Chapitre 1
Statistiques descriptives

L'objet de ce chapitre est de présenter brièvement la première étape de l'analyse des données : la description. L'objectif poursuivi dans une telle analyse est de 3 ordres : tout d'abord, obtenir un contrôle des données et éliminer les données aberrantes ensuite, résumer les données (opération de réduction) sous forme graphique ou numérique, enfin, étudier les particularités de ces données ce qui permettra éventuellement de choisir des méthodes plus complexes. Les méthodes descriptives se classent en deux catégories qui souvent sont complémentaires : la description numérique et la description graphique.

1.1 Description numérique

Avant de donner des définitions formelles de tous les indices, nous les calculerons sur la série de données suivante (GMQ de porcs exprimés en g):

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
737	630	573	615	718	620	820	763	786	529

Nous noterons *n* la taille de la série de données, ici *n* = 10

1.1.1 Les paramètres de position

Les paramètres de position, aussi appelés valeurs centrales, servent à caractériser l'ordre de grandeur des données.

- **La moyenne arithmétique :**

Elle est plus souvent appelée moyenne, et est en général notée $\bar{x}$, elle est calculée en utilisant la formule:

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

Dans notre exemple, $\bar{x}$ = 679.

- **La moyenne géométrique**

La moyenne géométrique ($\overline{x_g}$) est toujours inferieure (ou égale) à la moyenne arithmétique. Elle est donnée par :

$$\overline{x_g} = \left[\prod_{i=1}^{n} x_i\right]^{1/n}$$

Dans notre exemple, $\overline{x_g}$ = 672.6

On peut remarquer que

$$\log(\overline{x_g}) = \frac{1}{N}\sum_{i=1}^{n}\log(x_i)$$

en d'autres termes, le log de la moyenne géométrique est la moyenne arithmétique du log des données. Elle est très souvent utilisée pour les données distribuées suivant une loi log normale (par exemple les comptages cellulaires du lait).

- **La moyenne harmonique**

La moyenne harmonique ($\overline{x_h}$) est toujours inferieure (ou égale) à la moyenne géométrique, elle est en général utilisée pour calculer des moyennes sur des intervalles de temps qui séparent des évènements. Elle est donnée par :

$$\overline{x_h} = \frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}}$$

Dans notre exemple, $\overline{x_h}$ = 666.05

- **La médiane**

La médiane ***Mé*** est la valeur telle que la moitié des observations lui sont supérieures (ou égales) et la moitié inférieures (ou égales). Il est clair que la médiane existe pour toutes les distributions (ce qui n'est pas le cas de la moyenne) de plus, elle est peu sensible aux valeurs extrêmes.

Lorsque le nombre d'observations est pair, la médiane n'est pas définie de façon unique. La valeur usuellement retenue est la moyenne des observations de rang $\frac{n}{2}$ et de rang $\frac{n}{2}$ + 1 Dans notre exemple Mé = 674.

Dans le cas où les données sont regroupées en classe on détermine par interpolation linéaire la médiane qui est la valeur qui correspond à 50% des observations. On détermine la valeur ***x=Mé*** telle que ***F(Mé)=0.5***. Pour une valeur x comprise entre x_i et x_{i+1} (limites d'une classe) F(x) est donnée par la formule :

$$\frac{F(x)-F(x_i)}{x-x_i}=\frac{F(x_{i+1})-F(x_i)}{x_{i+1}-x_i}$$

Ce qui nous permet d'écrire :

$$x=x_i+\frac{F(x)-F(x_i)}{F(x_{i+1})-F(x_i)}(x_{i+1}-x_i)$$

- **Les quartiles**

Les quartiles sont au nombre de trois. La médiane est le deuxième.

Le premier quartile q_1 est la valeur telle que 75% des observations lui sont supérieures (ou égales) et 25% inférieures (ou égales).

Lorsqu'il n'est pas défini de fa, con unique, on utilise généralement la moyenne des observations qui l'encadrent pour le calculer. Dans notre exemple, q_1 = 615.

Le troisième quartile q_3 est la valeur telle que 25% des observations lui sont supérieures (ou égales) et 75% inferieures (ou égales).

Lorsqu'il n'est pas défini de façon unique, on utilise la moyenne des observations qui l'encadrent pour le calculer. Dans notre exemple, q_3 = 763.

- **Le mode**

Est la (ou les) valeur(s) pour laquelle les effectifs sont maximums, il est en général assez difficile de l'évaluer (quand il existe) sur des échantillons de petite taille.

Le mode est défini comme étant la valeur la plus fréquente.

1.1.2 Les paramètres de dispersion

Ces paramètres (comme leur nom l'indique) mesurent la dispersion des données. Les paramètres étudiés précédemment (moyenne, mode et médiane) ne suffisent pas pour avoir une idée concernant la répartition des valeurs d'une série statistique. Il s'agit de quantifier cette idée d'étalement des valeurs. On introduit d'autres mesures dites paramètres de dispersion.

- **la variance**

Elle est définie comme la moyenne des carrés des écarts à la moyenne. La variance fait intervenir toutes les données de la distribution. Elle est notée par S^2 ou V(x) et est égale à :

$$S^2=\frac{1}{N}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

Pour les séries statistiques :

$$S^2=\frac{1}{N}\sum_{i=1}^{n}n_i(x_i-\bar{x})^2$$

Pour les observations des fréquences. Pour les observations groupées, les deux formules sont équivalentes. Dans le cas des distributions groupées en K classes, on commet en général, comme pour la moyenne, une certaine erreur en remplaçant les valeurs observées par les centres de classes x_i.

- **L'écart type**

Est la racine carrée de la variance, noté par S, est définit par

$$S = \sqrt{S^2}$$

Il s'exprime dans la même unité que la variable. Il permet de mesurer la dispersion des observations autour de la moyenne $\bar{x}$. Un écart-type plus faible exprime une plus faible dispersion et donc une plus forte concentration autour de $\bar{x}$.

- **L'étendue ou amplitude**

Est définie comme la différence entre le maximum et le minimum, noté par w (les valeurs de la série statistique classées par ordre croissant)

$$w = x_{\max} - x_{\min}$$

- **Les moments**

Les moments d'ordre r sont définis par :

$$m_r = \frac{1}{N}\sum_{i=1}^{n} x_i^r$$

Ou par :

$$m_r = \frac{1}{N}\sum_{i=1}^{k} n_i x_i^r$$

Si les observations sont groupées (k est le nombre de groupes d'observations).
Les moments centrés d'ordre r sont définis par :

$$m_r = \frac{1}{N}\sum_{i=1}^{k} n_i (x_i - \bar{x})^r$$

Si les observations sont groupées.

Le moment d'ordre 1 se confond avec la moyenne ($m_1 = \bar{x}$) et le moment centré d'ordre 2 se confond avec la variance ($m_2 = S^2$)

- **Le coefficient de variation**

Est définie comme le rapport entre l'écart type et la moyenne.

$$CV = \sqrt{\frac{S^2}{\bar{x}}}$$

1.1.3 Les paramètres de forme

Les logiciels de statistiques fournissent généralement les paramètres Skewness et Kurtosis construits à partir des moments centrés d'ordre 2,3 et 4 qui mesurent respectivement la symétrie et l'aplatissement de la distribution dont l'échantillon est issu.
Pour une loi normale centrée réduite, ces coefficients sont nuls.
Les moments centrés d'ordre 3 et 4 sont définis par:

$$m_3 = \frac{1}{N}\sum_{i=1}^{k} n_i (x_i - \bar{x})^3$$

$$m_4 = \frac{1}{N}\sum_{i=1}^{k} n_i (x_i - \bar{x})^4$$

A partir de ces définitions, le coefficient d'asymétrie de Fisher est définit par :

$$\gamma_1 = \frac{m_3}{S^3}$$

Le coefficient d'aplatissement de Ficher est définit par :

$$\gamma_2 = \frac{m_4}{S^4}$$

Lorsque la statistique est symétrique, on a : $\gamma_1 = 0$. Pour une distribution normale réduite on a : $\gamma_2 = 3$. Les distributions symétriques pour lesquels γ_2 est supérieur à 3 sont plus pointues que la distribution normale réduite. Ces distributions sont dites leptocurtiques. Si γ_2 est inférieure à 3, la distribution est plus aplatie que la distribution normale réduite. Elle est dite platicurtique.

Résumé de chapitre