Chapitre 3 - Série statistique à deux variables

Table des matières

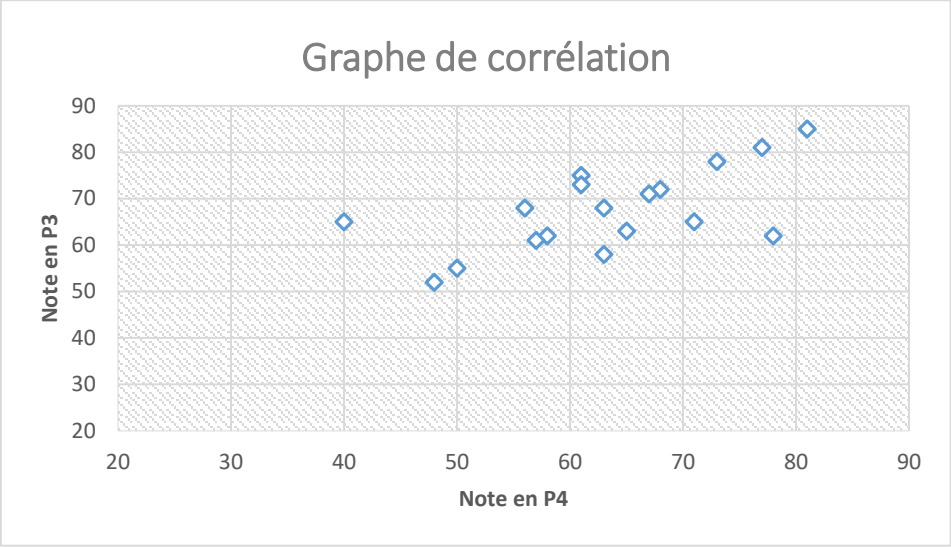
1. Définitions

L'objectif de l'analyse bivariée est d'étudier les éventuelles relations entre **deux variables statistiques**.

Exemple : xi = notes en BTS CG – P4 (sur 100) et yi = notes en BTS CG – P3 (sur 100) de 18 étudiants numérotés de 1 à 18 :

Étudiant	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Xi	71	40	48	81	58	63	56	68	61	67	57	73	77	63	78	65	61	50
yi	65	65	52	85	62	58	68	72	75	71	61	78	81	68	62	63	73	55

Graphe de corrélation : représentation dans un repère cartésien orthonormé de chaque couple de scores.



Une **série statistique à deux variables** $(x_i\ ;\ y_i)$ est constituée d'une liste de n couples de valeurs $(x_1\ ;\ y_1)\ ;\ (x_2\ ;\ y_2)\ ;\ \ldots\ ;\ (x_n\ ;\ y_n)$.

Le **nuage de points** de cette série est l'ensemble des points du plan de coordonnées $(x_1\ ;\ y_1)\ ;\ (x_2\ ;\ y_2)\ ;\ \ldots\ ;\ (x_n\ ;\ y_n)$.
Les coordonnées **du point moyen G** sont $(\overline{x}\ ;\ \overline{y})$, moyenne des xi, et moyenne des yi. Ici : $\overline{x}$: 63,17 et $\overline{y}$ = 67,44

Le nuage de points permet de constater qu'un des joueurs semble être en surcharge pondérale par rapport aux autres.

En observant ce nuage de points, on peut vérifier s'il y a **dépendance ou indépendance** entre les deux variables. Effectivement on observe une liaison forte et positive que l'on peut mesurer grâce au **coefficient de corrélation.**

La connaissance des paramètres suivants nous permettra de tracer la droite de régression de Y en X, encore appelée **droite des moindres carrés ordinaires.**

Entraînez-vous avec notre mise en situation réelle type examen. Sujet inédit + corrigé méthodologique étape par étape (sur le site - shop BTS CG).

2. Calcul du coefficient de corrélation linéaire

Coefficient de corrélation linéaire : coefficient qui quantifie la liaison (force et sens) entre deux variables x et y, on le note r_{xy} :

$$R(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}.$$

Variance (x) = $\sum_1^n (Xi - \overline{x})^2 \ / \ n$

Ecart Type (x) = **√(Variance(X))**

Ce coefficient peut prendre toute valeur entre –1 et +1.

Covariance : c'est un paramètre qui combine la dispersion de la variable x par rapport à sa moyenne mx et la dispersion de la variable y par rapport à sa moyenne my.

Elle se note :

$$\operatorname{cov}(X,Y) = \frac{(x_1 - \overline{x})(y_1 - \overline{y}) + (x_2 - \overline{x})(y_2 - \overline{y}) + \ldots + (x_N - \overline{x})(y_N - \overline{y})}{N},$$

cov_{xy} cov_{xy} = [somme $(x_i – m_x)$ $(y_i – m_y)$] / N ou encore cov_{xy} = [(somme $x_i y_i$) / N] - $m_x m_y$

Signification :

- Si r_{xy} est supérieur ou égal à 0.8 alors on peut conclure à une forte corrélation positive entre les deux variables.
- Si r_{xy} est inférieur ou égal à - 0.8 alors on peut conclure à une forte corrélation négative entre les deux variables.
- Si r_{xy} est proche de 0 alors on peut conclure à une absence de corrélation, donc à une indépendance des deux variables.

En dehors de ces bornes, on peut juste conclure qu'il n'y a pas de dépendance significative ou d'indépendance.

Exemple :

Rxy = 0,56 (niveau d'étude / salaire) ; Rxy = 0,86 (participation en classe et note générale).

Attention : ce coefficient sert à trouver la relation entre deux variables quantitatives sans pouvoir déduire les **relations causales.**

Exemple :

Étudiant	X_i	Y_i	$X_i - m_x$	$(X_i - m_x)^2$	$Y_i - m_y$	$(Y_i - m_y)^2$	$(X_i-m_x)(Y_i-m_y)$
1	71	65	7,83	61,36	-2,44	5,98	-19,15
2	40	65	-23,17	536,69	-2,44	5,98	56,63
3	48	52	-15,17	230,03	-15,44	238,53	234,24
4	81	85	17,83	318,03	17,56	308,20	313,07
5	58	62	-5,17	26,69	-5,44	29,64	28,13
6	63	58	-0,17	0,03	-9,44	89,20	1,57
7	56	68	-7,17	51,36	0,56	0,31	-3,98
8	68	72	4,83	23,36	4,56	20,75	22,02
9	61	75	-2,17	4,69	7,56	57,09	-16,37
10	67	71	3,83	14,69	3,56	12,64	13,63
11	57	61	-6,17	38,03	-6,44	41,53	39,74
12	73	78	9,83	96,69	10,56	111,42	103,80
13	77	81	13,83	191,36	13,56	183,75	187,52
14	63	68	-0,17	0,03	0,56	0,31	-0,09
15	78	62	14,83	220,03	-5,44	29,64	-80,76
16	65	63	1,83	3,36	-4,44	19,75	-8,15
17	61	73	-2,17	4,69	5,56	30,86	-12,04
18	50	55	-13,17	173,36	-12,44	154,86	163,85
Total				**1994,50**		**1340,44**	**1023,67**
Moyenne	**63,17**	**67,44**				**Covariance XiYi**	**56,87**
Variance	**110,81**	**74,47**					
Écart Type	**10,53**	**8,63**					

Variance X_i = 1994.50 / 18 => 110.8056,

Écart type X_i = racine carrée de 110.8056 = 10.53 Variance Y_i = 1340.44 / 18 => 74.4689,

Écart type Y_i = racine carrée de 74.4689 = 8.63 Covariance X_iY_i = 1023.67 / 18 => 56.87

Le coefficient de corrélation est donc $r_{xy} = cov_{xy} / \sigma_x \sigma_y$ => 56.87 / (10.53 * 8.63) = 0.626

Fonction Excel : COEFFICIENT.CORRELATION ; COVARIANCE.PEARSON

3. Droite des moindres carrés ordinaires (droite de régression)

La droite des moindres carrés, c'est la droite qui **résume** le mieux le nuage de points (graphique. C'est celle qui passe "au milieu" des points en minimisant l'écart avec chacun d'eux.

Droite de régression d'Y en X : quand la liaison est linéaire on peut remplacer le nuage de points par la ligne droite qui le résume le mieux, on dit qu'on ajuste le nuage ; on recherche la droite d'ajustement linéaire **Y = aX + b** qui résume le mieux le nuage.

Coefficient directeur de la droite :

$$a = \frac{\mathrm{cov}(X, Y)}{\mathrm{Var}(X)}$$

Variance (x) = $\sum_1^n (Xi - \overline{x})^2 \ / \ n$

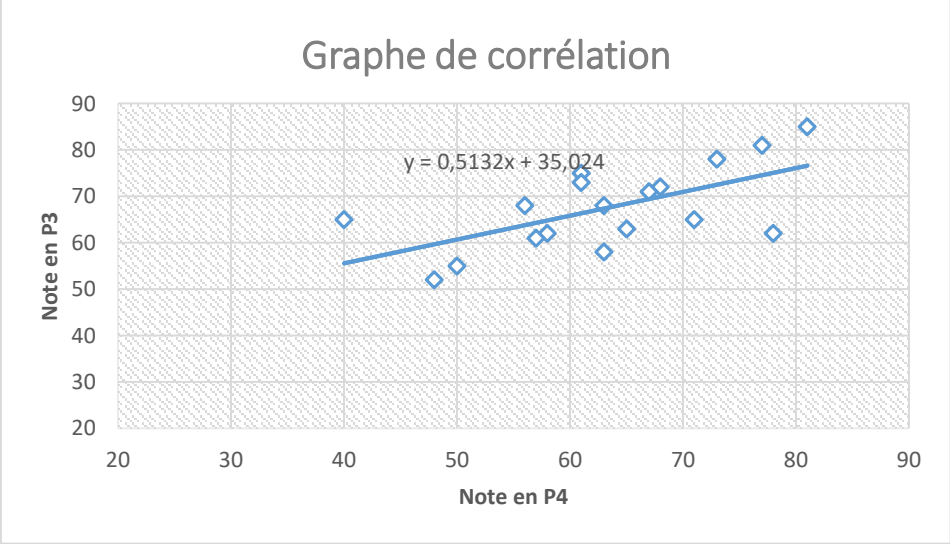
Dans l'exemple : 56.87 / 110.81 = 0.513

Ordonnée à l'origine : Y = aX + b => b = Y -aX

Pour retrouver b, il suffit de remplacer dans l'équation de la droite de régression **b = Y -aX ;** Y par m_y, X par m_x et **a** par la valeur trouvée.

Dans l'exemple : b = my – a*mx => 67.44 – 0.513 * 63.17 = 35.02

D'où, Y = 0,513X + 35,02



Cependant, avec un coefficient de corrélation < à 0,8, l'extrapolation par la droite risque d'être erronée.

Une mise en situation concrète pour valider vos acquis + la correction pas à pas pour ne plus faire d'erreurs (sur le site - shop BTS CG).