Statistique 1

UCAO-IEG : Licence 1

Année Universitaire 2019-2020

Prof. Armel YODE
Université Félix Houphouet Boigny
UFR Mathématiques et Informatique

Nous avons confiance en Dieu ; que tous les autres apportent des justificatifs. [Edwards Deming, Professeur de statistique, 1900 -1993]

Table des matières

Chapitre 1

Introduction

La statistique est l'ensemble des méthodes et des techniques destinées à la collecte, l'exploration, l'analyse et l'interprétation des données. Elle a pour objectif de mettre en évidence des informations cachées dans ces données en vue généralement de prendre une décision concernant le phénomène ayant généré ces données. La statistique se divise généralement en deux grandes parties :

- la statistique descriptive qui a pour but d'obtenir un résumé des données ;
- la statistique inférentielle qui a pour but d'utiliser les données afin de tester des hypothèses, de rechercher des modèles ou de faire des prévisions.

1.1 Terminologie de base

Population : C'est l'ensemble sur lequel porte l'étude statistique. La population que l'on envisage en statistique dépend du domaine que l'on traite, et peut donc aussi bien être constituée d'êtres humains que d'animaux, d'objets, voire d'événements.

Individu ou unité statistique : C'est un élément de la population.

Echantillon : C'est un sous-ensemble de la population ; l'échantillon doit être représentatif de la population, c'est à dire qu'il doit refléter fidèlement sa composition et sa complexité ; en effet, les informations obtenues à partir de l'échantillon doivent pouvoir être étendues, sans erreur grave, à l'ensemble de la population.

Enquête statistique : C'est l'opération consistant à collecter des données sur l'ensemble des individus d'un échantillon ou éventuellement la population entière.

Recensement : C'est la collecte des données effectuée sur toute la population.

Sondage : C'est la collecte des données effectuée sur un échantillon de la population.

Caractère : C'est une grandeur ou un attribut observable sur un individu. Parfois, on emploie le terme de variable statistique au lieu de caractère.

Modalité : C'est un état du caractère ; les modalités d'un caractère sont exhaustives et incompatibles, c'est à dire que chaque individu présente une et une seule modalité du caractère.

Série statistique : C'est la suite des valeurs du caractère observée sur chaque individu de l'ensemble étudié (population ou échantillon).

1.2 Caractères

On distingue deux types de caractères : le caractère qualitatif et le caractère quantitatif.

1.2.1 Caractère qualitatif

Le caractère est dit qualitatif si ses modalités sont non mesurables. Le caractère qualitatif est dit ordinal s'il existe un ordre entre ses modalités. Dans le cas contraire, il est dit qualitatif nominal.

Exemple 1. *Caractère qualitatif ordinal.*

- *Population : la classe.*
- *Individu : un étudiant*
- *Caractère : décision du jury*
- *Modalités : ajourné, passable, assez-bien, bien, très bien.*

Exemple 2. *Caractère qualitatif nominal.*

- *Population : la classe.*

- *Individu : un étudiant*
- *Caractère : groupe sanguin.*
- *Modalités : A, B, AB et O.*

1.2.2 Caractère quantitatif

Lorsque les modalités d'un caractère sont mesurables, on dit que ce caractère est quantitatif.

1.2.2.1 Caractère quantitatif discret

Le caractère quantitatif est dit discret lorsqu'il ne peut prendre que des valeurs isolées notées par exemple $x_1, x_2, \ldots, x_k$ où k est le nombre de modalités.

Exemple 3. - *Population : le personnel d'une entreprise*

- *Individu : un employé*
- *Caractère : nombre d'enfants*
- *Modalités :* 0, 1, 2, 3, 4, 5, 6 *et* 7.

1.2.2.2 Caractère quantitatif continu

Le caractère quantitatif est dit continu lorsqu'il peut prendre n'importe quelle valeur d'un intervalle de l'ensemble des nombres réels $\mathbb{R}$. Dans ce cas, l'intervalle des valeurs possibles est divisé en k classes

$$[a_0, a_1[,\ \ [a_1, a_2[, \ldots, [a_{k-1}, a_k[, \quad \text{où} \quad a_0 < a_1 < \ldots < a_{k-1} < a_k.$$

a_{j-1} et a_j sont les frontières de la j-ième classe, $c_j = \frac{a_{j-1}+a_j}{2}$ est le centre de celle-ci. L'amplitude de cette classe est $a_j - a_{j-1}$. Dans la suite, on supposera que les observations d'une classe sont concentrées au centre.

Exemple 4. - *Population : l'ensemble des ouvriers d'une entreprise*

- *Individu : un ouvrier*
- *Caractère : salaire mensuel net (en milliers francs)*
- *Modalités :* $[80, 100[$, $[100, 110[$, $[110, 120[$, $[120, 130[$ *et* $[130, 150[$.

La répartition en classes des données nécessite de déefinir a priori le nombre de classes J et donc l'amplitude de chaque classe. Il existe des formules qui nous permettent détablir le nombre de classes et l'intervalle de classe (l'amplitude) pour une série statistique de n observations.

— La règle de Sturge : $J = 1 + (3.3 * log_{10}(n)$

— La règle de Yule :$J = 2.5n^{1/4}$.

L'intervalle de classe est obtenue ensuite de la manière suivante : longueur de l'intervalle = $(x_{max} - x_{min})/J$, ou x_{max} (resp. x_{min}) désigne la plus grande (resp. la plus petite) valeur observée.

1.3 Effectif, fréquences

On observe un caractère X présentant k modalités, sur n individus. L'effectif n_i de la i-ème modalité du caractère est le nombre d'individus qui possède cette modalité. On a

$$n = n_1 + \ldots + n_k = \sum_{i=1}^{k} n_i.$$

On appelle fréquence de la i-ème modalité le rapport

$$f_i = \frac{n_i}{n}.$$

La fréquence est la proportion par rapport au nombre d'observations, des individus pour lesquels le caractère prend la valeur x_i ou appartient à la classe $[a_i, a_{i+1}[$. Elle est un nombre réel compris entre 0 et 1. Nous avons

$$\sum_{i=1}^{k} f_i = 1.$$

On exprime la fréquence souvent en pourcentage.

On suppose que les modalités du caractère quantitatif étudié sont rangées par ordre croissant. L'effectif cumulé croissant de la i-ème modalité est la somme des effectifs des modalités inférieures ou égales à cette modalité :

$$N_i = \sum_{j=1}^{i} n_j.$$

La fréquence cumulée croissante de la i-ème modalité est la somme des fréquences des modalités inférieures ou égales à cette modalité :

$$F_i = \sum_{j=1}^{i} f_j.$$

Cette fréquence représente la proportion des observation inférieures ou égales à la i-ème modalité du caractère quantitatif si il est discret ou bien inférieures à la borne supérieure du i-ème intervalle s'il est continu.

L'effectif cumulé décroissant de la i-ème modalité est la somme des effectifs des modalités supérieures ou égales à cette modalité :

$$D_i = \sum_{j=i}^{k} n_j$$

La fréquence cumulée décroissante de la i-ème modalité est la somme des fréquences des modalités supérieures ou égales à cette modalité :

$$G_i = \sum_{j=i}^{k} f_j.$$

1.4 Présentation générale des tableaux statistiques

On considère un échantillon de taille n issu d'une population. Pour chaque individu, on fait une observation concernant le caractère X comportant k modalités M_1, M_2,..., M_k. On obtient une série statistique $x_1, \ldots, x_n$. Les données recueillies, appelées données brutes, sont soumises à un premier traitement afin d'en faciliter à la fois la présentation et l'exploitation. Cela consiste à classer chacun des n individus dans les k sous-ensembles définis par les diverses modalités du caractère X. Pour chaque modalité M_i, on pourra inscrire dans le tableau statistique son effectif n_i, son effectif cumulé croissant ou décroissant, sa fréquence f_i et sa fréquence cumulée croissante ou décroissante. On prendra toujours soin de préciser dans la présentation du tableau :

- la population étudiée et le caractère ;
- l'origine du renseignement.

Modalité	Effectif	Effectif Cumulé	Fréquence	Fréquence cumulée
M_1	n_1	n_1	$f_1 = \frac{n_1}{n}$	$F_1 = f_1$
M_2	n_2	$n_1 + n_2$	$f_2 = \frac{n_2}{n}$	$F_2 = f_1 + f_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
M_j	n_j	$n = \sum_{i=1}^{j} n_i$	$f_j = \frac{n_j}{n}$	$F_j = \sum_{i=1}^{j} f_i$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
M_k	n_k	$n = \sum_{i=1}^{k} n_i$	$f_k = \frac{n_k}{n}$	$F_k = \sum_{i=1}^{k} f_i = 1$

TABLE 1.1 – Tableau statistique d'un caractère

La présentation des données sous forme de tableaux est intéressante car elle propose un premier résumé. On dégage ainsi les tendances de la population. Ces tableaux vont nous permettre de faire des représentations graphiques. L'idée sera de rendre compte visuellement du résumé que nous avons commencé. Ensuite, pour les caractères quantitatifs, nous chercherons à résumer numériquement l'information.

Chapitre 2

Représentations graphiques

2.1 Introduction

La représentation graphique a pour objectif de visualiser la distribution des données. Dans ce chapitre, nous passons en revue les principales représentations graphiques utilisées dans les analyses statistiques. Selon le type de variable statistique étudié, on a recours à des graphiques différents.

2.2 Diagrammes à secteurs

Les diagrammes à secteurs conviennent pour représenter les effectifs et les fréquences des caractères qualitatifs ou des caractères quantitatifs discrets. Un diagramme en secteurs est un graphique constitué d'un cercle divisé en secteurs dont les angles au centre sont proportionnels aux effectifs (ou aux fréquences). L'angle α_i d'une modalité d'effectif n_i ou de fréquence f_i est donné en dégrés par

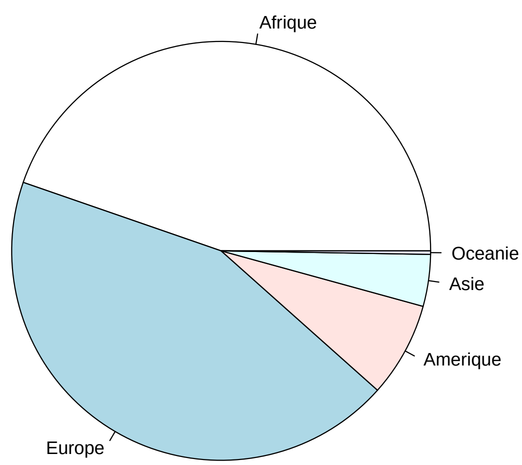
$$\alpha_i = \frac{n_i}{n} \times 360 = f_i \times 360.$$

Exemple 5. *Caractère qualitatif :*

Continent	Effectif	Fréquence (%)
Afrique	168238	44.7
Europe	164542	43.72
Amérique	27540	7.32
Asie	15058	4.00
Océanie	1014	0.27
Total	376392	100

TABLE 2.1 – Répartition des touristes et visiteurs arrivés à l'aéoroport Félix Houphouet-Boigny par continent de provenance en 1999

Source : Ministère de l'Economie et des Finances, 2007.



2.3 Diagramme en barres, diagramme en bâtons

Les diagrammes en barres et les diagrammes en bâtons conviennent pour représenter les fréquences des caractères qualitatifs ou quantitatifs discrets. Les modalités du caractère sont en abscisse et les fréquences sont en ordonné. Dans le cas d'un caractère qualitatif nominal, la position des modalités n'a pas de signification particulière. Si le caractère est qualitatif ordinal ou quantitatif discret, on placera les modalités dans leur ordre naturel.

— **Le digramme en barres :** à chaque modalité du caractère, on associe un rectangle de base constante dont la hauteur est proportionnelle à la fréquence.

— **Le diagramme en bâtons :** à chaque modalité du caractère, on fait correspondre un segment vertical de longueur proportionnelle à la fréquence de cette modalité.

Exemple 6. *Le tableau suivant donne la répartition selon le groupe sanguin de 50 individus pris au hasard dans une population :*

Groupe sanguin	*A*	*B*	*AB*	*O*
Effectif	25	*10*	12	3

1. *Déterminer la variable statistique et son type.*

   Variable statistique : groupe sanguin
   Nature : qualitative nominale.

2. *Donnez une représentation graphique qui fasse apparaître l'importance relative des différents groupes sanguins.*

   Nous pouvons faire un diagramme en barres ou un diagramme en secteurs

Exemple 7. *A Cauphygombokro, en vue d'instaurer la taxe d'habitation, une enquête portant sur le nombre de pièces du logement occupé a été réalisée auprès des ménages. Cette enquête a donné les résultats suivants :*

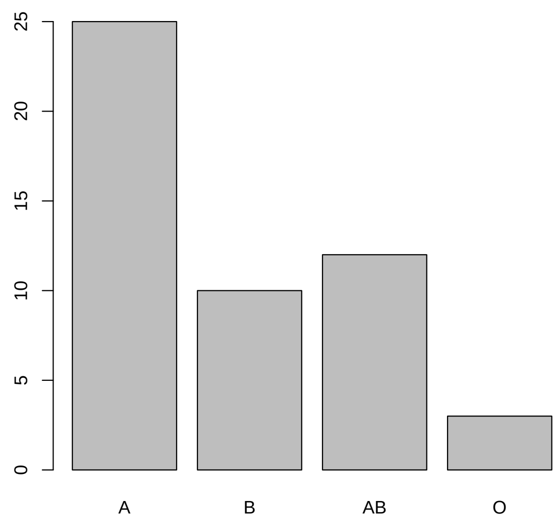


FIGURE 2.1 – Diagramme en barres de la répartition des individus selon le groupe sanguin

Nomrbe de pièces	*Nombre de ménages*
1	*20*
2	*40*
3	*40*
4	*60*
5	*40*

1. *Caractériser la distribution (population, individu, caractère, nature du caractère, modalités)*

   Population : l'ensemble des ménages de Cauphygombokro
   Individu : un ménage
   Caractère : nombre de pièces du logement occupé
   Modalités : 1,2,3,4,5.
2. *Tracer le diagramme en bâtons.*

2.4 Histogramme

L'histogramme est la représetation graphique de la distribution des effectifs ou des fréquences d'une variable statistique continue. Pour construire l'histogramme, on place en abscisse les différentes extrémités a_i des classes, puis on trace, pour chaque classe, un rectangle parallèle aux axes, de telle sorte que la partie parallèle à l'axe des abscisses ait une longueur correspondant à l'amplitude de la classe et que la surface du rectangle soit proportionnelle à l'effectif (ou à la fréquence) de la classe (ceci afin de bien visualiser l'importance de chaque classe).

- Choix de l'unité d'amplitude u : on retiendra par exemple le pgcd des diverses amplitudes.
- Expression des amplitudes dans cette nouvelle unité d'amplitude :
$$e_i = \frac{a_i - a_{i-1}}{u}$$
- La hauteur h_i de chaque rectangle est égale à
$$h_i = \frac{f_i}{a_i}$$
de telle sorte que la surface des rectangles représentatifs est égale à la fréquence de la classe correspondante.

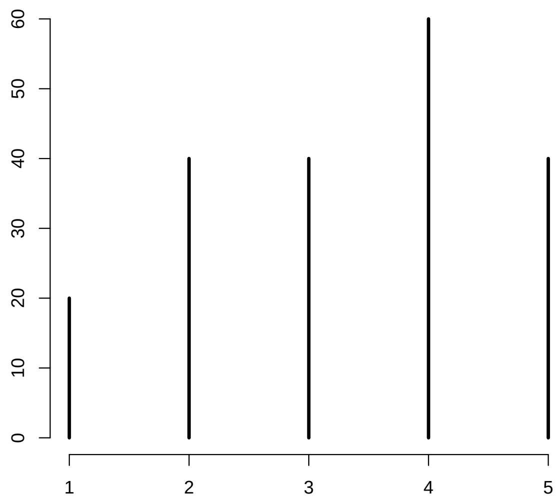


FIGURE 2.2 – Diagramme en bâtons de la répartition des ménages selon le nombre d'enfants

Salaire	Effectif	Fréquence (%)	Fréquence cumulées (%)
$[80, 100[$	26	18.6	18.6
$[100, 110[$	33	23.5	42.1
$[110, 120[$	64	45.8	87.9
$[120, 130[$	7	5.0	92.9
$[130, 150[$	10	7.1	100
Total	140	100	

TABLE 2.2 – Répartition des ouvriers selon leur salaire mensuel net (en milliers francs).

Exemple 8. *Traçons l'histogramme des fréquences.*

2.5 Diagramme de fréquences cumulées

2.5.1 Cas d'un caractère qualitatif ordinal

2.5.2 Cas d'un caractère quantitatif discret

C'est la représentation graphique de la fonction F_X définie par

$$F_X(x) = \begin{cases} 0 \quad \text{si } x < x_1 \\ F_i \quad \text{si } x_i \leq x < x_{i+1} \ i = 1, \ldots, k-1, \\ 1 \quad \text{si } x \geq x_k \end{cases}$$

ou

$$F_X(x) = \begin{cases} 0 \quad \text{si } x < x_1 \\ N_i \quad \text{si } x_i \leq x < x_{i+1} \ i = 1, \ldots, k-1, \\ n \quad \text{si } x \geq x_k \end{cases}$$

Exemple 9. *Répartition des ménages selon le nombre de pièces du logement occupé*

```
  Eff Freq FreqCum
1  20   10      10
2  40   20      30
3  40   20      50
4  60   30      80
5  40   20     100
```

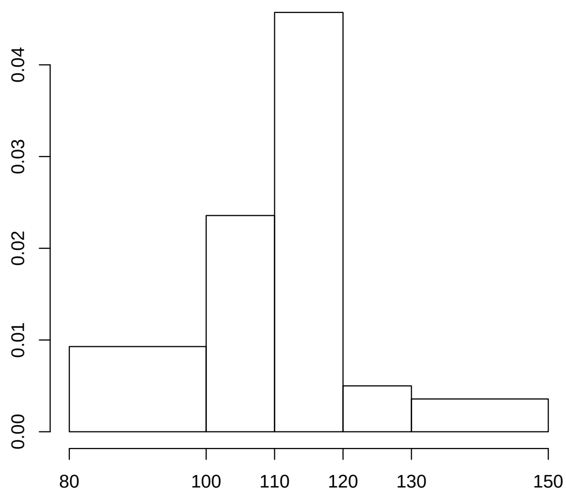


FIGURE 2.3 – Histogramme des fréquences de la répartition des ouvriers selon leur salaire mensuel net

$$F(x) = \begin{cases} 0 & si\ x < 1 \\ 10 & si\ 1 \leq x < 2 \\ 30 & si\ 2 \leq x < 3 \\ 50 & si\ 3 \leq x < 4 \\ 80 & si\ 4 \leq x < 5 \\ 100 & x \geq 5 \end{cases}$$

2.5.3 Cas d'un caractère quantitatif continu

La courbe cumulative est la représentation graphique de la fonction cumulative. Les observations étant groupées par classe, on ne connait de cette fonction que les valeurs qui correspondent aux extrémités supérieures de chaque classe et pour lesquelles elle est égale à la fréquence cumulée F_i :

$$F(a_i) = F_i$$

Exemple 10. *Dans notre exemple, nous avons :*

$$\begin{aligned} F(80) &= 0 \\ F(100) &= 0.186 \\ F(110) &= 0.421 \\ F(120) &= 0.879 \\ F(130) &= 0.929 \\ F(150) &= 1. \end{aligned}$$

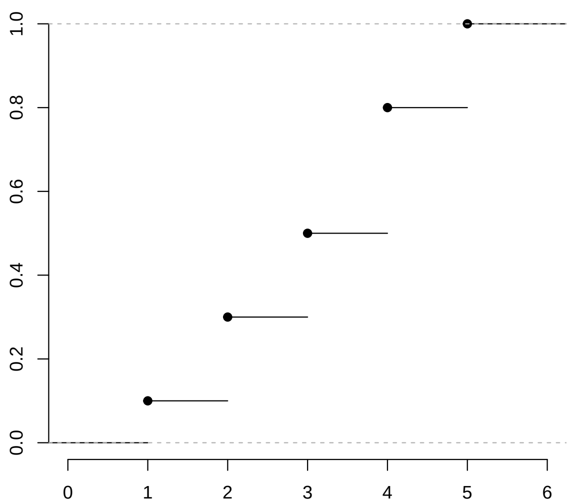


FIGURE 2.4 – Courbe cumulative de la répartition des ménages selon le nombre de pièces du logement occupé

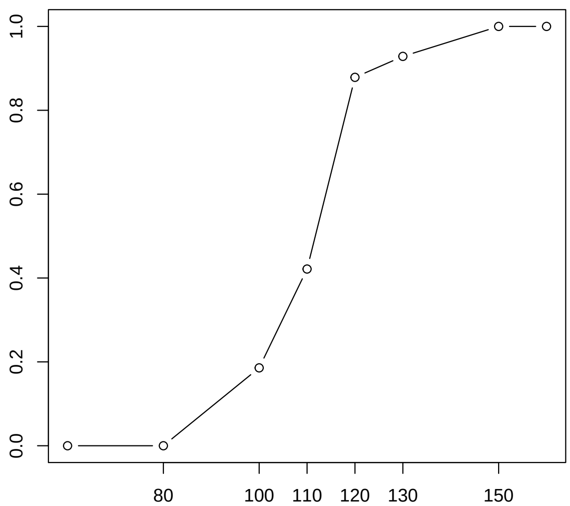


FIGURE 2.5 – Courbe cumulative de la répartition des ouvriers selon leur salaire mensuel net.

Chapitre 3

Paramètres numériques

On distingue les paramètres de tendance centrale (ou de position ou de localisation), les paramètres de dispersion, les paramètres de concentration et les paramètres de forme.

3.1 Paramètres de tendance centrale

Les paramètres de tendance centrale ont pour objet de résumer la série d'observations par une valeur considérée comme représentative. Selon les cas, certains sont plus appropriés que d'autres.

3.1.1 Le mode

Le mode peut être calculé pour tous les types de caractère (quantitatif ou qualitatif). Le mode n'est pas nécessairement unique.

3.1.1.1 Caractère quantitatif discret

Le mode d'un caractère quantitatif discret est la valeur pour laquelle la fréquence est la plus élevée. Graphiquement, le mode est la modalité qui correspond au sommet du diagramme en bâton.

3.1.1.2 Caractère quantitatif continu

Le mode est plus difficile à définir dans le cas d'un caractère quantitatif continu. Lorsque les données sont regroupées en classes, on définit

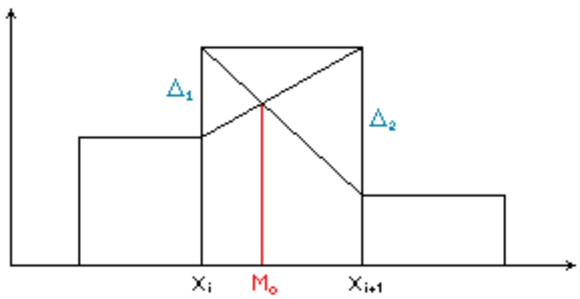


FIGURE 3.1 – Détermination du mode dans le cas d'un caractère continu

la classe modale. La classe modale n'est pas la classe de plus grande fréquence mais la classe de plus grande densité c'est à dire de plus grande fréquence par amplitude. Il est néanmoins possible de déterminer une valeur unique comme mode.

La classe modale $[x_i, x_{i+1}[$ étant déterminée, le mode M_0 est égale est :

$$M_0 = x_i + \frac{\Delta_1}{\Delta_1 + \Delta_2}(x_{i+1} - x_i).$$

Lorsque les classes adjacentes à la classe modale ont des densités de fréquences égales, le mode coincide avec le centre de la classe modale.

3.1.1.3 Remarques

- Le mode dépend beaucoup de la répartition en classes.
-

3.1.2 La moyenne arithmétique

3.1.2.1 Données brutes

Pour une série statistique x_1, $x_2, \ldots, x_n$, on définit la moyenne par

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

C'est la somme de toutes les observations divisée par le nombre total des observations.

3.1.2.2 Données rangées : caractère quantitatif discret

Pour un caractère quantitatif discret dont les n observations sont rangées selon ses k modalités $x_1, \ldots, x_k$ d'effectifs respectifs $n_1, \ldots, n_k$, la moyenne est

$$\bar{x}_n = \frac{1}{n} \sum_{i=1}^{k} n_i x_i.$$

3.1.2.3 Données rangées : caractère quantitatif continu

Pour un caractère quantitatif continu dont les n observations ont été réparties dans k intervalles $([a_{i-1}, a_i[)_{i=1,\ldots,k}$, la moyenne est

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{k} n_i c_i$$

où $c_i = \dfrac{a_{i-1} + a_i}{2}$ est le centre de la classe $[a_{i-1}, a_i[$.

3.1.2.4 Remarques

- La moyenne n'est pas nécessairement une valeur observable du caractère.
- La moyenne est sensible aux valeurs extrêmes ou atypiques.

3.1.3 La moyenne géométrique

La moyenne géométrique est définie par

$$G = \sqrt[n]{\prod_{i=1}^{n} x_i^{n_i}} = \Big(\prod_{i=1}^{n} x_i^{n_i}\Big)^{1/n} \qquad x_i \geq 0.$$

3.1.4 La moyenne harmonique

La moyenne harmonique, H, est l'inverse de la moyenne arithmétique des inverses des obervations :

$$H = \frac{n}{\sum_{i=1}^{k} \frac{n_i}{x_i}}.$$

3.1.5 La moyenne quadratique

La moyenne quadratique est définie par

$$Q = \sqrt{\frac{1}{n}\sum_{i=1}^{k} n_i x_i^2}.$$

Remarque 1. *Pour toute série statistique, l'inégalité suivante est vérifiée :*

$$H < G < \bar{x}_n < Q.$$

3.1.6 La médiane

La médiane M_e est la valeur du caractère pour laquelle la fréquence cumulée est égale à 0.5. Elle correspond donc au centre de la série statistique classée par ordre croissant ou à la valeur pour laquelle 50% des valeurs observées sont supérieures et 50% sont inférieures.

3.1.6.1 Caractère quantitatif discret

On procède ainsi après avoir rangé les n observations $x_1, x_2, \dots, x_n$ par ordre croissant $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$:

- si n est impair, alors $n = 2m + 1$ et la médiane est la valeur $M_e = x_{(m+1)}$.
- si n est pair, alors $n = 2m$ et une médiane est une valeur quelconque entre $x_{(m)}$ et $x_{(m+1)}$; $(x_{(m)}, x_{(m+1)})$ est appelé intervalle médian. Dans ce cas, on prend souvent le milieu comme médiane, c'est à dire
$$M_e = \frac{x_{(m)} + x_{(m+1)}}{2}.$$

3.1.6.2 Caractère quantitatif continu

On utilisera la méthode de l'interpolation linéaire exposée ci-dessous.

3.1.6.3 Remarques

- La médiane peut être calculée pour un caractère quantitatif et pour un caractère qualitatif ordinal.
- La médiane est plus robuste que la moyenne car elle n'est pas influencée par les valeurs extrêmes.
- La médiane est influencée par le nombre d'observations.

3.1.7 Les quantiles

Le quantile d'ordre α est la valeur x_α du caractère qui laisse une proportion α des observations en dessous et $1-\alpha$ des observations au dessus d'elle. Les fractiles sont les quantiles qui partitionnent les données triées en classes de taille égale. Les fractiles les plus utilisés sont les quartiles, les déciles et les centiles.

Les quartiles sont au nombre de trois.

- Le premier quartile Q_1 est le quantile d'ordre $\frac{1}{4}$; c'est la valeur du caractère telle qu'il ait 25% des observations qui lui soient inférieures et 75% supérieures.
- Le deuxième quartile Q_2 est le quantile d'ordre $\frac{1}{2}$, est la médiane.
- Le troisième quartile Q_3est la quantile d'ordre $\frac{3}{4}$; c'est la valeur du caractère telle que 75% des observations lui soient inférieures et 25% supérieures.

Les quartiles Q_1, Q_2 et Q_3 partagent la série ordonnée en quatre groupes de même effectif (25% chacun).

Remarque 2. *Un décile est l'une des neuf valeurs qui partagent la série ordonnée en* 10 *groupes de même effectif (*10% *chacun). Un centile est l'une des cent valeurs qui partagent la série ordonnée en* 100 *groupes de même effectif (*1% *chacun).*

Détermination pratique de la médiane
On utilise le tableau des effectifs cumulés ou des fréquences cumulées.

Caractère quantitatif discret : s'il existe une modalité x_j du caractère telle que $N_{j-1} < \alpha n \leq N_j$ ou $F_{j-1} < \alpha \leq F_j$ alors le quantile d'ordre α est x_j.
Caractère quantitatif continu : soit la première classe dont la fréquence empirique est supérieure ou égale à α. Notons là $C_i = [a_{i-1}, a_i[$ et appelons F_i sa fréquence cumulée. Si $F_i = \alpha$, le quantile est a_i. Dans le cas contraire, $F_i > \alpha$, considérons les points de coordonnées (a_{i-1}, F_{i-1}) et (a_i, F_i), F_{i-1} est la fréquence cumulée de la classe précédant C_i si elle existe, 0 sinon. La droite passant par ces deux points passe par un point d'ordonées α dont l'abscisse est x_α.

a_{i-1}	F_{i-1}
x_α	α
a_i	F_i

On tire x_α à partir de la formule suivante :

$$\frac{x_\alpha - a_{i-1}}{\alpha - F_{i-1}} = \frac{a_i - a_{i-1}}{F_i - F_{i-1}}.$$

Par suite

$$x_\alpha = a_{i-1} + (a_i - a_{i-1})\frac{\alpha - F_{i-1}}{F_i - F_{i-1}}.$$

Exemple 11. *Détermination des quartiles.*
On considère le tableau statistique 8.

100	18.6
Q_1	25
110	42.1

Par suite

$$Q_1 = 100 + (110-100)\frac{25-18.6}{42.1-18.6} = 102.72$$

110	42.1
Q_2	50
120	87.9

Par suite

$$Q_2 = 110 + (120-110)\frac{50-42.1}{87.9-42.1} = 111.72$$

110	42.1
Q_3	75
120	87.9

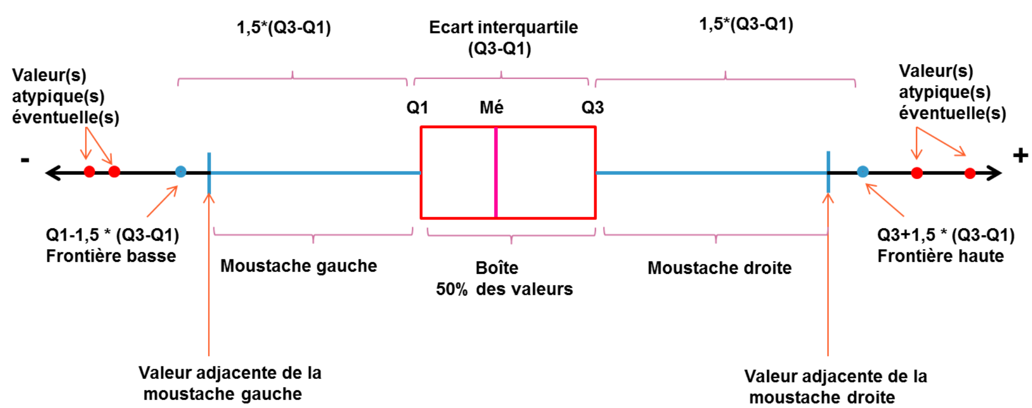
Par suite

$$Q_3 = 110 + (120-110)\frac{75-42.1}{87.9-42.1} = 117.18$$

3.1.8 Boîte à moustaches

La boîte à moustaches ou boxplot est un diagramme qui permet de représenter la distribution d'un caractère. Ce diagramme est composé de :

- un rectangle qui s'étend du premier au troisième quartile ; le rectangle est divisé par une ligne correspondant à la médiane ;



- ce rectangle est complété par deux segments de droites ; pour les dessiner, on calcule d'abord les bornes

$$b^- = Q_1 - 1.5(Q_3 - Q_1)$$

$$b^+ = Q_3 + 1.5(Q_3 - Q_1).$$

  Les valeurs au-délà des moustaches sont des valeurs hors norme éventuellement suspectes ou aberrantes mais pas nécessairement.

Ce diagramme est utilisé notamment pour comparer un même caractère dans deux ou plusieurs échantillons de tailles différentes.

3.2 Paramètres de dispersion

Exemple 12. *Deux groupes d'étudiants ont été observés selon la note obtenue en statistique descriptive :*

Groupe 1	*2*	*5*	*10*	*10*	*10*	*15*	*18*
Groupe 2	*8*	*9*	*10*	*10*	*10*	*11*	*12*

Pour le groupe 1 : $M_{01} = M_{e1} = \overline{X}_1 = 10$
Pour le groupe 2 : $M_{02} = M_{e2} = \overline{X}_2 = 10.$
On remarque que les deux séries présentent un même mode, une même

médiane et une même moyenne. Cependant, leur distribution se fait d'une manière nettement différente. En effet, contrairement au groupe 1, les notes du groupe 2 ne s'écartent pas trop des valeurs centrales ($Me = \bar{X} = 10$)*. Ainsi, les indicateurs de tendance centrale peuvent s'avèrer insuffisant pour permettre à eux seuls de résumer et de comparer deux ou plusieurs séries statistiques, d'où la nécessité de calculer d'autres indicateurs dits de dispersion.*

Les paramètres de dispersion servent à préciser la variabilité de la série statistique, c'est à dire à résumer l'éloignement de l'ensemble des observations par rapport à leur tendance centrale.

3.2.1 L'étendue

On appelle étendue l'écart entre la plus grande valeur et la plus petite valeur. Posons

$$x_{min} = \min(x_1, \dots, x_n) \quad x_{max} = \max(x_1, \dots, x_n).$$

L'etendue est définie par

$$E = x_{max} - x_{min}.$$

Plus l'étendue est faible, plus la série est moins dispersée. L'inconvénient majeur de l'étendue est qu'il ne dépend que des valeurs extrêmes qui sont souvent exceptionnelles et aberrantes.

3.2.2 L'écart moyen absolu

Pour un caractère quantitatif discret dont les n observations sont rangées selon ses k modalités $x_1, \dots, x_k$ d'effectifs respectifs $n_1, \dots, n_k$, l'écart absolu moyen est le nombre

$$EMA = \frac{1}{n} \sum_{i=1}^{k} n_i \mid x_i - \bar{x}_n \mid .$$

Pour un caractère quantitatif continu dont les n observations ont été réparties dans k intervalles $([a_i, a_{i+1}[)_{i=1,\dots,k}$, l'écart absolu moyen est le nombre

$$EMA = \frac{1}{n} \sum_{i=1}^{k} n_i \mid c_i - \bar{x}_n \mid,$$

où $c_i = \dfrac{a_i + a_{i+1}}{2}$ est le centre de la classe $[a_i, a_{i+1}[$.

Remarque 3. *On appelle écart absolu par rapport à la médiane M_e :*

$$EMA_1 = \frac{1}{n}\sum_{i=1}^{k} n_i \mid x_i - M_e \mid .$$

Cet indicateur de dispersion tient compte de tous les écarts entre les valeurs observées et la moyenne arithmétique. Son inconvénient est qu'il n'est pas commode pour le calcul algébrique vu la présence de l'expression de la valeur absolue. Une solution alternative consiste à considérer la moyenne des carrés des écarts et de calculer ensuite la racine carrée.

3.2.3 Variance, écart-type

Pour un caractère quantitatif discret dont les n observations sont rangées selon ses k modalités $x_1, \ldots, x_k$ d'effectifs respectifs $n_1, \ldots, n_k$,

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{k} n_i (x_i - \bar{x})^2.$$

Pour un caractère quantitatif continu dont les n observations ont été réparties dans k intervalles $([a_i, a_{i+1}[)_{i=1,\ldots,k}$, la variance est

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{k} n_i (c_i - \bar{x})^2.$$

où $c_i = \dfrac{a_i + a_{i+1}}{2}$ est le centre de la classe $[a_i, a_{i+1}[$.

L'écart-type σ est la racine carrée de la variance.

La variance mesure la dispersion des valeurs autour de la moyenne. La variance est exprimée dans le carré de l'unité de mesure de la variable. C'est la raison pour laquelle on ne doit pas interpréter la variance mais plutôt sa racine carrée : l'écart-type. L'écart-type est utilisé comme un indicateur de la dispersion de la série statistique. Plus il est grand, plus la dispersion des observations autour de la moyenne de la variable est forte, plus la population est hétérogène.

3.2.4 L'écart inter-quartile

L'intervalle interquartile est l'intervalle $[Q_1, Q_3]$. L'écart interquartile est défini par

$$IQ = Q_3 - Q_1.$$

Nous avons 50% des observations qui se trouvent entre Q_1 et Q_3. Ainsi, 50% des observations s'étalent sur un intervalle de longueur égale à $Q_3 - Q_1$. Plus l'intervalle interquartiles est petit, plus la dispersion est faible et plus la population est homogène.
Cette quantité mesure la dispersion autour de la médiane. Plus IQ est grand, plus il existe des valeurs éloignées de la médiane.

3.2.5 Le Coefficient de variation

Le coefficient de variation CV est défini comme le rapport de l'écart-type à la moyenne :

$$CV = \frac{\sigma}{\bar{x}}.$$

C'est un nombre sans dimension qui mesure la proportion de la moyenne expliquée par l'écart-type. Le coefficient de variation permet de comparer deux ou plusieurs distributions exprimées dans des unités différentes et qui n'ont pas le même ordre de grandeur (les moyennes sont différentes). Le coefficient de variation est souvent exprimé en pourcentage. Plus le coefficient de variation est faible, plus la dispersion est faible et plus la population est homogène.

3.3 Les paramètres de concentration

La notion de concentration tient une place importante dans les études économiques ; on parle de concentration des entreprises, de concentration du pouvoir ou de la richesse, etc. L'étude de concentration ne s'applique qu'à des variables statistiques continues à valeurs positives et cumulables. Il est clair qu'elle ne peut s'appliquer à des ensembles d'individus classés selon l'âge, la taille ou le poids, parce que la somme des âges par exemple d'une population n'a pas de signification. Elle a pour but de mesurer les inégalités de répartition d'une masse totale.

3.3.1 La médiale

La médiale est la valeur du caractère qui partage la valeur totale ou la masse totale en deux parties égales. La médiale se détermine par interpolation linéaire sur les valeurs globales relatives cumulées croissantes.
Soit X un caractère continu dont les observations sont rangées dans les classes $[a_{i-1}, a_i[$, $k = 1, \ldots, k$. Soit n_i l'effectif de la classe $[a_{i-1}, a_i[$ et $c_i = \dfrac{a_{i-1} + a_i}{2}$ son centre.

- On appelle $n_i c_i$ la valeur globale (v.g.) associée à la classe $[a_{i-1}, a_i[$.
- $\displaystyle\sum_{i=1}^{n} n_i c_i$ est appelée valeur totale ou masse totale du caractère étudié.
- $q_i = \dfrac{n_i c_i}{\displaystyle\sum_{i=1}^{n} n_i c_i}$ est la valeur globale relative (v.g.r.) associée à la classe $[a_{i-1}, a_i[$. q_i désigne la part, dans la valeur totale, detenur par les individus ayant une valeur du caractère appartenant à la classe $[a_{i-1}, a_i[$.
- $V(a_i) = V_i = \displaystyle\sum_{j=1}^{i} q_j$ est appelée valeur globale relative cumulée croissante (v.g.r.c.c). Elle indique la part, dans la valeur totale, detenue par les individus ayant une valeur du caractère appartenant à la classe $[a_{i-1}, a_i[$.

La médiale M vérifie $V(M) = 0.5$. La détermination de la médiale se fait en deux étapes :

1. Soit la première classe $[a_{i-1}, a_i[$ dont la valeur globale relative cumulée croissante V_i est supérieure ou égale à 0.5. Si $V_i = 0.5$ alors la médiale est $M = F_i$. Sinon, nous avons $V_{i-1} < 0.5 < V_i$.
2. Par interpolation linéaire, on calcule la valeur de la médiale :

a_{i-1}	V_{i-1}
M	0.5
a_i	V_i

Modalié	Effectif	Centre	Valeur globale	Valeur gloable relative	Valeur gloable relative cumulée
$[a_0, a_1[$	n_1	$c_1 = \frac{a_0+a_1}{2}$	n_1c_1	$q_1 = \dfrac{n_1c_1}{\sum_{i=1}^k n_1c_1}$	$V_1 = q_1$
$[a_1, a_2[$	n_2	$c_2 = \frac{a_1+a_2}{2}$	n_2c_2	$q_2 = \dfrac{n_2c_2}{\sum_{i=1}^k n_ic_i}$	$V_2 = q_1 + q_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[a_{i-1}, a_i[$	n_i	$c_i = \frac{a_{i-1}+a_i}{2}$	n_ic_i	$q_i = \dfrac{n_ic_i}{\sum_{i=1}^k n_ic_i}$	$V_i = \sum_{j=1}^i q_j$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[a_{k-1}, a_k[$	n_k	$c_k = \frac{a_0+a_1}{2}$	n_kc_k	$q_k = \dfrac{n_kc_k}{\sum_{i=1}^k n_ic_i}$	$V_k = \sum_{j=1}^k q_j = 1$
Total	n				

TABLE 3.1 – Tableau de calcul de la médiale

$$\frac{a_i - a_{i-1}}{V_i - V_{i-1}} = \frac{M - a_{i-1}}{0.5 - V_{i-1}} \Leftrightarrow M = a_{i-1} + \frac{0.5 - V_{i-1}}{V_i - V_{i-1}}(a_i - a_{i-1}).$$

Exemple 13. *La médiale est le niveau de salaire qui divise en deux la masse salariale : les salaires inférieurs à la médiale représentent la moitié de la masse salariale et ceux supérieurs à la médiale représentent aussi la moitié de la masse salariale.*

Classe de salaire (en milliers francs)	*Effectif*	*Centre de classe*	*Masse salariale*	*Valeur globale relative*	*Valeur glaobale relative cumulée*
$[80, 100[$	*26*	*90*	*2340*	*15.15*	*15.15*
$[100, 110[$	*33*	*105*	*3465*	*22.44*	*37.59*
$[110, 120[$	*64*	*115*	*7360*	*47.67*	*85.26*
$[120, 130[$	*7*	*125*	*875*	*5.67*	*90.93*
$[130, 150[$	*10*	*140*	*1400*	*9.07*	*100*
Total	*140*		*15440*		

110	37.59
M	50
120	85.26

$$M = 110 + (120 - 110) \times \frac{50 - 37.59}{65.26 - 37.59}.$$

3.3.2 L'écart entre médiane et médiale

On appelle écart médiale-médiane d'une série statistique, le nombre défini par :

$$\Delta M = M - M_e.$$

Cet écart nous fournit un premier renseignement sur la concentration d'une distribution statistique.

- Si $\Delta M = 0 \Leftrightarrow M = M_e$ alors alors la concentration est nulle et la répartition de la valeur totale est parfaitement égalitaire.
- Si $\Delta M \neq 0$ alors la répartition de la valeur totale n'est pas égalitaire. Cependant, aucune information sur l'intensité de cette inégalité ne peut être avancée.
- Pour comparer la concentration de deux ou plusieurs séries statistiques, on peut utiliser le rapport $\frac{\Delta M}{E}$. La concentration d'une série est d'autant plus forte que le rapport est élevé. (E représente l'etendue de la série).

3.3.3 La courbe de Lorenz

La courbe de Lorenz est obtenue en reliant, par des segments de droites, les points de coordonnées (F_i, V_i), $i = 0, \ldots, k$ avec $(F_0, V_0) = (0, 0)$. Plus la courbe de Lorenz s'éloigne de la première bissectrice, plus la concentration est forte et plus la répartition est inégalitaire.

3.3.4 L'indice de Gini

L'indice de Gini ou coefficient de Gini mesure le niveau d'inégalité de la répartition d'une variable dans la population. L'indice de Gini est compris entre 0 (égalité parfaite) et 1 (inégalité parfaite). Une baisse de l'indice de Gini indique une diminution globale des inégalités. A l'inverse, une élévation de l'indice reflète une augmentation globale des inégalités.

Le coefficient de Gini se calcule à partir de la courbe de Lorenz. Pour l'obtenir, il faut diviser l'aire de la zone hachuré en rouge, c'est-à-dire l'espace entre la première bissectrice (représente l'égalité parfaite) et la courbe de Lorenz (distribution observée des revenus disponible), par l'aire du rectangle bleu. On obtient ainsi un rapport qui est le coefficient de Gini (voir figure 3.2).

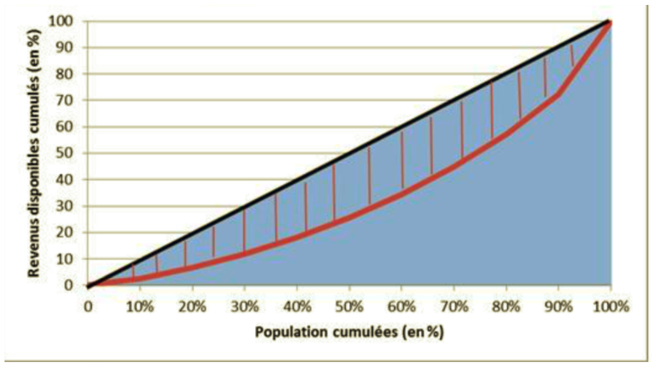


FIGURE 3.2 –

L'indice de Gini I est alors

$$I = 2S = 1 - \sum_{i=1}^{k} \frac{n_i}{n}(V_i + V_{i-1}).$$

3.4 Paramètres de forme

Les paramètres de forme permettent d'avoir une idée staisfaisante et plus précise sur la forme de la distribution. On distingue les coefficients d'asymétrie et les coefficients d'aplatissement.
Une distribution est dite symétrique si les observations également dispersées de part et d'autre de la valeur centrale. Dans le cas contraire, la distribution est dite asymétrique ou dissymétrique.

3.4.1 Moments

Pour un caractère quantitatif discret dont les n observations sont rangées selon ses k modalités $x_1, \ldots, x_k$ d'effectifs respectifs $n_1, \ldots, n_k$, le moment centré d'ordre r est défini par

$$\mu_r = \frac{1}{n} \sum_{i=1}^{k} n_i (x_i - \bar{x})^r.$$

Pour un caractère quantitatif continu dont les n observations ont été réparties dans k intervalles $([a_i, a_{i+1}[)_{i=1,\ldots,k}$, le moment centré d'ordre r est défini par

$$\mu_r = \frac{1}{n}\sum_{i=1}^{k} n_i(c_i - \bar{x})^r,$$

où $c_i = \dfrac{a_i + a_{i+1}}{2}$ est le centre de la classe $[a_i, a_{i+1}[$.

Remarque 4. $\mu_0 = 1$, $\mu_1 = 0$ *et* μ_2 *est la variance.*

3.4.2 Asymétrie

Le coefficient d'asymétrie de Pearson

Dans une distribution faiblement asymétrique, c'est la position du mode par rapport à la moyenne (ou à la médiane) qui caractérise l'asymétrie. Le coefficient d'asymétrie de Pearson est défini par :

$$s = \frac{\bar{X} - M_0}{\sigma}.$$

Le coefficient d'asymétrie de Fisher

Le Coefficient d'asymétrie de Fisher permet de quantifier le degré de déviation de la forme de la distribution par rapport à une distribution symétrique. Il est défini par

$$s = \frac{\mu_3}{\mu_2^{3/2}}.$$

Le coefficient d'asymétrie de Yule

On compare ici létalement de la courbe de distribution à gauche de la médiane et l'étalement à droite et à rapporter leur diiférence à leur somme. Le coefficient d'asymétrie de Yule est défini par :

$$s = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1} = \frac{Q_1 + Q_3 - 2Q_2}{Q_3 - Q_1}.$$

FIGURE 3.3 – $s = 0$: la distribution symétrique.

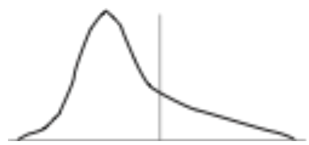


FIGURE 3.4 – $s > 0$: la distribution étalée à droite.

Interprétation

Quelque soit la formule adoptée, nous avons l'interprétation suivante. Ces coefficients n'ont d'intérât que dans la mesure où ils permettent de comparer les formes de deux ou plusieurs distributions ; bien entendu, les comparaison ne sont valables que si la même formule est retenue pour les diverses distributions.

1. $s = 0$ indique une distribution parfaitement symétrique. Dans ce cas $M_e = M_0 = \bar{x}$.
2. $s > 0$ indique une distribution unimodale étalée vers la droite. Dans ce cas $M_0 < M_e < \bar{x}$
3. $s < 0$ indique une distribution unimodale étalée vers la gauche. Dans ce cas $\bar{x} < M_e < M_0$

3.4.3 L'aplatissement

Le coefficient d'aplatissement (kurtosis) permet de mesurer le relief ou la platitude d'une courbe issue d'une distribution de fréquences. On compare la courbe de fréquence de la distribution à la courbe de fréquence de la distribution normale considérée comme la distribution

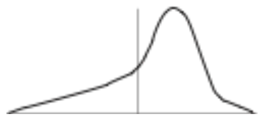


FIGURE 3.5 – $s < 0$: la distribution étalée à gauche.

FIGURE 3.6 – $\gamma = 0$: la distribution est normale.

idéale. On fait apparaître ainsi l'aplatissement ou l'allongement au voisinage du mode. Quand la courbe est plus aplatie que la courbe normale, on dit qu'elle est platicurtique ; quand elle est plus aigue, on dit qu'elle est leptocurtique ; une courbe normale est dite mésocurtique. Le coefficient d'aplatissement de Fisher est :

$$\gamma = \frac{\mu_4}{\mu_2^2} - 3 \quad \mu_2 \neq 0.$$

1. $\gamma = 0$: la distribution est normale
2. $\gamma > 0$: la distribution est aigue.
3. $\gamma < 0$: la distribution est aplatie.

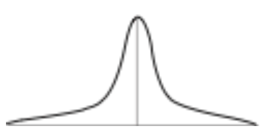


FIGURE 3.7 – $\gamma > 0$: la distribution est aigue.

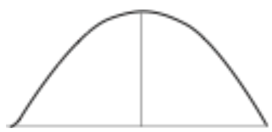


FIGURE 3.8 – $\gamma < 0$: la distribution est aplatie.

Remarque 5. *Les coefficients de kurtosis et de skewness peuvent être utilisés pour s'assurer que les variables suivent une distribution normale. On estime que le coefficient de symétrie ou skewness doit être inférieur à* 1 *et le coefficient d'aplatissement ou kurtosis doit être inférieur à* 1.5 *pour considérer que la variable suit bien une loi normale.*

Chapitre 4

Indices statistiques

4.1 Introduction

Un indice est un instrument statistique permettant de caractériser la variation relative d'un ensemble complexe entre deux situations de temps ou de lieu appelées date de référence et date courante. Deux catégories d'indices peuvent être distinguées selon le type de grandeur étudiée. Ainsi, si l'on considère le prix d'un produit, la production d'une entreprise donnée, le cours de l'action d'une société particulière, il s'agit de grandeurs simples au sens où la grandeur est un nombre ne prenant qu'une seule valeur dans une situation donnée. Les indices calculés sur la base de ces grandeurs sont appelés **indices élémentaires.** En revanche, le niveau général des prix, la production industrielle, le cours des actions sont des grandeurs complexes dans la mesure où leur calcul nécessite d'agréger un ensemble de valeurs hétérogènes (prix des différents produits, production de diverses industries, cours de différentes actions). Les indices calculés sur la base de ces grandeurs sont appelés **indices synthétiques.**

4.2 Indices élémentaires

4.2.1 Définitions

Soit X une grandeur prenant la valeur X_t à la date t. La date peut se rapporter au temps, à l'espace ou à l'évolution de tout autre critère.

Définition 1. On appelle indice élémentaire de la grandeur X à la date t par rapport à la date 0 la quantité définie par

$$I_{t/0} = \frac{X_t}{X_0}.$$

Exemple 14. *La populaion ivoirienne est passée de* 16 *millions en 1998 à* 22 *millions en 2013. L'indice de la population ivoirienne en 2013 par rapport à 1998 est* $I_{2013/1998} = 137.5$*, soit une augmentation de* 37.5% *en* 15 *ans.*

L'indice élémentaire est le plus simple de tous les indices. Il permet d'évaluer l'évolution de X entre la date de reférence 0 et la date courante t. Etant sans dimension, il permet aussi de comparer l'évolution de deux ou plusieurs grandeurs de nature éventuellement différentes, mesurées en unités différentes sur une même période.

Remarque 6. *Un indice élémentaire* $I_{t/0}$ *est équivalent à une variation pourcentage de* $(I_{t/0} - 1) * 100$.

Remarque 7. *Dans la pratique, on exprime un indice élémentaire en pourcentage :*

$$I_{t/0} = 100 \times \frac{X_t}{X_0}.$$

4.2.2 Propriétés d'un indice

4.2.2.1 Circularité (ou transférabilité ou transitivité)

L'indice de la date 0 par rapport à la date de référence t doit être égal au produit de l'indice de la date t par rapport à la date u par l'indice de la date u par rapport à la date 0 :

$$I_{t/0} = I_{t/u} \times I_{u/0}.$$

La circularité permet de changer de base en passant de la date de référence 0 à la date de référence u. En effet, nous obtenons

$$I_{t/u} = \frac{I_{t/0}}{I_{u/0}}.$$

La circularité entraîne la propriété d'enchaînement :

$$I_{t/0} = I_{t/t-1} \times I_{t-1/t-2} \times \cdots \times I_{1/0}.$$

On obtient l'indice à la date t par rapport à la date 0 en faisant le produit des indices intermédiaires d'une date par rapport à la précédente. On dit alors que l'on peut chaîner les évolutions.

4.2.2.2 Réversibilité

L'indice de la date de référence par rapport à la date t doit être égal à l'inverse de l'indice de la date t par rapport à la date de référence :

$$I_{1/0} = \frac{1}{I_{0/1}}.$$

Cette propriété est intéressante lorsqu'on se refère à un critère autre que le temps.

4.3 Indices synthétiques

Les indices élémentaires retracent l'évolution d'une seule grandeur parfaitement définie et homogène. Pour suivre les variations de grandeurs complexes qui sont composées d'un nombre plus ou moins important de grandeurs simples, on utilise les indices synthétiques. Un indice synthétique groupe en un nombre unique des indices élémentaires de même nature. Toute la difficulté réside dans le choix des règles qui détermineront l'indice synthétique. On établit alors une distinction entre l'indice simple et l'indice pondéré :

a) L'indice est simple lorsque les indices élémentaires qui composent l'indice synthétique entrent une fois et une fois seulement dans le calcul.

b) L'indice est pondéré lorsque les indices élémentaires composants n'entrent pas pour parties égales dans le calcul. Cette manière de faire est déterminée par le souci d'accorder plus d'importance à certains produits plutôt qu'à d'autres. On affecte donc à chaque indice élémentaire un poids ou coefficient de pondération différent, compte tenu de l'importance que l'ont veut attribuer à chaque produit, et l'on obtient un indice synthétique qui est une moyenne pondérée des indices élémentaires composants.

Un indice synthétique est une combinaison d'indices élémentaires. Les trois indices synthétiques classiques sont le Laspeyres, le Paasche et le Fisher.

Soit X une grandeur complexe constituée de k grandeurs simples

$$X^1, \ldots, X^k.$$

L'indice élémentaire de la grandeur simple X^i à la date t par rapport à la date de reférence 0 est défini par

$$I^i_{t/0} = \frac{X^i_t}{X^i_0}.$$

Soient

- ω^i_0 l'importance relative de la grandeur simple X^i à la date de référence 0
- ω^i_t l'importance relative de la grandeur simple X^i à la date courante t

4.3.1 Indice de Laspeyres

Définition 2. L'indice de Laspeyres est la moyenne arithmétique pondérée des indices élémentaires par les coefficients de pondération de la date de référence ω^i_0 :

$$L_{t/0} = \sum_{i=1}^{k} \omega^i_0 I^i_{t/0}.$$

L'indice de Laspeyres ne présente ni la propriété de circularité, ni celle de la reversibilité.

4.3.2 Indice de Paasche

Définition 3. L'indice de Paasche est la moyenne harmonique pondéré des indices élémentaires par les ω^i_t de la date courante :

$$\frac{1}{P_{t/0}} = \sum_{i=1}^{k} \frac{\omega^i_t}{I^i_{t/0}}$$

L'indice de Paasche ne possède ni la propriété de circularité ni celle de la réversibilité.

4.3.3 L'indice de Fisher

Définition 4. L'indice de Fisher est la moyenne géométrique des indices de Laspeyres et de Paasche :

$$F_{t/0} = \sqrt{L_{t/0} \times P_{t/0}}.$$

L'avantage de l'indice de Fisher est qu'il jouit de la propriété de réversibilité. Ce qui fait de lui un outil privilégié dans les comparaisons géographiques.

4.3.4 Comparaison

Il n'existe pas de critère général permettant de statuer sur la supériorité d'un indice synthétique par rapport à un autre. Il est cependant possible de présenter les principaux avantages et inconvénients de ceux-ci.

Supposons que l'on étudie l'évolution de la consommation d'un panier composé de plusieurs biens.

— **Indice de Laspeyres.** Les coefficients de pondération sont fixes, c'est-à-dire que l'on suppose que la structure de la consommation ne se modifie pas sur la période étudiée. En conséquence, si l'on considère que les coefficients de pondération sont fixés à la date de référence, plus la date courante est éloignée de cette date, plus il est probable que la structure du panier de biens du consommateur se soit modifiée et plus le risque que les coefficients de pondŕation soient obsolètes est important. Pour cette raison, le principal inconvénient attribué à l'indice de Laspeyres est qu'il tend à surestimer l'effet de l'évolution des prix sur le pouvoir d'achat du consommateur dans la mesure où il ne tient pas compte d'éventuelles substitutions entre les biens du panier considéré.

— **Indice de Paasche.** Les coefficients de pondération sont ceux de la date courante. Ceux-ci évoluent donc avec les prix, c'est-à-dire que la part des différents biens au sein du panier considéré évolue en même temps que les prix. Le calcul de l'indice de Paasche nécessite en conséquence de disposer simultanément des données relatives aux prix et aux quantités à chaque date considéré (et non plus seulement des prix comme dans le cas de l'indice de Laspeyres). Le principal inconvénient tient ici en une difficulté de calcul supplémentaire liée à la disponibilité des données, expliquant pourquoi l'indice de Laspeyres est plus fréquemment utilisé que l'indice de Paasche. Du fait de la variabilité des coefficients de pondération, l'indice de Paasche tend, au contraire de l'indice de Laspeyres, à sous-estimer l'effet de l'évolution des prix sur le pouvoir d'achat du consommateur. Il est important de souligner que les modifications de la structure de consommation ne dépendent évidemment pas que de l'évolution des prix relatifs des biens composant le panier.

— **Agrégation.** Les indices de Laspeyres et de Paasche ont des structures de moyenne. On peut calculer la moyenne arithmétique d'un ensemble à partir des moyennes des sous-ensembles qui le composent. Il en résulte que l'indice de Laspeyres (resp. de Paasche) d'un ensemble peut s'obtenir à partir des indices des groupes formant cet ensemble en leur appliquant la formule de Laspeyres (resp. de Paasche.

Exemple 15. *Entre janvier* 2006 *et janvier* 2010*, l'évolution des prix et du nombre d'exemplaires de journaux vendus en un mois par une société de presse éditant trois journaux mensuels A, B et C a été la suivante :*

	Jannvier 2013		**Janvier 2017**	
	Prix	*Quantité*	*Prix*	*Quantité*
Journal A	*1600*	*8000*	*1900*	*6500*
Journal B	*2600*	*4000*	*2950*	*5000*
Journal C	*3275*	*2000*	*4000*	*1500*

1. *La variation de recettes de la société de presse entre janvier 2013 et Janvier 2017 est de* 11.26%*, en effet :*

$$V_{2017/2013} \times 100 = \frac{33100000}{29750000} \times 100 \approx 111,26.$$

2. *Cette variation fait intervenir un effet quantité et un effet prix qu'on peut évaluer en calculant les indices des prix et des quantités de Paasche et Laspeyres :*
3. *La variation de la valeur globale peut être décomposée en ses deux effets prix et quantité. En effet, à partir de la formule*

   On peut établir le schéma de decomposition donné

Qualité	Laspeyres	Paasche	Fisher
Réversibilité	non mais : $L_{0/t} = \frac{1}{P_{t/0}}$	non mais : $P_{0/t} = \frac{1}{L_{t/0}}$	oui
Transitivité	non	non	non
Agrégation	oui	oui	non
Emploi	Couramment utilisé	peu utilisé	quasiment inusité

4.3.5 Indices de prix, de quantité et de valeur

On s'interesse à l'évolution des dépenses concernant un groupe de k biens de consommation étiquétés de 1 à k entre les dates 0 et 1. Nous notons

p_t^j : prix du bien j à la date t
q_t^j : quantité du bien j à la date t.

La dépense consacrée au bien j à la date t est $D_t^j = p_t^j q_t^j$. Le budget total consacré au groupe de k biens de consommation est

$$D_t = \sum_{j=1}^{k} p_t^j q_t^j.$$

On appelle coefficient budgétaire du bien j à la date t la part du budget total consacré au bien j :

$$\omega_t^j = \frac{p_t^j q_t^j}{\sum_{j=1}^{k} p_t^j q_t^j}$$

Le coefficient budgétaire mesure l'importance relative des différents biens dans le budget total.
Les indices élémentaires entre les dates 0 et 1 des grandeurs considérées sont par définition :

- $I_{1/0}(p^j) = \dfrac{p_1^j}{p_0^j}$: indice de prix du bien j
- $I_{1/0}(q^j) = \dfrac{q_1^j}{q_0^j}$: indice de quantité du bien j
- $I_{1/0}(D^j) = \dfrac{p_1^j q_1^j}{p_0^j q_0^j}$: indice de dépense du bien j

Ces trois indices sont liés par la relation suivante

$$I_{1/0}(D^j) = I_{1/0}(p^j) I_{1/0}(q^j).$$

L'indice de la dépense total est défini par

$$I_{1/0}(D) = \frac{D_1}{D_0} = \frac{\displaystyle\sum_{j=1}^{k} p_1^j p_1^j}{\displaystyle\sum_{j=1}^{k} p_0^j p_0^j}.$$

Pour les indices de Laspeyres et de Paasche, les coefficients budgétaires s'imposent comme coefficients de pondération.

Indice de	Prix	Quantité
Laspeyres	$L_{1/0}(p) = \frac{\sum_{j=1}^{k} p_1^j q_0^j}{\sum_{j=1}^{k} p_0^j q_0^j}$	$L_{1/0}(q) = \frac{\sum_{j=1}^{k} p_0^j q_1^j}{\sum_{j=1}^{k} p_0^j q_0^j}$
Paasche	$P_{1/0}(p) = \frac{\sum_{j=1}^{k} p_1^j q_1^j}{\sum_{j=1}^{k} p_0^j q_1^j}$	$P_{1/0}(q) = \frac{\sum_{j=1}^{k} p_1^j q_1^j}{\sum_{j=1}^{k} p_1^j q_0^j}$

Les indices de Laspeyres et de Paasche se présentent ainsi comme des rapports de dépenses où le facteur (prix ou quantité) autre que celui considéré est constant. L'indice de Laspeyres utilise les constantes de la date de référence tandis que l'indice de Paasche utilise celles de la date courante.