

STATISTIQUES : SERIES STATISTIQUES DOUBLES

A GENERALITES

I. Introduction :

Les premières notions de la statistique proviennent du recensement à but économique (payement des impôts) et militaire (connaître le nombre de bras valide pour la guerre).

II. QUELQUES DEFINITIONS :

1. Statistique :

La statistique se définit comme étant l'ensemble des méthodes qui permettent d'étudier les informations sur un phénomène donné par la réunion, la présentation et l'analyse à partir des données chiffrées

Distinction entre la statistique et les statistiques :

La statistique est la science qui étudie les chiffres et ces chiffres constituent les statistiques

Exemple : Les statistiques agricoles ; les statistiques démographiques ; les statistiques scolaires.etc.....

2. Population statistique-Unité statistique :

Les ensembles étudiés portent le nom d'ensembles statistiques ou d'univers statistiques ou populations statistiques.

Chaque élément de cet ensemble (population) est appelé une unité statistique.

3. Caractère et modalités :

Une unité statistique peut être étudiée suivant un ou plusieurs caractères.

Exemples : Pour étudier la population de Bamako, on peut s'intéresser à la taille ; aux poids ; aux sexes ; à l'ethnie ; à la religion ; à l'état matrimonial ;.....

Les modalités d'un caractère sont les différentes éventualités possibles de ce caractère.

Le caractère matrimonial à quatre(4) modalités à savoir : Célibataire ; marié(e) ; divorcé(e) et veuf (vé).

Les caractères sont classés en deux catégories :

- **Caractères qualitatifs** : Ce sont des caractères dont les modalités ne sont pas mesurables (ethnie ; sexe ; religion ;.....)
- **Caractères quantitatifs** : Ce sont des caractères dont les modalités sont mesurables (taille ; poids ;.....)

4. Variables statistiques :

On appelle variable statistique, un caractère quantitatif ; elle peut être discrète ou continue.

- **Variable discrète** : on dit qu'une variable est discrète lorsque ses valeurs sont des quantités dénombrables.
- **Variable continue** : C'est le cas où le caractère peut prendre théoriquement toutes les valeurs d'un intervalle donné.
Dans ce cas on définit les classes (intervalles de valeurs).

B.SERIES A DEUX CARACTERES

I ORGANISATION DES DONNEES :

1. Introduction :

On a relevé le poids (en kg) et la taille (en cm) de dix élèves .on a obtenue le tableau suivant :

Poids X_i	65	68	62	62	68	62	65	68	71	74
Taille Y_i	165	168	174	168	171	174	174	171	171	174

Ces données permettent de définir deux séries statistiques à un caractère

Série des poids :

Poids X_i	62	65	68	71	74
Effectif n_i	3	2	3	1	1

Série des tailles :

Taille Y_i	165	168	171	174
Effectif n_i	1	2	3	4

Posons $X = \{62, 65, 68, 71, 74\}$ et $Y = \{165, 168, 171, 174\}$

A chaque couple $(x; y)$ de $X * Y$, on associe le nombre de poids x et de taille y .

On définit ainsi une série statistique à deux caractères ou séries double.

Le nombre n est appelé effectif de la modalité $(x; y)$.

Séries marginales :

Regroupons les deux séries dans un même tableau en indiquant l'effectif n de la modalité $(x; y)$.

Poids \ Taille	62	65	68	71	74	Total
165	0	1	0	0	0	1

168	1	0	1	0	0	2
171	0	0	2	1	0	3
174	2	1	0	0	1	4
Total	3	2	3	1	1	

Les totaux obtenus sur la dernière ligne sont les effectifs de la série des poids.

Les totaux obtenus sur la dernière colonne sont des effectifs obtenus sur la série des tailles.

Ces effectifs apparaissent en marge du tableau du tableau à double entrée.

La série des poids et la série des tailles sont appelées les séries marginales de la série double.

Nuage de points :

Le plan est muni d'un repère orthogonal.

On représente chaque modalité $(x; y)$ par un disque dont l'aire est proportionnelle à l'effectif de $(x; y)$. Les disques d'aires nuls ne sont pas représentés.

Exemple : Reprenons la série de l'exemple introductif

Graphe1

Point moyen d'un nuage de point :

On appelle point moyen G d'un nuage de point, le point défini par $G(\bar{X}; \bar{Y})$

$$\text{Avec } \bar{X} = \frac{\sum n_i X_i}{N}; \bar{Y} = \frac{\sum n_i Y_i}{N}$$

Dans l'exemple introductif le point G est défini par :

$$\bar{X} = \frac{62(3) + 65(2) + 68(3) + 71(1) + 74(1)}{10} = 66,5$$

$$\bar{Y} = \frac{165(2) + 68(3) + 71(1) + 74(1)}{10} = 171$$

$$G(66,5; 171)$$

II. AJUSTEMENT LINEAIRE :

Faire un ajustement linéaire, c'est trouver la droite la plus proche possible des différents points.

1. Méthode de Mayer :

Pour déterminer la droite d'ajustement par la méthode de Mayer, on peut :

Partager le nuage en deux sous nuages disjoint E_1 et E_2 de même effectif dans l'ordre où les points se présentent ;

Déterminer respectivement les points moyens G_1 et G_2 des sous nuages E_1 et E_2 ;

Déterminer une équation de la droite $(G_1 G_2)$.

$G_1 G_2$ Est la droite de Mayer ajustant le nuage des points de la série double.

Exemple : Le tableau suivant donne le chiffre d'affaires d'une entreprise exprimé en million de francs CFA pendant huit années consécutives

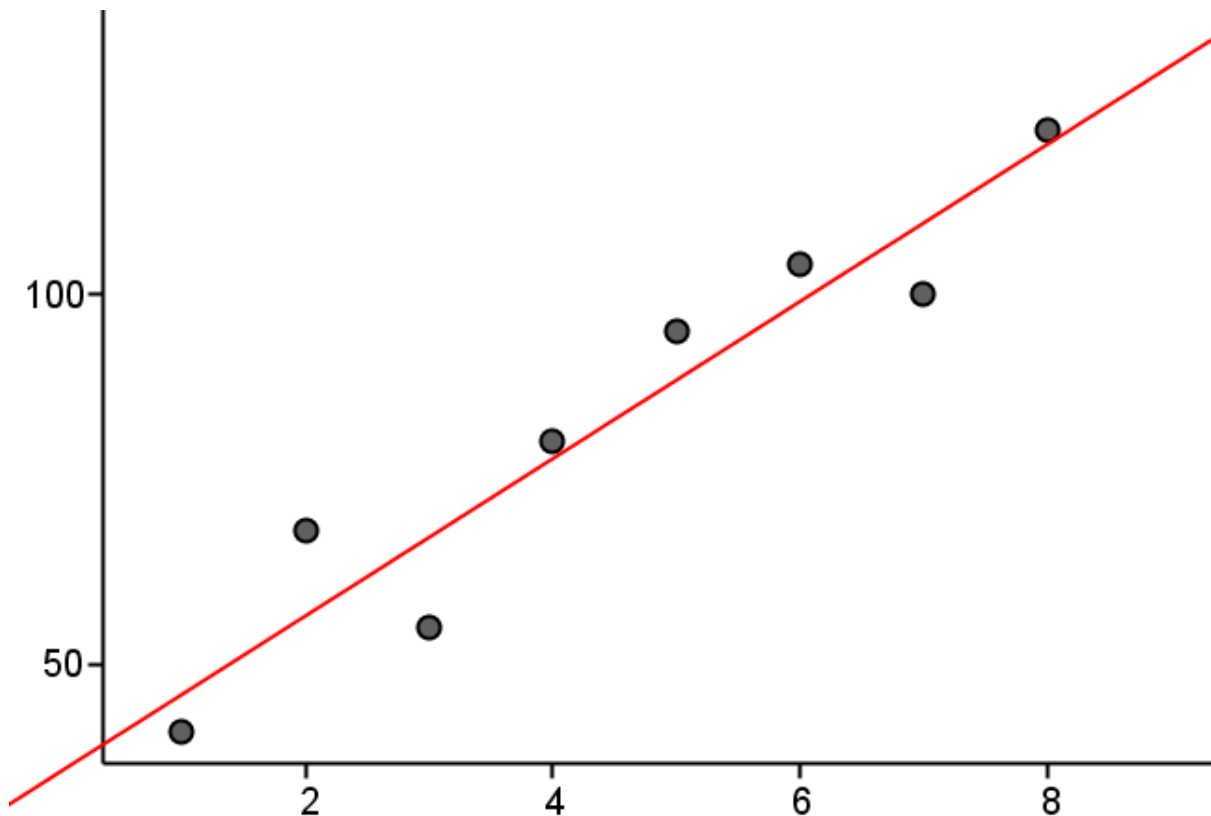
Nombre de l'année (X_i)	1	2	3	4	5	6	7	8
Chiffre d'affaire (Y_i)	41	68	55	80	95	104	100	122

Consigne

- Représenter le nuage des points de cette série double
- Calculer les coordonnées du point moyen G des nuages de points
- Déterminer une droite « la plus près possible » des points du nuage.

Solution

- Représentons le nuage des points de cette série double



b. Calculons les coordonnées du point moyen G du nuage des points

$$G \left(\frac{1 + 2 + 3 + 4 + 5 + 6 + 7 + 8}{8}, \frac{41 + 68 + 55 + 80 + 95 + 104 + 100 + 122}{8} \right) = \left(4,5, 83,125 \right)$$

c. Trouvons la droite « la plus près proche » des points du nuage

Soient E_1 et E_2 les deux sous nuages de même effectif dans l'ordre où les points se présentent

Sous nuage E_1

X_i	1	2	3	4
Y_i	41	68	55	80

Sous nuage E_2

X_i	5	6	7	8
Y_i	95	104	100	122

Déterminons respectivement les points moyens G_1 et G_2 des sous nuages E_1 et E_2

$$G_1 \left(\frac{\frac{1+2+3+4}{4}}{\frac{41+68+55+80}{4}} \right) = \begin{pmatrix} 2,25 \\ 61 \end{pmatrix}$$

$$G_2 \left(\frac{\frac{5+6+7+8}{4}}{\frac{95+104+100+122}{4}} \right) = \begin{pmatrix} 6,5 \\ 105,25 \end{pmatrix}$$

Déterminons l'équation de la droite ($G_1 G_2$).

L'équation d'une droite est de la forme $y = ax + b$

$$G_1 \in (G_1 G_2) \Rightarrow Y_{G_1} = aX_{G_1} + b \Rightarrow 61 = a(2,5) + b$$

$$G_2 \in (G_1 G_2) \Rightarrow Y_{G_2} = aX_{G_2} + b \Rightarrow 105,25 = a(6,5) + b$$

Formons le système avec ces deux solutions :

$$\begin{cases} 2,5a + b = 61 \\ 6,5a + b = 105,25 \end{cases}$$

Après résolution on trouve $a = 11,0625$ et $b = 33,36375$

D'où $Y = 11,0625X + 33,34375$ est l'équation de la droite ($G_1 G_2$)

2. Ajustement linéaire par la méthode des moindres carrés :

a- Variance ou fluctuation :

La variance d'une série est la moyenne des carrés des écarts à la moyenne c'est-à-dire :

Si $x_1, x_2, \dots, x_k$ sont les valeurs du caractère et $n_1 ; n_2 ; \dots ; n_k$ les effectifs.

La variance est notée : $V = \frac{1}{N} \sum_i^k n_i x_i^2 - (\bar{x})^2$ ou encore on la note : $V = \frac{1}{N} \sum_i^k (n_i x_i - \bar{x})^2$

b- Covariance :

La covariance d'une série statistique à deux caractères x et y , de moyennes respectives $\bar{x}$ et $\bar{y}$ et d'effectif total N est le nombre réel noté $\text{Cov}(X, Y)$ ou $\sigma_{x,y}$ tel que :

$$\text{Cov}(x,y) = \frac{\sum n_{ij}(x_i - \bar{X})(y_i - \bar{Y})}{\text{Effectif total}}$$

ou

$$\text{Cov}(x,y) = \frac{\sum n_{ij}x_i y_i}{\text{Effectif total}} - \bar{X} \cdot \bar{Y}$$

c- Droit de régression de y en x :

La droite (D) de régression de y en fonction de x passe par le point G ($\bar{X}, \bar{Y}$), point moyen du nuage de point son équation est $y = ax + b$ avec :

$$a = \frac{\text{Cov}(X,Y)}{V(X)} \text{ et } b = \bar{Y} - a\bar{X}$$

où V(X) est la variance et a pour formule : $V(X) = \frac{\sum X_i^2}{N} - \bar{X}^2$

d- Droite de régression de X en Y :

La droite (D') de régression de X en fonction de Y passe par le point G ($\bar{X}; \bar{Y}$), point moyen du nuage de point. Son équation est $x = a'y + b'$ avec :

$$a' = \frac{\text{Cov}(X,Y)}{V(Y)} \text{ et } b' = \bar{X} - a'\bar{Y}$$

où V(Y) est la variance et a pour formule : $V(Y) = \frac{\sum Y_i^2}{N} - \bar{Y}^2$

Remarque :

G ($\bar{X}; \bar{Y}$) appartient aux deux droites de régression pour tracer ces droites, il suffit donc de déterminer un autre point vérifiant l'équation.

e- Corrélation linéaire

Deux variables statistiques X et Y sont dites en corrélation linéaire lorsque la courbe de régression de Y en X et la courbe de régression de X en y sont des droites.

f- Coefficient de corrélation linéaire :

➤ **Définition :**

On appelle coefficient de corrélation linéaire d'une série statistique double de caractère (X ; Y) , le

nombre réel **r** définie par $\mathbf{r} = \frac{\text{Cov}(X;Y)}{\sqrt{V(X).V(Y)}}$

➤ **Propriétés de r :**

P₁ : Si $0,87 \leq |r| \leq 1$, on considère que la corrélation linéaire entre les deux variables est forte.

P₂ : si $r = 1$, on dit que la corrélation est parfaite.

g- Estimation et prévision

Lorsqu'il existe une bonne corrélation linéaire, il est possible de prévoir la valeur de y connaissant celle de x, en utilisant l'équation de la droite de régression de y en x ; et vice – versa.

Exercices d'application :

Exercice 1 :

Un pharmacien observe, durant les 6 premiers mois de l'ouverture de son officine, le chiffre d'affaire en millions de francs CFA. Le résultat de l'observation est résumé dans le tableau suivant où x désigne le numéro du mois et y le chiffre d'affaires – correspondant.

X	1	2	3	4	5	6
Y	12	13	15	19	21	22

- 1- Calculer $\bar{X}$ et $\bar{Y}$ les moyennes des variables X et Y.
- 2- Représenter graphiquement le nuage de points de cette série statistique double ainsi que le point G. on prendra : (Unités : 2 cm en abscisses et 1 cm pour 2 unités en ordonnées).
- 3- Calculer la variance $V(X)$ de x et la covariance $Cov(X,Y)$ de x et y.

N.B : Les résultats seront donnés sous forme de fractions irréductibles.

- 4- Démontrer qu'une équation de la droite de régression (D) de y en x est :
$$y = \frac{78}{35} x + 9,2$$
- 5- Tracer la droite (D).
- 6- En utilisant la droite (D), calculer une estimation du chiffre d'affaires de cette pharmacie à la fin du 7^{ème} mois.

Exercice 2 :

En prévision du lancement d'un nouveau produit, une société a effectué une enquête auprès de clients éventuels pour fixer le prix de vente (en milliers de francs) de ce produit. Les résultats sont donnés dans le tableau ci – dessous.

Prix de vente en milliers de francs xi	9	10	11	12	14	15	16	17
Nombre d'acheteurs potentiels yi	180	160	150	130	100	90	80	70

- 1- Représenter graphiquement le nuage de points de cette série statistique. On prendra pour unité :
1 cm pour 1000 francs en abscisse
1 cm pour 10 acheteurs en ordonnée.
- 2- Déterminer les coordonnées du point G
- 3- Déterminer le coefficient de corrélation linéaire r puis interpréter le résultat obtenu.
- 4- Déterminer une équation de la droite (D) de régression de y en x. Tracer (D)
- 5- Estimer graphiquement le prix maximum à fixer pour qu'il y ait au moins 50 acheteurs potentiels.
- 6- En utilisant l'équation de la droite (D) de régression de y en x, puis déterminer :
 - a) Le nombre d'acheteurs à prévoir si le prix est fixé à 13 000 F
 - b) Le prix à fixer pour que le nombre d'acheteurs potentiels soit supérieur à 250.

Solution

Exercice 1

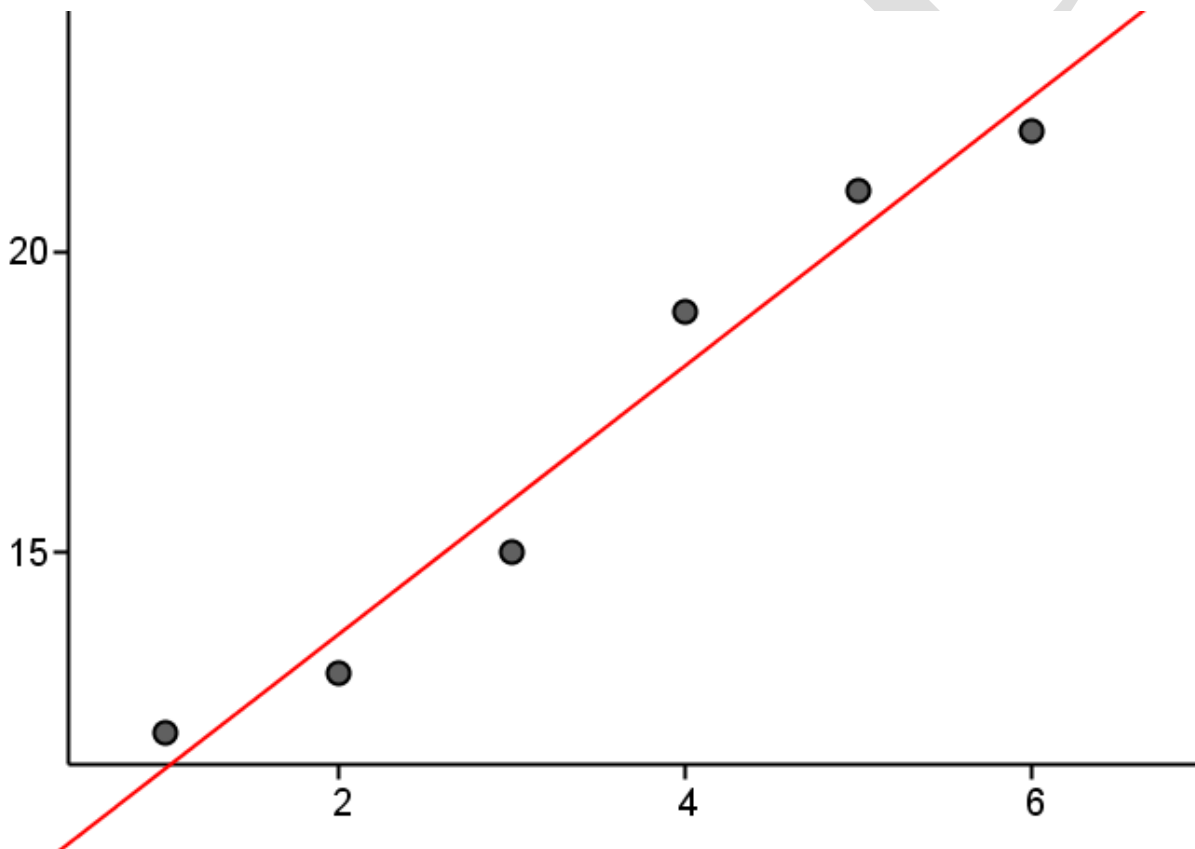
Pour le calcul, on pourra s'aider du tableau ci-dessous :

X_i	Y_i	X_i, y_i	X_i^2	Y_i^2	
1	12	12	1	144	
2	13	26	4	169	
3	15	45	9	225	
4	19	76	16	361	
5	21	105	25	441	
6	22	132	36	484	
21	102	396	91	2 331	Total

1- Calcul de $\bar{X}$ et $\bar{Y}$ les moyennes des variables x et y.

$$\bar{X} = \frac{\sum_{i=1}^6 x_i}{6} = \frac{21}{6} = \frac{7}{2} = 3,5 ; \bar{Y} = \frac{\sum_{i=1}^6 y_i}{6} = \frac{102}{6} = 17$$

2- Le nuage de points de la série statistique ainsi que le point moyen G.



3- Calcul de la variante $V(x)$ de x et la covariance $\text{Cov}(x, y)$ de x et y.

$$V(X) = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{70}{24} = \frac{35}{12} ; \text{cov}(X; Y) = \frac{396}{6} - \left(\frac{7}{2} \times 17\right) = \frac{39}{6} = \frac{13}{2}$$

4- Montons qu'une équation de la droite de régression (D) de y en x est : $y = \frac{78}{35}x + 9,2$

$$a = \frac{\text{cov}(X,Y)}{V(X)} = \frac{\frac{13}{2}}{\frac{35}{12}} = \frac{13}{2} \times \frac{12}{35} = \frac{156}{70} = \frac{78}{35} \text{ et } b = \bar{Y} - a\bar{X} = 17 - \frac{78}{35} \times \frac{7}{2} = 9,2$$

donc l'équation de la droite de régression est : $y = \frac{78}{35}x + 9,2$

5- Tracé de la droite (D). (Voir nuage de points)

La droite (D) passe par le point moyen point G (3,5 ; 17) et par un autre point que l'on détermine en remplaçant dans l'équation de la droite de régression, x par une valeur que l'on choisira.

Par exemple, si $x = 0$, on obtient $y = 9,2$. (D) passe donc également par le point A(0 ; 9,2).

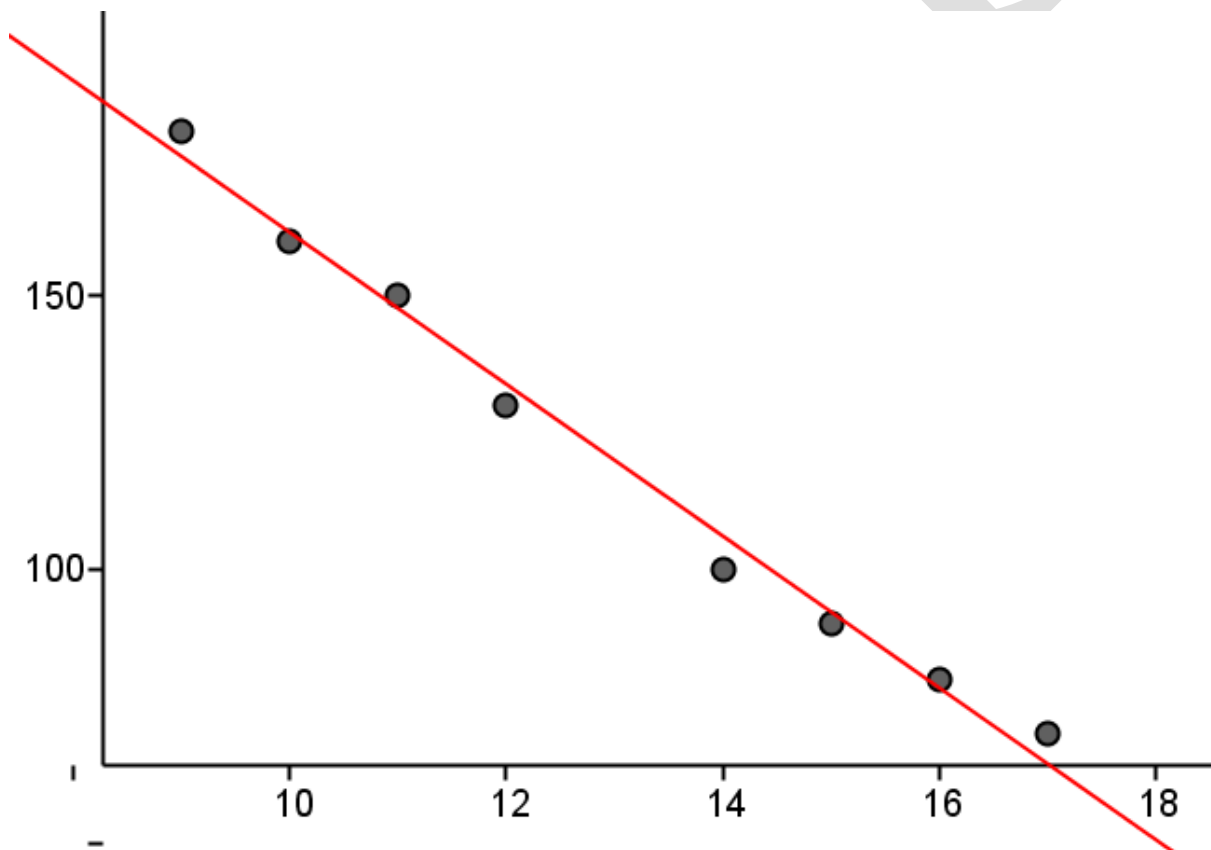
6- Calcul d'une estimation du chiffre d'affaires de cette pharmacie à la fin du 7^{ème} mois.

Le 7^{ème} mois correspond à $x = 7$. Il s'agit de déterminer la valeur de y associée à cette valeur de x.

Pour $x = 7$, on a : $y = \frac{78}{35} \times 7 + 9,2 = 24,8$ millions de francs CFA.

Exercice 2 :

- 1- Représentation graphique du nuage de points de cette série statistique (unité : 1 cm pour 1000 francs en abscisse, 1 cm pour 10 acheteurs en ordonnée).



- 2- Déterminons les coordonnées du point G.

$$\bar{X} = \frac{\sum_{i=1}^8 x_i}{8} = \frac{9+10+11+12+14+15+16+17}{8} = 13$$

$$\bar{Y} = \frac{\sum_{i=1}^8 y_i}{8} = \frac{180+160+150+130+100+90+80+70}{8} = 120$$

=> G(13 ; 120)

}

3-Détermination du coefficient de corrélation linéaire r .

$$\text{Cov}(X, Y) = \text{Cov} = \frac{\sum X_i \cdot Y_i}{N} - \bar{X} \cdot \bar{Y}$$

$$\Rightarrow \text{Cov}(X, Y) = \frac{9 \cdot 180 + 10 \cdot 160 + 11 \cdot 150 + 12 \cdot 130 + 14 \cdot 100 + 15 \cdot 90 + 16 \cdot 80 + 17 \cdot 70}{8} - 13 \cdot 120$$

$$\text{Cov}(X, Y) = -103,75$$

$$r = \frac{\text{cov}(X, Y)}{\sqrt{V(X) \cdot V(Y)}}$$

$$V(X) = \frac{\sum X_i^2}{N} - \bar{X}^2 = \frac{9^2 + 10^2 + 11^2 + 12^2 + 14^2 + 15^2 + 16^2 + 17^2}{8} - 13^2$$

$$V(X) = 7,499$$

$$V(Y) = \frac{\sum Y_i^2}{N} - \bar{Y}^2 = \frac{180^2 + 160^2 + 150^2 + 130^2 + 100^2 + 90^2 + 80^2 + 70^2}{8} - 120^2$$

$$V(Y) = 1450,003$$

$$r = \frac{\text{cov}(X, Y)}{\sqrt{V(X) \cdot V(Y)}} = \frac{-103,75}{\sqrt{7,499 \cdot 1450,003}} = -0,995$$

$0,87 \leq |r| \leq 1$: il existe alors une forte corrélation linéaire entre les variations X et Y . linéaire
L'on peut donc faire un ajustement linéaire entre X et Y .

4-Equation de la droite (D) de régression de Y en X

$$(D) : y = ax + b$$

$$A = \frac{\text{cov}(X, Y)}{V(X)} = \frac{-103,75}{7,499} = -13,88$$

$$\text{Alors on a } \bar{y} = a\bar{x} + b \Rightarrow b = \bar{y} - a\bar{x} = 120 - (-13,88) \times 13 = 299,83$$

$$(D) : y = -13,88x + 299,83$$

Tracé de (D) : (D) passe par le point moyen $G(13 ; 120)$. Déterminons un second en remplaçant par exemple x par 10. On trouve $y = 161,03$. On obtient : $A(10 ; 161,03)$

5-Pour qu'il ait au moins 50 acheteurs potentiels, on lit graphiquement que le prix à fixer ne doit pas dépasser 18000 francs.

6- a) le nombre d'acheteurs à prévoir, si le prix est fixé à 13000 F, correspond à $x = 13$.

$$Y = -13,88 \times 13 + 299,83 = 119,39$$

b)il s'agit de déterminer x pour y = 250

$$Y = -13,88x + 299,83 \text{ alors } x = \frac{y-299,83}{-13,88}$$

$$\text{Pour } y = 240, \text{ on a : } x = \frac{250-299,83}{-13,88} = 3,59 \text{ milliers de franc} = 3590 \text{ F}$$

Exercice 3 :

Pour préparer la retraite de ces membres a planté en 1991 des anacardiers qui sont rentrés en production 3 ans plutard.

Le tableau suivant donne l'évolution des productions depuis la 1^{ère} année des récoltes.

- 1.En quelle année cette cooperative a-t-elle obtenu 278 tonnes de production ?
- 2.Représenter le nuage de point de la serie double.
- 3.Déterminer l'arrondie d'ordre 2 du coefficient de corrélation linéaire de cette serie.

La corrélation entre les variables X et Y est elle bonne ?

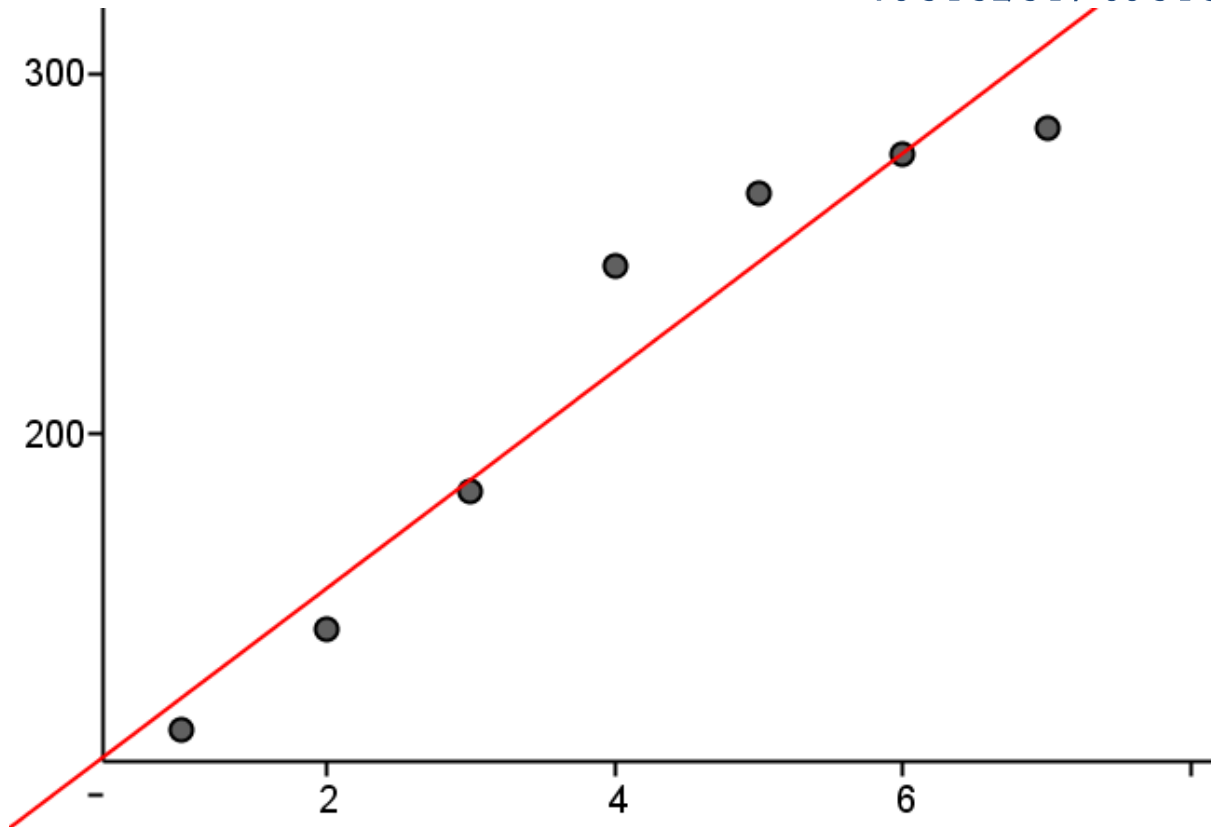
4. Déterminer par la moindre carrée, une équation de la droite de régression de Y enX.

En quelle année la coopérative produira t elle 350 tonnes ?

Ordre X_i de l'année de production	1	2	3	4	5	6	7
Quantité Y_i de production en (tonnes)	118	146	184	247	267	278	285

Solution :

1. Cette coopérative a obtenue 278 tonnes de sa production en 2000.
2. Représentons le nuage de point de la série double



3. Déterminons l'arrondie d'ordre 2 du coefficient de corrélation linéaire de cette série

X_i	Y_i	X_i^2	Y_i^2	$X_i Y_i$
1	118	1	13924	118
2	146	4	21316	292
3	184	9	33856	552
4	247	16	61009	988
5	267	25	71289	1335
6	278	36	77284	1668
7	255	49	65024	1785
Total	28	140	343703	6798

Le coefficient de corrélation linéaire de la série double est : $r = \frac{cov(X;Y)}{\sqrt{V(X)*V(Y)}}$

Avec $\bar{X} = \frac{1}{7} \sum_{i=1}^7 n_i X_i = \frac{28}{7} = 4$ et $\bar{Y} = \frac{1}{7} \sum_{i=1}^7 n_i Y_i = \frac{1495}{7}$

$$V(X) = \frac{1}{7} \sum_{i=1}^7 n_i (X - \bar{X})^2 = \frac{28}{7} = 4$$

$$V(Y) = \frac{1}{7} \sum_{i=1}^7 n_i (Y - \bar{Y})^2 = \frac{170896}{49}$$

La covariance de la serie double

$$\text{cov}(X; Y) = \frac{1}{N} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) = \frac{758}{7}$$

$$r = \frac{\frac{758}{7}}{\sqrt{4 \cdot \frac{170896}{49}}} = \frac{108,2857}{118,113} = 0,91679 \cong 0,92$$

La relation entre X et Y est bonne car $0,8 \leq 0,92 < 1$

4. Déterminons par la moindre carrée, l'équation de la droite de régression de Y en X

La droite (D) de regression de Y en X a pour équation :

$$Y - \bar{Y} = \frac{\text{cov}(X; Y)}{V(X)} (X - \bar{X})$$

$$Y - \frac{1495}{7} = \frac{\frac{758}{7}}{4} (X - 4)$$

$$Y = \frac{758}{28} (X - 4) + \frac{1495}{7}$$

$$Y = \frac{379}{14} X + \left(\frac{-758 + 1495}{7} \right)$$

$$Y = \frac{379}{14} X + \frac{737}{7}$$

Pour $Y = 350$; on aura :

$$350 = \frac{379}{14} X + \frac{737}{7} \Rightarrow X = \frac{350 - \frac{737}{7}}{\frac{379}{14}} = \frac{3476}{379} = 9,03 \cong 9$$

La coopérative produira 350 en 2003/au bout de 9 ans.

Exercice4 :

Le tableau ci-dessous donne les notes sur 20 obtenues en mathématiques et en sciences physiques par 8 candidats de la série TSS au baccalauréat malien en 2014.

X_i est la note de mathématiques et Y_i la note sciences physiques

X_i	4	6	7	9	11	14	12	17
Y_i	4	4	6	8	10	12	9	11

1. Représenter graphiquement le nuage de points de cette série statistique dans le plan muni d'un repère orthonormé $(O; \vec{i}; \vec{j})$. L'unité graphique est 7 cm.
2. Calculer les coordonnées point moyen G du nuage des points puis placer dans le repère.
3. a. Vérifier que la covariance $cov(X; Y)$ de la série statistique est égale à $\frac{87}{4}$
b. Calculer le coefficient de corrélation linéaire entre X et Y .

ADD