LECON N°7 Exposé 8

Séries statistiques à 2 variables numériques. Nuage de pts associé. Ajustement affine par la méthode des moindres carrés. Droites de regression. Applications

Niveau : TES avec des outils de TS

Prérequis : - notion de var
- $n \in \mathbb{N}^*$, $\Omega = \{\omega_1, \dots, \omega_n\}$ population de n individus
X var sur Ω. $\forall i \in \Delta_n$, $x_i = X(\omega_i)$
$\mu(X) = \bar{X} = \frac{1}{n} \sum_{i=1}^{n} x_i$
$V(X) = \sigma^2(X) = \mu((X-\bar{X})^2) = \mu(X^2) - (\mu(X))^2$

1. Séries statistiques à 2 variables numériques

11. Exemple introductif

On considère 5 élèves $1 \leq x_i \leq 5$

Note x_i au bac	7	10	11	13	16
Note y_i au concours	8	9	12	12	13

$\bar{X} = 11,4$ $V(X) = 9,04$ $\sigma(X) = 3,01$
$\bar{Y} = 10,8$ $V(Y) = 3,76$ $\sigma(Y) = 1,94$

12 Définition

$n \in \mathbb{N}^*$ $\Omega = \{\omega_1, \dots, \omega_n\}$ une population.
On appelle variable statistique quantitative (ou caractère quantitatif) toute application X Ω dans $\mathbb{R}$ On fait X est une var.
Une série statistique est la donnée d'un tableau où X et Y ont même ensemble Ω

X	x_1			x_n
Y	y_1			y_n

Rq Il existe des séries statistiques qui ont + de 2 caractères

1.3 Nuage de points associé

Définition Dans un repère orthogonal, $\{M_i(x_i, y_i), 1 \leq i \leq n\}$ est le nuage de points associé à la série.
$G(\bar{x}, \bar{y})$ est le point moyen du nuage

1.4 Covariance

La covariance permet d'étudier la dépendance entre X et Y
$Cov(X,Y) = \mu[(X-\bar{X})(Y-\bar{Y})] = \mu(XY) - \mu(X)\mu(Y)$
Dans l'exe. $Cov(X,Y) = 5,28$

2. Ajustement affine par la méthode des moindres carrés

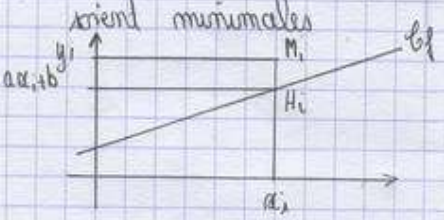
X et Y non constantes ie $\sigma(X) \neq 0$ et $\sigma(Y) \neq 0$

2.1 Principe

Soit $f: \mathbb{R} \to \mathbb{R}$, $\mathcal{C}_f : y = f(x)$ une fonction affine. f réalise un ajustement affine du nuage de points si $\mathcal{C}_f$ passe "le plus près possible" des points du nuage.

Méthode des moindres carrés

Cette méthode consiste à chercher a et b tels que $\mu(Z)$ et $\sigma^2(Z)$ soient minimales



$\mu(Z^2) = \mu^2(Z) + \sigma^2(Z)$
$Z = Y - aX - b$
$\mu(Z^2) = \frac{1}{n}\sum_{i=1}^{n}(y_i - ax_i - b)^2$
(somme des résidus en Y)

en fait on veut minimiser $\mu(Z^2)$
$\mu(Z^2) = \sum_{i=1}^{n} M_iH_i^2$

Rq Il existe d'autres méthodes pour minimiser notamment la méthode de la droite de Mayer

2.2 Théorème

Si $Z = Y - aX - b$ alors :

$\mu(Z)$ et $\sigma^2(Z)$ minimales $\iff$ $a = \frac{cov(X,Y)}{\sigma^2(X)}$ et $b = \overline{Y} - a\overline{X}$

Démonstration

$\mu(Z) = \mu(Y - aX - b) = \overline{Y} - a\overline{X} - b$

$Z - \overline{Z} = (Y - \overline{Y}) - a(X - \overline{X})$

donc $(Z - \overline{Z})^2 = (Y - \overline{Y})^2 - 2a(Y - \overline{Y})(X - \overline{X}) + a^2(X - \overline{X})^2$

$\sigma^2(Z) = \mu[(Z - \overline{Z})^2] = \sigma^2(Y) - 2a\,cov(X,Y) + a^2\sigma^2(X)$

$$= \underbrace{\left[a\sigma(X) - \frac{cov(X,Y)}{\sigma(X)}\right]^2}_{\geq 0} + \underbrace{\sigma^2(Y) - \frac{cov(X,Y)}{\sigma^2(X)}}_{c \geq 0}$$

$$\mu(Z^2) = (\overline{Y} - a\overline{X} - b)^2 + \left[a\sigma(X) - \frac{cov(X,Y)}{\sigma(X)}\right] + c$$

$$\mu(Z^2) \text{ minimale} \iff \begin{cases} a = \frac{cov(X,Y)}{\sigma^2(X)} \\ b = \overline{Y} - a\overline{X} \end{cases}$$

2.3 Droite de regression

D : $y = ax + b$ est la droite de regression de Y en X. D passe par le pt moyen.

Cqs : Elle permet d'estimer y en fct de x donné et inversement, on peut calculer la droite de regression de X en Y.

2.4 Coefficient de corrélation linéaire entre X et Y

le coefficient permet de mesurer la validité de l'ajustement affine, il est égal à $r(X,Y) = \frac{cov(X,Y)}{\sigma(X)}$

Rq On obtient une bonne corrélation si l'angle des droites de regression est $\leq 30°$

Propriété

On a $-1 \leq r(x,y) \leq 1$

$|r(x,y)| = 1 \iff$ les pts M_i sont alignés.

Dans l'ex : $r(x,y) = 0{,}91$

forte corrélation car $|r(x,y)| > 0{,}5$.

3. Application

3.1 Déterminer les droites de Y en X et de X en Y

- pour l'ex de la leçon
- si on considère un nuage pour les pts A(1,1) B(2,2) D(3,1) C(4,2)

3.2 Le tableau donne l'évolution des taux d'équipement y_i (en Y) en automobiles des familles françaises.

Année x_i	1975	1980	1985	1988	1991
taux y_i	64,1	69,3	73,4	74,6	76,8

Dans un repère représenter le nuage de pts associés à la série

La forme de ce nuage permet-elle d'envisager raisonnablement un ajustement affine ?

Donner les éq. des droites de régression de Y en X et de X en Y

les représenter.

Que constatez-vous expliquer ?

Si l'évolution se poursuit ainsi estimer le taux en 1993.

et l'année à partir de laquelle ce taux sera $\geq 85\%$.