24

Statistique descriptive

24.1 Statistique univariée

24.1.1 Définitions

Une étude statistique univariée consiste à traiter des données mesurées, qu'elles soient quantitatives (âge, taille, poids, précipitations, températures, longueurs, sommes d'argent...) ou qualitatives (sexe, couleur, opinion...). Ces données sont apelées des caractères, ou des variables.

Définition 1

• L'ensemble d'objets ou de personnes sur lequel porte une étude statistique est appelé population.
• Un élément de cette population est appelé un individu.
• L'effectif d'une population est le nombre d'individus total de cette population.
• La fréquence d'un caractère est le nombre d'individus possédant ce caractère divisé par l'effectif total de la population.
• Une variable est dite discrète si elle ne prend que des valeurs isolées (par exemple des nombres entiers, comme pour les âges).
• Une variable est dite continue si elle peut prendre toutes les valeurs d'un intervalle (comme pour les précipitations, la température, ou plus généralement toute mesure d'une grandeur physique).

24.1.2 Description d'une série statistique discrète

Considérons un effectif de taille n et une série statistique $(x_1, \dots, x_n)$ portant sur un caractère discret x, où pour tout $i \in [\![1, n]\!]$, x_i désigne la valeur du caractère mesurée pour l'individu i.

On peut présenter les données sous forme d'un tableau en regroupant les différentes valeurs du caratère prises par la population : si $\{x_1, \dots, x_n\} = \{y_1, \dots, y_p\}$ avec $y_1 < y_2 < \dots < y_p$, alors on peut écrire

Caractère	Effectif
y_1	n_1
y_2	n_2
$\dots$	$\dots$
y_p	n_p

où pour tout $i \in [\![1,p]\!], n_i$ est le nombre d'individus pour lesquels le caractère prend la valeur y_i et $\sum_{k=1}^{p} n_k = n$.

On représente généralement ces résultats sous la forme d'un diagramme à bâtons.

Définition 2: Fréquence

Avec les notations précédentes, la fréquence du caractère y_i est $f_i = \dfrac{n_i}{n}$.

Remarque 1. On a alors $\sum_{i=1}^{p} f_i = 1$.

Exemple 1. Dans une classe de 15 élèves, on relève les notes suivantes au dernier devoir de mathématiques :

Note	Nombre d'élèves
10	4
12	6
14	4
18	1

On a alors les fréquences suivantes :

Note	Fréquence
10	$\frac{4}{15}$
12	$\frac{2}{5}$
14	$\frac{4}{15}$
18	$\frac{1}{15}$

Définition 3: Effectifs et fréquences cumulés croissants

• Avec les notations précédentes, on définit pour tout $i \in [\![1,p]\!]$ les effectifs cumulés croissants

$$N_i = \sum_{k=1}^{i} n_i.$$

Autrement dit, N_i représente le nombre d'invidus pour lesquels la valeur du caractère est inférieure (ou égale) à y_i.

• On définit de même les fréquences cumulés croissants

$$F_i = \sum_{k=1}^{i} f_i.$$

Exemple 2. Avec l'exemple précédent, on obtient

Note	Effectifs cumulés croissants
10	4
12	10
14	14
18	15

Note	Fréquences cumulées croissantes
10	$\frac{4}{15}$
12	$\frac{2}{3}$
14	$\frac{14}{15}$
18	1

24.1.3 Description d'une série statistique continue

Prenons l'exemple de la taille d'élèves dans une classe de 14 élèves.

Taille (cm)	Effectif
$[150, 160[$	2
$[160, 170[$	6
$[170, 180[$	5
$[180, 190[$	0
$[190, 200[$	1

On dit qu'on a regroupé les caractères par classes.

On représente généralement ces résultats sous la forme d'un histogramme.

On définit alors comme précédemment les effectifs cumulés croissants et les fréquences cumulées croissantes.

24.1.4 Caractéristiques empiriques de position

Définition 4: Mode

Le mode d'une série statistique est le caratère (ou la classe de caractères) le plus fréquemment atteint.

Exemple 3. • Dans l'exemple des notes dans la classe de 15 élèves ci-dessus, le mode est la note 12.

• Dans le cas des tailles, on dit que la classe modale est l'intervalle $[160, 170[$.

Définition 5: Moyenne

La moyenne d'une série statistique $x = (x_1, \dots, x_n)$ est $\overline{x} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i$.

Remarque 2. Si $\{x_1, \dots, x_n\} = \{y_1, \dots, y_p\}$ avec $y_1 < y_2 < \cdots < y_p$ d'effectifs respectifs n_i et de fréquences respectives f_i, on a

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{p} n_i y_i = \sum_{i=1}^{p} f_i y_i.$$

Définition 6: Médiane

Soit $(x_1, \dots, x_n)$ une série statistique avec $x_1 \leqslant x_2 \leqslant \cdots \leqslant x_n$.
• Si $n = 2p + 1$ est impair, alors on définit la médiane de la série statistique x par la valeur

$$M = x_{p+1}.$$

• Si $n = 2p$ est pair, alors on définit la médiane de la série statistique x par la valeur

$$M = \frac{1}{2}(x_p + x_{p+1}).$$

Autrement dit, la médiane est la valeur du caractère qui partage l'effectif en deux.

Exemple 4. Reprenons l'exemple des notes du devoir de mathématiques. La classe étant constituée de 15 élèves, la médiane est la 8ème note dans l'ordre croissant. En considérant les effectifs cumulés croissants, on constate que la médiane est $M = 12$.

24.1.5 Caractéristiques empiriques de dispersion

Définition 7: Variance et écart-type

Soit $x = (x_1, \dots, x_n)$ une série statistique.

1. On définit la variance s_x^2 de la série x par

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

2. On définit l'écart-type s_x de la série x par

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

Remarque 3. • La variance de la série statisque x est en fait la moyenne de la série statistique $(x - \overline{x})^2$.

• On a $s_x^2 = 0 \Leftrightarrow \forall i \in [\![1, n]\!], x_i = \overline{x}$ donc une série statistique est de variance nulle si et seulement si elle est constante égale à sa moyenne.

On retrouve la formule de König-Huyens vue en probabilités :

Proposition 1: Formule de König-Huygens

Soit $x = (x_1, \dots, x_n)$ une série statistique.
Alors

$$s_x^2 = \overline{x^2} - \overline{x}^2$$

où x^2 est la série statistique $x^2 = (x_1^2, \dots, x_n^2)$.

Définition 8: Quartiles et déciles

Soit $x = (x_1, \dots, x_n)$ une série statistique avec $x_1 \leqslant x_2 \leqslant \cdots \leqslant x_n$.
• Les quartiles de x sont trois valeurs de caractères qui découpent la série x en quatre effectifs égaux : le premier quartile est la valeur séparant le quart constitué des plus faibles valeurs du reste, le deuxième quartile est la médiane, et le troisième quartile est la valeur séparant le quart constitué des plus fortes valeurs du reste.
• De même, les déciles de x sont 9 valeurs découpant la série statistique en 10 effectifs égaux.

Remarque 4. On représente souvent les quartiles d'une série statistique dans une boîte à moustaches.

Exemple 5. Reprenons l'exemple des notes du devoir de mathématiques. La classe est constituée de 15 élèves et $15/4 = 3,75$. On arrondit à l'entier du dessus pour trouver le premier quartile : c'est la 4ème note obtenue dans l'ordre croissant, c'est à dire 10. (Ici, le minimum et le premier quartile sont confondus.)

Le troisième quartile est alors la 12ème note obtenue dans l'ordre croissant, c'est à dire 14.

24.2 Statistique bivariée

24.2.1 Nuage de points et point moyen

Dans certaines situations, il peut être intéressant d'étudier une série statistique double de taille n portant sur deux caractères quantitatifs x et y (la taille et l'âge, la pression et la température...).

Dans ce cas, on représente les données par un n-uplet d'éléments de $\mathbb{R}^2((x_1, y_1), \dots, (x_n, y_n))$.

On représente graphiquement ces données par un nuage de points de $\mathbb{R}^2$.

Définition 9: Point moyen d'un nuage de points

On reprend les notations précédentes.
Soit $(x_i, y_i)_{1\leqslant i\leqslant n}$ une série statistique bivariée.
Soient $\overline{x}$ et $\overline{y}$ les moyennes respectives des séries statistiques $(x_i)_{1\leqslant i\leqslant n}$ et $(y_i)_{1\leqslant i\leqslant n}$.
On appelle point moyen du nuage de points $(x_i, y_i)_{1\leqslant i\leqslant n}$ le point $(\overline{x}, \overline{y})$.

24.2.2 Covariance et cœfficient de corrélation

Définition 10: Covariance

Soit $(x_i, y_i)_{1\leqslant i\leqslant n}$ une série statistique bivariée.
On appelle covariance de la série (x, y), notée s_{xy}, le nombre

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

où $\overline{x}$ et $\overline{y}$ désignent les moyennes respectives des séries statistiques $(x_i)_{1\leqslant i\leqslant n}$ et $(y_i)_{1\leqslant i\leqslant n}$.

Remarque 5. • La covariance s_{xy} est en fait la moyenne de la série statistique univariée

$$((x_i - \overline{x})(y_i - \overline{y}))_{1\leqslant i\leqslant n}.$$

• Si $y = x$, on a $s_{xx} = s_x^2$, c'est à dire que la covariance de la série (x, x) est égale à la variance de x.

Proposition 2

Soit $(x_i, y_i)_{1\leqslant i\leqslant n}$ une série statistique bivariée.
Alors

$$s_{xy} = \overline{xy} - \overline{x} \times \overline{y}$$

où $\overline{xy}$ désigne la moyenne de la série statistique $(x_i y_i)_{1\leqslant i\leqslant n}$.

Démonstration. On a

$$\begin{aligned} s_{xy} &= \frac{1}{n}\sum_{i=1}^{n}(x_i y_i - x_i\overline{y} - \overline{x}y_i + \overline{x}\times\overline{y}) \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{y}\frac{1}{n}\sum_{i=1}^{n} x_i - \overline{x}\frac{1}{n}\sum_{i=1}^{n} y_i + \frac{1}{n}\sum_{i=1}^{n}\overline{x}\times\overline{y} \\ &= \overline{xy} - \overline{y}\times\overline{x} - \overline{x}\times\overline{y} + \overline{x}\times\overline{y} \\ &= \overline{xy} - \overline{x}\times\overline{y}. \end{aligned}$$

■

Définition 11: Cœfficient de corrélation

Soit $(x_i, y_i)_{1\leqslant i\leqslant n}$ une série statistique bivariée.
On suppose que les séries statistiques $x = (x_i)_{1\leqslant i\leqslant n}$ et $y = (y_i)_{1\leqslant i\leqslant n}$ sont d'écart-types s_x et s_y non nuls.
On définit alors le cœfficent de corrélation de la série (x, y) par

$$r_{xy} = \frac{s_{xy}}{s_x s_y},$$

i.e. la covariance de (x, y) divisée par le produit des écart-types de x et y.

Remarque 6. On a toujours $r_{xy} \in [-1, 1]$.

En effet, $r_{xy} = \dfrac{\displaystyle\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\displaystyle\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\displaystyle\sum_{i=1}^{n}(y_i - \overline{y})^2}}$ et l'inégalité $|r_{xy}| \leqslant 1$ découle de l'inégalité de Cauchy-Schwarz.

Exemple 6. Soit $x = (x_1, \ldots, x_n)$ une série statistique. Soit $(a, b) \in \mathbb{R}^* \times \mathbb{R}$. On considère la série statistique $y = ax + b = (ax_1 + b, \ldots, ax_n + b)$.

On a $s_{xy} = \overline{xy} - \overline{x}\times\overline{y} = \overline{x(ax+b)} - \overline{x}\times\overline{ax+b} = a\overline{x^2} + b\overline{x} - a\overline{x}^2 - b\overline{x} = a(\overline{x^2} - \overline{x}^2) = as_x^2$.

D'autre part,

$$s_y^2 = \overline{(ax+b)^2} - \overline{ax+b}^2 = a^2\overline{x^2} + 2ab\overline{x} + b^2 - (a\overline{x} + b)^2 = a^2\overline{x^2} + 2ab\overline{x} + b^2 - a^2\overline{x}^2 - 2ab\overline{x} - b^2$$

d'où $s_y^2 = a^2 s_x^2$ puis $s_y = |a|s_x$.

Ainsi, $r_{xy} = \dfrac{s_{xy}}{s_x s_y} = \dfrac{as_x^2}{|a|s_x^2} = \dfrac{a}{|a|} = \begin{cases} 1 & \text{si } a > 0 \\ -1 & \text{si } a < 0 \end{cases}$.

24.2.3 Ajustement affine selon la méthode des moindres carrés

Soit $(x_i, y_i)_{1\leqslant i\leqslant n}$ une série statistique bivariée.

On suppose qu'il existe un couple $(i,j) \in [\![1,n]\!]^2$ tel que $x_i \neq x_j$, c'est à dire que les points ne sont pas tous alignés sur une droite verticale.

On se propose de trouver la droite d'équation $y = ax + b$ qui approche au mieux le nuage de points $(x_i, y_i)_{1 \leqslant i \leqslant n}$.

Nous allons chercher pour cela les réels $(a,b) \in \mathbb{R}^2$ qui minimisent la somme

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

C'est la somme des carrés des distances de chaque point du nuage de points au point de la droite de même abscisse.

On admet l'existence d'un tel couple (a,b). La droite d'équation $y = ax+b$ est alors appelée droite de régression linéaire de la série statistique (x,y).

Considérons la fonction $f : (a,b) \longmapsto \sum_{i=1}^{n} (y_i - (ax_i + b))^2$.

Cette fonction est de classe $\mathcal{C}^1$ sur $\mathbb{R}^2$ et pour tout $(a,b) \in \mathbb{R}^2$, on a

$$\frac{\partial f}{\partial a}(a,b) = \sum_{i=1}^{n} -2x_i(y_i - (ax_i + b)) = 2a\sum_{i=1}^{n} x_i^2 + 2b\sum_{i=1}^{n} x_i - 2\sum_{i=1}^{n} x_i y_i$$

et

$$\frac{\partial f}{\partial b}(a,b) = \sum_{i=1}^{n} -2(y_i - (ax_i + b)) = 2a\sum_{i=1}^{n} x_i + 2bn - 2\sum_{i=1}^{n} y_i.$$

Si (a,b) est un minimum de f, alors $\dfrac{\partial f}{\partial a} = \dfrac{\partial f}{\partial b} = 0$, d'où

$$\left\{ \begin{array}{rcl} a\displaystyle\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i & = & \displaystyle\sum_{i=1}^{n} x_i y_i \\ a\displaystyle\sum_{i=1}^{n} x_i + bn & = & \displaystyle\sum_{i=1}^{n} y_i. \end{array} \right. \Leftrightarrow \begin{pmatrix} \displaystyle\sum_{i=1}^{n} x_i^2 & \displaystyle\sum_{i=1}^{n} x_i \\ \displaystyle\sum_{i=1}^{n} x_i & n \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \displaystyle\sum_{i=1}^{n} x_i y_i \\ \displaystyle\sum_{i=1}^{n} y_i \end{pmatrix}.$$

Le déterminant de la matrice $\begin{pmatrix} \displaystyle\sum_{i=1}^{n} x_i^2 & \displaystyle\sum_{i=1}^{n} x_i \\ \displaystyle\sum_{i=1}^{n} x_i & n \end{pmatrix}$ vaut

$$n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 = n^2\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2\right) = n^2(\overline{x^2} - \overline{x}^2) = n^2 s_x^2.$$

Or, on a supposé que la série x n'était pas constante donc $s_x^2 > 0$ ce qui implique que le système est de Cramer et on a alors

$$\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \displaystyle\sum_{i=1}^{n} x_i^2 & \displaystyle\sum_{i=1}^{n} x_i \\ \displaystyle\sum_{i=1}^{n} x_i & n \end{pmatrix}^{-1} \begin{pmatrix} \displaystyle\sum_{i=1}^{n} x_i y_i \\ \displaystyle\sum_{i=1}^{n} y_i \end{pmatrix} = \frac{1}{n^2 s_x^2} \begin{pmatrix} n & -\displaystyle\sum_{i=1}^{n} x_i \\ -\displaystyle\sum_{i=1}^{n} x_i & \displaystyle\sum_{i=1}^{n} x_i^2 \end{pmatrix} \begin{pmatrix} \displaystyle\sum_{i=1}^{n} x_i y_i \\ \displaystyle\sum_{i=1}^{n} y_i \end{pmatrix}$$

d'où

$$a = \frac{1}{n^2 s_x^2}\left(n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i\right) = \frac{\overline{xy} - \overline{x} \times \overline{y}}{s_x^2} = \frac{s_{xy}}{s_x^2}$$

et

$$b = \frac{1}{n^2 s_x^2}\left(\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i\right) = \frac{1}{s_x^2}(\overline{y}\overline{x^2} - \overline{x} \times \overline{xy}).$$

Or,

$$\begin{aligned}
\overline{y} - a\overline{x} &= \overline{y} - \frac{s_{xy}\overline{x}}{s_x^2} \\
&= \frac{\overline{y}s_x^2 - (\overline{xy} - \overline{x} \times \overline{y})\overline{x}}{s_x^2} \\
&= \frac{\overline{y}(\overline{x^2} - \overline{x}^2) - \overline{x} \times \overline{xy} + \overline{x}^2\overline{y}}{s_x^2} \\
&= \frac{1}{s_x^2}(\overline{y}\overline{x^2} - \overline{x} \times \overline{xy}) \\
&= b.
\end{aligned}$$

Finalement, la droite de régression linéaire recherchée a pour équation $y = ax + b$ avec $a = \frac{s_{xy}}{s_x^2}$ et $b = \overline{y} - a\overline{x}$.