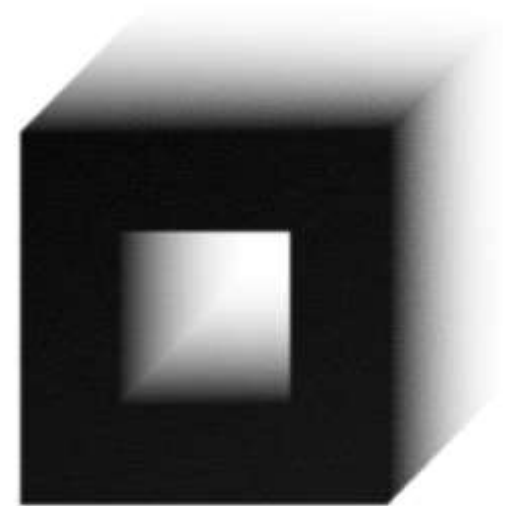


02

Statistique descriptive à deux dimensions



1. Introduction

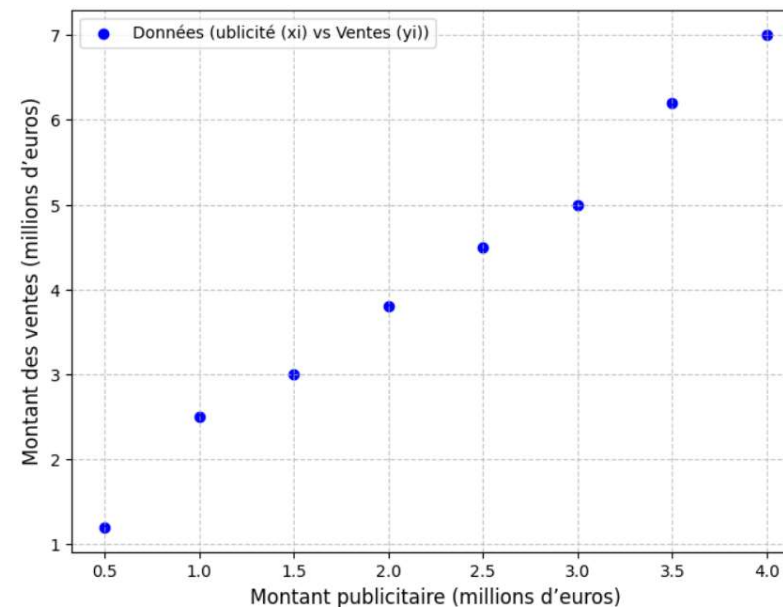
Dans les chapitres précédents, nous avons exploré les méthodes permettant de synthétiser et de représenter les informations associées à une seule variable. Cependant, un même individu ou une même situation peut être analysé(e) à travers plusieurs caractéristiques ou variables. Par exemple, en management, on peut évaluer la performance des employés en croisant leur niveau de satisfaction au travail et leur productivité. De même, on peut analyser la réussite d'un projet en prenant en compte à la fois le respect des délais et le budget alloué. Dans ce chapitre, nous introduirons une approche globale pour étudier les relations entre deux variables, en nous limitant au cas de deux variables à la fois.

2. Représentation graphique

Étant donné deux variables quantitatives X et Y. Un **nuage de points** est une représentation graphique où chaque point M_i a pour abscisse x_i (valeur de la première variable) et pour ordonnée y_i (valeur de la deuxième variable). Il permet de visualiser la relation entre deux variables quantitatives.

Exemple : Voici un exemple de série d'observations sur 8 mois (montant des publicités en millions d'euros et montant des ventes en millions d'euros) :

4	Publicité	Vente
1	0,5	1,2
2	1	1,5
3	1,5	3
4	2	3,8
5	2,5	4,5
6	3	5,0
7	3,5	6,2
8	4	7,0
Total		



3. Liaison entre deux variables quantitatives

3.1. Covariance entre deux variables statistiques

Soient X et Y deux variables quantitatives. **La covariance entre X et Y** est donnée par :

$$\text{Conv}(X, Y) = \left(\frac{1}{n} \sum_i^n x_i y_i \right) - \bar{x} \cdot \bar{y}$$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n} \sum_{i=1}^k n_i x_i$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = \frac{1}{n} \sum_{i=1}^k n_i y_i$$

3. Liaison entre deux variables quantitatives

3.1. Covariance entre deux variables statistiques

Propriétés de covariance :

- $\text{Cov}(X, Y) = \text{Cov}(Y, X)$;
- $\text{Cov}(X, X) = V(X)$;
- $V(X+Y) = V(X) + V(Y) + 2 \text{Cov}(X, Y)$;
- $\text{Cov}(aX, Y) = a \text{Cov}(X, Y)$;
- Si X et Y sont indépendantes, $\text{Cov}(X, Y) = 0$; la réciproque est fausse.

3. Liaison entre deux variables quantitatives

La covariance permet d'évaluer si les variables X et Y présentent une tendance à être liées de manière linéaire. Cependant, son principal inconvénient réside dans la difficulté de son interprétation, car sa valeur dépend des unités choisies pour X ou Y. C'est pourquoi il est généralement préférable d'utiliser le **coefficient de corrélation**.

3.2. Le coefficient de corrélation linéaire

Le coefficient de corrélation linéaire entre les deux variables statistiques X et Y s'exprime comme suit :

$$r(X; Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

Avec, $-1 \leq r(X; Y) \leq 1$

3. Liaison entre deux variables quantitatives

Propriétés de coefficient de corrélation linéaire

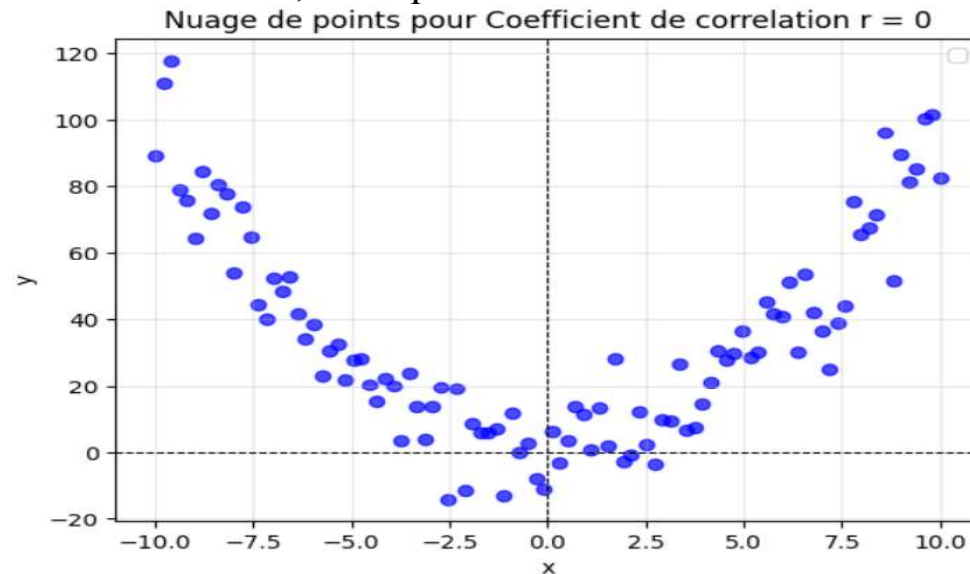
- $r(X, Y) = r(Y, X)$
- Lorsque $r(X, Y) < 0$, alors **la relation linéaire** entre X et Y **est négative**
- Lorsque $r(X, Y) > 0$, alors **la relation linéaire** entre X et Y **est positive**
- Lorsque $r(X, Y) = 0$, alors **la relation linéaire** entre X et Y **est nulle**; il existe peut être une relation non linéaire entre X et Y,

3. Liaison entre deux variables quantitatives

Propriétés de coefficient de corrélation linéaire

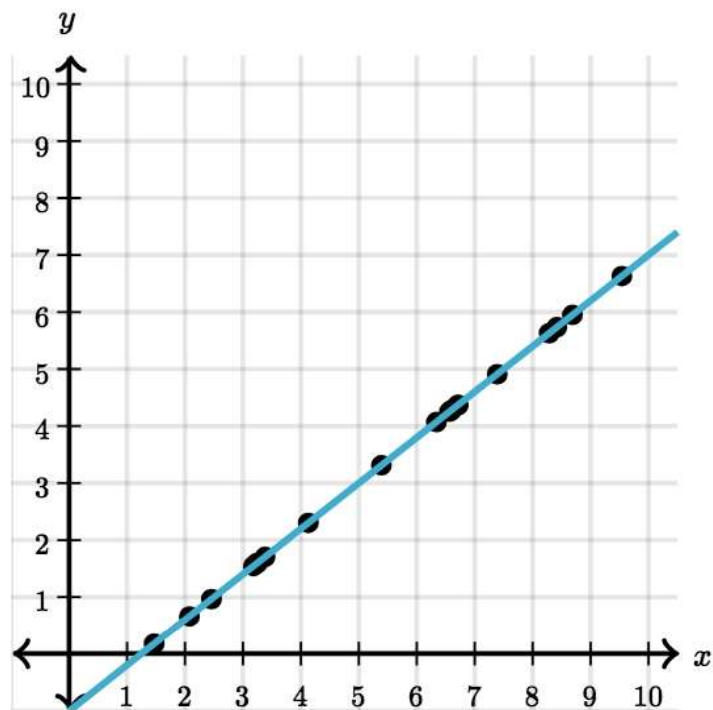
- Si $r = 1$, les deux variables sont dites **parfaitement corrélées**.
- Si $0,75 \leq r \leq 1$, les deux variables **sont fortement corrélées**.
- Si $0,5 \leq r \leq 0,74$, les deux variables **sont moyennement corrélées**.
- Si $0 \leq r \leq 0,49$, les deux variables **sont faiblement corrélées**.
- Si $r = 0$, cela indique une indépendance linéaire entre X et Y , bien qu'une autre forme de dépendance puisse exister.

Dans la figure : $Y = X^2$; Y et X sont dépendantes, mais le coefficient de corrélation : $r = 0$



3. Liaison entre deux variables quantitatives

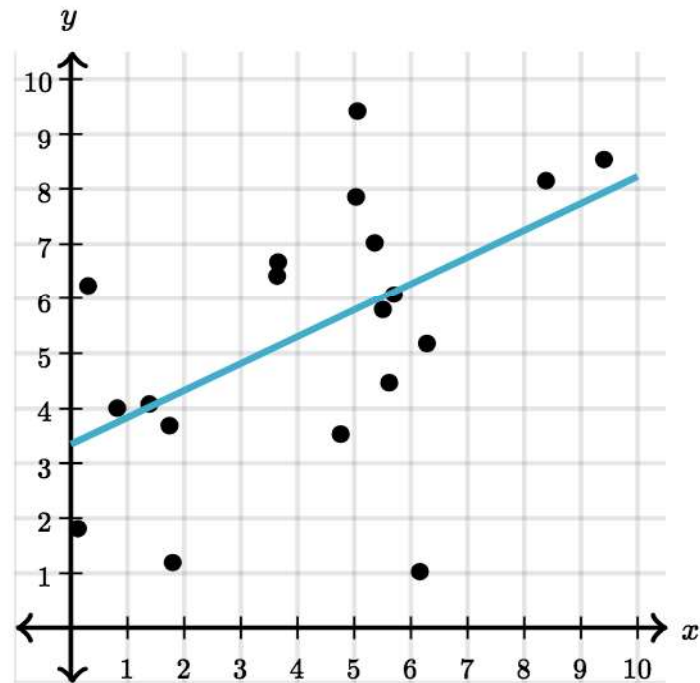
Exemple 1



$r = 1$: Corrélation positive parfaite entre les deux variables

3. Liaison entre deux variables quantitatives

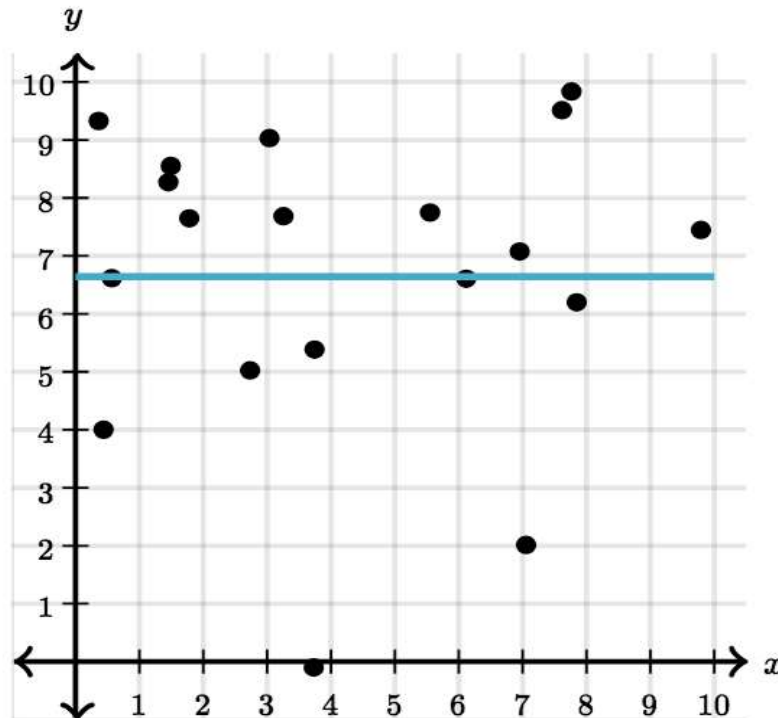
Exemple 2



$r = 0,5$: Corrélation positive faible entre les deux variables

3. Liaison entre deux variables quantitatives

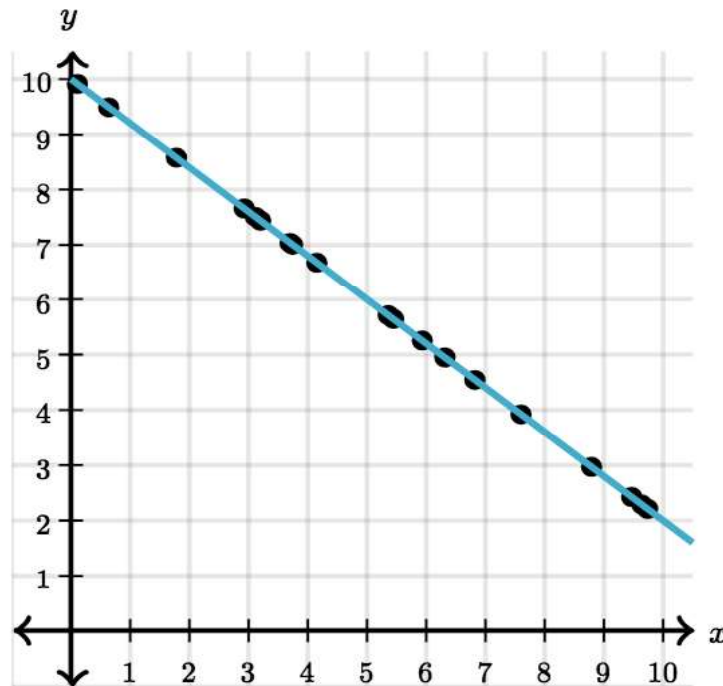
Exemple 3



$r = 0$: Absence totale de corrélation entre les deux variables. Les deux variables sont linéairement indépendantes

3. Liaison entre deux variables quantitatives

Exemple 4

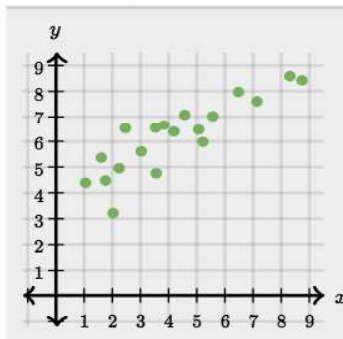


$r = -1$: Corrélation négative parfaite entre les deux variables

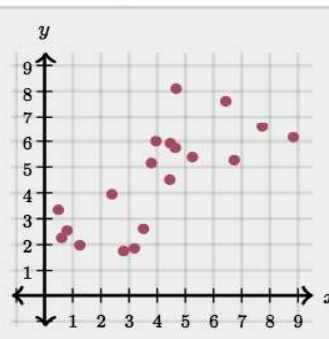
3. Liaison entre deux variables quantitatives

Exercice d'application 1 : Associez le coefficient de corrélation au nuage de points correspondant.

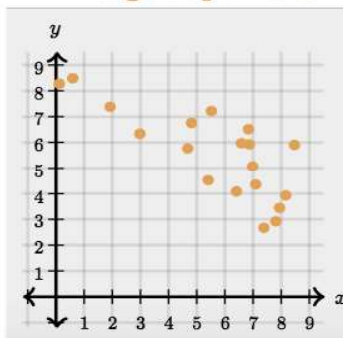
Nuage de points A



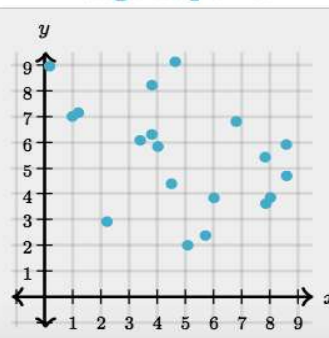
Nuage de points B



Nuage de points C



Nuage de points D



$$r = -0,77$$

$$r = 0,73$$

$$r = -0,42$$

$$r = 0,87$$

3. Liaison entre deux variables quantitatives

Exercice d'application 2 : On considère le tableau de donnée suivant :

X	2	2	2	2	4	4	4	4	4	4	4
Y	4	4	4	6	4	4	6	6	6	6	6

- Calculer le coefficient de corrélation r et interpreter le résultat

Solution : 1ere méthode

X	2	2	2	2	4	4	4	4	4	4	4	36
Y	4	4	4	6	4	4	6	6	6	6	6	56
XY	8	8	8	12	16	16	24	24	24	24	24	188
X²	4	4	4	4	16	16	16	16	16	16	16	128
Y²	16	16	16	36	16	16	36	36	36	36	36	296

3. Liaison entre deux variables quantitatives

Exercice d'application 2

Solution : 1ere méthode

Le coefficient de corrélation r est donnée par :

$$r = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y} = \frac{\frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y}}{\sqrt{\frac{1}{n} \sum_{i=1}^n x_i^2 - (\bar{x})^2} \sqrt{\frac{1}{n} \sum_{i=1}^n y_i^2 - (\bar{y})^2}} = \frac{\frac{188}{11} - \frac{36}{11} \times \frac{56}{11}}{\sqrt{\frac{128}{11} - \left(\frac{36}{11}\right)^2} \sqrt{\frac{296}{11} - \left(\frac{56}{11}\right)^2}} \cong 0,449$$

Interprétation :

Le coefficient de corrélation $r = 44,9\%$ est positif, ce qui signifie que X et Y varient dans le même sens (positif).

De plus, étant donné que $0 < |r| < 0,49$, on observe une corrélation faible entre les variables X et Y .

3. Liaison entre deux variables quantitatives

Exercice d'application 2

Solution : 2ème méthode en utilisant le tableau de contingence

X \ Y	Y		Total
	$y_1=4$	$y_2=6$	
$x_1=2$	$n_{11}=3$	$n_{12}=1$	$n_{1.}=4$
$x_2=4$	$n_{21}=2$	$n_{22}=5$	$n_{2.}=7$
Total	$n_{.1}=5$	$n_{.2}=6$	$n=11$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^2 n_{i.} \times x_i = \frac{4 \times 2 + 7 \times 4}{11} = \frac{36}{11} = 3,273$$

$$\Rightarrow V(X) = \frac{1}{n} \sum_{i=1}^2 n_{i.} \times (\bar{x} - x_i)^2 = 0,926 \Rightarrow \sigma_X = \sqrt{V(X)} = 0,962$$

$$\bar{y} = \frac{1}{n} \sum_{j=1}^2 n_{.j} \times y_j = \frac{5 \times 4 + 6 \times 6}{11} = \frac{56}{11} = 5,091$$

$$\Rightarrow V(Y) = \frac{1}{n} \sum_{j=1}^2 n_{.j} \times (\bar{y} - y_j)^2 = 0,992 \Rightarrow \sigma_Y = \sqrt{V(Y)} = 0,996$$

3. Liaison entre deux variables quantitatives

Exercice d'application 2

Solution : 2ème méthode en utilisant le tableau de contingence

$$\begin{aligned} \text{Cov}(X, Y) &= \frac{1}{n} \sum_{i=1}^2 \sum_{j=1}^2 n_{ij} x_i y_j - \bar{x} \bar{y} \\ &= \frac{188}{11} - \frac{36}{11} \frac{56}{11} \\ &= 0,430 \end{aligned}$$

Par suite le coefficient de corrélation linéaire est :

$$r(X; Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{0,430}{0,962 \times 0,996} = 0,449$$

4. Ajustement linéaire d'un nuage de points

On observe simultanément les variables quantitatives X et Y sur n individus, puis on trace l'ensemble des points $M_i(x_i, y_i)$. **L'objectif est de déterminer une droite d'équation $y = ax + b$ qui représente au mieux la relation linéaire entre les variables X et Y , de sorte à ce que cette droite soit la plus proche possible des points observés $M_i(x_i, y_i)$, en termes de la distance globale.**

En adoptant la méthode **des moindres carrés**, les coefficients de la droite $y = ax + b$ sont déterminés comme suit :

$$\begin{array}{ll} \blacktriangleright & a = \frac{\text{cov}(X,Y)}{V(X)} \end{array} \quad \blacktriangleright \quad b = \bar{y} - a\bar{x}$$

Cette droite est *la droite des moindres carrés de y en fonction de x* , ou encore *la droite de régression de y en x* .

Remarque :

On pourra définir la droite de moindres carrés de x en y par $x = a'y + b'$, Avec :

$$\begin{array}{ll} \blacktriangleright & a' = \frac{\text{cov}(X,Y)}{V(Y)} \end{array} \quad \blacktriangleright \quad b' = \bar{x} - a'\bar{y}$$

4. Ajustement linéaire d'un nuage de points

Exemple :

Reprenons le dernier exemple. La droite des moindres carrés de y en fonction de x est donnée par $y = ax + b$, où

$$a = \frac{\text{cov}(X,Y)}{V(X)} = \frac{0,430}{0,926} = 0,464$$

$$b = \bar{y} - a\bar{x} = 5,091 - 0,464 \times 3,273 = 3,572$$

On obtient :

$$\mathbf{y = 0,464\,x + 3,572}$$

4. Ajustement linéaire d'un nuage de points

Exercice d'application :

Le tableau suivant est fourni pour analyser la relation entre les deux variables X et Y :

X_i	2	3	1	5	9
Y_i	4	7	3	9	17

- 1) Déterminer l'équation de la droite de régression $y = ax + b$.
- 2) Représenter graphiquement la droite de régression.
- 3) Calculer le coefficient de corrélation r et interpréter sa valeur.
- 4) Calculer le résidu e_1 et e_3 .