FACULTE DE MEDECINE D'ORAN
LABORATOIRE DE BIOSTATISTIQUE

Corrélation et regression

Introduction :

D'habitude on effectue des expériences, des applications numériques sur les modèles théoriques et les théorèmes pour traduire sur le terrain les concepts, en passant de la THEORIE à LA PRATIQUE .
Dans le chapitre présent on procède dans le sens inverse. On détermine un modèle mathématique à partir des données échantillonnées :

DONNEES OBSERVEES → **MODELE MATHEMATIQUE**

ou encore :

ECHANTILLON → **RELATION THEORIQUE** → **PREVISION**

On se limite au cas de deux variables quantitatives et la corrélation linéaire.

Le poids et la taille, l'âge de la mère et le poids du nouveau-né, le taux d'alphabétisme des mères et le taux de mortalité infantile …..sont liés d'une certaine façon.

L'objet des techniques de corrélation et de régression est de vérifier l'existence ou l'absence de la relation entre ces deux variables, appelées variable explicative x indépendante et variable expliquée y dépendante

On suppose

Un échantillon de

n éléments

Sur chaque élément on effectue

l'expérience portant sur

deux caractères quantitatifs

dont les variables correspondantes sont **Xi** et **Yi**

(poids et taille par exemple)

Les observations forment

n couples **(Xi , Yi)**

Par exemple

à l'élément **5** correspond le couple **(X_5 , Y_5)**

où :

X_5 : est le poids de l'élément **5**

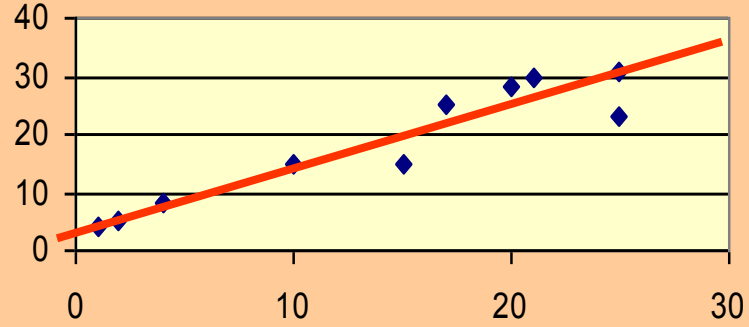
Y_5 : est la taille de l'élément **5**

Diagramme de dispersion:

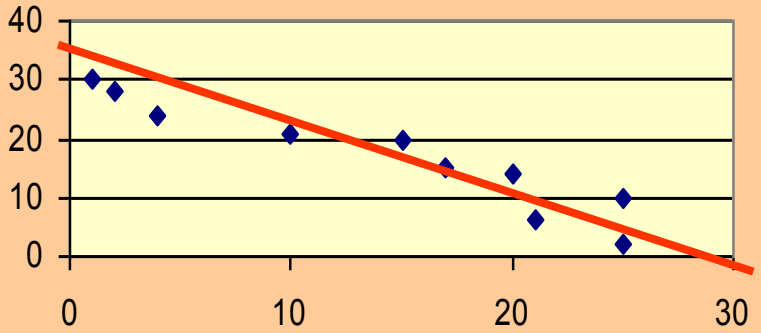
C’est un procédé graphique qui permet à première vue d’avoir une vision globale sur la nature et le sens de la relation . Il est appelé également nuage de points. Sa construction oriente et contribue au choix du type de modèle mathématique pouvant éventuellement lier les deux variables quantitatives Xi et Yi. Ce n’est nullement une démonstration ou une confirmation d’une quelconque relation.

Les n couples (Xi , Yi) observés sont représentés graphiquement dans un repère orthonormé. Plusieurs cas de figures peuvent se présenter :

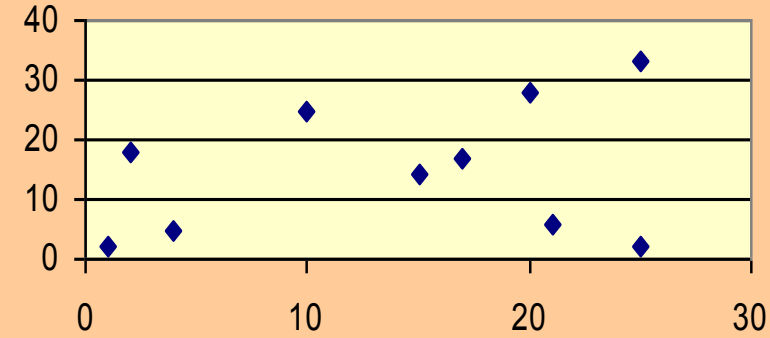
Corrélation linéaire positive



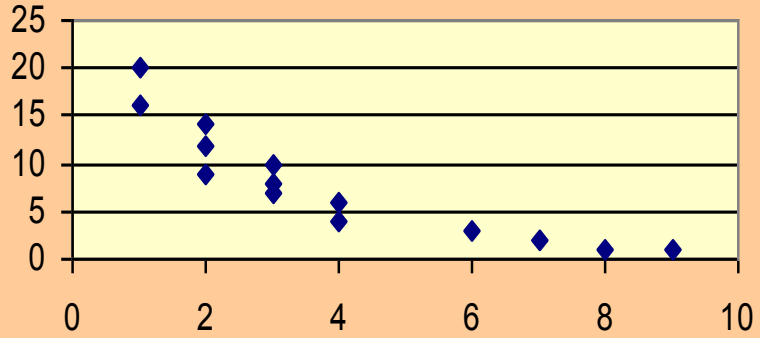
Corrélation linéaire négative



Pas de corrélation



corrélation non linéaire



ELEMENTS D'ANALYSE

La moyenne arithmétique en x et en y

$$\bar{X} = 1/n \sum xi \quad \text{et} \quad \bar{Y} = 1/n \sum Yi$$

La variance en x et en y

$$S^2x = 1/n \sum (Xi - \bar{X})^2$$

$$S^2y = 1/n \sum (Yi - \bar{Y})^2$$

$$S^2_x = 1/n \sum Xi^2 - (\bar{X})^2 \quad \text{et} \quad \mathbf{S^2_y = 1/n \sum Yi^2 - (\bar{Y})^2}$$

La covariance

C'est la variance commune entre les deux variables x et y

$$\text{Cov}(x, y) = 1/n \sum (Xi - \bar{X})(Yi - \bar{Y})$$

Ou encore

$$\text{Cov}(x, y) = 1/n \sum Xi.Yi - \bar{X}.\bar{Y}$$

La droite de régression

C'est le modèle mathématique linéaire représentant la relation qui lie la variable quantitative x_i à la variable quantitative y_i. Dans la corrélation linéaire le modèle mathématique est représenté par une droite dont la forme est :

$$y = a\,x + b$$

où $$a = Cov\,(x,y) / S^2_x$$

et $$b = \bar{y} - a\,\bar{x}$$

a représente le coefficient directeur de la droite de régression de y en x. On peut d'une manière similitude déterminer l'équation de la droite de régression de x en y.

Notons que la droite de régression passe par le point moyen($\bar{x}$; $\bar{y}$).

LE COEFFICIENT DE CORRELATION

C'est un paramètre statistique qui mesure
l'intensité de la linéarité et
le sens de la relation
Il est donné par :

$$R = Cov(x,y) / Sx \cdot Sy$$

Le coefficient de corrélation est compris entre **-1 et 1**,
en voici l'interprétation de ses différentes valeurs numériques
(à titre indicatif) :
On remarque que le coefficient de corrélation **R** (x,y) et la covariance Cov (x,y) ont le même signe (varient dans le même sens).

Coefficient de corrélation	Qualité de corrélation
$R = 1$	corrélation positive parfaite
$0,6 \leq R < 1$	bonne corrélation
$0,3 \leq R < 0,6$	corrélation médiocre
$0 < R < 0,3$	corrélation faible
$R = 0$	pas de corrélation

EXEMPLE

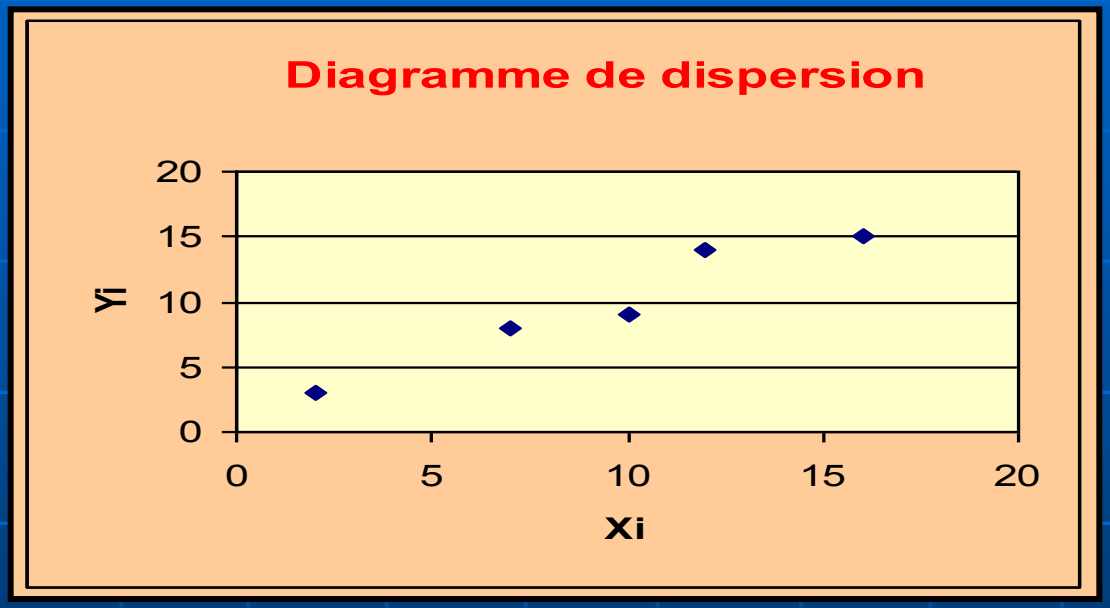
Soit un échantillon composé de 5 individus. Sur chaque individu a été relevé deux variables biologiques attachées aux variables x et y. Les résultats de l'expérience sont :

Xi	2	7	10	12	16
Yi	3	8	9	14	15

1) Construire le diagramme de dispersion.
   Commenter brièvement le graphe.
2) Calculer la covariance **cov (x, y),**
   en déduire le sens de la variation de **x** et de **y**.
3) Calculer le coefficient de corrélation **R(x,y)**
   et faire l'interprétation.
4) Etablir la droite de régression.
5) Prédire la valeur de **y** quand **x = 20**.

SOLUTION

1) Diagramme de dispersion :



Le nuage de points peut être ajusté à une droite.

Résumons les calculs dans un tableau :

Xi	Yi	Xi^2	Yi^2	Xi . Yi
2	7	4	49	14
7	8	49	64	56
10	9	100	81	90
12	14	144	196	168
16	15	256	225	240
$\sum Xi = 47$	$\sum Yi = 49$	$\sum Xi^2 = 553$	$\sum Yi^2 = 615$	$\sum Xi.Yi = 568$

2) Calcul de la covariance :

$$Cov(x, y) = 1/n \sum Xi.Yi - \bar{X} . \bar{Y}$$

$\bar{X} = 47/5 = 9,4$ et $\bar{Y} = 49/5 = 9,8$

$Sx = 4,71$ et $Sy = 5,19$

$Cov(x,y) = 568/5 - 9,4 . 9,8 = 21,48$

donc x et y varient dans le même sens

3)- Calcul du coefficient de corrélation :

R = cov(x,y) / Sx . Sy

R = 21,48 / 4,71 . 5,19

⇒ **R= 0,88**

Interprétation :

Très bonne corrélation linéaire positive

4) Droite de régression :

y = a x + b

avec **a = cov(x,y) / S^2_x**

a = 21,48 / $(4,71)^2$ = 0,968

a = 0,968

b = $\bar{y}$ – a $\bar{x}$ = 9,8 – 0,968 . 9,4 = 0,7008

b = 0,7008

L'équation de la droite :

y = 0,968 x + 0,7008

5) x = 20 ⇒ y = 0,968 . 20 + 0,7008

⇒ **y =20,06**