Chapitre 4

Statistiques descriptives :
Mesures de tendance centrale et de dispersion

On considère sur un échantillon de N individus la variable statistique définie par le tableau de valeurs (rangées dans l'ordre croissant) suivant :

Valeur	x_1	x_2	…	x_p
Effectif	n_1	n_2	…	n_p

L'effectif total est $N = \sum_{i=1}^{p} n_i$

I 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. Moyennes.

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

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

$$\bar{X} = \frac{n_1 x_1 + n_2 x_2 + \cdots + n_p x_p}{N} = \frac{\sum_{i=1}^{p} n_i x_i}{N}$$

Exemple fil rouge
Soit la série statistique suivante :

Notes	5	8	9	10	11	12	14	16	18
Effectif	1	2	6	7	5	4	3	2	1
Effectif cumulé croissant	1	3	9	16	21	25	28	30	31

$$\bar{X} = \frac{1\times5+2\times8+6\times9+7\times10+5\times11+4\times12+3\times14+2\times16+1\times18}{31} = \frac{340}{31} \approx 10{,}97$$

L'inconvénient principal de la moyenne empirique comme indicateur de tendance centrale est d'être assez sensible à la présence de valeurs «aberrantes». Un indicateur de tendance centrale plus robuste est donné par les moyennes tronquée: on retire un certain nombre de valeurs à la fin comme au début (ce sont les valeurs extrêmes) $\bar{X}_k = \frac{\sum_{i=d}^{f} n_i x_i}{\sum_{i=d}^{f} n_i}$ correspond à la moyenne des valeurs de rang d à celle de rang f.

Remarque : on parlera de moyenne tronquée d'ordre k , lorsque chaque valeur sera associée à un effectif de rang 1 et que l'on retirera les k premières et k dernières valeurs.

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 (Médiane)**.**
Si la série est de taille impaire la médiane est la donnée située au milieu de la liste (son rang est $\frac{N+1}{2}$)
Si la série est de taille paire on prendra comme médiane la moyenne des deux valeurs situées au milieu de la liste (c'est-à-dire celles de rang N/2 et N/2 +1)

Définition (Quantile d'ordre α)**.**
Soit α un pourcentage.
Q_α est la plus petite valeur des x_i telle qu'au moins α pourcent de la population ait une valeur inférieure ou égale à Q_α.

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 : $Q_{10\%}$,$Q_{20\%}$,. . .).

Remarque
Par abus de notation Q_1 =$Q_{25\%}$ et Q_3 =$Q_{75\%}$

Méthode
Pour trouver Q_1 et Q_3 je prends respectivement les valeurs de termes de rangs :
$\frac{N}{4}$ et $\frac{3N}{4}$ si ceux si sont des entiers,
ou de leurs arrondis à l'entier supérieur dans le cas contraire.

Exemple
N = 31 , N étant impair la médiane sera la valeur de rang $\frac{N+1}{2} = 16$, ainsi $M_e = 10$
$\frac{31}{4} = 7{,}75$ et $\frac{3\times31}{4} = 23{,}25$ donc je prendrais pour les premiers et troisièmes quartiles les valeurs de rang 8 et 24, $Q_1 = 9$ et $Q_3 = 12$.

Méthode
Dans le cas d'une répartion en classe, il faut faire un polygone des effectifs cumulés croissants (ou des fréquences cumulées croissantes), Q_1 et Q_3 seront les abscisses des points d'intersection entre le polygone et les horizontales d'équation $y = \frac{N}{4}$ et $y = \frac{3N}{4}$ (ou y=0,25 et y =0,75).

Définition Mode et classe modale**.**
Le mode d'une série statistique est la valeur la plus fréquente. Dans le cas d'une répartition en classe, la classe la plus fréquente sera dite modale.

Intuitivement, le «centre» d'une distribution doit «suivre» la transformation car celle-ci ne pertube pas la position relative des points observés.

Proposition
Si pour passer d'une série statistique à une autre on ajoute une valeur λ (on multiplie par k), alors pour passer des paramètres de tendance centrale et de position d'une série aux paramètre correspondant de l'autre j'ajoute λ(je multiplie par k).

Exemple
Si j'ajoute 1 point à toutes les notes d'une classe alors la moyenne et la médiane augmentent d'un point.
Si je multiplie toutes les notes par 0,5 alors la moyenne et la médiane sont aussi multipliées par 0,5.

II INDICATEURS DE DISPERTION

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

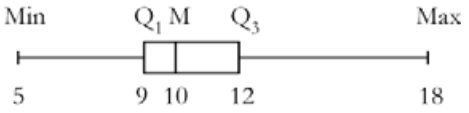
Définition (Étendue)**.**
L'étendu d'une série statistique est l'écart entre sa plus grande valeur et sa plus petite.

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

Définition (Intervalle interquartile)**.**
L'intervalle interquartile 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 dispersion des données autour de la moyenne est l'écart moyen absolu.

Boite à moustache (diagramme en boite)
Dans un repère orthonormé, on trace trois traits verticaux de centres alignés, dont les abscisses sont Min , Q_1, M_e, Q_3 et Max. on fera une boite de Q_1à Q_3, qu'on relira aux traits du Min et du Max
D'un seul coup d'œil je peux voir 5 indicateurs position, me faire une idée concernant l'intervalle interquartile.
Entre deux traits verticaux (Min et Q_1, Q_1 *et* M_e, …) il y a environ 25% de l'effectif.

Exemple



Définition (Variance empirique)**.**
On appelle variance empirique de la série statistique X la moyenne des carré des écarts des valeurs par rapport à la moyenne.

$$V = \frac{n_1(x_1 - \bar{x})^2 + n_2(x_2 - \bar{x})^2 + \cdots + n_p\left(x_p - \bar{x}\right)^2}{N} = \frac{\sum_{i=1}^{p} n_i(x_i - \bar{x})^2}{N}$$

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

Proposition
La variance est la différence entre la moyenne des carrés et le carré de la moyenne.

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

Démonstration

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

Exemple (fil rouge)

$$V = \frac{1\times5^2+2\times8^2+6\times9^2+7\times10^2+5\times11^2+4\times12^2+3\times14^2+2\times16^2+1\times18^2}{31} - \left(\frac{340}{31}\right)^2$$
$$= \frac{25+128+486+700+605+576+588+512+324}{31} - \frac{115600}{961} = \frac{3944}{31} - \frac{115600}{961} = \frac{122264-115600}{961} = \frac{6664}{961} \approx 6.93$$

Définition écart type**.**
L'écart type noté σ est la racine carrée de la variance, il permet de donner « une sorte » de moyenne entre les valeurs et leur moyenne.