Chapitre 5-Statistiques bivariées

1 Rappels

Vocabulaire

- L'ensemble Ω des éléments dont on étudie les données est appelé **population**. Ses éléments sont appelés les **individus**.
- Un **échantillon** est un sous-ensemble (fini) de la population.
- L'**effectif** d'une population, d'un échantillon est le nombre d'individus de cette population, cet échantillon.

Définition 1 (Variable, série statistique)

- Une **variable** (ou caractère) est une application définie sur la population Ω. Une variable est dite **quantitative** lorsqu'elle est à valeurs réelles et **qualitative** sinon.
- Les valeurs prises par X sont appelées les **modalités**.
- La liste des valeurs prises par X est appelée une **série statistique**.

On peut représenter une série statistique de deux manières :

1. en donnant la liste « brute » des valeurs prises par X : $[\mathrm{X}(\omega), \omega \in \Omega]$;
2. en donnant la liste des modalités **distinctes** prises par X affectées de leur effectif d'apparition (on dira que la série est groupée par modalités).

2 Série statistique à deux variables

2.1 Généralités

Définition 2 (Série statistique à deux variables)

Soient $\{\omega_1, \ldots, \omega_n\}$ un échantillon d'une population Ω et $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n)$ deux séries statistiques représentant les valeurs prises par deux caractères X et Y sur cet échantillon :

$$\forall i \in [\![1, n]\!] \quad x_i = \mathrm{X}(\omega_i) \quad \text{et} \quad y_i = \mathrm{Y}(\omega_i).$$

On appelle série statistique double la donnée de la liste

$$\big((x_1, y_1), \ldots, (x_n, y_n)\big).$$

Chaque couple est associé à un seul individu de l'échantillon : pour tout $i \in [\![1, n]\!]$, $(x_i, y_i) = (\mathrm{X}(\omega_i), \mathrm{Y}(\omega_i))$.

Lorsqu'on étudie un couple de caractères (X, Y), il est fréquent de se demander si l'un des deux caractères (par exemple X) est une cause de l'autre (par exemple Y). Dans ce cas, on dit que X est la **variable explicative** et que Y est la **variable à expliquer**.

Exemple 1

Un sociologue veut analyser s'il existe une relation entre le taux de criminalité dans des villes et la densité de population de ces villes. Entre la densité de population et le taux de criminalité, que choisiriez-vous comme variable explicative X *et comme variable à expliquer* Y *?*

2.2 Nuage de points

Définition 3 (Nuage de points)

On se place dans un repère du plan.
On appelle **nuage de points** associé à une série statistique double $\big((x_1,y_1),\dots,(x_n,y_n)\big)$ l'ensemble des points M_i de coordonnées (x_i,y_i) pour $i=1,\dots,n$.

Rappel(s) 1

Étant donnée un série statistique $x=(x_1,\dots,x_n)$ associée à un caractère X, *sa* ***moyenne*** *, notée $\bar{x}$, est le réel défini par :*

$$\bar{x}=\frac{1}{n}\sum_{k=1}^{n}x_k.$$

Définition 4 (Point moyen)

Soit $\big((x_1,y_1),\dots,(x_n,y_n)\big)$ une série statistique double associée à deux caractères X et Y.
On appelle **point moyen** le point de coordonnées $(\bar{x},\bar{y})$ où $\bar{x}$ est la moyenne de la série $(x_1,\dots,x_n)$ et $\bar{y}$ la moyenne de la série $\big(y_1,\dots,y_n\big)$.

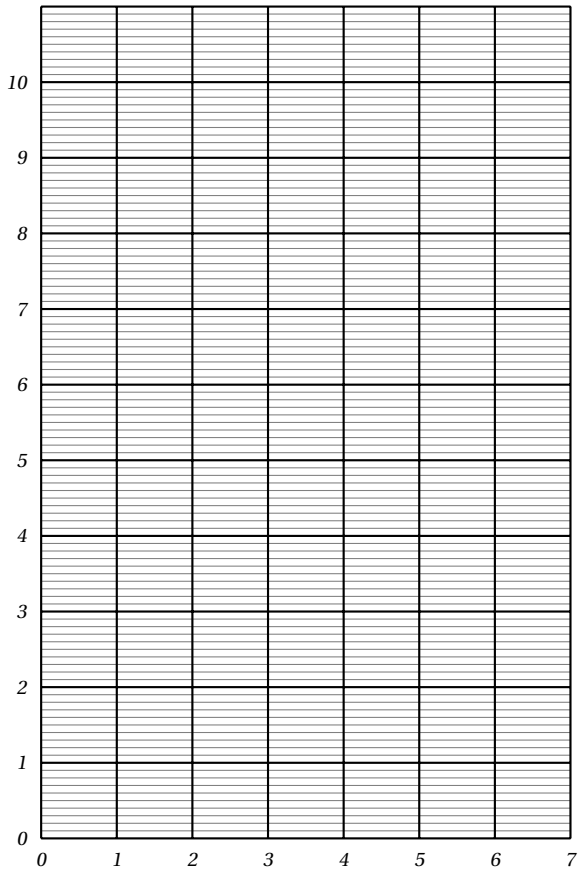
Exemple 2

On considère la série statistique à deux variables suivante :

x_i	0	1	2	3	4	5	6
y_i	4.1	5.0	5.9	7.4	8.3	9.7	10.7

1. *Déterminer les coordonnées du point moyen.*

2. *Tracer le nuage de point dans le repère ci-dessous.*



2.3 Covariance et corrélation

Rappel(s) 2

Étant donnée un série statistique $x = (x_1, \dots, x_n)$. On définit :

1. *sa **variance empirique**, notée s_x^2, par*

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{k=1}^{n} x_i^2 - \bar{x}^2.$$

2. *son **écart-type**, noté σ_x par :*

$$\sigma_x = \sqrt{s_x^2}.$$

Définition 5 (Covariance, coefficient de corrélation linéaire)

Étant donné une série statistique double associée à des séries $x = (x_1, \dots, x_n)$ et $y = (y_1, \dots, y_n)$.

- On appelle **covariance empirique** de la série double le réel :

$$s_{x,y}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})\left(y_i - \bar{y}\right).$$

- Lorsque σ_x et σ_y sont non nuls, on appelle **coefficient de corrélation linéaire** de la série double le réel :

$$\rho_{X,Y} = \frac{s_{x,y}^2}{\sigma_x \sigma_y}.$$

Proposition 1 (Formule de Koenig-Huygens)

Soient $x = (x_1, \ldots, x_n)$ et $y = (y_1, \ldots, y_n)$ deux séries statistiques. Alors :

$$s^2_{x,y} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}.$$

Démonstration : Soient $x = (x_1, \ldots, x_n)$ et $y = (y_1, \ldots, y_n)$ deux séries statistiques. Alors :

■

Théorème

Soient $x = (x_1, \ldots, x_n)$ et $y = (y_1, \ldots, y_n)$ deux séries statistiques. Alors :

$$\left(s^2_{x,y}\right)^2 \leqslant s^2_x s^2_y \quad \text{ou encore} \quad \left|s^2_{x,y}\right| \leqslant \sigma_x \sigma_y.$$

En particulier, lorsqu'il est défini le coefficient de corrélation linéaire vérifie :

$$\left|\rho_{x,y}\right| \leqslant 1.$$

Démonstration :

- Si l'une des deux séries est de variance nulle, disons x alors :

- Soient $x = (x_1, \ldots, x_n)$ et $y = (y_1, \ldots, y_n)$ deux séries statistiques de variances non nulles.

3 Régression

3.1 Généralités

On considère un couple de variables (X,Y) où X est la variable explicative et Y la variable à expliquer.
Déterminer un **modèle de régression** consiste à savoir si Y est une fonction de X :

$$Y = f(X)$$

où f est une fonction. En pratique, il est très rare que Y soit une fonction de X mais on peut souvent approcher Y par une variable aléatoire qui est une fonction de X :

$$Y = f(X) + \varepsilon$$

où la fonction f est appelée **la fonction de régression** et la variable aléatoire ε est appelée **l'erreur d'ajustement**.
Dans la suite, on s'intéressera à des modèles de **régression linéaire** c'est-à-dire modèle où la fonction f est **affine** :

$$f(X) = aX + b.$$

Plus précisément, on va chercher la droite qui « colle au mieux » au nuage de points.

3.2 Moindres carrés

Si le nuage de points associé à une série statistique double possède une forme étirée, on peut chercher une droite qui approche le mieux possible les points de ce nuage. Le problème consiste donc à déterminer une droite $y = ax + b$ qui ajuste bien le nuage de points.
Si on approche notre nuage par une droite $y = ax + b$, l'erreur que l'on commet en utilisant la droite de régression pour prédire y_i à partir de x_i est $y_i - (ax_i + b)$.

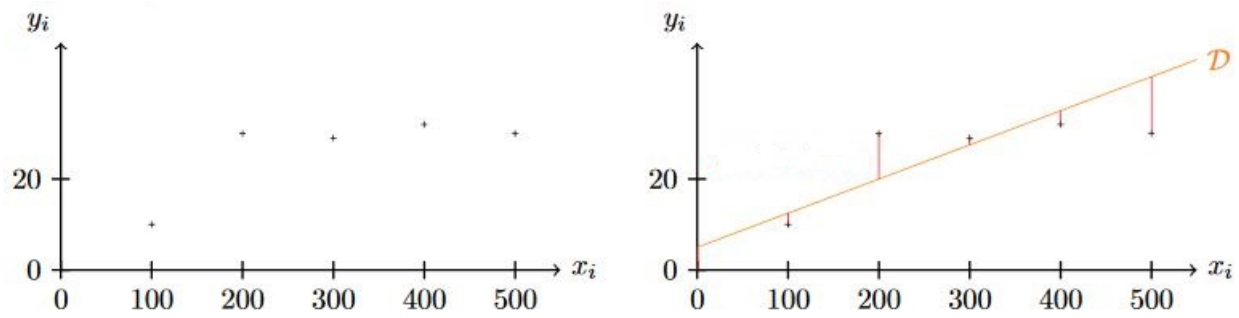


FIGURE 1 – À gauche : nuage de points ; à droite : le même nuage avec une droite d'ajustement et les erreurs (en rouge)

La **méthode des moindres carrés** consiste à trouver (s'ils existent) les coefficients a et b qui minimisent **l'erreur quadratique** :

$$\sum_{i=1}^{n} \big(y_i - (ax_i + b)\big)^2.$$

Théorème (Droite de régression linéaire)

Étant donné une série statistique double $\big((x_1, y_1), \ldots, (x_n, y_n)\big)$, l'unique droite minimisant l'erreur quadratique est la droite d'équation :

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

Cette droite est appelé **la droite de régression linéaire de y en x**.

Démonstration : On considère la fonction F de deux variables a et b :

$$\begin{aligned} F : \mathbb{R}^2 &\longrightarrow \mathbb{R} \\ (a,b) &\longmapsto \sum_{i=1}^{n} \big(y_i - (ax_i + b)\big)^2. \end{aligned}$$

Le but est de montrer que cette fonction possède un minimum et de le trouver. Nous verrons comment faire au second semestre. ■

Proposition 2

Soit $\big((x_1, y_1), \ldots, (x_n, y_n)\big)$ une série statistique double. La droite de régression linéaire de y en x passe par le point moyen.

3.3 Lien avec le coefficient de corrélation linéaire

Proposition 3

Étant donné une série statistique double $\big[(x_1, y_1), \ldots, (x_n, y_n)\big]$, on note $\rho_{X,Y}$ le coefficient de régression linéaire.

- Plus $|\rho_{X,Y}|$ est proche de 1 plus les points du nuage sont proches de l'alignement et plus les prévisions données par les droites de régression sont pertinentes.
- $|\rho_{X,Y}|$ vaut 1 si et seulement si les points du nuage sont alignés.
- Si $\rho_{X,Y} > 0$ la pente de la droite de régression est positive ; si $\rho_{X,Y} < 0$ la pente de la droite de régression est négative.

Démonstration :

■

4 Objectifs

- Savoir déterminer le moyenne, la variance d'une série statistique.
- Savoir déterminer le point moyen et tracer le nuage de points d'une série statistique à deux variables.
- Savoir déterminer la covariance et le coefficient de corrélation linéaire d'une série statistique à deux variables.
- Savoir déterminer la droite de régression linaire d'une série statistique à deux variables et le lien avec le coefficient de corrélation linéaire.