Cours 12 : Paramètres de dispersion

Étant donnée une variable statistique quantitative, nous définissons deux **paramètres de dispersion**, c'est-à-dire des valeurs mesurant l'étalement de la distribution de la variable (ou l'écart entre les valeurs prises par cette variable et la moyenne) :
— l'écart-type ;
— l'écart inter-quartiles.

Écart-type

Supposons que l'on dispose d'une série statistique brute $(x_1, \ldots, x_N)$, où l'effectif total est N. La **variance** σ^2 est définie comme la **moyenne des carrés des écarts à la moyenne** :

$$\sigma^2 := \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \ldots + (x_N - \overline{x})^2}{N},$$

où $\overline{x}$ est la moyenne.

En général, avec des variables discrètes, on ne travaille pas avec la série brute mais avec la série triée par classes : pour chaque modalité i, on calcule l'effectif n_i ou la fréquence $f_i = n_i/N$. La variance est alors :

$$\sigma^2 = \frac{n_1(x_1 - \overline{x})^2 + n_2(x_2 - \overline{x})^2 + \ldots + n_p(x_p - \overline{x})^2}{N} = f_1(x_1 - \overline{x})^2 + f_2(x_2 - \overline{x})^2 + \ldots + f_p(x_p - \overline{x})^2.$$

Dans le cas de variables continues triées par classes, on ne dispose pas des valeurs exactes x_i. La méthode la plus simple consiste à utiliser pour chaque classe la valeur centrale de la classe, c'est-à-dire la moyenne de ses extrémités [1].

Exemple : considérons l'exemple, utilisé les semaines passées, de la taille dans une promotion d'étudiants :

Taille (en cm)	[120,150[	[150,160[	[160,170[	[170,180[	[180,190[	[190, 210[
Centre (en cm)	135	155	165	175	185	200
Fréquence (en %)	1,7	12,8	37,8	36,0	9,3	2,3

Comme calculé en semaine 10, la taille moyenne vaut alors $169,31$ centimètres. Calculons la variance de la taille :

Taille (en cm)	[120,150[	[150,160[	[160,170[	[170,180[	[180,190[	[190, 210[
Centre (en cm)	135	155	165	175	185	200
Écart à la moyenne (en cm)	$135 - 169,31 = -34,31$	-14,31	-4,31	5,69	15,69	30,69
Carré de l'écart à la moyenne (en cm^2)	$(-34,31)^2 = 1177,18$	204,78	18,58	32,38	246,18	941,88
Fréquence (en %)	1,7	12,8	37,8	36,0	9,3	2,3

1. Il faudrait en toute rigueur rajouter un terme correctif pour prendre en compte la dispersion à l'intérieur de chaque classe (voir le polycopié de cours), mais s'il y a suffisamment de classes, ce terme correctif sera petit. Pour la suite du cours, nous ignorerons ce terme supplémentaire.

La variance est la moyenne des carrés des écarts à la moyenne :

$$\begin{aligned}\sigma^2 &= 0,017 \times 1177,18 + 0,128 \times 204,78 + 0,378 \times 18,58 + 0,36 \times 32,38 + 0,093 \times 246,18 + 0,023 \times 941,88 \\ &= 109,46 \text{ centimètres carrés.}\end{aligned}$$

Remarque : Si les observations ont une unité u, alors la variance a l'unité u^2. Par exemple, la variance est en centimètres carrés si on mesure des tailles en centimètres, en années carrées si l'on mesure des âges... La variance n'a pas d'interprétation simple, et n'est pas directement comparable aux données et aux autres paramètres (moyenne, médiane, etc.).

L'**écart-type** σ est définie comme la **racine carrée de la variance** :

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

L'écart-type, et non la variance, est un paramètre de dispersion que l'on peut comparer aux données.

Exemple : en gardant l'exemple de la taille dans une promotion d'étudiants : l'écart-type vaut [2] :

$$\sigma = \sqrt{109,46} = 10,5 \text{ centimètres.}$$

Les tailles des étudiants sont donc typiquement à une distance d'environ 10 centimètres de la taille moyenne.

Remarque : La définition de l'éart-type (racine carrée de la moyenne des carrés des écarts à la moyenne) peut paraître inutilement compliquée. Elle est en fait assez naturelle, et a de bonnes propriétés mathématiques. On peut la comparer à la distance euclidienne usuelle : si l'on considère deux points A et B dans l'espace, de coordonnées $(1, 2, 3)$ et $(2, 2, 2)$ respectivement dans un repère orthonormé, alors la distance entre A et B vaut, d'après le théorème de Pythagore,

$$\sqrt{(1-2)^2 + (2-2)^2 + (3-2)^2} = \sqrt{1+0+1} = \sqrt{2},$$

une formule que l'on pourrait décrire comme la racine carrée de la somme des carrés des écarts des coordonnées.

Écart inter-quartiles

L'**écart inter-quartiles** est simplement la différence entre les troisième et premier quartilee $Q3 - Q1$.

Exemple : en gardant l'exemple de la taille dans une promotion d'étudiants : le premier quartile est à 160 centimètres et le troisième quartile à 175 centimètres. L'écart inter-quartile est donc de $175 - 160 = 15$ centimètres.

Remarque : L'écart-type et l'écart-interquartiles ne sont pas directement comparables [3]. Il est plus pertinent de comparer les écarts-types (ou les écarts inter-quartiles) d'observations de la même variable dans des populations différentes.

2. Le terme correctif précédemment évoqué ferait passer la variance de $109,46$ centimètres carrés à $119,50$ centimètres carrés, et donc l'écart-type de $10,5$ centimètres à $10,9$ centimètres – une différence très faible.

3. Assez souvent, et en particulier pour les distributions gaussiennes, l'écart inter-quartile vaut environ $1,5$ fois l'écart-type. C'est le cas de l'exemple.