3 Analyse d'une série statistique

Il s'agit de caractériser une série par *une* ou *des* valeurs plus simples à manipuler que les tableaux et les graphiques et plus efficaces pour les comparaisons.

1- Valeurs typiques ou valeurs centrales

Les plus importantes de ces valeurs sont le mode, la médiane, les quartiles et la moyenne.

1.1 Le mode

Dans le cas discret, le mode (noté M_o) est la valeur de la variable pour laquelle l'effectif (ou la fréquence) est le plus élevé. Il se lit directement sur la série classée ou graphiquement sur le diagramme en bâtons.

Dans l'exemple 1, le mode (la valeur de la variable correspondant à l'effectif 74) vaut :

$$M_o = 1\,année \text{ (entière)},$$

ce qui signifie que ce sont les voitures âgées de $1\,an$ ou plus mais de moins de $2\,ans$ qui apparaissent en nombre le plus élevé.

Sur le diagramme en bâtons, le mode, (pour l'exemple 1 la valeur 1 de la variable) correspond au bâton le plus élevé. Il faut remarquer que :

- si dans la série apparaissent plusieurs valeurs **successives** de même effectif (ou de même fréquence) maximum, le mode est remplacé par un intervalle modal, comme le montre la *figure* 3.1.

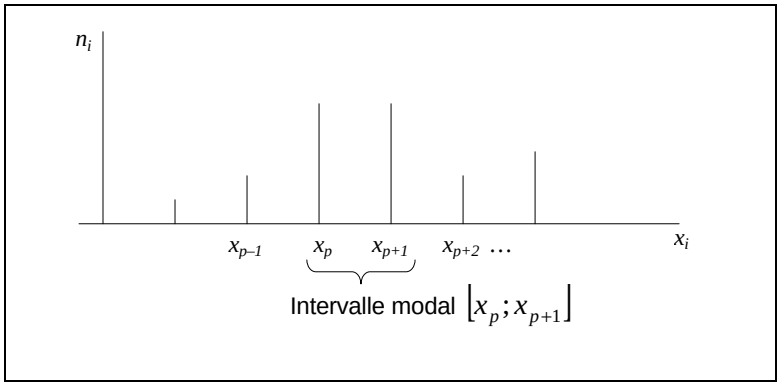

Figure 3.1

- Si deux ou plusieurs variables **non-successives** ont le même effectif (ou la même fréquence) maximum, la série est composée de deux ou plusieurs sous-populations.

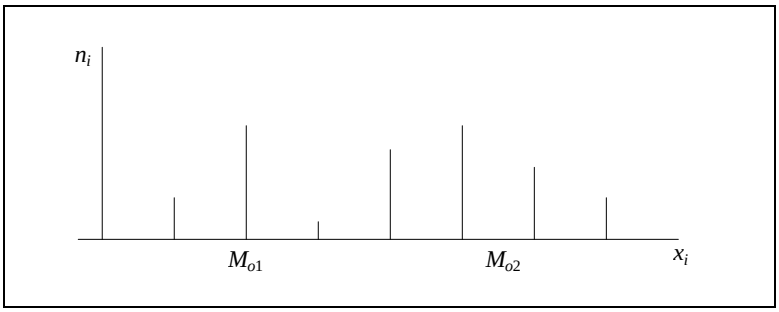

Figure 3.2

Dans le cas continu, la classe modale est celle dont l'effectif moyen (ou la fréquence moyenne) par unité d'intervalle de classe est le plus élevé. Il se lit directement sur la série classée ou graphiquement sur l'histogramme ou le polygone des effectifs ajustés (ou des fréquences ajustées).

- Si les classes situées immédiatement avant et après la classe modale ont sensiblement les mêmes effectifs, le mode est alors le centre de la classe modale.

  Dans l'exemple 2, la classe modale (d'effectif 12) est la classe $[425;475]$. Vu les valeurs des effectifs (7 et 8) des classes immédiatement voisines, le mode de la série de l'exemple 2 vaut :

$$M_o = centre\ de\ [425;475]$$
$$= 450\ euros$$

  La somme le plus souvent déboursée pour l'entretien annuel est 450 €.

- Si les classes adjacentes à la classe modale ont des effectifs très inégaux, il faut considérer que le mode se trouve plus proche de la classe dont l'effectif est le plus élevé.

  Le mode pourrait alors être calculé ou peut être lu sur l'histogramme.

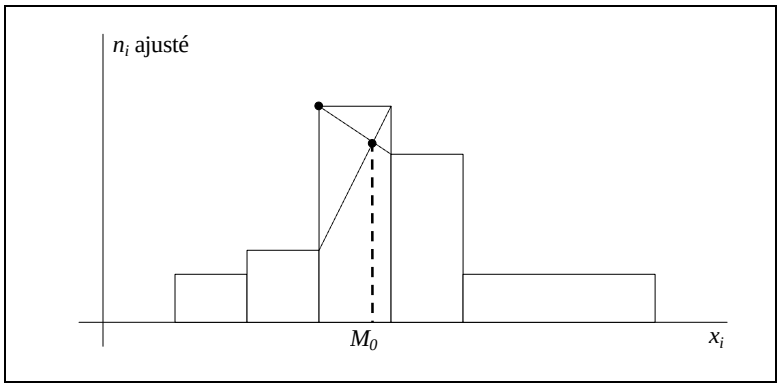

Figure 3.3

Il est encore possible de rencontrer des séries plurimodales.

Avantages et inconvénients

Le mode est de détermination aisée, que ce soit sur tableau ou sur graphique. Vu sa définition (valeur qui apparaît le plus souvent), son côté concret est utile aux économistes et aux études de marché.

En revanche, le mode n'est significatif que si l'effectif correspondant est nettement supérieur aux effectifs présentés par les autres valeurs de la variable et perd sa signification s'il n'est pas unique. De plus, dans le cas d'une même variable continue, deux répartitions différentes en classes pourraient conduire à des modes différents.

1.2 La médiane

La médiane est la valeur (quantitative) du caractère de l'élément qui se trouve « au milieu » de la série ordonnée.

Dans le cas discret

- Si le nombre d'éléments n est impair, la médiane (notée M_e) est la valeur qui occupe la valeur centrale. Par exemple, pour la série ordonnée de sept éléments :

$$\{5, 7, 8, ⑨, 10, 11, 14\}$$

L'élément central, c'est à dire la médiane vaut :

$$M_e = 9$$

Cet élément est, en fait, la valeur $x_{\frac{n+1}{2}}$ de numéro $\frac{n+1}{2}$.

- Si le nombre d'éléments n est pair, la médiane M_e est la moyenne arithmétique des deux éléments centraux de la série. Par exemple, pour la série ordonnée de huit éléments :

$$\{5, 7, 8, 9, 10, 11, 14, 16\}$$

Les deux éléments centraux sont 9 et 10 et la médiane vaut $M_e = \frac{9+10}{2} = 9{,}5$. Cette valeur n'est autre que la moyenne arithmétique des éléments $x_{\frac{n}{2}}$ et $x_{\frac{n}{2}+1}$:

$$M_e = \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1} \right)$$

Elle peut être fictive et ne pas se retrouver parmi les éléments de la série.

Il faut remarquer que la médiane peut se lire sur le tableau des effectifs cumulés (ou fréquences cumulées).

x_i	n_i	cn_i
0	60	60
1	74	134
2	72	206
3	27	
4	30	
5	10	
6	5	
7	1	
8	0	
9	1	280
	280	

Tableau 3.1

La médiane est la $\left(\frac{280}{2}+1\right)^{ième}$ ou la 141^e valeur, c'est à dire M_e=2 (la moitié des propriétaires de voiture ont un véhicule de moins de deux ans).

La médiane peut encore se lire sur le diagramme (en escaliers) des effectifs cumulés (ou fréquences cumulées) en repérant la ligne horizontale de $\frac{n}{2}$ (des $50\ \%$) et en retournant aux abscisses.

Dans le cas continu

La médiane M_e est la valeur de la variable qui atteint un effectif cumulé égal à $\frac{n}{2}$ (ou une fréquence cumulée égale à $50\ \%$).

Dans le tableau des effectifs cumulés (ou des fréquences cumulées), la classe médiane est la classe qui contient la médiane. Pour l'exemple 2, il faut considérer le *tableau* 3.2 :

Classes €	Centres des classes x_i	Effectifs n_i	Effectifs cumulés nc_i
[175 ; 225[	200	5	5
[225 ; 275[	250	4	9
[275 ; 325[	300	8	17
[325 ; 375[	350	7	24
[375 ; 425[	400	7	31
[425 ; 475[	450	12	43
[475 ; 525[	500	8	51
[525 ; 575[	550	8	59
[575 ; 625[	600	5	64
[625 ; 675[	650	2	66
[675 ; 725[	700	4	70
[725 ; 775[	750	2	72
[775 ; 825[	800	3	75
		75	

Tableau 3.2

$\frac{n}{2}=\frac{75}{2}=37{,}5$ et la médiane appartient à la 6e classe $[425;475[$. La M_e peut être prise, approximativement, égale au centre de cette classe (pour l'exemple 2, $M_e \approx 450$) ou pourrait être calculée par interpolation linéaire.

De nouveau, la médiane peut être lue sur le diagramme des effectifs cumulés (ou fréquences cumulées) en repérant l'horizontale des $50\ \%$ et en retournant aux abscisses, comme sur la *figure* 3.4.

Avantages et inconvénients

Le calcul de la médiane est facile, il aboutit à une valeur objective et concrète. La médiane donne une idée satisfaisante de la tendance générale de la série. De plus, elle n'est pas influencée par des valeurs aberrantes (anormalement grandes ou petites, hors norme par rapport au phénomène étudié).

En revanche, la médiane n'est pas un bon représentant des valeurs de la série puisqu'elle est uniquement basée sur les valeurs centrales et non sur toutes les observations. Elle ne prend pas non plus en compte l'étendue de la série. Elle se prête mal aux calculs ultérieurs.

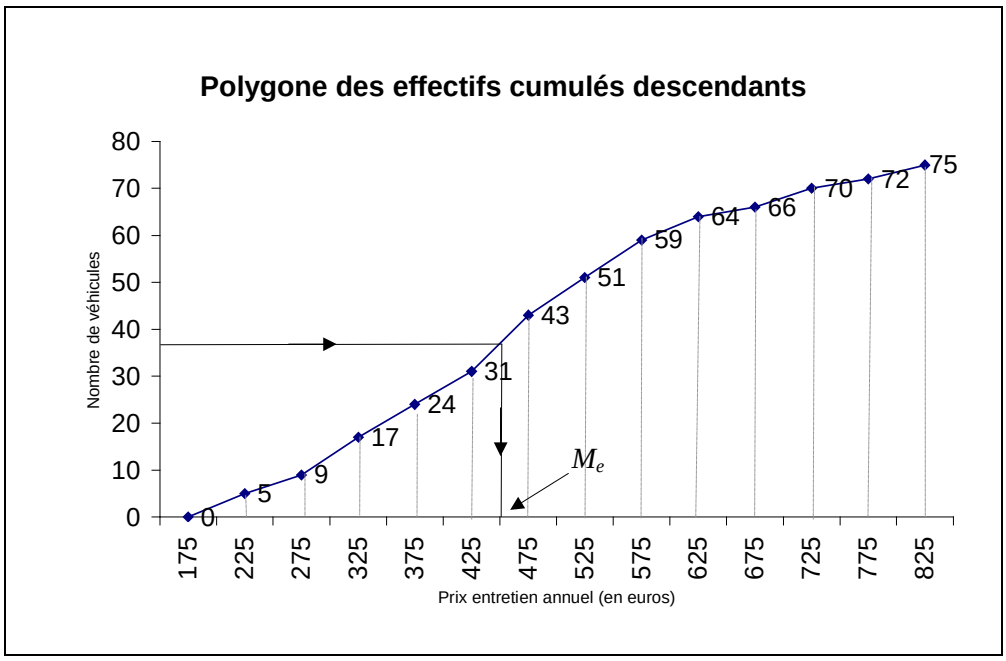

Figure 3.4

1.3 Les quartiles

Les quartiles sont les valeurs du caractère des éléments qui « divisent » en quatre parties égales la série ordonnée. Evidemment, la médiane est égale au 2e quartile Q_2.

Dans le cas discret

- Si le nombre d'éléments n est un multiple de 4, $n = 4k$, par exemple, pour la série ordonnée de 12 éléments $(12 = 4.3)$.

$$\{\underbrace{2\ 3\ 4}_{k=3}\ 5\ 6\ 7\ 8\ 9\ 10\ 11\ 12\ 13\}$$

Q_1 Q_2=M_e Q_3

$$Q_1 = \frac{x_k + x_{k+1}}{2} = 4{,}5$$

$$Q_3 = \frac{x_{3k} + x_{3k+1}}{2} = 10{,}5$$

- Si le nombre d'éléments n'est pas un multiple de 4, alors $n = 4k+1$ ou $n = 4k+2$ ou $n = 4k+3$ et :

$$Q_1 = x_{k+1}$$

$$Q_3 = x_{n-k}$$

Par exemple, dans la série ordonnée

$$\{2\ 3\ 4\ 5\ 6\ 7\ 8\ 9\ 10\ 11\}$$

Q_1 Q_3

$n = 10 = 4.2 + 2$ et $k = 2$

Ainsi : $Q_1 = x_{k+1} = x_3 = 4$

et $Q_2 = x_{n-k} = x_{10-2} = x_8 = 9$

Comme la médiane, les quartiles peuvent se lire sur le tableau des effectifs cumulés. Pour l'exemple 1, $n = 280 = 4.70$: $Q_1 = 1$ et $Q_3 = 3$ (le quart des propriétaires ont un véhicule de moins d'un an, les trois quarts ont un véhicule de moins de 3 ans).

Les quartiles peuvent aussi se lire sur le diagramme en escaliers des effectifs cumulés (ou des fréquences cumulées) en repérant la ligne de $\frac{n}{4}$ (des $25\,\%$) et la ligne de $\frac{3n}{4}$ (des $75\,\%$) et en retournant aux abscisses.

Dans le cas continu

Le premier quartile Q_1 est la valeur de la variable pour laquelle l'effectif cumulé égal à $\frac{n}{4}$ (ou une fréquence cumulée de $25\,\%$) est atteint. Le troisième quartile Q_3 est la valeur de variable pour laquelle l'effectif cumulé égal à $\frac{3n}{4}$ (ou une fréquence cumulée de $75\,\%$) est atteint. La détermination de Q_1 et Q_3 sur le tableau des effectifs cumulés (fréquences cumulées) se fait en utilisant la même méthode que pour déterminer la médiane.

1.4 La moyenne (arithmétique pondérée)

La moyenne arithmétique $\bar{x}$ est le quotient de la somme des valeurs des éléments de la série par le nombre d'éléments de la série. Ainsi, si la série est discrète et non recensée :

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Si la série est recensée :

$$\bar{x} = \frac{\sum_{i=1}^{J} n_i x_i}{n}$$

Si la série est répartie en classes :

$$\bar{x} = \frac{\sum_{i=1}^{J} n_i x_i}{n}$$

où x_i est le centre de la classe d'effectif n_i.

Pratiquement, le calcul de la moyenne peut être un peu simplifié en complétant le tableau des données par une colonne où sont calculées les valeurs successives $n_i x_i$. Il s'agit alors d'additionner les valeurs obtenues dans cette colonne et de diviser cette somme par l'effectif total n pour obtenir $\bar{x}$.

Avantages et inconvénients

Le calcul de la moyenne utilise toutes les valeurs de la série et est une valeur très stable (qui augmente avec n). Mais les calculs sont parfois très longs et les valeurs exceptionnelles peuvent influencer fortement ce nombre.

1.5 Position relative de M_0, M_e et $\bar{x}$

Dans le cas d'une série symétrique, comme le *tableau* 3.3, si la symétrie est parfaite, M_o, M_e et $\bar{x}$ sont égaux et la courbe des effectifs se présente comme une courbe en cloche (*figure* 3.5).

x_i	n_i
4	2
5	3
6	4
7	5
8	4
9	3
10	2

Tableau 3.3

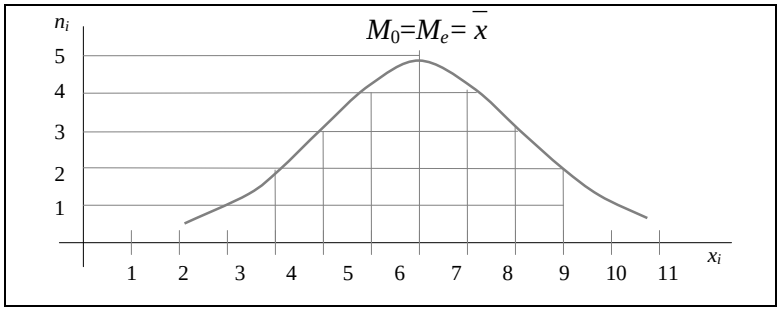

Figure 3.5

Dans la pratique, la distribution peut être presque symétrique et $M_o \approx M_e \approx \bar{x}$. Sinon, la distribution est dissymétrique, comme le montre la *figure* 3.6.

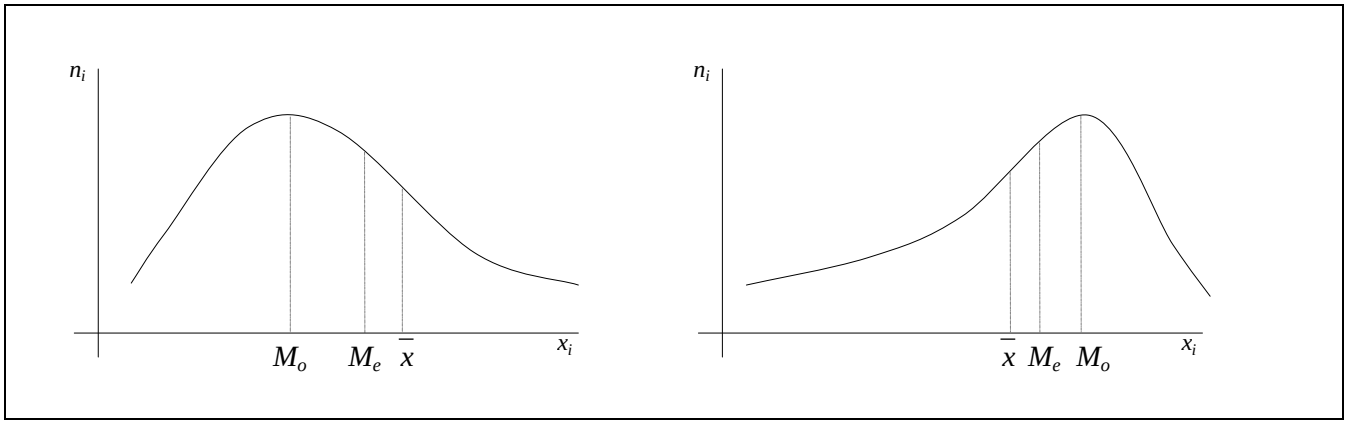

Figure 3.6

Et dans le cas des distributions légèrement dissymétriques, la formule de ***Pearson*** positionne M_o, $\bar{x}$ et M_e par :

$$M_o = \bar{x} - 3\left(\bar{x} - M_e\right)$$

ou

$$M_e = \bar{x} + \frac{1}{3}\left(M_o - \bar{x}\right)$$

2- Caractéristiques de dispersion

Les valeurs centrales ne suffisent pas pour caractériser une série statistique. Elles fournissent une idée de l'ordre de grandeur des observations mais ne donnent pas d'idée de la répartition des valeurs dans la série et ne suffisent pas à la comparaison de différentes séries.

Par exemple, soient les deux séries de valeurs :

$$\text{et} \quad \begin{array}{l} \text{série } A = \{78\,,\,79\,,\,79\,,\,80\,,\,80\,,\,80\,,\,81\,,\,81\,,\,82\} \\ \text{série } B = \{40\,,\,60\,,\,60\,,\,80\,,\,80\,,\,80\,,\,100\,,\,100\,,\,120\} \end{array}$$

Chacune de ces deux séries a une moyenne égale à 80 et un mode égal à 80. Pourtant les deux séries sont très différentes, la deuxième étant manifestement plus dispersée que la première. Elles sont représentées sur la *figure* 3.7.

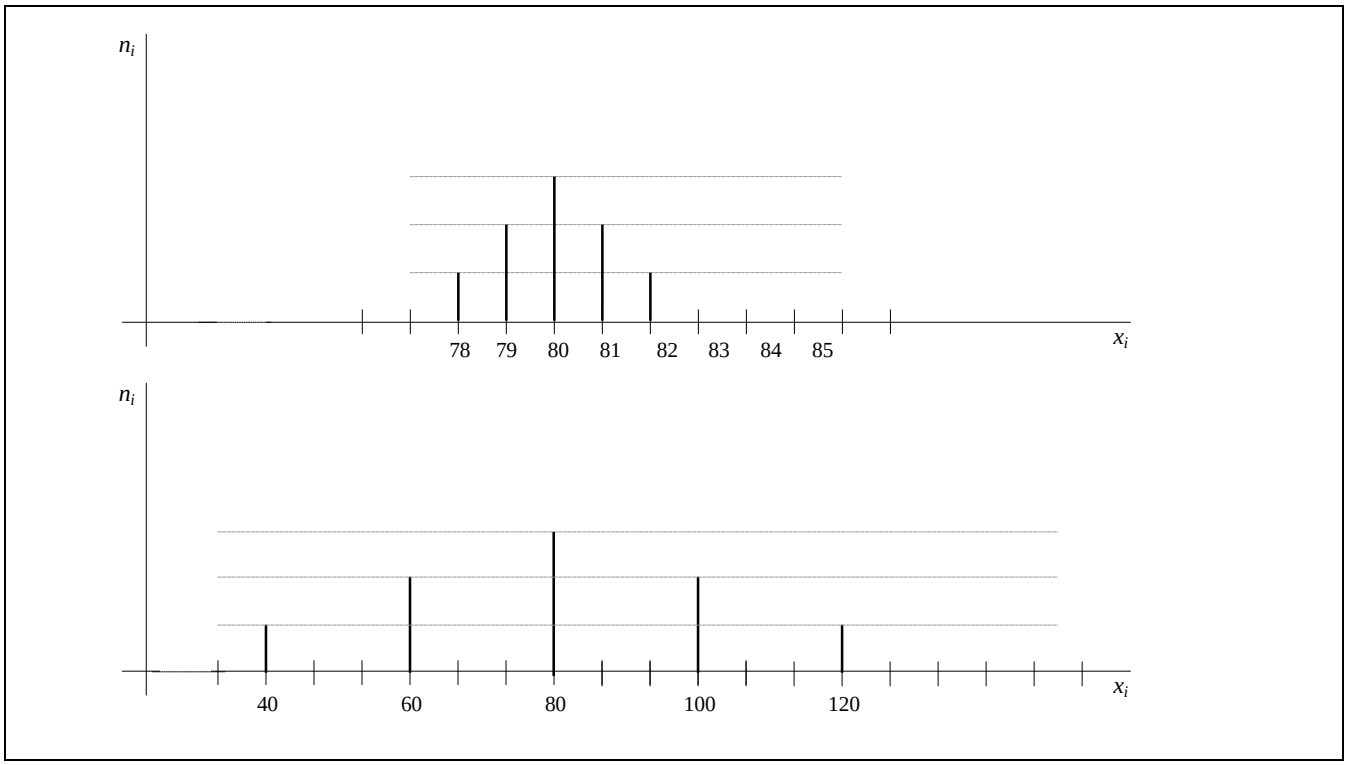

Figure 3.7

Il faut donc définir et calculer des paramètres de dispersion qui permettent de compléter l'analyse d'une série statistique. Certains de ces paramètres mesurent les écarts entre deux observations particulières, comme l'étendue ou l'intervalle interquartile ; d'autres, les écarts des observations à la moyenne, comme l'écart absolu moyen, la variance ou l'écart type.

2.1 L'étendue

L'étendue e d'une série est la différence entre la plus grande et la plus petite des valeurs de la série.

Pour la *série A* considérée dans les lignes qui précèdent, l'étendue $e_A = 82 - 78 = 4$ tandis que pour la *série B*, l'étendue devient : $e_B = 120 - 40 = 80$.

Avantages et inconvénients

L'étendue est le plus simple des paramètres de dispersion : il est plus simple à calculer et donne rapidement une idée de la série. Dans ce sens, il est utile dans les contrôles industriels de fabrication, dans les laboratoires, etc. Néanmoins, l'étendue ne dépend pas de tous les éléments de la série, ne tient pas compte de la répartition des éléments entre les valeurs extrêmes et, vu sa définition, est très sensible aux valeurs aberrantes.

2.2 L'intervalle interquartile

L'intervalle interquartile est la différence entre le 3e et le 1e quartile :

$$Q = Q_3 - Q_1$$

En comparant les intervalles interquartiles des *séries A* et *B*, $Q_A = 81 - 79 = 2$ et $Q_B = 100 - 60 = 40$, il apparaît que la différence entre ces étendues est moins sensible que celle entre les deux étendues des séries initiales. Ceci est dû au fait que l'intervalle interquartile a éliminé les valeurs hors normes 40 et 120.

2.3 L'écart absolu moyen

L'idée première a été d'évaluer les écarts algébriques des valeurs des observations par rapport à la position de la moyenne et d'en calculer la somme :

$$\sum_{i=1}^{J} n_i \left(x_i - \bar{x}\right)$$

Mais

$$\begin{aligned}\sum_{i=1}^{J} n_i \left(x_i - \bar{x}\right) &= \sum_{i=1}^{J} n_i.x_i - \sum_{i=1}^{J} n_i.\bar{x} \\ &= \sum_{i=1}^{J} n_i.x_i - n.\bar{x} \\ &= n.\bar{x} - n.\bar{x} \\ &= 0\end{aligned}$$

Les écarts à la moyenne peuvent être pris en valeur absolue et il s'agit alors de calculer l'écart absolu moyen (par rapport à la moyenne arithmétique) :

$$E = \frac{1}{n}\sum_{i=1}^{n} \left|x_i - \bar{x}\right|$$

ou pour une série recensée : $E = \frac{1}{n}\sum_{i=1}^{J} n_i \left|x_i - \bar{x}\right|$, c'est à dire la moyenne arithmétique des écarts absolus par rapport à la moyenne arithmétique $\bar{x}$.

Un tel calcul était historiquement difficile à réaliser, une valeur absolue étant un concept difficile à manier mathématiquement. Ainsi, malgré la signification concrète et facile à concevoir de cette grandeur, elle est peu utilisée. A la notion d'écart absolu moyen, il a été préféré celle d'écart type, moins facile à concevoir mais qui se prête mieux aux calculs algébriques.

2.4 La variance et l'écart type

L'option retenue a été de remplacer le calcul des valeurs absolues $\left|x_i - \bar{x}\right|$ par le calcul des valeurs de $\left(x_i - \bar{x}\right)^2$ et de définir la variance V d'une série statistique par :

$$V = \frac{1}{n}\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2$$

ou pour une série recensée :

$$V = \frac{1}{n}\sum_{i=1}^{J} n_i \left(x_i - \bar{x}\right)^2$$

La variance V est donc la moyenne arithmétique des carrés des écarts des valeurs de la série à la moyenne arithmétique de la série. Pour les *séries A* et *B* proposées dans les lignes précédentes :

$$A = \{78, 79, 79, 80, 80, 80, 81, 81, 82\} \quad \text{avec} \quad \bar{x}_A = 80$$

$$\begin{aligned} Variance = V_A &= \frac{1}{9}\left[1.(78-80)^2 + 2.(79-80)^2 + 3.(80-80)^2 + 2.(81-80)^2 + 1.(82-80)^2\right] \\ &= \frac{1}{9}(4+2+0+2+4) \\ &= \frac{12}{9} \\ &\approx 1{,}33 \end{aligned}$$

$$B = \{40, 60, 60, 80, 80, 80, 100, 100, 120\} \quad \text{avec} \quad \bar{x}_B = 80$$

$$\begin{aligned} Variance = V_B &= \frac{1}{9}\left[1.(80-40)^2 + 2.(60-80)^2 + 3.(80-80)^2 + 2.(100-80)^2 + 1.(120-80)^2\right] \\ &= \frac{1}{9}(1600+800+0+800+1600) \\ &= \frac{3360}{9} \\ &\approx 373{,}33 \end{aligned}$$

La variance ne s'exprime pas en unité de la variable mais en unité au carré. Pour en revenir à la notion de dispersion autour de la moyenne et donc à l'unité de la variable, il faut introduire l'écart type.

L'écart type d'une série statistique est, par définition, la racine carrée de la variance :

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

Ainsi, les écarts types des *séries A* et *B* sont respectivement :

$$\sigma_A = \sqrt{V_A} = \sqrt{\frac{12}{9}} = 1{,}1547... \text{ (unités de la variable)}$$

et

$$\sigma_B = \sqrt{V_B} = \sqrt{\frac{3360}{9}} = 19{,}3218... \text{ (unité de la variable)}$$

Ce paramètre (de dispersion) met bien en évidence que la *série B* est plus étalée que la *série A*.

Il est possible de simplifier le calcul de la variance en utilisant la formule :

$$V = \frac{1}{n}\sum_{i=1}^{J} n_i x_i^2 - \bar{x}^2$$

Cette dernière est équivalente à la définition de la variance. En effet :

$$
\begin{aligned}
V &= \frac{1}{n}\sum_{i=1}^{J} n_i\left(x_i - \bar{x}\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{J} n_i\left(x_i^2 - 2\bar{x}.x_i + \bar{x}^2\right) \\
&= \frac{1}{n}\sum_{i=1}^{J} n_i.x_i^2 - \frac{2\bar{x}}{n}\sum_{i=1}^{J} n_i.x_i + \frac{\bar{x}^2}{n}\sum_{i=1}^{J} n_i \\
&= \frac{1}{n}\sum_{i=1}^{J} n_i.x_i^2 - \frac{2\bar{x}}{n}.n.\bar{x} + \frac{\bar{x}^2}{n}.n \\
&= \frac{1}{n}\sum_{i=1}^{J} n_i.x_i^2 - \bar{x}^2
\end{aligned}
$$

qui est bien l'expression annoncée ci-dessus.

Le calcul pratique de la variance (donc de l'écart type) implique de compléter les tableaux principaux de calculs par une colonne des produits $n_i.x_i^2$.

Pour établir cette colonne facilement il est suggéré de multiplier les valeurs de la colonne $n_i.x_i$ déjà calculée (pour trouver $\bar{x}$) par les valeurs de la colonne x_i. Il suffit alors d'additionner les valeurs de cette nouvelle colonne et de diviser le résultat par n pour obtenir le premier terme de la formule de la variance. En soustrayant alors $\bar{x}^2$ de ce premier terme, la variance est calculée.

Pour l'exemple de référence 1 :

Valeurs x_i	Effectifs n_i	$n_i x_i$	$n_i x_i^2$
0	60	0	0
1	74	74	74
2	72	144	288
3	27	81	243
4	30	120	480
5	10	50	250
6	5	30	180
7	1	7	49
8	0	0	0
9	1	9	81
	280	515	1645

Tableau 3.4

$\bar{x} = 1{,}8393\ années$, $V = 2{,}492028...$ et $\sigma = 1{,}5786...\ années$

Pour l'exemple 2, il faut consulter le *tableau* 3.5

Classes	Centres des classes x_i	Effectifs n_i	$n_i x_i$	$n_i x_i^2$
[175 ; 225[	200	5	1000	200000
[225 ; 275[	250	4	1000	250000
[275 ; 325[	300	8	2400	720000
[325 ; 375[	350	7	2450	857500
[375 ; 425[	400	7	2800	1120000
[425 ; 475[	450	12	5400	2430000
[475 ; 525[	500	8	4000	2000000
[525 ; 575[	550	8	4400	2420000
[575 ; 625[	600	5	3000	1800000
[625 ; 675[	650	2	1300	845000
[675 ; 725[	700	4	2800	1960000
[725 ; 775[	750	2	1500	1125000
[775 ; 825[	800	3	2400	1920000
		75	34450	17647500

Tableau 3.5

$\bar{x} = 459{,}33\,euros$, $V = 24313$ et $\sigma = 155{,}9259\,euros$

Utilisation de l'écart type comme caractéristique de dispersion

Quelle que soit la distribution statistique étudiée, un intervalle dont les bornes sont $\bar{x} - 2\sigma$ et $\bar{x} + 2\sigma$ contient toujours au moins $75\,\%$ des unités constituant la population étudiée.

Plus généralement, un intervalle dont les bornes sont $\bar{x} - t\sigma$ et $\bar{x} + t\sigma$ contient toujours une proportion des effectifs de la population **au moins** égale à $1 - \frac{1}{t^2}$ (c'est la loi de ***Tchebicheff***).

Effet d'un changement de variable linéaire sur la moyenne et l'écart type (réduction des données)

Dans le but de comparer des séries de données différentes entre elles, c'est à dire de comparer ces séries sur une même échelle (de même origine et de même unité de mesure), il est possible d'effectuer le changement de variable

$$x_i \to u_i = \frac{x_i - \bar{x}}{\sigma}$$

qui possède deux propriétés remarquables :

$$\bar{u} = 0 \quad \text{et} \quad V_u = 1 \quad (\text{ou } \sigma_u = 1)$$

$$\begin{aligned}\bar{u} &= \frac{1}{n}\sum_i n_i.u_i \\ &= \frac{1}{n}\sum_i n_i\left(\frac{x_i-\bar{x}}{\sigma}\right) \\ &= \frac{1}{n\sigma}\sum_i n_i\left(x_i-\bar{x}\right) \\ &= \frac{1}{n\sigma}\left(\sum_i n_i.x_i-\sum_i n_i.\bar{x}\right) \\ &= \frac{1}{n\sigma}\left(n.\bar{x}-n.\bar{x}\right) \\ &= 0\end{aligned}$$

et

$$\begin{aligned}V_u &= \frac{1}{n}n_i.u_i^2-\bar{u}^2 \\ &= \frac{1}{n}\sum_i n_i\left(\frac{x_i-\bar{x}}{\sigma}\right)^2-0 \\ &= \frac{1}{n\sigma^2}\sum_i n_i\left(x_i-\bar{x}\right)^2 \\ &= \frac{1}{n\sigma^2}\sum_i n_i\left(x_i^2-2x_i.\bar{x}+\bar{x}^2\right) \\ &= \frac{1}{n\sigma^2}\left(\sum_i n_i.x_i^2-2\bar{x}\sum_i n_i.x_i+\bar{x}^2\sum_i n_i\right) \\ &= \frac{1}{n\sigma^2}\left(\sum_i n_i.x_i^2-2\bar{x}.n.\bar{x}+\bar{x}^2.n\right) \\ &= \frac{1}{n\sigma^2}\left(\sum_i n_i.x_i^2-n\bar{x}^2\right) \\ &= \frac{1}{\sigma^2}\left(\frac{1}{n}\sum_i n_i.x_i^2-\bar{x}^2\right) \\ &= \frac{1}{\sigma^2}.\sigma^2 \\ &= 1\end{aligned}$$

Exemple

Un groupe de 10 étudiants a été questionné au cours de mathématiques, en « théorie » et en « exercices ». En réduisant les deux séries de données (cotes de théorie et cotes d'exercices), les résultats peuvent être comparés sur une même échelle. Le *tableau* 3.6 affiche ces résultats.

Les cotes de l'étudiant 1 montrent qu'il est meilleur en théorie qu'en exercices, tout comme l'étudiant 6. En revanche, par rapport aux résultats moyens du groupe, l'étudiant 1 est toujours meilleur en théorie qu'en exercices, tandis que l'étudiant 6 est (toujours par rapport aux résultats moyens du groupe) est meilleur en exercices qu'en théorie. La situation de l'étudiant 6 se retrouve pour les étudiants 8 et 10.

Etudiants	Théorie /20	Exercices /20	Théorie réduite	Exercices réduite
1	12	8	0,198	–0,445
2	14	10	0,766	0,364
3	5	10	–1,788	0,364
4	8	4	–0,936	–2,066
5	10	8	–0,369	–0,445
6	12	10	0,198	0,364
7	18	12	1,901	1,175
8	12	11	0,198	0,770
9	8	6	–0,936	–1,256
10	14	12	0,766	1,175
	$\bar{x} = 11{,}3$	$\bar{y} = 9{,}1$	$\bar{u}_x = 0$	$\bar{u}_y = 0$
	$\sigma_x = 3{,}52$	$\sigma_y = 2{,}47$	$\sigma_{u_x} = 1$	$\sigma_{u_y} = 1$

Tableau 3.6

3- Les autres caractéristiques (caractéristiques de variation relative)

Lorsqu'il faut comparer des séries dont les moyennes sont identiques ou sensiblement les mêmes, l'écart type permet de connaître immédiatement la série dont la dispersion est la plus grande.

S'il faut comparer des séries dont les natures et les moyennes sont très différentes, il faut définir des nombres indépendants des unités (contrairement aux caractéristiques de dispersion qui dépendent des unités de mesure des observations).

Les deux caractères les plus simples sont la variation de la série et la symétrie de la série.

3.1 Le coefficient de variation

Le coefficient de variation CV est défini par :

$$CV = \frac{\sigma}{\bar{x}} \quad \text{ou} \quad CV = \frac{\sigma}{\bar{x}} \times 100 \ \text{en \%}$$

et est, par définition, indépendant de l'unité de mesure de la variable (comme rapport de deux grandeurs de même unité).

Il faut remarquer qu'il est nécessaire que la moyenne $\bar{x}$ soit non-nulle. Le CV est utilisé lorsque les données sont rapportées à une origine fixe et qu'elles sont en général toutes positives.

Exemple

Dans un pays déterminé, ont été relevées les ventes d'automobiles et les ventes de carburant entre 1977 et 1982, c'est sur le *tableau* 3.7 que ces valeurs sont présentées.

Années	Ventes d'automobiles (en 10^3 véhicules)	Ventes de carburant (en 10^3 m^3)
1977	100	420
1978	115	450
1979	80	410
1980	85	410
1981	106	446
1982	120	450
	$n_v = 606$	$n_c = 431$

Tableau 3.7

Le nombre moyen de véhicules vendus est de 101 *milliers de véhicules* ($\bar{x}_v = 101.10^3$ *voitures*) et l'écart type correspondant est de 14,6 *milliers de véhicules* ($\sigma_v = 14{,}6.10^3$ *voitures*).

Le volume moyen de carburant vendu est de 431 *milliers de* m^3 ($\bar{x}_c = 431.10^3\ m^3$) et l'écart type correspondant est de 18 *milliers de* m^3 ($\sigma_c = 18{,}02.10^3\ m^3$).

Les grandeurs ne sont pas mesurées dans les mêmes unités et ne sont pas comparables. En revanche, les coefficients de variation des séries « vente d'automobiles » et « vente de carburant » sont respectivement égaux à :

$$CV_v = \frac{\sigma_v}{x_v} \times 100 \approx 14{,}45\,\% \quad \text{et} \quad CV_c = \frac{\sigma_c}{x_\sigma} \times 100 \approx 4{,}18\,\%$$

La variation des ventes de voitures est nettement supérieure à la variation des ventes de carburant.

3.2 le coefficient de dissymétrie

Le coefficient de dissymétrie C est défini par :

$$C = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1}$$

et calculé à partir des quartiles.

Trois situations sont possibles :

- si $C = 0$, il peut être présumé que la série est symétrique ;
- si $C > 0$, alors $Q_3 - Q_2 > Q_2 - Q_1$ et la série est étalée à droite ;
- si $C < 0$, cela signifie que $Q_3 - Q_2 < Q_2 - Q_1$ et la série est étalée à gauche.