29 Statistique à deux variables

L'étude conjointe de deux variables statistiques sur une même population est fréquente dans le domaine des sciences exactes comme dans celui des sciences humaines.
On cherche alors à déterminer s'il existe un lien entre ces deux variables et, le cas échéant, quelle est la nature de ce lien. La première étape consiste à représenter sur un même graphique les deux variables statistiques. C'est ce que l'on appelle tracer un nuage de points. On regarde ensuite si ce nuage de points se rapproche d'une courbe connue, afin de déterminer la nature du lien (ou la **corrélation**) éventuel entre les deux variables statistiques.
La notion de corrélation semble avoir été esquissée pour la première fois par le britannique Francis Galton, (1822-1911), dans ses travaux sur l'hérédité.
En 1886, il examinait la taille des enfants en fonction de la taille moyenne des parents. Il nota que les enfants de parents de grande taille avaient tendance à être plus petits qu'eux. Il y avait donc régression du caractère "grande taille" : la droite d'ajustement de y en x qu'il utilisa fut nommée droite de régression. C'est pourquoi la droite d'ajustement affine est appelée droite de régression linéaire.

I - Statistique à une variable - Rappels

Moyenne - Variance - Écart-type

Soit X un caractère étudié dans une population d'effectif n, prenant les valeurs x_1 , x_2, x_3, $\cdots$, x_n.

Définition 1

- L'ensemble des réels x_i est appelé série statistique simple ou série statistique à une variable.
- La **moyenne** de la série statistique est le réel noté $\overline{x}$ ou $\overline{X}$ tel que :

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

- La **variance** de la série statistique est le réel noté $V(x)$ ou $V(X)$ tel que :

$$V(x) = \frac{x_1^2 + x_2^2 + \cdots + x_n^2}{n} - \overline{x}^2$$

L'**écart-type** est la racine carrée de la variance. On le note σ_x ou σ_X.

$$\sigma_x = \sqrt{V(x)}$$

Exemple 2
On considère la série de notes d'élèves de TS1, à un devoir de maths.

$Notesx_i$	7	14	13	15	10	8	9	11

On a :

$$\overline{x} = \frac{7+14+13+15+10+8+9+11}{8} = \frac{87}{8} = 10,875$$

La variance est alors :

$$V(x) = \frac{7^2+14^2+13^2+15^2+10^2+8^2+9^2+11^2}{8} - 10,875^2$$

$$V(x) = \frac{49+196+169+225+100+64+81+121}{8} - 118,26563$$

$$V(x) = \frac{1005}{8} - 118,26563$$

$$V(x) = 125,625 - 118,26563$$

$$V(x) = 7,35937$$

D'où l'écart-type :

$$\sigma_x = \sqrt{7,35937} \approx 2.713$$

II - Série statistique à deux variables

On considère dans une même population d'effectif n, deux caractères quantitatifs X et Y prenant respectivement les valeurs $x_1, x_2, \cdots, x_n$ et $y_1, y_2, \cdots, y_n$.
A chaque individu de la population , on associe un couple (x_i, y_i).
L'ensemble des couples (x_i, y_i) est appelé série statistique double ou à deux variables associée au couple de caractère (X, Y).

Exemple 3
Le tableau ci-dessous donne les notes X de maths et Y de français obtenues par 10 candidats au Bac L.

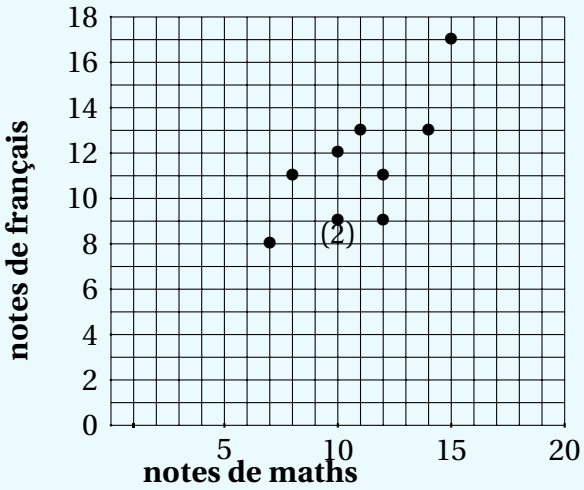
x_i	7	8	12	11	14	10	15	10	12	10
y_i	8	11	9	13	13	9	17	12	11	9

L'effectif du couple (11;13) est 1 . La fréquence du couple (11;13) est $\frac{1}{10} = 0,1$.
L'effectif du couple (10;9) est 2. La fréquence du couple (10;9) est $\frac{2}{10} = 0,2$.
Les modalités du caractère X sont : 7-8-10-11-12-14-15.
Les modalités du caractère Y sont : 8-9-11-12-13-17.

Nuage de points et point moyen

Soit (X, Y) une série statistique double.
Dans un plan muni d'un repère orthogonal, on représente les points de coordonnées (x_i, y_i).
L'ensemble de ces points est appelé **nuage** de la série double.

Exemple 4
Nuage de la série double de l'exemple précèdent.



Point moyen

Définition 5
$\overline{X}$ est la moyenne des valeurs de X et $\overline{Y}$ celle des valeurs de Y.

Le point G $(\overline{X}, \overline{Y})$ est appelé **point moyen**.

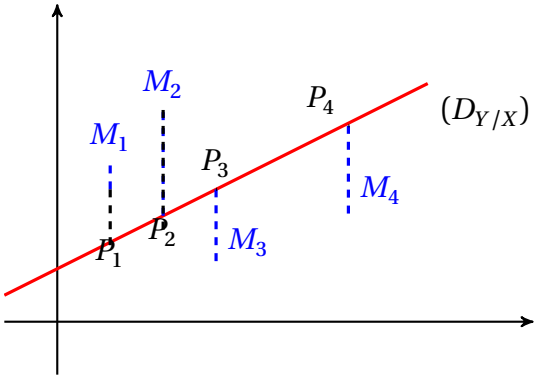
Exemple 6

$$\overline{X} = \frac{7+8+12+11+14+10+15+10+12+10}{10} = 10,9$$

$$\overline{Y} = \frac{8+11+9+13+13+9+17+12+11+9}{10} = 11,2 \qquad \text{d' où} \quad \text{G}\,(10,9;\ 11,2)$$

III - Ajustement linéaire par la méthode des moindres carrées

Lorsque le nuage de points semble présenter une forme allongée, c'est-à-dire que ses points paraissent sensiblement alignés suivant une direction de droite, cela suggère de trouver une fonction affine telle que $Y = aX + b$: on parle d'ajustement affine ou linéaire.
On utilise alors pour déterminer l'équation de la droite une méthode appelée méthode des moindres carrées, car la droite obtenue, parmi toutes les droites possibles pouvant approcher le nuage de points, est celle dont la somme des carrés des distances aux points du nuage est minimale.
Cette droite est appelée **droite de régression** de Y en X. On la note $D_{Y/X}$.



Covariance

Définition 7
Soit $\overline{X}$ et $\overline{Y}$ les moyennes des séries X et Y associées à la série double (X,Y) d'effectif n.
On appelle *covariance* de (X,Y) le réel noté $\text{cov}(X,Y)$ ou σ_{XY} défini par :

$$\text{cov}(X,Y) = \frac{x_1y_1 + x_1y_2 + \cdots + x_ny_n}{n} - \overline{XY}$$

Exemple 8
cov(X,Y)= $\frac{7\times8+8\times11+12\times9+11\times13+14\times13+10\times9+15\times17+10\times12+12\times11+10\times9}{10} - 10,9 \times 11,2 = 126,4 - 122,08 = 4,32$

Propriété 9
La droite de régression de Y en X passe par le point moyen et a pour équation :

$$y - \overline{Y} = a(x - \overline{X}) \quad \text{où} \quad a = \frac{\text{cov}(X,Y)}{V(X)}$$

Remarque 10
Cette équation permet de trouver par extrapolation à partir d'une valeur de x fixée, la valeur de y estimée et inversement.

Exemple 11
Calculons la variance de X. On a :
$V(X) = \frac{7^2+8^2+12^2+11^2+14^2+10^2+15^2+10^2+12^2+10^2}{10} - 10,9^2 = 5,49$
On a : $a = \frac{4,32}{5,49} \approx 0,787$ donc $y - 11,2 = 0.787\,(x - 10.9)$ c'est à dire $y = 0,787x + 2,622$ équation de la droite de régression de y en x.
Si cette tendance se maintient, on peut estimer la note de français d'un élève qui a eu 16 en maths.
On a ainsi : $y = 0,787 \times 16 + 2,622 = 15,214$ soit 15 en français.
Inversement quelle serait le note en maths d'un élève qui a eu 10 en français ?
Pour cela on résout l'équation d'inconnue x suivante : $10 = 0,787x + 2,622$.
On tire $x = \frac{10-2,622}{0,787} = 9,375$ soit 9,5 en maths.

Coefficient de corrélation linéaire

Lorsque les points du nuage sont groupés suivant une direction rectiligne, on a une dépendance statistique linéaire entre les caractères X et Y. On dit qu'il y a corrélation linéaire entre X et Y.

Définition 12
On appelle coefficient de corrélation linéaire d'une série statistique double $(X;Y)$, est le réel r défini par : $r = \dfrac{\text{cov}(X,Y)}{\sqrt{V(X)V(Y)}}$ ou $r = \dfrac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$

Exemple 13
On reprend l'exemple de la série statistique des notes X de maths et Y de français obtenues par 10 candidats au Bac L.

On a : $\text{cov}(X,Y) = 4,32$, $V(X) = 5,49$ et $V(Y) = 6,56$
On en déduit que : $r = \dfrac{4,32}{\sqrt{5,49 \times 6,56}} \approx 0.720$

Propriété 14

$$-1 \leq r \leq 1$$

Appréciation de la corrélation linéaire

Le réel $|r|$ permet d'apprécier la corrélation linéaire entre les variables X et Y.

- Si $0,87 \leq |r| \leq 1$ alors la corrélation linéaire entre les deux variables est forte.
- Si $|r| < 0,87$ la corrélation est faible.

Remarque 15
Si la corrélation est faible, un ajustement linéaire n'est pas justifié.