

**I. Nuage de points****1. Introduction**

Une série statistique à deux variables, X et Y, est le résultat de l'observation des deux caractères X et Y pour chaque individu d'une population.

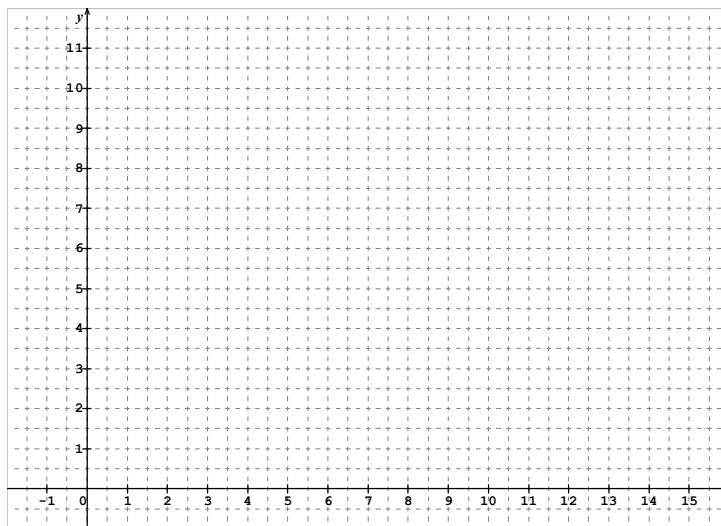
Lorsque les caractères sont quantitatifs, on peut associer, à chaque individu  $i$ , un couple de nombres réels noté  $(x_i, y_i)$ .

**Exemple :**

Le tableau suivant donne, en millions de dinars, le chiffre d'affaires  $x_i$  et la somme consacrée aux dépenses de publicité  $y_i$  pour cinq entreprises :

Entreprises	$x_i$	$y_i$
A	30	3
B	55	4,5
C	60	7
E	20	1,5
F	50	4

Placer les points  $M_i(x_i, y_i)$  dans le repère orthogonal ci – dessous :

**2. Définition**

Dans un repère orthogonal du plan, le nuage de points associés à la série statistique à deux variables, X et Y, est l'ensemble des points  $M_i$  de coordonnées  $(x_i, y_i)$  représentatifs de tous les individus  $i$  de la population.

**3. Point moyen**

On note X le caractère : « le chiffre d'affaires de chaque entreprise » et Y : « les dépenses de publicité ».  
Calculer la moyenne, la variance et l'écart – type de chaque caractère.

On rappelle que :

$$\bar{X} = \frac{1}{N} \sum_{i=1}^N n_i x_i, \quad V(X) = \frac{1}{N} \sum_{i=1}^N n_i (x_i - \bar{X})^2 = \frac{1}{N} \sum_{i=1}^N n_i x_i^2 - \bar{X}^2 \quad \text{et} \quad \sigma(X) = \sqrt{V(X)}.$$

- $x_1, x_2, \dots, x_N$  désignent les valeurs distinctes prises par la variable X si elle est discrète, ou les centres des classes si la variable X est continue
- N est la taille de l'échantillon
- Dans notre exemple  $n_i = 1$ , pour tout  $1 \leq i \leq N$

Entreprises	$x_i$	$y_i$	$x_i^2$	$y_i^2$	$x_i y_i$
A	30	3			
B	55	4,5			
C	60	7			
E	20	1,5			
F	50	4			
Somme					

$\bar{X} = \dots\dots\dots$  ;  $V(X) = \dots\dots\dots$  ;  $\sigma(X) = \dots\dots\dots$

$\bar{Y} = \dots\dots\dots$  ;  $V(Y) = \dots\dots\dots$  ;  $\sigma(Y) = \dots\dots\dots$

**Définition**

Le point  $G(\bar{X}, \bar{Y})$  est appelé point moyen du nuage de points associé à la série statistique à deux variables X et Y.

Exemple : Placer le point moyen G dans le repère précédent.

**II. Liaison entre deux caractères – Méthode d'ajustement par les moindres carrés**

**1. Droite d'ajustement**

Lorsque le nuage a tendance de s'accumuler autour d'une droite, alors on cherche une équation de la droite D qui approche le « mieux possible » les points du nuage, c'est ce qu'on appelle un ajustement linéaire. Cet ajustement est liée à deux paramètres appelés Covariance et Coefficient de corrélation linéaire.

**2. Covariance**

Définition

Soit (X,Y) une série statistique double à caractères quantitatifs donnés par des observations individuelles  $(x_i, y_i)$  où  $1 \leq i \leq n$ , n étant l'effectif de la population observée.

On appelle covariance de (X,Y) le réel noté  $cov(X, Y)$  défini par :  $cov(X, Y) = \frac{1}{N} \sum_i (x_i - \bar{x})(y_i - \bar{y})$  ou

encore  $cov(X, Y) = \frac{1}{N} \sum_i x_i y_i - \bar{x} \bar{y}$ .

Remarque

Comme pour le calcul de la variance, la formule  $cov(X, Y) = \frac{1}{N} \sum_i x_i y_i - \bar{x} \bar{y}$  est souvent la plus simple à utiliser pour les calculs.

Exemple

La covariance de la série statistique ci – dessus est :

Propriétés

- $cov(X, Y) = cov(Y, X)$ .
- $cov(X, X) = V(X)$ .
- $cov^2(X, Y) \leq V(X) \cdot V(Y) \Rightarrow |cov(X, Y)| \leq \sigma(X) \cdot \sigma(Y)$ .

**Avec la calculatrice :**

Pour programmer le tableau statistique de la page 1 :

. Choisir le mode de fonctionnement de statistique à deux variables

**MODE** **3** **Stat** **1**

. Entrer les couples  **$x_i$**  **STO**  **$y_j$**  **M+**

. Pour obtenir par exemple la moyenne  $\bar{X}$  appuyez sur **RCL**  **$\bar{X}$**

**3. Coefficient de corrélation linéaire**

Définition

Le coefficient de corrélation linéaire entre deux variables, X et Y, est le nombre r

défini par :  $r = \frac{cov(X, Y)}{\sigma(X) \cdot \sigma(Y)}$ .

Propriétés

- $|cov(X, Y)| \leq \sigma(X) \cdot \sigma(Y) \Rightarrow 0 \leq \frac{|cov(X, Y)|}{\sigma(X) \cdot \sigma(Y)} \leq 1 \Rightarrow \boxed{-1 \leq r \leq 1}$
- Les points du nuage sont alignés si et seulement si  $r = 1$  ou  $r = -1$ .
- Si  $|r| \geq \frac{\sqrt{3}}{2}$  alors la corrélation linéaire entre X et Y est forte. On peut trouver une droite approximative liant X et Y.
- Si  $|r| < \frac{\sqrt{3}}{2}$  alors la corrélation linéaire entre X et Y est faible. Il est inutile de chercher à exprimer Y comme fonction affine de X mais il peut exister d'autres types de relations.

Exemple

Le coefficient de corrélation linéaire est r = .....

On en déduit que : .....

.....

**4. Droites de régression**

Soient X et Y deux séries statistiques quantitatives non constantes et observées dans une population donnée.

On suppose que le coefficient de corrélation r vérifie :  $|r| \geq \frac{\sqrt{3}}{2}$  alors il est possible d'approcher la liaison entre X et Y par une relation affine de type Y en fonction de X ou aussi X en fonction de Y.

- La première droite est appelée droite de régression de Y en X, elle a pour équation :

$D : y = ax + b$  où  $a = \frac{cov(X, Y)}{V(X)}$  et  $b = \bar{Y} - a\bar{X}$ .

- La deuxième droite est appelée droite de régression de X en Y, elle a pour équation :  $D' : x = a'y + b'$

où  $a' = \frac{cov(X, Y)}{V(Y)}$  et  $b' = \bar{X} - a'\bar{Y}$ .

- Ces deux droites affines passent par le point  $G(\bar{x}, \bar{y})$ .

**Exemple**

1) Donner une équation de chacune des droites D et D'.

$a = \dots\dots\dots$  ;  $b = \dots\dots\dots$

D :  $\dots\dots\dots$

$a' = \dots\dots\dots$  ;  $b' = \dots\dots\dots$

D' :  $\dots\dots\dots$

2) Tracer D et D' dans le même repère R.

3) Quelle estimation peut on faire quant à la somme consacrée aux dépenses de publicité pour une entreprise ayant un chiffre d'affaire égal à 200 millions de dinars

.....  
 .....

**Exercice**

Le tableau ci-dessous donne l'évolution de la dette des pays du tiers-monde entre 1978 et 1992 ( en milliards de dollars ).

Année	1978	1982	1986	1990	1992
Rang de l'année $x_i$	0	4	8	12	14
Dette $y_i$	383	753	1089	1346	1510

Source : Banque mondiale, FMI, 1993.

- Le plan est rapporté à un repère orthogonal. Représenter le nuage de points  $(x_i, y_i)$ , et le point moyen G de cette série.
- a) Calculer le coefficient de corrélation linéaire de cette série double.  
 Un ajustement affine peut il être envisagé ? Pourquoi ?  
 b) Ecrire une équation de la droite de régression D de Y en X. Tracer D.  
 c) Estimer, à 1 milliard de dollars près, le montant prévisible de la dette des pays du tiers – monde en 2010.

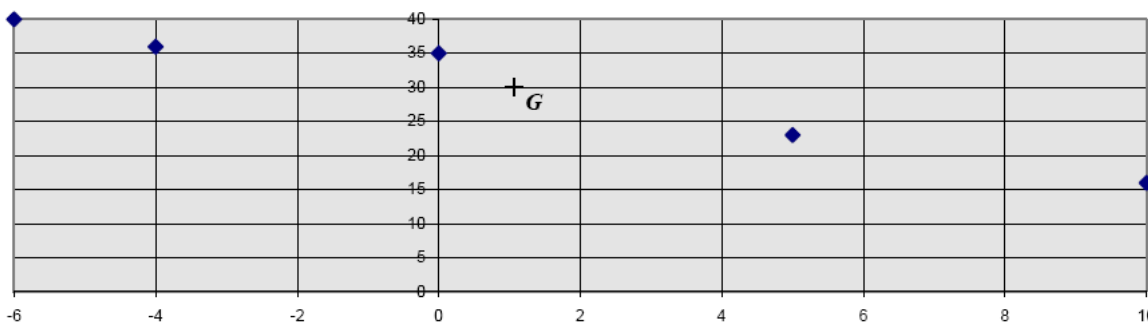
**Remarque :** il existe une autre méthode d'ajustement dite méthode de Mayer  
 ( Voir paragraphe du livre page (106 Sc) ou page (217 Maths) )

**Exercice :**

Le tableau ci-dessous donne la consommation quotidienne Y en fuel d'une chaudière ( en litres) en fonction des relevés de température extérieure X

X (en degré C)	$x_1 = -6$	$x_2 = -4$	$x_3 = 0$	$x_4 = 5$	$x_5 = 10$
Y (en litres)	$y_1 = 40$	$y_2 = 36$	$y_3 = 35$	$y_4 = 23$	$y_5 = 16$

On a représenté le nuage de points



1) Déterminer un ajustement affine de Y en X par la méthode de Mayer

.....  
 .....  
 .....  
 .....

.....  
.....  
2) Retrouver cet ajustement par la méthode des moindres carrés  
.....  
.....  
.....

3) A quelle température, la consommation en fuel dépassera elle 100 litres par jour  
.....  
.....  
.....

### III. Distributions marginales

#### 1. Construction d'un tableau à double entrée :

##### Activité

Dans une population de 50 ménages on a observé les deux caractères quantitatifs discrets suivants :

X nombre d'enfants dans chaque ménage

Y nombre de pièces du logement habité par chaque ménage

Les résultats de ces observations sont consignés dans les tableaux suivants :

Numéro du ménage	Nombre D'enfants	Nombre De pièces
1	0	1
2	3	3
3	0	2
4	3	2
5	1	1
6	3	3
7	2	2
8	1	2
9	4	2
10	3	2
11	2	1
12	1	4
13	1	2
14	1	3
15	0	1
16	3	3
17	2	2
18	2	3
19	0	2
20	1	2
21	2	3
22	2	2
23	4	3
24	1	2
25	2	3

Numéro du ménage	Nombre D'enfants	Nombre De pièces
26	4	4
27	2	2
28	2	4
29	2	3
30	3	3
31	1	1
32	1	3
33	4	3
34	2	3
35	2	2
36	3	4
37	0	1
38	4	4
39	0	3
40	4	4
41	2	3
42	3	3
43	5	4
44	1	2
45	3	4
46	5	3
47	2	3
48	2	4
49	3	2
50	2	3

1) Quelles sont les valeurs prises par X et par Y  
.....  
.....

2) Les données présentées dans le tableau précédent étant nombreuses , on se propose, dans la suite , de les présenter sous une forme plus réduite et à l'aide d'un tableau à double entrée . Compléter le tableau suivant :

**Tableau à double entrée ( tableau des effectifs ) :**

X \ Y	1	2	3	4	Total
0	3	2	1	0	6
1	2				
2	1				
3	0				
4	0				
5	0				
Total	6				50

Pour  $(1 \leq i \leq 6)$  et  $(1 \leq j \leq 4)$ , on a associé au couple  $(x_i, y_j)$  un nombre appelé effectif que l'on note  $n_{ij}$ .

Exemple : l'effectif correspondant au couple  $(x_4, y_2) = (3, 2)$  est  $n_{4,2} = 3$ .

Il indique le nombre de ménages ayant 3 enfants et un logement de 2 pièces.

Définition :

- $n_{ij}$  est appelé effectif associé au couple  $(x_i, y_j)$ .

**2. Distributions marginales :**

Du tableau précédent, on peut extraire deux séries statistiques à une variable. La première donne la répartition de 50 ménages selon le nombre d'enfants X dans chaque ménage et la seconde donne la répartition de 50 ménages selon le nombre de pièces Y du logement habité par chaque ménage.

On obtient alors les deux tableaux suivants :

Tableau 1 :

N <sup>bre</sup> d'enfants X	0	1	2	3	4	5	Total
Effectif $n_i$ des ménages	6	10	16	10	6	2	50

Tableau 2 :

N <sup>bre</sup> de pièces Y	1	2	3	4	Total
Effectif $n_j$ des ménages	6	16	19	9	50

Chacun des tableaux 1 et 2 définit une série statistique à une variable, appelée distribution marginale.

- Calculer  $\bar{X}$  et  $\bar{Y}$ , donner le point moyen G :

.....

.....

.....

.....

.....

- Calculer  $V(X)$ ,  $V(Y)$ ,  $\sigma(X)$  et  $\sigma(Y)$  :

.....

.....

.....

.....

.....

- La covariance de cette série est :

$$\text{cov}(X, Y) = \frac{1}{N} \sum_{i=1}^p \sum_{j=1}^q n_{ij} (x_i - \bar{X})(y_j - \bar{Y}) = \frac{1}{N} \sum_{i=1}^p \sum_{j=1}^q n_{ij} x_i y_j - \bar{X} \bar{Y}$$

$$\Rightarrow \text{cov}(X, Y) = \frac{1}{50} \sum_{i=1}^6 \sum_{j=1}^4 n_{ij} x_i y_j - \bar{X} \bar{Y}$$

Cette formule parait compliquée, pour cela on considère le tableau suivant qui nous facilite la tâche

X \ Y	Y				Totaux
	1	2	3	4	
0	3	2	1	0	
1	2	5	2	1	
2	1	5	8	2	
3	0	3	5	2	
4	0	1	2	3	
5	0	0	1	1	
Totaux					

.....

.....

.....

- Calculer le coefficient de corrélation linéaire ; que peut – on remarquer ?

**Avec la calculatrice :**

Pour programmer le tableau statistique à double entrée :

- Choisir le mode de fonctionnement de statistique à deux variables **MODE** **3** **Stat** **1**
- Entrer les triplets  **$x_i$**  **STO**  **$y_j$**  **STO**  **$n_{ij}$**  **M+**
- Pour obtenir par exemple la moyenne  $\bar{X}$  appuyez sur **RCL**  **$\bar{X}$**

**Exercice :**

Le tableau suivant donne la répartition de 800 agriculteurs suivant les deux variables statistiques suivantes :

X : superficie de l'exploitation agricole (en ha)

Y : âge de l'agriculteur

X \ Y	[20,30[	[30,40[	[40,50[	[50,60[	60 ans et plus
]0,5[	15	50	90	80	25
[5,15[	10	45	60	50	20
[15,25[	8	30	40	35	12
[25,35[	6	40	35	25	9
[35,45[	7	20	15	20	8
45 et plus	4	15	10	10	6

- 1) Déterminer les distributions marginales de X et Y
- 2) Calculer  $\bar{X}$  ,  $\bar{Y}$  ,  $V(X)$  ,  $V(Y)$  ,  $\sigma_X$  et  $\sigma_Y$
- 3) Calculer  $cov(X,Y)$  et le coefficient de corrélation du couple (X,Y). Peut on envisager un ajustement affine liant X et Y ?

**Cas d'un ajustement non affine :**

**Activité**

On a mesuré, entre 1989 et 1974, l'effet de la population sur une 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	1989	1990	1991	1992	1993	1994
Rang de l'année $x_i$	1	2	3	4	5	6
$y_i$	591,3	106,7	96,5	63,2	21	9,4

- 1) On considère la série statistique double  $(x_i, y_i)$ . Calculer le coefficient de corrélation entre x et y. Expliquer pourquoi un ajustement linéaire ne paraît pas bien adapté
- 2) On pose :  $z_i = \text{Log} y_i$ , pour  $i \in \{1, 2, 3, 4, 5, 6\}$ 
  - a) Calculer les nombres  $Z_i$
  - b) Représenter dans un repère orthogonal le nuage de points  $(x_i, z_i)$
- 3) Calculer le coefficient de corrélation de cette série. Justifier l'utilisation d'un ajustement affine pour la série  $(x_i, z_i)$
- 4) Déterminer l'équation de la droite de régression de z en x. Tracer cette droite dans le repère de la question 2) b)
- 5) On suppose que l'évolution de cette population se poursuit sur le même modèle
  - a) A partir de quelle année, cette population sera-t-elle strictement inférieure à 1000 ?
  - b) Donner une estimation de la population de cette rivière en l'an 2010.



## Statistiques à deux variables

### Exercices

#### Exercice n°1 :

Le tableau suivant donne la répartition de cent anciens élèves d'un établissement d'après les deux caractères : X le nombre d'années passées dans cet établissement et Y son âge

	X	2	3	4
Y				
20		0	8	30
25		5	20	7
30		25	3	2

- 1) a) Donner les distributions marginales de X et Y
- b) Calculer :  $\bar{X}$ ,  $\bar{Y}$ ,  $\sigma(X)$  et  $\sigma(Y)$

- 2) Calculer la covariance et le coefficient de corrélation linéaire entre X et Y, et interpréter le résultat obtenu
- 3) a) Déterminer la droite de régression de Y en X
- b) Donner une estimation de l'âge d'un élève qui a passé 5 ans dans cet établissement

#### Exercice n°2 :

Les mesures relatives à la vitesse d'un automobiliste et la distance nécessaire pour arrêter le véhicule sont réunies dans le tableau suivant .

$x_i$ ( Km / h )	27	43	62	80	98	115
$y_i$ ( distance d'arrêt en m )	6,8	20,5	35,9	67,8	101,2	135,8

- 1) Dans un repère orthogonal , construire le nuage des points associés à la série statistique  $(x_i , y_i)$ .
- 2) a) Calculez  $\bar{x}$  ,  $\bar{y}$  ,  $V(X)$  ,  $V(Y)$  et  $cov(X, Y)$  .
- b) Calculez le coefficient de corrélation linéaire .
- c) Déterminez et tracez la droite de régression de  $y$  en  $x$  .
- d) Déduire la valeur estimée de  $x$  correspondante à une distance d'arrêt de 180 m .
- e) Quelle est la distance d'arrêt en mètre , correspondante à une vitesse de 150 Km / h ?

#### Exercice n°3 :

L'étude de la population d'une ville a donné les résultats suivants : h est le nombre d'habitants par milliers.

Année	2001	2002	2003	2004	2005
Rang de l'année $x_i$	1	2	3	4	5
Nombre d'habitants $h_i$	24,5	27,38	31,205	36,125	37,845

1. Représenter dans un repère orthonormé le nuage de points de la série  $(x , h)$  dans un repère orthogonal  $R(O, \vec{i}, \vec{j})$  . Sur l'axe des abscisses placer 0 à l'origine 1 cm représente 1  
Sur l'axe des ordonnées placer 20 à l'origine 1 cm représente mille
2. a) Calculer le coefficient de corrélation linéaire entre x et h
- b) Justifier l'utilisation d'un ajustement affine pour la série  $(x , h)$
- c) Déterminer l'équation de la droite de régression de h en x
3. On suppose que l'évolution de cette population se poursuit sur le même modèle et que les dépenses de la municipalité de cette ville en une année en milliers des dinars noté y obéit à la loi :  $y = 10\sqrt{h_i} + 200$
- a) Donner une estimation de la population de cette ville à l'année 2010 ?
- b) Donner une estimation des dépenses de cette municipalité à l'année 2010

c) A partir de quelle année les dépenses de cette municipalité dépasseront-elles les 300 milles dinars.

#### Exercice n°4 :

Pour une série  $(x_i, y_i)$  avec  $(1 \leq i \leq 15)$  à deux variables, la méthode des moindres carrés a permis de trouver : L'équation de la droite (D) de régression de y par rapport à x.  $D : y = 0,37x + 12,59$

L'équation de la droite (D') de régression de x par rapport à y.  $D' : x = 0,9y + 14,07$

1. Déterminer le coefficient de corrélation linéaire de cette série
2. Déterminer les coordonnées du point moyen G associé à cette série
3. On suppose que  $\sum_{i=1}^{15} x_i^2 = 25205$ . Calculer :  $\text{Cov}(x, y)$ ;  $\sum_{i=1}^{15} x_i y_i$  et  $\sum_{i=1}^{15} y_i^2$

#### Exercice n°5 :

Le tableau suivant donne les résultats obtenus à partir de six essais de laboratoire concernant la charge de rupture d'un acier en fonction de sa teneur en carbone

Teneur en carbone $X$ ( pour 10000)	60	62	64	68	70	74
Charge de rupture $Y$ ( en Kg )	70	75	80	82	85	100

1. a) Représenter dans un repère orthogonal le nuage des points de la série  $(X, Y)$   
b) Déterminer le point moyen de ce nuage
2. a) Calculer  $V(X)$ ;  $V(Y)$  et  $COV(X, Y)$   
b) Est-il possible d'envisager un ajustement linéaire entre  $X$  et  $Y$ ? Justifier  
c) Déterminer et construire la droite de régression de  $Y$  en  $X$
3. Quelle pourrait être la charge de rupture d'un acier ayant une teneur en carbone de 90 pour 10000