Statistique descriptive

PROF. ARMEL YODÉ

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 voire d'objets.

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 une enquête statistique effectuée sur toute la population.

Sondage : C'est une enquête statistique 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 1.2.1. *Caractère qualitatif ordinal.*

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

Exemple 1.2.2. *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 1.2.3.

- *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 $a_j - a_{j-1}$. On supposera que les observations d'une classe sont concentrées au centre.

Exemple 1.2.4. - *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* $[130,150[$.

Le centre de la classe $[80,100[$ *est :*

$$\frac{80+100}{2} = 90.$$

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

— La règle de Sturge : $J = 1 + 3.3 \times \log_{10}(n)$
— La règle de Yule : $J = 2.5 \times n^{1/4}$.

L'amplitude de classe est obtenue ensuite de la manière suivante :

$$\text{amplitude} = \frac{x_{max} - x_{min}}{J}$$

où 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 :

$$f_i = \frac{n_i}{n} \times 100\,\%$$

Dans ce cas, nous avons :

$$\sum_{i=1}^{k} f_i = 100.$$

Exemple 1.3.1.

1.3.1 Effectifs cumulés, Fréquences cumulées

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

$$N_i = \sum_{j=1}^{i} n_j = n_1 + \cdots + n_i.$$

Le nombre n_i représente le nombre d'observations inférieures ou égale à x_i.

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

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

Cette fréquence représente la proportion (ou le pourcentage) des observation inférieures ou égales à la i-ème modalité x_i 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 x_i 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 :

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

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.

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.

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	$n = \sum_{i=1}^{k} n_i$	$f_k = \frac{n_k}{n}$	$F_k = \sum_{i=1}^{k} f_i = 1$
Total	n		1	

TABLE 1.1 – Tableau statistique d'un caractère

Caractère quantitatif discret

Nombre de pièces	Nombre de ménages	Fréquence	Fréquence cumulée croissante	Fréquence cumulée decroissante
1	20	$\frac{20}{200} = 0.1$	0.1	1
2	40	0.2	0.3	0.9
3	40	0.2	0.5	0.7
4	60	0.3	0.8	0.5
5	40	0.2	1	0.2
Total	200	1		

TABLE 1.2 – Répartition des ménages selon le nombre de pièces du logement occupé

Caractère quantitatif continu

Salaire	Effectif	Fréquence (%)	Fréquence cumulées croissante (%)	Fréquence cumulées decroissantes (%)
$[80, 100[$	26	18.6	18.6	100
$[100, 110[$	33	23.5	42.1	81.4
$[110, 120[$	64	45.8	87.9	57.9
$[120, 130[$	7	5.0	92.9	12.1
$[130, 150[$	10	7.1	100	7.1
Total	140	100		

TABLE 1.3 – Répartition des ouvriers selon leur salaire mensuel net (en milliers francs).

Caractère qualitatf ordinal

Diplôme	Nombre de personnes	Fréquence	Fréquence cumulée croissante	Fréquence cumulée decroissante
Sans diplôme	4	$\frac{4}{50}$ =		1
Primaire	11			
Secondaire	14			
Supérieur	21			
Total	50			

TABLE 1.4 – Répartition de 50 personnes selon le dernier diplôme obtenu

Caractère qualitatif nominal

Continent	Effectif	Fréquence (%)
Afrique	168238	
Europe	164542	
Amérique	27540	
Asie	15058	
Océanie	1014	
Total		100

TABLE 1.5 – 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.

Remarque 1.4.1. *Il s'agit d'un caractère qualitatif nominal. Les fréquences cumulées sont sans intérêt car il n'existe pas de relation d'ordre entre les modalités.*

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

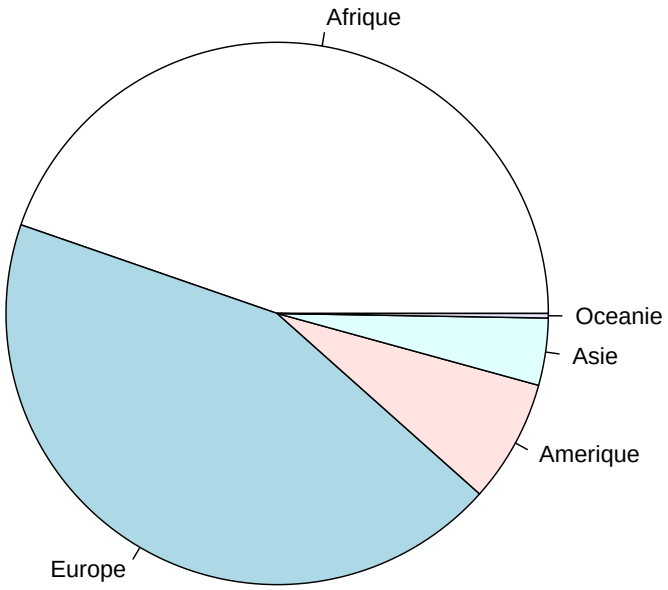
$$\alpha_i = \frac{n_i}{n} \times 360 = f_i \times 360.$$

Exemple 2.2.1. *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 2.3.1. *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 2.3.2. *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 :*

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). Deux classes de même amplitude sont directement comparables. Cette comparaison ne peut être étendue à des classes d'amplitude différente. Pour effectuer la comparaison correctement, nous allons construire les histogrammes en respectant le protocole ci-dessus :

- 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 ; h_i est la fréquence par unité d'amplitude de la classe i.

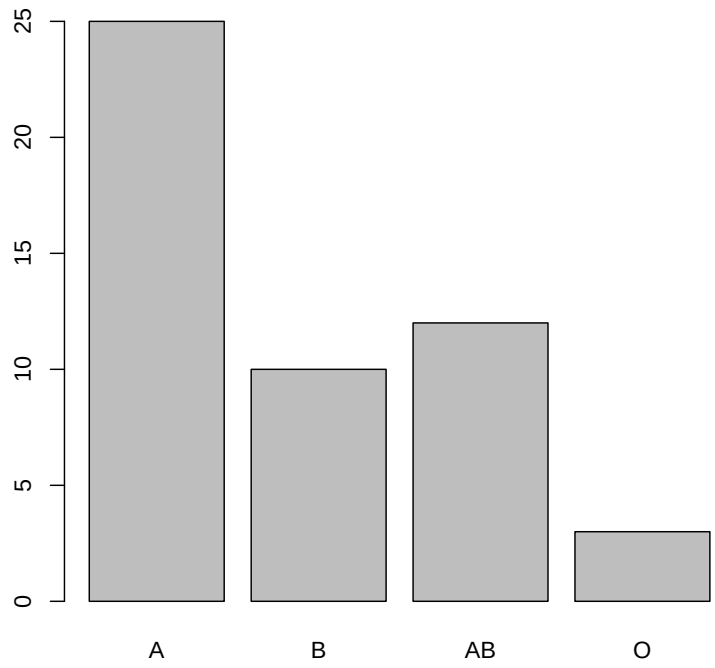


FIGURE 2.1 – Diagramme en barres de la répartition des individus selon le groupe sanguin

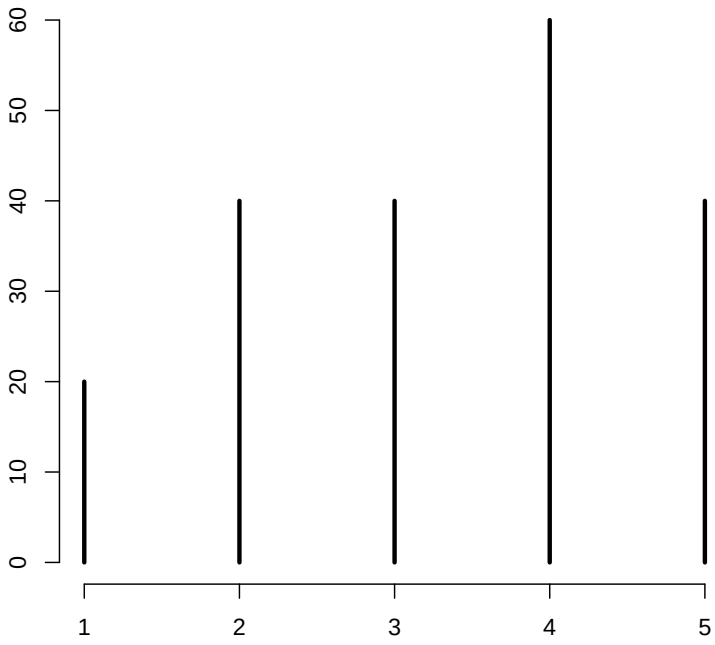


FIGURE 2.2 – Diagramme en bâtons de la répartition des ménages selon le nombre d'enfants

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 2.4.1. *Traçons l'histogramme des fréquences.*

2.5 Diagramme de fréquences cumulées

2.5.1 Cas d'un caractère qualitatif ordinal

Exemple 2.5.1. *On interroge* 50 *personnes sur leur dernier diplôme obtenu (Sans diplôme, Primaire, Secondaire, Supérieur non universitaire, Universitaire). On a obtenu la série statistique suivante :*

Sd Sd Sd Sd P P P P P P P P P P Se Se Se Se Se Se Se Se Se Se Se Se Su Su Su Su Su Su Su Su Su U U UUUUUUUUUU.

1. *Quelle est la population étudiée ? Quel est le caractère étudié ? Quelle est la nature de ce caractère ?*
2. *Construire le tableau statistique.*
3. *Tracer le diagramme des fréquences.*

2.5.2 Cas d'un caractère quantitatif discret

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

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

ou

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

Exemple 2.5.2. *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
```

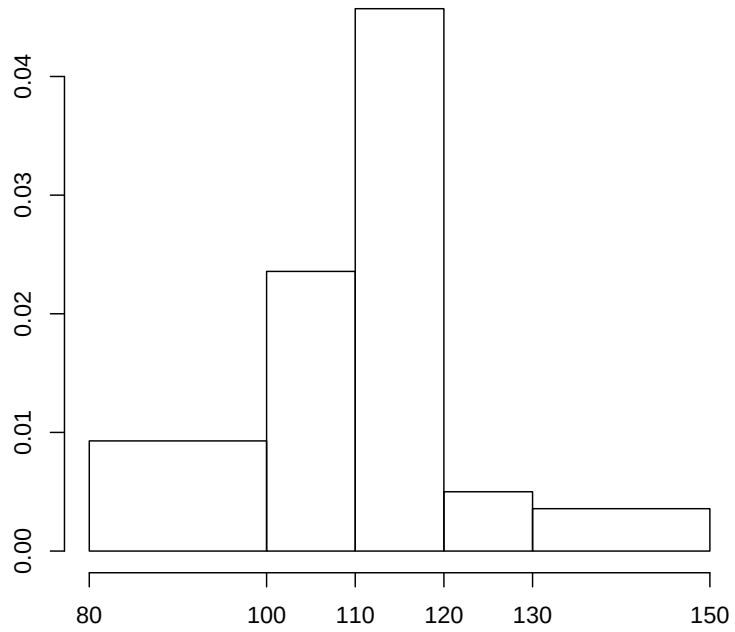


FIGURE 2.3 – Histogramme des fréquences de la répartition des ouvriers selon leur salaire mensuel net

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

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 2.5.3. *Dans notre exemple, nous avons :*

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

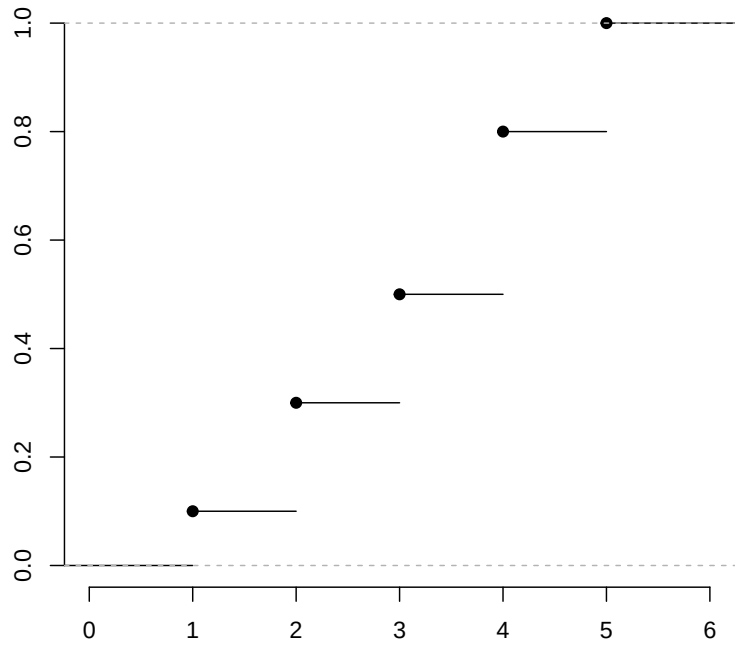


FIGURE 2.4 – Courbe cumulative de la répartition des ménages selon le nombre de pièces du logement occupé

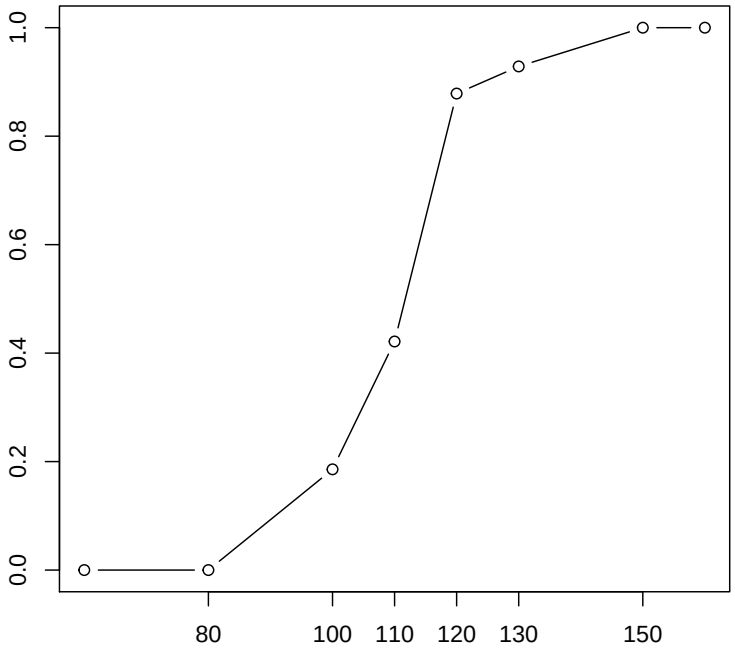


FIGURE 2.5 – Courbe cumulative de la répartition des ouvriers selon leur salaire mensuel net.

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 est la valeur la plus fréquente du caractère. Il 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 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. Le mode dépend beaucoup de la répartition en classes.

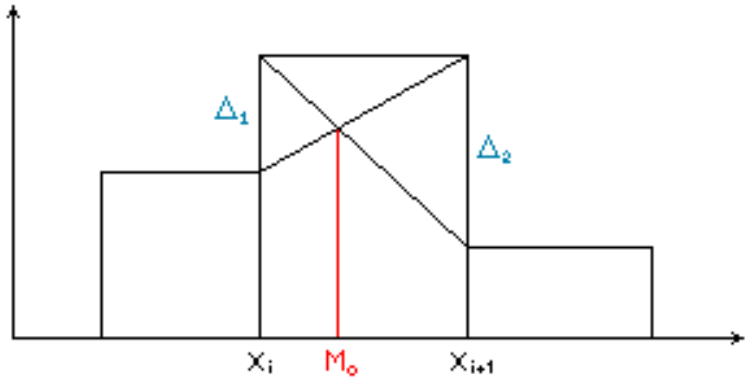


FIGURE 3.1 – Détermination du mode dans le cas d'un caractère continu

3.1.2 La moyenne arithmétique

3.1.2.1 Données brutes

Pour une série statistique x_1, x_2,...,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

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

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 3.1.1. *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 $x_1, x_2, \ldots, x_n$ par ordre croissant $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$:

- 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 calculéee 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 3.1.2. *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

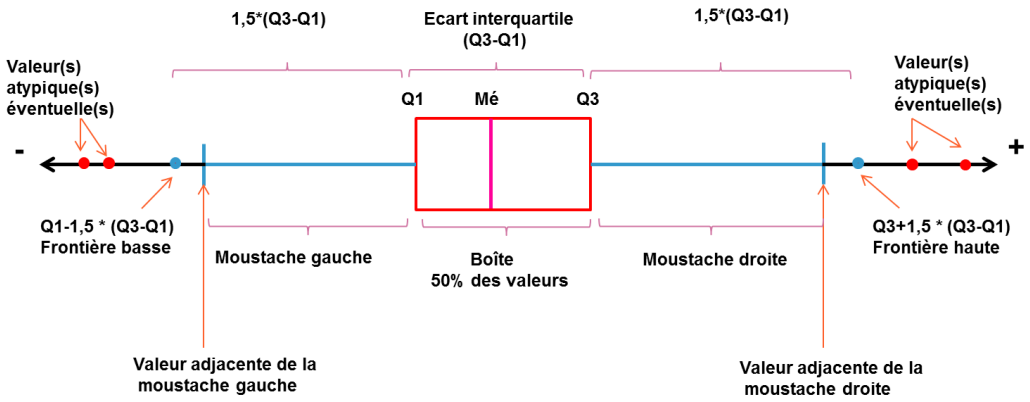
$$x_\alpha = a_{i-1} + (a_i - a_{i-1})\frac{\alpha - F_{i-1}}{F_i - F_{i-1}}.$$

Exemple 3.1.1. *Détermination des quartiles.*
On considère le tableau statistique 2.4.1.

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 3.2.1. *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

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

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 $x_1, \ldots, x_k$ d'effectifs respectifs $n_1, \ldots, n_k$, 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 $([a_i, a_{i+1}[)_{i=1,\ldots,k}$, 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 3.2.1. *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 $[a_{i-1}, a_i[$.
- $\sum_{i=1}^{n} n_i c_i$ est appelée valeur totale ou masse totale du caractère étudié.
- $q_i = \dfrac{n_i c_i}{\sum_{i=1}^{n} n_i c_i}$ 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 $[a_{i-1}, a_i[$.
- $V(a_i) = V_i = \sum_{j=1}^{i} q_j$ 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

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

Exemple 3.3.1. *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.*

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

TABLE 3.1 – Tableau de calcul de la médiale

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 est le double de l'aire comprise entrelacourbe de concentration et la première bissectrice (zone hachuré en rouge). Il 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.

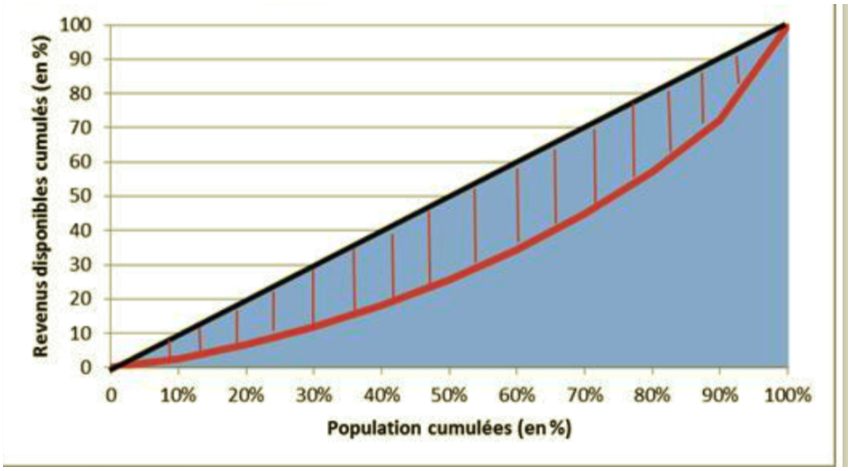


FIGURE 3.2 –

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 3.4.1. $\mu_0 = 1$, $\mu_1 = 0$ *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 :

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

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 $\bar{x} < M_e < M_0$

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

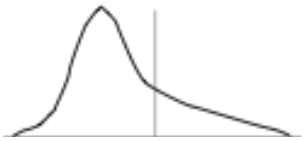


FIGURE 3.4 – $s > 0$: la distribution étalée à droite.

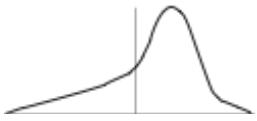


FIGURE 3.5 – $s < 0$: la distribution étalée à gauche.

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

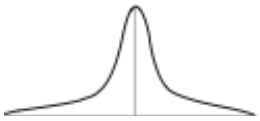


FIGURE 3.7 – $\gamma > 0$: la distribution est aigue.

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 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.

Remarque 3.4.2. *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.*

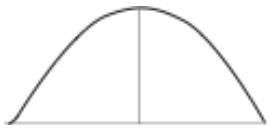


FIGURE 3.8 – $\gamma < 0$: la distribution est aplatie.

Chapitre 4

Indices statistiques

4.1 Introduction

Un indice est un instrument statistique permettant de caractériser la variation relative d'une grandeur 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 4.2.1. *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 4.2.1. *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 4.2.1. *Un indice élémentaire $I_{t/0}$ est équivalent à une variation pourcentage de $(I_{t/0} - 1) * 100$.*

Remarque 4.2.2. *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.
La circularité entraîne aussi la propriété de 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.

4.2.2.2 Produit

Supposons que la grandeur X soit telle que $X = Y \times Z$ où Y et Z sont deux grandeurs. Alors, nous avons :

$$I^X_{t/0} = I^Y_{t/0} \times I^Z_{t/0}.$$

4.2.2.3 Division

Supposons que la grandeur X soit telle que $X = \frac{Y}{Z}$ où Y et Z sont deux grandeurs. Alors, nous avons :

$$I^X_{t/0} = \frac{I^Y_{t/0}}{I^Z_{t/0}}.$$

4.3 Indices synthétiques

En économie, les grandeurs étudiées sont souvent complexes, c'est à dire composées de plusieurs grandeurs simples. Pour suivre les variations de grandeurs complexes, on utilise les indices synthétiques. Un indice synthétique est une combinaison d'indices élémentaires. 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.

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 4.3.1. *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 4.3.2. *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 4.3.3. *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 pondération 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 4.3.1. *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*

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

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

Cet indice est relativement peu informatif au sens où, s'il augmente, il n'est pas possible de distinguer si cette hausse provient d'une augmentation des prix accompagnée d'une baisse des quantités ou de toute autre combinaison. Pour palier à cette difficulté, on utilise fréquemment les indices de Laspeyres et de Paasche. 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.

4.4 Raccords d'indices

Considérons un indice I^X base 1 à la date 0, calculé pour la grandeur X jusqu'à la date d, date à laquelle il est remplacé par un indice J^X base 1 à la date f. Afin d'étudier l'évolution de la grandeur X sur l'ensemble de la période allant des dates 0 à t avec $t > d$,il convient de procéder à un raccord d'indices, c'est à dire de déterminer la valeur qu'aurait pris l'indice I^X à la date t. Notons I^{tX} l'indice ainsi raccordé. Nous avons alors

$$I_{t/0}^{tX} = J_{t/f}^X \times \frac{I_{k/0}^X}{J_{k/f}^X}$$

avec $f \leq k \leq d$. Le rapport $\frac{I_{k/0}^X}{J_{k/f}^X}$ est appelé coefficient de raccordement et correspond au coefficient par lequel on doit multiplier le nouvel indice afin d'en déduire la valeur prise par le précédent indice s'il avait continué à être calculé. Le choix de la date k est laissé au statisticien, mais l'on retient fréquemment la dernière date pour laquelle l'ancien indice est disponible.

4.5 Indices chaînes

Les mutations économiques ont pour conséquence que des indices dont la base reste fixe sur une longue période ne peuvent tenir compte de ces changements et ne sont en conséquence pas représentatifs de la réalité économique. Afin de pallier cette difficulté, il est possible de calculer des indices dont la base varie de date en date (ou de période en période). Cela consiste à généraliser le principe de raccord d'indices en définissant des indices chaînes.

Définition 4.5.1. *L'indice chaîne $C_{t/0}^X$ calculé pour une grandeur X à la date t par rapport à la date de référence 0 s'écrit :*

$$C_{t/0}^X = I_{t/t-1}^X \times C_{t-1/0}^X$$

soit encore

$$C_{t/0}^X = I_{t/t-1}^X \times I_{t-1/t-2}^X \times \ldots \times I_{1/0}^X.$$

4.6 Hétérogénéité et effet qualité

Dans les calculs d'indices que nous avons effectués jusqu'à présent, nous avons supposé implicitement que les biens ou produits considérés sont homogènes au sein d'une même classe. Or, cela n'est souvent pas le cas en pratique. La prise en compte de l'hétérogénéité, et donc de la qualité, des produits n'est pas neutre quant au calcul et à l'interprétation des indices.

Exemple 4.6.1. *Prix de vente (en euros par litre) et quantités produites (en millions de litres) de vin en France.*

Chapitre

5 Statistiques à deux variables

Dans ce chapitre, nous considérons simultanément deux caractères. L'idée est d'étudier la relation entre ces deux caractères.

5.1 Généralités

5.1.1 Distribution conjointe

Soit une population comprenant n individus pour chacun desquels on a fait une observation concernant simultanément les caractères X et Y. Le caractère X comporte les k modalités $X_1, \cdots, X_k$ et le caractère Y, les l modalités $Y_1, \cdots, Y_l$. L'opération préliminaire de mise en ordre des observations va consister à classer chacun des n individus dans les $k \times l$ sous-ensembles définis par le croisement des caractères X et Y. A chacun des sous-ensembles correspond une case du tableau statistique à double entrée où figurent en ligne les modalités de X et en colonne les modalités de Y (tableau à k lignes et l colonnes). Ce tableau est appelé tableau de contingence.
On note n_{ij} l'effectif des individus présentant à la fois la modalité X_i et la modalité Y_j. La fréquence des individus présentant à la fois la modalité X_i et la modalité Y_j est

$$f_{ij} = \frac{n_{ij}}{n}.$$

La distribution conjointe des caractères X et Y est donnée par le tableau de contingence

$X \setminus Y$	Y_1	Y_2	$\cdots$	Y_j	$\cdots$	Y_l
X_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1l}
X_2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2l}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{il}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_k	n_{k1}	n_{k2}	$\cdots$	n_{kj}	$\cdots$	n_{kl}

5.1.2 Distributions marginales

Le nombre d'individus présentant la modalité X_i du caractère X $n_{i\bullet}$ est

$$n_{i\bullet} = \sum_{j=1}^{l} n_{ij}.$$

La fréquence de la modalité X_i est donnée par

$$f_{i\bullet} = \frac{n_{i\bullet}}{n}.$$

Le nombre d'individus présentant la modalité Y_j du caractère Y est

$$n_{\bullet j} = \sum_{i=1}^{k} n_{ij}.$$

La fréquence de la modalité Y_j est donnée par

$$f_{\bullet j} = \frac{n_{\bullet j}}{n}.$$

Nous avons

$$n = \sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij} = \sum_{i=1}^{k} n_{i\bullet} = \sum_{j=1}^{l} n_{\bullet j}$$

$$\sum_{i=1}^{k}\sum_{j=1}^{l} f_{ij} = \sum_{i=1}^{k} f_{i\bullet} = \sum_{j=1}^{l} f_{\bullet j} = 1.$$

La distribution marginale de X est donnée par le tableau ci-dessous :

Modalités de X	Effectif	Fréquence
X_1	$n_{1\bullet}$	$f_{1\bullet}$
X_2	$n_{2\bullet}$	$f_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$
X_i	$n_{i\bullet}$	$f_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$
X_k	$n_{k\bullet}$	$f_{k\bullet}$
total	n	1

La distribution marginale de Y est donnée par le tableau ci-dessous :

Modalités de X	Effectif	Fréquence
Y_1	$n_{\bullet 1}$	$f_{\bullet 1}$
Y_2	$n_{\bullet 2}$	$f_{\bullet 2}$
$\vdots$	$\vdots$	$\vdots$
Y_j	$n_{\bullet j}$	$f_{\bullet j}$
$\vdots$	$\vdots$	$\vdots$
Y_l	$n_{\bullet l}$	$f_{\bullet l}$
total	n	1

5.1.3 Distributions conditionnelles

La distribution conditionnelle de Y sachant $X = X_i$ est donnée par

$Y \mid X = X_i$	Y_1	$\cdots$	Y_j	$\cdots$	Y_l	Total
Effectif	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{il}	$n_{i\bullet}$

Nous pouvons ainsi définir k distributions conditionnelles.

La distribution conditionnelle de X sachant $Y = Y_j$ est donnée par

$X \mid Y = Y_j$	X_1	$\cdots$	X_j	$\cdots$	X_k	Total
Fréquence	n_{1j}	$\cdots$	n_{ij}	$\cdots$	n_{kj}	$n_{\bullet j}$

De même, nous définissons aussi l distributions conditionnelles.

5.1.4 Indépendance

On dit que les caractères X et Y sont satistiquement indépendants dans l'ensemble des n individus considérés si toutes les distributions conditionnelles de X sont identiques à la distribution marginale en X.

$$\text{Indépendance entre } X \text{ et } Y \Longleftrightarrow \text{Pour tous } (i,j), \quad f_{i/j} = f_{i\bullet}$$

Puisque

$$f_{i/j} = \frac{n_{ij}}{n_{\bullet j}} = \frac{\frac{n_{ij}}{n}}{\frac{n_{\bullet j}}{n}} = \frac{f_{ij}}{f_{\bullet j}},$$

alors

$$f_{ij} = f_{\bullet j} f_{i/j}.$$

Ainsi, nous obtenons

$$\begin{aligned} \text{Indépendance entre } X \text{ et } Y &\Longleftrightarrow \text{Pour tous } (i,j), \quad f_{ij} = f_{i\bullet} f_{\bullet j} \\ &\Longleftrightarrow \text{Pour tous } (i,j), \quad n_{ij} = n_{i\bullet} n_{\bullet j} \end{aligned}$$

Par symétrie :

$$\text{Indépendance entre } X \text{ et } Y \Longleftrightarrow \text{Pour tous } (i,j), \quad f_{j/i} = f_{\bullet j}$$

Lorsque deux variables dépendent statistiquement l'une de l'autre, on cherche à évaluer l'intensité de leur liaison et, dans le cas de deux variables quantitatives, on examine si on peut les considérer liées par une relation linéaire.

5.2 Liaison entre deux caractères qualitatifs

5.2.1 Mesure de l'intensité de la liaison

L'intensité de la liaison entre deux caractères qualitatifs est mesurée par

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{\left(n_{ij} - \frac{n_{i\bullet} n_{\bullet j}}{n}\right)^2}{\frac{n_{i\bullet} n_{\bullet j}}{n}}.$$

Le χ^2 est toujours positif ou nul.

Définition 5.2.1.

$$X \text{ et } Y \text{ sont indépendants} \Leftrightarrow f_{ij} = f_{i\bullet} f_{\bullet j} \ i = 1,\dots,k, \ j = 1,\dots,l.$$

Remarque 5.2.1. *Nous avons*

$$\begin{aligned} f_{ij} = f_{i\bullet} f_{\bullet j} &\Leftrightarrow \frac{n_{ij}}{n} = \frac{n_{\bullet j}}{n} \times \frac{n_{i\bullet}}{n} \\ &\Leftrightarrow n_{ij} = \frac{n_{\bullet j} \times n_{i\bullet}}{n} \end{aligned}$$

De ce fait on a $\chi^2 = 0$ si et seulement si $n_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n}$. χ^2 mesure l'écart entre les effectifs observés n_{ij} et ceux attendus $\frac{n_{i\bullet} n_{\bullet j}}{n}$ sous l'hypothèse d'indépendance. On dira que X et Y ne sont pas indépendants si χ^2 est trop grand.

5.2.2 Coefficient de Cramer

Le coefficient de Cramer est défini par

$$C = \sqrt{\frac{\chi^2}{n \min(k-1, l-1)}}.$$

Nous avons $0 \leq C \leq 1$. Si $C \approx 0$, les deux caractères sont indépendants. Si $C = 1$, on parle de dépendance entre X et Y.

5.2.3 Exemple

Nous voulons étudier la liaison entre le type de musique X et l'âge Y. X a trois modalités (chansons, jazz, classique) et Y a quatre modalités (jeunes, adulte femme, adulte homme, vieux). Voici le tableau de contingence :

	Jeunes	Adulte femme	Adulte homme	Vieux	Total
Chansons	69	172	133	27	401
Jazz	41	84	118	11	254
Classique	18	127	157	43	345
Total	128	383	408	81	1000

Etudions la liaison entre X et Y. Nous avons

$$\chi^2 = \sum_{i=1}^{3} \sum_{j=1}^{4} \frac{\left(n_{ij} - \frac{n_{i\bullet} n_{\bullet j}}{n}\right)^2}{\frac{n_{i\bullet} n_{\bullet j}}{n}} = 52.9138.$$

Le coefficient de Cramer est

$$C = \sqrt{\frac{\chi^2}{1000 \min(2,3)}} = \sqrt{\frac{\chi^2}{2000}} \approx 0.16.$$

La dépendance entre X et Y est très faible.

5.3 Liaison entre deux caractères quantitatifs

5.3.1 Représentation graphique : nuage de points.

On suppose que les deux caractères X et Y sont quantitatifs. Pour chaque individu i, on connaît le couple de valeurs (X_i, Y_i) qui lui est attaché. Sur un graphique à axes de coordonnées rectangulaires, nous pouvons représenter chaque élément, par un point d'abscisse X_i et d'ordonnée Y_i. Ce graphique est appelé graphique de corrélation ou nuage de points. Schématiquement, le nuage peut revêtir trois aspects :

1. Les points représentatifs sont distribués sur toute la surface du graphique, à peu près comme s'ils avaient été placés au hasard. C'est le signe qu'il n'y a aucun lien entre les deux variables X et Y : on dit qu'elles sont indépendantes ;
2. Les points représentatifs sont , au contraire rangés le long d'une courbe (droite, arc de cercle,...). Une loi rigoureuse préside alors aux relations entre les deux variables. A chaque valeur de X correspond une seule valeur de Y. On dit qu'il y a liaison fonctionnelle entre Y et X
3. La plupart des phénomènes identifiés à des distributions à deux variables se trouvent entre ces deux extrèmes. Les points représentatifs se distribuent dans une région privilégiée du dessin. Moins le nuage de points a d'épaisseur et plus on se trouve proche de la liaison fonctionnelle : on dit qu'il y a une forte corrélation entre les deux variables. Inversement, plus le nuage de points s'étale, moins ses limites sont précises, plus on est proche de l'indépendance : la corrélation est faible.

5.3.2 Covariance, coefficient de correlation linéaire

La covariance entre les caractères X et Y est défini par

$$\begin{aligned} Cov(X,Y) &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)(Y_i - \overline{Y}_n) \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \overline{X}_n \overline{Y}_n. \end{aligned}$$

La covariance est un indice symétrique, c'est à dire, $Cov(X,Y) = Cov(Y,X)$ et peut prendre toute valeur (négative, nulle ou positive).
Le coefficient de correlation linéaire entre les caractères X et Y est défini par

$$\rho = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

où σ_X et σ_Y les écart-types respectifs de X et σ_Y, sont définis

$$\sigma_X = \left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2\right)^{1/2} \qquad \sigma_Y = \left(\frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2\right)^{1/2}.$$

Nous avons :

- $-1 \le \rho \le 1$
- $|\rho| = 1 \Longleftrightarrow$ les n points (X_i, Y_i) sont alignés
- $\rho = 0 \Longrightarrow$ Pas de liaison linéaire, mais possibilité d'une liaison d'un autre type.
- X et Y indépendantes$\Longrightarrow \rho = 0$

5.3.3 Regression linéaire

Si $|\rho| \simeq 1$, on peut supposer que X est cause de Y. Il est naturel de chercher, dans un ensemble donné de fonctions, la fonction de X approchant Y "le mieux possible" au sens d'un certain critère. On dit que l'on fait la regression de Y sur X. Si l'on choisit pour ensemble de fonctions celui des fonctions affines du type $(aX+b)$, on parle de regression linéaire. C'est le choix que l'on fait le plus fréquemment dans la pratique, le critère le plus usuel étant celui des moindres carrés.

Le critère des moindres carrés. Il consiste à minimiser la quantité

$$S(a,b) = \sum_{i=1}^{n} [Y_i - (aX_i + b)]^2.$$

Solution. La minimisation de S en a et b fournit la solution suivante :

$$\hat{a} = \frac{Cov(X,Y)}{\sigma_X^2} \qquad \hat{b} = \bar{y} - \hat{a}\bar{x}.$$

La droite d'équation $y = \hat{a}x + \hat{b}$ est appelée droite de régression de Y sur X. Elle passe par le point $(\overline{X}_n, \overline{Y}_n)$.

- On appelle valeur ajustée la quantité

5.3.4 Exemple

Sur un échantillon de 10 sujets d'âges différents, on a recueilli les données expérimentales suivant :

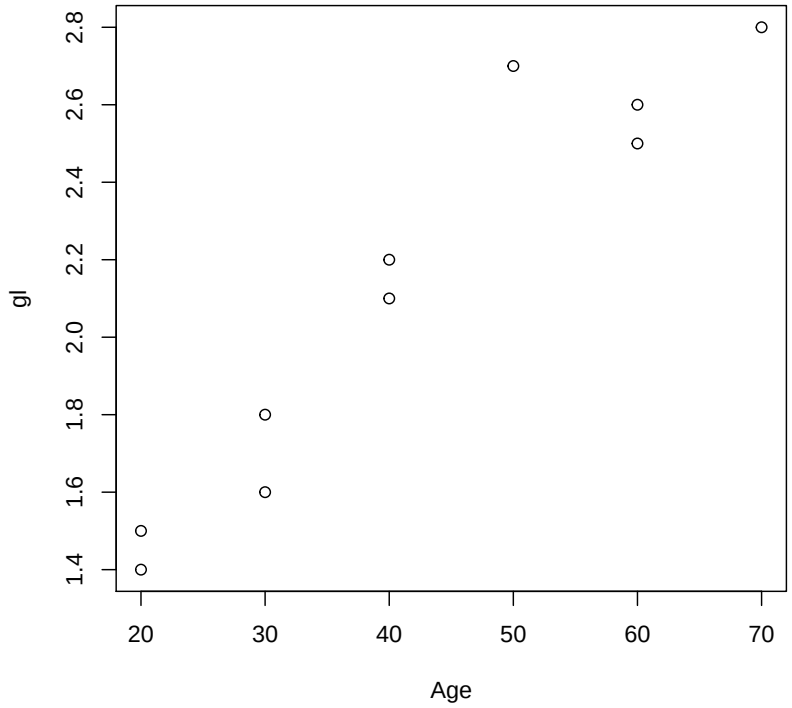
- âge en année
- la concentration sanguine du cholestérol (en g/L).

Age (X_i)	30	60	40	20	50	30	40	20	70	60
gl (Y_i)	1.6	2.5	2.2	1.4	2.7	1.8	2.1	1.5	2.8	2.6

Le taux de cholestérol est-il lié à l'âge ? La relation fonctionnelle est-elle linéaire ? Peut-on prévoir le taux de cholestérol attendu à 35 ans, 75 ans ?

1. Représentation du nuage de points.

```
> Age=c(30,60,40,20,50,30,40,20,70,60)
> gl=c(1.6,2.5,2.2,1.4,2.7,1.8,2.1,1.5,2.8,2.6)
> plot(Age,gl)
```



Les points sont rangés le long d'une droite. On peut donc supposer l'existence d'une relation linéaire entre l'age et le taux de cholesterol.

2. Coefficient de corrélation Le coefficient de corrélation est

```
> cor(Age,gl)

[1] 0.9546712
```

$$r = \approx 0.95$$

Le coefficient de corrélation est positif. Ce qui signifie que l'age et le taux de cholesterol évolue dans le même sens. De plus ils sont fortement corrélés ; ce qui confirme la relation linéaire entre l'age et le taux de cholesterol.

3. Estimation des paramètres

```
> lm(gl~Age)

Call:
lm(formula = gl ~ Age)

Coefficients:
(Intercept)          Age
    0.92391      0.02848
```

$$\hat{a} = 0.03 \quad \hat{b} = 0.92$$

4. La droite de regression est
$$y = 0.03x + 0.92$$
5. Prévisions A 35 ans le taux de cholestérol prédit est $gl = 0.03 * 35 + 0.92 = 1.97$
A 75 ans le taux de cholestérol prédit est $gl = 0.03 * 75 + 0.92 = 3.17$

5.4 Caractère quantitatif et caractère qualitatif

5.4.1 Rapport de correlation

Soient n observations portant simultanément sur un caractère qualitatif X à k modalités et sur un caractère quantitatif Y. Les observations du caractère quantitatif Y se répartissent dans les k modalités de X. Nous notons $n_{i\bullet}$ le nombre d'observations de Y relatifs à la i-ème modalité de X, Y_{ij} la j-ème mesure de Y pour la i-ème modalité de X et $\overline{Y}_i$ la moyenne des observations dans la i-ème modalité

$$\overline{Y}_i = \frac{1}{n_{i\bullet}} \sum_{j=1}^{n_{i\bullet}} Y_{ij}.$$

La moyenne des observations de Y dans la popultion entière est

$$\overline{Y}_n = \frac{1}{n} \sum_{i=1}^{n} n_{i\bullet} \overline{Y}_i.$$

On définit :

- la variance intra-groupe
$$V_{intra} = \frac{1}{n} \sum_{i=1}^{k} n_{i\bullet} \sigma_i^2$$
avec
$$\sigma_i^2 = \frac{1}{n_{i\bullet}} \sum_{j=1}^{n_{i\bullet}} (Y_{ij} - \overline{Y}_i)^2$$
- la variance inter-groupe
$$V_{inter} = \frac{1}{n} \sum_{i=1}^{k} n_{i\bullet} (\overline{Y}_i - \overline{Y}_n)^2$$

Formule de décomposistion de la variance totale σ^2 :

$$\sigma^2 = V_{intra} + V_{inter}.$$

Le rapport de corrélation est défini par

$$\eta^2_{Y|X} = \frac{V_{inter}}{\sigma^2}$$

$\eta^2_{X|Y}$ est un nombre compris entre 0 et 1.

- $\eta^2_{X|Y} = 0 \Rightarrow V_{inter} = 0 \Rightarrow \overline{Y}_i = \overline{Y}_n$. Ce qui signifie que les moyennes de Y sont les mêmes dans toutes les modalités de X. En moyenne, les données ne diffèrent pas selon qu'elles se trouvent dans telle ou telle modalité de X.
- $\eta^2_{X|Y} = 1 \Rightarrow V_{intra} = 0 \Rightarrow Y_{ij} = \overline{Y}_i$. Les données diffèrent d'un groupe à l'autre mais à l'interieur même de chaque groupe, il n'y a aucune variabilité.

Remarque 5.4.1. *Si $\eta^2_{X|Y}$ est proche de 1, c'est que le caractère X explique une grande partie de la variabilité des données alors que si sa valeur est proche de 0, elle n'en explique que très peu.*

5.4.2 Exemple

Liaison entre le sexe (caractère X) et le salaire (caractère Y).
Le caractère X admet deux modalités : femme et homme.

Salaire des femmes

1955	1764	1668	1441	1970	1795	1716	1911	1660	2001
1744	1676	1695	1652	1626	1698	1656	1739	1789	1716
1684	1445	1646	1617	1630	1440	1850	1252	1493	1537

Salaire des hommes

2283	2010	1970	2019	1941	2024	2046	1962	1948	2071
2108	1880	2008	2119	2030	2014	1919	1837	2094	2169

Soient n_F, l'effectif des femmes ; $\overline{Y}_F$ la moyenne des salaires des femmes ; σ_F^2 la variance des salaires des femmes ; n_H, l'effectif des hommes ; $\overline{Y}_H$ la moyenne des salaires des hommes ; σ_H^2 la variance des salaires des hommes ; $\overline{Y}$ la moyenne générale des salaires (hommes et femmes).
Nous avons

$$n_F = 30 \quad \overline{Y}_F = 1682.2 \quad \sigma_F^2 = 26959.56$$
$$n_H = 20 \quad \overline{Y}_H = 2022.6 \quad \sigma_F^2 = 9925.44$$
$$\overline{Y} = \frac{30\overline{Y}_F + 20\overline{Y}_H}{30+20} = 1818.36$$

La variance inter-groupe est :

$$V_{inter} = \frac{1}{50}\left\{n_F\left(\overline{Y}_F - \overline{Y}\right)^2 + n_H\left(\overline{Y}_H - \overline{Y}\right)^2\right\} = 27809.32$$

La variance totale est $\sigma^2 = 47955.23$. Le rapport de correlation est

$$\eta_{Y|X} = \frac{V_{inter}}{\sigma^2} \approx 0.58.$$

On peut considérer que le caractère sexe explique environ 58% de la variabilité des salaires observés.

5.5 Exercices

Exercice 3. Le tableau suivant donne la distance de freinage d'un véhicule roulant sur route sèche en fonction de sa vitesse.

Vitesse en km/h	40	50	60	70	80	90	100	110
Distance en m	8	12	18	24	32	40	48	58

1. Représenter cette série statistique par un nuage de points (**distance** en fonction de la **vitesse**). Commenter
2. Calculer le coefficient de corrélation linéaire. Commenter.
3. Calculer la vitesse moyenne et la distance moyenne de freinage.
4. Déterminer la droite de regression de la distance de freinage en fonction de la vitesse
5. Estimer, à l'aide de la droite de regression, la distance de freinage d'un véhicule roulant à $\mathbf{120}\boldsymbol{km/h}$?

Chapitre

6 Analyse descriptive d'une série chronologique

6.1 Présentation

6.1.1 Définitions

On s'intéresse à l'évolution au cours du temps d'un phénomène, dans le but de décrire, expliquer puis prévoir ce phénomène dans le futur. On dispose ainsi d'observations à des dates différentes, c'est à dire d'une suite de valeurs numériques indicées par le temps appelée série chronologique (chronique ou série temporelle). Les séries chronologiques sont présentes dans de nombreux domaines d'application (démographie,économie, écologie, finance, médecine, informatique...)

Exemple 6.1.1.
- *Température maximale journalière*
- *Chiffre d'affaire trimestriel d'une entreprise*
- *Indice mensuel des prix à la consommation*
-

Soit $(X_t, t \in \mathbb{T})$ une série chronologique ; l'ensemble $\mathbb{T}$ est appelé espace des temps. Nous avons en général $\mathbb{T} \subseteq \mathbb{N}$ ou $\mathbb{T} \subseteq \mathbb{Z}$.

On donne deux dimensions au temps :

- le mois, unité de référence correspondant aux dates d'observation ; le mois peut être le mois véritable mais également le trimestre, le semestre, etc.
- l'année composée d'un nombre p de mois ; le nombre p est appelé période ; par exemple, $p = 4$ pour les observations trimestrielles, $p = 12$ pour les observations mensuelles.

Soit X_t l'observation d'une grandeur X à la date t. Si les observations sont faites sur n années, et chaque année contenant p mois, on notera X_{ij} l'observation du mois j de l'année i. Nous avons

$$X_{ij} = X_t \quad \text{avec } t = (i-1)p + j.$$

Le mois t est le j-ème mois de la i-ème année. Si $\mathbb{T} = \{1, \ldots, T\}$ alors le nombre total d'observations est $T = np$. Nous avons donc deux façons de présenter unesérie chronologique sous forme de tableau :

t	1	2	...	T
X_t	X_1	X_2	...	X_T

Mois / Années	mois 1	mois 2	...	mois j	...	mois p
année 1	X_{11}	X_{12}	...	X_{1j}	...	X_{1p}
année 2	X_{21}	X_{22}	...	X_{ij}	...	X_{2p}
⋮						
année i	X_{i1}	X_{i2}	...	X_{ij}	...	X_{ip}
⋮						
année n	X_{n1}	X_{n2}	...	X_{nj}	...	X_{np}

Exemple 6.1.2. *Chiffre d'affaires trimestriel d'une entreprise (en millions de francs)*

t	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
X_t	2	8	6	12,5	5	10.5	9	15	7	12	10,5	17	8.5	14.5	12	19

Mois / Années	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4
1976	2	8	6	12.5
1977	5	10.5	9	15
1978	7	12	10.5	17
1979	8.5	14.5	12	19

Exemple 6.1.3. *Chiffre d'affaires trimestriel d'une entreprise (en millions de francs)*

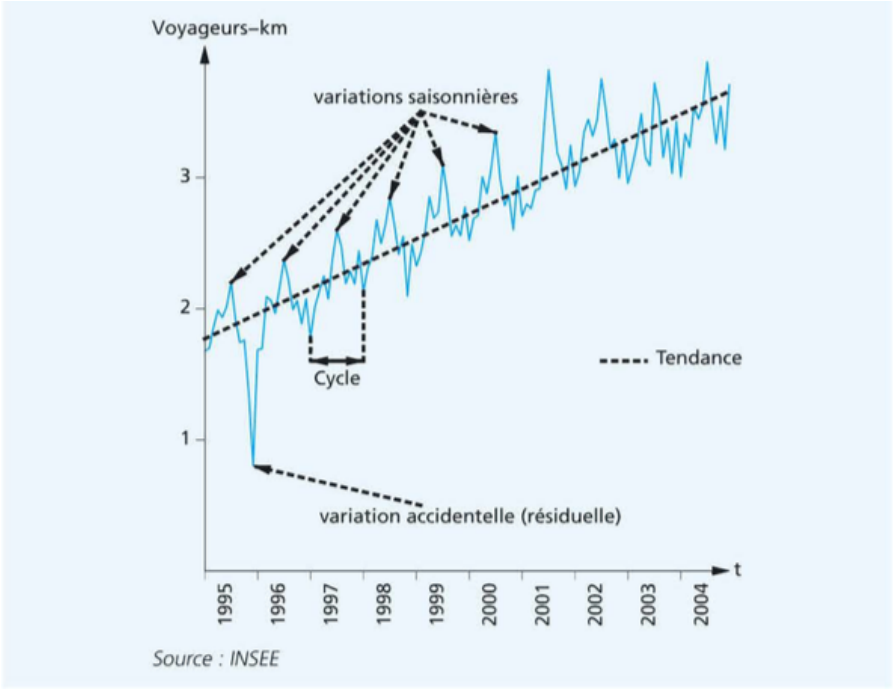
Mois / Années	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4
2000	5,5	6,3	16,9	32,4
2001	23,1	17,5	37,8	62,7
2002	40,6	28,7	58,7	93,3
2003	58,5	39,9	79,5	123,1

6.1.2 Les composantes d'une série chronologique

Soit $(X_t, t \in \mathbb{T})$ une série chronologique. On distingue différentes composantes fondamentales dans une série chronologique :

- **la tendance ou trend ou composante tendancielle T_t** indiquant l'évolution à long terme du phénomène. Elle traduit le comportement "moyen" de la série.
- **Le cycle (ou composante cyclique)**. Il s'agit d'un phénomène se répétant sur des durées qui ne sont pas fixes et généralement longues. Sans informations spécifiques, il est généralement très difficile de dissocier tendance et cycle.
- **la composante saisonnière ou saisonnalité S_t** correspond à un comportement qui se répète avec une certaine périodicité p ($p = 12$ pour des données mensuelles, $p = 4$ pour des données trimestrielles...). Ce sont des fluctuations s'inscrivant dans le cadre de l'année et qui se reproduisent de façon plus ou moins identiques d'une année à l'autre ; **la période notée p des variations saisonnières est la longueur exprimée en unité de temps séparant deux variations saisonnières dues à un même phénomène.**

- **la composante résiduelle** ε_t représentant des fluctuations irrégulières et imprévisibles ; ces fluctuations supposées en général de faible amplitude ; elles traduisent l'effet des facteurs perturbateurs non permanents (grèves, guerre, intempéries,...)



Trafic ferroviaire sur les trains à grande vitesse (TGV) sur la période janvier 1995-décembre 2005, données mensuelles

Remarque 6.1.1. *Ces trois composantes ne sont pas toujours simultanément présentes dans une série chronologique. Certaines séries n'ont pas de tendance, d'autres n'ont aucune composante saisonnière. D'autres n'ont pas de composantes résiduelle.*

Nous supposons que :

- le mouvement saisonnier est périodique de période p :

$$S_t = S_{t+p} = S_{t+2p} = \ldots ;$$

  le mouvement saisonnier relatif au mois j est $S_{ij} = S_j$ quelque soit l'année i.

- **Principe de conservation des aires :** sur une année, l'influence des variations saisonnières est nulle.

Le traitement des séries chronologiques peut avoir pour objectifs d'isoler et estimer une tendance, isoler et estimer une composante saisonnière, et désaisonnaliser la série, de réaliser une prévision, de construire un modèle explicatif en terme de causalité.

6.1.3 Représentations graphiques

Les deux représentations de la série temporelle conduisent à deux types de représentations graphiques :

– Le chronogramme : on représente dans un repère orthonormé les points (t, X_t) que l'on relie par des segments de droite ; ce graphique permet une analyse sur l'ensemble

des n années. l'étude d'une série chronologique commence par l'examen de son chronogramme ; Il en donne une vue d'ensemble, montre certains aspects, comme des valeurs atypiques, d'éventuelles ruptures, un changement dans la dynamique de la série.

– On représente les points (j, Y_{ij}) que l'on relie par des segments de droites, ceci pour chacune des années i; ce graphique permet une analyse année par année et une comparaison entre les différentes années

6.1.4 Modélisation d'une série chronologique

Un modèle est une image simplifiée de la réalité qui vise à traduire les mécanismes de fonctionnement du phénomène étudié et permet de mieux les comprendre.On distingue deux types de modèles : les modèles déterministes et les modèles stochastiques. Dans ce cours, nous nous limitons aux modèles déterministes. Les deux modèles déterministes les plus utilisés sont :

1. le modèle additif correspondant à des variations saisonnières dont la composition avec la tendance conduit à une modulation d'amplitude constante :

$$X_t = T_t + S_t + \varepsilon_t.$$

Principe de conservation des aires :

$$\sum_{j=1}^{p} S_j = 0.$$

2. le modèle multiplicatif correspondant à une modulation d'amplitude variable croissante avec la tendance :

$$X_t = T_t \times (1 + S_t) \times (1 + \varepsilon_t).$$

Principe de conservation des aires :

$$\sum_{j=1}^{p} S_j = 0.$$

Remarque 6.1.2. *Nous observons dans la litterature d'autre forme du modèle multiplicatif :*

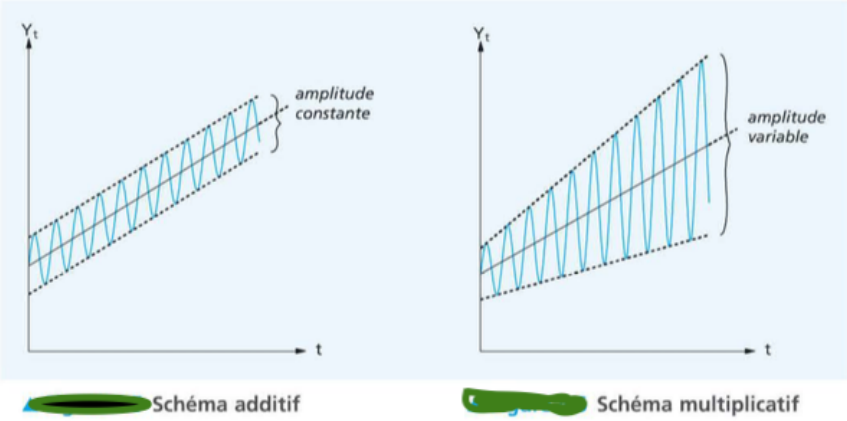
— $X_t = T_t \times S_t \times \varepsilon_t$

— $X_t = T_t \times S_t + \varepsilon_t$ *(souvent qualifié de modèle mixte). Le principde de conservation des aires dans ce cas est*

$$\sum_{j=1}^{p} S_j = p.$$

Dans la suite, lorsque nous parlerons de modèle multiplicatif nous considérons la forme :

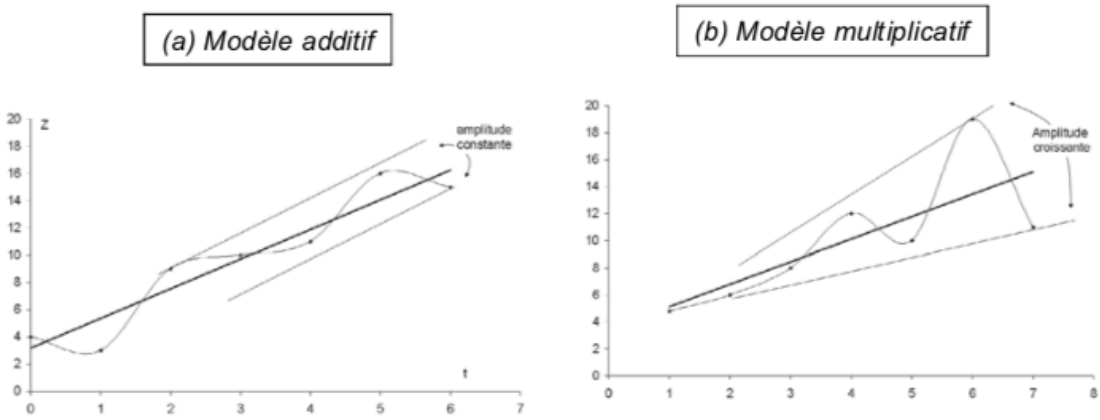
$$X_t = T_t \times (1 + S_t) \times (1 + \varepsilon_t).$$



6.1.5 Choix du modèle

6.1.5.1 Méthode de la bande

On utilise le graphe de la série et la droite passant par les minima et celle passant par les maxima. Si ces deux droites sont parallèles, le modèle est additif. Si les deux droites ne sont pas parallèles, le modèle est multiplicatif.



6.1.5.2 Méthode du profil

On utilise le graphique des courbes superposées. Si les différentes courbes sont parallèles, le modèle est additif. Sinon le modèle est multiplicatif.

6.1.5.3 Méthode du tableau de Buys et Ballot

On calcule les moyennes et écarts-types pour chacune des périodes considérées et on calcule la droite des moindres carrés $\sigma = a\bar{x} + b$. Si a est nul, c'est un modèle additif, sinon , le modèle est multiplicatif.

Exemple 6.1.4. *Nous allons une application de la méthode de Buys-Ballot avec le tableau suivant :*

Mois / Années	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4
1976	2	8	6	12.5
1977	5	10.5	9	15
1978	7	12	10.5	17
1979	8.5	14.5	12	19
Moyenne $\overline{x}$	5.625	11.25	9.375	15.875
Ecart-type σ	2.43349	2.358495	2.21853	2.40767

6.2 Estimation de la tendance

6.2.1 Moyennes mobiles

Le principe de cette technique est de construire une nouvelle série en calculant des moyennes arithmétiques successives de longueur p fixée à partir des données originales. Les moyennes mobiles de longueur égale à la période p permettent d'éliminer ou d'amortir les composantes saisonnière et résiduelle. On procède ainsi au lissage de la courbe pour mettre en évidence la tendance générale.

- On appelle moyenne mobile centrée de longueur impaire $p = 2k+1$ à l'instant t la valeur moyenne des observations

$$M_t = \frac{X_{t-k} + X_{t-k+1} + \ldots + X_{t-1} + X_t + X_{t+1} + \ldots + X_{t+k}}{p}$$

- On appelle moyenne mobile centrée de longueur paire $p = 2k$ à l'instant t la valeur moyenne

$$M_t = \frac{0.5X_{t-k} + X_{t-k+1} + \ldots + X_{t-1} + X_t + X_{t+1} + \ldots + 0.5X_{t+k}}{p}$$

$p = 3 = 2 \times 1 + 1 \Rightarrow k = 1$

$$M_t = \frac{X_{t-1} + X_t + X_{t+1}}{3}$$

$p = 4 = 2 \times 2 \Rightarrow k = 2$

$$M_t = \frac{0.5X_{t-2} + X_{t-1} + X_t + X_{t+1} + 0.5X_{t+2}}{4}$$

Remarque 6.2.1. *La tendance à la date t peut être estimée par la moyenne mobile centrée à la date t de longueur la période p si*

- *la tendance présente une faible courbure*
- *les variations saisonnières sont périodiques de période p et ont une influence nulle sur l'année*
- *les variations résiduelles sont de faible amplitude.*

Remarque 6.2.2. *Les moyennes mobiles peuvent être influencées par les valeurs extrêmes. Dans ce cas, on pourrait calculer les médianes mobiles de même ordre. Les moyennes mobiles donnent une meilleure estimation que les moindres carrés.*

6.2.2 Méthode de Mayer

On ajuste le nuage de points (t, X_t) à une droite passant par les deux points $(\bar{t}_1, \bar{X}_1)$ et $(\bar{t}_2, \bar{X}_2)$ calculés de la manière suivante :

- on découpe la série en deux parties de même effectif
- pour chacune des deux parties, on calcule la moyenne des t et celle des X_t : $(\bar{t}_1, \bar{X}_1)$ et $(\bar{t}_2, \bar{X}_2)$; on peut calculer les points médians au lieu des moyennes ; cela permet de limiter l'influence des valeurs extrêmes.
- il reste à tracer la droite passant par les deux points.

6.2.3 Méthode des moindres carrés

6.2.3.1 Tendance linéaire

On ajuste le nuage de points (t, X_t) à une droite d'équation $at + b$ où le couple (a, b) minimise la distance

$$\sum_{t=1}^{T} (X_t - (at + b))^2.$$

Nous obtenons

$$a = \frac{cov(t, X)}{var(t)} \quad b = \bar{X} - a\bar{t}$$

où

$$cov(t, X) = \frac{1}{T} \sum_{t=1}^{T} tX_t - \bar{t}\bar{X} \quad var(t) = \frac{1}{T} \sum_{t=1}^{T} t^2 - \bar{t}^2$$

$$\bar{X} = \frac{1}{T} \sum_{t=1}^{T} X_t \quad \bar{t} = \frac{1}{T} \sum_{t=1}^{T} t.$$

Remarque 6.2.3. *La droite des moindres carrés ajuste au mieux au sens des moindres carrés (c'est celle qui passe le plus près de l'ensemble des points), mais elle ne modélise pas toujours bien la tendance.*

6.2.3.2 Tendance polynomiale

On peut utiliser la méthode des moindres carrés afin d'ajuster le nuage de points (t, X_t) à un polynôme de degré choisi. L'observation du graphe de la série donne une idée du degré du polynôme (selon la forme de la courbe).

6.3 Variations saisonnières

6.3.1 Estimation des coefficients saisonniers du modèle additif

1. Calculer les moyennes mobiles : M_{ij}
2. Calculer les différences entre les observations et les moyennes mobiles : $X_{ij} - M_{ij}$.
3. Calculer la moyenne S'_j des $X_{ij} - M_{ij}$
4. Calculer la moyenne
$$M' = \frac{1}{p} \sum_{j=1}^{p} S'_j$$
5. Estimer S_j par $\widetilde{S}_j = S'_j - M'$ pour respecter le principe de conservation des aires.

6.3.2 Estimation des coefficients saisonniers du modèle multiplicatif

1. Calculer les moyennes mobiles : M_{ij}
2. Calculer les rapports des observations aux moyennes mobiles : $\dfrac{X_{ij}}{M_{ij}}$.
3. Calculer les moyennes des rapports S'_j des $\dfrac{X_{ij}}{M_{ij}}$ pour $j = 1, \ldots, p$.
4. Calculer la moyenne des moyennes
$$M' = \frac{1}{p} \sum_{j=1}^{p} S'_j.$$
5. Estimer S_j par $\widetilde{S}_j = \dfrac{S'_j}{M'} - 1$.

6.4 Désaisonnalisation

Désaisonnaliser une série chronologique, c'est éliminer la composante saisonnière sans modifier les autres composantes. On appelle observation corrigée des variations saisonnières ou observation désaisonnalisés, la valeur X^*_{ij} cdobtenue en éliminant l'effet saisonnier sur la valeur X_{ij}. On la notera X^*_t. La désaisonnalisation permet de comparer des observations dont les variations saisonnières sont différentes.

- Modèle additif : $X^*_{ij} = X_{ij} - \widetilde{S}_j$
- Modèle multiplicatif : $X^*_{ij} = \dfrac{X_{ij}}{1 + \widetilde{S}_j}$

Remarque 6.4.1. - *Les données X^*_t sont directement comparables car débarrassées de l'effet des saisons et donc du caractère propre de chaque mois. On peut donc comparer par exemple les données du mois de janvier à celles du mois d'août.*

- *On peut avoir une meilleure estimation de la tendance à partir de la série désaisonnalisée.*

6.5 Prévisions

- Modèle additif : la prévision est :

$$X_{ij}^{p} = a\Big[(i-1)p + j\Big] + b + \widetilde{S}_j.$$

- Modèle multiplicatif : la prévision est :

$$X_{ij}^{p} = \Big(a\Big[(i-1)p + j\Big] + b\Big)(1 + \widetilde{S}_j).$$

6.6 Approche générale de la modélisation d'une série chronologique

1. Tracer la série des données et on repère ses principales caractéristiques (tendance, composante saisonnière, observations aberrantes, ...).
2. Estimer la tendance, la composante saisonnière et la composnte résiduelle.
3. Choisir un modèle de série stationnaire pour les variations résiduelles (Chapitres suivants).
4. Prévisions.

6.7 Exemple : Modèle additif

Nous revenons sur le tableau concernant le chiffre d'affaire trimestriel d'une entreprise de 1976 à 1978.

Mois / Années	T_1	T_2	T_3	T_4
1976	2	8	6	12.5
1977	5	10.5	9	15
1978	7	12	10.5	17
1979	8,5	14.5	12	19

Nous avions montré par les méthodes précédentes que le modèle est additif.

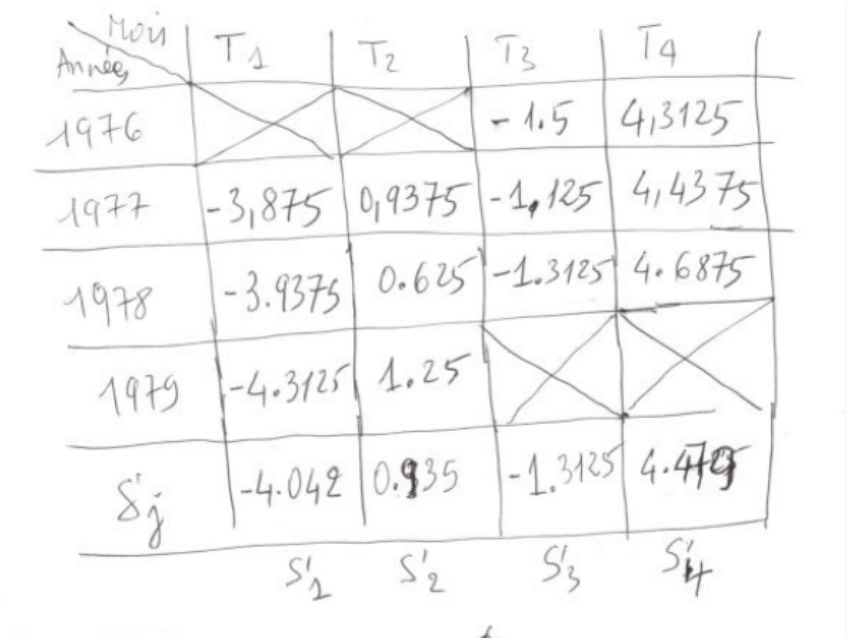
1. **Tableau des moyennes mobiles.** Nous utilisons la formule suivante :

$$M_t = \frac{0.5X_{t-2} + X_{t-1} + X_t + X_{t+1} + 0.5X_{t+2}}{4}$$

	T1	T2	T3	T4
1976			7.5	8.1875
1977	8.875	9.5625	10.125	10.5625
1978	10.9375	11.375	11.8125	12.3125
1979	12.8125	13.25		

2. **Tableau des différences bservations (X_{ij}) et moyennes mobiles (M_{ij})**

$$X_{ij} - M_{ij}$$



Mois / Année	T_1	T_2	T_3	T_4
1976			-1.5	4,3125
1977	-3,875	0,9375	-1,125	4,4375
1978	-3.9375	0.625	-1.3125	4.6875
1979	-4.3125	1.25		
S'_j	-4.042	0.935	-1.3125	4.479
	S'_1	S'_2	S'_3	S'_4

$$S'_1 = \frac{1}{3}\Big(-3.875 - 3.9375 - 4.3125\Big) = -4.042$$
$$S'_2 = \frac{1}{3}\Big(0.9375 + 0.625 + 1.25\Big) = 0.935$$
$$S'_3 = \frac{1}{3}\Big(-1.5 - 1.125 - 1.3125\Big) = -1.3125$$
$$S'_4 = \frac{1}{3}\Big(4.3125 + 4.4375 + 4.6875\Big) = 4.479$$

Comme

$$S'_1 + S'_2 + S'_3 + S'_4 = 0.0595 \neq 0,$$

le principe de conservation des aires n'est pas respectée. Nous passons à l'étape suivante.

3. **Principe de conservation des aires.** Posons

$$M' = \frac{1}{4}\left(S'_1 + S'_2 + S'_3 + S'_4\right)$$

$$= \frac{0.0595}{4} = 0.015375$$

Estimation des coefficients saisonniers

$$\tilde{S}_1 = S'_1 - M' = -4.057375$$

$$\tilde{S}_2 = S'_2 - M' = 0.919625$$

$$\tilde{S}_3 = S'_3 - M' = -1.327875$$

$$\tilde{S}_4 = S'_4 - M' = 4.463625$$

Les coefficients $\widetilde{S}'_1$, $\widetilde{S}'_2$, $\widetilde{S}'_3$ et $\widetilde{S}'_4$ respectent leprincipe de conservation des aires.

Interprétation des coefficients saisonniers : $\widetilde{S}'_1 = -4.057335$ signifie qu'en moyenne on a une baisse de 4.042 millions au trimestre 1 par rapport à l'ensemble de l'année ; $S'_4 = 4.479$ signifie qu'en moyenne on a une hausse de **4.463625** millions au trimestre 4 par rapport à l'ensemble de l'année.

4. **Séries désaisonnalisée :**

$$X^*_{ij} = X_{ij} - \widetilde{S}'_j.$$

Mois / Années	T_1	T_2	T_3	T_4
1976	6.057375	7.080375	7.327875	8,036375
1977	9,057375	9.580375	10.327875	10,536375
1978	11.057375	11.080375	11,827875	12.536375
1979	12.557375	13,580375	13,327875	14.536375

5. **Estimation de la composante résiduelle**

Comme $X_{ij} = T_{ij} + S_j + \varepsilon_{ij}$, on peut estimer la composante résiduelle par :

$$\tilde{\varepsilon}_{ij} = X_{ij} - M_{ij} - \tilde{S}_j \qquad i = 1, \dots, n \quad j = 1, \dots, p.$$

6. **Cas d'un modèle multiplicatif : A la main !**

- Tableau des données

Mois / Années	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4
2000	5,5	6,3	16,9	32,4
2001	23,1	17,5	37,8	62,7
2002	40,6	28,7	58,7	93,3
2003	58,5	39,9	79,5	123,1

- Tableau des moyennes mobiles $\boldsymbol{M_{ij}}$

Mois / Années	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4
2000			17.475	21.075
2001	25.0875	31.4875	37.4625	41.05
2002	45.0625	51.5	57.5625	61.2
2003	65.2	71.525		

- Tableau des $\frac{X_{ij}}{M_{ij}}$

Mois / Années	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4
2000			0.96709585	1.53736655
2001	0.92077728	0.5557761	1.00900901	1.5274056
2002	0.90097087	0.55728155	1.01976113	1.5245098
2003	0.89723926	0.55784691		
S'_j	0.9063291367	0.5569681867	0.9986219967	1.52976065

- Moyenne des S'_j

$$M' = \frac{0.9063291367 + 0.5569681867 + 0.9986219967 + 1.52976065}{4}$$
$$= 0.99792$$

- Estimation des coefficients saisonniers

$$\overline{S}_1 = \frac{S'_1}{M'} - 1 = -0.09178176$$
$$\overline{S}_2 = \frac{S'_2}{M'} - 1 = -0.4418709$$
$$\overline{S}_3 = \frac{S'_3}{M'} - 1 = 0.0007034674$$
$$\overline{S}_4 = \frac{S'_4}{M'} - 1 = 0.5329492$$

Bibliographie

[1] Christophe Hurlin, Valérie Mignon, Statistique et probabilités en économie-gestion, Openbook Licence/Bachelor, Dunod, 2015.

[2] Gérard Carlot, Cours de Statistique descriptive, Dunod, Paris, 1973.