Cours : Série statistique à deux variables (série statistique double)

Les statistiques à une variable s'intéressaient, pour une population donnée, à **un** caractère donné : les notes à un devoir surveillé d'une classe, les salaires dans une entreprise, etc…

Lorsque l'on s'intéresse à l'étude simultanée de **deux** caractères d'une même population, on fait ce que l'on appelle des **statistiques à deux variables**, en étudiant **des séries statistiques doubles**.

1 Définition d'une série statistique double

Définition 1 : On considère une population d'effectif n, si on étudie deux caractères X et Y de cette population, on dit que l'on étudie une série statistique double. Chaque individu de cette population est désigné par un nombre compris entre 1 et n. A chaque individu i ($1 \leq i \leq n$) correspond un couple $(x_i ; y_i)$, ou x_i est la modalité du caractère X et y_i est la modalité du caractère Y associé à l'individu i. L'ensemble des couples $(x_i ; y_i)$ définit une **série statistique à deux variables**.

2 Quelques exemples

Exemple 1 : Le tableau ci-dessous donne, pour chaque ville, le nombre moyen d'heures d'ensoleillement dans l'année, ainsi que la température moyenne :

Ville	JENDOUBA	TUNIS	KAIROUAN	KEF	BIZERTE	AIN DRAHAM	TALA
Ensoleillement	2790	2072	2763	1729	1574	1833	1685
Température	14,7	11,4	14,2	10,8	9,7	11,2	9,7

- Population : les sept villes étudiées
- Caractère n°1 : nombre moyen d'heures d'ensoleillement dans la ville
- Caractère n°2 : température moyenne dans la ville

Exemple 2 : Le tableau ci-dessous permet de suivre l'évolution de l'espérance de vie à la naissance (en années) en Tunisie de 1990 à 1999 pour les femmes:

Année	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
Espérance de vie	80,9	81,1	81,3	81,4	81,8	81,9	82,0	82,3	82,4	82,4

- population: les femmes en TUNISIE
- Caractère n°1 : l'année
- Caractère n°2 : l'espérance de vie

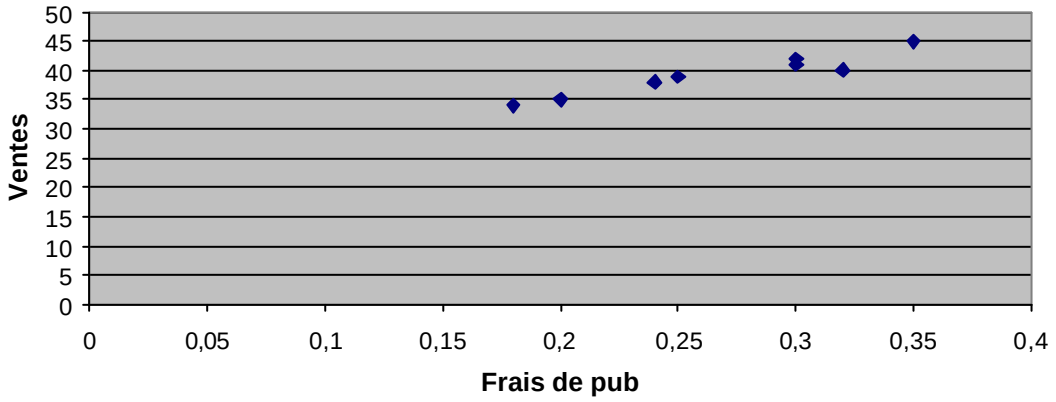
Définition 2 : Lorsque l'un des deux caractères est une année, une date, on dit que la série statistique double est une **série chronologique**.

3 Représentation graphique

Définition 3 : Si l'on appelle $x_1, x_2, \ldots, x_n$ les n valeurs du premier caractère (on notera cette série (x_i)), et si l'on appelle $y_1, y_2, \ldots, y_n$ les n valeurs du second caractère (on notera cette série (y_i)), alors on représente cette série statistique double par un **nuage de points** dans un repère du plan, constitué des points M_i de coordonnées (x_i, y_i)

Exemple 3 : Chaque mois, une entreprise consacre une somme à des opérations publicitaires. On met en regard le montant des ventes chaque mois. Une étude portant sur 8 mois a donné les résultats suivants exprimés en millions d'euros.

X=Frais de pub	0,24	0,3	0,25	0,32	0,35	0,2	0,18	0,3
Y=Ventes	38	42	39	40	45	35	34	41



Remarque 1 : De la donnée de la série statistique double, on peut déduire les séries statistiques simples décrivant séparément les caractères X et Y :

X=Frais de pub	0,18	0,2	0,24	0,25	0,3	0,32	0,35
Effectifs	1	1	1	1	2	1	1

Y=Ventes	34	35	38	39	40	41	42	45
Effectifs	1	1	1	1	1	1	1	1

Définition 4 : Soit $\overline{x}$ la moyenne de la série (x_i), et $\overline{y}$ la moyenne de la série (y_i). Le point G de coordonnées $(\overline{x}\ ;\ \overline{y})$ est appelé **point moyen** du nuage de points associé à cette série statistique double.

4 Covariance d'une série statistique double

Définition 5 : On appelle **covariance** d'une série statistique double (X ; Y) où les caractère X et Y sont quantitatifs le nombre noté cov(X, Y) ou σ_{XY} défini par :

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

Où $\overline{x}$ et $\overline{y}$ sont les moyennes des séries statistiques simples.

Théorème de Huyghens-König :

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x}\cdot\overline{y}$$

Propriété 1 : Soient α, β, α', β' des constantes réelles , U et V les caractères statistiques définis par: $U = \alpha X + \beta$ et $V = \alpha' X + \beta'$. C'est-à-dire tels que pour tout i tel que $1 \leq i \leq n$: $u_i = \alpha x_i + \beta$ et $v_i = \alpha' y_i + \beta'$

Alors : $\text{cov}(U, V) = \alpha\alpha' \text{cov}(X, Y)$

5 Ajustement affine

5.1 Principe

Soit (x_i, y_i) une série statistique double, avec un nuage de points $M_i(x_i, y_i)$ associé.

Lorsque les points du nuage paraissent presque alignés, on peut chercher une relation de la forme $y = ax + b$ qui exprime de façon approchée les valeurs de la série (y_i) en fonction des valeurs de la série (x_i) , autrement dit, une fonction affine f telle que l'égalité $y = f(x)$ s'ajuste au mieux avec les données.

Graphiquement, cela signifie qu'on cherche **une droite qui passe au plus près de tous les points du nuage**. Une telle relation permettrait notamment de faire des **prévisions**. Il existe de nombreuses manières d'obtenir un ajustement affine satisfaisant.

5.2 Méthodes

5.2.1 Méthode « au jugé »

A vous de tracer une droite qui passe le plus près possible de tous les points du nuage, si possible en la faisant passer par le point moyen du nuage. C'est peu précis, mais peut suffire dans certains cas.

5.2.2 Méthode de Mayer

Etape 1: On commence par « découper » la série statistique double en deux sous-séries bien distinctes, c'est-à-dire que l'on découpe le nuage de points M_i (x_i, y_i) en deux sous-nuages distincts et de même effectif (ou presque : si le nombre de points est pair, pas de souci. S'il est impair, on peut mettre le point surnuméraire dans n'importe lequel des deux sous-nuages)

Etape 2 : On calcule les coordonnées des deux points moyens G_1 et G_2 associés à ces deux sous-nuages, et on place ces deux points sur le graphique.

Etape 3 : On trace la droite $(G_1\ G_2)$?appelée **droite de Mayer** du nuage de points M_i (x_i, y_i), qui doit passer par le point moyen G du nuage de points M_i (x_i, y_i) .C'est cette droite qui constitue un ajustement affine tout à fait acceptable pour la série double (x_i, y_i)

5.2.3 Méthode des moindres carrés

5.2.3.1 Principe

On considère un nuage de points M_i (x_i, y_i) et soit (D) une droite d'équation $y = ax + b$ que l'on cherche à déterminer.

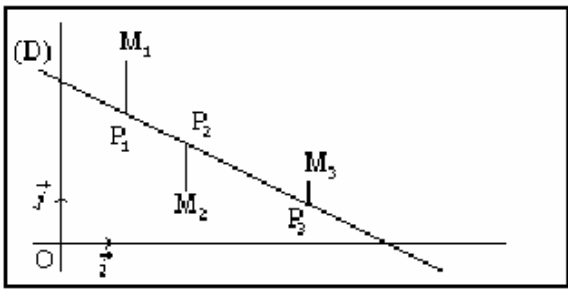
Définition 6 : On appelle **somme des résidus** associée à la droite (D), le nombre réel S défini par :

$$S = \sum_{i=1}^{n} \left(y_i - (a x_i + b)\right)^2$$

Si P_i désigne le point d'abscisses x_i sur la droite (D), on a :

$$S = \sum_{i=1}^{n} M_i P_i^2$$

Définition 7 : On appelle **méthode des moindres carrés** la méthode qui consiste à rechercher les coefficients *a* et *b* tels que la somme *S* soit minimale. Remarquons que *S* est une fonction des deux variables *a* et *b*.



5.2.3.2 Détermination des coefficients

Théorème 1 : Le nombre S est minimum pour

$$a = \frac{\mathrm{cov}(X,Y)}{V(X)} = \frac{\sigma_{XY}}{{\sigma_X}^2}$$

$$b = \overline{y} - a\overline{x}$$

Proposition 1 : La droite (D) d'équation y = ax+b où a et b sont déterminés d'après le théorème 1, est appelé **droite de régression de Y en X** et on dit qu'on a obtenu cette équation par la méthode des moindres carrés.

Proposition 2 : La droite (D') d'équation : x=a'y+b' avec :

$$a' = \frac{\mathrm{cov}(X,Y)}{V(Y)} = \frac{\sigma_{XY}}{{\sigma_Y}^2}$$

$$b' = \overline{x} - a'\overline{y}$$

est appelée droite de **droite de régression de X en Y** et on dit qu'on a obtenu cette équation par la méthode des moindres carrés.

Remarque : Les deux droites de régression de Y en X et de X en Y passent toutes deux par le point moyen de coordonnées ($\overline{x}$; $\overline{y}$)

6 Coefficient de corrélation linéaire

Définition 8 : On appelle **coefficient de corrélation linéaire** du couple (X, Y), le nombre réel, noté *r(X, Y)* tel que :

$$r(x,y) = \frac{Cov(x,y)}{\sigma_x \sigma_y}$$

Remarques :

- $-1 \leq r(X,Y) \leq 1$
- $aa' = (r(X,Y))^2$
- Lorsque la corrélation est forte ($r^2 \geq ¾$) les droites de régression sont très proches et le nuage peut être approximé par une droite.
- Lorsque la corrélation est faible, le nuage de points ne peut pas être ajusté par une droite, mais il se peut qu'une autre courbe permette un bon ajustement.