

CHAPITRE A3 - STATISTIQUES BIVARIÉES

1 Définitions et rappels

1.1 Série statistique double

Définition

Soit une population Ω et deux variables discrètes X et Y relatives à cette population.

Considérons $(x_i)_{i \in \llbracket 1, n \rrbracket}$ la série statistique des modalités de X observées sur un échantillon de taille n de la population $\Omega = (\omega_1, \dots, \omega_n)$ et $(y_i)_{i \in \llbracket 1, n \rrbracket}$ la série statistique des modalités de Y observées sur le même échantillon.

On appelle **série statistique double** la donnée de la liste $((x_i, y_i))_{i \in \llbracket 1, n \rrbracket}$, chaque (x_i, y_i) étant associé à un seul individu ω_i de la population.

Exemple : Par exemple, on peut collecter la taille et le poids de personnes dans une population.

1.2 Moyenne, variance

Définition : Moyenne empirique

$\bar{X}$ désigne la moyenne de X :

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

Définition : Variance empirique

s_X^2 désigne la variance de X :

$$s_X^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{X})^2.$$

Propriété : formule de Kœnig-Huygens.

$$s_X^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{X}^2.$$

2 Covariance et corrélation

2.1 Covariance

Définition : Covariance empirique

On appelle **covariance empirique** de la série statistique double $(x_i, y_i)_{i \in \llbracket 1, n \rrbracket}$ le réel :

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

Propriété : formule de Kœnig-Huygens.

$$s_{X,Y} = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{X} \bar{Y}.$$

2.2 Corrélation

Définition : Coefficient de corrélation linéaire

Le **coefficient de corrélation linéaire** de la série statistique double est le réel :

$$r_{X,Y} = \frac{s_{X,Y}}{s_X s_Y}$$

Remarques :

- s_X et s_Y sont les écart-types de X et Y , parfois notés σ_X et σ_Y . On a : $s_X = \sigma_X = \sqrt{s_X^2}$.
- On a toujours $|s_{X,Y}| \leq s_X s_Y$. Donc $|r_{X,Y}| \leq 1$.
- Si $r_{X,Y}$ est proche de ± 1 , alors X et Y sont fortement corrélées.
- Si $r_{X,Y}$ est proche de 0 alors X et Y sont faiblement (ou pas) corrélées.

3 Nuage de points, ajustement affine

3.1 Nuage de points

Définition : Nuage de points

Le **nuage de points** représentant la série statistique double $((x_i, y_i))_{i \in \llbracket 1, n \rrbracket}$ est l'ensemble des points de coordonnées (x_i, y_i) .

Définition : Point moyen

On appelle **point moyen** le point de coordonnées $(\bar{X}, \bar{Y})$.

3.2 Ajustement affine**Proposition : Méthode des moindres carrés**

L'unique droite rendant minimal $\sum_{i=1}^n (y_i - ax_i - b)^2$ est la droite d'équation :

$$y = \frac{s_{X,Y}}{s_X^2}(x - \bar{X}) + \bar{Y}.$$

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

Remarques :

- X est appelée la variable explicative, Y la variable expliquée.
- La droite de régression linéaire passe par le point moyen.
- Si $|r_{X,Y}|$ est proche de 1, alors les points du nuage sont proches de l'alignement. Si $|r_{X,Y}| = 1$ alors les points seront exactement sur la droite de régression.
- Si $r_{X,Y} > 0$ alors la droite est de pente positive. X et Y varient dans le même sens.
- Si $r_{X,Y} < 0$ alors la droite est de pente négative. X et Y varient dans des sens contraires.
- **Attention !** Corrélation ne vaut pas dire causalité !

3.3 Ajustement logarithmique, exponentiels et autres**Méthode**

Il arrive que certains nuages de points prennent une forme qui nous fait penser à une fonction de référence (exponentielle, logarithme...). Dans ce cas, l'énoncé nous invitera à "redresser" le nuage en appliquant une fonction adaptée à l'une ou l'autre des séries statistiques étudiée pour travailler avec un nuage, qui lui aura une forme rectiligne.