chapitre 7 : statistiques descriptives

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*.

I EFFECTIFS - FRÉQUENCES

1 DEFINITIONS

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)

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

2 EFFECTIFS CUMULÉS ET FRÉQUENCES CUMULÉES

On étudie un caractère quantitatif dans une population.

— Les différentes valeurs x_i du caractère quantitatif constituent une série statistique notée (x_i).

— On note n_i l'effectif de la valeur x_i.

L'effectif cumulé croissant de la valeur x_i est la somme des effectifs de toutes les valeurs inférieures ou égales à x_i.
La fréquence cumulée croissante de la valeur x_i est la somme des fréquences de toutes les valeurs inférieures ou égales à x_i.

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	total
Nombre de jours n_i	2	4	9	6	3	1	25

Le nombre total de journées d'intervention est $2+4+9+6+3+1=25$. Les fréquences des différentes valeurs du nombre d'intervention sont :

Nombre d'interventions x_i	3	5	6	7	8	9	total
Nombre de jours n_i	2	4	9	6	3	1	25
Fréquence $f_i = \frac{n_i}{25}$	0,08	0,16	0,36	0,24	0,12	0,04	1

Le tableau suivant donne les effectifs cumulés croissants ainsi que les fréquences cumulées croissantes :

Nombre d'interventions x_i	3	5	6	7	8	9	total
Nombre de jours n_i	2	4	9	6	3	1	25
Effectif cumulé	2	6	15	21	24	25	
Fréquence cumulée	0,08	0,24	0,6	0,84	0,96	1	

Interprétons les résultats de la colonne "7":

L'effectif est 6 : sur les 25 jours d'activité du service de maintenance, il y a eu 7 interventions par jour pendant 6 jours.
L'ECC est 21 : sur les 25 jours d'activité, 21 journées ont connu moins de 7 interventions par jour.
La FCC est 84% : les journées comptant moins de 7 interventions représentent 84% de l'activité du mois.

REMARQUE

Lorsque le caractère est **quantitatif et continu**, les modalités, *c'est à dire les différentes valeurs prises par le caractère* peuvent être regroupées en intervalles appelés **classes**. La valeur centrale de la classe est alors appelée **centre de la classe**.
Le centre de la classe est obtenu en effectuant la moyenne des bornes de la classe.

II CARACTÉRISTIQUES D'UNE SÉRIE STATISTIQUE

1 CARACTÉRISTIQUES DE POSITION

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.

$$\overline{x} = \frac{n_1 x_1 + n_2 x_2 + \cdots + n_p x_p}{N}$$

REMARQUE

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

EXEMPLE

Avec la série statistique précédente :

Nombre d'interventions x_i	3	5	6	7	8	9
Nombre de jours n_i	2	4	9	6	3	1
Fréquence f_i	0,08	0,16	0,36	0,24	0,12	0,04

Le nombre moyen d'interventions par jour est :

$$\overline{x} = \frac{2\times 3 + 4\times 5 + 9\times 6 + 6\times 7 + 3\times 8 + 1\times 9}{25} = 6{,}2$$

ou en utilisant les fréquences :

$$\overline{x} = 0{,}08\times 3 + 0{,}16\times 5 + 0{,}36\times 6 + 0{,}24\times 7 + 0{,}12\times 8 + 0{,}04\times 9 = 6{,}2$$

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.

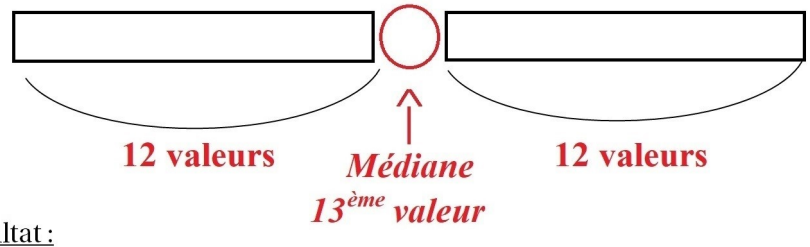
La médiane d'une série statistique de N valeurs rangées par ordre croissant est le nombre M_e défini par :

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

EXEMPLE

Dans la série précédente, l'effectif total $N = 25$ donc la médiane est la valeur du caractère de rang 13 soit $Me = 6$.



Interprétons le résultat :

La médiane est 6 : **au moins 50%** des journées d'interventions ont compté **au plus** 6 interventions.

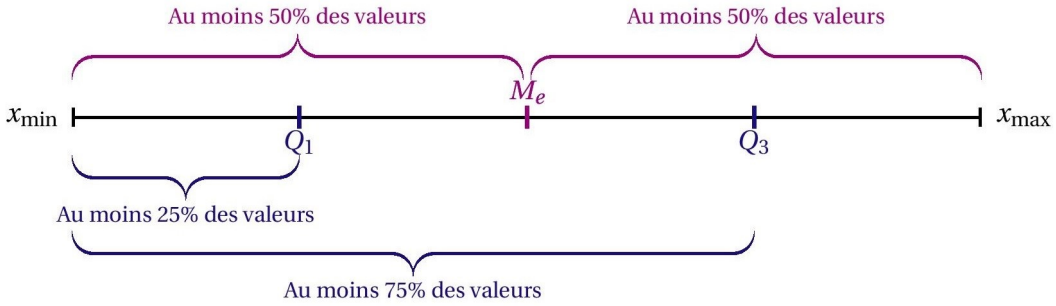
LES QUANTILES

1. 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.



2. LES DÉCILES

Les déciles au nombre de neuf $D_1, D_2, \ldots, D_9$ 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.

REMARQUE

Le deuxième quartile Q_2 et le cinquième décile D_5 sont égaux à la médiane.

VOCABULAIRE

La moyenne, la médiane, les quantiles sont appelés caractéristiques de position d'une série statistique.

2 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.

3 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