NOTION DE CORRELATION

Coefficient de corrélation de Bravais-Pearson, ou encore coefficient de corrélation linéaire.

Le coefficient de corrélation permet de mesurer la liaison ou le lien entre 2 ensembles de données.

Par exemple :

-Existe-t-il un lien entre le fait que les enfants mangent des sucreries et leur fréquentation des dentistes ?
-La satisfaction des clients est-elle liée à la température qui règne dans les magasins ?
-Est-ce que le niveau d'études atteint dépend du milieu social ?
-Est-ce que la mémorisation des mots d'un texte dépend de la longueur des mots ?
-Est-ce que l'impact d'une campagne publicitaire dépend du média choisi ?
-Est-ce que le cours du pétrole dépend de celui de l'euro ?
-Est-ce que le cours de l'euro dépend de celui du pétrole ?
-Est-ce que le loisir préféré des étudiants dépend de leur sexe ?

Toutes ces questions mettent en jeu deux variables. Ces deux variables sont observées sur la même population.

NOTION DE CORRELATION

Ce coefficient peut donc nous aider à émettre des pronostics mais **attention**, ce n'est pas parce qu'une corrélation existe entre 2 séries statistiques qu'il y a un lien de cause à effet entre les deux !

Exemple : une corrélation (probable) entre la taille des enfants entre 5 et 10 et leur score à un test en rapport avec le QI (non normalisé par tranche d'âge) ne signifie pas que « plus on est grand, plus on est intelligent ».

Autre exemple : s'il y a plus de naissances au printemps et en automne, on risque de trouver une corrélation entre le nombre de passage de cigognes et le nombre de naissance.
Dans ces 2 exemples, le lien provient d'une 3° variable (âge dans le 1°cas et saison dans le 2°cas).

NOTION DE CORRELATION

Exemple : Considérons une population de couples (femme , mari) et associons à chaque couple un couple d'observations (âge de la femme ; âge du mari) l'année de leur mariage.

Sur cette (même) population on observe deux variables :

Variable *X* : âge de la femme l'année de son mariage.
Variable *Y* : âge du mari l'année de son mariage

On obtient une série double (une série de 10 couples)

X	18	21	21	19	22	20	19	18	22	20
Y	20	24	26	20	24	26	24	20	26	24

NOTION DE CORRELATION

On peut évidemment étudier chaque variable indépendamment de l'autre. On se limitera au calcul de la moyenne et de la variance.

moyenne=mX=(18+21+.....)/10=20

variance=σ^2X=((18-21)²+...)/10=2

écart-type=racine carré(σ^2X)= σ X=1,42

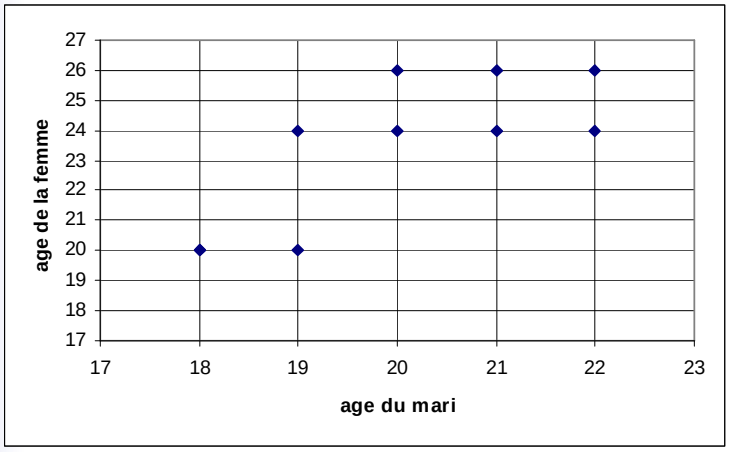
moyenne=mY=(20+24+.....)/10=23,4

variance= σ^2Y=((20-25)²+...)/10=5,64

écart-type =racine carré(σ^2Y)= σ Y=2,37

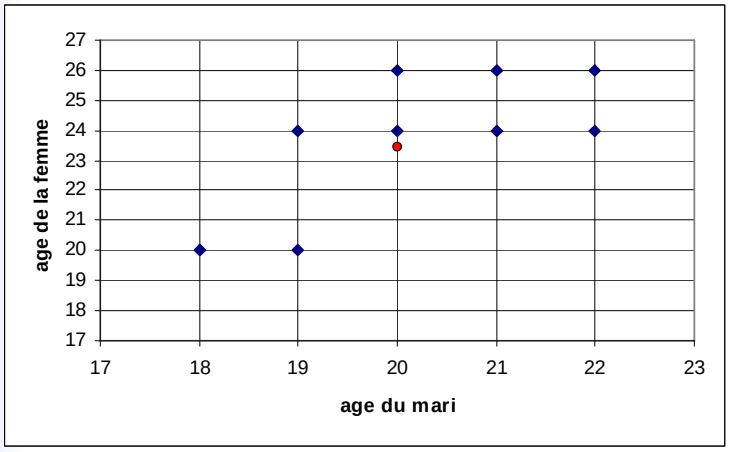
NOTION DE CORRELATION

A chaque couple d'observations (xi , yi) on associe, dans un repère cartésien, un point (géométrique) d'abscisse xi et d'ordonnée yi. L'ensemble des points ainsi obtenu est appelé nuage de points (ou nuage des individus).



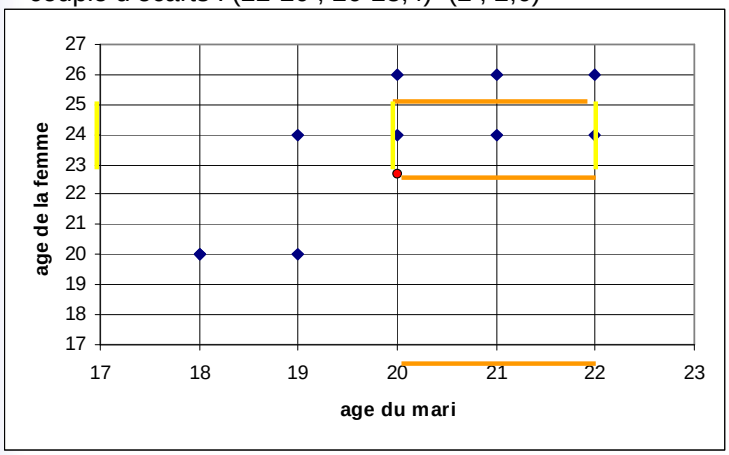
NOTION DE CORRELATION

A ce nuage de points on ajoute le point (mX , mY)=(20 ; 23,4) appelé centre de gravité du nuage ou plus simplement centre du nuage (on dit aussi point moyen).



NOTION DE CORRELATION

A chaque point (x_i , y_i) on associe son écart par rapport au point moyen (m_X , m_Y)=(20 ; 23,4). On obtient un couple d'écarts : (x_i - m_X , y_i - m_Y). Exemple au point (22,26) on associe le couple d'écarts : (22-20 , 26-23,4)=(2 , 2,6)



Le couple d'écart exprime les deux variances d'un point.

NOTION DE CORRELATION

CALCUL DU COEFFICIENT DE CORRÉLATION

Le coefficient de corrélation, noté r, se calcule à partir de 2 recueils de données différents.

$$r = \frac{\mathrm{cov}(x,y)}{\sigma_x \times \sigma_y}$$

Avec σ_x, σ_y = écart type de X et de Y

Et $\mathrm{cov}(x,y) = \frac{1}{N}\sum(x_i - \bar{x})(y_i - \bar{y}) = \frac{\sum x_i y_i}{N} - \bar{x}\bar{y}$

= covariance de X et Y

= variance (variation) commune aux deux variables/ produit des variances

= moyenne du produit des écarts de deux variables à leur moyenne respective

NOTION DE CORRELATION

											moyenne
x_i	18	21	21	19	22	20	19	18	22	20	**20**
y_i	20	24	26	20	24	26	24	20	26	24	**23,4**
x_i - moyenne X	-2	1	1	-1	2	0	-1	-2	2	0	
y_i - moyenne Y	-3,4	0,6	2,6	-3,4	0,6	2,6	0,6	-3,4	2,6	0,6	
(x_i - moyenne X) * (y_i - moyenne Y)	6,8	0,6	2,6	3,4	1,2	0	-0,6	6,8	5,2	0	**covariance=SOMME/10 = 2,6**

NOTION DE CORRELATION

La covariance permet d'estimer le sens de la variation entre deux variables numériques :

► Les variables varient dans le même sens (ou covarient) : les sujets qui ont des valeurs fortes (au dessus de la moyenne) sur une variable, présentent également des valeurs élevées sur l'autre variable. Autrement dit, les sujets les plus grands sont les sujets les plus lourds et, inversement, les sujets plus petits sont les plus légers : dans ce cas, la covariance est de signe positif (+66,831)

► Les variables varient en sens inverse : les sujets qui ont des valeurs fortes sur une des deux variables auront tendance à avoir des notes faibles sur l'autre variable. La valeur de la covariance sera alors de signe négatif

► Les variables ne covarient pas : Parmi les sujets présentant des valeurs fortes sur une variable, on peut observer que, sur l'autre variable, ces mêmes sujets obtiennent des notes fortes, faibles ou moyennes. La valeur de la covariance est proche de la valeur 0.

NOTION DE CORRELATION

Significativité du coefficient de corrélation

Si le coefficient de corrélation est proche de 1 ou de –1, cela signifie que les variables x et y sont très fortement liées (par une relation de la forme $y = ax + b$).

S'il est proche de 0, cela signifie que le lien est peu probable.

Un signe positif indique que x et y varient dans le même sens.

Mais on ne peut toujours se contenter d'une telle interprétation approximative. Il est nécessaire de savoir si ce lien est réel ou s'il est dû au simple hasard de nos mesures. Autrement dit peut-on le considérer comme reflétant un lien entre les 2 variables sur l'ensemble de la population et non pas seulement sur notre échantillon.

NOTION DE CORRELATION

1° étape : poser les hypothèses

Hypothèse nulle (Ho) : il n'y a pas de lien statistique entre les 2 séries
Hypothèse alternative (H1) : il existe un lien statistique entre les 2 séries

2° étape : calcul de **r (appelé r calculé).**

NOTION DE CORRELATION

Table du coefficient de corrélation.

Les lignes correspondent au nombre de degrés de liberté (ddl) :

avec 2 variables, ddl = N-2 (2 = nombre de variable)

Lecture de r **(appelé r lu)** correspondant à un risque d'erreur de 5% (ou probabilité de 0.05) de rejeter à tort Ho.

Si r calculé ≥ r lu alors on **rejette l'hypothèse nulle et accepte H1** : les deux distributions sont statistiquement liées. On peut l'affirmer avec un risque d'erreur < 5%.

Si r calculé < r lu alors on **ne rejette pas l'hypothèse nulle** : on ne peut pas dire que les deux distributions sont statistiquement liées. **Mais attention** : cela ne veut pas dire que l'on accepte Ho

NOTION DE CORRELATION

Notion du degré de liberté (ddl):

Le ddl reflète le nombre d'éléments indépendants entrant dans l'estimation d'une variance. C'est le nombre d'éléments pouvant varier librement, de là le terme de liberté.

Par exemple si nous voulons estimer la variance de 5 observations, nous devons dans un premier temps estimer la moyenne puisque la variance s'obtient à partir de l'écart à la moyenne. Or si nous fixons la somme des éléments, seuls quatre peuvent varier librement, la valeur du cinquième étant imposée par la valeur de la somme.

Par exemple la somme est 30 et les quatre premières valeurs sont 5, 6, 2 et 10: alors la valeur du cinquième est forcement 7 pour obtenir la somme de 30 (30-(5+6+2+10).

Donc pour estimer la moyenne d'une population à partir de la moyenne d'un échantillon, nous perdons un degré de liberté.

DDL = nombre d'observations – nombre de variables

NOTION DE CORRELATION

Notion d'hypothèses:

Nous souhaitons comparer des résultats obtenus à partir d'un échantillons (i.e. statistiques) à ceux d'une population (i.e. paramètres). Nous voulons faire des inférences d'un échantillon sur une population.
En gros ceci revient à dire que nous voulons savoir si notre échantillon est représentatif de la population.
Dans le cas de corrélation, nous voulons savoir si les deux séries sont linéairement liées ou pas. Nous avons deux possibilités, deux alternatives:

OUI: les deux séries sont liées. (H1)
NON: les deux séries ne sont pas liées. (H0)

On pose alors deux hypothèses et on va décider laquelle on accepte:

H1: Hypothèse alternative. C'est une hypothèse statistique inexacte (si l'effet n'est pas nul, alors son intensité est différente de zéro, et il existe une infinité de façon d'être différent de zéro)

H0: Hypothèse nulle. C'est une hypothèse statistique exacte (si l'effet est nul alors l'intensité de l'effet est de zéro, une seule possibilité)

NOTION DE CORRELATION

Notion d'hypothèses:
Les hypothèses concernent toujours la population, pas l'échantillon.

Comme il n'y a que H1 qui est une hypothèse inexacte alors :

On ne peut jamais rejeter H1.
Donc: On ne peut jamais accepter H0.

ON NE PEUT QUE REJETER ou NON H0.

Mais à partir de quel moment acceptons H1 ou rejetons H0. Il nous faut bien un critère pour prendre la décision.

Attention: ne pas rejeter H0 n'est pas équivalent à l'accepter.

NOTION DE CORRELATION

Notion de risque d'erreur:

Le seuil d'improbabilité est le seuil de significativité = α
Généralement on utilise comme seuil .05 ou .01.
α =.05 identifie la notion d'évènement improbable avec événements ayant moins de 5 chances sur 100 de se produire.
α est de l'ordre du subjectif, c'est notre seuil de décision, je suis sur de ce que je dis à 95% de chance.

Deux types d'erreurs:

-erreur de type I: accepter H1 alors qu'elle n'est pas vraie. Dire qu'il y a un effet alors que cet effet n'existe pas dans la réalité.
La probabilité de faire une erreur de type I correspond à α,

⇨Pour minimiser Erreur de type I: il faut un α petit.

-erreur de type II: rejeter H1 alors qu'elle est vraie. Ne pas accepter l'existence d'un effet alors que cet effet existe. La probabilité de faire une erreur de type II correspond à β, qui ne peut être connu car H1 est inexacte.

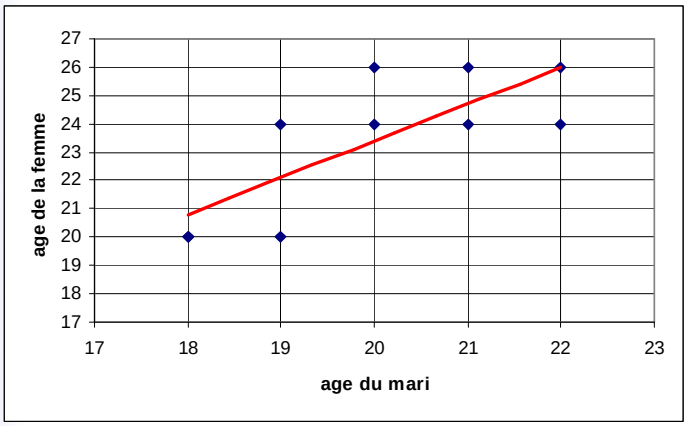
⇨Pour minimiser Erreur de type II: il faut un α grand.

NOTION DE CORRELATION

Les deux erreurs varient en sens inverse en fonction du seuil α

	Etat de la nature (inconnu)	
décision expérimentale	H0 vraie (H1 fausse)	H0 fausse (H1 vraie)
Non rejet de H0	Non rejet correct probabilité = 1-α	Erreur de type II probabilité = β
Rejet de H0	Erreur de type I probabilité = α	Détection correct probabilité = 1-β

NOTION DE CORRELATION



Covariance = 2,6

r calculé= 2,6 /(1,42 * 2,37) = 0,77

r lu (avec α = 0.05 et ddl =8) = 0,6319

r calculé > r lu donc rejet de HO, accepte H1

NOTION DE CORRELATION

Ex : Sur une population de 7 individus, on mesure 2 caractères X et Y qui prennent les valeurs suivantes :

X	5	-2	0	-1	1	2	-3
Y	9	4	0	1	1	4	9

x_i	y_i	$x_i y_i$	x_i^2	y_i^2
5	9	45	25	81
-2	4	-8	4	16
0	0	0	0	0
-1	1	-1	1	1
1	1	1	1	1
2	4	8	4	16
-3	9	-27	9	81
Σ=2	Σ=28	Σ=18	Σ=44	Σ=196

NOTION DE CORRELATION

x_i	y_i	x_iy_i	x_i^2	y_i^2
5	9	45	25	81
-2	4	-8	4	16
0	0	0	0	0
-1	1	-1	1	1
1	1	1	1	1
2	4	8	4	16
-3	9	-27	9	81
Σ=2	Σ=28	Σ=18	Σ=44	Σ=196

Calcul de cov (x,y) :

x = Σ xi/N = 2/7 = 0,2857 = 0,29

y = Σ yi/N =28/7 = 4

cov(x,y) = 18/7-0,29*4 = 2,57-1,16 = 1,41

Calcul de σx et σy :

σx2 = Σ xi2/N-x2 = 44/7-0,292 = 6,29-0,0841 = 6,21→ σx = √6,21 = 2,49

σy2 = Σ yi2/N-y2 = 196/7-42 = 28-16 = 12→ σy = √12 = 3,46

Calcul de r :

r calculé = 1,41/(2,49*3,46) = 0,16

Le degré de corrélation entre ces deux variables avoisine le 0, on peut donc penser qu'elles ne sont pas très liées mais on vérifie

ddl : N-2 = 7-2 = 5

r lu = 0,7545

donc r calculé = 0,16 < r lu : 0,7545

⇒ on ne rejette pas Ho : on ne peut conclure à un lien statistique entre X et Y

NOTION DE CORRELATION

Représentation graphique.

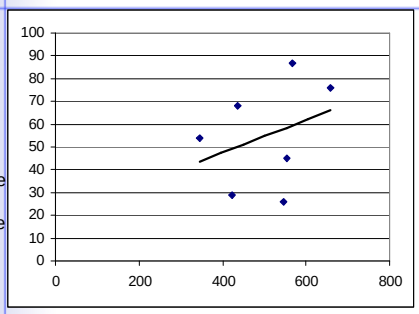
Si l'on représente sur un graphique les scores de chaque sujet dans un repère orthogonal dont les axes correspondent aux 2 variables :

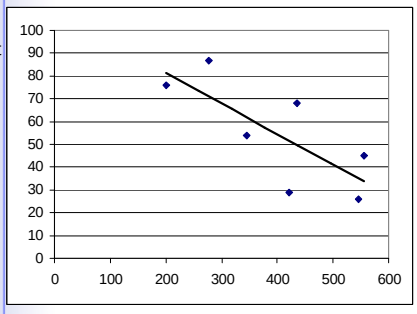
Le lien entre les 2 variables est visible dans le fait que le nuage de point se rapproche fortement d'une droite (appelée droite de régression) de type ax +b de pente non nulle.

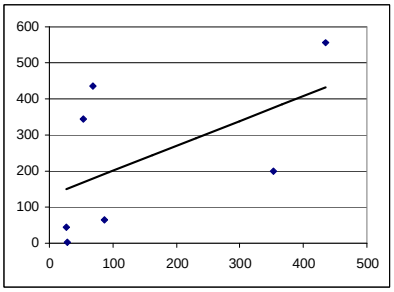
Si la pente est positive alors la corrélation est positive, si elle est négative alors la corrélation est négative.

Remarque : une droite horizontale ou verticale signifierait que les données ne dépendent que d'une seule des 2 variables.

NOTION DE CORRELATION







NOTION DE CORRELATION

REMARQUE:

Propriétés de la covariance :

Symétrie : cov(X,Y) = cov (X,Y)

Cov(X,X)=var(X)

Les vecteurs dont on veut connaître la covariance doivent être de taille identique

Linéarité:

On calcule la régression linéaire par la méthode des moindres carrés, c'est-à-dire qu'on cherche la droite y=a'x+b' qui minimise les distances entre la droite et tous les points. C'est-à-dire qu'on minimise $\Sigma (y_i-y'_i)^2$.

La régression linéaire Dx/y a pour forme: y=a'x+b'
Où a'= cov(x,y) / var(y) et b' = moyenne(X) –(a'*moyenne(Y))

La régression linéaire Dy/x a pour forme: y=a'x+b'
Où a'= cov(x,y) / var(x) et b' = moyenne(Y) –(a'*moyenne(X))