I GÉNÉRALITÉS

Les premières études statistiques étaient des recensements démographiques : on en a conservé le vocabulaire.

Population : C'est l'ensemble sur lequel porte l'étude statistique.

Individu : C'est un élément de la population.

Caractère : C'est l'aspect que l'on observe sur les individus. Un caractère permet de déterminer une partition de la population selon ses diverses valeurs (par exemple le genre est un caractère à deux modalités : masculin ou féminin).

Lorsque les différentes valeurs d'un caractère sont des nombres, le caractère est *quantitatif*. Dans le cas contraire, le caractère est *qualitatif*.

L'effectif d'une valeur du caractère étudié est le nombre d'individus de la population ayant cette valeur.

La fréquence d'une valeur est le quotient de l'effectif de cette valeur par l'effectif total de la population. (la fréquence peut être exprimée en pourcentage)

$$\text{féquence} = \frac{\text{effectif de la valeur}}{\text{effectif total}}$$

II MÉDIANE ET QUANTILES

1 LA MÉDIANE

La médiane d'une série statistique est une valeur telle qu'il y ait autant d'observations ayant une valeur supérieure à la médiane que d'observations ayant une valeur inférieure à la médiane.

La médiane d'une série statistique de N valeurs rangées par ordre croissant est le nombre M_e défini par :

— si l'effectif N est impair, la médiane M_e est la valeur centrale du caractère c'est à dire la valeur de rang $\frac{N+1}{2}$ de la série ordonnée.

— si l'effectif N est pair, la médiane M_e est la demi-somme des deux valeurs centrales du caractère c'est à dire la moyenne des valeurs de rangs $\frac{N}{2}$ et $\frac{N}{2}+1$ de la série ordonnée.

EXEMPLE

Dans un service de maintenance, on a répertorié le nombre d'interventions par jour sur un mois. On a obtenu la distribution suivante :

Nombre d'interventions x_i	3	5	6	7	8	9
Nombre de jours n_i	2	5	8	6	3	1

L'effectif total $N = 25$ donc la médiane est la valeur du caractère de rang 13 soit $Me = 6$.

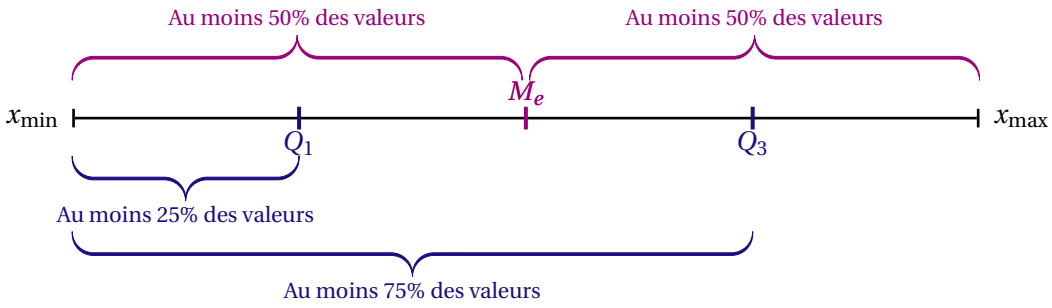
2 LES QUANTILES

LES QUARTILES

Les quartiles au nombre de trois Q_1, Q_2 et Q_3 partagent l'ensemble étudié de N éléments préalablement classés par valeurs croissantes, en quatre sous ensembles.

— Le premier quartile noté Q_1 est la plus petite valeur de la série statistique telle qu'au moins 25 % des valeurs de la série sont inférieures ou égales à Q_1.

— Le troisième quartile noté Q_3 est la plus petite valeur de la série statistique telle qu'au moins 75 % des valeurs de la série sont inférieures ou égales à Q_3.



REMARQUE

L'intervalle interquartile $[Q_1; Q_3]$ contient au moins 50% des valeurs de la série.

EXEMPLE

Dans la série précédente, l'effectif total $N = 25$.

— $25 \times \frac{1}{4} = 6{,}25$ donc le premier quartile est la valeur du caractère de rang 7 soit $Q_1 = 5$.

— $25 \times \frac{3}{4} = 18{,}75$ donc le troisième quartile est la valeur du caractère de rang 19 soit $Q_3 = 7$.

LES DÉCILES

Les déciles au nombre de neuf D_1, D_2, …, D_9 partagent l'ensemble étudié de N éléments préalablement classés par valeurs croissantes, en dix sous ensembles.

— Le premier décile noté D_1 est la plus petite valeur de la série statistique telle qu'au moins 10 % des valeurs de la série sont inférieures ou égales à D_1.

— Le neuvième décile noté D_9 est la plus petite valeur de la série statistique telle qu'au moins 90 % des valeurs de la série sont inférieures ou égales à D_9.

3 CARACTÉRISTIQUES DE DISPERSION

— L'étendue est la différence entre la plus grande et la plus petite valeur d'une série statistique.

— L'écart interquartile est égal à la différence entre le troisième et le premier quartiles.

— L'écart interdécile est égal à la différence entre le neuvième et le premier déciles.

4 DIAGRAMME EN BOÎTE

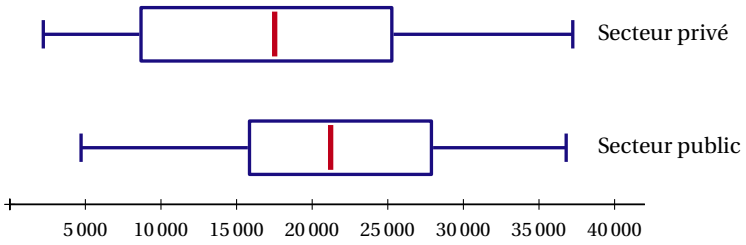
La représentation graphique de la dispersion d'une série statistique se fait à l'aide de diagramme en boîte appelés aussi « boîte à moustaches » ou « box-plot ».
Pour une catégorie donnée, on construit, en face d'un axe permettant de repérer les quantiles de la variable étudiée, un rectangle dont la longueur est égale à l'écart interquartile $Q_3 - Q_1$, la médiane est représentée par un trait. On ajoute alors des segments aux extrémités menant jusqu'aux valeurs extrêmes, ou jusqu'aux premier et neuvième déciles.

EXEMPLE

Le tableau suivant donne la distribution du revenu salarial par secteur d'activité en France en 2014.

	D1	Q1	Médiane	Q3	D9
Secteur privé	2 218	8 570	17 520	25 377	37 234
Secteur public	4 716	15 744	21 221	27 996	36 797

Source : INSEE



III MOYENNE, VARIANCE ET ÉCART-TYPE

1 LA MOYENNE

On considère la série statistique donnée par le tableau ci-contre.
On note $N = n_1 + n_2 + \cdots + n_p$ l'effectif total

Valeur x_i	x_1	x_2	$\cdots$	x_p
Effectif n_i	n_1	n_2	$\cdots$	n_p

La moyenne d'une série statistique est le quotient noté $\overline{x}$ de la somme de toutes les valeurs de cette série par l'effectif total.

$$\overline{x} = \frac{n_1 \times x_1 + n_2 \times x_2 + \cdots + n_p \times x_p}{N} = \frac{1}{N}\sum_{i=1}^{p} n_i x_i$$

REMARQUE

Soit $f_i = \dfrac{n_i}{N}$ la fréquence de la valeur x_i alors, la moyenne $\overline{x} = f_1 \times x_1 + f_2 \times x_2 + \cdots + f_p \times x_p$.

EXEMPLE

Avec la série statistique précédente :

Nombre d'interventions x_i	3	5	6	7	8	9
Nombre de jours n_i	2	5	8	6	3	1
Fréquence f_i	0,08	0,2	0,32	0,24	0,12	0,04

Le nombre moyen d'interventions par jour est :

$$\overline{x} = \frac{2\times 3 + 5\times 5 + 8\times 6 + 6\times 7 + 3\times 8 + 1\times 9}{25} = 6{,}16$$

ou en utilisant les fréquences :

$$\overline{x} = 0{,}08\times 3 + 0{,}2\times 5 + 0{,}32\times 6 + 0{,}24\times 7 + 0{,}12\times 8 + 0{,}04\times 9 = 6{,}16$$

2 VARIANCE ET ÉCART-TYPE

Soit $(x_i ; n_i)$, $1 \leqslant i \leqslant p$, une série statistique de moyenne $\overline{x}$ et d'effectif total N.

— La variance de cette série est le nombre V défini par :

$$V = \frac{n_1 \times (x_1 - \overline{x})^2 + n_2 \times (x_2 - \overline{x})^2 + \cdots + n_p \times (x_p - \overline{x})^2}{N} = \frac{1}{N}\sum_{i=1}^{p} n_i (x_i - \overline{x})^2$$

— L'écart-type, noté σ, de cette série est égal à la racine carrée de la variance :

$$\sigma = \sqrt{V}$$

REMARQUE

La variance est la moyenne des carrés des écarts à la moyenne $\overline{x}$. Les valeurs $\left(x_i - \overline{x}\right)$ sont les « écarts à la moyenne » ; les « carrés des écarts à la moyenne » sont donc $\left(x_i - \overline{x}\right)^2$.
En faisant la moyenne des carrés des écarts à la moyenne, on trouve la variance.

EXEMPLE

Dans la série précédente de moyenne $\overline{x} = 6{,}16$ la variance est :

$$V = \frac{2 \times (3-6{,}16)^2 + 5 \times (5-6{,}16)^2 + 8 \times (6-6{,}16)^2 + 6 \times (7-6{,}16)^2 + 3 \times (8-6{,}16)^2 + 1 \times (3-6{,}16)^2}{25} = 1{,}9744$$

L'écart-type de cette série est donc $\sigma = \sqrt{1{,}9744} \approx 1{,}4$

PROPRIÉTÉ

Soit $(x_i ; n_i)$, $1 \leqslant i \leqslant p$, une série statistique de moyenne $\overline{x}$ et d'effectif total N.
La variance de cette série est le nombre V défini par :

$$V = \frac{n_1 \times {x_1}^2 + n_2 \times {x_2}^2 + \cdots + n_p \times {x_p}^2}{N} - \overline{x}^2 = \frac{1}{N} \sum_{i=1}^{p} n_i {x_i}^2 - \overline{x}^2$$

✲ DÉMONSTRATION

$$\begin{aligned}
V &= \frac{n_1 \times \left(x_1 - \overline{x}\right)^2 + n_2 \times \left(x_2 - \overline{x}\right)^2 + \cdots + n_p \times \left(x_p - \overline{x}\right)^2}{N} \\
&= \frac{n_1 \times \left({x_1}^2 - 2x_1\overline{x} + \overline{x}^2\right) + n_2 \times \left({x_2}^2 - 2x_2\overline{x} + \overline{x}^2\right) + \cdots + n_p \times \left({x_p}^2 - 2x_p\overline{x} + \overline{x}^2\right)}{N} \\
&= \frac{n_1 \times {x_1}^2 + n_2 \times {x_2}^2 + \cdots + n_p \times {x_p}^2}{N} - 2\frac{n_1 \times x_1 + n_2 \times x_2 + \cdots + n_p \times x_p}{N} \times \overline{x} + \frac{n_1 + n_2 + \cdots + n_p}{N} \times \overline{x}^2 \\
&= \frac{n_1 \times {x_1}^2 + n_2 \times {x_2}^2 + \cdots + n_p \times {x_p}^2}{N} - 2\overline{x} \times \overline{x} + \overline{x}^2 \\
&= \frac{n_1 \times {x_1}^2 + n_2 \times {x_2}^2 + \cdots + n_p \times {x_p}^2}{N} - \overline{x}^2
\end{aligned}$$