Dr MERADSI Fouad

Modèle statistique n° 01

Méthode Statistique descriptive bi-variée : Corrélation linéaire

L'objectif de cette partie est d'étudier sur une même population de n individus, deux caractères différents X et Y et de rechercher s'il existe un lien entre ces deux variables.

Covariance

C'est une étape préliminaire pour étudier une relation entre deux variables quantitatives.

C'est la moyenne des produits des écarts pour chaque série d'observation, et donnée par la formule suivante :

$$Cov(x,y)= S_{xy} = 1/n \sum(x_i – m) (y_i - \bar{y})$$

La corrélation

La corrélation est la netteté ou l'intensité de la relation existante entre deux séries de données.

Le coefficient de corrélation (r)

Le coefficient de corrélation mesure la dépendance linéaire entre les variables X et Y. Le coefficient de corrélation, est précisément le rapport de la covariance sur le produit des écarts-types de deux variables X et Y.

$$r = Cov (x, y) / Sx \times Sy$$

On a $-1 < r < 1$. Si r est proche de 1 ou -1, les variables X et Y sont dits : fortement corrélées.

Propriétés

Si le coefficient de corrélation est positif, les points du nuage sont alignés le long d'une droite croissante. Dans ce cas X et Y évoluent dans le même sens (figure 1).

Si le coefficient de corrélation est négatif, les points sont alignés le long d'une droite décroissante. Dans ce cas X et Y évoluent dans des sens opposés (figure 1).

Si le coefficient de corrélation est nul ou proche de zéro, il n'y a pas de dépendance linéaire (figure 1).

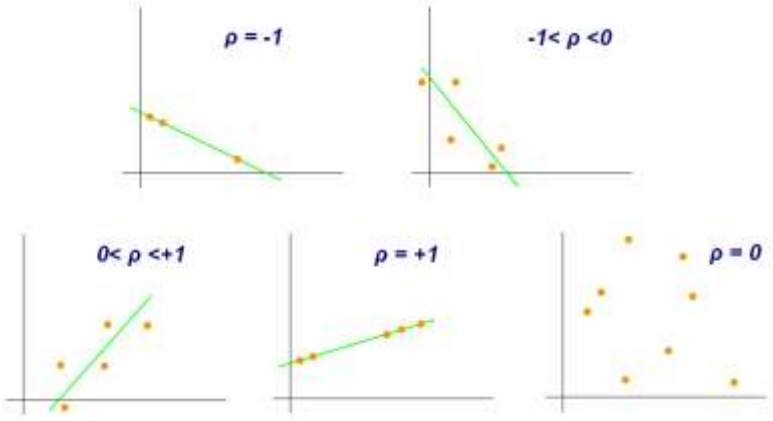


Figure 1 : Exemples de diagrammes de dispersion avec différentes valeurs de coefficient de corrélation.

Droite de régression

L'idée est de transformer un nuage de point en une droite. Celle-ci doit être la plus proche possible de chacun des points. On cherchera donc à minimiser les écarts entre les points et la droite. Cette méthode vise à expliquer un nuage de points par une droite qui lie Y à X.

$$Y = aX + b$$

$a = Cov(x, y) / S^2_X$ et $b = \bar{y} - a\,m$ (m : moyenne de premier paramètre).

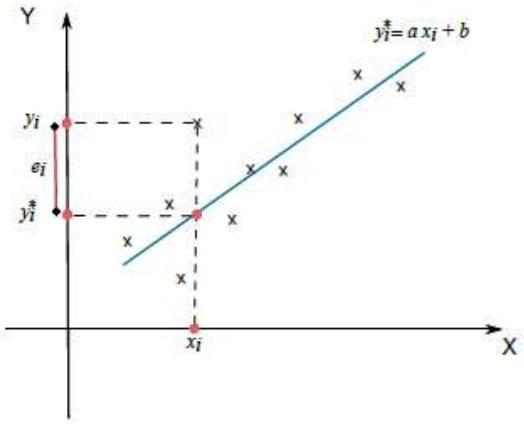


Figure 2 : La droite la plus proche possible de chacun des points.

Ajustement linéaire

L'ajustement linéaire consiste à remplacer le nuage de points par une droite à l'aide d'une équation de la régression.

Dans le cas où on peut mettre en évidence l'existence d'une relation linéaire significative entre deux caractères quantitatifs continus X et Y, on peut chercher à formaliser la relation moyenne qui unit ces deux variables à l'aide d'une équation de droite qui résume cette relation. Nous appelons cette démarche l'ajustement linéaire.

Remarque

Le coefficient de corrélation (r) permet de justifier le fait de l'ajustement linéaire. On adopte les critères numériques suivants (voir Figure 3).

Si $| r | < 0,7$; alors l'ajustement linéaire est refusé (droite refusée).

Si $| r | \geq 0,7$; alors l'ajustement linéaire est accepté (droite acceptée).

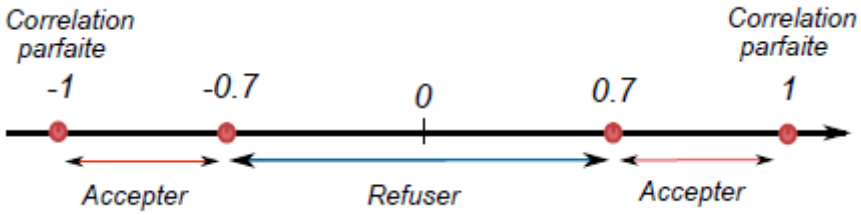


Figure 3 : La zone d'acceptation ou de refus de l'ajustement linéaire.

Exemple d'application 01

Une étude de la liaison entre l'âge (années) et le diamètre (mm) dans une population d'arbres a été conduite sur un échantillon de 7 plantes.

Age (années)	1	2	3	4	5	6	7
Diamètre (mm)	30	90	120	130	150	160	160

Etudiez la relation entre les deux variables.

Solution

m = 4 et Sx = 2,16

$\bar{y}$ = 120 et Sy = 46,90

Cov(xy) = 28,88.

Cov(xy) = 93,33

r = 0,92. Il existe une forte corrélation positive.

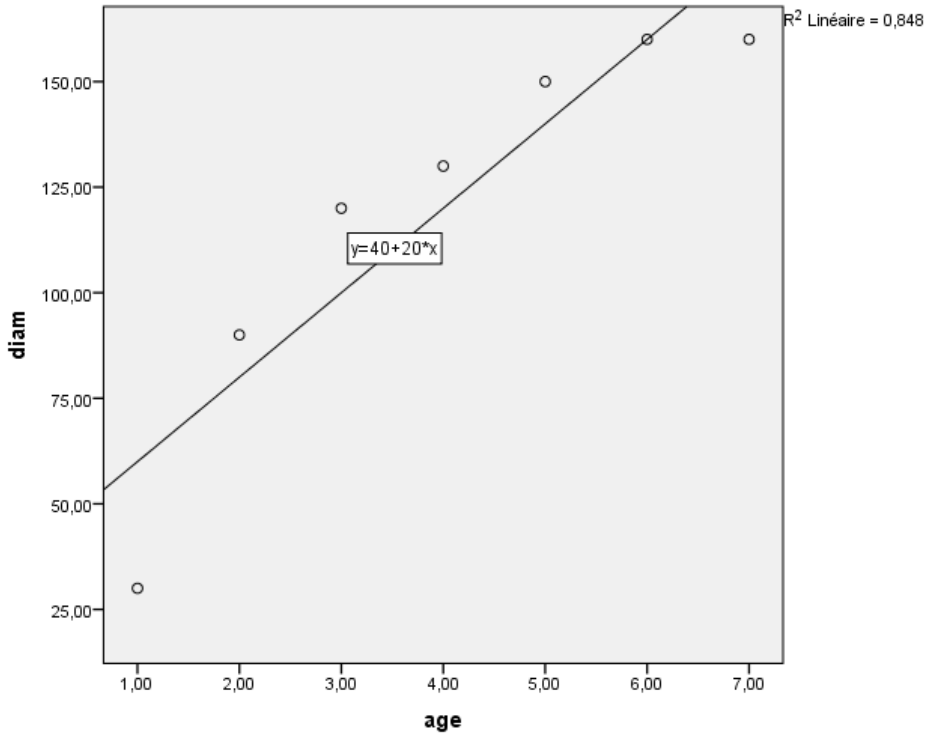
a = 20 et b = 40.

Y = (20) x + 40.

Corrélations

		âge	diamètre
âge	Coefficient de corrélation (r)	1	**0,921**
	Covariance		93,333
	N	7	7

Représentation graphique



Exemple d'application 02

L'étude de deux variables (X) le poids de 1000 grains en gramme et le rendement (Y) en quintaux par hectare chez le blé a donné les résultats suivants :

X (g)	43	46	48	50	52	55	56	58	60	62
Y (qx/ha)	30	33	35	36	37	39	39	42	43	45

1. Calculer la covariance Cov(xy) qui associe X et Y.
2. Calculer le coefficient de corrélation r_{xy} et quelle est votre conclusion ?
3. Déterminer l'équation de la régression qui lie Y à X.

Solution

m = 53 et Sx = 6,25

$\bar{y}$ = 37,5 Sy = 4,65

Cov(xy) = 28,88.

Cov(xy) = 28,88.

r = 0,99. Il existe une forte corrélation positive.

a = 0,73 et b = -1,24.

Y = (0,73) x – (1,24).

Statistiques descriptives

	Moyenne	Ecart-type	N
X	53,0000	6,25389	10
Y	37,9000	4,65355	10

Corrélations

		X	Y
X	Coefficient de corrélation (r)	1	**0,993**
	Covariance :		28,889
	N	10	10

Représentation graphique