Cours n° 5 : Corrélation et la Régression

- **Corrélation et régression linéaire : (Liaison entre deux variables)**

On constate très souvent dans la pratique qu'il existe une liaison entre deux (ou plusieurs) variables. Par exemple : la relation entre le poids et la taille des hommes adultes sont liés d'une certaine façon ; entre le revenus et les dépenses en nourriture.

Pour étudier la liaison entre deux variables quantitatives (discrètes), on commence par faire un graphique du type nuage de points. La forme générale de ce graphique indique s'il existe ou non une liaison entre les deux variables.

Pour préciser les choses, on calcule ensuite un indicateur de liaison. Pour cela, il faut d'abord introduire **la covariance**, généralisation bidimensionnelle de la variance. Comme elle dépend des unités de mesure des deux variables considérées, on doit la rendre intrinsèque en la divisant par le produit des **écarts-types**. On définit ainsi **le coefficient de corrélation linéaire**, indicateur de liaison chercher. Il est toujours compris entre 1 et +1, son signe indique le sens de la liaison, tandis que sa valeur absolue en indique l'intensité.

En complément, on explique ce qu'est la régression linéaire d'une variable sur une autre. Lorsqu'il existe une liaison causale entre les deux variables considérées, la régression linéaire permet d'approcher la variable réponse par une fonction de la variable causale.

- **Les relations entre deux variables quantitatives:**

Tout comme pour les variables quantitatives, on cherche á déterminer si il ya une relation de dépendance entre deux variables quantitatives, Xi et Yi. si on démontre qu´il existe un lien, on dit qu´il ya une corrélation entre Xi et Yi (corrélation ,lien de dépendance entre deux variables quantitatives)

- **Diagramme de dispersons ou nuage de points:**

On se sert notamment d´un graphique appelé nuage de point pour présenter une relation entre deux variables quantitatives.(Diagramme de dispersion ou nuage de points: ensemble de points représentants des couples de valeurs dans un système de coordonnées cartésiennes)

Pour étudier les relations ou corrélations entre deux variables statistiques, on peut les porter sur un graphique.

Exemple:

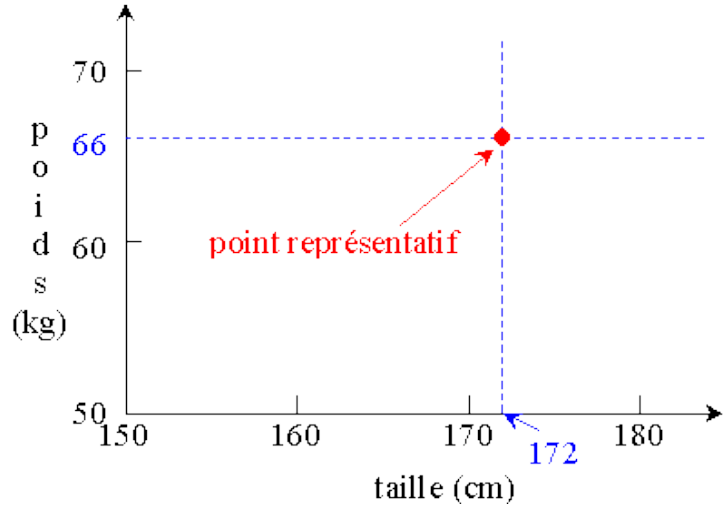
Relation entre la taille et le poids des individus.

Pour chaque individu de l'échantillon, on porte sur un graphique:

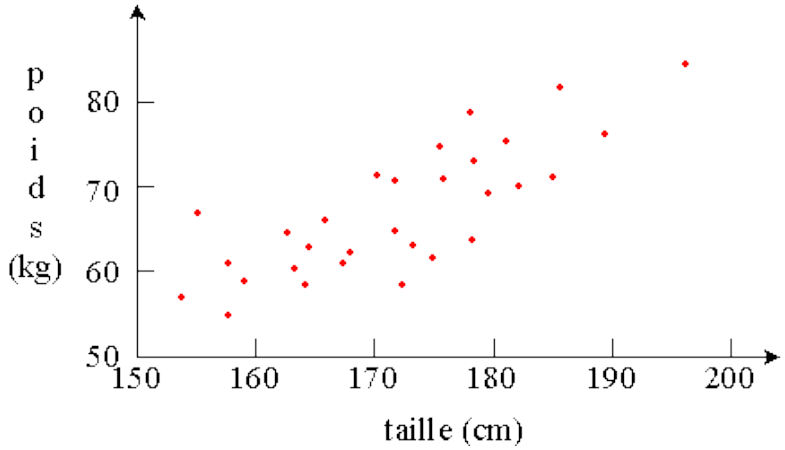
- sa taille en abscisse.
- son poids en ordonnée.

chaque individu est donc, dans ce graphique, représenté par un point (point représentatif).

Soit un individu mesurant 172 cm et pesant 66 kg:



Dans le graphe, il y aura donc autant de points qu'il y a d'individus dans l'échantillon.



➢ **Les nuages de points présente les caractéristiques suivantes:**

- ✓ il comporte autant de points qu´il ya d´unité statistique dans l´échantillon ou la population étudiée.
- ✓ L´axe des abscisses est réservé á la variable indépendante Xi. (La taille)
- ✓ L'axe des ordonnées est réservée á la variable dépendante Yi. (Le poids)

On peut (par la pensée ou réellement) tracer une droite qui passe au mieux par ces points (au milieu du "nuage" de points).

- Si cette droite **"monte",** on dira qu'il y a **corrélation positive** entre les deux variables.
- Si elle **"descend"**, c'est une **corrélation négative.**
- Si elle est **"horizontale",** ou si on ne peut pas décider, c'est qu'il y a **absence de corrélation**.

1) Détermination de coefficient de corrélation r de Person:

Dans le cas ou le nuage de points qui permet de croire à l'existence d'une corrélation entre les deux variables x et y prend une forme allongée telle que les points qui le constituent paraissent s'être regroupés au voisinage d'une droite.

Pour le calcul du coefficient de corrélation noté ***r***, nous utilisons la formule la plus simple et la plus pratique:

Un coefficient de corrélation linéaire, que nous désignerons par ***r***, peut être calculé comme suit :

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \times \sum(y_i - \bar{y})^2}} \quad \textbf{- Et : } (-1) \leq r \leq +1$$

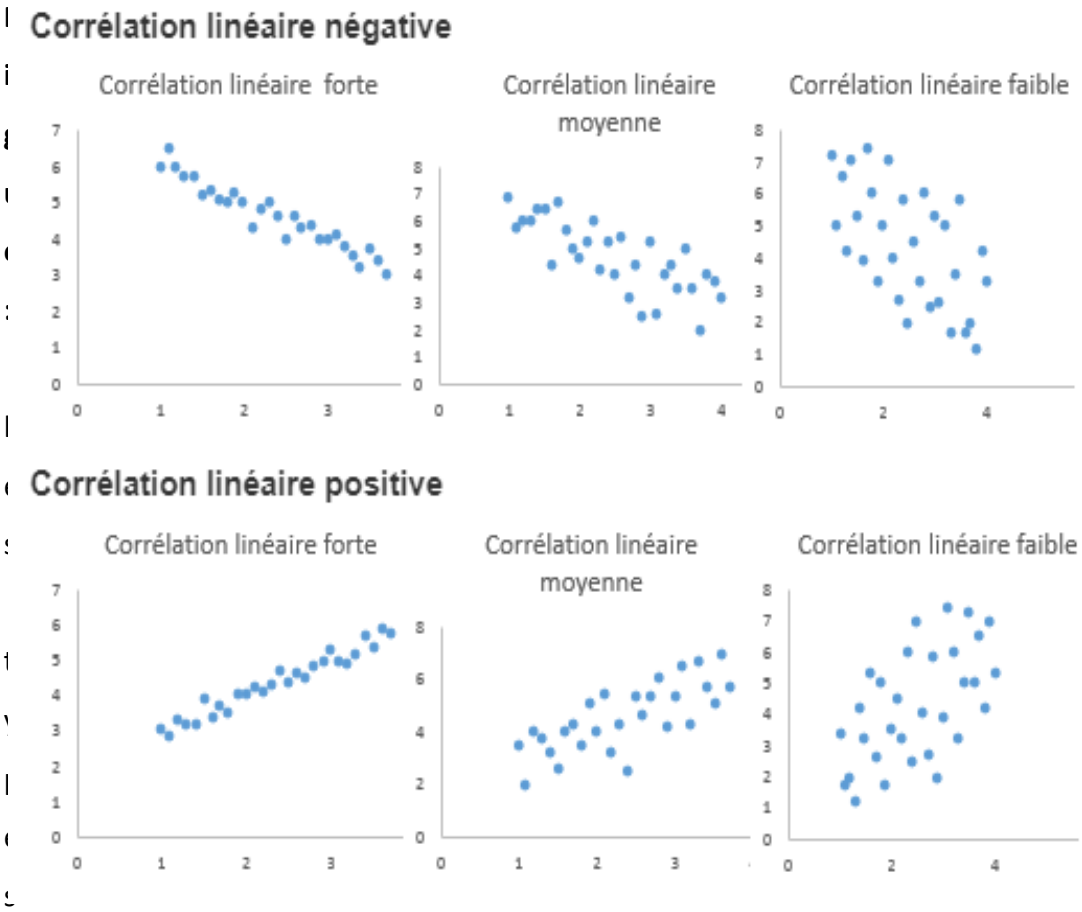
Où :

- $\bar{X}$: la moyenne de la première distribution
- $\bar{Y}$: la moyenne de la deuxième distribution
- Le signe de coefficient de corrélation r indique le sens positif ou négatif de la corrélation.
- La valeur absolue de r est inférieure ou égale à **1,** est la corrélation est d'autant plus serrée que cette valeur absolue est plus voisine de **1.**

Critères servant á interpréter les valeurs, qui seront utilisées pour qualifier la corrélation linéaire:

Valeur de r	Force du lien linéaire
Près de 0	Nulle
0< r ≤ 0.2	Faible
0.2< r≤ 0.5	Moyenne
0.5< r ≤ 0.8	Forte
r >0.8	Très forte
1	Parfaite

- **Le sens d´une correlation**



de corrélation entre deux variables.

- ✓ Une relation linéaire est dite **positive**,(ou direct) si la droite de régression qui la présente est croissante c.à.d. qu´elle s´élève de gauche a droite .Si Xi augmente alors Yi augmente aussi.
- ✓ Une relation linéaire est dite **négative**, (ou inverse)si la droite de régression qui la présente est décroissante c.à.d. qu´elle descente de gauche a droite. Si Xi augmente alors Yi diminue.
- ✓ Une droite de régression est dite **horizontale** (ou presque horizontale)indique l´absence de relation linéaire entre deux variables.(Xi et Yi)

Dans le cas ou le nuage de points qui permet de croire à l'existence d'une corrélation entre les deux variables x et y prend une forme allongée telle que les points qui le constituent paraissent s'être regroupés au voisinage d'une droite.

Exemple :

x_i	16	18	23	24	28	29	26	31	32	34
y_i	20	24	28	22	32	28	32	36	41	41

La représentation graphique a montrée que l'hypothèse d'une corrélation linéaire positive pouvait être retenue (nuage aplati, allongé, de forme linéaire, de pente positive)

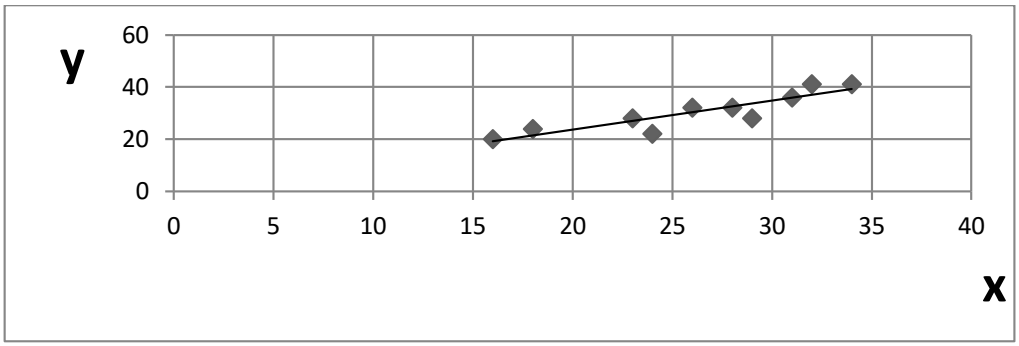


Tableau:

x_i	y_i	$x_i - \overline{x}$	$y_i - \overline{y}$	$(x_i - \overline{x})^2$	$(y_i - \overline{y})^2$	$(x_i - \overline{x})(y_i - \overline{y})$
16	20	-10,1	-10,4	102,01	108,16	105,04
18	24	-8,1	-6,4	65,61	40,96	51,84
23	28	-3,1	-2,4	9,61	5,76	7,44
24	22	-2,1	-8,4	4,41	70,56	17,64
28	32	1,9	1,6	3,61	2,56	3,04
29	28	2,9	-2,4	8,41	5,76	-6,96
26	32	-0,1	1,6	0,01	2,56	-0,16
31	36	4,9	5,6	24,01	31,36	27,44
32	41	5,9	10,6	34,81	112,36	62,54
34	41	7,9	10,6	62,41	112,36	83,74
261	304	0	0	314,9	492,4	365,84

- ✓ La moyenne arithmétique des x ; $\bar{x} = \frac{\sum x_i}{N} = \frac{261}{10} = 26{,}1$
- ✓ La moyenne arithmétique des y ; $\bar{y} = \frac{\sum y_i}{N} = \frac{304}{10} = 30{,}4$

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \times \sum(y_i - \bar{y})^2}} = \frac{\mathbf{351,60}}{\sqrt{\mathbf{314,90 \times 492,40}}}$$

$$r = \frac{\mathbf{351,60}}{\sqrt{\mathbf{155\,056,76}}} = \frac{\mathbf{351,60}}{\mathbf{393,77}} = +\mathbf{0,89}$$

Corrélation positive forte, le coefficient **r** ayant une valeur absolue voisine de 1 .

- ✓ **Droite de régression :**

Lorsque de variables sont en corrélation linéaire, il arrive que l'on puisse admettre que les variations de l'une des deux variables sont causes des variations de l'autre.

Il sera alors légitime d'essayer d'exprimer par une fonction linéaire les valeurs de la variable considérée comme conséquence, à partir des valeurs de la variable considérée comme cause.

Si la variable fonction (conséquence) est désignée par ***y***, la variable cause étant désignée par ***x***, il sera donc normal de rechercher une fonction ***y = ax + b*** qui permettra d'ajuster la variable ***y*** à partir de la variable ***x***.

La droite dont il est question est dite droite de régression, ou droite d'estimation de ***y*** à partir de ***x***.

Il suffit simplement de substituer aux deux colonnes ; « variable x_i » ; « effectif y_i » du tableau habituel le tableau à deux colonnes « variable x_i », « variable y_i ».

Les deux paramètres $a$et **b** sont de la fonction cherchée sont donnés par la formule suivante :

$$a = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \textbf{ Et b} = \bar{y} - a\bar{x}$$

Exemple : Reprenons le même exemple précédent.

x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
16	20	-10,1	-10,4	102,01	108,16	105,04
18	24	-8,1	-6,4	65,61	40,96	51,84
23	28	-3,1	-2,4	9,61	5,76	7,44
24	22	-2,1	-8,4	4,41	70,56	17,64
28	32	1,9	1,6	3,61	2,56	3,04
29	28	2,9	-2,4	8,41	5,76	-6,96
26	32	-0,1	1,6	0,01	2,56	-0,16
31	36	4,9	5,6	24,01	31,36	27,44
32	41	5,9	10,6	34,81	112,36	62,54
34	41	7,9	10,6	62,41	112,36	83,74
261	304	0	0	314,9	492,4	365.84

$\boldsymbol{a = \frac{365{,}84}{314{,}9} = 1{,}161}$ **Et b** $\boldsymbol{= 30{,}4 - (1{,}161) \times (26{,}1) = 0{,}10}$

Equation de régression **y** $\boldsymbol{= 1{,}161x + 0{,}10}$

Par analogie on pourra chercher l'équation de la droite de régression de x par rapport à y, équation de la forme **:** $\boldsymbol{x = a'y + b'}$ **avec :** $\boldsymbol{a' = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(y_i - \bar{y})^2}}$ **et b'** $\boldsymbol{= \bar{x} - a'\bar{y}}$

Dans le même exemple :$\boldsymbol{a' = \frac{365{,}84}{492{,}40} = 0{,}742}$ $\boldsymbol{et\ b' = 26{,}1 - (0{,}742 \times 30{,}4) = 3{,}55}$

Equation de régression de x par rapport à y :

$$\boldsymbol{x = 0{,}742y + 3{,}55}$$

Exercice:

La société anonyme par action R augmente son capital. On a relevé pendant 10 jours le cours en bourse de l'action de celui du droit de souscription.

x_i	98	94	97	98	100	102	102	104	104	101
y_i	6.5	5.4	6.1	6.4	6.9	8	7.5	7.5	7.4	7.3

- ✓ Calculer le coefficient de corrélation linéaire entre les deux variables x et y.
- ✓ Donner l'équation de la droite de régression qui permet d'estimer le cours du droit de souscription à partir du cours d'action.

Solution :

x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
98	6,5	-2	-0,4	4	0,16	0,8
94	5,4	-6	-1,5	36	2,25	9
97	6,1	-3	-0,8	9	0,64	2,4
98	6,4	-2	-0,5	4	0,25	1
100	6,9	0	0	0	0	0
102	8	2	1,1	4	1,21	2,2
102	7,5	2	0,6	4	0,36	1,2
104	7,5	4	0,6	16	0,36	2,4
104	7,4	4	0,5	16	0,25	2
101	7,3	1	0,4	1	0,16	0,4
1000	69	0	0	94	5,64	21,4

- ✓ La moyenne arithmétique des x ; $\bar{x} = \frac{\sum x_i}{N} = \frac{1000}{10} = \mathbf{100}$
- ✓ La moyenne arithmétique des y ; $\overline{y} = \frac{\sum y_i}{N} = \frac{69}{10} = \mathbf{6.9}$

- $\boldsymbol{r} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \times \sum(y_i - \bar{y})^2}} = \frac{\mathbf{21.4}}{\sqrt{\mathbf{94 \times 5.64}}} = \frac{\mathbf{21.4}}{\sqrt{\mathbf{530.16}}} = \mathbf{+0.93}$

Corrélation positive et assez serrée, le coefficient ***r*** ayant une valeur absolue voisine de 1.

Equation de régression : $\boldsymbol{y = ax + b}$

$\boldsymbol{a = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}}$ **Et b** $= \boldsymbol{\bar{y} - a\bar{x}}$

$\boldsymbol{a = \frac{21.4}{94} = 0.228}$ **Et b** $= \mathbf{6.9 - (0.228) \times (100) = -15.9}$

Equation de régression **y** $= \mathbf{0.228x + (-15.9) = 0.228x - 15.9}$

$$\boldsymbol{y = 0.228x - 15.9}$$

Cours n° 6

2) Détermination de coefficient de corrélation de Spearman(rho)

Le coefficient de corrélation des rangs de Charle Spearman est une mesure de corrélation non paramétrique, il sert á déterminer la relation qui existe entre deux séries de données.

Le coefficient de corrélation de rang (appelé coefficient de Spearman) examine aussi s'il existe une relation entre le rang des observations pour deux caractères X et Y, ce qui permet de détecter l'existence de relations (croissante ou décroissante), quelle que soit leur forme précise (linéaire, exponentielle, puissance, ...)

Le coefficient de corrélation des rangs de Spearman généralement désigné par(rs) ou (rho)

$$r_s = 1 - \frac{6 \sum(x_i - y_i)^2}{n(n^2 - 1)}$$

D´ou: di= Rxi-Ryi

- **Les étapes pour évaluer les rangs de Spearman:**

- ✓ On attribue á chaque valeur de **Xi** un rang **Rxi.**

- ✓ On attribue á chaque valeur de **yi** un rang **Ryi**
- ✓ On ordonne les valeurs selon un ordre croissant ou décroissant.
- ✓ On calcule la différences entre les deux variables **di= Rxi-Ryi.**

Exemple:

Le tableau suivant donne la répartition de 06 étudiants ,selon leur mentions á deux examens de fin d´année.

n	1	2	3	4	5	6
Xi	Faible	Excellent	Bien	T. Faible	Passable	T. Bien
Yi	Passable	T. Bien	Bien	Faible	T. Faible	Excellent

n	**Xi**	**yi**	**Rxi**	**Ryi**	**/Rxi-Ryi/**	**(Rxi-Ryi)²**
1	Faible	Passable	5	4	1	1
2	Excellent	T. Bien	1	2	1	1
3	Bien	Bien	3	3	0	0
4	T. Faible	Faible	6	5	1	1
5	Passable	T. Faible	4	6	2	4
6	T. Bien	Excellent	2	1	1	1
∑	**/**	**/**	**/**	**/**	**/**	**8**

-On ordonne les valeurs selon un ordre décroissant.(rang)

Rxi: (1) Excellent - (2) T. Bien -(3) Bien - (4) Passable- (5) Faible- (6) T. Faible.

Ryi: (1) Excellent - (2) T. Bien -(3) Bien - (4) Passable- (5) Faible- (6) T. Faible.

$$rs = 1 - \frac{6\sum(Xi-yi)^2}{n\,(n^2-1)}$$

$$rs = 1 - \frac{6 \times 8}{6\,(36-1)}$$

$$rs = 1 - \frac{48}{210}$$

rs= 1 - 0,22 = +**0,78**

rs = **0,78** → On déduit qu´il existe une corrélation positive forte entre les (2) variables .

Remarque:

Dans le cas ou les variables sont identique , on ordonne les rangs égale la moyenne arithmétique qui leur a effectuer au départ.

Exercice 1: Huit échantillons de l'huile d'olive ont été notes par deux gastronomes. On obtient les classements suivants:

Echantillons	A	B	C	D	E	F	G	H
1er gastronome **Xi**	6	17	14	15	8	9	7	16
2éme gastronome **yi**	4	13	8	14	12	10	7	11

Question: Calculer le coefficient de corrélation des rangs de Spearman. Que concluez vous?

N	**xi**	**Yi**	**Rxi**	**Ryi**	**/Rxi-Ryi/**	**(Rxi-Ryi)²**
A	6	4	8	8	0	0
B	17	13	1	2	1	1
C	14	8	4	6	2	4
D	15	14	3	1	2	4
E	8	12	6	3	3	9
F	9	10	5	5	0	0
G	7	7	7	7	0	0
H	16	11	2	4	2	4
∑	/	/	/	/	/	**22**

On ordonne les valeurs selon un ordre décroissant.(rang)

Rxi: (1) 17 - (2) 16 -(3) 15 - (4) 14- (5) 9- (6) 8-(7) 7 -(8) 6.

Ryi: (1) 14 - (2) 13 -(3) 12 - (4) 11- (5) 10- (6) 8-(7) 7 -(8) 4.

$$rs = 1 - \frac{6\sum(Xi-yi)^2}{n\,(n^2-1)}$$

$$rs = 1 - \frac{6 \times 22}{8\,(64-1)}$$

$$rs = 1 - \frac{132}{504}$$

rs= 1 - 0,26 = +**0,74**

rs = **0,74** → On déduit qu´il existe une corrélation positive forte entre les (2) variables .

Exercice 2: Le tableau donne le résultat d´un test d´aptitude auquel se sont soumis 13 pairs de jumeaux .

1	2	3	4	5	6	7	8	9	10	11	12	13
27	16	15	13	10	21	23	22	11	24	17	14	12
25	11	13	14	15	22	18	19	17	23	12	16	24

Question: Calculer le coefficient de corrélation des rangs de Spearman. Que concluez vous?

N	**Xi**	**yi**	**Rxi**	**Ryi**	**/Rxi-Ryi/**	**(Rxi-Ryi)²**
1	27	25	1	1	0	0
2	16	11	7	13	6	36
3	15	13	8	11	3	9
4	13	14	10	10	0	0
5	10	15	13	9	4	16
6	21	22	5	4	1	1
7	23	18	3	6	3	9
8	22	19	4	5	1	1
9	11	17	12	7	5	25
10	24	23	2	3	1	1
11	17	12	6	12	6	36
12	14	16	9	8	1	1
13	12	24	11	2	9	81
Total	/	/	/	/	/	**216**

On ordonne les valeurs selon un ordre décroissant.(rang)

Rxi: (1) 27 -(2) 24 -(3) 23 - (4) 22- (5) 21- (6) 17- (7) 16 - (8) 15- (9)- 14 (10) 13- (11) 12-(12) 11- (13) 10.

Ryi: (1) 25 -(2) 24 -(3) 23 - (4) 22- (5) 19- (6) 18-(7) 17 -(8) 16- (9) 15- (10) 14- (11) 13-(12) 12- (13) 11.

$$rs = 1 - \frac{6\sum(Xi-yi)^2}{n\,(n^2-1)}$$

$$rs = 1 - \frac{6 \times 216}{13\,(169-1)}$$

$$rs = 1 - \frac{1296}{2184}$$

rs= 1 - 0,59 = +**0,41**

rs = **0, 41** → On déduit qu´il existe une corrélation positive moyenne entre les (02) tests variables .