MESURES DE TENDANCE CENTRALE ET DE DISPERSION

On considère sur un échantillon de N individus la variable statistique $X = (X_1, X_2, \ldots, X_N)$.

1. INDICATEURS DE TENDANCE CENTRALE

Les mesures de tendance centrale permettent de résumer un ensemble de données relatives à une variable quantitative. Elles permettent de déterminer une valeur «typique» ou centrale autour de laquelle des données ont tendance à se rassembler.

1.1. **Moyennes.** L'indicateur le plus couramment utilisé est la moyenne empirique ou moyenne arithmétique.

Définition 1 (Moyenne arithmétique)**.** *On appelle moyenne arithmétique de X la quantité*

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_N}{N} = \frac{\sum_{n=1}^{N} X_n}{N} .$$

Elle possède, entre autre, la propriété importante suivante :

Proposition 1. *La somme des écarts à la moyenne empirique est nulle.*

Démonstration.

$$\sum_{n=1}^{N} (X_n - \bar{X}) = \sum_{n=1}^{N} X_n - N\bar{X} = 0 .$$

∎

L'inconvénient principal de la moyenne empirique comme indicateur de tendance centrale est d'être assez sensible à la présence de valeurs «abérantes». Un indicateur de tendance centrale plus robuste est donné par la moyenne tronquée d'ordre k :

Définition 2 (Moyenne tronquée d'ordre k)**.** *On appelle moyenne tronquée d'ordre k de X la quantité*

$$\bar{X}_k = \frac{1}{N - 2 * k} \sum_{n=k+1}^{N-k} X_k .$$

Cette moyenne s'obtient en fait en supprimant les k plus petites valeurs et les k plus grandes valeurs d'une observations.

Il existe d'autres moyennes, dont on donne la définition pour les plus courantes.

Définition 3 (Moyenne géométrique)**.** *On appelle moyenne géométrique de X la quantité*

$$M_g(X) = \sqrt[N]{X_1.X_2 \ldots .X_N} = \sqrt[N]{\prod_{n=1}^{N} X_n} .$$

L'utilisation de la moyenne géométrique fait sens si les valeurs ont un caractère multiplicatif.

Définition 4 (Moyenne harmonique)**.** *On appelle moyenne harmonique de X la quantité*

$$M_h(X) = \frac{N}{\frac{1}{X_1} + \ldots + \frac{1}{X_N}} = \frac{N}{\sum_{n=1}^{N} \frac{1}{X_n}} .$$

On utilise la moyenne harmonique lorsqu'on veut déterminer un rapport moyen dans des domaines ou ils existent des liens de porportionnalité inverse.

Définition 5 (Moyenne quadratique)**.** *On appelle moyenne quadratique de X la quantité*

$$M_q(X) = \sqrt{\frac{X_1^2 + X_2^2 + \ldots + X_N^2}{N}} = \sqrt{\frac{1}{N} \sum_{n=1}^{N} X_n^2} .$$

Définition 6 (Généralisation de la moyenne)**.** *On peut généraliser la notion de moyenne de X de la façon suivante, pour $m \in \mathbb{R}$*

$$M_m(X) = \sqrt[m]{\frac{1}{N} \sum_{n=1}^{N} X_n^m} .$$

Remarque 1. *On retrouve les moyennes définies précédemment avec cette définition très générale :*

- *Pour $m = 1$, $M_1(X)$ est la moyenne arithmétique ;*
- *Pour $m = -1$, $M_{-1}(X)$ est la moyenne harmonique ;*
- *Pour $m = 2$, $M_2(X)$ est la moyenne quadratique ;*
- *Lorsque $m \to 0$ $M_m(X)$ tend vers la moyenne géométrique.*

Théorème 1 (Inégalité des moyennes)**.** *Soit $a \in \mathbb{R}$ et $b \in \mathbb{R}$. Soit une variable statistique X sur N individus. On note $M_0(X)$ la moyenne géométrique.*

Si $a < b$, alors

$$M_a(X) < M_b(X) .$$

1.2. **Quantiles.** Les quantiles permettent de donner des indications du type «1 personne sur 10 a moins de tel âge».

La médiane est un indicateur de tendance centrale (plus robuste que la moyenne empirique) qui divise la population en deux parties, qui ont le même nombre d'individus. Autrement dit, elle sépare l'échantillon en deux parties égales.

Définition 7 (Médiane)**.**

$$M = \begin{cases} X_{N/2} & \text{si } N \text{ est pair} \\ X_{\lfloor N/2 \rfloor + 1} & \text{si } N \text{ est impair} \end{cases}$$

Plus généralement, on peut définir une valeur qui sépare l'échantillon en deux parties de tailles approximativement égale à αN, où $\alpha \in]0, 1[$. Une telle valeur est appelée quantile ou fractile empirique d'ordre α. Plusieurs définitions existent, et l'on donne la suivante :

Définition 8 (Quantile d'ordre α)**.** *Soit* $\alpha \in]0,1[$.

$$Q_\alpha = \begin{cases} X_{\alpha N} & \text{si } \alpha N \in \mathbb{N} \\ X_{\lfloor \alpha N \rfloor + 1} & \text{sinon .} \end{cases}$$

Les quantiles les plus utilisés sont les quartiles et les déciles. Les quartiles divisent les observations en 4 parties ($Q_{25\%}$,$Q_{50\%}$,$Q_{75\%}$). Les déciles divisent l'ensemble des observations en 10 parties : $Q_{10\%}$,$Q_{20\%}$,...).

Enfin, un indicateur de position souvent utilisé dans le cas d'un caractère discret est le mode, défini comme la valeur la plus fréquente dans la série d'observation (cette valeur n'est pas nécessairement unique). Dans le cas d'un caractère continu, cette notion ne s'applique pas directement, mais on peut définir une *classe modale*, lorsque les données ont été préalablement catégorisées.

Les mesures données ci-dessus possèdent les deux propriétés suivantes, qui permettent de savoir comment les données se comportent si elles subissent une translation ou un changement d'échelle. Intuitivement, le «centre» d'une distribution doit «suivre» la transformation car celle-ci ne pertube pas la position relative des points observés.

Proposition 2 (Translation)**.** *Soit* $a \in \mathbb{R}$ *et la variable statistique* Y *définie comme* $Y = X + a$*. Alors on a* $\mu_Y = \mu_X + b$*, où* μ *désigne une mesure de tendance centrale (par exemple, la moyenne ou la médiane).*

Proposition 3 (Changement d'échelle)**.** *Soit* $a \in \mathbb{R}$ *et* $Y = aX$*. On a alors* $\mu_Y = a\mu_X$.

Enfin, on peut se demander quels relations il existent entre la moyenne et la médianne. De manière générale, *il n'existe pas de lien entre la moyenne et la médianne*. Cependant, on comparera souvent la moyenne et la médianne pour caractériser la distribution d'une série statistique :

- Si la moyenne est supérieure à la médianne, on dit que la distribution des valeurs observées présente une dissymétrie positive.
- Si la moyenne est inférieure à la médianne, on dit que la distribution des valeurs observées présente une dissymétrie négative.
- Si la moyenne est égale à la médianne, on dit que la distribution des valeurs observées est symétrique.

2. INDICATEURS DE DISPERTION

Comme le nom l'indique, les indicateurs de dispertions permettent de mesurer comment les données se «répartissent». On peut définir deux types de mesure de dispertions :

- Les mesures définies par la distance entre deux valeurs représentatives de la distribution.
- Les mesures calculées en fonction de la déviation par rapport à une valeur centrale.

Définition 9 (Étendue)**.** *L'étendu d'une série statistique est l'écart entre sa plus grande valeur et sa plus petite.*

$$e = \max X - \min X \ .$$

Ce dernier indicateur est très peu robuste. On lui préferera souvent l'intervalle interquartile :

Définition 10 (Intervalle inter-quartile)**.** *L'intervalle inter-quartile est la différence entre le troisième et le premier quartile.*

On peut remarquer que cet intervalle contient 50% des données.

Un premier moyen de mesurer la dispertion des données autour de la moyenne est l'écart moyen absolu.

Définition 11 (Écart moyen absolu)**.** *L'écart moyen absolu est définie par la quantité*

$$\frac{1}{N}\sum_{n=1}^{N}|X_n - \bar{X}|\,.$$

Cette mesure à l'inconvénient mathématique de ne pas être dérivable partour (la valeur absolue n'est pas dérivable en 0). On corrige ce problème en mesurant la moyenne des écarts élevés au carré. On obtient alors définition de la variance empirique :

Définition 12 (Variance empirique)**.** *On appelle variance empirique de la série statistique X la quantité*

$$\sigma^2 = \frac{1}{N}\sum_{n=1}^{N}(X_n - \bar{X})^2$$

Un moyen pratique de calculer la variance empirique est donné par la proposition suivante

Proposition 4.

$$\sigma^2 = \frac{1}{N}\sum_{n=1}^{N}X_n^2 - \bar{X}^2$$

Démonstration.

$$\begin{aligned}\sigma^2 = \frac{1}{N}\sum_{n=1}^{N}(X_n - \bar{X})^2 &= \frac{1}{N}\sum_{n=1}^{N}(X_n^2 - 2 * X_n\bar{X} * + \bar{X}^2)\\ &= \frac{1}{N}\sum_{n=1}^{N}X_n^2 - 2\bar{X}\frac{1}{N}\sum_{n=1}^{N}X_n + \frac{\bar{X}}{N}\sum_{n=1}^{N}1\\ &= \frac{1}{N}\sum_{n=1}^{N}X_n^2 - \bar{X}^2\end{aligned}$$

∎

Cet estimateur pose un autre problème : il est *biaisé*. On utilise alors en pratique une version corrigée

Définition 13 (Variance empirique corrigée)**.**

$$\sigma^{*^2} = \frac{1}{N-1}\sum_{n=1}^{N}(X_n - \bar{X})^2$$

Proposition 5.

$$\sigma^{*^2} = \frac{N}{N-1}\sigma^2 \, .$$

Enfin, pour avoir une quantité qui s'exprime dans la même unité que la moyenne (l'unité de la variance est l'unité de la moyenne élevée au carré), on utilise l'écart-type.

Définition 14 (Écart-type)**.** *On définit l'écart type empirique comme la racine de la variance empirique :*

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(X_n - \bar{X})^2} \, .$$

Les mesures de dispertions possèdent notamment les propriétés suivantes :

Proposition 6 (Invariance par translation)**.** *Les quantités de mesure de dispertion définies ci-dessus sont invariantes par translation.*

Proposition 7 (Changement d'échelle)**.** *Soit $a \in \mathbb{R}$ et $Y = aX$. On note σ_Y^2 (resp. σ^2) la variance de Y (resp. de X). On a $\sigma_Y^2 = a^2\sigma_X^2$ et $\sigma_Y = a\sigma_X$.*