Chapitre 5

Etude de corrélation et régression

Contenu

Série statistique double

Définition 1. *Lorsqu'on étudie deux caractères statistiques sur une population donnée on obtient une série statistique double.*

Nuage de points

Définition 2. *L'ensemble de points M_i de coordonnées (x_i, y_i).*

On distingue trois cas

5.1 Cas 1 : tableau de deux lignes

x_i	x_1	x_2	...	x_n
y_i	y_1	y_2	...	y_n

Définition 3. *Les moyennes marginales*

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{n} x_i$$

$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{n} y_i$$

Les variances marginales

$$V(X) = \left(\frac{1}{N}\sum_{i=1}^{n} x_i^2\right) - \bar{X}^2$$

$$V(Y) = \left(\frac{1}{N}\sum_{i=1}^{n} y_i^2\right) - \bar{Y}^2$$

Les écarts-types marginaux

$$\delta_X = \sqrt{V(X)}$$

$$\delta_Y = \sqrt{V(Y)}$$

La covariance de X et Y

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

5.1.1 Droite de régression(méthode des moindres carrées)

Théorème 5.1.1. *La droite de régression de Y en fonction de X notée par $D_Y(X)$ a pour équation $Y = aX + b$ tels que*

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

$$b = \bar{Y} - a\bar{X}$$

Propriété 5.1.1. ❶ *C'est une droite unique*

❷ *Elle passe toujours par le point moyen $(\bar{X}, \bar{Y})$*

5.1.2 Coefficient de corrélation linéaire

Définition 4. *Le coefficient de corrélation linéaire d'une série statistique double est le nombre*

$$r = \frac{cov(X,Y)}{\delta_X \delta_Y}$$

Remarque 5.1.1. ❶ $-1 \leq r \leq 1$

❷ *Si $r = 0$ alors il n y a pas de corrélation entre X et Y (X et Y sont indépendantes).*

❸ *Si $0 < r < 1$ alors il y a une corrélation positive faible, moyenne ou forte entre X et Y .*

❹ *Si $-1 < r < 0$ alors il y a une corrélation négative faible, moyenne ou forte entre X et Y .*

Example 5.1. *On considère la série double suivante*

x_i	2	5	6	10	12
y_i	83	70	70	54	49

Les moyennes marginales

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{n} x_i = 7 \quad \bar{Y} = \frac{1}{N}\sum_{i=1}^{n} y_i = 65.2$$

Les variances marginales

$$V(X) = \left(\frac{1}{N}\sum_{i=1}^{n} x_i^2\right) - \bar{X}^2 = 12.8 \quad V(Y) = \left(\frac{1}{N}\sum_{i=1}^{n} y_i^2\right) - \bar{Y}^2 = 150.16$$

Les écarts-types marginaux

$$\delta_X = \sqrt{V(X)} = 3.578 \quad \delta_Y = \sqrt{V(Y)} = 12.25$$

La covariance de X et Y

$$cov(X,Y) = \left(\frac{1}{N}\sum_{i=1}^{n} x_i y_i\right) - \bar{X}\bar{Y} = -43.6$$

La droite de régression de Y en fonction de X, $Y = aX + b$ tels que

$$a = \frac{cov(X,Y)}{V(X)} = -3.4$$

$$b = \bar{Y} - a\bar{X} = 89$$

donc $Y = -3,4X + 89$

Le coefficient de corrélation linéaire

$$r = \frac{cov(X,Y)}{\delta_X \delta_Y} = -0.99$$

donc il y a une corrélation linéaire négative forte entre X et Y.

5.2 Cas 2 : tableau de trois lignes

x_i	x_1	x_2	...	x_k
y_i	y_1	y_2	...	y_k
n_i	n_1	n_2	...	n_k

Définition 5. *Les moyennes marginales*

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{k} n_i x_i$$

$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{k} n_i y_i$$

Les variances marginales

$$V(X) = \left(\frac{1}{N}\sum_{i=1}^{k} n_i x_i^2\right) - \bar{X}^2$$

$$V(Y) = \left(\frac{1}{N}\sum_{i=1}^{k} n_i y_i^2\right) - \bar{Y}^2$$

Les écarts-types marginaux

$$\delta_X = \sqrt{V(X)}$$

$$\delta_Y = \sqrt{V(Y)}$$

La covariance de X et Y

$$cov(X,Y) = \left(\frac{1}{N}\sum_{i=1}^{k} n_i x_i y_i\right) - \bar{X}\bar{Y}$$

5.3 Cas 3 : tableau de contingence

XY	y_1	y_2	...	y_l
x_1	n_{11}	n_{12}	...	n_{1l}
x_2				
$\vdots$				
x_k	n_{k1}			n_{kl}

Définition 6. *Les distributions marginales*

x_i	x_1	x_2	...	x_k
n_i	n_1	n_2	...	n_k

y_j	y_1	y_2	...	y_l
n_j	n_1	n_2	...	n_l

Les moyennes marginales

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{k} n_i x_i \qquad \bar{Y} = \frac{1}{N}\sum_{j=1}^{l} n_j y_j$$

Les variances marginales

$$V(X) = \left(\frac{1}{N}\sum_{i=1}^{k} n_i x_i^2\right) - \bar{X}^2 \qquad V(Y) = \left(\frac{1}{N}\sum_{j=1}^{l} n_j y_j^2\right) - \bar{Y}^2$$

Les écarts-types marginaux

$$\delta_X = \sqrt{V(X)} \qquad \delta_Y = \sqrt{V(Y)}$$

La covariance de X et Y

$$cov(X,Y) = \left(\frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij} x_i y_i\right) - \bar{X}\bar{Y}$$

Example 5.2. *On considère la série double suivante*

XY	1	2	4	n_i
3	2	0	3	5
5	4	6	1	11
6	5	1	7	13
n_j	11	7	11	N=29

Les distributions marginales

x_i	3	5	6
n_i	5	11	13

y_j	1	2	4
n_j	11	7	11

Les moyennes marginales

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{k} n_i x_i = 5,10 \qquad \bar{Y} = \frac{1}{N}\sum_{j=1}^{l} n_j y_j = 2.38$$

Les variances marginales

$$V(X) = \left(\frac{1}{N}\sum_{i=1}^{k} n_i x_i^2\right) - \bar{X}^2 = 1.13 \qquad V(Y) = \left(\frac{1}{N}\sum_{j=1}^{l} n_j y_j^2\right) - \bar{Y}^2 = 1.75$$

Les écarts-types marginaux

$$\delta_X = \sqrt{V(X)} = 1.06 \qquad \delta_Y = \sqrt{V(Y)} = 1.32$$

La covariance de X et Y

$$cov(X,Y) = \left(\frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij} x_i y_i\right) - \bar{X}\bar{Y} = 0$$