Terminale D

Mathématiques

CÔTE D'IVOIRE – ÉCOLE NUMÉRIQUE

THÈME : ORGANISATION ET TRAITEMENT DES DONNÉES

Durée : 10 **heures** **Code :**

Leçon 11 : STATISTIQUE À DEUX VARIABLES

A. SITUATION D'APPRENTISSAGE

Dans le cadre des recherches pour un exposé, des élèves d'une classe de Terminale ont été accrochés par les informations suivantes :
La prévision météorologique est une science en pleine évolution. Elle a pour objectif de prédire un ensemble de paramètres comme la pluviométrie, la pression atmosphérique, la température, etc.
Le tableau suivant donne les pluviométries et températures moyennes de septembre 2018 à août 2019 dans une ville.

	Sept 2018	Oct. 2018	Nov. 2018	Déc 2018	Jan 2019	Fév. 2019	Mar 2019	Avr. 2019	Mai 2019	Juin 2019	Juillet 2019	Août 2019
Pluviométrie (en mm)	13	23	49	49	50	64	79	48	40	10	5	6
Température (en °C)	23	17	14	10	10	11	13	15	17	23	27	28

Dans l'affiche la température moyenne d'octobre 2019 était de 32 °C. Les élèves veulent connaitre la pluviométrie du mois d'octobre 2019. Un des élèves affirme que la pluviométrie n'est pas liée à la température et qu'on ne peut savoir la pluviométrie d'octobre. Ce que réfutent certains. Toute la classe ayant été saisi, décide de chercher à savoir si la pluviométrie est liée à la température et si c'est le cas, de prévoir la pluviométrie d'octobre 2019.

B. CONTENU DE LA LEÇON

I. Série statistique double

1. Définition

On considère deux caractères quantitatifs X et Y sur une même population de n individus.
On note : x_1 , x_2 , x_3 ,…, x_p les valeurs(ou les modalités)du caractère X ; y_1 , y_2 , y_3 ,…, y_p les valeurs du caractère Y et n_{ij} l'effectif du couple (x_i , y_j).
On appelle série statistique double de caractère (X, Y), l'ensemble des triplets (x_i , y_j , n_{ij}).

Exemple

Une étude statistique porte sur une population de 100 ménages. Deux caractères X et Y sont étudiés :
- le caractère X est le nombre d'enfants.
- le caractère Y est le nombre de pièces de l'appartement occupé.

On obtient le tableau ci-dessous qui représente la série statistique de caractère(X ,Y)

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

2	3	11	10	5	1	0
3	1	3	16	13	4	1
4	0	1	3	5	8	4

Sur la 1ère ligne, sont inscrites les valeurs du caractère X, soit $x_1 = 0$; $x_2 = 1$; $x_3 = 2$ $x_4 = 3$; $x_5 = 4$; $x_6 = 5$.
La 1ère colonne affiche les valeurs du caractère Y qui sont y_1 =1 ;y_2 =2 ; y_3 =3;y_4 =4.
Les nombres qui ne sont pas dans cette ligne et cette colonne, représentent les différents n_{ij}.
Ainsi considérons le nombre 4 dans ce tableau. On constate qu'il est dans la colonne de la valeur 1 du caractère X et dans la ligne de la valeur 1du caractère Y. On dit alors il y a 4 ménages qui ont un enfant et occupent un appartement d'une pièce.
Ainsi le couple (x_2, y_1)= (1 ; 1) a pour effectif $n_{21} = 4$.
Combien de ménages ont deux enfants et occupent un appartement de quatre pièces ?
On va donc considérer la colonne ayant la valeur 2 du caractère X et la ligne ayant la valeur 4 du caractère Y. L'intersection de cette ligne et de cette colonne est 3.
3 ménages ont donc deux enfants et occupent un appartement quatre pièces.
Ce tableau à double entrée ci-dessus est appelé **tableau de contingence.**

2. Taleau de séries marginales

Reprenons l'exemple précédent.
Il est question de trouver l'effectif de chaque modalité du caractère X et l'effectif de chaque valeur du caractère Y

X / Y	0	1	2	3	4	5	Total
1	6	4	1	0	0	0	11
2	3	11	10	5	1	0	30
3	1	3	16	13	4	1	38
4	0	1	3	5	8	4	21
Total	10	19	30	23	13	5	100

Considérons le caractère X.
Pour trouver l'effectif de la valeur 0, on additionne tous les n_{ij}qui se trouvent dans la colonne de la valeur 0 c'est-à-dire 6+3+1+0=10. 10 ménages n'ont donc pas d'enfants.

Combien de ménages ont-ils quatre enfants ? Il est question donc d'additionner tous les n_{ij} se trouvant dans la colonne de la valeur 4 du caractère X.
On a donc 0+1+4+8=13 ménages.
On procède de la même manière pour trouver l'effectif des autres modalités du caractère X. Ainsi à chaque valeur on a son effectif dans la dernière ligne.

D'où le tableau linéaire associé à X :

x_i	0	1	2	3	4	5
n_i	10	19	30	23	13	5

La série ainsi obtenue est appelée série marginale du caractère X.

En faisant de même avec les lignes, on obtient l'effectif de la modalité 1 du caractère Y en additionnant les n_{ij} de la ligne où se trouve cette modalité. Soit 6+4+1+0+0+0=11.
On obtient ainsi l'effectif de chaque modalité du caractère Y dans la dernière colonne du tableau.

D'où le tableau linéaire associé à Y :

y_i	1	2	3	4
n_i	11	30	38	21

La série ainsi obtenue est appelée série marginale du caractère Y.

Dressons le tableau des fréquences marginales du caractère X.
On rappelle que la fréquence est l'effectif de la modalité sur l'effectif total.
On obtient le tableau suivant :

x_i	0	1	2	3	4	5
f_i	$\frac{10}{100}$	$\frac{19}{100}$	$\frac{30}{100}$	$\frac{23}{100}$	$\frac{13}{100}$	$\frac{5}{100}$

De la même manière, définis le tableau des fréquences marginales du caractère Y.

y_i	1	2	3	4
f_i	$\frac{11}{100}$	$\frac{30}{100}$	$\frac{38}{100}$	$\frac{21}{100}$

3. Nuage de points

Définition

On considère deux caractères quantitatifs X et Y sur une même population de *n* individus.
On note x_1 , x_2 , x_3 ,…, x_p les valeurs du caractère X,
y_1, y_2 , y_3 ,…, y_p les valeurs du caractère Y.
On appelle nuage de points associé à la série statistique double de caractère (X, Y) les points de couple de coordonnées $(x_i; y_j)$ d'effectifs non nuls.

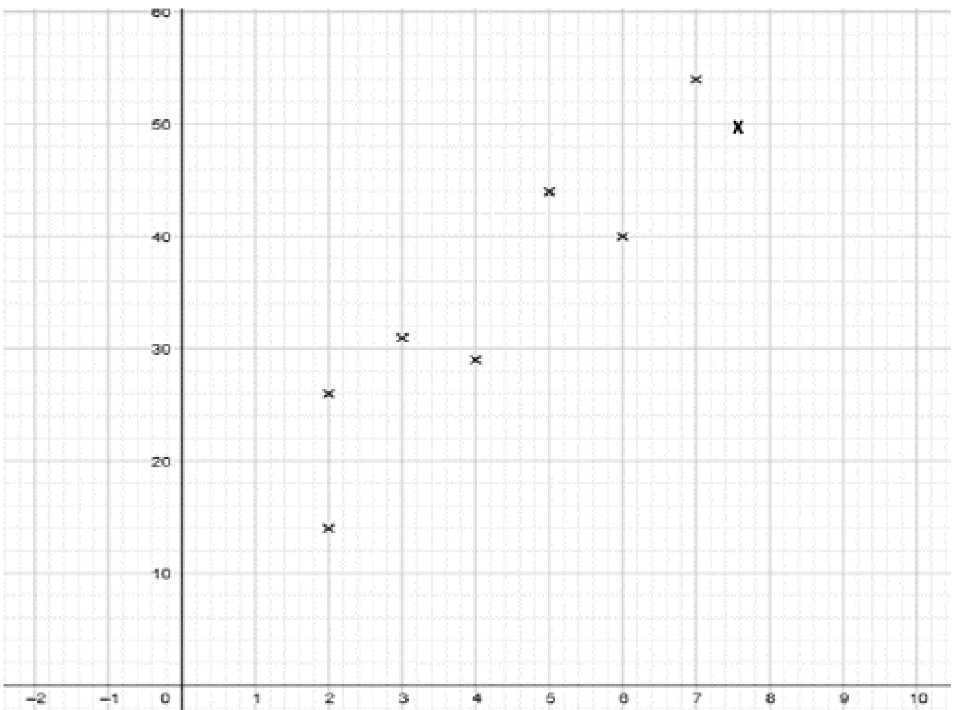
Exercice de fixation

Le tableau suivant donne le nombre d'exploitations agricoles d'une région selon leur superficie en hectares.

Superficie X	2	2	3	4	5	6	7	7,6
Nombre d'exploitations Y	14	26	31	29	44	40	54	50

Représente le nuage de points associé à cette série.

Solution



Remarque

Dans la suite, les séries doubles considérées seront comme la série de l'exemple précédent ; c'est-à-dire l'effectif n_{ij}du couple $(x_i\ ,y_j)$ vaut 1.

4. Point moyen

Définition

On appelle point moyen d'un nuage de n points M_i de coordonnées $(x_i; y_i)$ le point G de coordonnées $(x_G; y_G)$ telles que :

$x_G = \overline{X} = \frac{x_1+x_2+\cdots+x_n}{n}$; $y_G = \overline{Y} = \frac{y_1+y_2+\cdots+y_n}{n}$.

Exercice de fixation

Détermine les coordonnées du point moyen du nuage de points de la série statistique suivante :

Superficie X	2	2	3	4	5	6	7	7,6
Nombre d'exploitations Y	14	26	31	29	44	40	54	50

Réponse

C'est le point de coordonnées $(\overline{\mathrm{X}}\ ;\overline{\mathrm{Y}})$.

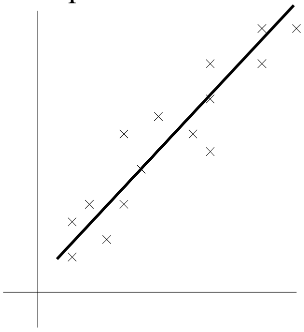
On a : $\overline{\mathrm{X}} = \frac{2+2+3+4+5+6+7+7{,}6}{8} = \frac{36{,}6}{8} = 4{,}575$

et $\overline{\mathrm{Y}} = \frac{14+26+31+29+44+40+54+50}{8} = \frac{288}{8} = 36$

Donc : G (4,575 ; 36)

II. Ajustement linéaire par la méthode des moindres carrés

Soit un nuage de points associé à une série statistique double représenté dans un repère orthogonal.
Faire un ajustement de ce nuage de points, c'est trouver une courbe qui passe le plus près « possible » du maximum de points de ce nuage.
Lorsque cette courbe est une droite, on dit que l'ajustement est affine ou linéaire.



Exemple d'ajustement par une droite.

1. Covariance

Définition

On appelle covariance de la série statistique double de caractère (X ;Y),
le nombre réel noté COV(X ; Y) tel que :

$\mathrm{COV}(X,Y) = \frac{1}{n}\sum n_{ij}(x_i - \overline{x})(y_j - \overline{y})$ ou $\mathrm{COV}(X,Y) = \frac{\sum n_{ij}x_i y_j}{n} - \overline{\mathrm{X}}\,\overline{\mathrm{Y}}$.

Exercice de fixation

Calcule la covariance de la série statistique précédente

Superficie X	2	2	3	4	5	6	7	7,6
Nombre d'exploitations Y	14	26	31	29	44	40	54	50

Solution

La covariance $\mathrm{COV}(\mathrm{X},\mathrm{Y})$de cette série statistique est $\frac{\sum x_i y_j}{n} - \overline{\mathrm{X}}\,\overline{\mathrm{Y}}$

$\mathrm{COV}(\mathrm{X},\mathrm{Y}) = \frac{2\times14+2\times26+3\times31+4\times29+5\times44+6\times40+7\times54+7{,}6\times50}{8}$ - 4,575×36

$\mathrm{COV}(\mathrm{X},\mathrm{Y}) = \frac{1503}{8} - 164{,}7$.

Donc :$\mathrm{COV}(\mathrm{X},\mathrm{Y})$ = 23,675.

2. Coefficient de corrélation linéaire

Définition

Soit V(X) la variance de la série statistique de caractère X, V(Y) la variance de la série statistique de caractère Y et COV (X ; Y) la covariance de la série statistique (X ; Y).

On appelle coefficient de corrélation linéaire de la série statistique double (X ; Y), le nombre réel noté r tel que : $r = \frac{\mathrm{COV(X,Y)}}{\sqrt{\mathrm{V(X)}}\sqrt{\mathrm{V(Y)}}}$.

Exercice de fixation

Calcule le coefficient de corrélation linéaire de la série statistique du B.1.3.

Superficie X	2	2	3	4	5	6	7	7,6
Nombre d'exploitations Y	14	26	31	29	44	40	54	50

Solution

Le coefficient de corrélation linéaire $r$de cette série statistique est: $r = \frac{\mathrm{COV(X,Y)}}{\sqrt{\mathrm{V(X)}}\sqrt{\mathrm{V(Y)}}}$

On a:

• $\mathrm{V(X)} = \frac{\sum x_i{}^2}{n} - \left(\overline{\mathrm{X}}\right)^2 = \frac{2^2+2^2+3^2+4^2+5^2+6^2+7^2+(7{,}6)^2}{8} - 4{,}575^2$

$\mathrm{V(X)} = \frac{200{,}76}{8} - 4{,}575^2 \approx 4{,}16$

• $\mathrm{V(Y)} = \frac{\sum y_i{}^2}{n} - \left(\overline{\mathrm{Y}}\right)^2 = \frac{14^2+26^2+31^2+29^2+44^2+40^2+54^2+50^2}{8} - 36^2$

$\mathrm{V(Y)} = \frac{11626}{8} - 36^2 = 157{,}25$

Donc :$r = \frac{23{,}675}{\sqrt{4{,}16\times157{,}25}}$ ≈0,92.

Remarques

• Le coefficient de corrélation linéaire permet de voir la dépendance linéaire des deux caractères X et Y.

•Le coefficient de corrélation linéaire r est un nombre réel de même signe que COV (X, Y) et on a : $-1 \leq r \leq 1$.

• Si $|r|$ est proche de 1, c'est-à-dire en pratique : $0{,}87 \leq r < 1$ ou $-1 < r \leq -0{,}87$, alors on dit qu'il y a une bonne corrélation linéaire ou une forte corrélation linéaire entre les deux caractères X et Y.

Exemple
Interprète le coefficient de corrélation linéaire ci-dessus.

Solution
On a : $r = 0{,}92$.
Comme $0{,}87 \leq r < 1$, il y a une forte corrélation entre la superficie et le nombre d'exploitations agricoles de cette région.

3. Droites de régressions

a) Propriété

Soit V(X) la variance de la série statistique de caractère X, V(Y) la variance de la série statistique de caractère Y et COV (X, Y) la covariance de X et Y.
i. Droite de régression de Yen X.
En supposant qu'il y ait une forte corrélation entre les caractères X et Y alors, la droite (D) d'équation : $y = ax + b$ où $a = \frac{\mathrm{Cov(X,Y)}}{\mathrm{V(X)}}$ et $b = \overline{y} - a\overline{x}$ est appelée la droite de régression de Yen X par la méthode des moindres carrés.
ii. Droite de régression de X en Y.
La droite (D') d'équation : $x = a'y + b'$ avec : $a' = \frac{\mathrm{Cov(X,Y)}}{\mathrm{V(Y)}}$ et $b' = \overline{x} - a'\overline{y}$ est appelée la droite de régression de X en Y par la méthode des moindres carrés.

Exercice de fixation
On considère la série statistique précédente.
On sait que : $0{,}87 \leq r < 1$.
1. Détermine une équation de la droite d'ajustement linéaire de Y en X par la méthode des moindres carrés. On donnera les arrondis d'ordre 2 de a et b.
2. Détermine une équation de la droite d'ajustement linéaire de X en Y par la méthode des moindres carrés. On donnera les arrondis d'ordre 2 de a' et b'.

Solution
1. C'est la droite (D) d'équation : $y = ax + b$ où $a = \frac{\mathrm{Cov(X,Y)}}{\mathrm{V(X)}}$ et $b = \overline{\mathrm{Y}} - a\overline{\mathrm{X}}$

$a = \frac{\mathrm{Cov(X,Y)}}{\mathrm{V(X)}} = \frac{23{,}675}{4{,}16} = 5{,}69$ et $b = \overline{\mathrm{Y}} - a\overline{\mathrm{X}} = 36 - 5{,}69 \times 4{,}575 = 9{,}97$

Donc (D) : $y = 5{,}69x + 9{,}97$.

2. C'est la droite (D') d'équation : $x = a'y + b'$ avec $a' = \frac{\mathrm{Cov(X,Y)}}{\mathrm{V(Y)}}$ et $b' = \overline{\mathrm{X}} - a'\overline{\mathrm{Y}}$

$a' = \frac{\mathrm{Cov(X,Y)}}{\mathrm{V(Y)}} = \frac{23{,}675}{157{,}25} = 0{,}15$ et $b' = \overline{\mathrm{X}} - a'\overline{\mathrm{Y}} = 4{,}575 - 0{,}15 \times 36 = -0{,}825$

Donc : (D') : $x = 0{,}15y - 0{,}825$.

Remarques

- Les droites (D) et (D') passent par le point moyen G du nuage de points.
- Si r est le coefficient de corrélation linéaire on a :

• $aa' = r^2$ et $|r| = \sqrt{aa'}$
• Si $a > 0$ et $a' > 0$, alors $r = \sqrt{aa'}$.
• Si $a < 0$ et $a' < 0$, alors $r = -\sqrt{aa'}$.
• Si $r^2 = 1$, alors $a = \frac{1}{a'}$ et les deux droites sont confondues.

b) Estimation

•La droite d'ajustement tracée du nuage de points permet graphiquement une estimation de y connaissant x (resp. x connaissant y).
•L'équation de la droite d'ajustement permet de calculer une estimation de y connaissant x (resp.x connaissant y).

Exercice de fixation
On considère la série statistique précédente.
En considérant que la tendance se poursuit ainsi, détermine le nombre d'exploitations agricoles pour une superficie de 9 ha.

Solution
Une superficie de 9 ha correspond à $x = 9$.
En utilisant l'équation de la droite par la méthode des moindres carrés, on a : y = 5,69 x + 9,97
$y = 5{,}69 \times 9 + 9{,}97 = 61{,}8$
Donc pour une superficie de 9 ha, le nombre d'exploitations agricoles est estimé à 62.

C. SITUATION COMPLEXE

Le tableau ci-dessous donne le nombre total d'adhérents au club littéraire d'un lycée au cours de l'année civile 2020.

Mois	Janv	Fév	Mars	Avr	Mai	Juin	Juil	Août	Sept	Oct	Nov	Déc
Rang x_i	1	2	3	4	5	6	7	8	9	10	11	12
Nombre d'adhérents y_i	1100	1160	1220	1370	1620	1550	1600	1500	1790	1940	2060	1980

Une Organisation Non Gouvernementale promet d'octroyer une aide financière considérable au club si le nombre d'adhérents dépasse les 3000 élèves. L'élève de la Terminale A qui dirige le club désire connaître la date à laquelle ce don pourra se faire. Il te sollicite pour l'aider.

Détermine la date (mois et année) probable de la réception de ce don.

Solution.

- **Pour trouver la date, nous allons utiliser les statistiques à deux variables,**
- **Je détermine la droite de régression linéaire,**
- **J'estime la date.**

- **Je détermine une équation de la droite de régression de Y en X.**

Soit (D) cette droite.
Une équation de (D) est sous la forme: $y = ax + b$ $avec$ $a = \frac{\mathrm{Cov(X,Y)}}{\mathrm{V(X)}}$ et $b = \overline{Y} - a\overline{X}$.

- Les coordonnées du point moyen G
On a:

$$\bar{X} = \frac{\sum x_i}{n} = \frac{78}{12} = 6{,}5$$

$$\bar{Y} = \frac{\sum y_i}{n} = \frac{18890}{12} = 1574{,}167$$

- La variance de X

$$V(X) = \frac{\sum {x_i}^2}{n} - \overline{\mathrm{X}}^2$$

$$V(X) = \frac{1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2 + 7^2 + 8^2 + 9^2 + 10^2 + 11^2 + 12^2}{12} - 6{,}5^2$$

$$\mathrm{V(X)} = \frac{650}{12} - (6.5)^2 = 11{,}917$$

- La variance de Y

$$V(Y) = \frac{\sum {y_i}^2}{n} - \overline{\mathrm{Y}}^2$$

$$V(Y) = \frac{1100^2 + 1160^2 + 1220^2 + 1370^2 + 1620 + 1550^2 + 1600^2 + 1500^2 + 1790^2 + 1940^2 + 2060^2 + 1980^2}{12} - (1574{,}167)^2$$

$$V(Y) = \frac{30889500}{12} - (1574{,}167)^2 = 96123{,}256$$

- La covariance de X et Y

$$\mathrm{Cov}(X,Y) = \frac{\sum x_i y_i}{n} - \overline{\mathrm{X}}\,\overline{\mathrm{Y}}$$

$$\mathrm{Cov}(X,Y) = \frac{1100 + 2320 + 3660 + 5480 + 8100 + 9300 + 11200 + 12000 + 16110 + 19400 + 22660 + 23760}{12} - 6{,}5 \times 1574{,}167$$

$$\mathrm{COV(X,Y)} = \frac{135090}{12} - 6{,}5 \times 1574{,}167 = 1025{,}4145$$

- Une équation de la droite (D): *y* = *ax* + *b*

$$a = \frac{1025{,}4145}{11{,}917} = 86{,}046$$

$$b = 1574{,}167 - 86{,}046 \times 6{,}5 = 1014{,}868$$

D'où (D): $\boldsymbol{y} = 86{,}046x + 1014{,}868$

- Je déduis le rang du mois pour $y = 3000$

$y = 3000$ équivaut à $x = \frac{3000 - 1014{,}868}{86{,}046} = 23{,}071$

Le rang cherché est sensiblement égal à 24.

- Je donne la date (mois et année) probable de la réception de ce don.

La date probable de la réception de ce don est Décembre 2021.

D. EXERCICES

1. Exercices de fixation

Exercice 1

On considère la série statistique suivante :

x_i	1	4	7	8	10
y_i	2	7	8	10	13

Détermine les coordonnées du point moyen G.

Solution

On a : $x_G = \bar{x} = \frac{1+4+7+8+10}{5} = \frac{30}{5} = 6$

$y_G = \bar{y} = \frac{2+7+8+10+13}{5} = \frac{40}{5} = 8$; donc G (6 ; 8)

Exercice 2

Détermine la covariance de la série statistique de l'exercice 1

Solution

On a : $\mathrm{COV}(X, Y) = \frac{\sum n_{ij} x_i y_j}{n} - \bar{X}\bar{Y} = \frac{1\times2+4\times7+7\times8+8\times10+10\times13}{5} - 6\times 8 = \frac{296}{5} - 48 = 11{,}2$

2. Exercices de renforcement

Exercice 1

La tension artérielle est une donnée médicale correspondant à la pression du sang dans les artères. On la mesure chez les patients car une tension anormale peut être le symptôme de pathologies cardiovasculaires comme l'hypertension artérielle.

La tension artérielle d'une personne comporte deux mesures :

- la Tension Artérielle Systolique (notée TAS) ;
- la Tension Artérielle Diastolique (notée TAD).

Le tableau suivant regroupe les mesures de la tension artérielle pour un groupe de personnes saines :

Ages	26	39	40	50	53	56
TAS (en mm Hg)	128	126	118	136	142	145
TAD (en mm Hg)	80	83	92	91	87	93

On s'intéresse à l'évolution de la TAS en fonction de l'âge.

Pour cela, on symbolise les données du tableau à l'aide de points de coordonnées (x;y) où x est l'âge de la personne et y sa TAS.

Détermine les coordonnées du point moyen des 3 points dont l'âge est le plus petit.

Détermine les coordonnées du point moyen des 3 autres points.

Solution

Désignons par G_1 le point moyen des trois points dont l'âge est le plus petit et par G_2 celui des trois autres points. On a :

$$x_{G_1} = \frac{26+39+40}{3} = 35$$

$$y_{G_1} = \frac{128+126+118}{3} = 124$$

Donc $G_1(35 ; 124)$

$$x_{G_2} = \frac{50+53+56}{3} = 53$$

$$y_{G_2} = \frac{136+142+145}{3} = 141$$

Donc $G_2(53 ; 141)$

Exercice 2

On considère la série statistique suivante :

X	0	1	2	3	4		5	6	7	8
Y	160	110	100	72	36		29	20	10	3

1) Détermine la covariance de la série statistique.
2) Détermine le coefficient de corrélation linéaire de cette série. Interprète ce coefficient de corrélation linéaire.
3) Détermine une équation de la droite d'ajustement linéaire de Y en X du nuage de points de la série par la méthode des moindres carrés.
4) Détermine une équation de la droite d'ajustement linéaire de X en Y.

Solution

1) On a : $\bar{X} = \frac{0+1+2+3+4+5+6+7+8}{9} = 4$

$$\bar{Y} = \frac{160+110+100+72+36+29+20+10+3}{9} = 60$$

$$COV(X,Y) = \frac{\sum n_{ij}x_iy_j}{n} - \bar{X}\bar{Y} = \frac{1\times110+2\times100+3\times72+4\times36+5\times29+6\times20+7\times10+8\times3}{9} - 4\times 60 = -125{,}67$$

2) On a : : $r = \frac{\text{COV(X,Y)}}{\sqrt{\text{V(X)}}\sqrt{\text{V(Y)}}}$; or $\text{V(X)} = \frac{\sum x_i^2}{n} - (\bar{\text{x}})^2 = \frac{1^2+2^2+3^2+4^2+5^2+6^2+7^2+8^2}{9} - 4^2 = 6{,}67$ et

• $\text{V(Y)} = \frac{\sum y_i^2}{n} - (\bar{\text{y}})^2 = \frac{160^2+110^2+100^2+72^2+36^2+29^2+20^2+10^2+3^2}{9} - 60^2 = 2570$

Donc $r = \frac{-125{,}67}{\sqrt{6{,}67}\times\sqrt{2570}} = -0{,}96$

Comme $-1 < r \leq -0{,}87$ alors il y a une forte corrélation linéaire entre les caractères X et Y.

3) C'est la droite (D) d'équation : $y = ax + b$ où $a = \frac{\text{Cov(X,Y)}}{\text{V(X)}}$ et $b = \bar{\text{Y}} - a\bar{\text{X}}$

$a = \frac{-125{,}67}{6{,}67} = -18{,}84$ et $b = 60 + 18{,}84\times4 = 135{,}36$ donc(D) : $y = -18{,}84x + 135{,}36$

4). C'est la droite (D') d'équation : $x = a'y + b'$ avec $a' = \frac{\text{Cov(X,Y)}}{\text{V(Y)}}$ et $b' = \bar{\text{X}} - a'\bar{\text{Y}}$

$a' = \frac{-125{,}67}{2570} = -0{,}049$ et $b' = 4+0{,}049\times60 = 6{,}94$ donc (D') : $x = -0{,}049y + 6{,}94$

3. Exercice d'approfondissement

Dans le cadre d'un recensement portant sur le nombre de travailleurs dans les champs d'hévéa, un agent a visité huit (8) exploitations .Un exploitant voudrait estimer le nombre de travailleurs que prendrait une exploitation de 16ha d'hévéa. Pour cela, l'agent recenseur a recueilli les informations consignées dans le tableau ci-dessous.

Nombre x de travailleurs	2	4	4	5	7	7	8	8

Superficie exploitée y(en ha)	3	5	6	7	10	11	8	12

1) Représente le nuage de points correspondant à la série statistique double (X, Y) dans le plan muni d'un repère orthonormé.
On prendra sur l'axe des abscisses 1cm pour 1 travailleur et sur l'axe des ordonnées 1cm pour une superficie de 1ha.

Pour les questions 2) 3) 4) et 5), les résultats seront arrondis à l'ordre 2.

2) Justifie que le point moyen a pour coordonnées (5,63; 7,75).

3) On note $V(X)$ la variance de $X, V(Y)$ la variance de Y et $Cov\ (X\ ;\ Y)$ la covariance de X et Y.
Justifie que $V(X)\ =\ 4{,}18$ et $Cov\ (X, Y)\ =\ 5.57$.

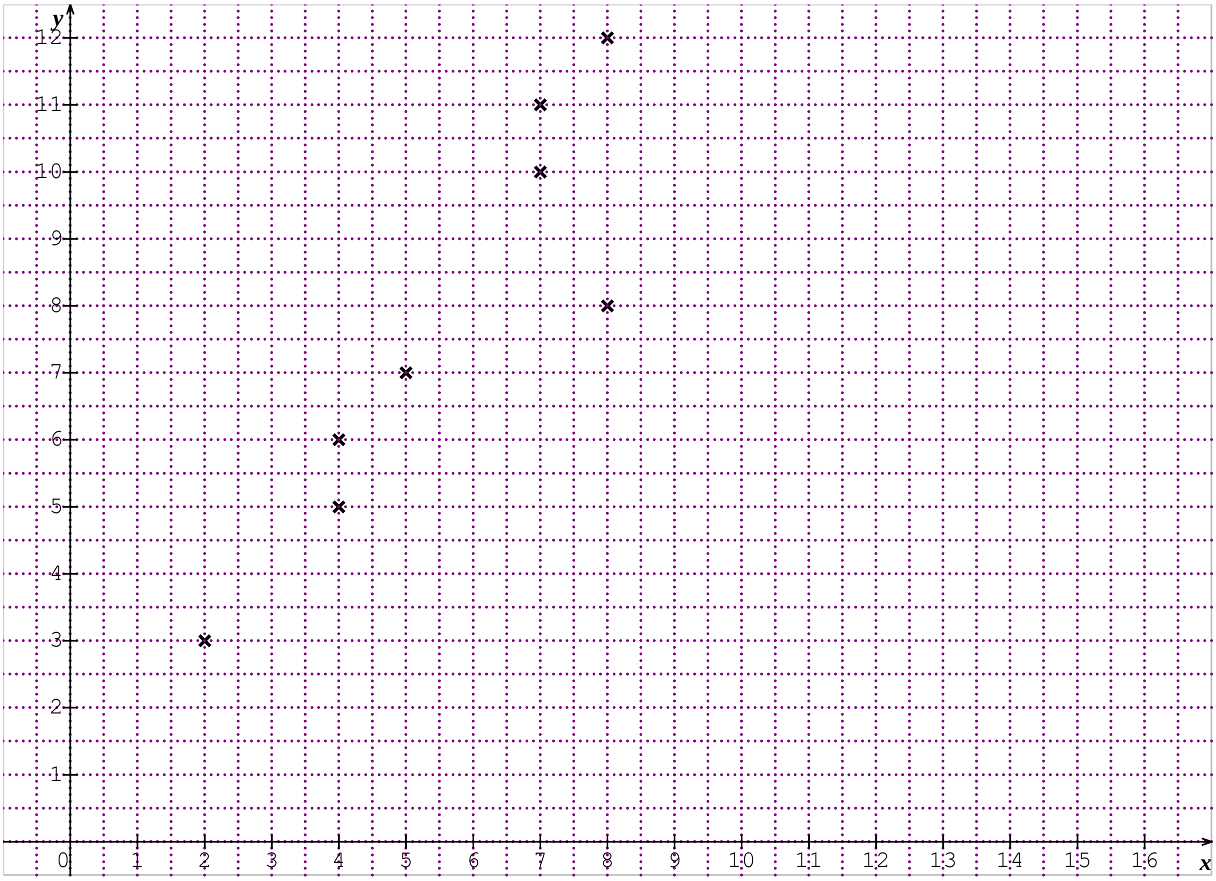
4) a) Calcule le coefficient de corrélation linéaire r de la série (X, Y).
b) Interprète le résultat obtenu précédemment.

5) a) Justifie qu'une équation de la droite (D) d'ajustement de X en Y par la méthode des moindres carrés est : y =1,28x +0,54.

b) Trace(D) sur le graphique précédent.

6) Utilise l'ajustement précédent pour répondre à la préoccupation de l'exploitant.
On donnera l'arrondi d'ordre zéro du résultat.

Solution
<u>1) Représentation du nuage de points associé à la série</u>



2) Justifions que le point moyena pour coordonnées (5,63 ;7,75).
Soit G($\bar{X}$; $\bar{Y}$)le point moyen du nuage réprésentant cette série statistique.
On a :

$$\bar{X} = \frac{2+4+4+5+7+7+8+8}{8} = 5\,,625 \approx 5{,}63$$

$\bar{X} = 5{,}63.$

$$\bar{Y} = \frac{3+5+6+7+10+11+8+12}{8} = 7{,}75$$

Donc : G(5,63;7,75).

3) Justifions que V(X)=4,18, V(Y)=8 ,44 et Cov (X, Y) =5.57

$$V(X) = \frac{\sum {x_i}^2}{n} - \overline{\mathrm{X}}^2 = \frac{2^2+4^2+4^2+5^{2+}7^2+7^2+8^2+8^2}{8} - 5{,}63^2$$

$V(X) = 4{,}178$; donc V(X) =4,18.

$$V(Y) = \frac{\sum {y_i}^2}{n} - \overline{\mathrm{Y}}^2 = \frac{3^2+5^2+6^2+7^{2+}10^2+11^2+8^2+12^2}{8} - 7{,}75^2$$

$V(Y)$ =8,437 ; donc V(Y)=8 ,44.

$$\mathrm{COV}(X,Y) = \frac{\sum x_i y_j}{n} - \overline{\mathrm{X}}\overline{Y}.$$

$$COV(X,Y) = \frac{2\times3+4\times5+4\times6+5\times7+7\times10+7\times11+8\times8+8\times12}{8} - 5{,}63\times7{,}75.$$

$COV\ (X,Y)$ =5,37

4) a) Calculons le coefficient de corrélation linéaire r de la série (X,Y).

$$r = \frac{\mathrm{COV(X,\ Y)}}{\sqrt{\mathrm{V(X)}}\sqrt{\mathrm{V(Y)}}} = \frac{5{,}37}{\sqrt{4{,}18}\times\sqrt{8{,}44}}$$

$r\ = 0.904\ soit\ r = 0{,}90.$

b) Interprète le résultat obtenu précédemment.
On remarque que : 0,87$\leq r < 1$, ainsi, on peut conclure qu'il y a une forte corrélation entre le nombre de travailleurs et la superficie exploitée sur les 8 différentes exploitations.

5) a)Justifie qu'une équation de la droite (D) d'ajustement de X en Y par la méthode des moindres carrés est : y =1,28x+0,54

Puisqu'il y a une forte corrélation entre le nombre de travailleurs et la superficie exploitée alors

(D) a pour équation : $y = a\,x + b$ où $a = \frac{\mathrm{Cov(X,Y)}}{\mathrm{V(X)}}$ et $b = \overline{\mathrm{Y}} - a\overline{\mathrm{X}}$

$a = \frac{5{,}37}{4{,}18} = 1{,}28$ et $b = 7{,}75 - 1{,}28\times5{,}63 = 0\,,54.$

Soit (D) : $y\ =\ 1{,}28x\ + 0{,}54.$