Chapitre 3 : Séries statistiques doubles (à deux variables)

2024/2025

Sommaire

1. Définition

Dans ce chapitre on considèrera que les données sont connues individuellement.

Définition 1.1

On appelle distribution statistique double une série statistique à deux variables X et Y représentée par le tableau ci-dessous :

X	x_1	x_2	...	x_i	...	x_N
Y	y_1	y_2	...	y_i	...	y_N

Exemple 1.1

Le tableau suivant donne l'âge X et la moyenne Y des maxima de tension artérielle en fonction de l'âge d'une population féminine.

Age X	36	42	48	54	60	66
Tension Y	$11,8$	14	$12,6$	15	$15,5$	$15,1$

Exemple 1.2

Pour des emplois analogues, diverses entreprises proposent les salaires notés x_i en euros. On a vu se présenter pour ces emplois le nombre de candidats noté y_i.

Salaires (x_i)	2 200	2 250	2 300	2 350	2 400
Nombre de candidats (y_i)	10	13	17	19	21

2. Nuage de points

Définition 2.1

Le nuage de points associé à la série statistique double définie par le tableau ci-dessous est l'ensemble des N points de coordonnées (x_i, y_i) dans un repère du plan.

X	x_1	x_2	...	x_i	...	x_N
Y	y_1	y_2	...	y_i	...	y_N

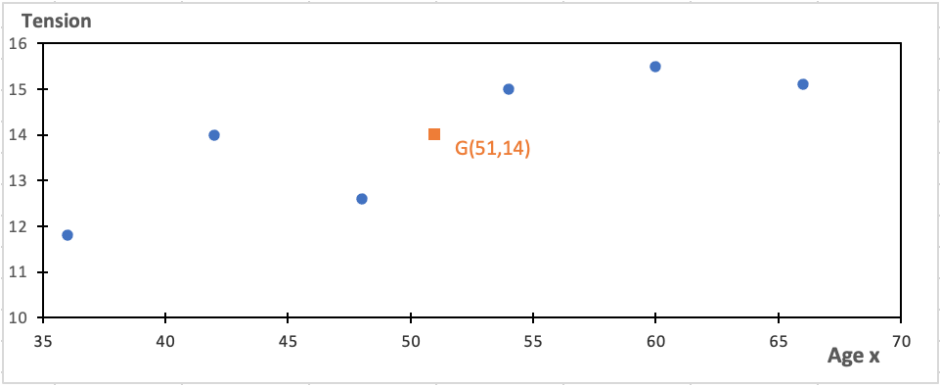
Définition 2.2

Le point moyen G est le point de coordonnées $\bar{x}$ et $\bar{y}$, où $\bar{x}$ et $\bar{y}$ sont les moyennes des variables X et Y.

2. Nuage de points

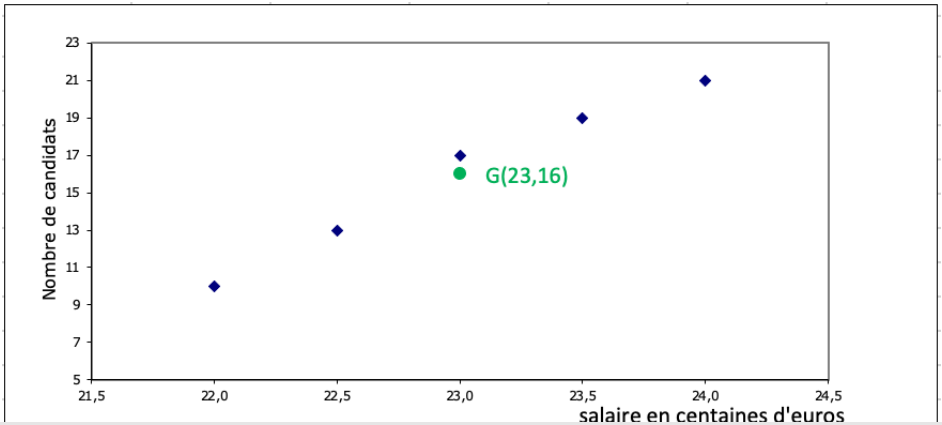
Exemple 2.1

Nuage de points associé à la série double âge/tension.



Exemple 2.2

Nuage de points associé à la série double salaire/nombre de candidats



2. Nuage de points

Remarque 2.1

Les nuages de points peuvent avoir différentes formes. Le cas le plus simple est celui où les points sont répartis assez régulièrement autour d'une droite, comme dans les exemples précédents. On pourra dire alors que X et Y sont liés par une relation du type $Y = aX + b$ (équation de droite).

Remarque 2.2

Dans le cas où le nuage de points laisse entrevoir une dépendance entre X et Y de la forme $Y = f(X)$, on fera un "ajustement" à l'aide de la courbe $(\mathcal{C})$ représentative de f. Un tel ajustement permet d'effectuer des prévisions "dans un domaine raisonnable" (estimation des naissances dans un proche avenir, de la croissance d'une entreprise ...). Il suffira pour cela de dire que la courbe $(\mathcal{C})$ est une bonne approximation de la série statistique en une valeur choisie.

Définition 3.1

Soit la série double

X	x_1	x_2	...	x_N
Y	y_1	y_2	...	y_N

.

On appelle covariance de X et Y le réel

$$cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

Théorème 3.1

On a aussi :

$$cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N} x_i y_i - \bar{x}\bar{y}$$

Cette formule est plus pratique pour les calculs.

Remarque 3.1

La covariance caractérise les variations simultanées de deux variables : elle est positive lorsque les écarts entre les variables et leurs moyennes ont tendance à être de même signe, négative dans le cas contraire.

Définition 3.2

Lorsque $\sigma_X \neq 0$ et $\sigma_Y \neq 0$, le coefficient de corrélation linéaire r entre les variables X et Y est défini par

$$r = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sigma_X \sigma_Y}$$

Théorème 3.2

On montre que $-1 \leqslant r \leqslant 1$. Plus r est proche de -1 ou 1, plus la dépendance linéaire est forte entre les variables X et Y.
Plus r est proche de 0, plus cette dépendance est faible.

Remarque 3.2

Lorsqu'il y a une corrélation linéaire entre deux variables on peut considérer l'une des variables comme conséquence et l'autre variable comme cause. Si la variable "conséquence" est notée Y, et la variable "cause" X, il est donc normal de rechercher une fonction $y = ax + b$ qui ajuste la variable Y à partir de la variable X. C'est pour cela que la droite d'ajustement est dite droite de régression (ou d'estimation) de Y en X.

Exemple 3.1

Calculons le coefficient de corrélation linéaire de la série âge/tension.

Age X	Tension Y	x_i^2	y_i^2	$x_i y_i$
36	11,8	1296	139,24	424,8
42	14	1764	196	588
48	12,6	2304	158,76	604,8
54	15	2916	225	810
60	15,5	3600	240,25	930
66	15,1	4356	228,01	996,6
306	84	16236	1187,26	4354,2

On a $\bar{x} = 306/6 = 51$ et $\bar{y} = 84/6 = 14$.

$\sigma_X = \sqrt{\frac{1}{6} * 16\ 2336 - 51^2} = 10,25$ et

$\sigma_Y = \sqrt{\frac{1}{6} * 1\ 187,26 - 14^2} = 1,37.$

Les valeurs sont arrondies à 10^{-2} près.

Alors $r = \left(\frac{1}{6} * 4\ 354,2 - 51 * 14\right)/(10,25 * 1,37) = 0,83.$
r étant proche de 1, la corrélation linéaire est forte entre les caractères X et Y.

Exemple 3.2

Calculons le coefficient de corrélation linéaire de la série salaire/candidats.

x_i	y_i	$x_i * y_i$	x_i^2	y_i^2
22,0	10	220	484	100
22,5	13	292,5	506	169
23,0	17	391	529	289
23,5	19	446,5	552	361
24,0	21	504	576	441
115,0	80	1 854	2 648	1 360

On a $\bar{x} = 115/5 = 23$ et $\bar{y} = 80/5 = 16$.

$\sigma_X = \sqrt{\frac{1}{5} * 2\ 648 - 23^2} = 0,707$ et

$\sigma_Y = \sqrt{\frac{1}{5} * 1\ 360 - 16^2} = 4.$

Les valeurs sont arrondies à 10^{-3} près.

Alors $r = \left(\frac{1}{5} * 1\ 854 - 23 * 16\right)/(0,707 * 4) = 0,990.$
r étant très proche de 1, la corrélation linéaire est forte entre les caractères X et Y. Elle l'est davantage que dans l'exemple 3.1.

Il est donc tout-à-fait cohérent d'approcher les nuages de points représentant ces deux séries par une droite.

4. Méthode des moindres carrés

On dispose de la distribution suivante :

X	x_1	x_2	...	x_N
Y	y_1	y_2	...	y_N

Supposons que le nuage de points associé ait montré qu'il était légitime d'envisager un ajustement des données à l'aide d'une droite.

Théorème 4.1

Par la méthode des moindres carrés, la droite d'ajustement (Δ) a pour équation $y = ax + b$ où :

$$a = \frac{cov(X,Y)}{V(X)} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$ coefficient directeur de (Δ).

$b = \bar{y} - a\bar{x}$ qui traduit que $G(\bar{x}, \bar{y})$ est élément de (Δ).

Cette droite est appelée **droite de régression de Y en X** .

Théorème 4.2

On montre aussi que $a = \dfrac{\sum_{i=1}^{N} x_i y_i - N\bar{x}\bar{y}}{\sum_{i=1}^{N} x_i^2 - N\bar{x}^2}$.

Exemple 4.1

Avec les données du tableau de l'exemple 3.1 (âge/tension) on obtient :

$$a = \frac{4\ 354,2 - 6 * 51 * 14}{16\ 236 - 6 * 51^2} = 0,11 \text{ et } b = 14 - a * 51 = 8,32.$$

La droite de régression (Δ) a donc pour équation : $y = 0,11x + 8,32$.
Cette droite permet de conjecturer la tension maximale moyenne d'une personne de 70 ans par exemple : $y = 0,11 * 70 + 8,32 = 16,02$.

Exemple 4.2

Avec les données du tableau de l'exemple 3.2 (salaires/nombre de candidats) on obtient :

$$a = \frac{1\ 854 - 5 * 23 * 16}{2\ 648 - 5 * 23^2} = 5,6 \text{ et } b = 16 - a * 23 = -112,8.$$

La droite de régression (Δ) a donc pour équation :

$$y = 5,6x - 112,8.$$

Cette droite permet de conjecturer le nombre de candidats qui se seraient présentés si le salaire était de 2 600 euros par exemple :

$$y = 5,6 * 26 - 112,8 = 32,8$$

.

Exemple 4.3 : Nuages de points et droites de régressions