le nombre de paramédicaux en 2010 ?

Solution

1.

x_i	y_i	$z_i = \ln(y_i)$	x_i^2	z_i^2	$x_i z_i$
0	23743	10.075	0	101.505	0
1	24555	10.108	1	102.171	10.108
2	25070	10.129	4	102.597	20.258
3	25291	10.138	9	102.778	30.414
4	25466	10.145	16	102.921	40.580
5	25874	10.160	25	103.230	50.800
6	26130	10.170	36	103.430	61.020
7	26369	10.179	49	103.612	71.253
8	26676	10.191	64	103.860	81.528
9	27050	10.205	81	104.142	91.845
10	27392	10.218	100	104.410	102.180
11	30292	10.318	121	106.461	113.498
12	28629	10.262	144	105.310	123.144
13	29976	10.308	169	106.254	134.004
14	29584	10.294	196	105.970	144.116
15	29607	10.295	225	105.990	154.425

2. a. Le calcul donne $\overline{X} = 7.5$, $\sigma_X \simeq 4.6$, $\overline{Z} \simeq 10.199$, $\sigma_Z \simeq 0.074$.

$\text{Cov}(X,Z) \simeq 0.326$, $\rho_{XZ} \simeq 0.960$.

Le coefficient de corrélation est très proche de 1

L'ajustement est donc justifié.

b. La droite de régression est d'équation $z = \dfrac{0.326}{(4.61)^2}(x - 7.5) + 10.199$.

Soit $z = 0.015(x - 7.5) + 10.199$.

3. Le nombre de paramédicaux sera de $e^{0.015(20-7.5)+10.199} \simeq 32419$.

Utilisation d'une calculatrice

Dans cet exercice la série est à données simples.

- Pour entrer les données taper $\boxed{x_i}$ $\boxed{\text{STO}}$ $\boxed{y_i}$ $\boxed{M+}$.
- Pour afficher la valeur du coefficient de corrélation, appuyer sur $\boxed{\text{RCL}}$ $\boxed{r}$.
- Pour afficher la valeur de la pente de la droite de régression de Y en X, appuyer sur $\boxed{\text{RCL}}$ $\boxed{b}$.
- Pour afficher la valeur de l'ordonné à l'origine de la droite de régression de Y en X, appuyer sur $\boxed{\text{RCL}}$ $\boxed{a}$.

Activité 6

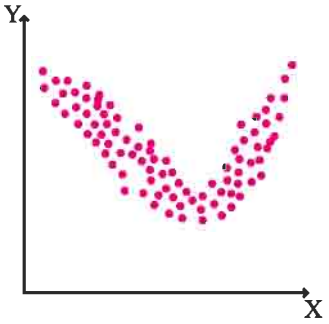
La résistance à l'avancement d'un poids lourd est une fonction de la vitesse. L'objet de cette activité est de déterminer la meilleur expression possible de cette fonction dans un intervalle de vitesse compris entre 10 km/h et 100 km/h. Cette résistance est mesurée en kW.
Les résultats de ces mesures sont consignés dans le tableau ci-dessous.

V (km/h)	10	20	30	40	50	60	70	80	90
R (kW)	2.6	5.8	9.9	15.4	23.6	34.5	49	67.2	89.1

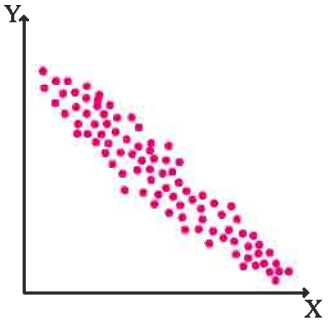
1. Dresser le tableau des valeurs de la série (X,Y) où $X=V^2$ et $Y=\frac{R}{V}$.
2. Donner le coefficient de corrélation linéaire entre X et Y et une équation de la droite de régression de Y en X.
3. En déduire une relation entre R et V.
4. Donner une évaluation de la valeur de R pour une vitesse de 100 km/h.

1 Pour chacun des graphiques suivants, indiquer si le nuage de points justifie la recherche d'un ajustement affine.

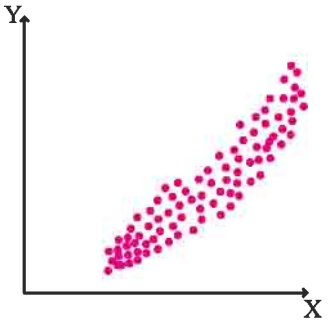
1.



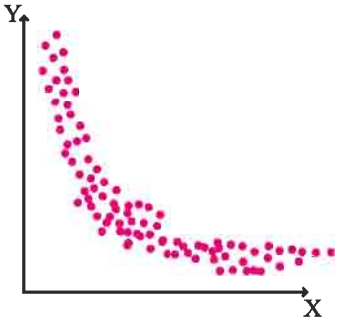
2.



3.



4.



2 On a relevé dans le tableau ci-dessous les poids (en kg) respectif de 12 pères et de leurs fils aînés.

X Poids du père	Y Poids du fils
65	63
63	61
67	66
64	62
68	67
62	60
70	69
66	65
68	67
67	67
69	66
71	70

1. Tracer le nuage de la série (X, Y).
2. Déterminer un ajustement affine par la méthode de Mayer.
3. Quel poids devrait avoir le fils aîné d'un homme qui pèse 77 kg ?

3 Le tableau ci-dessous indique l'évolution du nombre de médecins en Tunisie de l'année 1990 à l'année 2003.

Année	Rang de l'année (X)	Nombre de médecins (Y)
1990	1	4425
1991	2	4500
1992	3	5099
1993	4	5257
1994	5	5344
1995	6	5965
1996	7	6177
1997	8	6464
1998	9	6819
1999	10	7149
2000	11	7444
2001	12	7767
2002	13	7964
2003	14	8189

1. Tracer le nuage de la série (X,Y).

2. Déterminer le point moyen $G(\overline{X},\overline{Y})$

3. Déterminer un ajustement affine par la méthode de Mayer.

4. Donner une estimation du nombre de médecin en Tunisie dans l'année 2010 ?

4 Le tableau suivant donne la répartition d'une population de 100 ménages selon les deux caractères X le nombre de pièces habitées et Y le nombre d'enfants.

X \ Y	0	1	2	3	4	Total
1	6	2	1	0	0	9
2	5	12	8	1	1	27
3	2	7	15	11	3	38
4	0	1	8	14	3	26
Total	13	22	32	26	7	100

1. a. Calculer la moyenne $\overline{X}$ et l'écart type σ_X de la variable X.

b. Calculer la moyenne $\overline{Y}$ et l'écart type σ_Y de la variable Y.

2. a. Calculer le coefficient de corrélation entre X et Y.

b. Un ajustement affine de la série (X,Y) est-il justifié ?

5 Le tableau ci-dessous donne la charge maximale Y, en tonnes, qu'une grue peut lever pour une longueur X, en mètres, de la flèche.

X	Y
9	1.4
10	1.25
12	1
14	0.84
16	0.7
18	0.62
20	0.55
22	0.5

1. Les réponses numériques à cette question seront données à 10^{-2} près.

a. Représenter le nuage de points dans un repère orthogonal.

b. Déterminer le coefficient de corrélation linéaire entre X et Y.

c. Déterminer une équation de la droite de régression de Y en X.

Construire cette droite sur le graphique précédent.

d. Utiliser cette équation pour déterminer la charge maximale que peut lever la grue avec une flèche de 23 mètres.

6 Le tableau suivant recense par clinique le nombre de postes du personnel non médical en fonction du nombre de lits de la clinique.

Clinique	Nombre de lits (X)	Nombre de postes (Y)
C_1	122	185
C_2	177	221
C_3	77	114
C_4	135	164
C_5	109	125
C_6	88	118
C_7	185	193
C_8	128	160
C_9	120	151
C_{10}	146	172
C_{11}	100	150

1. Représenter le nuage de points associé à la série statistique (X,Y) dans le plan rapporté à un repère orthogonal.

2. Déterminer le coefficient de corrélation linéaire entre X et Y.

3. Donner une équation de la droite de régression de Y en X .

Pour les coefficients, on prendra les valeurs décimales arrondies à 10^{-1} près.

Tracer cette droite dans le repère précédent.

4. Une clinique possède 35 lits.

En utilisant les résultats obtenus dans la question 3, combien devrait-elle embaucher de personnel occupant un poste non médical ?

7 A/ Un club sportif a été crée en 1999, à l'origine le nombre d'adhérents était égal à 600. On donne dans le tableau suivant le nombre d'adhérents de 1999 à 2005.

Année	Rang (X)	Nombre d'adhérents (Y)
1999	0	600
2000	1	690
2001	2	794
2002	3	913
2003	4	1045
2004	5	1207
2005	6	1380

On pose $Z = \ln Y$.

1. a. Vérifier qu'on peut réaliser un ajustement affine par la méthode des moindres carrés de la série (X, Z).

b. Déterminer une prévision du nombre d'adhérents en 2006.

2. Justifier que $Y \simeq 602 \times (1.15)^X$.

B/ En fait le club a compté 2400 adhérents lors de l'année 2006.

Soit f la fonction définie sur $\mathbb{R}_+$ par

$$f(x) = \frac{3600}{1 + 0.5e^{-x}}$$

On suppose que le nombre d'adhérents en $(2006+n)$ est égal à $f(n)$ où n est un entier naturel.

1. Déterminer la limite de $f(n)$ lorsque n tend vers $+\infty$ et interpréter le résultat.

2. a. Reproduire et compléter le tableau suivant :

Année	2007	2008	2009	2010	2011
n	1	2	3	4	5
$f(n)$	3040				

b. Calculer la moyenne M du nombre prévisionnel d'adhérents entre 2007 et 2011.

3. Calculer la valeur moyenne $\bar{f}$ de f sur l'intervalle $[0.5, 5.5]$.

8 On a relevé la taille et le poids de 16 jeunes filles. Les résultats obtenus sont résumés dans le tableau suivant.

Tailles X (en cm)	Poids Y (en kg)
160	46
165	48
167	48
160	46
168	49
170	51
160	45
162	45
165	48
170	49
170	51
168	50
172	50
165	48
165	47
170	50

1. a. Construire, dans un repère orthogonal, le nuage de points de la série (X, Y).

b. Un ajustement affine est-il justifié ?

2. a. Déterminer le coefficient de corrélation entre X et Y.

b. Ecrire une droite de régression de Y en X.

c. Donner une estimation de la masse d'une jeune fille mesurant 180 cm.

3. Un journal de santé publie la loi de Lorentz qui donne une relation entre le poids M et la taille T pour une jeune fille, $M = (T - 100) - \frac{T - 130}{2}$.

Utiliser cette relation pour estimer la masse d'une jeune fille mesurant 180 cm.

9 Onze élèves de $7^{ème}$ année de base travaillent sur la proportionnalité.
Ils mesurent le rayon d'un disque puis évaluent l'aire de ce disque.
Les rayons, exprimés en cm, forme la série (X).

Les aires correspondantes, exprimées en cm^2, forment la série (Y).

Les résultat de cette expérience sont donnés dans le tableau suivant.

X (en cm)	Y (en cm^2)
2	12
2.5	20
3	28
3.5	38
4	50
4.5	64
5	78
5.5	95
6	113
6.5	133
7	154

1. Les deux séries sont-elles proportionnelles ?
2. On pose $Z=\sqrt{Y}$.
a. Construire, dans un repère orthogonal, le nuage de points de la série (X,Z). (Les valeurs de Z seront arrondis à 10^{-1} près).
b. Calculer le coefficient de corrélation ρ_{XZ}.
Interpréter le résultat.
c. Déterminer une équation de la droite de régression de Z en X.
(Les coefficients seront arrondis à 10^{-1} prés).
d. En déduire une valeur approchée de π.

10 Une entreprise envisage la fabrication d'un nouveau produit.
Une étude a permis d'établir le tableau suivant où, pour différentes observations, X désigne la quantité de produit que la clientèle est disposée à acheter, et Y le prix de vente (en DT) d'une unité.

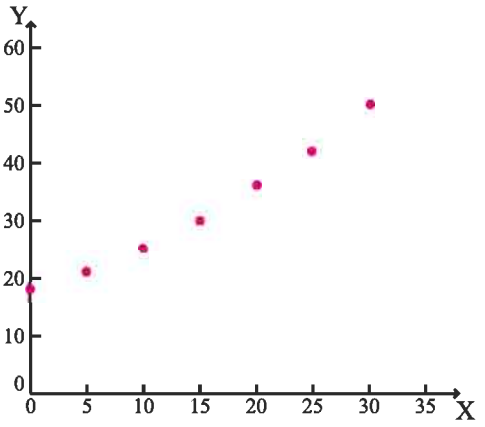
X	350	400	450	500	550	600
Y	140	120	100	95	85	70

1. Calculer le coefficient de corrélation ρ_{XY}.
2. Déterminer une équation de la droite de régression de Y en X. (les coefficients seront arrondis à 10^{-1} prés).
3. Soit $r(x)$ la recette correspondant à la vente de x articles au prix unitaire y.
a. Montrer que $r(x)=(226.5-0.3x)x$.
b. Etudier le variations de la fonction f définie sur $\mathbb{R}_+$ par $f(x)=-0.3x^2+226.5x$.
c. En déduire le prix de vente pour lequel la recette est maximale. Calculer cette recette maximale.

11 Le tableau suivant donne la population d'une ville entre les années 1975 et 2005.

Année	Rang de l'année (X)	Population (en milliers d'habitants) (Y)
1975	0	18
1980	5	21
1985	10	25
1990	15	30
1995	20	36
2000	25	42
2005	30	50

Le nuage de points associé à ce tableau est représenté graphiquement ci-après.



A/ 1. Calculer le coefficient de corrélation entre X et Y.
2. a. Déterminer une équation de la droite de régression de Y en X.
b. Tracer cette droite sur le graphique ci-dessus.
c. En déduire une estimation de la population en 2008 à un millier prés.

B/ 1. L'allure du nuage suggère à chercher un ajustement par une fonction f 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 $f(30)=50$.

On donnera une valeur arrondie de b au millième.
2. Déduire de cet ajustement une estimation de la population en 2008 à un millier prés.
3. Tracer la courbe de f sur le même graphique.
4. La population en 2008 était de 55 milliers.
Lequel des deux ajustement vous semble le plus pertinent ? Justifier votre choix.

C/ On considère maintenant que, pour une année, la population est donnée en fonction du rang x par $f(x)=18e^{0.034x}$.

1. Calculer la valeur moyenne $\bar{f}$ de la fonction f sur $[0,30]$. On donnera le résultat arrondi au dixième.
2. A l'aide d'une lecture graphique, déterminer l'année au cours de laquelle la population atteint cette valeur moyenne.

12 Le tableau ci-dessous donne l'évolution de la population d'un pays de 1965 à 2000.
T désigne le rang de l'année et P la population en millions d'habitants.

Année	Rang de l'année (T)	P
1965	0	8
1970	5	8.9
1975	10	9.9
1980	15	11
1985	20	12
1990	25	13.5
1995	30	15
2000	35	16.6

A/ 1. Représenter le nuage de points associé à la série statistique (T,P) dans un repère orthogonal.

• Sur l'axe des abscisses, choisir 2 cm pour 5 unités (5 ans).
• Sur l'axe des ordonnées, placer 8 à l'origine, puis choisir 2 cm pour une unité (1 million d'habitants).
2. Les experts cherchent à modéliser cette évolution par une fonction f dont la courbe est voisine du nuage de points.
On pose $Y=\ln P$.

a. Donner une valeur approchée à 10^{-3} près par défaut du coefficient de corrélation linéaire de la série (T,Y).

b. Déterminer une équation de la droite de régression de Y en T. (Les coefficients seront arrondis à 10^{-3} près).
c. En déduire l'expression de la population P en fonction du rang T de l'année.

B/ On admet que la fonction f définie sur $[0,35]$ par $f(t)=8.e^{0.02t}$ est une modélisation satisfaisante de l'évolution de la population (en millions d'habitants) de 1965 à 2000.
1. Étudier le sens de variation de f sur $[0,35]$ et dresser le tableau de variation de f sur cet intervalle.
2. Construire la courbe représentative de f, notée (C), dans le repère de la partie A.

3. On pose $I=\int_0^{35} f(t)dt$.

a. Donner une valeur approchée de I arrondie à 10^{-2} près.
b. En déduire la population moyenne m du pays durant ces 35 années et la représenter sur le graphique.

4. Calculer le rapport $\frac{f(t+1)-f(t)}{f(t)}$ et en donner une interprétation en terme de pourcentage.
5. Si le modèle exponentiel étudié dans la partie B restait valable après 2000, en quelle année la population aurait-elle dépassé les 19 millions d'habitants ?

15 On étudie la croissance d'une plante à partir d'un instant considéré comme initial.
Le tableau ci-dessous indique le diamètre D de la tige après T semaines.

Temps T en semaines	Diamètre D en centimètres
0	0.4
2	1.2
6	5.4
8	5.8
10	6.4
12	6.9

1. Représenter le nuage de points associé à cette série statistique.

2. On pose $U = \ln\left(\frac{8}{D} - 1\right)$.

a. Calculer le coefficient de corrélation linéaire de la série (T, U).

b. Déterminer par les moindres carrés une équation de la droite d'ajustement de U en T.

c. Vérifier que pour cette plante, le diamètre de sa tige principale est donné par une relation de la forme $D(t) = \frac{8}{1 + ce^{-at}}$ où a et c sont deux réels que l'on précisera.

3. a. Pour les valeurs de a et c trouvées, tracer dans le repère précédent la fonction $f : t \mapsto D(t)$ pour $t \geq 0$

b. Le diamètre de la plante dépassera-t-il 8 cm ?