Chapitre 2

CARACTÉRISTIQUES STATISTIQUES

1. CARACTÉRISTIQUES DE TENDANCE CENTRALE.

1.1 Notion de distance.

Pour évaluer la proximité entre une valeur x et n observations x_i , ι= 1, ..., n, deux méthodes sont possibles :

- on considère la somme e_x des valeurs absolues des différences entre x et les x_i :

$$e_x = \Leftrightarrow | x - x_1 | + | x - x_2 | + | x - x_3 | + \ldots = \sum_{i=1}^{n} | x - x_i |$$

- on considère la somme d définie par son carré :

$$d_x^2 = [x - x_1]^2 + [x - x_2]^2 + [x - x_3]^2 + \ldots = \sum_{i=1}^{n} [x - x_i]^2$$

1.2 Caractéristiques de tendance centrale ; médiane, moyenne.

Pour déterminer l'ordre de grandeur des observations x_i, il suffit de calculer la valeur x qui en est la plus proche possible. Chaque distance précédente conduit à un paramètre :

Définition de la médiane :

La médiane est égale à toute valeur x telle que la somme des valeurs absolues des différences e_x soit minimale :

propriété caractéristique :

médiane : valeur *mé* telle que la moitié des observations x_i lui soit inférieure et l'autre moitié supérieure.

Définition de la moyenne :

La moyenne est la valeur m de l'inconnue x telle que la somme des carrés des différences d_x^2 soit minimale (c'est le critère des moindres carrés)

Propriété caractéristique :

$$x = (x_1 + x_2 + x_3 + x_4 + \ldots) / n$$

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

propriétés de la médiane et de la moyenne :

- médiane : peu sensible aux valeurs numériques de la série. A choisir dans le cas de données peu nombreuses, certaines observations très élevées en valeur absolue, d'un risque d'erreur de mesure non négligeable
- moyenne : propriétés mathématiques intéressantes, facile à manipuler mathématiquement. A choisir dans le cas d'observations précises, fiables, relativement nombreuses, réparties plus ou moins symétriquement.

2. CARACTÉRISTIQUES DE DISPERSION.

2.1 Écart absolu moyen, écart type :

- écart absolu moyen e_{am} , $m_é$ étant la médiane :

$$e_{am} = \frac{1}{n} \sum_{i=1}^{n} | x_i - m_é |$$

- variance , m étant la moyenne :

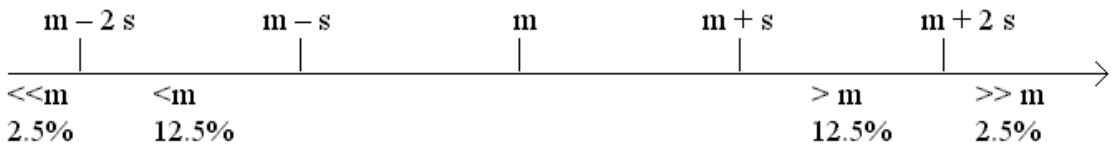
$$s^2 = \frac{1}{n} \sum_{i=1}^{n} [x_i - m]^2$$

- écart type : racine carrée s de la variance s^2.
- Propriété de la variance : la variance est égale à la moyenne des carrés moins le carré de la moyenne :

$$s^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - m^2$$

L'écart absolu moyen et l'écart type sont les distances de la médiane et de la moyenne aux données suivant les deux critères.

2.2 Comparaison d'une valeur à la moyenne (répartion proche de la courbe en cloche).



Première règle de classification

x – m > 2 s >0	(x – m) / s > 2	x est particulièrement grande
2 s > x - m > s	2 > (x – m) / s > 1	x est grande
x – m <- 2 s < 0	(x – m) / s < -2	x est particulièrement petite
- 2 s < x - m < - σ	- 2 > (x – m) / s < -1	x est petite

2.3 Valeurs centrées réduites

Définition :

On appelle valeur centrée réduite d'une observation x_i la valeur x_i' définie par :

$$x_i' = (x_i - m) / s$$

où m est la moyenne des valeurs observées et s leur écart type.

Propriété caractéristique : la valeur centrée réduite est indépendante de l'unité de mesure des observations.

3. AUTRES PARAMÈTRES.

3.1 Coefficient de variation (données positives)

Coefficient de variation :
$c_v = (s/m) \times 100\%$

Supposons que le coefficient de de variation d'une série d'observations positives soit égal à 20% et que la moyenne soit égale à à 12.5 : les valeurs observées sont de l'ordre de 12.5 plus ou moins 20%.

3.2 Coefficients de forme

$$c_{as} = \frac{1}{n} \sum_{i=1}^{n} [(x_i - m) / s]^3$$

Coefficient d'asymétrie

$$c_{ap} = \frac{1}{n} \sum_{i=1}^{n} [(x_i - m) / s]^4$$

Coefficient d'aplatissement

Utilisation des coefficients d'asymétrie et d'aplatissement :

En pratique , ces coefficients servent à contrôler la proximité de la répartition des données à celle de la loi normale qui est une répartition de référence dont la forme est proche d'une « courbe en cloche » :

- $c_{as} \cong 0$ et $c_{ap} \cong 3$: la répartition des données est plus ou moins normale ;
- $c_{as} \neq 0$ ou $c_{ap} \neq 3$: la répartition des données est différente de la loi normale.

4. FONCTION DE RÉPARTITION. QUANTILES.

4.1 Fonction de répartition :

$x \in \mathbf{R} \rightarrow F(x)$: proportion d'observations inférieures ou égales à x.

Exemple : fonction de répartition de la série (x_i) :
$x_1 = 10, x_2 = 11, x_3 = 12, x_4 = 13, x_5 = 14, x_6 = 15, x_7 = 16$

valeurs x_i	effectifs	F(x)
$x < x_1$	0	0
$x_1 \leq x < x_2$	1	1/6
$x_2 \leq x < x_3$	2	2/6
$x_3 \leq x < x_4$	3	3/6
$x_4 \leq x < x_5$	4	4/6
$x_5 \leq x < x_6$	5	5/6
$x_6 \leq x$	6	6/6

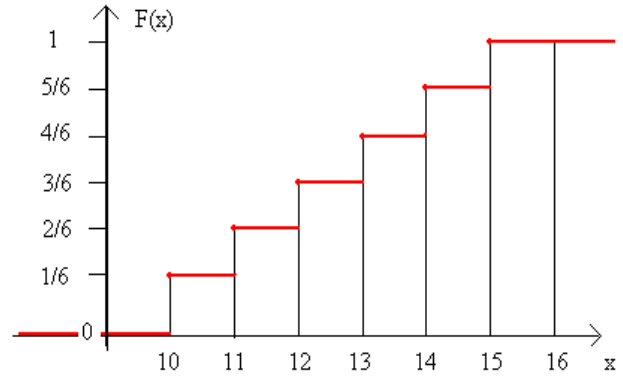
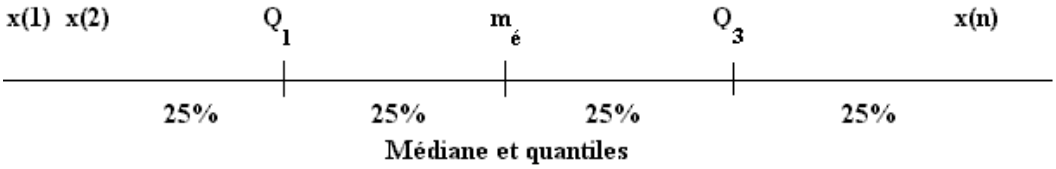


Figure 1.2 : Fonction de répartition

4.2 Quantiles, quartiles, déciles ...

Classement des observations suivant les valeurs croissantes :

dans l'ordre des observations $x_1, x_2, x_3, \ldots, x_{n-1}, x_n$
dans l'ordre croissant : $x(1) \leq x(2) \leq x(3) \leq x(4) \ldots \leq x(n-1) \leq x(n)$



Médiane et quantiles

médiane	$m_é$	deux classes d'effectifs n/2	(50%)	n≥10
quartiles	$q_1, q_2 = m_é, q_3$	quatre classes d'effectifs n/4	(25%)	n≥20
quintiles	r_1, r_2, r_3, r_4, r_5	cinq classes d'effectifs n/5	(20%)	n≥25
déciles	$d_1, d_2, ..., d_9$	dix classes d'effectifs n/10	(10%)	n≥50
centiles	$c_1, c_2, ..., c_{99}$	cent classes d'effectifs n/100	(1%)	n≥500
etc.				

4.3 Seconde règle de classification :

$F(x) < 0.025$	x est particulièrement petite
$0.025 < F(x) < 0.15$	x est petite
$0.95 < F(x) < F(0.975)$	x est grande
$F(x) > 0.975$	x est particulièrement grande.

4.4 Concentration

Au nombre k on associe la somme des k plus petites valeurs x(i), i = 1, …, k..

$$k \in \mathbf{N} \longrightarrow \sum_{i=1}^{k} x(i)$$

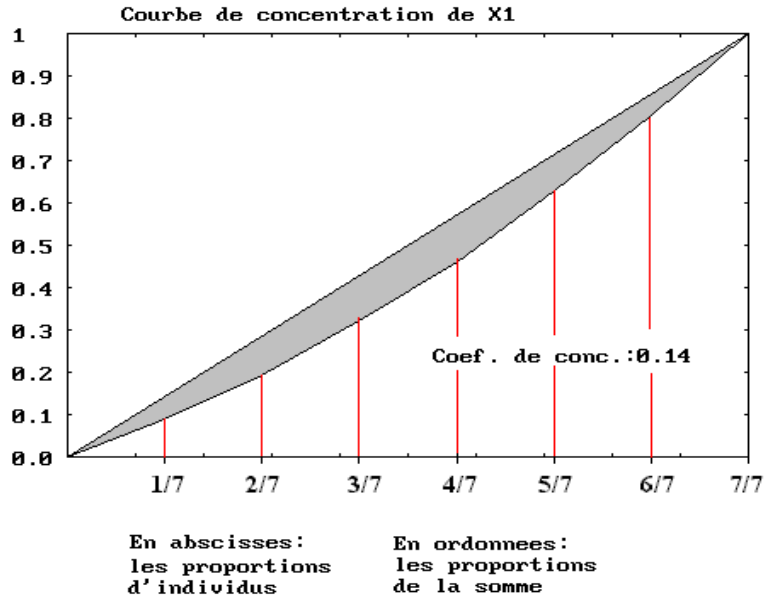
Au nombre n on associe donc la somme des n valeurs :

$$n \longrightarrow S = \sum_{i=1}^{n} x(i)$$

A la proportion p = k/n, on associe la proportion de la somme des k plus petites valeurs, et l'on définit ainsi la fonction de concentration :

$$p = k/n \longrightarrow C(p) = \frac{\sum_{i=1}^{k} x(i)}{\sum_{i=1}^{n} x(i)}$$

Exemple : Série des sept observations : 7, 8, 10, 11, 13, 14, 15. Somme : 78



n°	Proportion	valeur	part de la somme totale	concentration
1	1/7	7	7	7/78 = 0.08974
2	2/7	8	15	15/78 = 0.19231
3	3/7	10	25	25 / 78 = 0.32051
4	4/7	11	36	36 / 78 = 0.46154
5	5/7	13	49	49 / 78 = 0.62821
6	6/7	14	63	63 / 78 = 0.80769
7	7/7	15	78	78 / 78 = 1.00000

Définition :

On appelle coefficient de concentration g de Gini un coefficient variant entre 0 et 1 mesurant la concentration des observations :

- Plus le coefficient est proche de 1, plus la somme dépend d'un petit nombre des plus grandes valeurs.
- Plus le coefficient est proche de 0, moins la somme dépend d'un petit nombre des plus grandes valeurs.

5. CAS DES DONNÉES CLASSÉES OU GROUPÉES.

Les données groupées se présentent sour la forme (n_k, x_k) : la valeur x_k a été observée n_k fois. Le nombre de valeurs distinctes est noté p et le nombre total d'observations est donc :

$$n = \sum_{k=1}^{p} n_k$$

Il suffit de tenir compte des effectifs dans les formules. Les quantiles (médiane, quartiles etc.) peuvent être calculés par interpolation linéaire.

- moyenne :

$$x = (n_1 x_1 + n_2 x_2 + ... + n_p x_p)/n$$

$$m = \frac{1}{n} \sum_{k=1}^{p} n_k x_k$$

- variance :

$$s^2 = [n_1 (x_1 - m)^2 + n_2 (x_2 - m)^2 ... + n_p (x_p - m)^2] /n$$

$$s^2 = \frac{1}{n} \sum_{k=1}^{p} n_k (x_k - m)^2 = \frac{1}{n} \sum_{k=1}^{p} n_k x_k^2 - m^2$$