Les caractéristiques de tendance centrale (2)

1. La moyenne arithmétique

1.1 Définition

La ***moyenne arithmétique*** $\bar{x}$ est égale à la somme des valeurs observées $x_1, x_2, \ldots, x_n$ divisée par le nombre d'observations *n* :

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

Dans cette formule, une valeur x_i est prise en compte autant de fois qu'elle a été observée. Dès lors, en utilisant les fréquences cette formule peut se réécrire :

$$\bar{x} = \frac{\sum_{i=1}^{p} n_i x_i}{n} \text{ ou encore } \bar{x} = \sum_{i=1}^{p} f_i x_i \ .$$

Dans le cas d'une distribution groupée, la formule de définition de la moyenne ne peut être directement appliquée car on ne connaît pas les valeurs exactes prises par la variable statistique.

Par convention, on choisit comme valeurs les centres de chaque classe (c_i) :

$$c_i = \frac{borne\ gauche + borne\ droite}{2}$$

Cette convention revient à supposer soit que toutes les observations à l'intérieur d'une classe sont groupées en son centre, soit que les observations sont réparties uniformément à l'intérieur de chaque classe.

Lorsque ces hypothèses ne sont pas vérifiées, le calcul résultant de cette convention ne donne pas nécessairement le résultat qui aurait été obtenu sur données individuelles. Cependant, l'erreur reste faible lorsqu'on se trouve en présence d'une distribution symétrique car les erreurs commises de part et d'autre de la valeur centrale se compensent. Ce n'est pas le cas dans les distributions non-symétriques : on peut alors réduire cet inconvénient en diminuant l'amplitude de chaque classe.

1.2 Propriétés

P1 *La somme des écarts entre valeurs observées et la moyenne arithmétique est nulle*

$$\sum_{i=1}^{p} n_i\left(x_i - \overline{x}\right) = 0$$

<u>Démonstration</u>

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

Comme $\overline{x}$ ne dépend pas de *i*, on peut écrire:

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

Par définition on a $\sum_{i=1}^{p} n_i = n$, on peut écrire:

$$\sum_{i=1}^{p} n_i\left(x_i - \overline{x}\right) = \sum_{i=1}^{p} n_i x_i - \overline{x} \cdot n$$

De même comme $\overline{x} = \dfrac{\sum_{i=1}^{p} n_i x_i}{n}$, on a $\sum_{i=1}^{p} n_i x_i = n \cdot \overline{x}$, d'où

$$\sum_{i=1}^{p} n_i\left(x_i - \overline{x}\right) = n\overline{x} - n\overline{x} = 0$$

P2 *C'est par rapport à la moyenne que la somme des carrés des écarts est la plus petite.*

$$\sum_{i=1}^{p} n_i\left(x_i - a\right)^2 \text{ est minimale pour } \overline{x} = a\,.$$

Démonstration

$$\sum_{i=1}^{p} n_i (x_i - a)^2 = \sum_{i=1}^{p} n_i \left[(x_i - \bar{x}) + (\bar{x} - a) \right]^2$$

$$\sum_{i=1}^{p} n_i (x_i - a)^2 = \sum_{i=1}^{p} n_i (x_i - \bar{x})^2 + \sum_{i=1}^{p} n_i (\bar{x} - a)^2 + 2\sum_{i=1}^{p} n_i \left[(x_i - \bar{x}) \cdot (\bar{x} - a) \right]$$

comme $\sum_{i=1}^{p} n_i (x_i - \bar{x}) = 0$, $2\sum_{i=1}^{p} n_i \left[(x_i - \bar{x}) \cdot (\bar{x} - a) \right] = 0$, et

$$\sum_{i=1}^{p} n_i (x_i - a)^2 = \sum_{i=1}^{p} n_i (x_i - \bar{x})^2 + \sum_{i=1}^{p} n_i (\bar{x} - a)^2$$

pour rendre minimale cette expression où la variable est *a*, il faut annuler le deuxième terme, d'où $\bar{x} = a$.

P3 ***Si a et b sont des constantes et si pour toutes les valeurs observées ou tous les points centraux*** c_i ***on vérifie :*** $x'_i = a + bx_i$ ***, on peut également démontrer que :*** $\bar{x}' = a + b\bar{x}$**.**

Démonstration

$$\begin{aligned} \bar{x}' &= \sum_{i=1}^{p} \frac{n_i x'_i}{n} \\ &= \sum_{i=1}^{p} \frac{n_i (a + bx_i)}{n} \\ &= \frac{na}{n} + b\sum_{i=1}^{p} \frac{n_i x_i}{n} \\ &= a + b\bar{x} \end{aligned}$$

Conséquences : La moyenne est influencée à la fois par une translation ou un changement d'origine (*a*) et par un changement d'unité (*b*).

P4 ***La moyenne $\bar{x}$ d'une population P composée de h sous populations P_h, ayant pour effectif n_h et pour moyenne $\bar{x}_h$, s'exprime simplement en fonction des moyennes des sous populations :***

$$\bar{x} = \frac{1}{n}\sum_h n_h \bar{x}_h$$

<u>Démonstration</u>

Montrons que la moyenne $\bar{x}$ d'une population P composée de 2 sous populations P_1 et P_2 s'exprime en fonction de $\bar{x}_1$ et $\bar{x}_2$.

$$\bar{x} = \frac{1}{n}\sum_i n_i x_i \qquad \bar{x}_1 = \frac{1}{n_1}\sum_i n_{1i} x_i \qquad \bar{x}_2 = \frac{1}{n_2}\sum_i n_{2i} x_i \qquad \text{avec} \qquad n_i = n_{1i} + n_{2i}$$

$$\bar{x} = \frac{1}{n_1 + n_2}\sum_i (n_{1i} + n_{2i}) x_i = \frac{1}{n_1 + n_2}\sum_i n_{1i} x_i + \frac{1}{n_1 + n_2}\sum_i n_{2i} x_i$$

$$\bar{x} = \frac{n_1}{n_1 + n_2}.\underbrace{\frac{1}{n_1}\sum_i n_{1i} x_i}_{\bar{x}_1} + \frac{n_2}{n_1 + n_2}.\underbrace{\frac{1}{n_2}\sum_i n_{2i} x_i}_{\bar{x}_2}$$

$$\bar{x} = \frac{n_1}{n}.\bar{x}_1 + \frac{n_2}{n}\bar{x}_2 = \frac{1}{n}(n_1\bar{x}_1 + n_2\bar{x}_2)$$

1.3 Utilisation

La moyenne constitue probablement le paramètre de position le plus utilisé. Elle dépend cependant de valeurs extrêmes souvent aberrantes. Pour une bonne interprétation de la série statistique, il est ainsi nécessaire de la compléter avec un paramètre de dispersion tel l'écart type.

1.4 Positions respectives du mode, de la médiane et de la moyenne

Lorsque la distribution est symétrique, les trois caractéristiques de tendance centrale (mode, médiane et moyenne arithmétique) sont confondues.

Dans le cas d'une distribution non symétrique, la position respective de ces trois caractéristiques permet de donner l'allure de la courbe des fréquences (oblique à droite, oblique à gauche).

2. La moyenne géométrique

2.1 Définition

La ***moyenne géométrique*** G (ou $\bar{x}_g$) d'une série statistique composée de n valeurs positives $x_1,...,x_n$ est la racine nième du produit de ces n valeurs. En effet :

$$G = x_1^{\frac{1}{n}} \times ... \times x_n^{\frac{1}{n}} = \left[\prod_{i=1}^{n} x_i\right]^{\frac{1}{n}} = \sqrt[n]{x_1 ... x_n}$$

En prenant en compte les fréquences, la moyenne géométrique d'une distribution de fréquences, de valeurs positives x_i et de fréquences n_i (i=1, ..., p) peut être définie comme suit:

$$G = \sqrt[n]{x_1^{n1} ... x_p^{np}} = \left[\prod_{i=1}^{p} x_i^{ni}\right]^{\frac{1}{n}} \text{ ou encore } G = x_1^{f1} ... x_p^{fp} = \prod_{i=1}^{p}\left[x_i^{f_i}\right]$$

Car $\sqrt[n]{x_i^{f_i}} = \left[x_i^{n_i}\right]^{\frac{1}{n}} = x_i^{\frac{n_i}{n}} = x_i^{f_i}$

2.2 Propriétés

P1 ***La moyenne géométrique est toujours inférieure ou égale à la moyenne arithmétique, l'égalité n'étant vérifiée que si les valeurs x_i et x_j sont égales entre elles, $\forall i$ et $\forall j$.***

Démonstration

On montre que la propriété est vérifiée pour deux valeurs x_1 et x_2 :

$$\bar{x} - G = \frac{x_1 + x_2}{2} - \sqrt{x_1 x_2} = \frac{x_1 + x_2 - 2\sqrt{x_1 x_2}}{2} = \frac{\left(\sqrt{x_1} - \sqrt{x_2}\right)^2}{2}$$

d'où on déduit:

$$\bar{x} - G > 0 \text{ si } x_1 \neq x_2$$
$$\bar{x} - G = 0 \text{ si } x_1 = x_2$$

Par récurrence, on peut étendre cette propriété à un nombre quelconque d'observations.

P2 ***La variable*** $z_i = x_i \times y_i$ ***a pour moyenne géométrique le produit des moyennes géométriques des variables*** x ***et*** y**.**

<u>Démonstration</u>

$$G_z = \sqrt[n]{z_1^{n_1} \cdots z_p^{n_p}}$$

$$G_z = \sqrt[n]{(x_1 y_1)^{n_1} \cdots (x_p y_p)^{n_p}}$$

$$G_z = \sqrt[n]{x_1^{n_1} \cdots x_p^{n_p}} \cdot \sqrt[n]{y_1^{n_1} \cdots y_p^{n_p}}$$

$$G_z = G_x \cdot G_y$$

P3 ***La moyenne géométrique des rapports*** $\frac{x}{y}$ ***est égale au rapport des moyennes géométriques de x et de y.***

Cette propriété s'avérera particulièrement intéressante pour le calcul d'indices.

<u>Démonstration</u>

Soit $q_i = \dfrac{x_i}{y_i}$	$$\begin{aligned} G_q &= \sqrt[n]{q_1 q_2 \ldots\ldots\ldots q_n} \\ &= \sqrt[n]{\frac{x_1}{y_1}\cdot\frac{x_2}{y_2}\cdot\ldots\cdot\frac{x_n}{y_n}} \\ &= \frac{\sqrt[n]{x_1 x_2 \ldots\ldots x_n}}{\sqrt[n]{y_1 y_2 \ldots\ldots y_n}} \\ &= \frac{G_x}{G_y} \end{aligned}$$

2.3 Utilisation

La moyenne géométrique est utilisée dans la définition de certains nombres-indices et d'une manière générale pour calculer les moyennes des taux d'accroissement.

3. La moyenne harmonique

3.1 Définition

La ***moyenne harmonique*** H (ou $\bar{x}_h$) d'une série statistique composée de *n* valeurs positives $x_1, \ldots, x_n$ est égale à l'inverse de la moyenne arithmétique des inverses :

$$H = \frac{n}{\frac{1}{x_1} + \ldots \frac{1}{x_n}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

La moyenne harmonique pour les distributions de fréquences est définie comme suit (avec les x_i supposés positifs):

$$H = \frac{n}{\sum_{i=1}^{p} \frac{n_i}{x_i}} \quad \text{ou encore} \quad H = \frac{n}{\sum_{i=1}^{p} \frac{n_i}{x_i}}$$

3.2 Propriétés

P1 ***La moyenne harmonique est toujours inférieure ou égale à la moyenne géométrique et donc aussi à la moyenne arithmétique***.

3.3 Utilisation

On utilise la moyenne harmonique pour le calcul des moyennes de % et de rapport.

4. La moyenne quadratique

4.1 Définition

La ***moyenne quadratique*** Q est la racine carrée de la moyenne arithmétique des carrés des valeurs obsrvées:

$$Q = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2} \text{ et } Q = \sqrt{\frac{1}{n}\sum_{i=1}^{p} n_i x_i^2}$$

4.2 Propriétés

P1 ***Si elle peut être calculée pour des valeurs quelconques (positives, nulles ou négatives) la moyenne quadratique n'a un sens comme indice de position que si tous les*** x_i ***sont positifs ou nuls.***

4.3 Utilisation

On utilise la moyenne quadratique pour calculer des moyennes d'écarts à une valeur centrale.

5. Propriétés comparées des différentes moyennes

La définition des différentes moyennes procède du même principe. Pour chacune d’elles, les observations sont introduites sous une fonction particulière (inverse pour la moyenne harmonique, racine pour la moyenne géométrique, carré pour la moyenne quadratique et identité pour la moyenne arithmétique).

De toutes les moyennes, c’est la moyenne arithmétique qui remplit le mieux les conditions de Yule (calcul aisé, signification simple).

La moyenne géométrique et harmonique tendent à réduire l'influence des observations les plus grandes et à relever celle des plus petites.

Les différentes moyennes d’une même distribution se répartissent dans l’ordre suivant :

$$H \leq G \leq \overline{x} \leq Q$$