I/Les tableaux de contingence :

$X \backslash y$	y_1	y_2	...	y_p	total ($n_{i.}$)
x_1	n_{11}	n_{12}	...	n_{1p}	$n_{1.}$
x_2	n_{21}	n_{22}	...	n_{2p}	$n_{2.}$
⋮	⋮	⋮		⋮	⋮
x_k	n_{k1}	n_{k2}	...	n_{kp}	$n_{k.}$
Total $n_{.j}$	$n_{.1}$	$n_{.2}$	...	$n_{.p}$	$n_{..}$=N

Distribution marginale de x :

x_i	x_1	x_2	...	x_k	total
$n_{i.}$	$n_{1.}$	$n_{2.}$	...	$n_{k.}$	$n_{..} = N$

Distribution marginale de y :

y_j	y_1	y_2	...	y_p	total
$n_{.j}$	$n_{.1}$	$n_{.2}$	...	$n_{.p}$	$n_{..} = N$

C/ Distributions conditionnelles :

La distribution conditionnelle correspondant à une modalité x_i de la variable x suivant les modalités de y est appelée distribution conditionnelle de y pour $= x_i$

.

$y / x = x_i$	y_1	y_2	...	y_p	total
n_{ij}	n_{i1}	n_{i2}	...	n_{ip}	$n_{i.}$

distribution conditionnelle de x pour $y = y_j$

$x / y = y_j$	x_1	x_2	...	x_k	total
n_{ij}	n_{1j}	n_{2j}	...	n_{kj}	$n_{.j}$

Application :Le tableau suivant donne la répartition de 1000 familles selon l'âge du père (X_i) et le nombre d'enfants (Yj)

Xi \ Yj	Moins de 2 enfants	[2-5[	5 et plus	**Totaux ($n_{i.}$)**
Moins de 25 ans	100	20	5	125
[25 -30[	50	25	15	90
[30-40[	30	100	100	230
40 et plus	20	200	335	555
Totaux ($n_{.j}$)	200	345	455	1000

Distribution marginale de X :

X âge du père	Moins de 25 ans	[25-30[	[30-40[	40 et plus	Total
$n_{i.}$	125	90	230	555	1000

Distribution marginale de Y :

Y nombre d'enfants	Moins de 2 enfants	[2-5[	5 et plus	Total
$n_{.j}$	200	345	455	1000

Distribution conditionnelle de Y selon X∈ [30 − 40[:

Y/X∈ [30-40[	Moins de 2 enfants	[2-5[	5 et plus	Total
n_{3j}	30	100	100	230 ($n_{3.}$)

Distribution conditionnelle de X selon Y∈ [2-5[:

X/Y∈ [2-5[	Moins de 25 ans	[25-30[	[30-40[	[40 et plus	Total
n_{i2}	20	25	100	200	345 ($n_{.2}$)

Fréquences relatives partielles sur l'effectif total :

$f_{ij} = {}^{n_{ij}}/_{n_{..}}$ et $\sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij} = 1$

Fréquences relatives marginales

Pour la distribution marginale de x: $f_{i.} = \frac{n_{i.}}{n_{..}}$

Pour la distribution marginale de y : $f_{.j} = \frac{n_{.j}}{n_{..}}$

Et $\sum_{i=1}^{k} f_{i.} = 1 = \sum_{j=1}^{p} f_{.j}$.

Fréquences relatives conditionnelles

On a p fréquences relatives conditionnelles de x selon y puisque j varie de 1 jusqu'à p :

f de i si j $\quad f_{i/j} = \frac{n_{ij}}{n_{.j}}$

On a k fréquences relatives conditionnelles de y selon x puisque i varie de 1 jusqu'à k : f de j si i $\quad f_{j/i} = \frac{n_{ij}}{n_{i.}}$

1/ Le nombre de familles ayant de 2 à 5 enfants et dont l'âge du père est compris entre 30 et 40 ans est égal à 100 familles (l'effectif partiel **n_{32}**).

2/ Le nombre de famille ayant moins de 2 enfants est égal à 200 (l'effectif marginal de Y : **$n_{.1}$**).

3/ Le nombre de familles dont l'âge du père est égal à 40 ans et plus est 555 (L'effectif marginal de X : **$n_{4.}$**)

4/ Le nombre 5 de la première ligne et de la troisième colonne (**n_{13}**) représente le nombre de famille ayant 5 enfants ou plus et dont l'âge du père est inférieur à 25 ans.

5/ **n_{11}**= 100 ; **n_{23}**= 15 ; **$n_{2.}$** = 90 ; **$n_{.3}$**= 455

Les fréquences relatives :

f_{33} $= \frac{n_{33}}{n_{..}} = \frac{100}{1000}$ = 0, 10 ou 10% (Fréquence partielle sur l'effectif total). Cela signifie qu'il y a 10% de familles ayant 5 enfants et plus et dont l'âge du père est compris entre 30 et 40.

$f_{2.}$ $= \frac{n_{2.}}{n_{..}} = \frac{90}{1000}$ = 0,09 ou 9% (Fréquence marginale de X). Cela signifie qu'il y a 9% de familles dont l'âge du père est compris entre 25 et 30 ans quel que soit le nombre d'enfants.

$f_{.3}$ $= \frac{n_{.3}}{n_{..}} = \frac{455}{1000}$ = 0,455 ou 45,5% (Fréquence marginale de Y). Cela veut dire qu'il y a 45,5% de familles ayant 5 enfants et plus quel que soit l'âge du père.

$f_{1/2}$ avec i fixé $= \frac{n_{21}}{n_{2.}} = \frac{50}{90}$ = 0,5555 ou 55,55% (la fréquence conditionnelle de Y avec j = 1 si i = 2. Cela signifie que parmi les familles dont l'âge du père est compris entre 25 et 30 ans, 55,55% ont moins de deux enfants.

$f_{3/2}$ avec j fixé $= \frac{n_{32}}{n_{.2}} = \frac{100}{345}$ = 0,2898 ou 28,98% (la fréquence conditionnelle de X avec i = 3 si j = 2). Cela veut dire que parmi les familles ayant de 2 à 5 enfants, 28,98% des pères ont l'âge compris entre 30 et 40 ans.

Les paramètres des lois marginales selon x

a) La moyenne marginale de x *est* $\bar{x}$. Elle est définie comme suit :

$$\bar{x} = \frac{1}{n_{..}} \sum_{i=1}^{k} n_{i.} x_i = \sum_{i=1}^{k} f_{i.} x_i$$

b) La variance marginale de x notée $V(x)$.

Formule de définition : $V(x) = \frac{1}{n_{..}} \sum_{i=1}^{k} n_{i.} (x_i - \bar{x})^2 = \sum_{i=1}^{k} f_{i.} (x_i - \bar{x})^2$

Formule développée : $V(x) = \frac{1}{n_{..}} \sum_{i=1}^{k} x_i^2 n_{i.} - \bar{x}^2 = \sum_{i=1}^{k} f_{i.} x_i^2 - \bar{x}^2$

Les paramètres des lois marginales de y :

a) La moyenne marginale $\bar{y}$: $\bar{y} = \frac{1}{n_{..}} \sum_{j=1}^{p} n_{.j} y_j = \sum_{j=1}^{p} f_{.j} y_j$

b) La variance $V(y)$:

Par définition : $V(y) = \frac{1}{n_{..}} \sum_{j=1}^{p} n_{.j} (y_j - \bar{y})^2 = \sum_{j=1}^{p} f_{.j} (y_j - \bar{y})^2$

Formule développée : $V(y) = \frac{1}{n_{..}} \sum_{j=1}^{p} (n_{.j} y_j^2) - \bar{y}^2 = \sum_{j=1}^{p} f_{.j} y_j^2 - \bar{y}^2$

On peut définir **la covariance(xy)** comme suit :

a) Formule de définition $Cov(xy) = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{p} [(x_i - \bar{x})(y_j - \bar{y})] n_{ij}$

$$= \sum_{i=1}^{k} \sum_{j=1}^{p} [(x_i - \bar{x})(y_j - \bar{y})] f_{ij}$$

b) Formule développée : $Cov(xy) = \left(\frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{p} n_{ij} x_i y_j\right) - \bar{x}\bar{y}$

$$= \left(\sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij} x_i y_j\right) - \bar{x}\bar{y}$$

Application : Le tableau suivant donne la répartition des dépenses mensuelles (10^3 DA), (notées Y_j), des employés d'une entreprise selon le nombre d'enfants (noté X_i)

Y_j / X_i	[0 ; 20[	[20 ; 40[	[40 ; 60[	[60 ; 80[	[80 ; 100[	Totaux
[0 ; 2[	10	6	4	2	0	22
[2 ; 4[	8	6	4	1	0	19
[4 ; 6[	1	2	6	4	3	16
[6; 8 [	0	1	2	4	6	13
[8 ; 10[	0	0	1	1	3	5
Totaux	19	15	17	12	12	75

1-Calculer la dépense moyenne.
2- Calculer la variance marginale de X.
3-Quelle est la dépense moyenne des employés ayant entre deux et quatre enfants ?
4-Quel est le nombre d'enfants moyen pour les salariés qui dépensent entre 40 000 DA et 60 000 DA ?

1/ La dépense moyenne $\bar{Y}$

Yj (dépenses)	[0-20[	[20-40[	[40-60[	[60-80[	[80-100[	Total
n.j	19	15	17	12	12	75
Yj	10	30	50	70	90	
n.j×Yj	190	450	850	840	1080	3410

$\bar{Y} = \frac{\sum Yj \times n.j}{n_{..}} = \frac{\sum Yj \times n.j}{N} \Longrightarrow \bar{Y} = \frac{3410}{75} = 45{,}46 \times 10^3$ DA.

2/ La variance marginale du caractère X : $V(x) = \frac{\sum X_i^2 n_{i.}}{N} - \bar{X}^2$

Xi	[0-2[	[2-4[	[4-6[	[6-8[	[8-10[	Total
ni.	22	19	16	13	05	75
Xi	1	3	5	7	9	
Xi×ni.	22	57	80	91	45	295
X_i^2	1	9	25	49	81	
X_i^2 **×ni.**	22	171	400	637	405	1635

Paramètres des distributions conditionnelles de x selon y

a) Les moyennes conditionnelles de x selon y, $y = y_j$ (y_j fixe)

$$\bar{x}_j = \frac{1}{n_{.j}} \sum_{j=1}^{k} n_{ij} x_i = \sum_{i=1}^{k} f_{i/j}\, x_i$$

b) Les variances conditionnelles de x selon y ($y = y_j$)

Par définition : $V_j(x) = \frac{1}{n_{.j}} \sum_{i=1}^{k} \left[(x_i - \bar{x}_j)^2 n_{ij} \right]$ Ou= $\sum_{i=1}^{k} (x_i - \bar{x}_j)^2 f_{i/j}$

Formule développée: $V_j(x) = \frac{1}{n_{.j}} \sum_{i=1}^{k} n_{ij} x_i^2 - \bar{x}_j^2 = \sum_{i=1}^{k} x_i^2 f_{i/j} - \bar{x}_j^2$

Paramètres des distributions conditionnelles de y selon x

a. Les moyennes conditionnelles de y selon x

$$\bar{y}_i = \frac{1}{n_{i.}} \left[\sum_{j=1}^{P} n_{ij} y_j \right] = \sum_{j=1}^{P} f_{j/i} y_j$$

b. Les variances de y selon x

Par définition $V_i(y) = \frac{1}{n_{i.}} \sum_{j=1}^{P} \left[(y_j - \bar{y}_j)^2 n_{ij} \right]$

ou $= \sum_{j=1}^{P} (y_j - \bar{y}_j)^2 f_{j/i}$

Formule développée : $V_i(y) = \frac{1}{n_{i.}} \sum_{j=1}^{P} (n_{ij} y_j^2) - \bar{y}_j^2$

Ou $= \sum_{j=1}^{P} (f_{j/i} y_j^2) - \bar{y}_j^2$

On calcule d'abord la moyenne $\bar{X}$: $\bar{X} = \frac{\sum X i n i.}{n..} = \frac{295}{75} = 3,93$ enfants.

$V(x) = \frac{1635}{75} - (3,93)^2 = 6,36$ enfants.

3/La dépense moyenne des employés ayant entre 2 et 4 enfants : il s'agit de calculer la moyenne conditionnelle $\bar{Y}/ X \in$ [2-4[:

Y/X∈ [2-4[	[0-20[	[20-40[	[40-60[	[60-80[	[80-100[	Total
n_{2j}	8	6	4	1	0	19
Yj	10	30	50	70	90	
$n_{2j} \times Y_j$	80	180	200	70	0	530

$\bar{Y}/ X \in$ [2-4[$= \frac{\sum n_{2j} \times Yj}{\sum n_{2j}} = \frac{530}{19} = 27,89.10^3$DA.

4/ Le nombre d'enfants moyen pour les salariés qui dépensent entre 40000 DA et 60000DA : il s'agit de calculer la moyenne conditionnelle $\bar{X}$/Y∈ [40-60[:

X/ Y ∈ [40-60[	[0-2[	[2-4[	[4-6[	[6-8[	[8-10[	Total
n_{i3}	4	4	6	2	1	17
X_i	1	3	5	7	9	
$n_{i3} \times X_i$	4	12	30	14	9	69

$\bar{X}$/Y∈ [40-60[$= \frac{\sum n_{i3} \times X_i}{\sum n_{i3}} = \frac{69}{17} = 4,06$ enfants.

II/ AJUSTEMENT, REGRESSION ET CORRELATION

On s'interroge sur la relation qui peut exister entre deux grandeurs. Trois types de problèmes peuvent apparaitre :

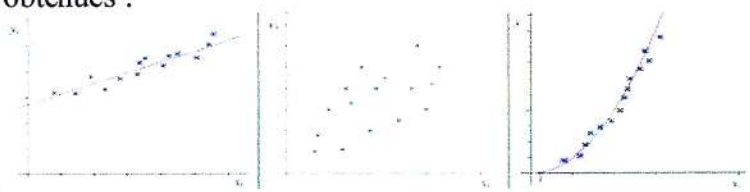
- ✓ Problème d'ajustement analytique.
- ✓ Analyse de la régression
- ✓ Problème de corrélation.

Trois types de liaisons entre les caractères x et y

- ✓ *Indépendance totale (absence de liaison)*
- ✓ *Liaison fonctionnelle ou dépendance totale*
- ✓ *La liaison relative*

L'ajustement graphique

D'abord, on porte nos données dans un graphe appelé: **nuage de points.** Ensuite, à main levée, nous traçons une courbe qui passe au plus près de l'ensemble des points. Plusieurs formes peuvent être obtenues :



Si le nuage de points forme une droite comme dans le premier graphe, on parle d'une liaison linéaire entre les deux variables.

Ajustement mécanique : Dans ce cas, deux méthodes peuvent être utilisées.

- ✓ **Méthode des moyennes échelonnées :** elle consiste à diviser la série statistique en plusieurs groupes, pour chaque groupe on calcule la Médiane (Me) pour les valeurs de la variable x et la Moyenne arithmétique ($\bar{Y}$) pour les valeurs de la variable y.

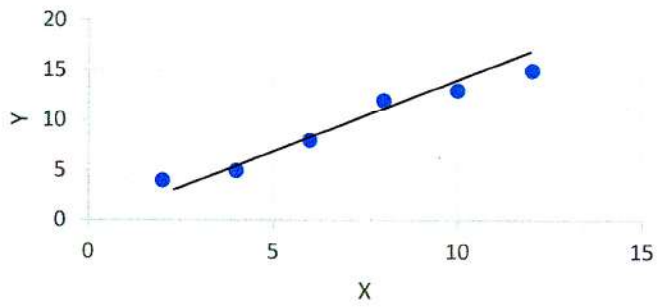
Exemple : soit la série bi-variée suivante :

X	2	4	6	8	10	12
Y	4	5	8	12	13	15

Question : Déterminer l'ensemble des points correspondant aux couples (x_i ; y_i) par la méthode de moyennes échelonnées (ordre 3).

Solution :

- ✓ On forme des sous-ensembles composés de 3 valeurschacun.
- ✓ On calcule les Médianes pour les sous-ensembles de la variable x_i
  - Pour le premier groupe de valeurs x_i ; c'est-à-dire (2 ;4 ;6) : M_e= 4.
  - Pour le deuxième groupe de valeurs x_i ; c'est-à-dire (8 ;10 ;12) : M_e= 10.
- ▪ On calcule les moyennes arithmétiques pour les sous-ensembles de la variable yi. :
  - Pour le premier groupe de valeurs y_i ; c'est-à-dire (4 ; 5 ; 8) : $\bar{Y}$= 5,66.
  - Pour le deuxième groupe de valeurs y_i; c'est-à-dire (12 ;13,15): $\bar{Y}$= 13,33.
- ▪ On déduit alors les coordonnés des deux points déjà calculés : P_1 (4 ;5,66) et P_2(10 ;13,33).



- ✓ **La méthode des moyennes mobiles :** le principe de calculressemble à celui des moyennes échelonnées (Médiane pour les x_i et moyenne arithmétique pour les y_i). La différence se situe dans la formation des sous-ensembles qui ne sont pas strictementdistinctsles uns des autres. Autrement dit, les valeurs se répètent dans plusieurs sous-ensembles.

Exemple : Soit la série bi-variée suivante :

X_i	2	4	6	8	10	12	14	16	18
Y_i	4	5	8	12	13	15	18	21	24

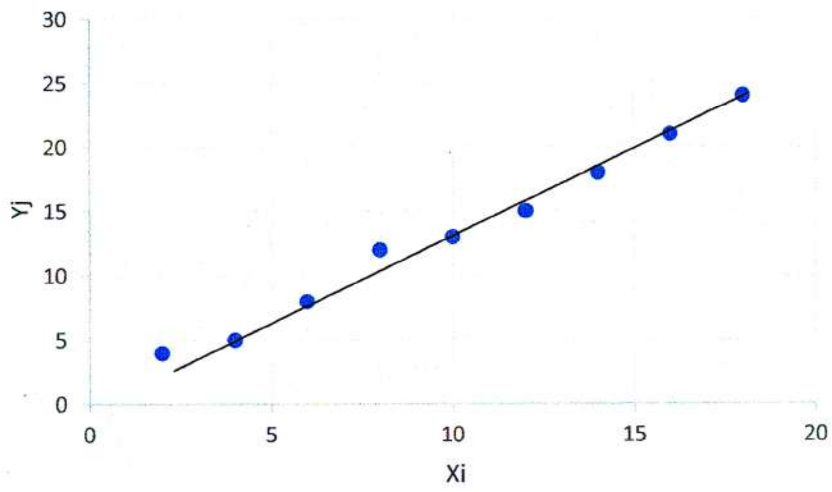
Question : Déterminer l'ensemble des points correspondant aux couples (x_i ; y_i) par la méthode de moyennes mobiles (ordre 3).

Solution :

- ✓ On forme des sous-ensembles composés de 3 valeurs chacun. On calcule, alors les médianes pour les sous-ensembles de la variable x_i et les moyennes arithmétiques pour les sous-ensembles de la variable yi. Cela nous permet de déduire les coordonnés des points correspondant aux couples (x_i ; y_i).

Sous-ensembles x_i	Sous-ensembles y_i	Coordonnées (x_i ; y_i)
- 2 ; 4 ;6 M_e=4	- 4 ;5 ;8 $\bar{Y}$=5,66	(4 ; 5,66)
- 4 ; 6 ;8 M_e=6	- 5 ;8 ;12 $\bar{Y}$ = 8,33	(6 ; 8,33)
- 6 ; 8 ;10 M_e=8	- 8 ;12 ;13 $\bar{Y}$ = 11	(8 ;11)
- 8 ;10 ;12 M_e=10	- 12 ;13 ;15 $\bar{Y}$ = 13,33	(10 ;13,33)
- 10 ;12 ;14 M_e=12	- 13 ;15 ;18 $\bar{Y}$ = 15,33	(12 ; 15,33)
- 12 ;14 ;16 M_e=14	- 15 ;18 ;21 $\bar{Y}$ = 18	(14 ; 18)
- 14 ;16 ;18 M_e=16	- 18 ;21 ;24 $\bar{Y}$ = 21	(16 ;21)

- On réalise, ensuite, la représentation graphique qui reprend les données du tableau (nuage des points) sur lequel nous traçons une droite qui passe par les points moyens précédemment calculés.



AJUSTEMENT ANALYTIQUE (DROITE DE REGRESSION)

On désire ici déterminer et tracer une droite qui représente au mieux la relation de dépendance de Y par rapport à X.

L'équation de cette droite est du type Y=a X + b.

Avec - "a" comme coefficient directeur de la droite.
- "b" ordonnée à l'origine.

Plusieurs méthodes de détermination sont possibles, mais, la plus utilisée **est la méthode des moindres carrés.**

La droite **Y= *a* X + *b***est la droite d'ajustement de Y en fonction de X.

On peut également chercher à exprimer X en fonction de Y. On cherche alors la droite d'ajustement de X en Y d'équation **X= *á*Y+*b́*.**

1- Droite d'ajustement de Y en fonction de X

La méthode des moindres carrés repose sur le principe de la minimisation des écarts entre les points observés et les points de la droite,

$$a = \frac{cov(xy)}{v(x)}$$

Si les données ne sont pas pondérées, on calcule la covariance et la variance de la manière suivante :

$$\text{Cov}(xy) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{N} \text{ (formule de définition)}$$

$$\text{ou } \text{Cov}(xy) = \frac{\sum_{i=1}^{n}x_iy_i}{N} - \bar{x}\bar{y} \text{ (formule développée).}$$

La variance $v(x) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{N}$ (formule de définition)

ou $V(x)= \frac{\sum_{i=1}^{n}x_i^2}{N} - \bar{x}^2$ (formule développée).

En simplifiant, on peut calculer le « a » comme suit : $a = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$

ou encore $a = \frac{\sum_{i=1}^{n}x_i\,y_i - N\bar{x}\,\bar{y}}{\sum_{i=1}^{n}(x_i)^2 - N\bar{x}^2}$ Par ailleurs $b = \bar{y} - a\bar{x}$

2- Droite d'ajustement de X en fonction de Y

On peut trouver l'équation de la droite de x en y, c'est- à-dire x devient la variable dépendante ou expliquée et y devient la variable indépendante. l'équation s'écrit X= *á* Y+*b́*

$$\grave{a} = \frac{cov(xy)}{v(y)},$$

$v(y) = \frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{N}$ (formule de définition) ou $v(y) = \frac{\sum_{i=1}^{n}y_i^2}{N} - \bar{y}^2$ (formule développée).

En simplifiant, on calcule le coefficient *á* comme suit :

$$\acute{a} = \frac{\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \quad ou \quad a' = \frac{\sum_{i=1}^{n}x_i\,y_i - N\bar{x}\,\bar{y}}{\sum_{i=1}^{n}(y_i)^2 - N\bar{y}^2} \quad et \quad \acute{b} = \bar{x} - \acute{a}\bar{y}$$

LA CORRELATION

Pour mesurer l'intensité de la relation entre deux variables x et y, nous utilisons un indicateur appelé coefficient de corrélation.

Le coefficient de corrélation linéaire *r* se calcule par la formule :

$r = \sqrt{a.\acute{a}}$

Ou encore par la formule

$r = \frac{Cov\,(x,y)}{\sqrt{V(x).V(y)}}$ ou $r = \frac{Cov\,(x,y)}{\sigma(X).\sigma(Y)}$ ou $r = a \times \frac{\sigma(X)}{\sigma(Y)}$

Autrement dit, $r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 . \sum_{i=1}^{n}(y_i - \bar{y})^2}}$

Ou $r = \frac{\sum_{i=1}^{n} x_i y_i - N\bar{x}\bar{y}}{\sqrt{\sum x_i^2 - N\bar{x}^2}.\sqrt{\sum y_i^2 - N\bar{y}^2}}$

Le coefficient de corrélation varie entre -1 et +1.

-Si r =0, il y a absence de corrélation entre x et y.

-Si r = +1 ou -1, il y a une corrélation maximale entre x et y,

-Si r est proche de +1 ou de -1 : très forte corrélation linéaire.

-Si r est proche de zéro, : une faible corrélation linéaire.

Le signe positif (+) signifie que les deux variables varient dans le même sens.

Le signe négatif (-) signifie que les deux variables varient en sens inverse.

Nous pouvons calculer **le coefficient de détermination r^2**, il exprime le pourcentage de variation de la variable y expliquée par la variable x.

$r^2 = a \times \acute{a}$

Application : Soit la série bi-variée suivante où X représente les résultats au test (noté sur 10) de six (6) employés et Y les rendements (en douzaine d'unités).

X_i	2	3	5	7	9	10
Y_i	1	3	7	11	15	17

1/ Représenter le nuage de points.

2/ Trouver l'équation de la droite de régression de Yen X par la méthode des moindres carrés.

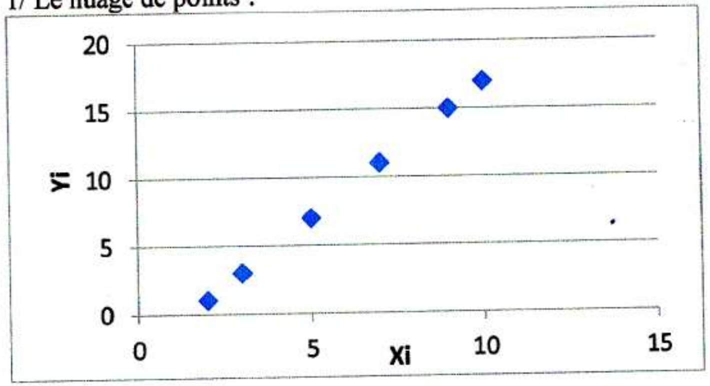
3/ Trouver l'équation de la droite de régression de X en Y.

4/ Calculer les coefficients de corrélation et de détermination.

5/ Estimer le rendement d'un employé ayant obtenu un résultat de 4 sur 10.

Solution :

1/ Le nuage de points :



2/ L'équation de la droite de régression de Yen X : Y = aX+b

X_i	Y_i	X_iY_i	X_i^2	Y_i^2
2	1	2	4	1
3	3	9	9	3
5	7	35	25	49
7	11	77	49	121
9	15	135	81	225
10	17	170	100	289
$\sum = 36$	$\sum = 54$	428	268	694

$a = \frac{cov\,(xy)}{v\,(x)}$ avec Cov (xy) $= \frac{\sum X_i Y_i}{N} - \bar{X}\bar{Y}$ et V(x) $= \frac{\sum x_i^2}{N} - \bar{X}^2$

Calculons d'abord les moyennes marginales : $\bar{X}$ et $\bar{Y}$

$\bar{X} = \frac{\sum X_i}{N} = \frac{36}{6} = 6$ et $\bar{Y} = \frac{\sum Y_i}{N} = \frac{54}{6} = 9$

Cov (xy)$= \frac{428}{6} - (6)\times(9) = 17{,}33$ et V(x) $= \frac{268}{6} - (6)^2 = 8{,}66$

$a = \frac{17{,}33}{8{,}66} = 2$. On trouve le coefficient b comme suit : on a Y= aX+b comme la droite d'ajustement passe par le point moyen ($\bar{X}, \bar{Y}$) $\Longrightarrow \bar{Y} = a\bar{X} + b \Longrightarrow$ $b = \bar{Y} - a\bar{X}$

b= 9− (2) ×(6) = −3

L'équation est Y= 2X−3 (on peut la représenter sur le nuage de points)

3/ La droite de régression de X en Y : X= $\acute{a}$Y +$\acute{b}$ avec $\acute{a} = \frac{cov\,(xy)}{v(y)}$, calculons la variance de Y : v(y) $= \frac{\sum Y_i^2}{N} - \bar{Y}^2 = \frac{694}{6} - (9)^2 = 34{,}66$

$\acute{a} = \frac{17{,}33}{34{,}66} = 0{,}5$ et $\acute{b} = \bar{X} - \acute{a}\bar{Y}$; $\acute{b} = 6 - 0{,}5(9) = 1{,}5$

L'équation est X= 0,5 Y+1,5

4/ Coefficients de corrélation (r) et de détermination r^2 :

$$r = \frac{cov\,(xy)}{\sigma(x)\times\sigma(y)} = \frac{17{,}33}{\sqrt{v(X)}\sqrt{v(Y)}} = \frac{17{,}33}{\sqrt{8{,}66}\sqrt{34{,}66}} = \frac{17{,}33}{2{,}943\times 5{,}887} \cong 1 \text{ ou } r = \sqrt{a\acute{a}} = \sqrt{2\times 0{,}5} = 1$$

r est égal à 1, il y a une corrélation maximale entre les résultats du test et le rendement des employés.

$r^2 = (1)^2 = 1$ ou 100%. Cela signifie que le rendement des employés est expliqué totalement (à 100%) par les résultats du test.

5/ Si X=4 ; Y= ?, nous avons Y=2X−3 donc Y= 2(4) −3 = 5.