

## Statistiques

L'étude d'une distribution statistique suppose une *population* dont on analyse un *caractère* au travers d'un *échantillon* prenant un nombre fini de *valeurs*, chacune de ces valeurs ayant un *effectif* donné.

Ainsi, pour étudier le *nombre de voitures par famille* dans la commune de *Cléry en Vexin (95)*.

La *population* (ici confondue avec l'échantillon étudié) est l'ensemble des familles de la commune.

Le *caractère* étudié est le *nombre de voitures par famille*.

L'enquête menée a permis d'établir le tableau suivant :

Nbre Voitures	0	1	2	3
Effectif	12	65	45	8

Les *valeurs du caractère* sont ( 0 ; 1 ; 2 ; 3 ), notées (  $x_1 ; x_2 ; x_3 ; x_4$  ), nombre de voitures par familles.

Les *effectifs relatifs à chaque valeur* sont respectivement ( 12 ; 65 ; 45 ; 8 ), notés (  $n_1 ; n_2 ; n_3 ; n_4$  ).

### Fréquence de chaque valeur du caractère

Nbre voitures	0	1	2	3
Effectif	12	65	45	8
Fréquence	12/130	65/130	45/130	8/130

La *fréquence* d'une *valeur du caractère* est le *pourcentage* (probabilité) que représente l'effectif de cette valeur par rapport à l'*effectif total* de l'échantillon étudié.

### Calcul de la moyenne et de l'écart-type.

Une distribution statistique est connue par sa *moyenne arithmétique* (indicateur de *valeur centrale*) qui donne un *point moyen* représentatif de l'ensemble de l'échantillon, mais également par son *écart-type* (indicateur de *dispersion*) qui exprime l'étalement de l'échantillon autour de sa moyenne.

Un élève qui a pour notes ( 6 ; 12 ; 15 ) a une *moyenne* de 11, tout comme un élève noté ( 10 ; 11 ; 12 ), mais la *dispersion* des notes du second élève est moindre que celle du premier.

**Les deux informations (moyenne ; dispersion) sont nécessaires à la connaissance d'une distribution statistique.**

Soit un *caractère*  $X$  dont on veut étudier une *répartition statistique* :

Soient (  $x_1, x_2, x_3, \dots, x_k$  ) les  $k$  *valeurs* possibles du *caractère* étudié.

Soient (  $n_1, n_2, n_3, \dots, n_k$  ) les  $k$  *effectifs* correspondant à chacune des valeurs du caractère.

La *moyenne arithmétique pondérée* du caractère  $X$  est égale à la somme des produits des diverses valeurs du caractère par leur effectif propre  $S = n_1 x_1 + n_2 x_2 + n_3 x_3 + \dots + n_k x_k$ .

Cette somme est ensuite divisée par l'effectif total  $N = n_1 + n_2 + n_3 + \dots + n_k$ .

$$\bar{X} = \frac{S}{N} = \frac{n_1 x_1 + n_2 x_2 + \dots + n_k x_k}{n_1 + n_2 + \dots + n_k}$$

Un savant calcul mathématique estime l'écart-type  $\sigma$  (*sigma*), qui mesure la dispersion de l'échantillon autour de sa *moyenne*, par la formule :

$\sigma = \sqrt{V}$  ou  $V = \overline{X^2} - (\overline{X})^2$  est appelée *variance*, formule où  $\overline{X^2} = \frac{n_1 x_1^2 + n_2 x_2^2 + \dots + n_k x_k^2}{n_1 + n_2 + \dots + n_k}$ , est la *moyenne des carrés des valeurs*, et où  $(\overline{X})^2$  est le *carré de la moyenne* des valeurs vues précédemment.

On dit souvent  $\sigma = \sqrt{\overline{X^2} - (\overline{X})^2} = \text{Racine de (moyenne des carrés moins carré de la moyenne)}$ .

$x_i$	$n_i$	$n_i x_i$	$n_i x_i^2$
0	12	0	0
1	65	65	65
2	45	90	180
3	8	24	72
	<b>130</b>	<b>179</b>	<b>317</b>

L'effectif total est  $N = \sum_{i=1}^4 n_i = 130$ , d'où  $\sum_{i=1}^4 n_i x_i = 179$ , d'où :

$$\overline{X} = \frac{1}{N} \left( \sum_{i=1}^4 n_i x_i \right) = \frac{179}{130} = 1,38 \text{ voitures par famille.}$$

$$\sum_{i=1}^4 n_i x_i^2 = 317. \text{ La moyenne des carrés est : } \overline{X^2} = \frac{1}{N} \left( \sum_{i=1}^4 n_i x_i^2 \right) = \frac{317}{130} = 2,44.$$

$$\text{La variance : } V = \overline{X^2} - (\overline{X})^2 = 2,44 - (1,38)^2 = 0,536.$$

$$\text{L'écart-type : } \sigma = \sqrt{V} = \sqrt{0,536} = 0,73 \text{ voiture.}$$

### **Comparaison de la dispersion de deux caractères statistiques dissemblables – Coefficient de Variation**

Pour comparer la dispersion de deux caractères *identiques* dans deux populations différentes, il suffit de comparer les *écarts-type* de ces deux populations.

Ainsi, dans l'exemple précédent :  $\sigma_1 = 0,73$  voiture à Cléry en Vexin, et  $\sigma_2 = 0,81$  voiture à Pontoise, permettent de conclure en termes de dispersion.

Par contre : Comparer les dispersions entre "le nombre de voitures par foyer" et "le nombre d'arbres par jardin" ne peut se limiter à la consultation des écarts-type :  $\sigma_1 = 0,73$  voiture et  $\sigma_2 = 1,23$  arbres.

Pour comparer la dispersion de deux caractères différents, on utilise le Coefficient de Variation :  $CV = \frac{\sigma}{X}$ .

Le Coefficient de Variation est un nombre sans unité.

Nombre de voitures par foyer :  $\overline{X_1} = 1,38$  voitures ,  $\sigma_1 = 0,73$  voiture  $\Rightarrow CV_1 = \frac{\sigma_1}{\overline{X_1}} = \frac{0,73}{1,38} \approx 0,529$  .

Nombre d'arbres par jardin :  $\overline{X_2} = 2,41$  arbres ,  $\sigma_2 = 1,23$  arbres  $\Rightarrow CV_2 = \frac{\sigma_2}{\overline{X_2}} = \frac{1,23}{2,41} \approx 0,510$  .

Contrairement à ce que l'on pouvait penser, ce second caractère est le moins dispersé autour de sa moyenne.

### **Autres indicateurs de Valeur Centrale :**

- Le Mode  $Mo$  : Valeur du caractère de plus forte fréquence (valeur à la mode)

Lorsque la population est rangée en classes, on parlera de Classe Modale .

- La Médiane  $Me$  : Valeur du caractère qui sépare la population en deux sous-populations de même effectif.

50% de la population telle que  $x_i < Me$  et 50% de la population telle que  $x_i > Me$  .

Pour déterminer graphiquement la *Médiane* , on trace le *polygone des fréquences cumulées* .

La valeur de  $Me$  correspondant à la valeur d'ordonnée de  $0,50 = 50\%$  .

### **Autres indicateurs de Position :**

- Les Quartiles  $Q_1$  ,  $Q_3$  :

$Q_1$  , 1<sup>er</sup> quartile d'une série statistique, plus petite valeur du caractère étudié telle qu'au moins 25% des données soient inférieures à ce nombre.

$Q_3$  , 3<sup>ème</sup> quartile d'une série statistique, plus petite valeur du caractère étudié telle qu'au moins 75% des données soient inférieures à ce nombre.

On remarquera que  $Q_2 = Me$  (médiane), 2<sup>nd</sup> quartile d'une série statistique, plus petite valeur du caractère étudié telle qu'au moins 50% des données soient inférieures à ce nombre.

- Les Déciles  $D_1$  ,  $D_2$  , ... ,  $D_9$  :

$D_1$  , 1<sup>er</sup> décile d'une série statistique, plus petite valeur du caractère étudié telle qu'au moins 10% des données soient inférieures à ce nombre.

$D_2$  , 2<sup>ème</sup> décile d'une série statistique, plus petite valeur du caractère étudié telle qu'au moins 20% des données soient inférieures à ce nombre.

.....

$D_9$  , 9<sup>ème</sup> décile d'une série statistique, plus petite valeur du caractère étudié telle qu'au moins 90% des données soient inférieures à ce nombre.

On remarquera que  $D_5 = Me$  (médiane), 5<sup>ème</sup> décile d'une série statistique, plus petite valeur du caractère étudié telle qu'au moins 50% des données soient inférieures à ce nombre.

### **Autres indicateurs de Dispersion :**

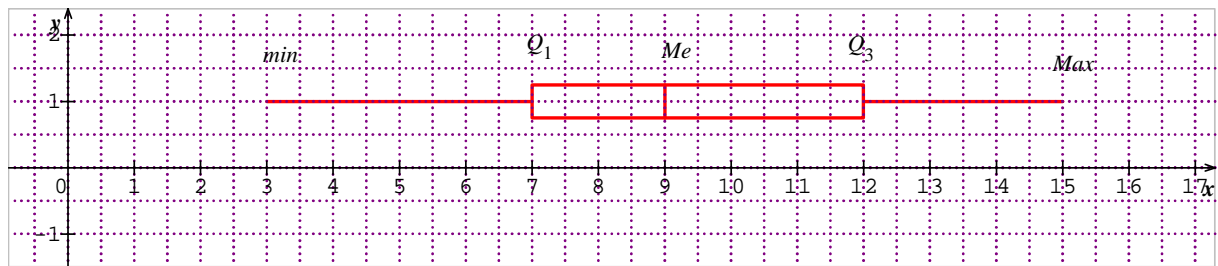
- Etendue  $e$  : Différence  $e = Max - min$  , entre la valeur Maximum et la valeur minimum du caractère étudié.

- Intervalle Interquartile  $Q_3 - Q_1$  : Différence entre le 1<sup>er</sup> quartile et le 3<sup>ème</sup> quartile, écart dans lequel se trouvent les 50% centrales des valeurs du caractère (25% avant  $Q_1$  et 25% après  $Q_3$ ).

**Diagramme en Boite (boite à moustache ou diagramme de Tuckey) :**

Le diagramme en boite représente graphiquement les principaux indicateurs de position d'une série statistique :

$min$  ,  $Q_1$  ,  $Me$  ,  $Q_3$  ,  $Max$  .



Le diagramme en boite peut aussi inclure le 1<sup>er</sup> décile  $D_1$  (10% des valeurs lui sont inférieures, et le 9<sup>ème</sup> décile  $D_9$  (10% des valeurs lui sont supérieures):

