CHAPITRE 3

INDICATEURS DE DISPERSION ET D'ASYMÉTRIE

- → *Étendue*
- → *Quartiles, déciles, centiles*
- → *Intervalle interquartile relatif*
- → *Variance et écart-type*
- → *Coefficient de variation*
- → *Coefficients d'asymétrie*

REPÈRES

La tendance centrale est la première caractéristique attachée à une série statistique, mais — quel que soit l'indicateur choisi — elle ne mesure qu'un aspect de la distribution.

Ainsi deux groupes de travailleurs peuvent avoir un même salaire moyen de 8 000 F par mois avec des répartitions très différentes : dans le premier groupe, chacun gagne exactement 8 000 F, tandis que dans le second, les salaires vont de 3 500 F à 60 000 F. Les deux séries diffèrent par leur *dispersion*.

Les indicateurs présentés quantifient la notion de dispersion d'une série statistique numérique.

1. Étendue - Intervalle interquartile

a) Étendue

L'*étendue* d'une série statistique est l'écart entre la plus grande et la plus petite valeur.

Exemple. L'étendue de la série « 5, 8, 2, 9, 5, 23, 11 » est 21 (23 – 2).

b) Quartiles - Intervalle interquartile

Pour limiter l'effet des valeurs les plus marginales, on préfère à l'étendue l'*intervalle interquartile*, qui est l'étendue de la série, privée de ses deux quarts extrêmes. Il contient donc la « moitié centrale » des observations.

Plus précisément, soit une série statistique numérique ordonnée par valeurs croissantes ; on définit d'abord les *quartiles :*

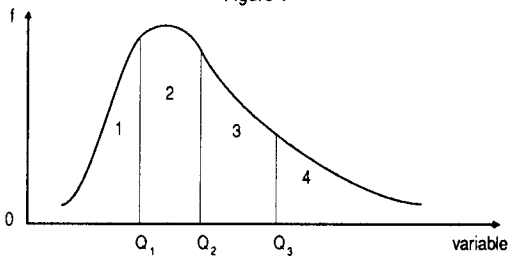
– le premier quartile, noté Q_1, est la valeur de la série telle qu'il y ait un quart des observations qui lui soient inférieures et les trois quarts qui lui soient supérieures ;

– le second quartile, noté Q_2, est la valeur de la série qui sépare les deux premiers quarts des deux derniers, c'est donc la médiane ;

– le troisième quartile, noté Q_3, est la valeur de la série telle qu'il y ait les trois quarts des observations qui lui soient inférieures et un quart qui lui soient supérieures.

L'intervalle interquartile est l'écart : $Q_3 - Q_1$ (fig. 1).

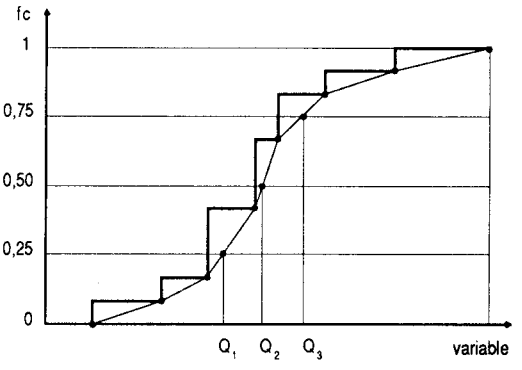
Figure 1



Nota : les surfaces 1, 2, 3 et 4 sont égales.

Le plus souvent, les séries étant données en classes, on estime les quartiles Q_1 et Q_3 par interpolation, c'est-à-dire comme les points où le polygone de fréquences cumulées coupe les lignes horizontales : fc = 0,25 et fc = 0,75 (fig.2).

Figure 2



c) Autres coefficients - Déciles, centiles

Les quartiles permettent de construire d'autres coefficients autorisant dans certaines conditions la comparaison des séries statistiques d'échelle ou de nature différentes.

● L'*intervalle interquartile relatif* $\frac{Q_3 - Q_1}{Q_2}$ donne une mesure de la dispersion d'une série, indépendante de l'unité employée.

● Le coefficient $\frac{Q_3 - Q_2}{Q_2 - Q_1}$ donne une mesure de l'asymétrie, également indépendante de l'unité employée.

● Selon la même méthode que pour les quartiles, on définit les *déciles* : $D_1, D_2, \ldots, D_9$ qui séparent les observations ordonnées en dixièmes successifs. Pour des statistiques très abondantes, on définit de même des *centiles* : $C_1, C_2, C_3, \ldots, C_{99}$ qui séparent les centièmes de la population observée.

Ces divisions définissent des découpages standard des séries statistiques. Elles permettent par exemple un examen ou des comparaisons fines de séries différentes.

2. Variance - Écart-type

Si l'on mesure la tendance centrale d'une série statistique par sa moyenne, il est naturel de considérer la dispersion par rapport à cette dernière. Cela peut être fait de différentes façons.

a) L'écart moyen

C'est la moyenne arithmétique de la valeur absolue des écarts entre les observations et leur moyenne arithmétique.

$$E = \frac{1}{N} \sum_i n_i \left| x_i - \bar{x} \right|.$$

b) Variance

Le maniement des valeurs absolues n'étant pas toujours commode, il est préférable de mesurer l'apport d'une observation à la dispersion par le carré de son écart à la moyenne ; et la *variance*, notée « var », est la moyenne des carrés des écarts à la moyenne.

Nous avons deux formules de définition, selon qu'il s'agit d'une série simple ou d'une série regroupée en classes.

● Pour une série simple : $x_1, x_2, \ldots, x_N$, la variance s'écrit :

$$\text{var}(x) = \frac{1}{N} \cdot \sum \left(x_i - \bar{x}\right)^2,$$

en notant comme à l'accoutumée $\bar{x}$ la moyenne de la série.

● Pour une série regroupée par classes de valeurs : $x_1, x_2, \ldots, x_k$ et d'effectifs associés : $n_1, n_2, \ldots, n_k$, la variance s'écrit :

$$\text{var}(x) = \frac{\sum n_i \cdot \left(x_i - \bar{x}\right)^2}{\sum n_i} ;$$

(expression que nous appellerons « formule de définition » de la variance), ou bien, en utilisant les fréquences f_i :

$$\text{var}(x) = \sum f_i \cdot \left(x_i - \bar{x}\right)^2.$$

c) L'écart-type

La variance n'est pas mesurée dans la même unité que la série étudiée, mais dans son carré ; ainsi, si les x_i sont des mètres, var(x) est en mètres carrés. Pour cette raison, on utilise plutôt comme indicateur de la dispersion la racine carrée de la variance ou *écart-type*, noté σ (la lettre grecque « sigma minuscule », sans autre rapport avec le symbole « Σ » vu au chapitre précédent) :

$$\sigma(x) = \sqrt{\text{var}(x)}.$$

L'écart-type est donc exprimé dans la même unité que la série étudiée.

d) Le coefficient de variation

Le *coefficient de variation* est défini comme le rapport de l'écart-type à la moyenne :

$$V = \frac{\sigma(x)}{\bar{x}}.$$

Cet indicateur, tel ceux définis ci-dessus (cf. partie **1-c** et suivantes), est un nombre « sans dimension », c'est-à-dire indépendant des unités de mesure utilisées. Il permet de ce fait de comparer des séries d'échelle, voire de nature différentes.

e) Coefficients d'asymétrie

Outre la mesure de l'asymétrie proposée dans la partie **1-c**, on rencontre fréquemment deux autres coefficients utilisant l'écart-type :

● *Le coefficient de Pearson*

Dans le cas des séries unimodales, l'asymétrie peut être mesurée par le rapport :

$$CP = \frac{\bar{x} - Mo}{\sigma(x)},$$

c'est-à-dire l'écart de la moyenne au mode rapporté à l'écart-type.

● *Le coefficient de Fisher*

Ce coefficient d'asymétrie, d'usage plus général mais moins simple à calculer, a pour expression, dans le cas d'une série simple :

$$CF = \frac{\Sigma(x_i - \bar{x})^3}{[\sigma(x)]^3}.$$

Une valeur positive de ces coefficients indique une asymétrie à droite ; une valeur négative implique au contraire une asymétrie à gauche.

f) Calculs et formule développée

Si on doit calculer un écart-type à l'aide d'une calculatrice, il est nécessaire de commencer par calculer la variance de la série.

Soit le cas d'une série regroupée, de valeurs : $x_1, x_2, \ldots, x_k$ et de fréquences associées : $f_1, f_2, \ldots, f_k$. Ayant préalablement déterminé la moyenne $\bar{x}$, on peut remplir un tableau dont les colonnes successives donnent les valeurs : x_i, les fréquences : f_i, les écarts à la moyenne : $(x_i - \bar{x})$, puis les carrés de ces écarts : $(x_i - \bar{x})^2$ et enfin les produits : $f_i \cdot (x_i - \bar{x})^2$.

La variance est la somme de cette dernière colonne ; enfin, il ne faut pas oublier d'en prendre la racine carrée pour obtenir l'écart-type cherché.

Si on dispose d'un micro-ordinateur, un tel calcul par colonnes se fait très simplement avec un tableur ; mais certains tableurs possèdent même les fonctions variance et écart-type.

Le calcul à la main est parfois simplifié par l'emploi d'une autre expression de la variance, que nous qualifierons de « formule développée » de la variance.

La variance est encore égale à la moyenne des carrés des valeurs, diminuée du carré de la moyenne des observations ; ce qui s'écrit, par exemple, dans le cas d'une série regroupée :

$$\text{var}(x) = \Sigma f_i \cdot x_i^2 - \bar{x}^2.$$

Le calcul des carrés des observations et de leur moyenne est plus rapide que l'élaboration du tableau décrit plus haut.

En fait, la plupart des calculatrices statistiques ou mathématiques calculent directement l'écart-type à partir des valeurs des observations.