

Sciences de gestion

Synthèse  
de cours &  
exercices  
corrigés

# Statistique descriptive

avec Excel  
et la calculatrice



- Pour les étudiants en sciences de gestion, en économie et en sciences humaines
- Près de 40 problèmes et exercices corrigés avec Excel ou la calculatrice
- Retrouvez les données Excel et des exemples supplémentaires sur [www.pearson.fr](http://www.pearson.fr)

collection  
**Synthex**

**PEARSON**  
Education

**Étienne BRESSOUD**  
**Jean-Claude KAHANÉ**

**Sciences de gestion**

Synthèse & exercices  
de cours corrigés

# Statistique descriptive

## Applications avec Excel et la calculatrice

**Étienne Bressoud**

Université Paris 8 Vincennes-Saint-Denis

**Jean-Claude Kahané**

Université Paris 8 Vincennes-Saint-Denis

**Directeur de collection : Roland Gillet**

Université Paris 1 Panthéon-Sorbonne

collection  
**Synthex**



ISBN : 978-2-7440-4052-8

ISSN : 1768-7616

Copyright© 2009 Pearson Education France

Tous droits réservés

Mise en page : [edito.biz](http://edito.biz)

Aucune représentation ou reproduction, même partielle, autre que celles prévues à l'article L. 122-5 2° et 3° a) du code de la propriété intellectuelle ne peut être faite sans l'autorisation expresse de Pearson Education France ou, le cas échéant, sans le respect des modalités prévues à l'article L. 122-10 dudit code.

# Sommaire

<b>Les auteurs.....</b>	<b>IV</b>
<b>Préface .....</b>	<b>V</b>
<b>Introduction .....</b>	<b>VII</b>
<b>Chapitre 1 • Introduction à la statistique descriptive .....</b>	<b>1</b>
<b>Chapitre 2 • Les caractéristiques de tendance centrale .....</b>	<b>35</b>
<b>Chapitre 3 • Les caractéristiques de dispersion .....</b>	<b>63</b>
<b>Chapitre 4 • Les caractéristiques de forme et de concentration ....</b>	<b>83</b>
<b>Chapitre 5 • Les séries bivariées.....</b>	<b>107</b>
<b>Chapitre 6 • La régression .....</b>	<b>145</b>
<b>Chapitre 7 • Les séries chronologiques .....</b>	<b>185</b>
<b>Chapitre 8 • Les indices.....</b>	<b>219</b>
<b>Index.....</b>	<b>246</b>

# Les auteurs

**Étienne Bressoud**, docteur ès sciences de gestion et normalien agrégé en sciences économiques, est maître de conférences à l'université Paris 8 Vincennes-Saint-Denis et professeur associé de marketing à l'*European Business School* (EBS) Paris. Il enseigne la statistique descriptive, les études quantitatives appliquées au marketing, et assure des formations professionnelles sur un logiciel d'analyse de données et de statistiques pour Microsoft Excel.

Contact : <http://bressoud.blogspot.com>

**Jean-Claude Kahané** est professeur agrégé de mathématiques à l'université Paris 8 Vincennes-Saint-Denis et professeur associé à l'École nationale d'assurance (ENASS, un institut du CNAM), en formation initiale et continue. Membre du jury de CAPES externe de sciences économiques et sociales, il enseigne les statistiques, les probabilités et les mathématiques.

# Préface

Née voici cinq millénaires pour dénombrer les richesses et les hommes en état de porter des armes, la statistique est de plus en plus une science de pleine actualité, quand elle ne la fait pas. Il ne se passe pas une semaine, voire une journée, sans que nous en lisions ou évoquions des utilisations, que ce soit pour mesurer la santé de notre économie, la cote de popularité d'un homme politique, l'avis de l'opinion sur tel ou tel sujet, le succès d'un média ou d'une émission, ou autre.

Nous pouvons même affirmer que la science statistique devrait faire partie du bagage intellectuel minimal de « l'honnête homme » de notre époque, dont la caractéristique essentielle est la profusion d'informations de tout ordre, qui plus est accessibles pratiquement en temps réel grâce à Internet et à la convergence numérique. Pour leur gestion, les entreprises élaborent des entrepôts de données – des *datawarehouses* – qui se remplissent automatiquement et systématiquement, au point d'ailleurs de devenir difficilement exploitables de façon directe, sans recours à l'analyse. Le succès actuel du *datamining* n'est rien d'autre que celui de la pensée statistique, revue avec des notions de marketing.

Devant une telle accumulation d'informations, il est nécessaire, indispensable, pour chacun, de posséder les clés pour structurer, hiérarchiser, présenter, illustrer, comprendre, expliquer.

La statistique est la science de l'apprentissage du doute, cette notion parfois si importante, y compris dans le pays de Descartes, de l'appréhension de l'incertain, du refus de la pensée unique. En un mot, la statistique est la science de la diversité. Et cette diversité s'exprime – forcément – par trois voies principales : diversité de domaines, diversité d'approches, diversité d'objectifs.

Les domaines d'abord : au fil des siècles, les champs d'application de la statistique se sont multipliés. Au comptage initial des ressources, pour la gestion de l'État, se sont ajoutés l'astronomie (Tycho Brahe, Johannes Kepler, Galilée), l'agronomie, la démographie, la biométrie (Galton), l'économie (création en 1933 de la Société d'économétrie), le marketing, la gestion d'entreprise, la finance, la mesure d'audience, etc. Chacun de ces champs de recherche et d'application a apporté ou apporte encore ses innovations, tant il est de sujets non ou mal résolus ou nouveaux.

Les approches ensuite : à l'origine science du dénombrement et de la description, la statistique s'est lentement enrichie d'éléments plus conceptuels. Une illustration marquante en est, au XVII<sup>e</sup> siècle, la recherche de constantes de comportements – des paramètres

comme le nombre moyen d'enfants par femme, ou le nombre moyen de personnes par logement – et la technique du multiplicateur pour permettre des estimations et des extrapolations : c'est le règne de l'école anglaise dite de l'arithmétique politique de John Graunt et de William Petty.

Le cadre théorique s'affirme plus tard, soit avec une vision géométrique, comme les moindres carrés de Carl Friedrich Gauss, soit avec une optique probabiliste, merveilleusement utilisée dans les années 1920 par sir Ronald Fisher avec une présentation innovante et générale de la théorie statistique.

Géométrie et probabilités se trouvent également derrière les techniques dites d'analyse des données – projection de nuages complexes de points sur des plans adéquats – comme l'analyse en composantes principales ou l'analyse des correspondances, d'une part, et les modèles stochastiques de plus en plus sophistiqués tels les Arima et leurs divers petits cousins Starima, Arch, Garch, etc., d'autre part.

Les objectifs enfin : nous en distinguerons deux principaux.

Le premier consiste à avancer sur le chemin de la théorie « pure », de la recherche pilotée par les mathématiques, de la conceptualisation. Axe fondamental s'il en est, ne serait-ce que pour fonder la statistique comme une théorie scientifique et la faire progresser intrinsèquement et en liaison avec les autres théories mathématiques.

Le second repose sur la volonté d'application, quel qu'en soit le domaine, la confrontation aux données ; on est dans le domaine de la description, de la visualisation, de la mesure de paramètres – le principe de réduction de Fisher –, permettant de caractériser le phénomène étudié dans ses principales lignes. La statistique appliquée est aussi noble que la statistique théorique, et très proche des origines historiques mêmes. Il est vrai que l'informatique est un appui majeur pour la manipulation des fichiers de données et la mise en œuvre des méthodes.

Le présent ouvrage d'Étienne Bressoud et de Jean-Claude Kahané relève ouvertement de la statistique appliquée et procède d'une volonté claire d'aborder l'opérationnalité des méthodes.

Le livre joue ainsi sur deux tableaux complémentaires.

En premier lieu, les concepts de base sont développés dans le corps des chapitres : caractéristiques de tendance centrale ou de dispersion, de forme ou de concentration, indices, tableaux croisés, modèle linéaire ou régression, séries temporelles. Ensuite, après les présentations des outils de référence, chaque chapitre est suivi d'exercices et de problèmes sur de vraies données, avec traitement et correction, à partir des possibilités d'outils comme la calculatrice graphique et le tableur Excel, qui possèdent un grand nombre de fonctionnalités pour passer à l'application concrète et, somme toute, simple.

C'est ce qui en fait l'originalité, et aidera à vulgariser la pensée statistique auprès de nombreux étudiants de l'enseignement supérieur. Que les auteurs en soient remerciés.

*Philippe Tassi*

*Directeur Général Adjoint de Médiamétrie*

*Professeur Associé à l'Université Paris 2*

# Introduction

Ce livre est avant tout l'histoire d'une rencontre entre deux enseignants et la mise en commun de leur pratique et de leur écoute auprès des étudiants.

Les statistiques sont aujourd'hui incontournables et leur enseignement s'est généralisé. Il existe de nombreux ouvrages de statistiques, souvent de qualité, mais il nous a paru intéressant d'en concevoir un qui mette en avant le côté actuel et opérationnel de la statistique.

## **De la statistique à sa mise en œuvre par calculatrice graphique et tableur Excel**

Cet ouvrage veut proposer aux étudiants d'économie, de gestion, de marketing, des secteurs de l'assurance, un outil qui soit avant tout une aide pour leur « pratique de la statistique ».

La partie théorique des chapitres est assez synthétique et expose les concepts en présentant succinctement les grands noms qui ont fait la statistique.



La partie pratique comporte de nombreux exercices qui sont corrigés avec deux outils fondamentaux en statistique : la calculatrice et le tableur. Nous avons choisi la calculatrice graphique Texas Instrument TI-84 Plus Silver Edition et le tableur Excel, car il est couramment utilisé par les étudiants. En dépit de ces choix, les exercices peuvent être effectués à l'aide d'autres tableurs et calculatrices graphiques, notamment les calculatrices Casio. Pour chaque exercice, nous précisons si la correction est proposée avec la calculatrice ou avec le tableur grâce à un pictogramme dans la marge.

Ces deux outils, comme de nombreux rappels de techniques mathématiques, font l'objet de développements construits comme autant d'aides à leur mise en œuvre. Par ailleurs, les exercices sont présentés à partir de données réelles et récentes obtenues auprès des grands organismes de statistique.

Les parties théoriques peuvent être prolongées par des exemples complémentaires disponibles sur le site Internet de l'éditeur, [www.pearsoneducation.fr](http://www.pearsoneducation.fr). Sur ce même site se trouvent également les tableaux de données et les corrections de chaque exercice, au format Excel.

## **Un cours de statistique descriptive élargi**

En ce qui concerne le contenu, cet ouvrage est avant tout conçu comme un ouvrage de statistique descriptive ; oui, mais... il nous a paru difficile de parler de statistique sans faire quelques incursions en probabilité, sans apporter une initiation à la statistique inférentielle qui donne son vrai sens à la statistique.

Le lecteur trouvera dans les trois premiers chapitres les bases de la statistique descriptive : le vocabulaire, les principaux graphiques, ainsi que les paramètres de position et de dispersion des séries univariées.

Le quatrième chapitre, qui traite des caractéristiques de forme et de concentration, débute par une introduction à la loi normale (loi de Laplace-Gauss).

Les chapitres 5, 6 et 7 traitent respectivement des séries bivariées, de la régression linéaire et des séries chronologiques. Les chapitres 5 et 6 sont l'occasion d'introduire les tests d'hypothèses, et notamment les tests du khi-deux, de Student et de Fisher.

Enfin, le chapitre 8 est consacré aux indices élémentaires et synthétiques.

### **Merci**

Nous tenons à remercier ici vivement Philippe Tassi, pour ses conseils et sa relecture méticuleuse et éclairée. Également un grand merci à Christine Dhers, enseignante de mathématiques, pour sa disponibilité et sa passion pour les statistiques.

Nous espérons que ce manuel transmettra aux étudiants notre engouement pour la statistique et l'envie de découvrir les ouvrages cités dans la bibliographie propre à chaque chapitre, et qu'il sera pour eux un compagnon efficace de leur réussite.

# Introduction à la statistique descriptive

1. Terminologie.....	2
2. Présentation des données.....	8
3. Représentations graphiques des séries à une variable .....	10

## Problèmes et exercices

1. De la série brute à la présentation des statistiques ..	18
2. Représentations graphiques simples .....	22
3. L'histogramme .....	25
4. Discréttisation des données ....	26
5. Les polygones .....	29

Les méthodes de la statistique descriptive (statistique déductive) permettent de mener des études à partir de données exhaustives, c'est-à-dire concernant tous les individus de la population concernée par l'étude. Comme le rappelle André Vessereau (voir bibliographie), l'idée première et toujours fondamentale de la statistique descriptive est celle de dénombrement.

Quand les données ne concernent qu'un échantillon de la population, comme dans le cas des sondages, on a recours à la statistique inférentielle (statistique inductive), qui utilise la théorie des probabilités.

Globalement, la statistique reste très liée à la science du hasard, puisque les recensements nous fournissent des fréquences d'apparition auxquelles on fait jouer le même rôle qu'à la probabilité. Déjà, les manuscrits de Gottfried Leibniz, rédigés au début des années 1680, se situaient, à partir des travaux de John Graunt, dans la perspective d'une « synthèse entre science de la population et calcul des probabilités ».

Ce premier chapitre présente les principales clefs de lecture de la statistique. La terminologie usuelle y est exposée, ainsi que la forme et le contenu des tableaux de données.

Deux annexes, proposées en fin de chapitre, sont consacrées à la prise en main d'Excel (annexe 1.1), ou de tout autre tableur équivalent, et d'une calculatrice graphique, Texas Instrument (annexe 1.2), ou de toute autre calculatrice approchante. L'utilisation de ces outils facilitera la compréhension et la résolution de tous les exemples numériques des parties théoriques et des problèmes et exercices qui suivent.

# 1 Terminologie

Comme toute science, la statistique a son vocabulaire, qu'il est primordial de définir de façon rigoureuse afin d'indiquer le groupe sur lequel porte l'étude, les caractères ou variables relevés sur chacun des individus et les différents types de caractères.

## 1.1 LA POPULATION

Le terme de population statistique est antérieur à la démographie et s'appliquait à l'origine à des catégories d'humains. Les populations n'étaient en effet pas pensées en bloc, leurs membres n'étant pas considérés comme égaux. Par exemple, on comptait les hommes en état de porter des armes, les individus soumis à l'impôt, etc. La démographie est venue plus tard, avec l'idée d'égalité des individus, qui a mené à la notion de recensement.

En statistique, le terme de population est plus général et peut désigner des humains, mais aussi des objets, des villes, des pays, des entreprises, des logements, etc., l'essentiel étant, comme pour la définition d'un ensemble en mathématiques, que l'on puisse dire clairement de tout élément qu'il appartient ou n'appartient pas à la population.

Les villes européennes de plus de 100 000 habitants, les voitures immatriculées en France, les départements français d'outre-mer sont autant d'exemples de population.

### Définition

La **population statistique** est l'ensemble des éléments sur lesquels porte l'étude. Les éléments de la population sont appelés **individus statistiques** ou unités statistiques. La population constitue l'univers de référence de l'étude. Si la population comporte N individus, on notera  $\Omega = \{\omega_1 ; \omega_2 ; \dots ; \omega_N\}$  les N individus qui la composent. Un **échantillon** de taille n est un sous-ensemble formé de n individus de la population ( $n \leq N$ ).

La notion d'échantillon est fondamentale, car, en règle générale, la population entière n'est pas disponible ou observable. Dans ce cas, seul un échantillon est étudié et les résultats obtenus sont extrapolés à la population (voir P. Roger, chapitre 5). Par exemple, lorsqu'un magazine souhaite connaître la personnalité préférée des Français, il interroge seulement un échantillon de Français, généralement 1 000 individus, et non toute la population résidant en France métropolitaine, soit plus de 60 millions d'individus.

## 1.2 NOTION DE CARACTÈRE OU VARIABLE STATISTIQUE

Chaque individu d'une population peut être décrit relativement à un ou plusieurs caractères ou variables statistiques.

### Définition

Une **variable statistique**, ou **caractère statistique**, est une application définie sur une population statistique et à valeurs dans un ensemble  $M$ , appelé ensemble des modalités. Les **modalités** correspondent aux valeurs possibles de la variable statistique. Une variable statistique définit une partition sur une population, chaque individu appartenant à une et une seule modalité.

Si le nombre de modalités est noté  $r$ , l'ensemble des modalités de la variable  $X$  sera noté :  $M = \{x_1 ; x_2 ; \dots ; x_r\}$ .

### Exemple 1.1

#### Une population statistique

Considérons les données suivantes concernant le nombre de femmes et d'hommes dans la population résidant en France métropolitaine en 2006 (en milliers) :

Femmes	Hommes
31 444	29 722

Source : Insee, recensement de la population, 2007 (champ : France métropolitaine)

La population étudiée est la population résidant en France métropolitaine recensée en 2006 et la variable étudiée est le sexe. Cette variable peut prendre deux valeurs possibles appelées modalités : féminin ou masculin. Ces modalités sont en général numérotées : si la variable étudiée, ici le sexe, est notée  $X$ , les deux modalités seront respectivement notées  $x_1$  (pour féminin) et  $x_2$  (pour masculin).

Une des premières opérations de la statistique consiste à recenser le nombre et/ou le pourcentage d'individus qui présentent une modalité déterminée d'une variable. C'est ainsi qu'à chaque modalité est associé un effectif et/ou une fréquence.

### Définitions

L'**effectif** (aussi appelé fréquence absolue) de la modalité  $x_i$  est noté  $n_i$  et désigne le nombre d'individus de la population présentant la modalité  $x_i$ . L'effectif total de la population  $n$  est alors :  $n = n_1 + n_2 + \dots + n_r$ , soit  $n = \sum_{i=1}^r n_i$  (la somme des  $n_i$  pour  $i$  variant de 1 à  $r$ , et la lettre grecque sigma,  $\Sigma$ , désignant la somme).

La **fréquence** (par défaut fréquence relative) de la modalité  $x_i$  est notée  $f_i$  et est définie par :  $f_i = n_i / N$  ; la fréquence exprime la proportion d'individus présentant une modalité donnée. Elle peut s'exprimer sous la forme d'un nombre décimal (en général avec une précision de quatre chiffres après la virgule) ou sous la forme d'un pourcentage.

<b>Propriété</b>	Soit $X$ une variable à $r$ modalités :
	$0 \leq f_i \leq 1$
	$\sum_{i=1}^r f_i = 1$ (ou, en pourcentage : $\sum_{i=1}^r f_i = 100$ )
<b>Exemple 1.2</b>	<p><b>Effectifs et fréquences</b></p> <p>Reprendons l'exemple précédent sur le sexe des individus de la population résidant en France métropolitaine. Les effectifs respectifs de ces modalités sont notés <math>n_1 = 31\ 444</math> et <math>n_2 = 29\ 722</math>, avec <math>n = n_1 + n_2 = 61\ 166</math> milliers, effectif total de la population.</p> <p>Les fréquences sont telles que : <math>f_1 = n_1 / n = 31\ 444 / 61\ 166 = 0,5141</math> et <math>f_2 = n_2 / N = 29\ 722 / 61\ 166 = 0,4859</math>, soit 51,41 % de femmes et 48,59 % d'hommes.</p>

L'exemple 1.1 a mis en évidence une des deux natures des variables statistiques : la variable qualitative. Le sexe est une variable qualitative, car ses modalités ne sont pas des nombres. Une variable quantitative est une variable dont les modalités sont numériques. Le poids d'un individu, l'âge, le nombre d'enfants par ménage, le salaire constituent des exemples de variables quantitatives.

## 1.3 LES VARIABLES QUALITATIVES

---

<b>Définition</b>	Une variable statistique est dite de nature <b>qualitative</b> si ses modalités ne sont pas mesurables. Les modalités d'une variable qualitative sont les différentes catégories d'une nomenclature. Ces catégories doivent être exhaustives (chaque individu est affecté à une modalité) et incompatibles (un individu ne peut être affecté à plusieurs modalités) de façon à créer une partition.
	<p>Le sexe, la profession, l'état matrimonial sont quelques exemples de variables qualitatives. Pour ses enquêtes auprès des ménages, l'Insee utilise la nomenclature des Professions et catégories socioprofessionnelles (PCS-2003).</p> <p>Les modalités d'une variable qualitative peuvent être classées sur deux types d'échelle : nominale ou ordinale. À ces deux types d'échelle correspondent deux types de variables qualitatives.</p>
	<p><b>Variables qualitatives nominales</b></p> <p>Les variables qualitatives nominales ne se mesurent pas. Cependant, leurs modalités peuvent être codées. L'ordre et l'origine de la codification sont arbitraires, cette codification pouvant être numérique, alphabétique ou alphanumérique. Les individus d'une même catégorie sont réputés « équivalents » pour la variable étudiée.</p>
<b>Définition</b>	Une variable statistique qualitative est dite définie sur une échelle <b>nominale</b> si ses catégories ne sont pas naturellement ordonnées.

**Exemple 1.3****Codage d'une variable qualitative nominale**

Le tableau suivant indique les différentes catégories de la variable nominale Professions et catégories socioprofessionnelles (CSP) :

Code	Catégorie
1	Agriculteurs exploitants
2	Artisans, commerçants et chefs d'entreprise
3	Cadres et professions intellectuelles supérieures
4	Professions intermédiaires
5	Employés
6	Ouvriers
7	Retraités
8	Autres personnes sans activité professionnelle

Source : Insee, PCS-2003 (niveau 1 de la nomenclature)

Dans cet exemple, il n'y a pas d'ordre naturel entre les huit catégories, ou modalités, qui sont de simples étiquettes ; la variable qualitative « CSP » est définie sur une échelle nominale.

### Variables qualitatives ordinaires

Une échelle ordinaire suppose l'existence d'une relation d'ordre total entre les catégories, c'est-à-dire que l'on peut opérer un classement de l'ensemble des catégories, de la plus petite à la plus grande (ou, inversement, de la plus grande à la plus petite).

Contrairement à ce qui se passe avec une échelle nominale, les expressions telles que « plus grand que », « précède », « se place après », etc. prennent un sens dans une échelle ordinaire.

La codification peut être numérique, alphabétique ou alphanumérique, en association avec un sens de lecture. En cas de codage numérique, les opérations mathématiques sont dénuées de sens et l'écart entre les valeurs ne revêt aucune signification.

**Définition**

Une variable statistique qualitative est dite définie sur une échelle **ordinale** si l'ensemble de ses catégories peut être doté d'une relation d'ordre.

## 1.4 LES VARIABLES QUANTITATIVES

Toute variable qui n'est pas qualitative ne peut être que quantitative. Les différentes modalités d'une variable quantitative constituent l'ensemble des valeurs numériques que peut prendre la variable.

### Définition

Une variable statistique est dite de nature **quantitative** si ses modalités sont mesurables. Les modalités d'une variable quantitative sont des nombres liés à l'unité choisie, qui doit toujours être précisée.

Il existe deux types de variables quantitatives : les variables discrètes et les variables continues.

Ces variables ont en commun des modalités clairement ordonnées, pour lesquelles l'écart entre les valeurs possède une signification, et sur lesquelles il est possible de réaliser des opérations mathématiques telles que des calculs de moyennes, etc. Néanmoins, elles ont des propriétés et des traitements spécifiques qui nécessitent une étude séparée.

### Variables quantitatives discrètes

Lorsque les modalités sont des valeurs numériques isolées, comme le nombre d'enfants par ménage, on parle de variable discrète<sup>1</sup>.

### Définition

Une variable statistique quantitative est dite **discrète** si l'ensemble de ses modalités est un ensemble fini ou dénombrable. Ainsi, l'ensemble des modalités peut être donné sous la forme d'une liste de nombres,  $M = \{x_1 ; x_2 ; \dots ; x_i ; \dots\}$ , finie ou infinie.

Le plus souvent, les modalités appartiennent à l'ensemble  $N$  des entiers naturels ( $N = \{0 ; 1 ; 2 ; \dots\}$ ). Cependant, une variable discrète peut prendre des valeurs non entières.

### Variables quantitatives continues

Lorsque la variable, par exemple la taille d'un individu, peut prendre toutes les valeurs d'un intervalle, ces valeurs peuvent alors être regroupées en classes, et on parle dans ce cas de variable continue.

### Définitions

Une variable statistique quantitative est dite **continue** si l'ensemble de ses modalités n'est pas dénombrable. Ainsi, une variable continue peut prendre toutes les valeurs d'un intervalle.

Pour étudier une variable statistique continue, on définit des classes ou intervalles de valeurs possibles. On peut ainsi **discretiser** une variable continue (voir section 2.1). Les classes retenues constituent les modalités de la variable.

On appelle **amplitude de la classe**  $[a_i ; b_i]$  le réel noté  $A_i$  représentant la longueur de l'intervalle et défini par :  $A_i = b_i - a_i$ .  $a_i$  et  $b_i$  sont respectivement les bornes inférieure et supérieure de la classe  $n_i$ .

Le **centre de classe** de la classe  $[a_i ; b_i]$  est le réel noté  $x_i$  représentant le milieu de l'intervalle et donné par :  $x_i = (a_i + b_i) / 2$ ; c'est la moyenne arithmétique des bornes de la classe.

1. Du latin *discretus*, qui signifie « séparé » ; dans un ensemble discret, on peut séparer les éléments.

Le centre de classe est appelé à jouer un grand rôle dans les calculs, car le regroupement en classes constitue une perte d'information importante ; nous prendrons l'hypothèse de répartition uniforme à l'intérieur d'une classe, c'est-à-dire de concentration au centre des classes (voir chapitre 2).

### Exemple 1.4

#### Calculs d'amplitudes et centres de classes

Le tableau suivant indique la structure par âges de la population féminine en France métropolitaine :

Âge	$f_i (\%)$
Moins de 15 ans	17,5
15-24 ans	12,3
25-34 ans	12,7
35-44 ans	14,0
45-54 ans	13,6
55-64 ans	11,1
65-74 ans	8,6
75 ans ou +	9,1

Source : Insee, bilan démographique, 2006

Les modalités sont des intervalles qui, par convention, sont – à part pour la dernière classe – fermés à gauche et ouverts à droite. Ainsi, la première classe se note aussi : [0 ; 15[, la deuxième [15 ; 25[, etc.

Les classes ne sont pas de même amplitude, la première classe ayant une amplitude de 15 ans et les suivantes de 10 ans. Pour la dernière classe, dont l'amplitude n'est pas définie explicitement, la convention suivante est adoptée : en l'absence d'information, il lui est attribué l'amplitude de la classe précédente, [65 ; 75[, donc 10 ans, et elle est donc écrite : [75 ; 85[.

Le centre de la première classe est :  $x_1 = (a_1 + b_1) / 2 = (0 + 15) / 2 = 7,5$  ans.

Cette distinction entre variable discrète et variable continue est parfois arbitraire, toute mesure étant discrète du fait de la précision limitée des instruments de mesure ou des arrondis. Cependant, la taille d'un individu, par exemple, est une variable continue du fait que, indépendamment de la mesure, toute valeur de l'intervalle [140 ; 150[ peut représenter en centimètres la taille d'un individu. De même, il arrive qu'une variable discrète, comme le nombre d'habitants d'un pays, qui peut prendre un grand nombre de valeurs dans un intervalle soit considérée comme une variable continue.

En conclusion, toute étude de variable statistique devra être précédée d'une identification claire de la population, du caractère étudié et de sa nature, à savoir qualitatif ou quantitatif et, dans le cas quantitatif, discret ou continu.

## 2 Présentation des données

Les données statistiques sont issues de données brutes présentées sous forme de tableaux statistiques dans lesquels sont indiqués les effectifs et/ou les fréquences.

### 2.1 DISTRIBUTION DES EFFECTIFS OU DES FRÉQUENCES

Les tableaux statistiques contenant les effectifs et/ou les fréquences sont une première exploitation des données brutes.

#### Des données brutes au tableau statistique

Il est primordial de définir la population et de préciser avec rigueur la ou les variables relevées sur chacun des individus de la population ou de l'échantillon la représentant. Ensuite, quand les observations ont été recueillies, le premier travail consiste à les présenter, aussi clairement que possible, sous forme de tableau statistique. Ce tableau révèle la distribution statistique en présentant les couples de type  $(x_i ; n_i)$ , où les  $x_i$  sont les modalités et les  $n_i$  leurs effectifs respectifs,  $i$  entier variant de 1 à  $r$ , si  $r$  désigne le nombre de modalités du caractère. Il est également possible de présenter la distribution des fréquences, c'est-à-dire les couples de type  $(x_i ; f_i)$ .

#### Définitions

On appelle **données brutes** ou tableau élémentaire le tableau relevant pour chaque unité statistique la modalité de la variable étudiée.

Le **tri à plat** est la transformation qui permet de passer du tableau des données brutes au tableau de la distribution statistique présentant les modalités et les effectifs, les modalités étant classées par ordre croissant.

#### Discrétisation

Dans le cas d'une variable statistique quantitative continue, il est nécessaire de définir des classes pour pouvoir proposer un tri à plat.

#### Définition

On appelle **discrétisation** le découpage en classes d'une série statistique quantitative.

Ce découpage en classes pose de nombreuses questions : choix des amplitudes, amplitudes constantes ou variables, nombre de classes, etc. Nous ne rentrerons pas ici dans le détail de ces opérations (voir l'exercice 4 de ce chapitre).

### 2.2 VARIABLES QUANTITATIVES : DISTRIBUTION DES EFFECTIFS ET DES FRÉQUENCES CUMULÉS

Cette section concerne les variables quantitatives pour lesquelles le tableau statistique est réalisé, les modalités étant ordonnées dans l'ordre croissant. Les notions que nous allons définir sont liées à la notion de fonction de répartition, fondamentale en probabilité pour les variables aléatoires continues et sur laquelle nous reviendrons dans la section 3.3.

Reprendons l'exemple 1.4 et proposons de répondre à la question suivante : quelle proportion de la population féminine en France métropolitaine a moins de 35 ans ?

Nous pouvons affirmer que 42,5 % de la population féminine en France métropolitaine a moins de 35 ans, soit 17,5 % + 12,3 % + 12,7 %. Pour obtenir ce résultat, nous avons cumulé les fréquences des modalités inférieures ou égales à 34 ans.

## Définitions

**Effectifs cumulés croissants sur variable discrète :** Si  $X$  désigne une variable quantitative discrète, on appelle effectif cumulé croissant, noté  $n_{cc}$ , le nombre d'individus statistiques pour lesquels  $X$  est inférieur ou égal à  $x_i$ .

On a :  $n_{cc} = n_1$  et  $n_{cc} = n_1 + n_2 + \dots + n_i = \sum_{k=1}^i n_k$ .

Si la série possède  $r$  modalités,  $x_r$  désignant alors la plus grande valeur de  $X$ , on a :  $n_{cc} = n_1 + n_2 + \dots + n_r = \sum_{k=1}^r n_k = n$ , où  $n$  désigne l'effectif total de la série.

**Fréquences cumulées croissantes sur variable discrète :** Avec les mêmes hypothèses, on définit la fréquence cumulée croissante, notée  $f_{cc}$ , représentant la proportion d'individus statistiques pour lesquels  $X$  est inférieur ou égal à  $x_i$ .

On a :  $f_{cc} = f_1$  et  $f_{cc} = f_1 + f_2 + \dots + f_i = \sum_{k=1}^i f_k$ , ou encore  $f_{cc} = \frac{n_{cc}}{n}$ .

Si la série possède  $r$  modalités,  $x_r$  désignant alors la plus grande valeur de  $X$ , on a :  $f_{cc} = f_1 + f_2 + \dots + f_r = \sum_{k=1}^r f_k = 1$  (ou 100 si les fréquences sont exprimées en pourcentage).

Dans le cas d'une **variable quantitative continue**, les données sont groupées en classes  $[a_i ; b_i]$ , et on définit, de même que pour une variable discrète,  $n_{cc}$  le nombre d'individus statistiques pour lesquels  $X$  est inférieur ou égal à  $b_i$ , et  $f_{cc}$  la proportion d'individus statistiques pour lesquels  $X$  est inférieur ou égal à  $b_i$ .

Il est également possible de cumuler les effectifs et les fréquences dans le sens décroissant.

## Définitions

**Effectifs cumulés décroissants sur variable discrète :** Si  $X$  désigne une variable quantitative discrète, on appelle effectif cumulé décroissant, noté  $n_{cd}$ , le nombre d'individus statistiques pour lesquels  $X$  est supérieur ou égal à  $x_i$ .

(Certains auteurs adoptent une convention différente : le nombre d'individus statistiques pour lesquels  $X$  est strictement supérieur à  $x_i$ ).

On a :  $n_{cd} = n$  ;  $n_{cd} = n_i + n_{i+1} + \dots + n_r = \sum_{k=i}^r n_k$ ,  $r$  désignant le nombre de modalités, et  $n_{cd} = n_r$ .

**Fréquences cumulées décroissantes sur variable discrète :** Avec les mêmes hypothèses, on définit la fréquence cumulée décroissante, notée  $f_{cd}$ , représentant la proportion d'individus statistiques pour lesquels  $X$  est supérieur ou égal à  $x_i$ .

On a :  $f_{cd} = 1$  ;  $f_{cd} = f_i + f_{i+1} + \dots + f_r = \sum_{k=i}^r f_k$ , et  $f_{cd} = f_r$ , ou encore  $f_{cd} = \frac{n_{cd}}{n}$ .

Dans le cas d'une **variable quantitative continue**, les données sont groupées en classes  $[a_i ; b_i]$ , et on définit, de même que pour une variable discrète,  $n_{cd}$  le nombre d'individus statistiques pour lesquels  $X$  est supérieur ou égal à  $a_i$ , et  $f_{cc}$  la proportion d'individus statistiques pour lesquels  $X$  est supérieur ou égal à  $a_i$ .

### Exemple 1.5

#### Calculs d'effectifs et fréquences cumulés croissants et décroissants

Le tableau suivant recense les enfants de moins de 6 ans en France métropolitaine :

Année	Moins de 3 ans	De 3 à 5 ans
2006	2 294 846	2 317 874

Source : Insee, bilan démographique, 2006

Les effectifs cumulés croissants ( $n_{i,cc}$ ), décroissants ( $n_{i,cd}$ ), et les fréquences cumulées croissantes ( $f_{i,cc}$ ), décroissantes ( $f_{i,cd}$ ), correspondants sont les suivants :

Âge	$n_i$	$n_{i,cc}$	$n_{i,cd}$	$f_{i,cc}$	$f_i$	$f_{i,cd}$
[0 ; 3[	2 294 846	2 294 846	4 612 720	0,4975	0,4975	1
[3 ; 6[	2 317 874	4 612 720	2 317 874	1	0,5025	0,5025
Total	4 612 720				1,0000	

## 3

# Représentations graphiques des séries à une variable

L'apparition des graphiques statistiques, liée à l'utilisation des coordonnées, doit essentiellement son origine au philosophe et mathématicien René Descartes (1596-1650). Ces graphiques constituent une synthèse visuelle indispensable de l'information contenue dans le tableau statistique.

Les graphiques utilisés dépendent de la nature de la variable. Nous utiliserons, pour représenter les distributions d'effectifs (ou de fréquences), les diagrammes circulaires (ou secteurs), les diagrammes en tuyaux d'orgue, les diagrammes en bâtons, les histogrammes et le polygone des effectifs. Pour les distributions cumulées, nous utiliserons les polygones des effectifs (ou des fréquences) cumulés croissants et décroissants.

## 3.1 GRAPHIQUES POUR VARIABLES QUALITATIVES

Les variables qualitatives – nominales ou ordinaires – peuvent être représentées au choix à l'aide d'un diagramme circulaire ou à l'aide d'un diagramme en tuyaux d'orgue.

### Diagramme circulaire

Le diagramme circulaire, également appelé « camembert », permet une représentation de la distribution d'une variable dans un cercle qui représente 100 % des modalités (voir figure 1.1).

**Définition**

Un **diagramme circulaire** est un graphique constitué d'un cercle divisé en secteurs dont les angles au centre sont proportionnels aux effectifs (ou aux fréquences). De fait, les aires des secteurs sont proportionnelles aux effectifs. L'angle  $\alpha_i$  d'une modalité d'effectif  $n_i$  est donné en degrés par :  $\alpha_i = \frac{n_i}{n} \times 360 = f_i \times 360$ .

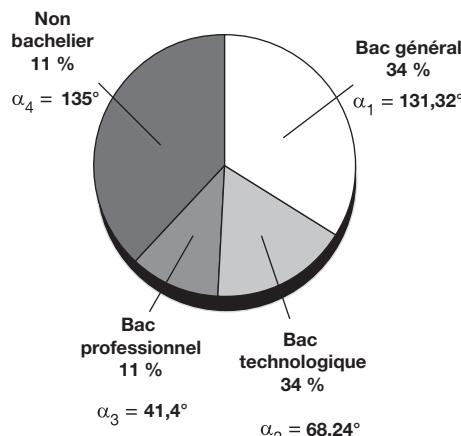
Il est également possible d'utiliser un graphique semi-circulaire formé d'un demi-cercle ( $180^\circ$ ).

**Diagramme en tuyaux d'orgue (en barres)**

Le diagramme en tuyaux d'orgue est une représentation de la distribution d'une variable selon des rectangles horizontaux ou verticaux ayant tous une même base, de largeur arbitraire (voir figure 1.2).

**Figure 1.1**

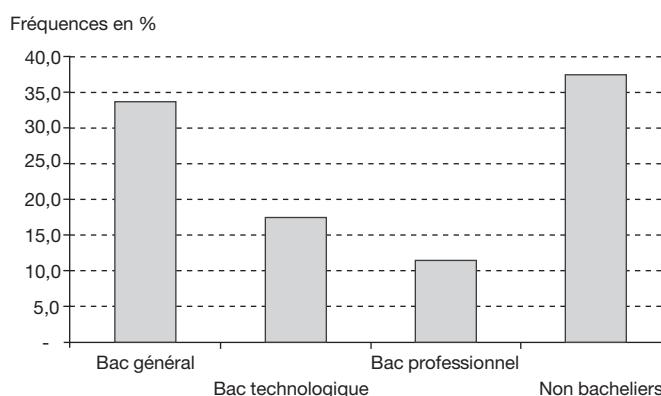
**Diagramme circulaire :**  
proportion (en pourcentage) de bacheliers et non-bacheliers dans une génération en France métropolitaine et DOM, 2005.

**Définition**

Un **diagramme en tuyaux d'orgue** est un graphique qui à chaque modalité d'une variable qualitative associe un rectangle de base constante dont la hauteur est proportionnelle à l'effectif (ou à la fréquence). De fait, les aires des secteurs sont proportionnelles aux effectifs. Les rectangles sont en général disjoints, verticaux ou horizontaux.

**Figure 1.2**

**Diagramme en tuyaux d'orgue :**  
proportion (en pourcentage) de bacheliers et non-bacheliers dans une génération en France métropolitaine et DOM, 2005.



## 3.2 GRAPHIQUES POUR VARIABLES QUANTITATIVES

La représentation graphique d'une variable quantitative dépend de sa nature : discrète ou continue.

### Variables discrètes : diagramme en bâtons

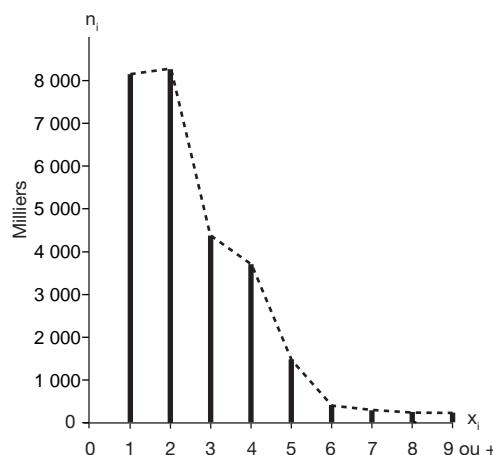
La distribution d'une variable quantitative discrète peut être représentée par un diagramme en bâtons (voir figure 1.3).

#### Définition

On appelle **diagramme en bâtons** un graphique qui à chaque modalité d'une variable quantitative discrète associe un segment (bâton) dont la hauteur est proportionnelle à l'effectif (ou à la fréquence).

Figure 1.3

Diagramme en bâtons et polygone des effectifs : nombre de personnes par ménage, France, 1999.



### Variables continues : histogramme

En 2005 Monaco avait 32 543 habitants et le Japon 127 417 244 (source : Institut national d'études démographiques). Bien sûr, les démographes diront que ces renseignements sont très largement insuffisants pour comparer la démographie des deux pays : il faut au minimum s'intéresser aux superficies de ces deux pays et calculer pour chacun d'entre eux la densité de population, c'est-à-dire le nombre d'habitants au kilomètre carré. Avec une superficie de 2,02 km<sup>2</sup> pour Monaco et de 378 000 km<sup>2</sup> pour le Japon, les densités sont respectivement  $d_1 = 32\ 543 / 2,02 = 16\ 110,40$  h/km<sup>2</sup> pour Monaco et  $d_2 = 127\ 417\ 244 / 378\ 000 = 337$  h/km<sup>2</sup> pour le Japon. Autrement dit, alors que la population de Monaco est la moins importante en taille, sa densité de population est plus importante que celle du Japon.

Cette notion de densité est essentielle pour les variables continues : il est absurde de comparer ou de représenter côté à côté des classes qui n'ont pas la même amplitude sans faire intervenir la densité. Ce principe est omniprésent lors de la réalisation d'un histogramme.

#### Définitions

Un **histogramme** est un diagramme composé de rectangles contigus dont les aires sont proportionnelles aux effectifs (ou aux fréquences) et dont les bases sont déterminées par les intervalles de classes.

Dans le cas d'une variable quantitative continue, on définit la **densité** d'effectif  $d_i$  d'une classe d'effectif  $n_i$  et d'amplitude  $a_i$  par :  $d_i = n_i / a_i$  (ou, dans le cas des fréquences,  $f_i / a_i$ ).

Lors de la réalisation d'un histogramme, il est indispensable de distinguer deux cas.

1. Si les amplitudes de classes sont égales, la hauteur des rectangles correspondra aux effectifs (ou aux fréquences) des classes.
2. Si les amplitudes sont différentes, afin de constituer l'histogramme, il est nécessaire de :
  - calculer, pour chaque classe, l'amplitude  $a_i$  ;
  - calculer la densité  $d_i = n_i / a_i$  pour un histogramme des effectifs, et  $d_i = f_i / a_i$  pour un histogramme des fréquences ;
  - affecter à chaque rectangle une hauteur proportionnelle à la densité  $d_i$  de la classe correspondante.

Soit  $\min(a_i)$  l'amplitude minimale de classe, la hauteur est alors appelée « effectif corrigé » et notée  $n_i c = d_i \times \min(a_i)$  ; cette convention revient à adopter  $\min(a_i)$  comme unité d'amplitude de classe. Les classes ayant pour amplitudes  $\min(a_i)$  sont alors représentées par des rectangles dont la hauteur est l'effectif. De même, il est possible de retenir comme hauteur la fréquence corrigée  $f_i c = d_i \times \min(a_i)$ , avec  $d_i = f_i / a_i$  dans le cas d'un histogramme des fréquences. L'utilisation de  $\min(a_i)$  est une convention facultative ; un histogramme est correct dès lors que les effectifs (ou les fréquences) corrigés sont proportionnels aux densités.

### Exemple 1.6

### Réalisation d'un histogramme et d'un polygone des effectifs

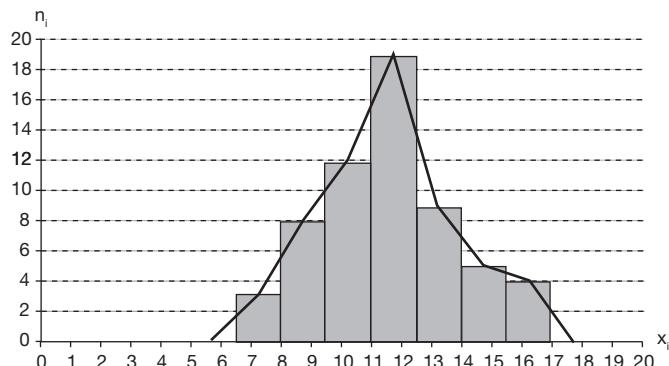
Le responsable des ressources humaines d'une entreprise a relevé la distribution statistique suivante correspondant à l'ancienneté du personnel cadre dans l'entreprise, exprimée en années :

Classes	Effectifs
[6,5 ; 8[	3
[8 ; 9,5[	8
[9,5 ; 11[	12
[11 ; 12,5[	19
[12,5 ; 14[	9
[14 ; 15,5[	5
[15,5 ; 17[	4
<b>Total</b>	<b>60</b>

L'histogramme des effectifs est présenté avec, sur le même graphique, le polygone des effectifs tracé en courbe pleine (voir figure 1.4). Ce polygone permet de représenter la distribution sous la forme d'une courbe ; quand les amplitudes de classes sont égales, on l'obtient en joignant les milieux des bases supérieures de chaque rectangle de l'histogramme par des segments de droite. On adjoint généralement une classe d'effectif nul, de part et d'autre de l'histogramme, afin de respecter la règle de compensation des aires : l'aire totale du domaine situé entre l'axe des x et le polygone est égale à la somme des aires des rectangles de l'histogramme. Elle représente l'effectif total.

**Figure 1.4**

**Histogramme et polygone des effectifs, classes de même amplitude : ancienneté du personnel cadre de l'entreprise.**



Modifions légèrement cet exemple en regroupant les deux dernières classes en une seule. Ce regroupement permet de traiter le cas de classes d'amplitudes différentes, puisque ainsi la dernière classe est d'amplitude 3 contre 1,5 pour toutes les autres classes.

Classes	Effectifs
[6,5 ; 8[	3
[8 ; 9,5[	8
[9,5 ; 11[	12
[11 ; 12,5[	19
[12,5 ; 14[	9
[14 ; 17[	9
<b>Total</b>	<b>60</b>

Les classes étant d'amplitudes inégales, il est nécessaire de calculer les amplitudes ( $a_i$ ), les densités ( $d_i$ ) puis les effectifs corrigés ( $n_{i,c}$ ) pour chaque classe. Les résultats de ces calculs sont présentés dans la figure 1.5.

**Figure 1.5**

**Calcul des effectifs corrigés dans le cas de classes d'amplitudes inégales.**

	A	B	C	D	E
1 Classes	$n_i$	$a_i$	$d_i$	$n_{i,c}$	
2 [6,5 ; 8[	3	1,50	2,00	3,0	
3 [8 ; 9,5[	8	1,50	5,33	8,0	
4 [9,5 ; 11[	12	1,50	8,00	12,0	
5 [11 ; 12,5[	19	1,50	12,67	19,0	
6 [12,5 ; 14[	9	1,50	6,00	9,0	
7 [14 ; 17[	9	3,00	3,00	4,5	
8 Total	60				

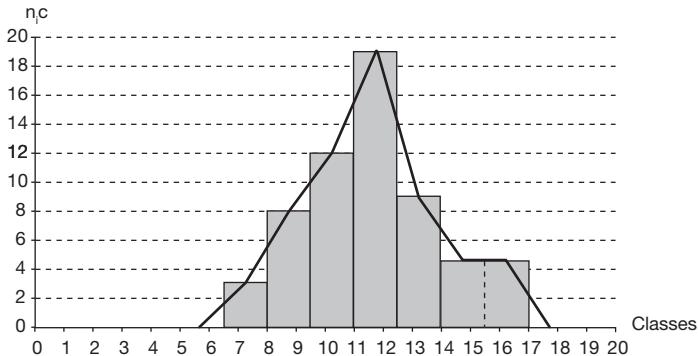
On peut alors tracer l'histogramme de la figure 1.6 à partir des effectifs corrigés, ainsi que le polygone des effectifs, en trait continu.

Pour tracer le polygone des effectifs, nous avons effectué un découpage artificiel en pseudo-classes d'amplitude 1,5, dont nous avons pris les milieux des bases supérieures de façon à respecter la règle de compensation des aires : les aires des triangles extérieurs au domaine délimité par le polygone sont égales à celles des triangles qui sont situés sous le polygone. Ainsi, l'aire totale du domaine situé sous le polygone des effectifs est égale à l'aire totale des rectangles de l'histogramme.

Ce qui est fait dans cet exemple à partir des effectifs peut également être réalisé à partir des fréquences, afin de tracer l'histogramme et le polygone des fréquences.

**Figure 1.6**

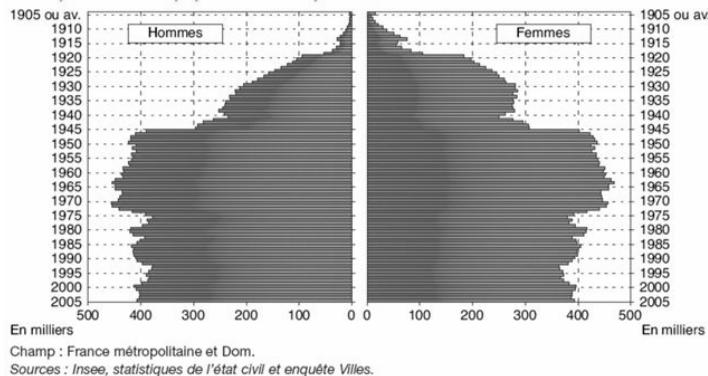
**Histogramme et polygone des effectifs : classes d'amplitudes inégales.**



Enfin, il serait inconcevable de ne pas évoquer une variété d'histogramme, la pyramide, dont l'exemple le plus célèbre est la pyramide des âges (voir figure 1.7). Cette variété d'histogramme, où les axes ont été modifiés (classes en ordonnées et effectifs en abscisses), est largement utilisée en démographie. Les classes sont annuelles. Les aires des rectangles représentent le nombre d'hommes ou de femmes vivants et nés l'année considérée, en lecture sur l'axe des abscisses.

**Figure 1.7**

**Pyramide des âges.**



### 3.3 DIAGRAMMES CUMULATIFS

Les notions d'effectifs et de fréquences cumulés nous ont donné l'occasion d'introduire la notion de fonction de répartition, que nous définissons ci-après avant d'évoquer sa représentation graphique.

#### Définition

Si  $X$  est une variable quantitative, on introduit la **fonction de répartition**, qui à tout nombre réel  $x$  associe la proportion des individus de la population pour lesquels  $X$  est inférieur ou égal à  $x$ .

Pour tout  $x$  réel,  $0 \leq F(x) \leq 1$  (les valeurs de  $F$  peuvent également être exprimées en pourcentage).

La première étape de la construction d'une fonction de répartition consiste donc à calculer les fréquences cumulées croissantes, en distinguant deux cas : le discret et le continu.

### Fonction de répartition d'une variable discrète

La fonction de répartition d'une variable quantitative discrète est une fonction en escalier, c'est-à-dire constante par intervalle. De plus, elle est croissante de 0 à 1 et définie par :

- Si  $x < x_i$ ,  $F(x) = 0$
- Si  $x = x_i$ ,  $F(x) = f_{i,cc}$
- Si  $x_i \leq x < x_{i+1}$ ,  $F(x) = f_{i,cc}$
- Si  $x \geq x_r$ ,  $F(x) = 1$

### Fonction de répartition d'une variable continue

*A priori*, la fonction de répartition d'une variable continue n'est connue que pour les extrémités de classes. Cependant, si l'on admet l'hypothèse de répartition uniforme des observations au sein de chaque classe, on peut estimer les valeurs de  $F(x)$  par interpolation linéaire. Cela revient à approximer la représentation graphique par une fonction affine par morceaux : concrètement, on trace la courbe en joignant deux points consécutifs connus par un segment de droite (cette courbe est aussi appelée ogive de Galton<sup>1</sup>).

Avec cette hypothèse,  $F(x)$  représente l'aire située sous l'histogramme des fréquences, à gauche de la valeur  $x$ .

### Polygones des effectifs cumulés croissants et décroissants

Dans le cas d'une variable continue, on définit les polygones des effectifs (ou des fréquences) cumulés croissants et décroissants ; ils seront utilisés notamment pour déterminer la médiane de la série (voir chapitre 2).

Le polygone des fréquences cumulées croissantes commence au point de coordonnées  $(a_1 ; 0)$ , car la proportion de valeurs inférieures à  $a_1$  est nulle. Il est obtenu en joignant les points de coordonnées  $(b_i ; f_{i,cc})$  – il correspond à la restriction de la fonction de répartition aux valeurs de  $x$  inférieures ou égales à la borne supérieure de la dernière classe.

Le polygone des fréquences cumulées décroissantes s'obtient de la même façon, en adjointant le point de coordonnées  $(b_r ; 0)$ , car,  $b_r$  désignant la borne supérieure de la dernière classe, la proportion de valeurs supérieures à  $b_r$  est nulle.

#### Exemple 1.7

#### Réalisation des polygones des fréquences cumulées croissantes et décroissantes

Le tableau suivant donne la structure de la population chinoise suivant l'âge :

0-14 ans	15-24 ans	25-59 ans	60 ans et plus
21,4 %	16,6 %	51,1 %	10,9 %

Source : ONU, 2005

1. Francis Galton (1822-1911) fut l'un des fondateurs de la biométrie et collabora avec son ami Karl Pearson (1857-1936).

Les fréquences cumulées croissantes et décroissantes sont calculées puis organisées pour correspondre aux bornes des classes (voir figure 1.8).

**Figure 1.8**

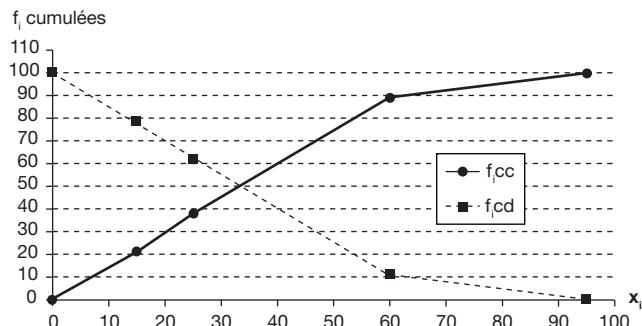
**Plages de données des polygones des fréquences cumulées croissantes et décroissantes.**

	A	B	C	D	E
1	Classes	$f_i$	Borne de classe	$f_{i,cc}$	$f_{i,cd}$
2			0	0,00	100,00
3	[0 ; 15[	21,40	15	21,40	78,60
4	[15 ; 25[	16,60	25	38,00	62,00
5	[25 ; 60[	51,10	60	89,10	10,90
6	60 et +	10,90	95	100,00	0,00

À partir de ces données, il est possible de tracer les polygones des fréquences cumulées croissantes et décroissantes (voir figure 1.9).

**Figure 1.9**

**Polygones des fréquences cumulées croissantes et décroissantes de l'âge de la population chinoise.**



## Conclusion

On retiendra de ce premier chapitre l'importance de la terminologie. On devra savoir identifier, dans un exercice, la population, les variables étudiées et leur nature : qualitative, quantitative discrète ou quantitative continue. On notera que le discret et le continu, en statistique comme en probabilité, nécessitent des traitements différents ; dans le cas continu, on retiendra l'importance de la notion de densité. Par ailleurs, on n'insistera jamais assez sur l'importance des représentations graphiques en statistique ; à l'issue de ce chapitre, on devra maîtriser notamment les histogrammes et les polygones des effectifs (ou des fréquences) cumulés croissants et décroissants.

# Problèmes et exercices

Les problèmes et exercices suivants proposent la mise en application des notions exposées dans la première partie de ce chapitre.

- L'exercice 1 traite du passage d'une série brute à un tableau statistique.
- Les exercices 2, 3 et 5 s'attachent aux graphiques associés aux différentes natures de variables statistiques.
- L'exercice 4 s'intéresse à la discréétisation des données.



## EXERCICE 1 DE LA SÉRIE BRUTE À LA PRÉSENTATION DES STATISTIQUES

### Énoncé

La liste suivante est composée de prénoms d'un groupe d'étudiants, suivis entre parenthèses du nombre de films que chacun d'entre eux a vus au cours du mois dernier :

Pierre (3), Paul (2), Jacques (2), Ralph (3), Abdel (1), Sidonie (2), Henri (0), Paulette (1), Farida (2), Laure (2), Kevin (0), Carole (3), Marie-Claire (0), Jeanine (3), Julie (2), Ernest (3), Cindy (3), Vanessa (2), José (1), Aurélien (1).

1. Déterminez :
  - a. la population étudiée ;
  - b. la variable étudiée.
2. Précisez :
  - a. la nature de la variable ;
  - b. les modalités de la variable.
3. Construisez le tableau statistique associé à la distribution des effectifs.
4. Représentez la distribution des effectifs par un diagramme en bâtons.
5. Calculez les effectifs :
  - a. cumulés croissants ;
  - b. cumulés décroissants.
6. Calculez les fréquences :
  - a. cumulées croissantes ;
  - b. cumulées décroissantes.

**Solution**

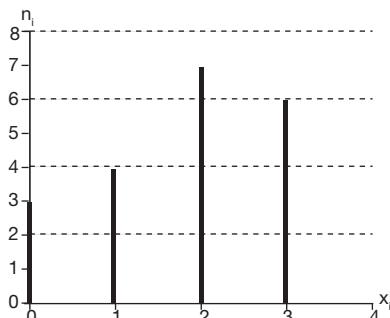
- 1. a.** La population étudiée est le groupe d'étudiants.  
**b.** La variable étudiée est  $X = \text{« nombre de films que chacun d'entre eux a vus au cours du mois dernier »}$ .
- 2. a.** La variable étudiée est quantitative discrète.  
**b.** L'ensemble  $M$  des modalités est  $M = \{0 ; 1 ; 2 ; 3\}$ .  
**3.** Le tableau statistique associé est composé de deux colonnes :  
  - la première colonne comporte les modalités  $x_i$  de  $X$  ;
  - la seconde colonne comporte les effectifs  $n_i$  associés à chacune de ces modalités.
Le tableau statistique associé à  $X$  est le suivant.

<b><math>x_i</math></b>	<b><math>n_i</math></b>
0	3
1	4
2	7
3	6

L'effectif total est  $n = \sum_{i=1}^4 n_i$ , soit  $n = 20$ .

**4.****Figure 1.10**

**Diagramme en bâtons des effectifs.**



Le même diagramme en bâtons peut être réalisé sous Excel. Pour cela, cliquez sur Insertion/Graphique dans la barre de menus d'Excel.

L'assistant graphique apparaît. Dans l'assistant graphique, choisissez le type de graphique Histogramme et cliquez sur Suivant. Notez que le mot « histogramme » est employé par Excel comme un terme générique désignant des barres verticales et non un histogramme au sens statistique.

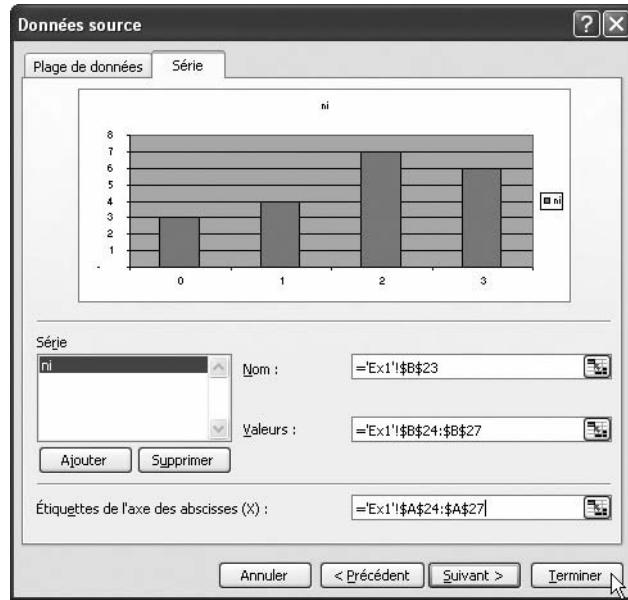
L'assistant graphique propose de saisir les données du graphique. Cliquez sur l'onglet Série et indiquez dans les champs correspondants les plages où se trouvent les données. Pour cela, sélectionnez-les à l'aide de la souris, comme indiqué sur la figure 1.11 :

- la cellule B23 de la feuille Ex1 pour le nom ;

- la plage B24:B27 de la feuille Ex1 pour les valeurs ;
  - la plage A24:A27 de la feuille Ex1 pour les graduations de l'axe des abscisses.
- Cliquez sur le bouton Terminer.

**Figure 1.11**

**Sélection des données à représenter dans l'assistant graphique.**



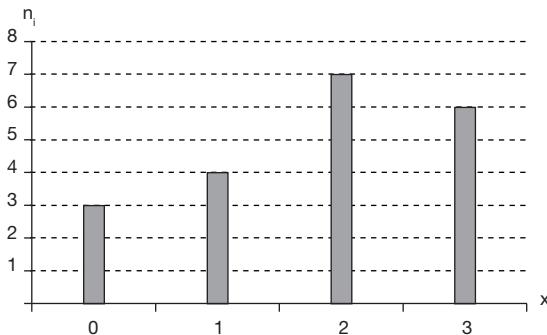
L'assistant graphique se ferme et le graphique apparaît (voir figure 1.12). Vous pouvez modifier les options d'affichage du graphique en appelant un menu par un clic droit sur la zone de graphique.

**a.** Soit  $n_{i,cc}$  l'effectif cumulé croissant de la modalité  $i$ :  $n_{1,cc} = n_1 = 3$ , soit  $n_{1,cc} = 3$ ;  $n_{2,cc} = n_{1,cc} + n_2 = 3 + 4$ , soit  $n_{2,cc} = 7$ ;  $n_{3,cc} = n_{2,cc} + n_3 = 7 + 7$ , soit  $n_{3,cc} = 14$ ;  $n_{4,cc} = n_{3,cc} + n_4 = 14 + 6$ , soit  $n_{4,cc} = 20$ .

**b.** Soit  $n_{i,cd}$  l'effectif cumulé décroissant de la modalité  $i$ :  $n_{1,cd} = n$ , soit  $n_{1,cd} = 20$ ;  $n_{2,cd} = n_{1,cd} - n_1 = 20 - 3$ , soit  $n_{2,cd} = 17$ ;  $n_{3,cd} = n_{2,cd} - n_2 = 17 - 4$ , soit  $n_{3,cd} = 13$ ;  $n_{4,cd} = n_{3,cd} - n_3 = 13 - 7$ , soit  $n_{4,cd} = 6$ .

Les résultats des effectifs cumulés croissants et décroissants se présentent dans un tableau obtenu en ajoutant deux colonnes au tableau statistique initial : les effectifs cumulés croissants  $n_{i,cc}$  et les effectifs cumulés décroissants  $n_{i,cd}$ .

<b><math>x_i</math></b>	<b><math>n_i</math></b>	<b><math>n_{i,cc}</math></b>	<b><math>n_{i,cd}</math></b>
0	3	3	20
1	4	7	17
2	7	14	13
3	6	20	6

**Figure 1.12****Diagramme en bâtons sous Excel.**

**6.** Pour pouvoir calculer les fréquences cumulées croissantes  $f_{iCC}$  et décroissantes  $f_{iCD}$ , il convient de calculer les fréquences  $f_i$ .

Soit  $f_i$  la fréquence de la classe  $i$ :  $f_1 = \frac{n_1}{n} = \frac{3}{20}$ , soit  $f_1 = 0,15$ ;  $f_2 = \frac{n_2}{n} = \frac{4}{20}$ , soit  $f_2 = 0,20$ ;

$f_3 = \frac{n_3}{n} = \frac{7}{20}$ , soit  $f_3 = 0,35$ ;  $f_4 = \frac{n_4}{n} = \frac{6}{20}$ , soit  $f_4 = 0,30$ .

**a.** Soit  $f_{iCC}$  la fréquence cumulée croissante de la classe  $i$ :  $f_{iCC} = f_i = 0,15$ , soit  $f_{iCC} = 0,15$ ;  $n_{iCC} = f_{iCC} + f_1 = 0,15 + 0,20$ , soit  $f_{iCC} = 0,35$ ;  $f_{iCC} = f_{iCC} + f_3 = 0,35 + 0,35$ , soit  $f_{iCC} = 0,70$ ;  $f_{iCC} = f_{iCC} + f_4 = 0,70 + 0,30$ , soit  $f_{iCC} = 1$ .

**b.** Soit  $f_{iCD}$  la fréquence cumulée décroissante de la classe  $i$ :  $f_{iCD} = 1$ , soit  $f_{iCD} = 1$ ;  $f_{iCD} = f_{iCD} - f_i = 1 - 0,15$ , soit  $f_{iCD} = 0,85$ ;  $f_{iCD} = f_{iCD} - f_2 = 0,85 - 0,20$ , soit  $f_{iCD} = 0,65$ ;  $f_{iCD} = f_{iCD} - f_3 = 0,65 - 0,35$ , soit  $f_{iCD} = 0,30$ .

Les résultats des fréquences cumulées croissantes et décroissantes se présentent dans un tableau obtenu en ajoutant deux colonnes au tableau statistique : les fréquences cumulées croissantes  $f_{iCC}$  et les fréquences cumulées décroissantes  $f_{iCD}$ .

x <sub>i</sub>	n <sub>i</sub>	n <sub>iCC</sub>	n <sub>iCD</sub>	f <sub>i</sub>	f <sub>iCC</sub>	f <sub>iCD</sub>
0	3	3	20	0,15	0,15	1,00
1	4	7	17	0,20	0,35	0,85
2	7	14	13	0,35	0,70	0,65
3	6	20	6	0,30	1,00	0,30



## EXERCICE 2 REPRÉSENTATIONS GRAPHIQUES SIMPLES

### Énoncé

Le tableau suivant indique la répartition des familles de l'île de La Réunion selon leur nombre d'enfants :

Nombre d'enfants	Nombre de familles
0	31 038
1	54 812
2	51 252
3	26 613
4 ou +	16 162

Source : Insee, recensement, 1999

1. Déterminez :
  - a. la population étudiée ;
  - b. la variable étudiée.
2. Précisez :
  - a. la nature de la variable ;
  - b. les modalités de la variable.
3. Représentez la distribution par diagramme circulaire.
4. À la suite de la question précédente :
  - a. Calculez les effectifs cumulés croissants et décroissants.
  - b. Représentez la fonction de répartition.
5. Combien de familles sont composées de :
  - a. au moins 1 enfant ?
  - b. au plus 2 enfants ?

### Solution

1. a. La population étudiée est composée des familles de La Réunion.  
b. La variable étudiée est  $X = \text{« nombre d'enfants »}$ .
2. a. La variable étudiée est quantitative discrète.  
b. L'ensemble des modalités de la variable étudiée est  $M = \{0 ; 1 ; 2 ; 3 ; 4 \text{ ou } +\}$ .
3. Pour réaliser un diagramme circulaire, il convient de tracer un cercle et de retenir pour chaque modalité  $i$  un secteur d'angle au centre :  $\alpha_i = 360 \times f_i$  exprimé en degrés.

Pour la modalité 1,  $f_1 = \frac{n_1}{n} = \frac{31038}{179877} = 0,1726$ , donc  $\alpha_1 = 360 \times 0,1726$ , soit  $\alpha_1 = 62,12^\circ$ .

Pour la modalité 2,  $f_2 = \frac{n_2}{n} = \frac{54812}{179877} = 0,3047$ , donc  $\alpha_2 = 360 \times 0,3047$ , soit  $\alpha_2 = 109,7^\circ$ .

Pour la modalité 3,  $f_3 = \frac{n_3}{n} = \frac{51252}{179877} = 0,2849$ , donc  $\alpha_3 = 360 \times 0,2849$ , soit  $\alpha_3 = 102,57^\circ$ .

Pour la modalité 4,  $f_4 = \frac{n_4}{n} = \frac{26613}{179877} = 0,1480$ , donc  $\alpha_4 = 360 \times 0,1480$ , soit  $\alpha_4 = 53,26^\circ$ .

Pour la modalité 5,  $f_5 = \frac{n_5}{n} = \frac{16162}{179877} = 0,0899$ , donc  $\alpha_5 = 360 \times 0,0899$ , soit  $\alpha_5 = 32,35^\circ$ .

On vérifie que la somme des angles est bien de  $360^\circ$ .

Ces calculs sont effectués sous Excel, dans le tableau présenté à la figure 1.13, colonnes C et D.

**Figure 1.13**

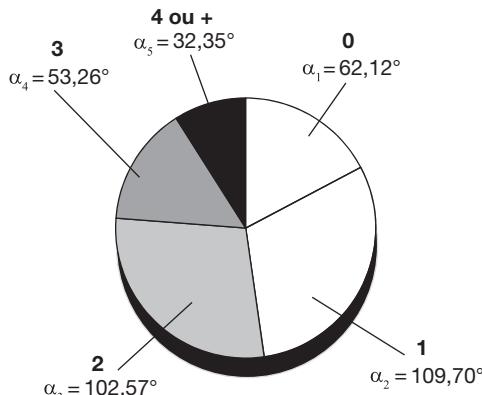
Résultats sous Excel.

	A	B	C	D	E	F	G
1	x <sub>i</sub>	n <sub>i</sub>	f <sub>i</sub>	α <sub>i</sub>	n <sub>cc</sub>	n <sub>cd</sub>	f <sub>cc</sub>
2	0	31 038	0,1726	62,12	31 038	179 877	0,1726
3	1	54 812	0,3047	109,70	85 850	148 839	0,4773
4	2	51 252	0,2849	102,57	137 102	94 027	0,7622
5	3	26 613	0,1480	53,26	163 715	42 775	0,9101
6	4 ou +	16 162	0,0899	32,35	179 877	16 162	1,0000

Le diagramme circulaire de la figure 1.14 est réalisé à partir de ces résultats.

**Figure 1.14**

Réalisation d'un diagramme circulaire : répartition des familles de La Réunion selon leur nombre d'enfants.

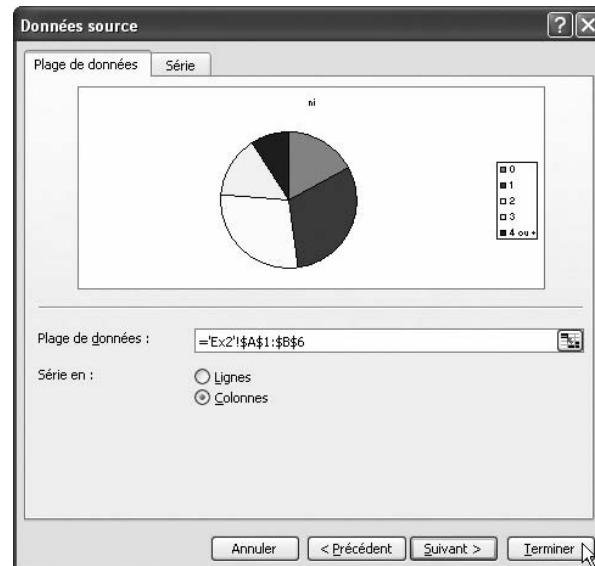


Pour réaliser un diagramme circulaire sous Excel, cliquez sur Insertion/Graphique dans la barre de menus.

L'assistant graphique apparaît. Dans l'assistant graphique, choisissez le type de graphique Secteurs et cliquez sur Suivant.

L'assistant graphique propose de saisir les données du graphique. Indiquez dans le champ Plage de données la plage où se trouvent les données en les sélectionnant à l'aide de la souris (voir figure 1.15). Il s'agit ici de la plage A2:B6 sur la feuille Ex2. Cliquez sur le bouton Terminer.

**Figure 1.15**  
**Sélection des données à représenter dans l'assistant graphique.**



L'assistant graphique se ferme et le graphique apparaît. Vous pouvez modifier les options d'affichage du diagramme en appelant un menu par un clic droit sur la zone de graphique.

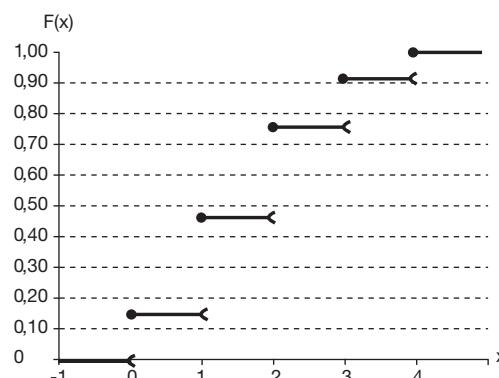
**4. a.** Soit  $n_{i,cc}$  l'effectif cumulé croissant de la classe  $i$  :  $n_{1,cc} = n_1 = 31\ 038$ , soit  $n_{1,cc} = 31\ 038$  ;  $n_{2,cc} = n_{1,cc} + n_2 = 31\ 038 + 54\ 812$ , soit  $n_{2,cc} = 85\ 850$  ;  $n_{3,cc} = n_{2,cc} + n_3 = 85\ 850 + 51\ 252$ , soit  $n_{3,cc} = 137\ 102$  ;  $n_{4,cc} = n_{3,cc} + n_4 = 137\ 102 + 26\ 613$ , soit  $n_{4,cc} = 163\ 175$  ;  $n_{5,cc} = n_{4,cc} + n_5 = 163\ 175 + 16\ 162$ , soit  $n_{5,cc} = 179\ 877$ .

Soit  $n_{i,cd}$  l'effectif cumulé décroissant de la classe  $i$  :  $n_{1,cd} = n$ , soit  $n_{1,cd} = 179\ 877$  ;  $n_{2,cd} = n_{1,cd} - n_1 = 179\ 877 - 31\ 038$ , soit  $n_{2,cd} = 148\ 839$  ;  $n_{3,cd} = n_{2,cd} - n_2 = 148\ 839 - 54\ 812$ , soit  $n_{3,cd} = 94\ 027$  ;  $n_{4,cd} = n_{3,cd} - n_3 = 94\ 027 - 51\ 252$ , soit  $n_{4,cd} = 42\ 775$  ;  $n_{5,cd} = n_{4,cd} - n_4 = 42\ 775 - 26\ 613$ , soit  $n_{5,cd} = 16\ 162$ .

Les résultats des effectifs cumulés croissants et décroissants se présentent dans un tableau obtenu en ajoutant deux colonnes au tableau statistique précédent : les effectifs cumulés croissants  $n_{i,cc}$  en colonne E et les effectifs cumulés décroissants  $n_{i,cd}$  en colonne F (voir figure 1.13).

**b.** La fonction de répartition est réalisée à partir des fréquences cumulées croissantes ( $f_{i,cc}$ ), calculées en colonne G du tableau statistique précédent (voir figure 1.13), sur du papier millimétré (voir figure 1.16).

**Figure 1.16**  
**Fonction de répartition du nombre d'enfants des familles de La Réunion.**



**5. a.** « Au moins 1 enfant » correspond aux familles qui ont 1, 2, 3 ou 4 et + enfants, ou encore toutes les familles sauf celles qui ont 0 enfant, c'est-à-dire toutes les familles sauf celles qui présentent la modalité  $x_1$  de X. Le nombre de ces familles est l'effectif cumulé décroissant  $n_{2cd} = 148\ 839$ , soit  $179\ 877 - 31\ 038$ . Ainsi, **148 839 familles** sont composées d'au moins 1 enfant.

**b.** « Au plus 2 enfants » correspond aux familles qui ont 0, 1 ou 2 enfants, c'est-à-dire les familles qui présentent les modalités  $x_1$ ,  $x_2$  ou  $x_3$  de X. Le nombre de ces familles est l'effectif cumulé croissant  $n_{3cc} = 137\ 102$ , soit  $31\ 038 + 54\ 812 + 51\ 252$ . Ainsi, **137 102 familles** sont composées d'au plus 2 enfants.



### EXERCICE 3 L'HISTOGRAMME

#### Énoncé

La Sécurité routière étudie l'accidentologie des passagers des véhicules de tourisme, âgés de 18 à 65 ans. Le tableau suivant indique le nombre de tués par tranches d'âge en 2005 :

Âge	Effectif
[18 ; 25[	790
[25 ; 35[	545
[35 ; 45[	377
[45 ; 65[	606

Source : ONISR, 2006

1. Déterminez :
  - a. la population étudiée,
  - b. la variable étudiée.
2. Précisez :
  - a. la nature de la variable ;
  - b. les modalités de la variable.
3. Dessinez l'histogramme de la distribution.

#### Solution

1. **a.** La population étudiée est composée des passagers des véhicules de tourisme, âgés de 18 à 65 ans.
- b.** La variable étudiée est  $X = \text{« âge des tués »}$ .
2. **a.** La variable étudiée est quantitative continue.
- b.** Les modalités de la variable étudiée sont les quatre classes suivantes : [18 ; 25[ ; [25 ; 35[ ; [35 ; 45[ ; [45 ; 65[.
3. Nous calculons les amplitudes de classes ( $A_i$ ), soit :
 
$$A_1 = 25 - 18 = 7 ; A_2 = 35 - 25 = 10 ; A_3 = 45 - 35 = 10 ; A_4 = 65 - 45 = 20.$$

Puisqu'elles sont différentes, il est nécessaire d'utiliser les densités pour réaliser l'histogramme.

Conformément à la figure 1.17, saisissez les effectifs ( $n_i$ ) dans la colonne L1 et les amplitudes ( $A_i$ ) dans la colonne L2.

Pour calculer les densités ( $d_i$ ) dans la colonne L3, placez le curseur sur l'en-tête de colonne L3. Indiquez  $L3=L1\div L2$ . Puis appuyez sur **ENTER**. La colonne L3 fait alors apparaître les densités (voir figure 1.17).

Les effectifs corrigés ( $n_{i,c}$ ) sont obtenus en multipliant ces densités par l'effectif minimal, soit 7. Pour calculer les effectifs corrigés ( $n_{i,c}$ ) dans la colonne L4, placez le curseur sur l'en-tête de colonne L4. Indiquez  $L4=L3\times 7$ . Puis appuyez sur **ENTER**. La colonne L4 fait alors apparaître les effectifs corrigés (voir figure 1.18).

**Figure 1.17 (gauche)**

**Calcul des densités avec la calculatrice.**

L1	L2	L3	L4
790	7	112.86	890
545	10	54.5	381.5
377	10	37.7	263.9
606	20	30.3	212.1

**Figure 1.18 (droite)**

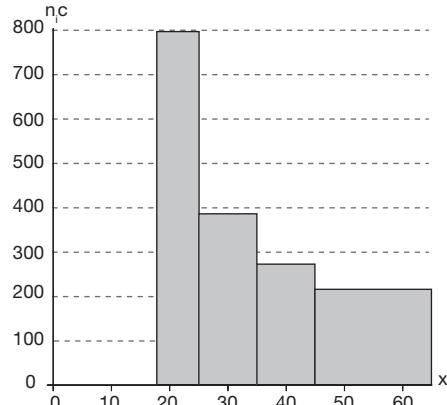
**Calcul des effectifs corrigés avec la calculatrice.**

$L3(1)=112.8571428...$	$L4(1)=789.9999999...$
------------------------	------------------------

L'histogramme des effectifs est ensuite tracé sur une feuille de papier millimétré (voir figure 1.19).

**Figure 1.19**

**Histogramme des tués par tranches d'âge.**



## EXERCICE 4 DISCRÉTISATION DES DONNÉES

### Énoncé

L'Agence de l'environnement et de la maîtrise de l'énergie (ADEME) vous informe sur les émissions de CO<sub>2</sub> par habitant dans le monde en 2002 :

Pays	Émissions de CO <sub>2</sub> (tonnes de CO <sub>2</sub> par habitant)
Asie du Sud	5
Afrique	1,39

Pays	Émissions de CO <sub>2</sub> (tonnes de CO <sub>2</sub> par habitant)
Amérique latine	2,79
Chine	3,05
Europe centrale	5,68
CEI	5,97
Moyen-Orient	6,04
Europe de l'Ouest	8,28
Japon	9,14
Asie (NPI)	10,46
Australasie	12,2
Amérique du Nord	20,02

Source : ADEME, 2002

L'ADEME souhaite distinguer trois classes de pays, selon leur niveau d'émissions de CO<sub>2</sub> :

- ceux qui émettent moins de 6 tonnes par habitant ;
  - ceux qui émettent de 6 à moins de 10 tonnes par habitant ;
- ceux qui émettent de 10 à moins de 22 tonnes par habitant.

Déterminez :

- a. la population étudiée ;
- b. la variable étudiée.
2. Précisez :
  - a. la nature de la variable ;
  - b. les modalités de la variable.
3. Construisez le tableau statistique associé. Pour cela, discrétez le caractère étudié selon la classification souhaitée par l'ADEME.
4. Dessinez l'histogramme de la distribution.

### Solution

1. a. La population étudiée est composée des régions du monde énumérées.
- b. La variable étudiée est X = « émissions de CO<sub>2</sub> ».
2. a. La variable étudiée est quantitative continue.
- b. L'ensemble des modalités de la variable étudiée est M = {0,82 ; 1,39 ; 2,79 ; 3,05 ; 5,68 ; 5,97 ; 6,04 ; 8,28 ; 9,14 ; 10,46 ; 12,2 ; 20,02}.
3. Le tableau statistique associé est composé de deux colonnes :
  - la première colonne comporte les classes d'émission de CO<sub>2</sub> ;
  - la seconde colonne comporte les effectifs n<sub>i</sub> affectés à chacune de ces classes.

Le tableau statistique associé à X est le suivant.

<b>Émissions de CO<sub>2</sub></b>	<b>n<sub>i</sub></b>
[0 ; 6[	6
[6 ; 10[	3
[10 ; 22[	3

4. Nous calculons ensuite les amplitudes de classes ( $A_i$ ), soit :

$$A_1 = 6 - 0 = 6 ; A_2 = 10 - 6 = 4 ; A_3 = 22 - 10 = 12.$$

Conformément à la figure 1.20, saisissez les effectifs ( $n_i$ ) dans la colonne L1 et les amplitudes ( $A_i$ ) dans la colonne L2.

Pour calculer les densités ( $d_i$ ) dans la colonne L3, placez le curseur sur l'en-tête de colonne L3. Indiquez L3=L1÷L2. Puis appuyez sur **ENTER**. La colonne L3 fait alors apparaître les densités (voir figure 1.20).

Figure 1.20 (gauche)

Calcul des densités avec la calculatrice.

L1	L2	L3	L4
6	6	1	4
3	4	.75	1
3	12	.25	1

**L3(1)=1**      **L4(1)=4**

Figure 1.21 (droite)

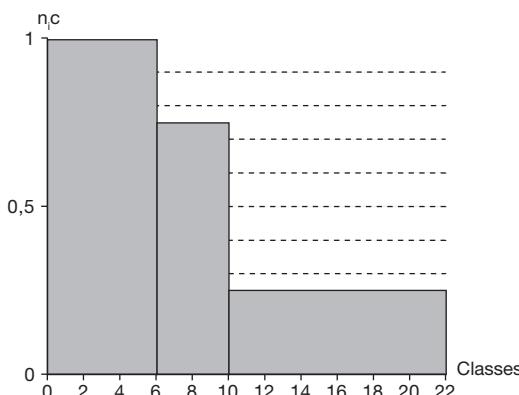
Calcul des effectifs corrigés avec la calculatrice.

Les effectifs corrigés ( $n_{i,c}$ ) sont obtenus en multipliant ces densités par l'effectif minimal, soit 4. Pour calculer les effectifs corrigés ( $n_{i,c}$ ) dans la colonne L4, placez le curseur sur l'en-tête de colonne L4. Indiquez L4=L3×4. Puis appuyez sur **ENTER**. La colonne L4 fait alors apparaître les effectifs corrigés (voir figure 1.21).

L'histogramme des effectifs est ensuite tracé sur une feuille de papier millimétré (voir figure 1.22).

Figure 1.22

Histogramme des pays selon leurs émissions de CO<sub>2</sub>.



## EXERCICE 5 LES POLYGONES

### Énoncé

L'ADEME vous transmet le tableau suivant, qui recense les individus dans le monde selon le niveau de CO<sub>2</sub> qu'ils émettent :

Émission moyenne de CO <sub>2</sub> (tonnes CO <sub>2</sub> par habitant)	Population (millions)
[0 ; 2[	2 205,79
[2 ; 4[	1 809,21
[4 ; 6[	401,26
[6 ; 8[	172,46
[8 ; 10[	590,05
[10 ; 16[	112,48
[16 ; 22[	319,84

Source : ADEME, 2002

1. Sur un même graphique :
  - a. Dessinez l'histogramme des fréquences de la distribution.
  - b. Dessinez le polygone des fréquences de la distribution.
2. À la suite de la question précédente :
  - a. Calculez les fréquences cumulées croissantes et décroissantes.
  - b. Représentez les polygones des fréquences cumulées croissantes et décroissantes sur un même graphique.

### Solution

1. a. Une simple lecture du tableau permet de voir que les amplitudes de classes ne sont pas constantes, ce qui est confirmé par leur calcul en colonne C (voir figure 1.23). Les fréquences sont calculées en colonne D, puis les densités ( $d_i$ ) en colonne E, en effectuant le rapport des fréquences sur les amplitudes. Enfin, les fréquences corrigées ( $f_{ic}$ ) sont obtenues en colonne F en multipliant ces densités par l'effectif minimal.

Figure 1.23

Résultats sous Excel.

	A Emissions de CO <sub>2</sub>	B n	C $a_i$	D $f_i$	E $d_i$	F $f_{ic}$	G $f_{cc}$	H $f_{icd}$
1	Emissions de CO <sub>2</sub>							
2	[0,2[	2 205,79	2	0,3931	0,1966	0,3931	0,39	1,00
3	[2;4[	1 809,21	2	0,3224	0,1612	0,3224	0,72	0,61
4	[4;6[	401,26	2	0,0715	0,0358	0,0715	0,79	0,28
5	[6;8[	172,46	2	0,0307	0,0154	0,0307	0,82	0,21
6	[8;10[	590,05	2	0,1052	0,0526	0,1052	0,92	0,18
7	[10;16[	112,48	6	0,0200	0,0033	0,0067	0,94	0,08
8	[16;22[	319,84	6	0,0570	0,0095	0,0190	1,00	0,06

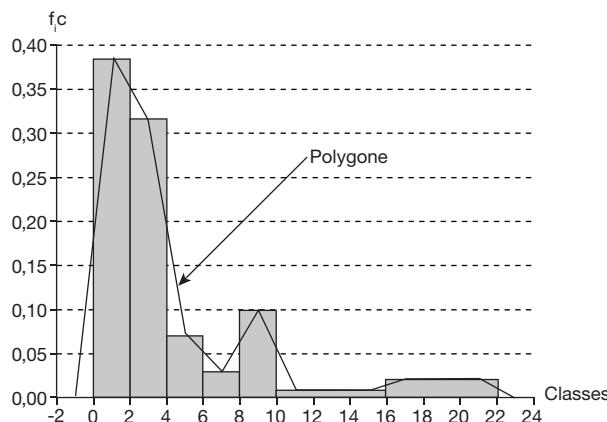
À partir de ces fréquences corrigées, il est possible de tracer l'histogramme des fréquences sur une feuille de papier millimétré (voir figure 1.24).

b. Les classes sont d'amplitudes inégales. On procède à un découpage artificiel en prenant l'amplitude minimale, soit 2, pour unité d'amplitude. Le polygone des fréquen-

ces est alors obtenu en joignant à la règle les milieux des bases supérieures des rectangles du découpage précédent (voir figure 1.24).

**Figure 1.24**

**Histogramme et polygone des fréquences des pays selon leurs émissions de CO<sub>2</sub>.**



**2. a.** À la suite du tableau Excel précédent, les fréquences cumulées croissantes ( $f_{cc}$ ) sont calculées dans la colonne G et les fréquences cumulées décroissantes ( $f_{cd}$ ) dans la colonne H (voir figure 1.23).

Ces calculs sont effectués selon le même principe que pour les effectifs cumulés croissants et décroissants, en remplaçant les effectifs par les fréquences.

**b.** La présentation de ces résultats est légèrement modifiée pour faire apparaître dans un même tableau les fréquences cumulées croissantes et décroissantes de chacune des bornes des classes (voir figure 1.25).

**Figure 1.25**

**Données pour les polygones de fréquences cumulées.**

	Bornes de classes	$f_{cc}$	$f_{cd}$
34			
35	0	0,00	1,00
36	2	0,39	0,61
37	4	0,72	0,28
38	6	0,79	0,21
39	8	0,82	0,18
40	10	0,92	0,08
41	16	0,94	0,06
42	22	1,00	0,00

Les courbes des fréquences cumulées croissantes et décroissantes de la figure 1.26 sont réalisées à partir de ce dernier tableau.

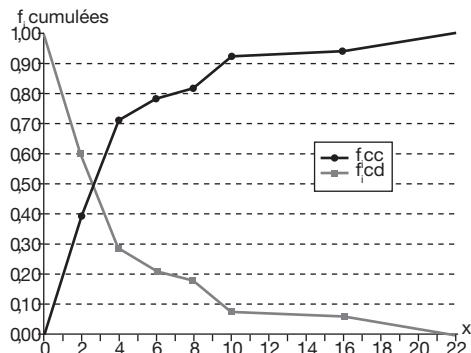
Pour réaliser ces courbes des effectifs cumulés sous Excel, cliquez sur Insertion/Graphique dans la barre de menus d'Excel.

L'assistant graphique apparaît. Dans l'assistant graphique, choisissez le type de graphique Nuages de points, puis, dans Sous-type de graphique, sélectionnez l'image représentant le Nuage de points reliés par une courbe. Cliquez sur Suivant.

L'assistant graphique propose de saisir les données du graphique.

**Figure 1.26**

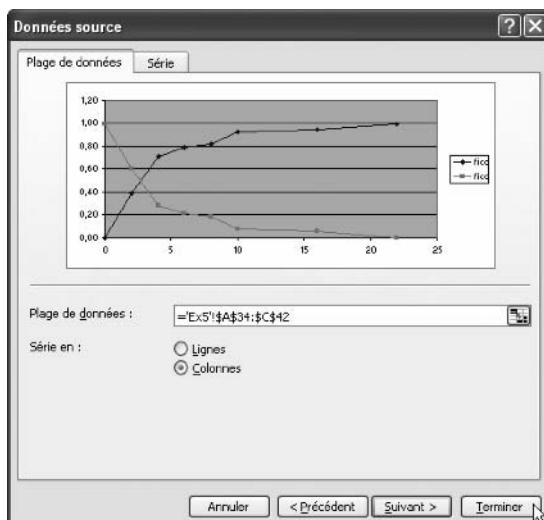
**Polygones des fréquences cumulées croissantes et décroissantes des pays selon leurs émissions de CO<sub>2</sub>.**



Dans l'onglet Plage de données, indiquez dans le champ correspondant la plage où se trouvent les données permettant de tracer les courbes correspondant aux polygones des effectifs cumulés croissants et décroissants. Pour cela, sélectionnez à l'aide de la souris la plage A34:C42 de la feuille Ex5 comme indiqué sur la figure 1.27, puis cliquez sur Terminer.

**Figure 1.27**

**Sélection des données à représenter dans l'assistant graphique.**



L'assistant graphique se ferme et le graphique apparaît (voir figure 1.26). Vous pouvez modifier les options d'affichage du graphique en appelant un menu par un clic droit sur la zone de graphique.

# Annexe 1.1

## Présentation du tableur (Excel)

Quand vous ouvrez Excel, la zone de travail située au centre s'appelle le CLASSEUR.

La BARRE DE TITRE de la fenêtre affiche le nom du classeur – par exemple, « Classeur1 » – que vous devez renommer et enregistrer.

Un classeur comporte par défaut trois feuilles, dont le nom figure sur un ONGLET – par exemple, « Feuil2 ». Il est possible de renommer, d'insérer ou de supprimer une feuille en faisant un clic droit sur un des onglets et en choisissant Insérer, Supprimer ou Renommer dans le menu.

L'intersection d'une ligne et d'une colonne s'appelle une CELLULE. Une cellule est caractérisée par sa RÉFÉRENCE, colonne-ligne – par exemple, « B4 ».

La BARRE DE MENUS permet d'accéder aux différents menus déroulants : Fichier, Edition, Affichage, Insertion, Format, Outils, Tableau, Fenêtre, ?.

Sous la barre de menus se trouvent les BARRES D'OUTILS, accessibles uniquement avec la souris. Lorsqu'on pointe sans cliquer sur les différents boutons, une info-bulle affiche le nom du bouton et sa fonction.

Sous les barres d'outils se trouve la BARRE DE FORMULE. Dans sa partie gauche apparaît la référence de la cellule active et dans la partie droite apparaissent les données, lors de leur saisie. Entre les deux, le symbole  $f_x$  (Insérer une fonction) désigne l'assistant fonction. Il comprend toutes sortes de fonctions, notamment statistiques, et sera extrêmement précieux pour les problèmes et exercices.

Pour saisir des données dans une cellule, placez la souris dessus, cliquez et entrez les chiffres ou les lettres voulus. Passez d'une cellule à une autre grâce à la souris ou aux touches  $\uparrow$ ,  $\downarrow$ ,  $\leftarrow$  et  $\rightarrow$  du clavier.

Pour effectuer une opération mathématique, cliquez sur une cellule, tapez le signe  $=$  pour indiquer qu'il s'agit d'une formule de calcul, puis faites l'opération en utilisant les signes mathématiques du clavier :  $+$ ,  $-$ ,  $*$  et  $/$ . Par exemple, pour additionner une cellule à une autre, cliquez sur la cellule qui doit accueillir le résultat, tapez  $=$ , cliquez sur la première cellule, tapez  $+$  puis cliquez sur la seconde cellule à additionner. Validez avec **ENTRÉE** pour faire apparaître le résultat. L élévation à la puissance s obtient en appuyant sur la touche accent grave,  $\wedge$ , suivie du nombre de la puissance désirée, ou en utilisant la fonction Puissance de l'assistant fonction.

Il existe trois types de références de cellules : pour passer d'un type à l'autre, utilisez la touche **F4**, qui procède par permutation circulaire, comme le montre cet exemple : saisissez **A1** dans la cellule A2, placez le curseur de la souris à la suite de A1, contre le 1, et appuyez sur **F4**. Vous voyez alors apparaître : **\$A\$1** (référence absolue). Si vous appuyez de nouveau sur **F4**, vous voyez apparaître successivement : **A\$1**, **\$A1** (références mixtes) et enfin **A1** (référence relative).

**Références relatives** : par défaut, sous Excel, les références des cellules sont « relatives ». Lorsqu'on recopie une formule d'une cellule à une autre, elle s'adapte automatiquement en fonction du déplacement en ligne ou en colonne.

- Si la formule  $=B2 + B3$  est saisie en B4 puis recopiée en C4, elle devient  $=C2 + C3$ .
- Si la formule  $=B2 + C2$  est saisie en D2 puis recopiée en D3, elle devient  $=B3 + C3$ .
- Si la formule  $=B2 + C2$  est saisie en D2 puis recopiée en E3, elle devient  $=C3 + D3$ .

**Références absolues** : on peut figer la colonne et la ligne d'une cellule, en mettant le signe « \$ » devant la lettre de la colonne et devant le nombre de la ligne, afin que la cellule concernée reste identique en cas de recopie d'une formule. Cette cellule est alors définie par une référence absolue dans la formule.

- Si la formule  $=B2 + \$C\$2$  est saisie en D2 puis recopiée en E3, elle devient  $=C3 + \$C\$2$ .

**Références mixtes** : on peut aussi décider de ne figer que la colonne ou que la ligne d'une cellule, en positionnant le symbole « \$ » uniquement devant la lettre ou le nombre de la cellule. La cellule est alors définie par une référence mixte.

Si la formule  $=B2 + \$C2$  est saisie en D2 puis recopiée en E3, elle devient  $=C3 + \$C3$ .

- Si la formule  $=B2 + C\$2$  est saisie en D2 puis recopiée en E3, elle devient  $=C3 + D\$2$ .

**La notion de fonction** : Excel comporte des fonctions intégrées, identifiées par des noms de fonctions – par exemple, SOMME, PRODUIT, MOYENNE, RACINE... Les éléments sur lesquels porte la fonction sont appelés ARGUMENTS, se placent entre parenthèses et sont séparés par des points-virgules.

**Pour utiliser une fonction** : placez le curseur dans la cellule où vous souhaitez faire apparaître le résultat. Cliquez sur Insertion/Fonction (ou utilisez directement  $\mathbb{F}_2$ ), sélectionnez la catégorie de fonction souhaitée (dans cet ouvrage, Statistique ou Math & Trigo), puis la fonction désirée. Entrez les arguments en vous laissant guider par la boîte de dialogue Excel. Validez en cliquant sur **OK**.

**Remarque** : pour faire une somme, il est possible de se servir de l'icône  $\Sigma$  proposée par défaut dans la barre d'outils. Cliquez sur la cellule où vous souhaitez faire apparaître la somme, cliquez sur l'icône  $\Sigma$ , puis sélectionnez les cellules dont vous souhaitez faire la somme, et validez avec **ENTRÉE**.

Les fonctions statistiques seront explorées lors de la correction des exercices.

## Annexe 1.2

### Présentation de la calculatrice (Texas Instrument)

**Notations** : les colonnes sont notées L1, L2, L3, L4, L5, L6. Les cellules sont identifiées par leur colonne, suivie de leur ligne entre parenthèses – par exemple, L1(2) indique la cellule figurant dans la première colonne, à la deuxième ligne.

**Pour saisir un tableau** : appuyez sur la touche **STAT**. Éditez le tableau en appuyant sur la touche **1**. Saisissez les données (validez chacune par la touche **ENTER**) en vous déplaçant avec le curseur.

**Pour quitter l'éditeur de tableau** : appelez la fonction **QUIT** par l'appui successif sur les touches **2ND** et **MODE**.

**Pour effacer une colonne entière** : placez le curseur sur l'en-tête de colonne **Li** que vous souhaitez effacer. Appuyez sur les touches **CLEAR** et **ENTER**.

**Pour effectuer la somme des termes d'une colonne** : placez le curseur dans la cellule (1) où vous souhaitez faire apparaître la somme. Appuyez sur les touches **2ND** et **LIST**, puis, dans le menu MATH,appelez la fonction **sum()**. Indiquez la colonne **Lj** dont vous souhaitez faire la somme (par exemple, **L1** est obtenu par **2ND** et **1**) et validez avec **ENTER**.

**Pour effectuer la somme cumulée d'une colonne** : placez le curseur sur l'en-tête de colonne **Li** dans laquelle vous souhaitez obtenir les effectifs cumulés. Appuyez sur les touches **2ND** et **LIST**, puis, dans le menu OPS,appelez la fonction **cumSum()**. Indiquez la colonne **Lj** dont vous souhaitez faire la somme cumulée et validez avec **ENTER**.

## Bibliographie

BOLL M., *L'exploitation du hasard*, Que sais-je ?, PUF, 1947.

CALOT G., *Cours de statistique descriptive*, Dunod, Paris, 1969.

CHAREILLE P. et PINAULT Y., *Statistique descriptive*, Collection AES, Montchrestien, Paris, 1996.

DODGE Y., *Premiers pas en statistiques*, Springer, 2005.

DROESBEKE J.-J., *Éléments de statistiques*, Éditions de l'université de Bruxelles, Ellipses, 2001.

LE BRAS H., *Naissance de la mortalité. L'origine politique de la statistique et de la démographie*, Gallimard/Le Seuil, Paris, 2000.

LIORZOU A., *Initiation à la pratique statistique*, Eyrolles, 1985.

PIATIER A., *Statistique descriptive et initiation à l'analyse*, Thémis, PUF, 1962.

ROGER P., *Probabilités, statistique et processus stochastiques*, Collection Synthex, Pearson Education, 2004.

ROHRBASSER J.-M. et VERON J., *Leibniz et les raisonnements sur la vie humaine*, INED, Paris, 2001.

VESSERAU A., *La statistique*, Que sais-je ?, PUF, 1962.

# Les caractéristiques de tendance centrale

1. Le mode .....	36
2. Les moyennes.....	39
3. Les quantiles .....	44
<b>Problèmes et exercices</b>	
1. Lecture de tendances centrales sur série brute .....	54
2. Tendances centrales sur tableau statistique, caractère discret.....	55
3. Tendances centrales sur tableau statistique, caractère continu .....	56
4. Visualisation graphique des tendances centrales .....	59
5. Moyenne géométrique.....	60
6. Moyenne harmonique .....	61

L'objectif de ce chapitre est de présenter les principaux paramètres qui permettent de résumer une série statistique d'observations et d'éclairer sur la position du noyau (centre) de la série. Ces paramètres sont appelés caractéristiques de position ou de tendance centrale de la série statistique à une variable. Nous présenterons ici le mode, la moyenne, la médiane, les quartiles et, plus généralement, les quantiles. Le statisticien anglais George Yule (1871-1951) a défini en 1911 les conditions idéales souhaitables pour une valeur centrale :

- être définie objectivement à partir de la série ;
- dépendre de tous les termes de la série ;
- être compréhensible par des non-spécialistes ;
- être simple à calculer ;
- être peu sensible aux fluctuations d'échantillonnage ;
- se prêter à des calculs algébriques.

Aucune des valeurs centrales définies ci-après n'est « parfaite » au sens de Yule.

# 1 Le mode

Lors de l'observation de la représentation graphique d'une distribution statistique (diagramme en bâtons ou histogramme), l'œil est souvent attiré par le bâton ou le rectangle le plus haut. Une des valeurs typiques d'une série statistique est le mode (valeur dominante).

Ce mot semble inspiré de « la mode », car il met en évidence la valeur la plus probable de la série.

La courbe « en cloche » de la distribution normale (voir chapitre 4, section 1) en donne une bonne vision.

## 1.1 PRÉSENTATION

**Définition** | Le **mode** est la valeur de la variable qui a l'effectif (ou la fréquence) le plus grand. On le note  $Mo$ .

En économie, dans les problèmes d'alimentation, de revenu, de logement, etc., le groupe qui a le plus grand poids est celui du mode. Il situe bien la position des valeurs les plus fréquemment rencontrées.

Le repérage du mode n'est pas un problème complexe, mais il faut distinguer le cas d'une variable qualitative ou quantitative discrète du cas d'une variable continue.

Il existe des séries unimodales (un mode) et des séries plurimodales (plusieurs modes).

## 1.2 VARIABLE QUALITATIVE OU QUANTITATIVE DISCRÈTE

Si la variable est qualitative ou quantitative discrète, on détermine le mode directement en identifiant la modalité de la variable qui correspond à l'effectif maximal (ou à la fréquence maximale).

Le mode d'une série discrète est une valeur de la série. Graphiquement, le mode correspond au bâton le plus long (aux bâtons les plus longs dans le cas des séries plurimodales).

## 1.3 VARIABLE QUANTITATIVE CONTINUE

Si la variable est quantitative continue, il faut procéder en deux étapes :

1. Détermination de la classe modale (elle n'est pas nécessairement unique), c'est-à-dire celle qui est représentée dans l'histogramme par le rectangle le plus haut : c'est la classe de plus grande densité. On notera que, si les classes sont de même amplitude, la classe modale est celle qui a le plus grand effectif (ou la plus grande fréquence).

**Rappel** (voir chapitre 1, section 3.2) : la densité d'effectif de la classe  $i$ , notée  $d_i$ , est le rapport  $d_i = \frac{n_i}{a_i}$ , avec  $n_i$  l'effectif et  $a_i$  l'amplitude de la classe. Cette densité représente le nombre d'individus par unité d'amplitude.

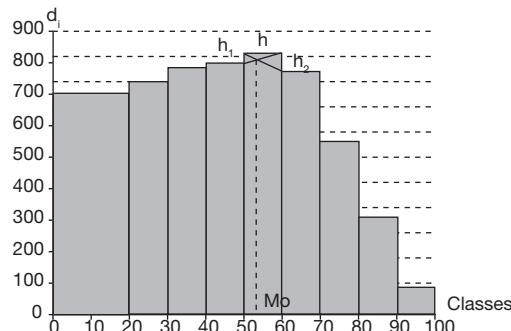
## 2. Détermination du mode à l'intérieur de cette classe modale.

Pour une première estimation, le mode peut être approché par le centre de la classe modale. En fait, le mode est l'abscisse du point où la courbe de densité de fréquence atteint son maximum.

Dans la pratique, nous disposons uniquement de l'histogramme. Le mode peut être estimé par la méthode des diagonales : on utilise le trapèze mis en évidence par les deux rectangles encadrant le rectangle modal (voir figure 2.1).

**Figure 2.1**

**Histogramme des effectifs et détermination du mode : structure des âges en 2020, France, prévisions.**



Graphiquement, la classe modale est le pic de l'histogramme corrigé (amplitudes égales) et le mode correspond à l'abscisse du point d'intersection des deux diagonales. Il se calcule donc sur les effectifs corrigés (c'est le seul indicateur qui se calcule sur effectifs corrigés).

Dans le cas où les amplitudes de classes sont égales, la classe modale est celle qui a le plus grand effectif (ou la plus grande fréquence). La suite de la démarche est identique.

Soit  $[x_1 ; x_2]$  la classe modale,  $h_1$  et  $h_2$  les hauteurs (effectifs corrigés ou densités) des rectangles encadrant le rectangle modal,  $h$  la hauteur du rectangle modal et  $Mo$  le mode.

Afin de calculer le mode, l'idée est d'abandonner l'hypothèse de répartition uniforme à l'intérieur de la classe modale, qui conduit à retenir le centre de classe. L'hypothèse privilégiée est celle d'une répartition influencée par les valeurs  $h_1$  et  $h_2$ , le mode étant « attiré » du côté du rectangle voisin de plus grande densité. Il est supposé que la densité croît de la valeur  $h_1$  à son maximum  $h$  et décroît de  $h$  à  $h_2$  avec la même vitesse, ce qui donne, avec les taux d'accroissement :  $\frac{h-h_1}{Mo-x_1} = \frac{h-h_2}{x_2-Mo}$ .

Soit  $\frac{k_1}{Mo-x_1} = \frac{k_2}{x_2-Mo}$ , avec  $k_1$  et  $k_2$  les différences :  $k_1 = h - h_1$  et  $k_2 = h - h_2$ .

En effectuant le produit en croix :  $Mo = \frac{k_2 x_1 + k_1 x_2}{k_1 + k_2}$ .

Le mode apparaît comme la moyenne pondérée de  $x_1$  et  $x_2$  respectivement affectés des coefficients  $h_2$  et  $k_1$ .

Une formule équivalente du mode est donnée par :  $Mo = x_1 + \frac{k_1}{k_1 + k_2}(x_2 - x_1)$ . Cette formule montre bien, par exemple, le déplacement du mode vers  $x_1$  dans le cas où  $k_1 < k_2$ , donc où  $\frac{k_1}{k_1 + k_2} < 0,5$ .

### Exemple 2.1

#### Calcul du mode sur variable quantitative continue

Considérons les prévisions de la structure démographique de la France en 2020 :

Âge	<b>n<sub>i</sub></b>	<b>a<sub>i</sub></b>	<b>d<sub>i</sub></b>
0-19 ans	14 115	20	705,75
20-29 ans	7 403	10	740,3
30-39 ans	7 842	10	784,2
40-49 ans	7 967	10	796,7
50-59 ans	8 281	10	828,1
60-69 ans	7 716	10	771,6
70-79 ans	5 521	10	552,1
80-89 ans	3 074	10	307,4
90-99 ans	878	10	87,8

Source : Insee, projections des ménages à l'horizon 2020 pour la France métropolitaine, juillet 2006

Les amplitudes de classes étant différentes, nous utilisons les densités pour déterminer la classe modale et représenter l'histogramme (voir figure 2.1). La classe modale est donc la classe des 50-59 ans soit [50 ; 60[ avec une densité de 828,1.

$$x_1 = 50 ; \quad x_2 = 60 ; \quad h = 828,1 ; \quad h_1 = 796,7 ; \quad h_2 = 771,6 ; \quad k_1 = 828,1 - 796,7 = 31,4 ; \\ k_2 = 828,1 - 771,6 = 56,5.$$

En appliquant la formule du mode :

$$Mo = \frac{k_2 x_1 + k_1 x_2}{k_1 + k_2} = \frac{56,5 \times 50 + 31,4 \times 60}{31,4 + 56,5}, \text{ soit } Mo = 53,57 \text{ ans, soit environ 52 ans et } 7 \text{ mois.}$$

Le mode est très peu conforme aux conditions de Yule. Il ne se prête pas aux calculs algébriques, et ne dépend pas de tous les termes de la série. Cependant, il reste une valeur centrale importante pour les distributions ayant un effectif important, car il donne la valeur la plus typique.

## 2 Les moyennes

« Si un individu possédait à une époque donnée toutes les qualités de l'homme moyen, il représenterait tout ce qui est grand, bon et beau », disait Adolphe Quetelet<sup>1</sup>.

Sans nous attacher à la notion contestée d'« homme moyen » de Quetelet, gardons à l'esprit que l'idée de moyenne est une notion abstraite. Quand le statisticien calcule une moyenne, il fabrique en général une grandeur nouvelle, qui a la vocation d'être représentative de toutes les grandeurs considérées, mais qui n'a en général aucune existence réelle. Nous imaginons mal un fabricant de chaussures qui fabriquerait des chaussures correspondant à la taille moyenne.

Quatre types de moyennes sont définies ici : les moyennes arithmétiques et celles, moins utilisées, que sont les moyennes géométriques, harmoniques et quadratiques.

La moyenne arithmétique garde un rôle primordial du fait de sa simplicité de calcul, mais surtout du fait de sa place fondamentale dans la théorie des erreurs d'observation (loi de Laplace-Gauss<sup>2</sup>) et dans la théorie de la régression (voir chapitre 6).

L'idée fondamentale de la notion de moyenne est que cette dernière vise à représenter des grandeurs inégales par une grandeur unique qui ne change pas la globalité de la situation. Ainsi, dans une entreprise où les personnels ont des salaires différents, la masse salariale resterait inchangée si tous les personnels percevaient le même salaire moyen.

### 2.1 LA MOYENNE ARITHMÉTIQUE

C'est en astronomie, avec Tycho Brahe<sup>3</sup>, que la moyenne arithmétique s'impose. Johann Bernoulli<sup>4</sup> la qualifie dans *l'Encyclopédie* comme le « milieu à prendre entre les observations ».

Cette moyenne, liée à l'addition, est la moyenne la plus couramment utilisée. Elle représente bien l'idée de milieu, d'équilibre, symbolisée par la place du zéro dans les nombres.

#### Définitions

La moyenne arithmétique est la somme des valeurs observées rapportée au nombre d'observations. Elle se note  $\bar{x}$ .

La **moyenne arithmétique simple** de  $n$  réels (données en tableau brut) correspond à la division de leur somme par leur nombre. Soit  $x_1, x_2, \dots, x_n$  les  $n$  observations de la variable  $X$  (non nécessairement distinctes) : la moyenne arithmétique se note  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ . Cette formule

implique que  $\sum_{i=1}^n x_i = n\bar{x}$ .

1. Adolphe Quetelet (1796-1874), astronome, statisticien belge.

2. Pierre Simon de Laplace (1749-1827), mathématicien, astronome français. Carl Friedrich Gauss (1777-1855), astronome, mathématicien allemand.

3. Tycho Brahe (1546-1601), astronome danois.

4. Johann Bernoulli (1667-1748), mathématicien suisse.

La **moyenne arithmétique pondérée** de  $r$  réels (distincts)  $x_1, x_2, \dots, x_r$  (données en tableau statistique), affectés respectivement des coefficients  $n_i$ , tels que  $\sum_{i=1}^r n_i = n$ , se note  

$$\bar{x} = \frac{1}{n} \sum_{i=1}^r n_i x_i .$$
 Ou encore  $\bar{x} = \sum_{i=1}^r f_i x_i .$

Les probabilistes parlent d'espérance d'une variable aléatoire et notent sa moyenne  $E(X) = \sum_{i=1}^r p_i x_i$ , les probabilités  $p_i$  se substituant aux fréquences  $f_i$ .

### Calcul de la moyenne arithmétique dans le cas d'une variable discrète

#### Exemple 2.2

#### Calcul d'une moyenne arithmétique pondérée

La répartition des effectifs du préélémentaire dans les établissements publics par âges en 2005-2006 est la suivante :

Âge	$n_i$	$f_i (\%)$
2 ans	154 141	0,0702
3 ans	667 328	0,3038
4 ans	685 158	0,3119
5 ans	680 202	0,3097
6 ans et plus	9 683	0,0044

Source : ministère de l'Éducation nationale, 2007

On notera qu'*a priori* l'âge est une variable continue ; cependant, l'Éducation nationale présente ici cette variable comme une variable discrète et nous la traiterons ainsi. Par ailleurs, nous prendrons 6 pour la dernière modalité.

Pour calculer la moyenne, il est nécessaire de calculer chacun des  $n_i x_i$ , avant d'en calculer la somme (voir figure 2.2).

La moyenne est obtenue en divisant la somme des  $n_i x_i$  par l'effectif total. L'âge moyen dans les établissements publics de maternelle est :  $\bar{x} = \frac{8 510 006}{2 196 512} = 3,87$  ans. On peut également retrouver cette valeur en calculant chacun des  $f_i x_i$  et en effectuant leur somme.

Figure 2.2

**Calcul des  $n_i x_i$  sous Excel (établissements publics).**

	A	B	C	D	E
1	Âge ( $x_i$ )	$n_i$	$n_i x_i$	$f_i$	$f_i x_i$
2	2	154 141	308 282	0,0702	0,1404
3	3	667 328	2 001 984	0,3038	0,9114
4	4	685 158	2 740 632	0,3119	1,2477
5	5	680 202	3 401 010	0,3097	1,5484
6	6	9 683	58 098	0,0044	0,0265
7	Somme	2 196 512	8 510 006	1,00	3,87

De même, en calculant chacun des  $f_i x_i$  et en effectuant leur somme, on trouve que l'âge moyen dans les établissements privés de maternelle est :  $\bar{x} = 3,8$  ans (voir figure 2.2).

## Calcul de la moyenne arithmétique dans le cas d'une variable continue

Les définitions et formules des moyennes arithmétiques simple et pondérée sont les mêmes que celles utilisées dans le cas d'une variable discrète. La méthode reste identique à l'exception de l'utilisation de l'hypothèse de répartition uniforme à l'intérieur des classes et de concentration au centre des classes, ce qui autorise le calcul de la moyenne à partir des centres de classes.

### Exemple 2.3

#### Calcul d'une moyenne arithmétique pondérée sur variable continue en classe

Reprendons les prévisions de l'Insee à l'horizon 2020 (voir exemple 2.1) et calculons l'âge moyen prévisible. Pour calculer les  $n_i x_i$ , il faut préalablement calculer les centres de classes  $x_i$ . Si  $a_i$  et  $b_i$  représentent respectivement les bornes inférieure et supérieure des classes, alors le centre de classe  $x_i = \frac{a_i + b_i}{2}$ . Une fois les  $x_i$  connus, il convient de calculer chacun des  $f_i x_i$ , avant d'en faire la somme (voir figure 2.3).

**Figure 2.3**

#### Calcul des $f_i x_i$ sous Excel.

	A	B	C	D	E
1	$a_i$	$b_i$	$x_i$	$f_i$	$f_i x_i$
2	0	20	10	0,225	2,25
3	20	30	25	0,118	2,95
4	30	40	35	0,124	4,34
5	40	50	45	0,127	5,72
6	50	60	55	0,132	7,26
7	60	70	65	0,123	8,00
8	70	80	75	0,088	6,60
9	80	90	85	0,049	4,17
10	90	100	95	0,014	1,33
11	<b>Somme</b>			1,000	42,61

$\bar{x} = 42,61$ , l'âge moyen est de 42,61 ans.

## Propriétés de la moyenne arithmétique

La moyenne arithmétique possède la propriété de linéarité :  $\overline{x+y} = \bar{x} + \bar{y}$  et  $\overline{ax} = a\bar{x}$ ,  $a$  étant une valeur constante.

Par exemple, soit une entreprise dans laquelle le revenu des personnels se compose d'un salaire  $x$  et d'une prime  $y$ , le salaire moyen mensuel étant de 3 500 euros et la prime moyenne mensuelle de 200 euros. Le revenu moyen mensuel sera de 3 700 euros. De même, si tous les salaires sont augmentés de 5 %, le salaire moyen deviendra :  $3\,500 \times 1,05 = 3\,675$  euros.

Si toutes les valeurs des observations sont identiques, la moyenne de ces observations est égale à cette valeur commune. Autrement dit, la moyenne d'une variable statistique constante est égale à elle-même.

D'où :  $\overline{ax+b} = a\bar{x} + b$ ,  $a$  et  $b$  étant des valeurs constantes. Cela permet notamment de changer d'unité, ou d'origine, toute transformation linéaire effectuée sur la variable étant répercutée sur la moyenne.

La moyenne des écarts à la moyenne est nulle.

$$\sum_{i=1}^r n_i(x_i - \bar{x}) = \sum_{i=1}^r n_i x_i - \sum_{i=1}^r n_i \bar{x} = \sum_{i=1}^r n_i x_i - n \bar{x} = 0, \text{ car selon la formule de la moyenne}$$

$$\frac{1}{n} \sum_{i=1}^r n_i x_i = \bar{x}, \text{ soit } \sum_{i=1}^r n_i x_i = n \bar{x}.$$

Cela explique pourquoi nous choisirons la moyenne des écarts au carré pour mesurer la dispersion, encore appelée variance.

La moyenne arithmétique dépend de tous les termes de la série, elle se prête bien aux calculs, c'est un bon indicateur de tendance centrale au sens de Yule. En revanche, elle présente l'inconvénient d'être très sensible aux valeurs extrêmes. C'est pourquoi elle est qualifiée d'indicateur peu robuste.

## 2.2 LES AUTRES MOYENNES

---

### La moyenne géométrique : moyenne de la multiplication

Introduisons cette moyenne par un exemple : soit une pièce rectangulaire de 16 mètres sur 9 mètres. Quelle serait la dimension du côté d'une pièce carrée de même aire ?

Si  $g$  désigne notre inconnue,  $g^2 = 16 \times 9$  soit  $g$  étant un réel positif,  $g = \sqrt{16 \times 9} = 5$  ; ce nombre est appelé la moyenne géométrique de 16 et 9.

#### Définitions

**La moyenne géométrique simple**, notée  $g$ , de  $n$  réels positifs est la racine  $n^{\text{ième}}$  de leur produit :  $G = \sqrt[n]{\prod_{i=1}^n X_i}$ . Ou encore  $G = \sqrt[n]{x_1 \times x_2 \times \dots \times x_n} = (x_1 \times x_2 \times \dots \times x_n)^{\frac{1}{n}}$ .

**La moyenne géométrique pondérée** de  $r$  réels positifs, affectés respectivement des coefficients  $n_i$ , tels que  $\sum_{i=1}^r n_i = n$ , se note  $G$ , tel que  $G = \sqrt[n]{\prod_{i=1}^r X_i^{n_i}}$ .

Ou encore  $G = \sqrt[n]{x_1^{n_1} \times x_2^{n_2} \times \dots \times x_r^{n_r}} = (x_1^{n_1} \times x_2^{n_2} \times \dots \times x_r^{n_r})^{\frac{1}{n}} = x_1^{f_1} \times x_2^{f_2} \times \dots \times x_r^{f_r}$ .

La moyenne géométrique sera utilisée dans le chapitre 8 sur les indices (indice de Fisher). Par ailleurs, elle est indispensable dans les calculs de taux de croissance ; elle donne le coefficient multiplicateur moyen.

#### Exemple 2.4

#### Calcul d'une moyenne géométrique pondérée

Supposons que la population d'un pays ait augmenté trois années de suite de 4 % et deux années de suite de 5 %, l'augmentation moyenne sera donnée par  $1 + t = \sqrt[5]{1,04^3 \times 1,05^2}$ , soit un taux de croissance annuel moyen  $t = (1,04^3 \times 1,05^2)^{\frac{1}{5}} - 1$ , soit environ 4,40 % par an.

La moyenne géométrique est très liée à la moyenne arithmétique. En effet :  $\ln(g) = \frac{1}{n} \sum_{i=1}^r n_i \ln(x_i)$ . Ainsi, la moyenne géométrique est égale à la moyenne arithmétique pondérée des logarithmes népériens.

Nous noterons également que, sur la courbe de la fonction exponentielle, en prenant deux points d'abscisses respectives  $a$  et  $b$ , l'ordonnée du point d'abscisse  $\frac{a+b}{2}$  est  $e^{\frac{a+b}{2}} = \sqrt{e^a \times e^b}$ , soit une moyenne géométrique.

### La moyenne harmonique : moyenne de l'inverse

Si la moyenne arithmétique s'impose dans de nombreuses situations, le recours à d'autres moyennes est parfois indispensable.

Prenons un exemple classique : supposons qu'un aller-retour Paris-Deauville soit effectué avec une vitesse moyenne de 130 km/h à l'aller et de 80 km/h au retour. Que penser de la vitesse moyenne sur l'aller-retour ?

Soit  $d$  la distance Paris-Deauville,  $t$  le temps du trajet et  $v$  la vitesse. Alors  $v = \frac{d}{t}$ . D'où

$$v = \frac{2d}{\frac{d}{130} + \frac{d}{80}} = \frac{2}{\frac{1}{130} + \frac{1}{80}} = 99,04 \text{ km/h, et non } 105 \text{ km/h comme le donnerait la}$$

moyenne arithmétique. Nous pouvons également écrire :  $\frac{2}{v} = \frac{1}{130} + \frac{1}{80}$ ,  $v$  s'appelant la moyenne harmonique des vitesses.

#### Définitions

**La moyenne harmonique simple** de  $n$  nombres réels non nuls est le réel noté  $H$  et défini par :

$$\frac{n}{H} = \sum_{i=1}^n \frac{1}{x_i} \text{ soit } H = \frac{1}{\frac{1}{n} \sum_{i=1}^n \frac{1}{x_i}}$$

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

**La moyenne harmonique pondérée** de  $r$  nombres réels non nuls, affectés respectivement des coefficients  $n_i$ , tels que  $\sum_{i=1}^r n_i = n$ , est le réel noté  $h$  et défini par :  $\frac{n}{H} = \sum_{i=1}^r \frac{n_i}{x_i}$  soit

$$H = \frac{1}{\frac{1}{n} \sum_{i=1}^r \frac{n_i}{x_i}}, \text{ soit l'inverse de la moyenne arithmétique pondérée de leurs inverses. Ou}$$

$$\text{encore } H = \frac{n}{\sum_{i=1}^r \frac{n_i}{x_i}}.$$

La moyenne harmonique sera également utilisée dans le chapitre 8 sur les indices (indice de Paasche).

## La moyenne quadratique

Le mot quadratique, qui vient du latin, évoque le carré et est utilisé pour désigner la puissance deux.

Partons d'un exemple simple : supposons un appartement composé de deux pièces carrées de côtés respectifs  $a$  et  $b$  ( $a \neq b$ ) et cherchons la mesure du côté  $Q$  des pièces d'un appartement de même surface, mais composé de deux pièces identiques carrées. On

$$\text{aura : } 2Q^2 = a^2 + b^2 \text{ soit } Q = \sqrt{\frac{a^2 + b^2}{2}}.$$

La moyenne quadratique, ou moyenne d'ordre 2, est la moyenne qui sert à définir l'écart-type d'une variable statistique, que nous utiliserons lors de l'étude de la dispersion.

### Définitions

La **moyenne quadratique simple** de  $n$  nombres réels, notée  $Q$ , correspond à la moyenne arithmétique de leurs carrés :  $Q = \sqrt{\frac{1}{n} \sum_{i=1}^n x_i^2}$ . Ou encore  $Q = \sqrt{\frac{x_1^2 + x_2^2 + \dots + x_n^2}{n}}$ .

La **moyenne quadratique pondérée**, notée  $Q$ , de  $r$  nombres réels, affectés respectivement des coefficients  $n_i$ , tels que  $\sum_{i=1}^r n_i = n$ , correspond à la moyenne arithmétique pondérée de leurs carrés :  $Q = \sqrt{\frac{1}{n} \sum_{i=1}^r n_i x_i^2}$ . Ou encore  $Q = \sqrt{\frac{n_1 x_1^2 + n_2 x_2^2 + \dots + n_r x_r^2}{n}}$ .

### Focus 2.1

#### Remarques sur les moyennes

1. Une moyenne représente toujours un « centre » d'une série de données. Soit  $x_{\min}$  et  $x_{\max}$  respectivement la plus petite et la plus grande valeur de la série. Les moyennes de la série statistique appartiennent toujours à l'intervalle  $[x_{\min}; x_{\max}]$ .
2. Il est important de retenir l'ordre de ces moyennes :  $x_{\min} \leq H \leq G \leq \bar{x} \leq Q \leq x_{\max}$ . Cette remarque, qui servira notamment pour comparer les indices synthétiques, est aisément démontrable pour deux réels  $a$  et  $b$ . Vérifions-le avec un exemple :  $a = 9$  et  $b = 16$  donne :  $H = \frac{288}{25} = 11,52$  ;  $G = 12$  ;  $\bar{x} = 12,5$  et  $Q = \sqrt{\frac{256 + 81}{2}} = 12,98$ .
3. Les moyennes sont des indicateurs qualifiés de peu robustes en ce sens qu'ils sont sensibles aux valeurs extrêmes.

## 3 Les quantiles

Partons d'un exemple : « En 2005, 10 % des salariés à temps complet du secteur privé et semi-public gagnent un salaire annuel net inférieur à 12 506 € » (source : Insee, DADS, 2005).

On dit que 12 506 constitue le quantile d'ordre 0,10 de la série des salaires considérée.

Si  $p$  est un réel de l'intervalle  $]0 ; 1[$ , on lui associe la valeur de la série, notée  $Q(p)$ , appelée quantile d'ordre  $p$ . La proportion des valeurs de la série inférieures ou égales à  $Q(p)$  est supérieure ou égale à  $p$ .

La médiane est un quantile particulier qui sépare la population en deux groupes d'effectifs égaux.

### 3.1 LA MÉDIANE

---

Il est clair que l'idée de partager la série en deux groupes ayant exactement le même effectif n'est pas toujours réalisable, aussi la définition de la médiane doit-elle être affinée.

#### Définition

La **médiane**, notée  $M_e$ , est la plus petite valeur de la série\* pour laquelle le nombre d'observations inférieures ou égales à cette valeur représente au moins 50 % de l'effectif total de la série.

C'est le quantile d'ordre 0,5.

\* Convention : dans le cas d'une série discrète comportant un nombre pair d'observations, la médiane n'est pas nécessairement une valeur observée (voir exemple 2.6).

Ainsi, il y a au moins 50 % des observations ayant une valeur inférieure ou égale à la médiane et au moins 50 % des observations ayant une valeur supérieure ou égale à la médiane.

On détermine la médiane à l'aide des effectifs cumulés croissants, à partir de la série des valeurs ordonnées dans l'ordre croissant. Il convient de distinguer le cas d'une variable présentée sous forme de données brutes du cas d'une variable présentée dans un tableau statistique. Dans ce dernier cas, on distinguera le cas discret et le cas continu.

#### La médiane d'une série de données brutes

Tout d'abord la série doit être classée dans l'ordre croissant des valeurs.

La détermination directe ou non de la médiane dépend du nombre de données brutes.

1. Si ce nombre est impair, il est possible de déterminer directement la médiane.
2. Si ce nombre est pair, la médiane est déduite de l'intervalle médian.

Calcul 1 : si la série brute comporte un nombre *impair* d'observations, noté  $n = 2p + 1$ , la médiane est la valeur centrale de la série (ordonnée en sens croissant), donc la  $(p + 1)^{\text{ème}}$  observation.

#### Exemple 2.5

#### Calcul de la médiane, nombre impair de données brutes

Le tableau suivant donne le taux d'emploi (en pourcentage) des jeunes de 15 à 24 ans, en 2005, dans les sept pays de l'Union européenne ayant le plus fort taux.

Pays	Taux d'emploi
Allemagne	42
Pays-Bas	65,2
Autriche	53,1

Pays	Taux d'emploi
Irlande	48,7
Royaume-Uni	54
Danemark	62,3
Finlande	40,5

Source : Insee, juillet 2006

Classons tout d'abord les modalités par ordre croissant. Dans notre exemple, ces modalités sont au nombre de  $n = 7$ , c'est-à-dire un nombre impair, et  $p = 3$ , donc la médiane est la valeur centrale de la série ordonnée, c'est-à-dire la 4<sup>e</sup> observation : 40,5 – 42 – 48,7 – 53,1 – 54 – 62,3 – 65,2. La médiane est  $Me = 53,1$ .

Calcul 2 : si la série brute comporte un nombre *pair* d'observations, noté  $n = 2p$ , il convient de déterminer l'intervalle médian, constitué par les observations de rang  $p$  et  $p + 1$  de la série ordonnée. Par convention, la médiane est le milieu de cet intervalle médian.

#### Exemple 2.6

#### Calcul de la médiane, nombre pair de données brutes

Reprendons l'exemple précédent (voir exemple 2.5) et rajoutons la France avec un taux de 30,1 %. Le nombre de modalités devient  $n = 8$ , donc  $p = 4$ . L'intervalle médian est constitué de la 4<sup>e</sup> et de la 5<sup>e</sup> observation, c'est donc l'intervalle médian [48,7 ; 53,1]. Par convention,  $Me = \frac{48,7 + 53,1}{2} = 50,9$ .

#### La médiane dans un tableau statistique

Pour calculer la médiane à partir d'un tableau statistique, il convient de distinguer deux cas :

1. Soit la variable est présentée comme un caractère discret.
2. Soit la variable est présentée comme un caractère continu.

Calcul 1 : Dans le premier cas, les modalités de la variable sont des valeurs isolées. La détermination de la médiane se fait directement à l'aide des effectifs cumulés croissants (voir figure 2.4).

#### Exemple 2.7

#### Calcul de la médiane pour une variable présentée comme un caractère discret

Le tableau suivant donne le nombre d'enfants de moins de 25 ans par famille, en France métropolitaine en 2005 :

Nombre d'enfants	$n_i$ (milliers)
1	3 714
2	3 369
3	1 237
4 ou +	410

Source : Insee, enquêtes de recensement, 2004-2006

**Figure 2.4**  
**Effectifs cumulés croissants.**

	A	B	C
1	Age ( $x_i$ )	$n_{\text{Public}}$	$n_{\text{cc}}$
2	2	154 141	154 141
3	3	667 328	821 469
4	4	685 158	1 506 627
5	5	680 202	2 186 829
6	6	9 683	2 196 512
7	<b>Somme</b>	2 196 512	

Le nombre d'observations est pair, donc l'intervalle médian est constitué par les deux observations centrales, c'est-à-dire de rangs respectifs  $p = \frac{n}{2} = \frac{8730000}{2} = 4365000$  et  $p + 1 = 4365001$ . Les effectifs cumulés croissants nous montrent que ces observations sont dans la modalité 2, donc que la médiane, leur moyenne arithmétique, est 2. Il y a au moins 50 % des familles ayant un nombre d'enfants inférieur ou égal à 2 et au moins 50 % des familles ayant un nombre d'enfants supérieur ou égal à 2.

Calcul 2 : Dans le second cas, les modalités de la variable sont des classes. La détermination de la médiane repose sur l'hypothèse que les observations sont réparties uniformément au sein de chaque classe. La médiane est alors définie par  $F(Me) = 0,50$ , où  $F$  désigne la fonction de répartition. Son calcul se fait en deux temps :

1. Localisation de la classe médiane à l'aide des effectifs cumulés croissants ou des fréquences cumulées croissantes.
2. Calcul de la médiane par interpolation linéaire (voir focus 2.2).

## Focus 2.2

### Interpolation linéaire

Le mot « inter » signifie que nous opérons entre deux valeurs connues, appelées pôles. Le mot « linéaire » évoque la droite.

Supposons une fonction  $f$  définie sur un segment  $[a ; b]$ , et dont nous connaissons les valeurs  $f(a)$  et  $f(b)$ , le problème étant d'estimer la valeur de  $f$  en un point  $x$  du segment  $[a ; b]$ . Le principe de l'interpolation linéaire est donc de supposer l'alignement des points A, B et M dont les coordonnées sont  $A(x_A ; y_A)$  ;  $B(x_B ; y_B)$  ;  $M(x_M ; y_M)$ .

Cet alignement des points A, B et M est représenté sur la figure 2.5.

L'alignement des points A, B et M se traduit par l'égalité des coefficients directeurs, ou encore par l'égalité des rapports des distances, en utilisant le théorème de Thalès.

$$\frac{AB}{AM} = \frac{AB'}{AM'} = \frac{AB''}{AM''}, \text{ soit } \frac{y_B - y_A}{x_B - x_A} = \frac{y_M - y_A}{x_M - x_A}, \text{ ce qui donne, après un produit en croix :}$$

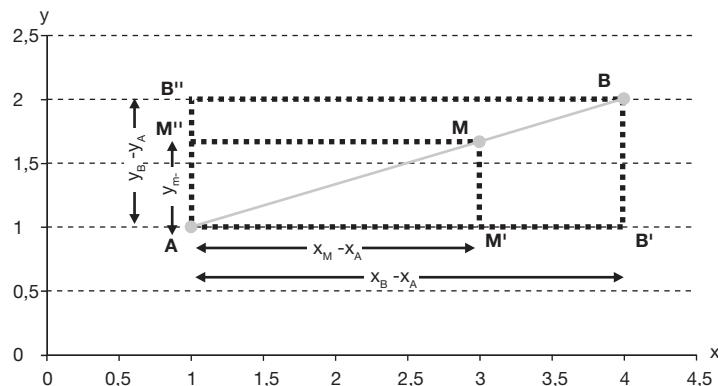
$$y_M = y_A + \frac{y_B - y_A}{x_B - x_A}(x_M - x_A).$$

Par exemple, déterminons une valeur approchée de  $\sqrt{3}$  ( $= 1,732$ ) par interpolation linéaire sur le segment  $[1 ; 4]$  (voir figure 2.5). L'interpolation linéaire donne :  $\frac{2-1}{4-1} = \frac{\sqrt{3}-1}{3-1}$  soit :

$$\sqrt{3} = 1 + \frac{2}{3} = \frac{5}{3} = 1,667.$$

**Figure 2.5**

**Alignement et égalité des coefficients directeurs.**



**Exemple 2.8**

**Calcul de la médiane pour une variable présentée comme un caractère continu**

Soit le nombre de personnes de plus de 15 ans ayant un niveau d'études supérieures (voir figure 2.6).

**Figure 2.6**

**Calcul des  $n_{iCC}$  sous Excel.**

	A	B	C	D
1	a <sub>i</sub>	b <sub>i</sub>	n <sub>i</sub>	n <sub>iCC</sub>
2	15	20	7 573	7 573
3	20	25	563 044	570 617
4	25	30	1 594 191	2 164 808
5	30	40	2 488 412	4 653 220
6	40	60	3 063 199	7 716 419
7	60	80	974 740	8 691 159
8	Somme		8 691 159	

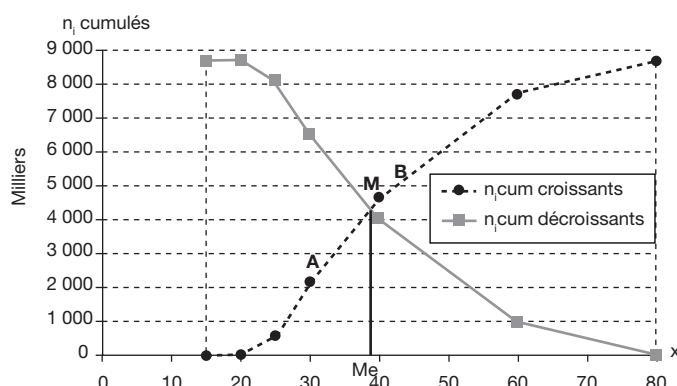
Source : Insee, recensement de la population, 1999

Le calcul de l'effectif moitié, 4 345 579,5, et les effectifs cumulés croissants permettent de localiser la médiane dans l'intervalle des 30-40 ans.

Le polygone des effectifs cumulés croissants permet une visualisation graphique de la médiane (voir figure 2.7). Soit les trois points A (30 ; 2 164 808), B (40 ; 4 653 220) et M ( $M_e$  ; 4 345 579,5).

**Figure 2.7**

**Médiane et effectifs cumulés par âges des personnes de plus de 15 ans ayant un niveau d'études supérieures.**



Nous pouvons écrire l'alignement de ces trois points par égalité des coefficients directeurs (interpolation linéaire ; voir focus 2.2) :  $\frac{4\,653\,220 - 2\,164\,808}{40 - 30} = \frac{4\,345\,579,5 - 2\,164\,808}{Me - 30}$ , soit  $\frac{2\,488\,412}{10} = \frac{2\,180\,771,5}{Me - 30}$ , ce qui donne, en effectuant le produit en croix :  $Me = \frac{2\,180\,771,5}{248\,841,2} + 30 = 38,76$  ans.

La médiane est à relier à la notion de fonction de répartition, fonction définie de R dans  $[0 ; 1]$ , extrêmement importante en probabilité. Pour une variable statistique continue, la fonction de répartition se définit par :  $F(x) = P(X \leq x)$ , qui donne la proportion des individus de la population pour lesquels la variable statistique prend une valeur inférieure ou égale à x. Ainsi :  $F(Me) = 0,50$ .

La médiane ne satisfait pas bien aux conditions de Yule. Elle dépend du nombre de termes, mais pas de leur grandeur, et est inadaptée aux calculs. Elle présente cependant le grand avantage d'être insensible à l'influence des termes extrêmes, et donc d'être robuste.

## 3.2 LES QUANTILES : GÉNÉRALISATION DE LA MÉDIANE

### Les quantiles

#### Définition

On suppose que les modalités de la série statistique sont rangées dans l'ordre croissant.

Soit  $p$  un réel tel que  $0 < p < 1$ , on lui associe la valeur de la série\*, notée  $Q(p)$ , appelée quantile d'ordre  $p$ .  $Q(p)$  est la plus petite valeur de la série pour laquelle la proportion des observations inférieures ou égales à  $Q(p)$  est au moins égale à  $p$ .

\* Convention : dans le cas d'une série discrète comportant un nombre pair d'observations, le quantile d'ordre 0,50 sera pris égal à la médiane.

La proportion d'observations inférieures ou égales à  $Q(p)$  est au moins égale à  $p$  et la proportion d'observations supérieures ou égales à  $Q(p)$  est au moins égale à  $(1 - p)$ .

En plus de la médiane, fréquemment utilisée, nous présentons ici les quantiles les plus courants :

- les trois quartiles partagent la série en quatre groupes comprenant chacun 25 % des observations ;
- les neuf déciles partagent la série en dix groupes comprenant chacun 10 % des observations ;
- les quatre-vingt-dix-neuf centiles partagent la série en cent groupes comprenant chacun 1 % des observations.

## Les quartiles

### Définition

Les **quartiles** partagent la population ou l'échantillon en quatre groupes comprenant chacun 25 % des observations.

Au nombre de trois, ils se notent  $Q_1$ ,  $Q_2$  et  $Q_3$ .

- $Q_1$  est le quantile d'ordre 0,25 : au moins 25 % des observations sont inférieures ou égales à  $Q_1$  et au moins 75 % supérieures ou égales à  $Q_1$ .
- $Q_2$  est le quantile d'ordre 0,50 : au moins 50 % des observations sont inférieures ou égales à  $Q_2$  et au moins 50 % supérieures ou égales à  $Q_2$ ;  $Q_2$  est égal à la médiane.
- $Q_3$  est le quantile d'ordre 0,75 : au moins 75 % des observations sont inférieures ou égales à  $Q_3$  et au moins 25 % supérieures ou égales à  $Q_3$ .
- Dans le cas continu, on se réfère à la fonction de répartition :  $F(Q_1) = 0,25$  ;  $F(Q_2) = 0,5$  et  $F(Q_3) = 0,75$ .

La détermination des quartiles se fait comme pour la médiane, avec une interpolation linéaire dans le cas continu, les quartiles pouvant être déterminés grâce au polygone des fréquences ou des effectifs cumulés croissants.

### Exemple 2.9

#### Calcul d'un quartile dans un tableau statistique contenant une variable continue

Reprendons l'exemple 2.8, traité pour la médiane, concernant le niveau d'études des personnes de plus de 15 ans, et déterminons  $Q_1$ . Après avoir calculé  $\frac{n}{4} = 2\,172\,789,75$ , nous en déduisons que  $Q_1$  appartient à la classe des 30-40 ans. Il reste à effectuer l'interpolation linéaire qui donne :

$$\frac{4\,653\,220 - 2\,164\,808}{40 - 30} = \frac{2\,172\,789,75 - 2\,164\,808}{Q_1 - 30}, \text{ soit } Q_1 = 30,03 \text{ ans, ce qui signifie que}$$

25 % de cette population a un âge inférieur ou égal à 30,03 ans.

## Les déciles

### Définition

Les **déciles** partagent la population ou l'échantillon en dix groupes comprenant chacun 10 % des observations.

Au nombre de neuf, ils se notent :  $D_1$ ,  $D_2$ ,  $D_3$ ,  $D_4$ ,  $D_5$ ,  $D_6$ ,  $D_7$ ,  $D_8$  et  $D_9$ .

- $D_1$  est le quantile d'ordre 0,10 : au moins 10 % des observations sont inférieures ou égales à  $D_1$  et au moins 90 % des observations sont supérieures ou égales à  $D_1$ .
- $D_2$  est le quantile d'ordre 0,20 : au moins 20 % des observations sont inférieures ou égales à  $D_2$  et au moins 80 % des observations sont supérieures ou égales à  $D_2$ .
- $D_9$  est le quantile d'ordre 0,90 : au moins 90 % des observations sont inférieures ou égales à  $D_9$  et au moins 10 % des observations sont supérieures ou égales à  $D_9$ .

Dans le cas continu, on se réfère à la fonction de répartition :  $F(D_1) = 0,1$  ;  $F(D_2) = 0,2$  ; ... ;  $F(D_9) = 0,9$ .

La détermination des déciles est faite selon le même processus que celui utilisé pour les quartiles.

## Les centiles

### Définition

Les **centiles** partagent la population ou l'échantillon en cent groupes comprenant chacun 1 % des observations.

Au nombre de quatre-vingt-dix-neuf, ils se notent :  $C_1, C_2, \dots, C_{99}$ .

- $C_1$  est le quantile d'ordre 0,01 : au moins 1 % des observations sont inférieures ou égales à  $C_1$  et au moins 99 % des observations sont supérieures ou égales à  $C_1$ .
- $C_{99}$  est le quantile d'ordre 0,99 : au moins 99 % des observations sont inférieures ou égales à  $C_{99}$  et au moins 1 % des observations sont supérieures ou égales à  $C_{99}$ .
- Dans le cas continu :  $F(C_1) = 0,01 ; F(C_2) = 0,02 ; \dots ; F(C_{99}) = 0,99$ .

La détermination des centiles est faite selon le même processus que celui utilisé pour les quartiles.

### Focus 2.3

#### Positions relatives de la moyenne arithmétique, du mode et de la médiane

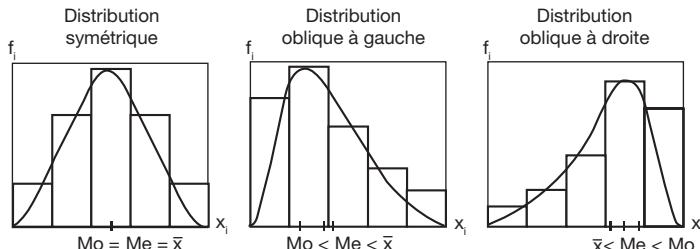
La moyenne arithmétique, le mode et la médiane sont trois paramètres de position qui permettent de préciser la forme de la distribution (voir figure 2.8) :

- Lorsque le diagramme de la distribution est symétrique, ces trois paramètres de position sont confondus, comme dans le cas d'une distribution probabiliste normale ou gaussienne (voir chapitre 4, section 1) où la valeur centrale s'impose.
- Lorsque la distribution est asymétrique, ou oblique, le mode est par définition au sommet de la courbe des fréquences. La moyenne, comme un centre d'inertie, attirée par les termes extrêmes, se déplace vers la zone où la courbe est le plus étirée. La médiane reste située entre ces deux valeurs. Ce type de graphique évoque notamment la distribution binomiale. Dans ce cas le choix d'une valeur centrale est beaucoup moins évident et exige de la circonspection.

Nous n'oublierons jamais en statistiques que l'interprétation et la présentation des calculs exigent une grande honnêteté intellectuelle : « Les chiffres sont des innocents, qui, sous la sollicitation, sous la torture, avouent très vite ce qu'on leur demande, quitte à se rétracter plus tard » (Alfred Sauvy<sup>1</sup>).

**Figure 2.8**

**Histogramme, densité de probabilité et valeurs centrales.**



1. Alfred Sauvy (1898-1990), économiste et sociologue français, fut directeur de l'INED (Institut national d'études démographiques).

Karl Pearson<sup>1</sup> a introduit, à la fin du XIX<sup>e</sup> siècle, la relation empirique suivante :  $Me - Mo = 2 \times (\bar{x} - Me)$ . Elle est valable pour les distributions unimodales, pas trop asymétriques, et permet une estimation rapide d'un paramètre à partir des deux autres.

## Focus 2.4

### Les fonctions Excel

**Pour faire la moyenne arithmétique d'une variable** : appelez la fonction MOYENNE dans la cellule où vous souhaitez faire apparaître le résultat. Puis, à l'aide de votre curseur, sélectionnez les valeurs dans la liste d'arguments. Cette fonction permet d'effectuer uniquement une moyenne simple.

**Pour faire la moyenne harmonique d'une variable** : appelez la fonction MOYENNE.HARMONIQUE dans la cellule où vous souhaitez faire apparaître le résultat. Puis, à l'aide de votre curseur, sélectionnez les valeurs dans la liste d'arguments. Cette fonction permet d'effectuer uniquement une moyenne simple.

**Pour faire la moyenne géométrique d'une variable** : appelez la fonction MOYENNE.GEOMETRIQUE dans la cellule où vous souhaitez faire apparaître le résultat. Puis, à l'aide de votre curseur, sélectionnez les valeurs dans la liste d'arguments. Cette fonction permet d'effectuer uniquement une moyenne simple.

## Focus 2.5

### Les fonctions de la calculatrice

**Avant tout calcul statistique, n'oubliez pas d'effacer les listes** : appuyez sur la touche **STAT**, puis **ClrList L1,L2,L3**, et validez avec **ENTER**. La calculatrice indique alors « done », pour signifier qu'elle a effacé ces trois listes.

**Pour effectuer la moyenne arithmétique simple d'une variable** : saisissez les modalités dans la colonne L1 du tableau. Appuyez sur la touche **STAT**, puis, dans le menu CALC,appelez la fonction 1-Var Stats. Validez avec **ENTER**. La moyenne s'affiche sur l'écran, entre autres résultats.

**Pour effectuer une moyenne pondérée** : saisissez les modalités dans la colonne L1, saisissez les effectifs ou les fréquences dans la colonne L2, puis, dans le menu CALC,appelez la fonction 1-Var Stats, puis indiquez dans l'ordre L1, L2, et validez avec **ENTER**.

**Pour effectuer le produit de deux colonnes** de même dimension, L1 et L2 : mettez par exemple la colonne L3 en surbrillance et tapez **L1×L2**. Le produit des deux colonnes s'affiche dans la colonne L3.

**Pour effectuer le produit des éléments d'une colonne** : placez le curseur dans la cellule L1 (1) (première cellule de la liste i où vous souhaitez faire apparaître le produit). Appuyez sur les touches **2ND** et **LIST**, puis, dans le menu MATH,appelez la fonction **Prod(**. Indiquez la colonne L1 dont vous souhaitez calculer le produit des éléments, fermez la parenthèse et validez avec **ENTER**.

1. Karl Pearson (1857-1936), mathématicien, statisticien anglais, fondateur avec Galton de la revue *Biometrika*.

**Pour calculer la racine  $n^{\text{ième}}$  d'une cellule :** placez le curseur dans la cellule L1 (1) où vous souhaitez faire apparaître la racine  $n^{\text{ième}}$ . Appuyez sur la touche **MATH**,appelez la fonction  $\sqrt[n]{\phantom{x}}$ . Indiquez la cellule dont vous souhaitez calculer la racine  $n^{\text{ième}}$  et validez avec **ENTER**.

Vous pouvez également utiliser la propriété  $\sqrt[n]{x} = x^{\frac{1}{n}}$  et vous ramener à un calcul de puissance, en tapant  $x^{(1/n)}$ .

---

## Conclusion

Nous voyons ainsi que nous serons amenés à faire de nombreux calculs de valeurs centrales pour analyser une série statistique. Nous devrons choisir parmi ces valeurs celles qui par leurs qualités correspondent au contexte de l'étude.

La moyenne arithmétique est généralement pertinente si la série est suffisamment longue et homogène. Elle varie peu d'un échantillon à l'autre. La médiane est très simple à calculer, mais est plus sensible aux fluctuations d'échantillonnage. Elle participe bien à la description de la série et élimine l'effet des valeurs aberrantes. Le mode a un but pratique évident : il indique la valeur la plus typique. Par ailleurs, il est incontournable pour les séries asymétriques. Ces paramètres qui participent à une description synthétique de la série doivent toujours être visualisés sur les différentes représentations graphiques. Nous reviendrons dans le chapitre suivant sur l'importance des quartiles et leur rôle dans la représentation graphique des séries par des boîtes à moustaches.

# Problèmes et exercices

La mise en œuvre des caractéristiques de tendance centrale diffère selon la nature des données.

- Les exercices 1, 2 et 3 proposent la détermination de caractéristiques de tendance centrale pour des variables de diverse nature.
- L'exercice 4 fait appel à une approche graphique des caractéristiques de tendance centrale.
- Les exercices 5 et 6 approfondissent la notion de moyenne, grâce aux moyennes géométriques et harmoniques.

## EXERCICE 1 LECTURE DE TENDANCES CENTRALES SUR SÉRIE BRUTE

### Énoncé

La liste ci-après est composée des vingt-cinq pays de l'Union européenne. Les nombres entre parenthèses indiquent le nombre de médecins pour 100 000 habitants :

Allemagne (350) ; Autriche (300) ; Belgique (400) ; Chypre (270) ; Danemark (340) ; Espagne (440) ; Estonie (310) ; Finlande (310) ; France (300) ; Grèce (390) ; Hongrie (360) ; Irlande (230) ; Italie (570) ; Lettonie (310) ; Lituanie (390) ; Luxembourg (250) ; Malte (260) ; Pays-Bas (250) ; Pologne (230) ; Portugal (310) ; République tchèque (310) ; Royaume-Uni (160) ; Slovaquie (320) ; Slovénie (220) ; Suède (310).

Source : PNUD, Rapport mondial sur le développement humain, 2003

1. Déterminez le mode de cette série.
2. Déterminez la médiane.

### Solution

1. On classe le nombre de médecins pour 100 000 habitants par ordre croissant :

160 ; 220 ; 230 ; 230 ; 250 ; 250 ; 260 ; 270 ; 300 ; 300 ; 310 ; 310 ; 310 ; 310 ; 310 ; 320 ; 340 ; 350 ; 360 ; 390 ; 390 ; 400 ; 440 ; 570.

**Mo = 310.** Le mode est la valeur la plus représentée, soit 310 médecins pour 100 000 habitants, valeur observée dans 6 pays.

2. L'effectif total  $n$  est impair, avec ici  $n = 25$ . Or,  $n = 2p + 1$ , donc  $p = 12$ . La valeur centrale est la  $(p + 1)^{\text{ème}}$  observation, soit la 13e. Il s'agit de 310. Donc **Me = 310**. Le nombre médian de médecins pour 100 000 habitants est 310. Douze pays, soit la moitié, ont moins de 310 médecins pour 100 000 habitants et 12 pays, soit l'autre moitié, ont plus de 310 médecins pour 100 000 habitants.



## EXERCICE 2 TENDANCES CENTRALES SUR TABLEAU STATISTIQUE, CARACTÈRE DISCRET

### Énoncé

Le tableau ci-après recense le nombre de résidences principales en France, selon le nombre de pièces :

Nombre de pièces	Nombre de résidences principales
1	1 526 573
2	3 028 244
3	5 299 675
4	6 418 808
5	4 432 943
6	3 103 918

Source : Insee, recensement de la population, 1999

1. Déterminez le mode.
2. Déterminez la médiane.
3. Déterminez les quartiles.
4. Calculez la moyenne.

### Solution

1. **Mo = 4.** Ce sont les résidences principales de 4 pièces qui sont le plus fréquentes, avec un effectif de 6 418 808.

2. Nous cherchons le nombre de pièces en dessous duquel se trouvent 50 % des résidences principales. Nous calculons donc les effectifs cumulés croissants, selon les étapes suivantes, sous Excel (voir figure 2.9) : l'effectif total ( $n$ ) en cellule B8, les fréquences ( $f_i$ ) en colonne C, puis les fréquences cumulées croissantes ( $f_{cc}$ ) en colonne D.

L'effectif total est impair, donc la médiane est l'observation centrale, de rang  $(p + 1)$ , avec  $p = \frac{23\,810\,160}{2}$ .

Figure 2.9

Résultats sous Excel.

	A	B	C	D	E
1	$x_i$	$n_i$	$f_i$	$f_{cc}$	$n_i x_i$
2	1	1 526 573	6,41%	6,41%	1 526 573
3	2	3 028 244	12,72%	19,13%	6 056 488
4	3	5 299 675	22,26%	41,39%	15 899 025
5	4	6 418 808	26,96%	68,35%	25 675 232
6	5	4 432 943	18,62%	86,96%	22 164 715
7	6	3 103 918	13,04%	100,00%	18 623 508
8	<b>Somme</b>	23 810 161	100%		89 945 541

Ou encore : à partir de la colonne des fréquences cumulées croissantes ( $f_{cc}$ ), nous lisons que 41 % des résidences principales ont 3 pièces et moins ; 68 % des résidences principale

pales ont 4 pièces et moins. Donc, entre ces deux valeurs, 50 % des résidences principales ont moins de 4 pièces. Soit **Me = 4**.

3. À partir du tableau utilisé pour la médiane, il est possible de déterminer que :

- **Q<sub>1</sub> = 3** : 19 % des résidences principales ont 2 pièces et moins ; 41 % des résidences principales ont 3 pièces et moins. Donc, entre ces deux valeurs, 25 % des résidences principales ont moins de 3 pièces.
- **Q<sub>2</sub> = 4**, car Q<sub>2</sub> = Me.
- **Q<sub>3</sub> = 5** : 68 % des résidences principales ont 4 pièces et moins ; 87 % des résidences principales ont 5 pièces et moins. Donc, entre ces deux valeurs, 75 % des résidences principales ont moins de 5 pièces.

4. À la suite du tableau précédent, nous calculons les  $n_i x_i$  en colonne E puis leur somme en cellule E8, sous Excel (voir figure 2.10).

La moyenne est égale à  $\bar{x} = \frac{1}{n} \sum_{i=1}^6 n_i x_i = \frac{89\,945\,541}{23\,810\,161}$ , soit  $\bar{x} = 3,78$  pièces. La moyenne du nombre de pièces dans les résidences principales est de 3,78.



### EXERCICE 3 TENDANCES CENTRALES SUR TABLEAU STATISTIQUE, CARACTÈRE CONTINU

#### Énoncé

Le tableau ci-après indique la structure des entrées dans les salles de cinéma en France, selon les tranches d'âge des spectateurs de moins de 25 ans :

Âge	Nombre d'entrées (millions)
[5 ; 10[	7,632
[10 ; 15[	12,316
[15 ; 20[	26,192
[20 ; 25[	24,631

Source : CNC, 2005

1. Calculez le mode.
2. Calculez la médiane.
3. Calculez les quartiles.
4. Calculez les déciles :
  - a. Calculez D<sub>1</sub>.
  - b. Calculez D<sub>9</sub>.
5. Calculez les centiles :
  - a. Calculez C<sub>1</sub>.
  - b. Calculez C<sub>99</sub>.
6. Calculez la moyenne.

**Solution**

1. Nous vérifions en premier lieu que les amplitudes de classes sont égales, ici de valeur 5. Il n'est donc pas nécessaire de corriger les effectifs en passant par les densités. La classe modale est celle de plus grand effectif, soit la classe [15 ; 20[.

$$\text{Le mode est donc égal à } Mo = \frac{k_2 x_1 + k_1 x_2}{k_1 + k_2} = \frac{(26,192 - 12,316) \times 20 + (26,192 - 24,631) \times 15}{(26,192 - 12,316) + (26,192 - 24,631)},$$

soit **Mo = 19,49**. L'âge modal de la population étudiée est 19,49 ans, soit 19 ans et 6 mois.

2. La première étape consiste à calculer les centres de classes ( $x_i$ ).

$$x_1 = \frac{5+10}{2} = 7,5; x_2 = \frac{10+15}{2} = 12,5; x_3 = \frac{15+20}{2} = 17,5; x_4 = \frac{20+25}{2} = 22,5.$$

Saisissez les centres de classes ( $x_i$ ) dans la colonne L1 de la calculatrice et les effectifs ( $n_i$ ) dans la colonne L2 (voir figure 2.10).

**Figure 2.10**

**Saisie du tableau de données avec la calculatrice.**

L1	L2	L3	L4
7,5	7,632	-----	
12,5	12,316		
17,5	26,192		
22,5	24,631		
-----			
			L2(5) =

Pour calculer les fréquences ( $f_i$ ) dans la colonne L3, placez le curseur sur l'en-tête de colonne L3. Indiquez  $L3=L2/\text{sum}(L2)$ , en appelant la fonction SUM (voir chapitre 1, annexe 1.2). Puis appuyez sur **ENTER**. La colonne L3 fait alors apparaître les fréquences.

Pour obtenir les fréquences cumulées croissantes ( $f_{cc}$ ) dans la colonne L4 (voir figure 2.11a), placez le curseur sur l'en-tête de colonne L4, puis entrez la formule  $L4=\text{CumSum}(L3)$ , en appelant la fonction CUMSUM (voir chapitre 1, annexe 1.2), puis appuyez sur **ENTER**.

**Figure 2.11a**

**Calcul des fréquences et des fréquences cumulées croissantes avec la calculatrice**

L3	L4	L4
.10784	.10784	
.17403	.28187	
.3701	.65196	
.34804	1	
-----	-----	

28,2 % des entrées sont faites par les moins de 15 ans ; 65,2 % des entrées sont faites par les moins de 20 ans. Donc la médiane appartient à la classe [15 ; 20[.

Par interpolation linéaire,  $Me = \frac{0,5 - 0,28187}{0,65196 - 0,28187} \times (20 - 15) + 15$  ; soit **Me = 17,95**. La moitié de la population étudiée a moins de 17,95 ans, soit environ 17 ans et 11 mois.

3. 10,8 % des entrées sont faites par les moins de 10 ans ; 28,2 % des entrées sont faites par les moins de 15 ans. Donc  $Q_1$  appartient à la classe [10 ; 15[.

Par interpolation linéaire,  $Q_1 = \frac{0,25 - 0,10784}{0,28187 - 0,10784} \times (15 - 10) + 10$  ; soit  $Q_1 = 14,08$ . Un

quart de la population étudiée a moins de 14,08 ans, soit environ 14 ans et 1 mois.

$Q_2 = Me$ , donc  $Q_2 = 17,95$ . La moitié de la population étudiée a moins de 17,95 ans, soit environ 17 ans et 11 mois.

65,2 % des entrées sont faites par les moins de 20 ans ; 100 % des entrées sont faites par les moins de 25 ans. Donc  $Q_3$  appartient à la classe [20 ; 25[.

Par interpolation linéaire,  $Q_3 = \frac{0,75 - 0,65196}{1 - 0,65196} \times (25 - 20) + 20$  ; soit  $Q_3 = 21,41$ . Trois

quarts de la population étudiée ont moins de 21,41 ans, soit environ 21 ans et 5 mois.

4. a. 0 % des entrées sont faites par les moins de 5 ans ; 10,8 % des entrées sont faites par les moins de 10 ans. Donc  $D_1$  appartient à la classe [5 ; 10[.

Par interpolation linéaire,  $D_1 = \frac{0,1 - 0}{0,10784 - 0} \times (10 - 5) + 5$  ; soit  $D_1 = 9,64$ . 10 % de la

population étudiée a moins de 9,64 ans, soit environ 9 ans et 8 mois.

b. 65,2 % des entrées sont faites par les moins de 20 ans ; 100 % des entrées sont faites par les moins de 25 ans. Donc  $D_9$  appartient à la classe [20 ; 25[.

Par interpolation linéaire,  $D_9 = \frac{0,9 - 0,65196}{1 - 0,65196} \times (25 - 20) + 20$  ; soit  $D_9 = 23,56$ . 90 % de

la population étudiée a moins de 23,56 ans, soit environ 23 ans et 7 mois.

5. a. 0 % des entrées sont faites par les moins de 5 ans ; 10,8 % des entrées sont faites par les moins de 10 ans. Donc  $C_1$  appartient à la classe [5 ; 10[.

Par interpolation linéaire,  $C_1 = \frac{0,01 - 0}{0,10784 - 0} \times (10 - 5) + 5$  ; soit  $C_1 = 5,46$ . 1 % de la

population étudiée a moins de 5,46 ans, soit environ 5 ans et 5 mois.

b. 65,2 % des entrées sont faites par les moins de 20 ans ; 100 % des entrées sont faites par les moins de 25 ans. Donc  $C_{99}$  appartient à la classe [20 ; 25[.

Par interpolation linéaire,  $C_{99} = \frac{0,99 - 0,65196}{1 - 0,65196} \times (25 - 20) + 20$  ; soit  $C_{99} = 24,86$ . 99 % de

la population étudiée a moins de 24,86 ans, soit environ 24 ans et 10 mois.

6. Pour calculer les  $n_i x_i$  dans la colonne L5, placez le curseur sur l'en-tête de colonne L5. Indiquez  $L5=L2\times L1$ . Puis appuyez sur **ENTER**. La colonne L5 fait alors apparaître les  $n_i x_i$  (voir figure 2.11b).

Pour en faire la somme, placez le curseur sur la cellule L5(5), et indiquez  $L5(5)=sum(L5)$ , en appelant la fonction SUM (voir annexe 1.2). Puis appuyez sur **ENTER**. La cellule L5(5) fait alors apparaître la somme des  $n_i x_i$ .

Pour connaître l'effectif total, placez le curseur sur la cellule L2(5), et indiquez L2(5)=sum(L2), en appelant la fonction SUM. Puis appuyez sur **ENTER**. La cellule L2(5) fait alors apparaître la somme des  $n_i$ .

**Figure 2.11b**

**Calcul des  $n_i x_i$  et de la somme des colonnes avec la calculatrice.**

L2	L5	5
7.632	57.24	
12.316	153.95	
26.192	458.36	
24.631	554.2	
70.771	1223.7	
<b>L2(6) =</b>		

La moyenne est donc égale à  $\bar{x} = \frac{1}{n} \sum_{i=1}^4 n_i x_i = \frac{1223,7}{70,771}$ , soit  $\bar{x} = 17,29$ . L'âge moyen de la population étudiée est 17,29 ans, soit environ 17 ans et 3 mois.

## EXERCICE 4 VISUALISATION GRAPHIQUE DES TENDANCES CENTRALES

### Énoncé

À partir des données de l'exercice 3 :

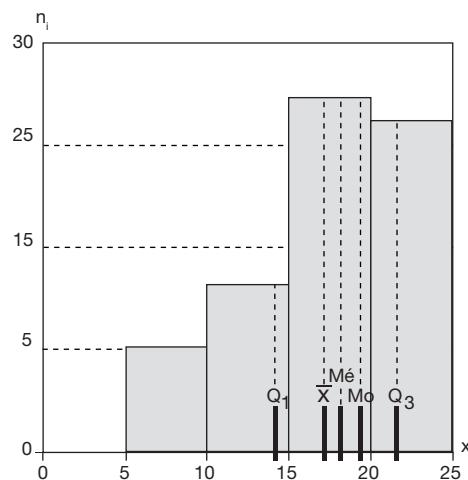
1. Dessinez un histogramme. Positionnez sur cet histogramme le mode, la médiane, les quartiles et la moyenne.
2. Retrouvez la valeur de la médiane à l'aide des polygones des effectifs cumulés.

### Solution

1. Les amplitudes de classes sont toutes identiques. Il est donc inutile de passer par les densités des effectifs afin de respecter le rapport entre l'aire du rectangle et sa hauteur (voir figure 2.12).

**Figure 2.12**

**Histogramme des entrées cinématographiques par âges et tendances centrales.**



2. La médiane se trouve à l'intersection des polygones des effectifs cumulés croissants et décroissants. Afin de pouvoir tracer graphiquement ces polygones, il convient de calculer les effectifs cumulés croissants  $n_{cc}$  en colonne D et les effectifs cumulés décroissants  $n_{cd}$  en colonne E (voir figure 2.13).

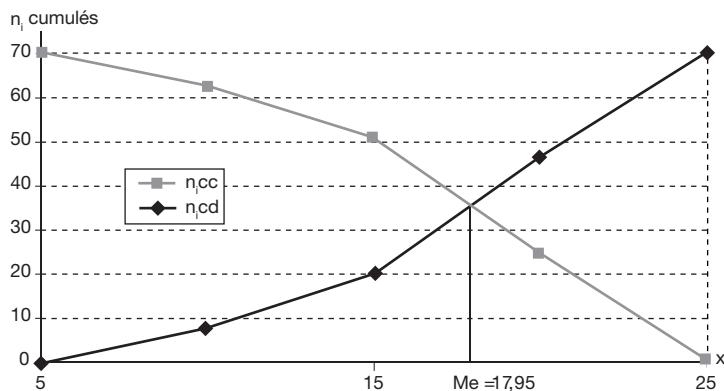
**Figure 2.13**

Résultats sous Excel.

	A	B	C	D	E
1	Age	$x_i$	$n_i$	$n_{cc}$	$n_{cd}$
2	[5;10[	7,50	7,632	7,632	70,771
3	[10;15[	12,50	12,316	19,948	63,139
4	[15;20[	17,50	26,192	46,140	50,823
5	[20;25]	22,50	24,631	70,771	24,631

**Figure 2.14**

Effectifs cumulés par âges des entrées cinématographiques.



## EXERCICE 5 MOYENNE GÉOMÉTRIQUE

### Énoncé

Un jeune diplômé est augmenté de 5 % la première et la deuxième année de sa vie professionnelle. La troisième année, son augmentation de salaire est de 3 %. Il change d'entreprise au début de la quatrième année, et négocie un salaire de 12 % plus élevé que celui qu'il avait.

Déterminez la moyenne de ses augmentations de salaire sur les quatre années.

### Solution

Pour une augmentation de  $x_i = 5\% = 0,05$ , la croissance se traduit par un coefficient multiplicateur de  $y_i = 1 + 0,05 = 1,05$ . Ainsi, nous savons que le coefficient multiplicateur moyen est la moyenne géométrique pondérée des coefficients multiplicateurs affectés des durées. Nous allons donc introduire la série des  $y_i = 1 + x_i$ .

Saisissez les  $y_i$  dans la colonne L1 de la calculatrice et les effectifs ( $n_i$ ) dans la colonne L2 (voir figure 2.15).

Pour calculer les  $y_i^{n_i}$  dans la colonne L3, placez le curseur sur l'en-tête de colonne L3. Indiquez L3=L1^L2, puis appuyez sur **ENTER**. La colonne L3 fait alors apparaître les  $y_i^{n_i}$ .

Pour en faire le produit, placez le curseur sur la cellule L3(4) et indiquez L3(4)=prod(L3), en appelant la fonction PROD. Puis appuyez sur **ENTER**. La cellule L3(4) fait alors apparaître le produit des  $y_i^{n_i}$ , soit 1,2718.

Pour faire la racine 4<sup>e</sup> du résultat, placez le curseur sur la cellule L3(5), et indiquez L3(5)= L3(4)<sup>(1 / 4)</sup>. Puis appuyez sur **ENTER**. La cellule L3(5) donne 1,062.

**Figure 2.15**

Saisie du tableau de données et calcul avec la calculatrice.

L1	L2	L3	3
1.05	2	1.1025	
1.03	1	1.03	
1.12	1	1.12	
-----		1.2718	
		1.062	
<b>L3(6) =</b>			

La moyenne géométrique est  $G = \sqrt[4]{\prod_{i=1}^3 y_i^{n_i}} = 1,062$ . L'augmentation moyenne du salaire sur les quatre années est 6,20 %.



## EXERCICE 6 MOYENNE HARMONIQUE

### Énoncé

Christophe Moreau est arrivé premier Français et 20<sup>e</sup> au classement du Tour de France 2005. Le tableau ci-après indique sa vitesse moyenne (km/h) sur chaque étape, ainsi que la distance de l'étape (km).

Jour	Étape	Vitesse moyenne (km/h)	Distance de l'étape (km)
Mardi 19 juillet 2005	Mourenx > Pau	38,40	180,5
Mercredi 20 juillet 2005	Pau > Revel	39,48	239,5
Jeudi 21 juillet 2005	Albi > Mende	39,10	189
Vendredi 22 juillet 2005	Issoire > Le Puy-en-Velay	42,44	153,5
Samedi 23 juillet 2005	Saint-Étienne > Saint-Étienne	44,40	55
Dimanche 24 juillet 2005	Corbeil-Essonnes > Paris Champs-Élysées	39,23	144

Source : <http://www.letour.fr/2005>

Calculez la vitesse moyenne de Christophe Moreau sur la dernière semaine du Tour de France 2005.

### Solution

Si  $H$  désigne la vitesse moyenne, alors :  $H = \frac{d}{t} = \frac{\sum_{i=1}^r n_i}{\sum_{i=1}^r \frac{n_i}{x_i}}$ .

Nous cherchons donc la moyenne harmonique des vitesses ( $x_i$ ), chaque vitesse ayant pour poids la distance de l'étape ( $n_i$ ). Nous calculons les  $n_i / x_i$  en colonne E puis leur somme en cellule E8, sous Excel (voir figure 2.16).

**Figure 2.16**

Résultats sous Excel.

	A Jour	B Etape	C $x_i$	D $n_i$	E $n_i / x_i$
1					
2	Mardi 19 juillet 2005	Mourenx > Pau	38,40	180,5	4,70
3	Mercredi 20 juillet 2005	Pau > Revel	39,48	239,5	6,07
4	Jeudi 21 juillet 2005	Albi > Mende	39,10	189,0	4,83
5	Vendredi 22 juillet 2005	Issoire > Le Puy-en-Velay	42,44	153,5	3,62
6	Samedi 23 juillet 2005	Saint-Etienne > Saint-Etienne	44,40	55,0	1,24
7	Dimanche 24 juillet 2005	Corbeil-Essonnes > Paris Champs-Élysées	39,23	144,0	3,67
8	<b>Somme</b>			962	24,13

$$\text{La moyenne harmonique est : } H = \frac{\sum_{i=1}^6 n_i}{\sum_{i=1}^6 \frac{n_i}{x_i}} = \frac{962}{24,13}, \text{ soit } H = 39,85 \text{ km/h.}$$

moyenne de Christophe Moreau sur la dernière semaine du Tour de France 2005 est 39,85 km/h.

## Bibliographie

ANTOINE C., *Les moyennes*, Que sais-je ?, PUF, 1998.

CHAREILLE P. et PINAULT Y., *Statistique descriptive*, Collection AES, Montchrestien, Paris, 1996.

DELMAS B., *Statistique descriptive*, Armand Colin, 2005.

DROESBEKE J.-J., *Éléments de statistiques*, Éditions de l'université de Bruxelles, Ellipses, 2001.

LIORZOU A., *Initiation à la pratique statistique*, Eyrolles, 1985.

PIATIER A., *Statistique descriptive et initiation à l'analyse*, Thémis, PUF, 1962.

ROGER P., *Probabilités, statistique et processus stochastiques*, Collection Synthex, Pearson Education, 2004.

SCHLACHTHER D., *De l'analyse à la prévision*, Ellipses, 1986.

YULE G., *An Introduction to the Theory of Statistics*, Griffin, 1911.

# Les caractéristiques de dispersion

- 1. Les caractéristiques simples ... 64
- 2. Variance et écart-type ..... 66

## Problèmes et exercices

- 1. Caractéristiques simples de dispersion ..... 73
- 2. Boîte à moustaches..... 75
- 3. Variance et écart-type sur caractère discret ..... 76
- 4. Comparaison de distributions sur caractère continu ..... 77
- 5. Manipulations de formules.... 79

Dans son incontournable livre *Le jeu de la science et du hasard*, Daniel Schwartz<sup>1</sup> raconte cette anecdote : « Les mauvaises langues prétendent qu'un statisticien se noya dans un cours d'eau dont la profondeur moyenne était de 20 cm. C'est qu'à l'endroit où il souhaitait patauger, elle atteignait 2 m. »

Dans le chapitre 2, nous avons vu comment une série statistique pouvait être résumée par ses caractéristiques de position. Cependant, ces dernières ne renseignent pas sur la structure interne de la distribution, sur la variabilité de la série autour de sa moyenne. C'est pourquoi il convient de compléter ce travail en introduisant les caractéristiques de dispersion.

Nous en étudierons cinq : l'étendue, les intervalles interquantiles, l'écart absolu moyen, l'écart-type (lié à la variance) et le coefficient de variation.

---

<sup>1</sup>. Daniel Schwartz, polytechnicien, est le fondateur du Centre d'enseignement de la statistique appliquée à la médecine (CESAM). Il a été le pionnier de l'introduction de la statistique dans la médecine en France.

# 1 Les caractéristiques simples

L'étendue, les intervalles interquartiles et l'écart absolu moyen sont qualifiés de simples, car ces caractéristiques restent limitées dans leur construction et leur utilisation, au regard de la notion de variance (exposée dans la seconde partie de ce chapitre).

## 1.1 L'ÉTENDUE

La première mesure de la dispersion d'une distribution est l'étendue. Cette mesure est la plus simple des caractéristiques de dispersion ; dans le langage courant, on parle d'éventail, ou de fourchette, ou d'intervalle de variation de la série.

### Définition

L'**étendue** d'une série est la différence entre la plus grande et la plus petite valeur observée.

Elle est notée :  $E = \text{Max}(x_i) - \text{Min}(x_i)$ .

L'étendue permet une approche aisée de la dispersion d'une variable, mais sa signification reste très limitée, car elle ne prend en compte que les deux valeurs extrêmes de la série. Or, ces valeurs extrêmes peuvent être mal connues, voire aberrantes ou erronées.

Par ailleurs, l'étendue n'est pas indépendante de l'effectif observé et peut donner une vision faussée de la dispersion.

Enfin, dans le cas de séries continues, l'étendue n'est pas connue avec exactitude, puisque la perte d'information due au regroupement en classes ne permet pas de connaître les valeurs minimales et maximales réellement prises par la variable.

## 1.2 LES INTERVALLES ET ÉCARTS INTERQUANTILES

### Définitions

Il existe trois intervalles et écarts interquartiles :

- L'intervalle interquartile  $[Q_1 ; Q_3]$  représente la zone centrale de la population comprenant 50 % de la série ; l'amplitude de cet intervalle est appelée **écart interquartile** et on note :  $EIQ = Q_3 - Q_1$ .
- L'intervalle interdécile  $[D_1 ; D_9]$  représente la zone centrale de la population comprenant 80 % de la série ; l'amplitude de cet intervalle est appelée **écart interdécile** et on note :  $EID = D_9 - D_1$ .
- L'intervalle intercentile  $[C_1 ; C_{99}]$  représente la zone centrale de la population comprenant 98 % de la série ; l'amplitude de cet intervalle est appelée **écart intercentile** et on note :  $EIC = C_{99} - C_1$ .

### Exemple 3.1

#### Calcul de l'écart interquartile

Reprendons l'exemple 2.8 du chapitre précédent concernant le niveau d'études supérieures des personnes de plus de 15 ans. Dans cet exemple,  $Q_1 = 30,03$  ans. En procédant au calcul de  $Q_3$ , nous trouvons  $Q_3 = 52,18$  ans. Ainsi,  $EIQ = 52,18 - 30,03 = 22,15$  ans, soit environ 22 ans et 2 mois.

Par rapport à l'étendue, l'écart interquartile présente l'avantage d'écarter les valeurs extrêmes, mais l'inconvénient de laisser de côté 50 % des données. C'est pourquoi on préfère habituellement l'intervalle interdécile,  $EID = D_9 - D_1$ , qui comprend 80 % de la population.

### 1.3 LA BOÎTE À MOUSTACHES (BOX PLOT)

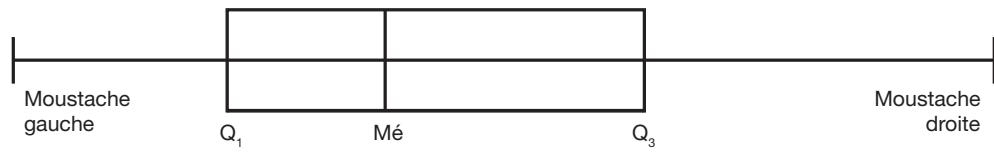
La boîte à moustaches est souvent appelée « *box plot* » dans les logiciels statistiques.

#### Définition

Les quantiles permettent une représentation de la distribution statistique par le **diagramme de Tukey**<sup>1</sup>, ou **boîte à moustaches**. Il s'agit d'une boîte délimitée par les quartiles  $Q_1$  et  $Q_3$ , coupée en deux parties par la médiane et prolongée de chaque côté par des moustaches (voir figure 3.1).

Figure 3.1

Schéma de la boîte à moustaches ou diagramme de Tukey.



Il existe plusieurs conventions permettant de fixer la valeur des moustaches :

- **Termes extrêmes** : la méthode classique consiste à démarrer la moustache de gauche à la plus petite des valeurs,  $\text{Min}(x_i)$ , et à finir celle de droite par  $\text{Max}(x_i)$ . Dans ce premier cas, si la série a des valeurs extrêmes isolées, les moustaches de la série seront très longues et fausseront l'interprétation.
- **Moustaches limitées à  $1,5 \times \text{EIQ}$**  : pour éviter le problème évoqué ci-dessus, un calcul permet de limiter la taille des moustaches à une fois et demie l'écart interquartile. La moustache de gauche est égale à la plus grande des valeurs entre  $\text{Min}(x_i)$  et  $Q_1 - 1,5 \times (Q_3 - Q_1)$ . La moustache de droite est composée de la plus petite des valeurs entre  $\text{Max}(x_i)$  et  $Q_1 + 1,5 \times (Q_3 - Q_1)$ .
- **Centiles** : une méthode simple consiste à utiliser les centiles pour fixer la valeur des moustaches. Le centile  $C_{10}$  est utilisé pour la moustache de gauche, et le centile  $C_{90}$  pour la moustache de droite.

La boîte à moustaches permet une bonne visualisation de la zone centrale de la série et de la dispersion. Ce diagramme est extrêmement précieux pour comparer diverses séries statistiques.

1. John Wilder Tukey (1915-2000) : mathématicien et statisticien, il fut le premier directeur du département statistique de l'université de Princeton.

## 1.4 L'ÉCART ABSOLU MOYEN

L'écart absolu moyen est le paramètre de dispersion le plus simple qui mesure les fluctuations de la série par rapport à la moyenne.

### Définition

L'**écart absolu moyen** de  $n$  observations est la moyenne arithmétique des valeurs absolues des écarts à la moyenne :  $e_a = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|$ .

L'**écart absolu moyen** de  $n$  observations, ordonnées dans un tableau statistique  $(x_i ; n_i)$ , présentant  $r$  modalités, est la moyenne arithmétique pondérée des valeurs absolues des écarts à la moyenne :  $e_a = \frac{1}{n} \sum_{i=1}^r n_i |x_i - \bar{x}|$ ,  $r$  désignant le nombre de modalités, avec  $n = \sum_{i=1}^r n_i$ .

La valeur absolue des écarts à la moyenne est utilisée afin d'empêcher que les écarts positifs ne se compensent avec les écarts négatifs. En effet, par cette compensation, la somme des écarts à la moyenne est nulle :  $\sum_{i=1}^r n_i (x_i - \bar{x}) = 0$ .

L'écart absolu moyen présente l'avantage de prendre en compte toutes les valeurs de la série. Il a été introduit par Laplace avant la variance et est utilisé notamment dans la méthode d'estimation L1, méthode alternative à la méthode des moindres carrés.

## 2 Variance et écart-type

### 2.1 PRÉSENTATION

L'écart-type ou écart quadratique moyen est de loin l'indicateur de dispersion le plus utilisé. L'introduction en 1893 de son nom anglais *standard deviation* est due à Karl Pearson, mathématicien, statisticien et philosophe. La variance, qui est le carré de l'écart-type, a été introduite en statistique par le statisticien et généticien anglais Ronald Fisher.

### Définitions

L'**écart-type**, noté  $\sigma_x$ , est la racine carrée de la variance.

Dans le cas de  $n$  observations, la variance est donnée par : 
$$\begin{cases} V(x) = \sigma_x^2 = \frac{1}{n} \sum (x_i - \bar{x})^2 \\ \sigma(x) = \sqrt{\frac{1}{n} \sum (x_i - \bar{x})^2} \end{cases}$$

Dans le cas de  $n$  observations, ordonnées dans un tableau statistique  $(x_i ; n_i)$ , présentant  $r$  modalités :

$$\begin{cases} V(x) = \sigma_x^2 = \frac{1}{n} \sum_{i=1}^{i=r} n_i (x_i - \bar{x})^2 \\ \sigma(x) = \sqrt{\frac{1}{n} \sum_{i=1}^{i=r} n_i (x_i - \bar{x})^2} \end{cases} .$$

La **variance** (ou fluctuation) est la moyenne arithmétique des carrés des écarts à la moyenne. Elle se note  $V(x)$ .

L'écart-type peut également se définir comme la moyenne quadratique des écarts à la moyenne.

### Exemple 3.2

#### Calcul de variance et d'écart-type

La série suivante donne le salaire minimal de croissance pour 169 heures de travail dans vingt pays d'Europe en 2006. La valeur du SMIC est indiquée entre parenthèses :

Belgique (1 234) ; Bulgarie (81,8) ; République tchèque (261,3) ; Estonie (191,7) ; Irlande (1 293) ; Grèce (667,7) ; Espagne (631) ; France (1 218) ; Lettonie (129,2) ; Lituanie (159,3) ; Luxembourg (1 503) ; Hongrie (247) ; Malte (580) ; Pays-Bas (1 273) ; Pologne (233,5) ; Portugal (450) ; Roumanie (90,2) ; Slovénie (511,9) ; Slovaquie (183,2) ; Royaume-Uni (1269).

Source : Eurostat, 2006

Calculons la variance et l'écart-type à l'aide d'Excel (voir figure 3.2).

**Figure 3.2**

**Calcul des  $(x_i - \bar{x})^2$  sous Excel.**

	A	B	C
	Pays	$x_i$	$n_i (x_i - \bar{x})^2$
1			
2	Belgique	1 234,00	388 889,43
3	Bulgarie	81,80	279 407,39
4	République tchèque	261,30	121 863,83
5	Estonie	191,70	175 301,32
6	Irlande	1 293,00	465 956,41
7	Grèce	667,70	3 284,44
8	Espagne	631,00	424,77
9	France	1 218,00	369 189,91
10	Lettonie	129,20	231 543,82
11	Lituanie	159,30	203 482,19
12	Luxembourg	1 503,00	796 752,61
13	Hongrie	247,00	132 052,29
14	Malte	580,00	923,55
15	Pays-Bas	1 273,00	439 052,01
16	Pologne	233,50	142 046,07
17	Portugal	450,00	25 724,95
18	Roumanie	90,20	270 597,64
19	Slovénie	511,90	9 700,28
20	Slovaquie	183,20	182 491,30
21	Royaume-Uni	1 269,00	433 767,13
22	<b>Somme</b>	12 207,80	4 672 451,34

Le calcul de la moyenne donne  $\bar{x} = \frac{1}{20} \sum_{i=1}^{20} x_i = \frac{12 207,8}{20} = 610,39$ . De là, après calcul de

chacun des écarts à cette moyenne, et leur élévation au carré,  $V(x) = \frac{1}{20} \sum_{i=1}^{20} (x_i - \bar{x})^2 = \frac{4 672 451,34}{20} = 233 622,57$ . La variance de la valeur du SMIC des différents pays européens est de 233 622,57. D'où l'écart-type  $\sigma_x = \sqrt{V(x)} = \sqrt{233 622,57} = 483,35$  €.

Afin de faciliter les différentes étapes de calcul de la variance, il est possible d'utiliser la formule développée de la variance. Cette formule est issue du théorème de Koenig.

## Définitions

### Formules développées de la variance :

- Cas de n observations :  $V(x) = \left( \frac{1}{n} \sum_{i=1}^n x_i^2 \right) - \bar{x}^2$ .
- Cas d'un tableau statistique avec r modalités :  $V(x) = \left( \frac{1}{n} \sum_{i=1}^r n_i x_i^2 \right) - \bar{x}^2$ .

Démonstration (dans le cas de n observations, ordonnées dans un tableau statistique  $(x_i; n_i)$ , comprenant r modalités) :

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

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

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

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

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

Par cette formule, la variance est égale à la moyenne des  $x_i$  au carré moins le carré de la moyenne des  $x_i$ . Le procédé mnémotechnique suivant est parfois utilisé :  $V(x) = MC - CM$ , soit **la variance est égale à la Moyenne des Carrés moins le Carré de la Moyenne**.

## Exemple 3.3

### Calcul de la variance par la formule développée

Reprendons les données de l'exemple 3.2 et calculons la variance avec la formule développée, à l'aide de la calculatrice : saisissez les valeurs du SMIC dans la colonne L1 du tableau (voir figure 3.3) en appuyant sur la touche **STAT** puis en éditant le tableau par appui sur la touche **1**.

**Figure 3.3**

Extrait de la saisie du tableau de données avec la calculatrice.

L1	L2	L3	1
1234	-----	-----	
81.8			
261.3			
191.7			
1293			
667.7			
131			
<b>L1(7)=631</b>			

Lorsque les vingt valeurs sont saisies, appuyez sur la touche **STAT**, puis, dans le menu **CALC**, appelez la fonction **1-Var Stats**. Validez avec **ENTER**. Les résultats présentés figure 3.4 s'affichent.

**Figure 3.4**

Résultats de l'analyse statistique effectuée avec la calculatrice.

### 1-Var Stats

$\bar{x}=610,39$   
 $\sum x=12207,8$   
 $\sum x^2=12123970,4$   
 $S_x=495,9016954$   
 $\sigma_x=483,345184$   
 $\downarrow n=20$

Lecture des résultats : on notera que la calculatrice désigne par  $\sum x$  et  $\sum x^2$  les sommes des valeurs ou de leurs carrés, que l'on ait affaire à n observations brutes ou à n observations ordonnées dans un tableau statistique. Par défaut, comme dans cet exemple, les  $n_i$  sont pris égaux à 1. Enfin, l'écart-type est  $\sigma_x = 483,35$ . Il ne faut pas le confondre avec  $S_x = 495,9$  appelé écart-type d'échantillon (supérieur à  $\sigma_x$ ), qui permet d'estimer l'écart-type d'une population à partir d'un échantillon de cette population (voir P. Roger, chapitre 5).

À partir de ces résultats il est possible de calculer directement la variance :

$$V(x) = \frac{1}{20} \sum_{i=1}^{20} x_i^2 - \bar{x}^2 = \frac{1}{20} \times 12123970,4 - 610,39^2 = 233\,622,57, \text{ soit la même valeur que}$$

par la formule classique de la variance, conformément à la démonstration du théorème de Koenig. Ce résultat peut également être obtenu en élevant l'écart-type au carré :

$$V(x) = \sigma_x^2 = 483,34^2 = 233\,622,57.$$

## 2.2 CAS D'UN CARACTÈRE CONTINU

Dans le cas d'un caractère continu, le calcul se fait en remplaçant chaque classe par sa valeur centrale,  $x_i$ . Cette méthode, dite du centre de classe, tend à augmenter l'écart-type, notamment dans le cas d'une distribution unimodale où les effectifs diminuent rapidement quand on s'écarte de la moyenne (distribution proche de la distribution normale). Une correction empirique, dite correction de Sheppard, est parfois utilisée.

## 2.3 PROPRIÉTÉS DE CALCUL DE LA VARIANCE ET DE L'ÉCART-TYPE

La variance et l'écart-type ne sont pas linéaires comme la moyenne, mais possèdent des propriétés très importantes.

### Propriétés

$V(x + a) = V(x)$ , donc  $\sigma(x + a) = \sigma(x)$  : ajouter une constante ne change pas la dispersion.

$V(ax) = a^2 V(x)$ , donc  $\sigma(ax) = |a| \times \sigma(x)$  : multiplier la série par un réel positif multiplie la variance par le carré de ce nombre et l'écart-type par la valeur absolue de ce nombre.

Démonstrations dans le cas de n observations, ordonnées dans un tableau statistique ( $x_i$ ;  $n_i$ ) :

$$V(x+a) = \frac{1}{n} \sum_{i=1}^r n_i [(x_i + a) - (\bar{x} + a)]^2$$

$$V(x+a) = \frac{1}{n} \sum_{i=1}^r n_i (x_i + a - \bar{x} - a)^2, \text{ avec la propriété de la moyenne } (\bar{x+a}) = \bar{x} + a$$

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

Soit  $V(x+a) = V(x)$

$$V(ax) = \frac{1}{n} \sum_{i=1}^r n_i (ax_i)^2 - (\bar{ax})^2$$

$$V(ax) = a^2 \frac{1}{n} \sum_{i=1}^r n_i x_i^2 - a^2 \bar{x}^2, \text{ avec la propriété de la moyenne } (\bar{ax}) = a\bar{x}$$

$$V(ax) = a^2 \left( \frac{1}{n} \sum_{i=1}^r n_i x_i^2 - \bar{x}^2 \right)$$

Soit  $V(ax) = a^2 V(x)$

#### Exemple 3.4

#### Applications des propriétés de la variance

Supposons que, dans une entreprise, le salaire moyen soit de 2 500 € avec un écart-type de 500 €.

- Ajout d'une constante** : si tous les salaires augmentent de 200 €, la moyenne augmente également de 200 €, mais l'écart-type reste constant. Autrement dit, la dispersion des salaires sera toujours mesurée par un écart-type de 500 € autour du salaire moyen de 2 700 €.
- Multiplication par une constante** : si tous les salaires augmentent de 5 %, le salaire moyen sera de  $2500 \times 1,05 = 2625$  € et l'écart-type deviendra :  $500 \times 1,05 = 525$  €.

#### Focus 3.1

#### L'écart-type

- L'écart-type est conforme à trois des conditions de Yule : il est défini de façon rigoureuse, il dépend de toutes les valeurs de la série et se prête bien aux calculs algébriques. Il a le défaut d'être sensible aux valeurs aberrantes, mais cette influence est limitée, les écarts exceptionnels étant pondérés par des effectifs faibles.
- On notera que l'écart-type, qui représente l'écart moyen d'une unité statistique à la moyenne, s'exprime dans les mêmes unités que la variable, ce qui n'est pas le cas de la variance (si la variable est une longueur exprimée en centimètres, la variance est exprimée en centimètres carrés).
- Population et échantillon : dans le cadre de la statistique inférentielle, on cherche à préciser les paramètres d'une population à partir d'un échantillon ; on rappelle (voir

exemple 3.4) que la calculatrice donne deux paramètres notés respectivement  $\sigma_x$  et  $S_x$  :  $\sigma_x$  désigne l'écart-type calculé sur les données considérées comme constituant la population et  $S_x$  une estimation ponctuelle de l'écart-type de la population, obtenue à partir d'un échantillon ( $S_x \geq \sigma_x$ ).

- Additivité des variances : en général, la variance ne possède pas la propriété d'additivité. « Les variances ne s'additionnent que si les éléments constituant la somme ou la différence sont prélevés au hasard » (voir A. Liorzou). On dit alors que les variables sont indépendantes, et dans ce cas on a alors :  $x, y$  étant des variables quantitatives indépendantes et  $z$  leur somme,  $V(z)=V(x)+V(y)$ , ce qui donne pour les écarts-types une relation de Pythagore :  $\sigma_z^2=\sigma_x^2+\sigma_y^2$  soit  $\sigma_z = \sqrt{\sigma_x^2 + \sigma_y^2}$ .
  - L'écart absolu moyen est toujours inférieur ou égal à l'écart-type.
- 

## 2.4 LE COEFFICIENT DE VARIATION

La comparaison directe de deux écarts-types peut donner une impression fausse concernant la dispersion des deux séries dont les valeurs des moyennes sont différentes. De plus, l'écart-type dépend de l'unité choisie. C'est pourquoi le coefficient de variation, qui mesure la dispersion relative à la moyenne, est utilisé pour comparer la dispersion de plusieurs séries.

### Définition

Le **coefficient de variation** est le rapport noté  $CV(x)$  et défini par :  $CV(x)=\frac{\sigma_x}{\bar{x}}$  ; ce coefficient s'exprime en pourcentage de la moyenne.

### Exemple 3.5

#### Écart-type et coefficient de variation

Considérons qu'à la suite d'une étude statistique portant sur le poids  $x$  des voyageurs et sur celui  $y$  des bagages, une compagnie aérienne ait obtenu les résultats suivants :

Paramètres	$x$	$y$
Moyenne	70 kg	15 kg
Écart-type	8 kg	6 kg

Pour la série des voyageurs  $CV(x)=\frac{8}{70}=0,1143$ , soit 11,43 %, et pour la série des bagages :  $CV(y)=\frac{6}{15}=0,40$ , soit 40 %.

Alors que l'écart-type de la série des voyageurs est plus grand que celui des bagages ( $\sigma_x > \sigma_y$ ), la série des poids des bagages est plus dispersée que celle des poids des voyageurs, car  $CV(y) > CV(x)$ .

Le coefficient de variation est un nombre sans dimension, indépendant de l'unité de mesure ; il permet de mesurer la dispersion de séries exprimées en unités ou ordres de grandeur différents. Il mesure l'homogénéité des données.

## Conclusion

Ce chapitre nous a enseigné que les valeurs centrales ne suffisent jamais à décrire une série statistique et que les paramètres de dispersion sont incontournables pour appréhender la structure interne de la série.

On notera le rôle prépondérant de la variance et de l'écart-type et on s'attachera à retenir leurs propriétés algébriques.

On retiendra que le coefficient de variation et la boîte à moustaches sont des outils extrêmement précieux dans le cadre de la comparaison des séries.

Enfin, ces paramètres vont nous permettre d'aller plus loin et de nous intéresser à la forme des distributions et notamment à la plus célèbre des lois de probabilité, la loi normale.

# Problèmes et exercices

Aux côtés des caractéristiques de tendance centrale, les caractéristiques de dispersion fournissent une seconde série d'indicateurs permettant de caractériser une distribution statistique.

- Les exercices 1, 2 et 3 mettent en œuvre le calcul des indicateurs de dispersion, ainsi que leur représentation graphique sous forme de boîte à moustaches.
- L'exercice 4 montre comment deux distributions peuvent être comparées au regard des caractéristiques de tendance centrale et de dispersion.
- L'exercice 5 permet une familiarisation avec les propriétés des caractéristiques de tendance centrale et de dispersion.



## EXERCICE 1 CARACTÉRISTIQUES SIMPLES DE DISPERSION

### Énoncé

Le tableau ci-après recense la population de la France métropolitaine par tranches d'âge en 2007 (données provisoires) :

Âge	Population
0-14 ans	11 275 845
15-24 ans	7 806 706
25-34 ans	8 022 951
35-44 ans	8 733 224
45-54 ans	8 428 982
55-64 ans	7 166 591
65-74 ans	4 929 936
75-112 ans	5 173 765

Source : Insee, recensement de la population, bilan démographique, 2007

1. Calculez l'étendue.
2. Calculez les écarts interquartiles.
3. Calculez l'écart absolu moyen.

### Solution

1. L'étendue est la différence entre l'âge maximal et l'âge minimal.

$$\text{Etendue} = \text{Max}\{x_i\} - \text{Min}\{x_i\} = 112 - 0.$$

**Etendue = 112.** La distribution des âges en France métropolitaine se répartit sur 112 ans.

2. Afin de pouvoir déterminer l'ensemble des quantiles, puis les intervalles correspondants, nous calculons les effectifs cumulés croissants, selon les étapes suivantes, sous Excel (voir figure 3.5) : l'effectif total n ( $\sum n_i$ ) en cellule B10, les fréquences (f<sub>i</sub>) en colonne C puis les fréquences cumulées croissantes (f<sub>cc</sub>) en colonne D.

**Figure 3.5**

**Résultats sous Excel.**

	A	B	C	D	E	F	G
1	Âge	n <sub>i</sub>	f <sub>i</sub>	f <sub>cc</sub>	x <sub>i</sub>	n <sub>i</sub> x <sub>i</sub>	n <sub>i</sub>  x <sub>i</sub> - x̄
2	0-14 ans	11 275 845	18,32%	18,32%	7	78 930 915,00	376 006 762,46
3	15-24 ans	7 806 706	12,89%	31,01%	19,5	152 230 767,00	162 740 279,18
4	25-34 ans	8 022 951	13,04%	44,09%	29,5	236 677 054,50	87 018 699,15
5	35-44 ans	8 733 224	14,19%	58,24%	39,5	344 962 348,00	7 390 193,50
6	45-54 ans	8 426 982	13,70%	71,94%	49,5	417 234 609,00	77 157 080,94
7	55-64 ans	7 166 591	11,65%	83,58%	59,5	426 412 164,50	137 267 336,35
8	65-74 ans	4 929 936	8,01%	91,59%	69,5	342 630 552,00	143 726 289,49
9	75-112 ans	5 173 765	8,41%	100,00%	93,5	483 747 027,50	275 005 187,50
10	<b>Somme</b>	<b>61 538 000</b>				<b>2 482 825 437,50</b>	<b>1 266 311 788,57</b>

Avec les mêmes méthodes de calcul que dans l'exercice 3 (interpolation linéaire) du chapitre 2 et à partir de la colonne des fréquences cumulées croissantes (f<sub>cc</sub>), nous pouvons déterminer que :

**Q<sub>1</sub> = 19,74** : 25 % des Français ont moins de 19,74 ans, soit environ 19 ans et 9 mois.

**Q<sub>3</sub> = 57,37** : 75 % des Français ont moins de 57,37 ans, soit environ 57 ans et 4 mois.

Donc l'écart interquartile **EIQ** est **Q<sub>3</sub> - Q<sub>1</sub> = 37,63** : 50 % des Français ont des âges répartis sur 37,63 ans, soit environ 37 ans et 8 mois.

**D<sub>1</sub> = 7,64** : 10 % des Français ont moins de 7,64 ans, soit environ 7 ans et 8 mois.

**D<sub>9</sub> = 72,21** : 90 % des Français ont moins de 72,21 ans, soit environ 72 ans et 3 mois.

Donc l'écart interdécile **EID** est **D<sub>9</sub> - D<sub>1</sub> = 64,57** : 80 % des Français ont des âges répartis sur 64,57 ans, soit environ 64 ans et 7 mois.

**C<sub>1</sub> = 0,76** : 1 % des Français ont moins de 0,76 an, soit environ 9 mois.

**C<sub>99</sub> = 107,6** : 99 % des Français ont moins de 107,6 ans, soit environ 107 ans et 7 mois.

Donc l'écart intercentile **EIC** est **C<sub>99</sub> - C<sub>1</sub> = 106,84** : 98 % des Français ont des âges répartis sur 106,84 ans, soit environ 106 ans et 10 mois.

3. Pour calculer l'écart absolu moyen, nous avons besoin de connaître la moyenne. Les centres de classes (x<sub>i</sub>) sont calculés en colonne E, les (n<sub>i</sub>x<sub>i</sub>) et leur somme en colonne F, à la suite du tableau précédent (voir figure 3.5).

La moyenne est égale à  $\bar{x} = \frac{1}{61 538 000} \sum_{i=1}^8 n_i x_i = \frac{2 482 825 437,5}{61 538 000}$ , soit  $\bar{x} = 40,35$ . L'âge

moyen de la population est d'environ 40 ans et 4 mois. Une fois la moyenne connue, les  $n_i |x_i - \bar{x}|$  et leur somme sont calculés en colonne G, à la suite du tableau précédent (voir figure 3.5).

L'écart absolu moyen est égal à  $e_a = \frac{1}{61 538 000} \sum_{i=1}^8 n_i |x_i - \bar{x}| = \frac{1 266 311 788,57}{61 538 000}$ , soit

**e<sub>a</sub> = 20,58 ans**. La moyenne des écarts à la moyenne est d'environ 20 ans et 7 mois.

## EXERCICE 2 BOÎTE À MOUSTACHES

### Énoncé

À partir des données et des résultats de l'exercice précédent, et en effectuant les calculs complémentaires nécessaires :

1. Recensez et donnez la valeur des indicateurs nécessaires au diagramme « boîte à moustaches ».
2. Dessinez le diagramme « boîte à moustaches ».

### Solution

1. Pour dessiner le diagramme « boîte à moustaches », nous avons besoin des indicateurs suivants :  $Q_1$  ;  $Me$  ;  $Q_3$  ;  $Q_3 + 1,5 (Q_3 - Q_1)$  et  $Q_1 - 1,5 (Q_3 - Q_1)$ .

D'après les résultats de l'exercice précédent :

$$Q_1 = 19,74.$$

$$Q_3 = 57,37.$$

$$Q_3 - Q_1 = 37,63.$$

Par interpolation linéaire, en utilisant le tableau construit pour l'exercice précédent (voir figure 3.5), notamment la colonne des fréquences cumulées croissantes ( $f_{cc}$ ), nous pouvons déterminer :  $Me = \frac{0,5 - 0,4405}{0,5824 - 0,4405} \times (44 - 35) + 35$  ; soit  $Me = 38,78$ . La moitié de la population étudiée a moins de 38,78 ans, soit environ 38 ans et 9 mois.

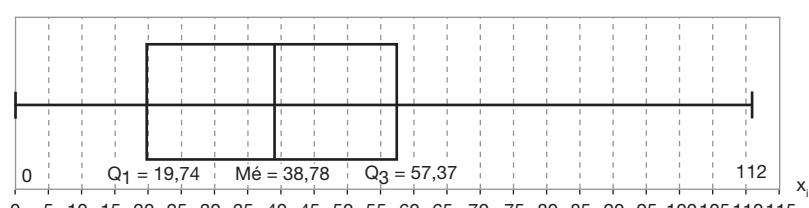
$Q_1 - 1,5 (Q_3 - Q_1) = 19,74 + 1,5 \times 37,63$ , soit  $Q_1 - 1,5 (Q_3 - Q_1) = -36,71$ . La moustache inférieure commence donc à 0, car un âge ne peut pas être négatif. Aucune valeur extrême inférieure à  $Q_1 - 1,5 (Q_3 - Q_1)$  n'est recensée.

$Q_3 + 1,5 (Q_3 - Q_1) = 57,37 + 1,5 \times 37,63$ , soit  $Q_3 + 1,5 (Q_3 - Q_1) = 113,82$ . La moustache supérieure finit donc à 112 qui est l'âge maximal.

2.

Figure 3.6

Boîte à moustaches.





## EXERCICE 3 VARIANCE ET ÉCART-TYPE SUR CARACTÈRE DISCRET

### Énoncé

Un enseignant de statistique demande à ses étudiants le nombre de films qu'ils ont vus au cinéma au cours des deux derniers mois. Les résultats sont reportés dans le tableau suivant :

Nombre de films vus	Nombre d'étudiants
0	6
1	4
2	9
3	7
4	3
5	2

1. Calculez la moyenne du nombre de films vus au cinéma.
2. Calculez :
  - a. la variance du nombre de films vus au cinéma ;
  - b. l'écart-type du nombre de films vus au cinéma.
3. Calculez le coefficient de variation.

### Solution

1. Saisissez les modalités dans la colonne L1 et les effectifs dans la colonne L2 (voir figure 3.7).

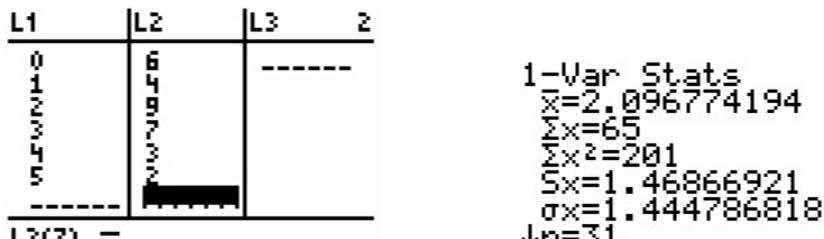
Dans le menu CALC de STAT,appelez la fonction 1-Var Stats, puis indiquez dans l'ordre L1, L2 comme suit : 1-Var Stats L1,L2. Validez avec **ENTER**. Les résultats de la figure 3.8 s'affichent.

Figure 3.7 (gauche)

Saisie du tableau de données avec la calculatrice.

Figure 3.8 (droite)

Résultats de l'analyse statistique effectuée avec la calculatrice.



La moyenne est  $\bar{x}=\frac{1}{31} \sum_{i=1}^6 n_i x_i = \frac{65}{31}$ , soit  $\bar{x} = 2,096$ . Le nombre moyen de films vus au cinéma par étudiant au cours des deux derniers mois est de 2,1 films.

**2. a.** La variance est égale à  $V(x) = \sigma_x^2 = 1,4447^2$ , soit  $\mathbf{V(x) = 2,087}$ . Ou encore, par la formule développée,  $V(x) = \frac{1}{31} \sum_{i=1}^6 n_i x_i^2 - \bar{x}^2 = \frac{201}{31} - 2,1^2 = 2,087$  (aux arrondis près). La variance du nombre de films vus au cinéma par étudiant au cours des deux derniers mois est de 2,1.

**b.** L'écart-type est égal à  $\sigma_x = \sqrt{V(x)} = 1,44$ . L'écart-type du nombre de films vus au cinéma par étudiant au cours des deux derniers mois est de 1,44 film.

**3.** Le coefficient de variation est égal à  $CV(x) = \frac{\sigma_x}{\bar{x}} = \frac{1,44}{2,1}$ , soit  $\mathbf{CV(x) = 0,69}$ . L'écart-type est inférieur à la moyenne.

## EXERCICE 4 COMPARAISON DE DISTRIBUTIONS SUR CARACTÈRE CONTINU



### Énoncé

Le tableau ci-après recense la population féminine et masculine de la France métropolitaine par tranches d'âge en 2007 (données provisoires) :

Âge	Femmes	Hommes
0-14 ans	5 503 794	5 772 051
15-24 ans	3 858 982	3 947 724
25-34 ans	3 985 506	4 037 445
35-44 ans	4 396 709	4 336 515
45-54 ans	4 301 816	4 127 166
55-64 ans	3 637 565	3 529 026
65-74 ans	2 657 004	2 272 932
75-112 ans	3 289 624	1 884 141

Source : Insee, recensement de la population, bilan démographique, 2007

1. Pour les femmes, calculez :

- a. la moyenne ;
- b. l'écart-type ;
- c. le coefficient de variation.

2. Pour les hommes, calculez :

- a. la moyenne ;
- b. l'écart-type ;
- c. le coefficient de variation.

3. Comparez les deux distributions.

## Solution

1. Pour les femmes, les centres de classes ( $x_i$ ) sont calculés en colonne B, les ( $n_i x_i$ ) et leur somme en colonne D, puis les ( $n_i x_i^2$ ) et leur somme en colonne E (voir figure 3.9).

**Figure 3.9**

Résultats sous Excel.

	A	B	C	D	E
1	Age	$x_i$	$n_i$ (Femmes)	$n_i x_i$	$n_i x_i^2$
2	0-14 ans	7	5 503 794	38 526 558	269 685 906
3	15-24 ans	19,5	3 858 982	75 250 149	1 467 377 906
4	25-34 ans	29,5	3 985 506	117 572 427	3 488 386 597
5	35-44 ans	39,5	4 396 709	173 670 006	6 859 965 217
6	45-54 ans	49,5	4 301 816	212 939 892	10 540 524 654
7	55-64 ans	59,5	3 637 565	216 435 118	12 877 889 491
8	65-74 ans	69,5	2 657 004	184 661 778	12 833 993 571
9	75-112 ans	93,5	3 289 624	307 579 844	28 758 715 414
10	Somme		31 631 000	1 326 635 771	77 076 538 756

a. La moyenne est égale à  $\bar{x} = \frac{1}{31631000} \sum_{i=1}^8 n_i x_i = \frac{1326635771}{31631000}$ , soit  $\bar{x} = 41,94$ . L'âge moyen des femmes est d'environ 41 ans et 11 mois.

b. Par la formule développée, la variance est égale à :

$V(x) = \frac{1}{31631000} \sum_{i=1}^8 n_i x_i^2 - \bar{x}^2 = \frac{77076538756}{31631000} - 41,94^2$ , soit  $V(x) = 677,69$ . La variance de l'âge des femmes est de 677,69.

c. L'écart-type est égal à  $\sigma_x = \sqrt{V(x)} = \sqrt{677,69}$ , soit  $\sigma_x = 26,03$ . L'écart-type de l'âge des femmes est de 26,03 ans, soit environ 26 ans.

Le coefficient de variation pour les femmes est égal à  $CV(x) = \frac{\sigma_x}{\bar{x}} = \frac{26,03}{41,94}$ , soit  $CV(x) = 0,621$ . L'écart-type est inférieur à la moyenne.

2. En procédant de la même manière pour les hommes, on obtient sous Excel la figure 3.10.

**Figure 3.10**

Résultats sous Excel.

	A	B	C	D	E
21	Age	$x_i$	$n_i$ (Hommes)	$n_i x_i$	$n_i x_i^2$
22	0-14 ans	7	5 772 051	40 404 357	282 830 499
23	15-24 ans	19,5	3 947 724	76 980 618	1 501 122 051
24	25-34 ans	29,5	4 037 445	119 104 628	3 513 506 511
25	35-44 ans	39,5	4 336 515	171 292 343	6 766 047 529
26	45-54 ans	49,5	4 127 166	204 294 717	10 112 588 492
27	55-64 ans	59,5	3 529 026	209 977 047	12 493 634 297
28	65-74 ans	69,5	2 272 932	157 968 774	10 978 829 793
29	75-112 ans	93,5	1 884 141	176 167 184	16 471 631 657
30	Somme		29 907 000	1 156 189 667	62 120 270 828

a. La moyenne est égale à  $\bar{x} = \frac{1}{29907000} \sum_{i=1}^8 n_i x_i = \frac{1156189667}{29907000}$ , soit  $\bar{x} = 38,66$ . L'âge moyen des hommes est d'environ 38 ans et 8 mois.

b. Par la formule développée, la variance est égale à :

$V(x) = \frac{1}{29\,907\,000} \sum_{i=1}^8 n_i x_i^2 - \bar{x}^2 = \frac{62\,120\,270\,828}{29\,907\,000} - 38,66^2$ , soit  $V(x) = 582,56$ . La variance de l'âge des hommes est de 582,56.

L'écart-type est égal à  $\sigma_x = \sqrt{V(x)} = \sqrt{582,56}$ , soit  $\sigma_x = 24,14$ . L'écart-type de l'âge des hommes est de 24,14 ans, soit environ 24 ans et 2 mois.

c. Le coefficient de variation pour les hommes est égal à  $CV(x) = \frac{\sigma_x}{\bar{x}} = \frac{24,14}{38,66}$ , soit  $CV(x) = 0,624$ . L'écart-type est inférieur à la moyenne.

3. Les hommes sont en moyenne plus jeunes que les femmes (âge moyen : 38,66 contre 41,94).

Dans l'absolu, l'âge des hommes est légèrement moins dispersé que celui des femmes (écart-type : 24,14 contre 26,03).

En rapportant cette dispersion à l'âge moyen, nous pouvons cependant conclure que, par rapport à leur âge moyen, l'âge des hommes est légèrement plus dispersé que celui des femmes (coefficient de variation : 0,624 contre 0,621).

## EXERCICE 5 MANIPULATIONS DE FORMULES

### Énoncé

Afin de mieux servir ses clients, un magasin a mesuré le temps d'attente, noté  $x$ , au guichet de son service après-vente. Le temps d'attente est mesuré en minutes. La personne en charge du traitement de l'étude vous communique les données suivantes :

$$V(x) = 17,18$$

$$\sum_{i=1}^k f_i x_i^2 = 50,17$$

$$\sum_{i=1}^k n_i x_i = 425$$

1. Déterminez l'effectif total à partir duquel l'enquête a été réalisée. Indiquez les valeurs de :
  - a. la moyenne ;
  - b. l'écart-type.
2. L'objectif de la direction est de diminuer le temps d'attente de 30 %. Calculez :
  - a. le temps d'attente moyen correspondant ;
  - b. l'écart-type correspondant.
3. En effectuant une vérification du chronomètre utilisé, le directeur du magasin s'aperçoit que ce dernier accuse un retard de 5 % par rapport au temps réel. Calculez :
  - a. la vraie moyenne ;
  - b. le vrai écart-type.

## Solution

1.

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

$$V(x) = \sum_{i=1}^k f_i x_i^2 - \left( \frac{1}{n} \sum_{i=1}^k n_i x_i \right)^2$$

Soit, en remplaçant par les valeurs connues :  $17,18 = 50,17 - \left( \frac{425}{n} \right)^2$ , donc

$n = \frac{425}{\sqrt{50,17 - 17,18}}$ , soit  $n = 74$ . **L'effectif total est de 74**, ce qui signifie que 74 temps d'attente ont été observés.

a. La moyenne est égale à  $\bar{x} = \frac{1}{n} \sum_{i=1}^k n_i x_i = \frac{425}{74}$ , soit  $\bar{x} = 5,74$ . Le temps d'attente moyen est d'environ 5 minutes et 44 secondes.

b. L'écart-type est égal à  $\sigma_x = \sqrt{V(x)} = \sqrt{17,18}$ , soit  $\sigma_x = 4,14$ . L'écart-type du temps d'attente est d'environ 4 minutes et 8 secondes.

2. La base d'application du pourcentage est le temps d'attente mesuré. Les objectifs de temps d'attente, notés  $y_i$ , sont égaux aux temps d'attente actuels, notés  $x_i$ , auxquels sont retirés 30 % des temps d'attente actuels. Soit  $y_i = x_i - 0,3 \times x_i = 0,7 \times x_i$ .

a. Grâce aux propriétés de la moyenne, nous pouvons en conclure que  $\bar{y} = 0,7 \times \bar{x}$ , soit  $\bar{y} = 4,02$ . L'objectif de réduction de 30 % du temps d'attente ramène la **moyenne** de ce dernier à environ **4 minutes et 1 seconde**.

b. Grâce aux propriétés de l'écart-type, nous pouvons en conclure que  $\sigma_y = 0,7 \times \sigma_x$ , soit  $V(y) = 2,90$ . L'objectif de réduction de 30 % du temps d'attente ramène l'**écart-type** de ce dernier à environ **2 minutes et 54 secondes**.

3. La base d'application du pourcentage est le temps réel. Les temps d'attente réels, notés  $z_i$ , sont égaux aux faux temps d'attente, notés  $x_i$ , auxquels sont ajoutés 5 % des temps d'attente réels. Soit  $z_i = x_i + 0,05 \times z_i$ ; c'est-à-dire  $z_i = \frac{x_i}{0,95}$ .

a. Grâce aux propriétés de la moyenne, nous pouvons en conclure que  $\bar{z} = \frac{\bar{x}}{0,95}$ , soit  $\bar{z} = 6,05$ . Le temps d'attente réel a une **moyenne** d'environ **6 minutes et 3 secondes**.

b. Grâce aux propriétés de l'écart-type, nous pouvons en conclure que  $\sigma_z = \frac{\sigma_x}{0,95}$ , soit  $\sigma_z = 19,04$ . Le temps d'attente réel a un **écart-type** d'environ **4 minutes et 2 secondes**.

## Bibliographie

- CHAREILLE P. et PINAULT Y., *Statistique descriptive*, Collection AES, Montchrestien, Paris, 1996.
- DELMAS B., *Statistique descriptive*, Armand Colin, 2005.
- DROESBEKE J.-J., *Éléments de statistiques*, Éditions de l'université de Bruxelles, Ellipses, 2001.
- LIORZOU A., *Initiation à la pratique statistique*, Eyrolles, 1985.
- PIATIER A., *Statistique descriptive et initiation à l'analyse*, Thémis, PUF, 1962.
- ROGER P., *Probabilités, statistique et processus stochastiques*, Collection Synthex, Pearson Education, 2004.
- SCHLACHTHER D., *De l'analyse à la prévision*, Ellipses, 1986.
- GRENON G. et VIAU S., *Méthodes quantitatives en sciences humaines*, Gaëtan Morin, 1999.
- HAUCHECORNE B., *Les mots et les maths*, Ellipses, 2003.



# Les caractéristiques de forme et de concentration

1.	La courbe de la loi normale ..	84
2.	Les caractéristiques de forme ..	85
3.	Les caractéristiques de concentration .....	89
<b>Problèmes et exercices</b>		
1.	Caractéristiques d'asymétrie ..	95
2.	Caractéristiques d'aplatissement .....	98
3.	Caractéristiques de forme et médiale .....	100
4.	Caractéristique de concentration : l'indice de Gini .....	104

Ce chapitre prolonge et complète la description d'une série statistique amorcée dans les chapitres 2 et 3, en précisant les notions de tendance centrale et de dispersion, autour de la courbe de la loi normale. Cette courbe est à rattacher aux modèles théoriques des distributions de probabilité. La loi normale, dite loi de Laplace-Gauss, en est le modèle phare, et sa fameuse « courbe en cloche » sert de référence.

Dans un premier temps, nous donnerons un aperçu rapide de la loi normale.

Dans un deuxième temps, nous définirons différents coefficients, introduits par Karl Pearson, le père de la statistique moderne, George Yule et Ronald Fisher, permettant de caractériser la forme d'une distribution.

Enfin, nous terminerons ce chapitre par la notion de concentration, introduite par le statisticien et démographe Corrado Gini, à propos de distributions de salaires et de revenus. Ce sera l'occasion de prolonger l'analyse de la dispersion relative et de rendre compte des inégalités éventuelles de répartition.

# 1 La courbe de la loi normale

Nous avons vu que, selon son caractère discret ou continu, une série statistique peut être représentée par un diagramme en bâtons ou un histogramme des fréquences que l'on complète en général par le tracé du polygone des fréquences. Il faut garder à l'esprit que l'histogramme des fréquences est un bon estimateur de la densité et qu'en lissant le polygone des fréquences on peut représenter la série statistique par une distribution continue. La loi normale, également appelée loi de Laplace-Gauss, est le modèle fondamental des distributions continues.

La loi normale représente la distribution des valeurs d'une grandeur soumise à l'influence d'un grand nombre de facteurs indépendants les uns des autres, chacun exerçant des actions de faible intensité dont les effets tendent à se compenser.

## 1.1 PRÉSENTATION DE LA LOI NORMALE

De nombreux caractères quantitatifs du monde réel suivent une loi normale : les tailles des individus, les poids, la pression sanguine, les notes à un examen, etc.

Quand on désire mesurer une grandeur, par exemple une longueur, dont la vraie valeur est L, on opère n mesures,  $x_1, x_2, \dots, x_n$ , et la variable X dont les modalités sont les  $(x_i - L)$ , représente l'erreur commise dans la mesure de L. Cette variable suit une loi normale.

Aussi cette distribution est-elle souvent appelée « loi des erreurs », parce que les erreurs aléatoires dans les résultats de mesures sont souvent normalement distribuées.

### Définitions

La **loi normale** est entièrement déterminée par deux paramètres : sa moyenne ( $m$ ) et son écart-type ( $\sigma$ ).

La **loi normale centrée réduite** constitue le modèle de référence ; sa moyenne est 0 (centrée) et son écart-type 1 (réduite). Sa densité est donnée par :  $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$  et sa représentation graphique est la célèbre courbe en cloche (voir figure 4.1). On dit que X suit la loi  $N(0 ; 1)$ .

Si une variable X suit une loi normale de paramètres  $m$  et  $\sigma$ , notée  $N(m ; \sigma)$ , alors  $Z = \frac{X-m}{\sigma}$  suit la loi normale centrée réduite de paramètres 0 et 1. On dit que l'on a standardisé X.

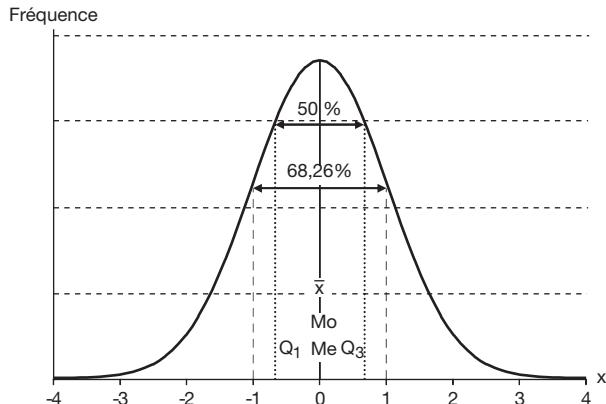
La courbe représentant la distribution  $N(0 ; 1)$  est symétrique, avec :  $\bar{x} = Mo = Me = 0$ . Elle est « normalement aplatie ».

Avec  $\bar{x} = 0$  et  $\sigma = 1$ , l'intervalle  $[\bar{x} - \sigma ; \bar{x} + \sigma]$  qui correspond à  $[-1 ; 1]$  représente 68,26 % des observations et l'intervalle  $[\bar{x} - 2\sigma ; \bar{x} + 2\sigma]$  qui correspond à  $[-2 ; 2]$  représente 95,44 % des observations.

Les deux quartiles  $Q_1$  et  $Q_3$  sont opposés et valent respectivement -0,67 et 0,67.

**Figure 4.1**

**La courbe en cloche de la loi normale centrée réduite.**

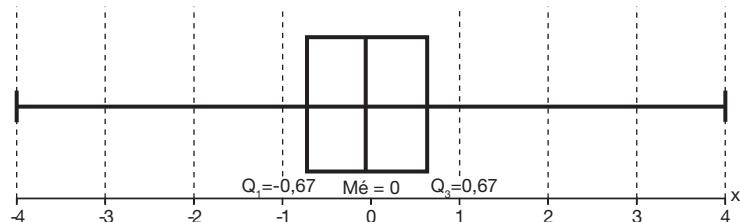


## 1.2 LOI NORMALE ET BOÎTE À MOUSTACHES

La boîte à moustaches d'une distribution statistique conforme à une distribution normale mettra en évidence la symétrie :  $Q_1$  et  $Q_3$  sont équidistants de la médiane (Me) qui est dans ce cas la moyenne arithmétique et le mode (voir figure 4.2).

**Figure 4.2**

**Boîte à moustaches de la loi normale centrée réduite.**



## 2 Les caractéristiques de forme

### 2.1 L'ASYMÉTRIE (SKEWNESS)

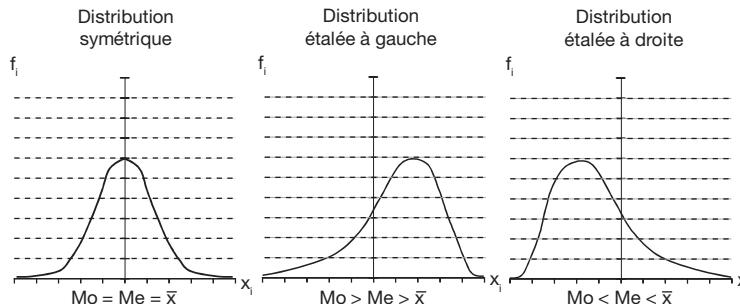
Une distribution est dite symétrique, comme la loi normale, si les valeurs observées se répartissent de façon uniforme autour des trois valeurs centrales alors égales : la moyenne, le mode et la médiane.

Pour mesurer l'asymétrie d'une distribution, on dispose de différents coefficients. Le but est de comparer les formes de plusieurs distributions, ces comparaisons n'ayant de sens que si elles sont faites à partir des mêmes coefficients appliqués aux différentes distributions.

La figure 4.3 montre les trois formes de symétrie et asymétrie possibles.

**Figure 4.3**

**Symétrie et asymétrie.**



### Le coefficient de Yule et Kendall

Le coefficient de Yule et Kendall – couramment appelé coefficient de Yule – compare l'étalement de la courbe à droite et à gauche de la médiane.

**Définition**

Le **coefficient de Yule** sert à mesurer l'asymétrie de la distribution en tenant compte des positions relatives des quartiles par rapport à la médiane. Il est défini par :  $C_y = \frac{Q_1 + Q_3 - 2Me}{Q_3 - Q_1}$ , ou de manière équivalente par  $C_y = \frac{Q_1 - Me + Q_3 - Me}{Q_3 - Q_1}$ .

Ce coefficient permet de localiser la médiane dans la boîte à moustaches, par rapport au milieu du segment formé par  $Q_1$  et  $Q_3$ .

Dans le cas d'une distribution symétrique, comme la loi normale, ce coefficient est nul, les quartiles  $Q_1$  et  $Q_3$  étant équidistants de la médiane.

Ce coefficient  $C_y$  est indépendant de l'unité de mesure. En outre, il est toujours compris entre  $-1$  et  $1$ , car la médiane est située entre  $Q_1$  et  $Q_3$ .

- Si  $C_y = 0$ , la distribution est symétrique.
- Si  $C_y > 0$ , la distribution est étalée à droite.
- Si  $C_y < 0$ , la distribution est étalée à gauche.

### Les coefficients de Pearson

Les coefficients de Pearson étudient l'étalement de la courbe à partir des valeurs de la moyenne, du mode et de l'écart-type.

**Définition**

Le **coefficient S de Pearson** mesure l'asymétrie d'une distribution par une comparaison entre les valeurs de la moyenne et du mode. Il se note  $S = \frac{\bar{x} - Mo}{\sigma}$ . Il s'agit d'un coefficient sans dimension.

L'interprétation de la valeur du S de Pearson se fait comme suit :

- Si  $S = 0$ , la distribution est symétrique.
- Si  $S > 0$ , la distribution est étalée à droite.
- Si  $S < 0$ , la distribution est étalée à gauche.

**Définition**

Le coefficient d'asymétrie  $\beta_1$  de Pearson est défini par :  $\beta_1 = \frac{\mu_3^2}{\mu_2^3}$ .

$\mu_3$  désigne le moment centré d'ordre 3, soit  $\mu_3 = \frac{1}{n} \sum_{i=1}^r n_i (x_i - \bar{x})^3$ .

$\mu_2$  est le moment centré d'ordre 2, c'est-à-dire la variance.

L'interprétation de la valeur du  $\beta_1$  de Pearson se fait comme suit :

- Si  $\beta_1$  est proche de 0, la distribution est approximativement symétrique.
- Si  $\beta_1 > 0$ , elle est étalée à droite pour  $\mu_3 > 0$  et étalée à gauche pour  $\mu_3 < 0$ .

**Le coefficient de Fisher****Définition**

Le coefficient d'asymétrie  $\gamma_1$  de Fisher est défini par :  $\gamma_1 = \frac{\mu_3}{\sigma^3}$ .

$\mu_3$  désigne le moment centré d'ordre 3, soit  $\mu_3 = \frac{1}{n} \sum_{i=1}^r n_i (x_i - \bar{x})^3$ .

Ce coefficient, sans dimension, a le même signe que  $\mu_3$ .

L'interprétation de la valeur du  $\gamma_1$  de Fisher se fait comme suit :

- Si  $\gamma_1$  est proche de 0, la distribution est approximativement symétrique.
- Si  $\gamma_1 > 0$ , la distribution est étalée à droite.
- si  $\gamma_1 < 0$ , la distribution est étalée à gauche.

**Exemple 4.1****Calculs des coefficients d'asymétrie**

Le tableau suivant donne une estimation de la répartition par âges des assurés obligatoires de plus de 20 ans et de moins de 60 ans, en France, en 1921 :

Âge (années)	Effectif (milliers)
[20 ; 25[	1 275
[25 ; 30[	1 080
[30 ; 35[	890
[35 ; 40[	805
[40 ; 45[	745
[45 ; 50[	675
[50 ; 55[	610
[55 ; 60[	505

Source : Bureau international du travail, 1921

Calculons les différents coefficients d'asymétrie à l'aide d'Excel (voir figure 4.4).

**Figure 4.4**

**Calcul des coefficients d'asymétrie sous Excel.**

	A	B	C	D	E	F	G	H	I
1	a <sub>i</sub>	b <sub>i</sub>	n <sub>i</sub>	f <sub>i</sub>	x <sub>i</sub>	n <sub>i</sub> Cum	n <sub>i</sub> x <sub>i</sub>	n <sub>i</sub> x <sub>i</sub> <sup>2</sup>	n <sub>i</sub> (x <sub>i</sub> - x̄) <sup>3</sup>
2	20	25	1 275	0,1936	22,5	1 275	28 687,5	645 468,75	-3 723 555,35
3	25	30	1 080	0,1640	27,5	2 355	29 700,0	816 750,00	- 866 983,20
4	30	35	890	0,1352	32,5	3 245	29 925,0	940 062,50	- 70 458,03
5	35	40	605	0,1222	37,5	4 050	30 187,5	1 132 031,25	283,46
6	40	45	745	0,1131	42,5	4 795	31 562,5	1 345 656,25	138 415,86
7	45	50	675	0,1026	47,5	5 470	32 062,5	1 522 968,75	828 330,75
8	50	55	610	0,0926	52,5	6 080	32 025,0	1 681 312,50	2 363 410,09
9	55	60	605	0,0767	57,5	6 585	29 037,5	1 689 656,25	4 483 213,97
10	Somme		6 585	1,0000			242 287,5	9 753 906,25	3 152 657,56

$$\text{Le calcul de la moyenne donne } \bar{x} = \frac{1}{6\ 585} \sum_{i=1}^8 n_i x_i = \frac{242\ 287,5}{6\ 585} = 36,79.$$

La variance, ou moment centré d'ordre 2, est :

$$V(x) = \frac{1}{6\ 585} \sum_{i=1}^8 n_i x_i^2 - \bar{x}^2 = \frac{1}{6\ 585} \times 9\ 753\ 906,25 - 36,79^2 = 127,44, \text{ et l'écart-type :}$$

$$\sigma_x = \sqrt{V(x)} = \sqrt{127,44} = 11,29.$$

$$\text{Le moment centré d'ordre 3 est } \mu_3 = \frac{1}{6\ 585} \sum_{i=1}^8 n_i (x_i - \bar{x})^3 = \frac{3\ 152\ 657,56}{6\ 585} = 478,76.$$

À partir de la colonne des n<sub>i</sub> cumulés croissants et par interpolation linéaire, on obtient Q<sub>1</sub> = 26,72 ; Me = 35,30 et Q<sub>3</sub> = 46,06.

$$\text{Le mode est égale à } Mo = \frac{k_2 x_1 + k_1 x_2}{k_1 + k_2} = \frac{195 \times 20 + 1275 \times 25}{1275 + 195}, \text{ soit } Mo = 24,33.$$

Suite à ces calculs, nous pouvons déterminer l'ensemble des coefficients d'asymétrie.

$$C_Y = \frac{26,72 + 46,06 - 2 \times 35,3}{46,06 - 26,72}, \text{ soit } C_Y = 0,11; \quad S = \frac{36,79 - 24,33}{11,29}, \text{ soit } S = 1,10;$$

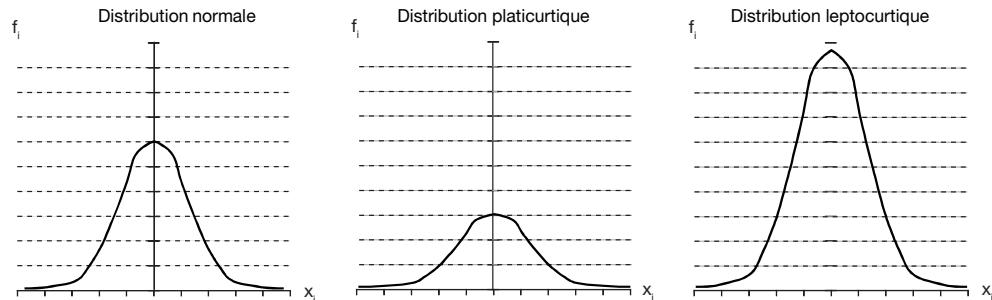
$$\beta_1 = \frac{478,76^2}{127,44^3}, \text{ soit } \beta_1 = 0,11; \quad \gamma_1 = \frac{478,76}{11,29^3}, \text{ soit } \gamma_1 = 0,33.$$

Les coefficients mettent en évidence une distribution asymétrique étalée à droite, ce que confirme la réalisation de l'histogramme.

## 2.2 L'APLATISSEMENT (KURTOSIS)

L'aplatissement d'une distribution est un indicateur de la dispersion autour des valeurs centrales. Plus la dispersion est grande, plus la courbe sera « plate ». On définira deux coefficients, celui de Pearson et celui de Fisher, ces coefficients étant des coefficients de comparaison par rapport à la distribution normale.

La figure 4.5 montre les trois formes d'aplatissement possibles.

**Figure 4.5****Aplatissement.**

### Le coefficient de Pearson

#### Définition

Le **coefficient  $\beta_2$  de Pearson** sert à mesurer l'aplatissement. Il est défini par  $\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{\mu_4}{\sigma^4}$ .

Il s'agit d'un coefficient sans dimension.  $\beta_2 \geq 1$  et dans le cas d'une distribution normale  $\beta_2 = 3$ .

Interprétation :

- Si  $\beta_2 < 3$ , la courbe est dite platicurtique, c'est-à-dire plus plate que la loi normale.
- Si  $\beta_2 = 3$ , la courbe est proche de la courbe normale.
- Si  $\beta_2 > 3$ , la courbe est leptocurtique, c'est-à-dire plus pointue que la loi normale.

### Le coefficient de Fisher

#### Définition

Le **coefficient  $\gamma_2$  de Fisher** sert à mesurer l'aplatissement. Il est défini par  $\gamma_2 = \frac{\mu_4}{\mu_2^2} - 3 = \frac{\mu_4}{\sigma^4} - 3$ . Ou encore, de manière équivalente,  $\gamma_2 = \beta_2 - 3$ .

La constante 3 est choisie de façon à obtenir un coefficient nul pour une distribution normale ; par ailleurs,  $\gamma_2 \geq -2$ .

Interprétation :

- Si  $\gamma_2 < 0$ , la courbe est dite platicurtique, c'est-à-dire plus plate que la loi normale.
- Si  $\gamma_2 = 0$ , la courbe est proche de la courbe normale.
- Si  $\gamma_2 > 0$ , la courbe est leptocurtique, c'est-à-dire plus pointue que la loi normale.

On notera que  $\gamma_2$  mesure l'importance des « queues de distribution ».

## 3 Les caractéristiques de concentration

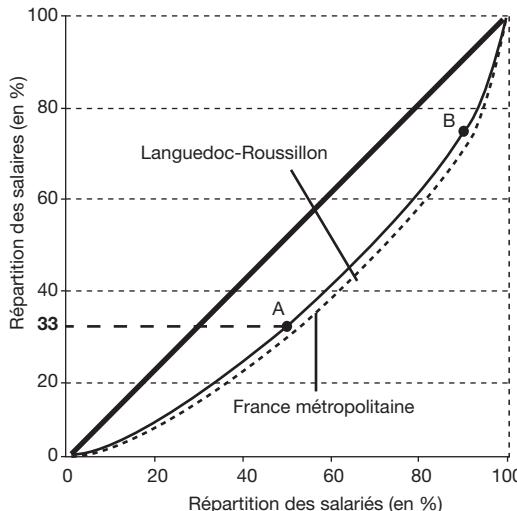
La mesure de la concentration concerne les caractères statistiques quantitatifs représentant une grandeur positive cumulable. Il s'agit de traduire la densité des données autour de la valeur centrale. Sont principalement étudiés la concentration des salaires, des revenus, de l'emploi, ou encore le degré de concentration dans une branche d'un secteur économique.

Afin de mesurer la concentration, il convient de définir les valeurs globales, la médiale, l'indice de Gini et la courbe de concentration, appelée courbe de Lorentz.

Un exemple de courbe de concentration des salaires, proposée par l'Insee, est donné figure 4.6.

**Figure 4.6**

**Concentration des salaires du secteur privé en Languedoc-Roussillon : une répartition inégalitaire.**



Source : Insee – DADS, novembre 2003

**Note de lecture :** si la répartition des salaires était totalement égalitaire, la courbe de concentration se confondrait avec la bissectrice en noir. Dans la région, les 50 % des salariés les moins rémunérés se partagent 33 % de la masse salariale (point A) ; les 10 % les mieux rémunérés concentrent 25 % des salaires (point B). La courbe de concentration pour la France métropolitaine est en dessous de celle de la Région, la distribution des salaires y est donc plus inégalitaire.

### 3.1 LES VALEURS GLOBALES

#### Définitions

Étant donnée une série statistique comportant  $n$  observations ordonnées dans un tableau statistique  $(x_i ; n_i)$ , présentant  $r$  modalités, on appelle :

- **masse** associée à la modalité  $x_i$  d'effectif  $n_i$ , la quantité définie par  $n_i x_i$  ;
- **masse relative** associée à la modalité  $x_i$ , notée  $q_i$ , la quantité définie par  $q_i = \frac{n_i x_i}{\sum_{k=1}^r n_k x_k}$ .

Généralement, les masses relatives  $q_i$  sont exprimées en pourcentage de la masse totale

$$S = \sum_{i=1}^r n_i x_i \quad (\text{appelée masse salariale dans le cas des salaires}).$$

Les masses relatives cumulées croissantes sont notées  $q_{i,cc}$  et définies par  $q_{i,cc} = \sum_{k=1}^i q_k$ .

**Exemple 4.2****Calcul des masses relatives**

Le tableau suivant indique les réserves de pétrole, en milliards de barils, dont disposent les pays producteurs :

Réserves de pétrole	Nombre de pays
[0 ; 10[	10
[10 ; 50[	8
[50 ; 100[	3
[100 ; 275[	4

Source : Energy Information Administration, Department of Energy, janvier 2004

À partir de la série ordonnée par ordre croissant sont effectués les calculs des centres de classes  $x_i$ , des fréquences  $f_i$  et  $f_{cc}$  cumulées croissantes, ainsi que ceux des masses relatives  $q_i$  et  $q_{cc}$  cumulées croissantes (voir figure 4.7). Ces calculs permettent de tracer la courbe de Lorentz et de calculer l'indice de concentration de Gini que nous allons définir ci-après (voir section 3.4).

**Figure 4.7**

**Calcul des masses relatives sous Excel.**

	A	B	C	D	E	F	G	H
1	Réserves	$n_i$	$x_i$	$f_i$	$f_{cc}$	$n_i x_i$	$q_i$	$q_{cc}$
3	[0 ; 10[	10	5	40,00%	40,00%	50	3,95%	3,95%
4	[10 ; 50[	8	30	32,00%	72,00%	240	18,97%	22,92%
5	[50 ; 100[	3	75	12,00%	84,00%	225	17,79%	40,71%
6	[100 ; 275[	4	188	16,00%	100,00%	750	59,29%	100,00%
7	Somme	25		100,00%		1 265	100,00%	

## 3.2 LA MÉDIALE

### Définition

La **médiale** est la valeur du caractère qui partage en deux parties égales la masse totale du caractère.

La médiale est notée  $Ml$ , elle s'exprime dans la même unité que le caractère, et correspond à une valeur de la masse relative cumulée croissante  $q_{cc}$  de 50 %.

La médiale est, d'une certaine façon, une médiane et sa détermination en est similaire :

- Dans le cas discret, la médiale est la plus petite valeur du caractère dont la masse relative cumulée croissante est inférieure ou égale à 50 %.
- Dans le cas continu, on peut opérer de deux façons : soit graphiquement à l'aide du polygone des masses relatives cumulées croissantes, soit algébriquement par interpolation linéaire.

L'écart entre la médiale et la médiane ( $Ml \geq Me$ ) donne une première indication sur la concentration de la série. Plus cet écart est important par rapport à l'étendue de la série, plus la concentration est forte.

### **Exemple 4.3**

#### **Calcul de la médiale**

Reprenez les données de l'exemple 4.2.

Dans cet exemple, par interpolation linéaire, la médiane est 22,5. 50 % des pays ont une réserve de pétrole inférieure ou égale à 22,5 milliards de barils.

La médiale se calcule comme la médiane, en utilisant les  $q_{cc}$  au lieu des  $f_{cc}$ ;  $q_{cc} = 50\%$  pour l'intervalle [100 ; 275]. La médiale est 127,42 ; c'est la plus petite valeur telle que les pays ayant une réserve inférieure ou égale à cette valeur se partagent au moins 50 % des réserves totales.

L'écart  $Ml - Me$  vaut  $127,42 - 22,5 = 104,92$ , l'étendue étant de  $275 - 0 = 275$ , soit à peine trois fois plus grande, ce qui traduit une forte concentration.

## **3.3 LA COURBE DE CONCENTRATION**

---

La courbe de concentration est réalisée à partir des calculs précédents. On la dessine en utilisant les fréquences cumulées croissantes ( $f_{cc}$ ) et les masses relatives cumulées croissantes ( $q_{cc}$ ). Cette représentation permet de comparer la distribution observée à la distribution théorique d'égale répartition, celle où, pour chaque modalité,  $f_{cc} = q_{cc}$ .

Les fréquences cumulées croissantes sont portées en abscisses et les masses relatives cumulées croissantes en ordonnées.

La distribution théorique d'égale répartition correspond à la bissectrice du repère.

L'aire comprise entre la distribution théorique et la courbe de concentration s'appelle surface de concentration.

### **Exemple 4.4**

#### **Réalisation de la courbe de concentration**

Reprenez les données de l'exemple 4.2.

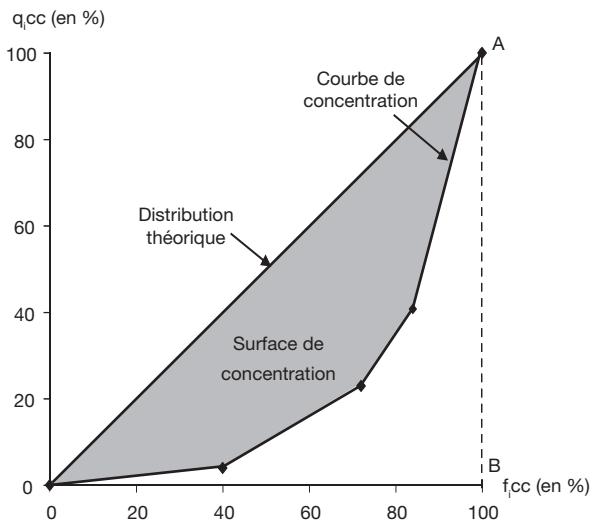
À partir des calculs des fréquences cumulées croissantes ( $f_{cc}$ ) et des masses relatives cumulées croissantes ( $q_{cc}$ ) présentés figure 4.7, il est possible de dessiner la courbe de concentration (voir figure 4.8).

La courbe de Lorentz est inscrite dans le carré de côté 100, quand les fréquences sont exprimées en pourcentage. Plus la courbe de Lorentz est éloignée de la diagonale, qui représente la distribution théorique d'égale répartition, plus la concentration est forte.

La surface de concentration est comprise entre la courbe de Lorentz et la diagonale. Plus cette surface est grande, plus la concentration est forte.

**Figure 4.8**

**Courbe de concentration des réserves de pétrole.**



### 3.4 L'INDICE DE GINI

**Définition**

La **surface de concentration** est le domaine compris entre la diagonale [OB] du carré de concentration et la courbe de concentration.

L'aire de la surface de concentration est égale à l'aire du triangle rectangle OAB diminuée de l'aire du domaine situé sous la surface de concentration. Le triangle OAB est formé des points de coordonnées O(0 ; 0), A(100 ; 100) et B(100 ; 0) (voir figure 4.8). Avec les  $f_{cc}$  et les  $q_{cc}$  exprimées en pourcentages, l'aire du triangle OAB est de  $100 \times 100 / 2$ . Dans le cas où les  $f_{cc}$  et les  $q_{cc}$  sont exprimées en nombres décimaux, cette aire de 0,5.

**Définition**

L'**indice de Gini** est le rapport de l'aire de la surface de concentration à l'aire de la surface du triangle rectangle OAB. Il est noté  $I_G = \frac{\text{aire de la surface de concentration}}{\text{aire du triangle OAB}}$ .

L'indice de Gini est un nombre sans dimension, compris entre 0 et 1, que l'on exprime parfois en pourcentage.

- Si  $I_G$  est proche de 0, la courbe de Lorentz est proche de la diagonale, la concentration est faible ; la concentration nulle correspond à la distribution égalitaire.
- Si  $I_G$  est proche de 1, la courbe de Lorentz est proche des côtés OA et AB, la concentration est forte ; si la concentration est proche de 1, cela signifie qu'une très faible fraction de modalités se partage la quasi-totalité de la masse totale.

### Exemple 4.5

#### Calcul de l'indice de Gini

Reprendons les données de l'exemple 4.2.

Nous rappelons qu'on obtient l'aire d'un trapèze en appliquant la formule suivante : aire = hauteur  $\times$  (grande base + petite base) / 2.

Les aires des trapèzes sont calculées dans la dernière colonne du tableau de la figure 4.9. Les valeurs  $f_i(q_{i-1}cc + q_i cc) / 2$  correspondent aux aires des trapèzes rectangles situés entre l'axe des abscisses et la courbe de Lorentz (le premier étant en fait un triangle rectangle). Leur somme indique l'aire du domaine situé sous la courbe de Lorentz.

**Figure 4.9**

**Calcul de l'aire sous la courbe de Lorentz sous Excel.**

	A	G	H	I
1	Réserve	$q_i$	$q_i cc$	$f_i(q_{i-1}cc + q_i cc)/2$
3	[0 ; 10[	3,95%	3,95%	0,0079
4	[10 ; 50[	18,97%	22,92%	0,0430
5	[50 ; 100[	17,79%	40,71%	0,0382
6	[100 ; 275[	59,29%	100,00%	0,1126
7	Somme	100,00%		0,2017

Ainsi, l'aire de la surface de concentration est égale à l'aire de OAB diminuée de la somme des aires des trapèzes.

Aire de la surface de concentration :  $0,5 - 0,2017 = 0,2983$ .

L'indice de Gini est  $I_G = 0,2983 / 0,5 = 2 \times 0,2983$ , soit  $I_G = 0,5967$ , ce qui traduit une forte concentration.

## Conclusion

Ce chapitre complète la première démarche qui a consisté à ordonner les observations et à les résumer à l'aide de graphiques et de paramètres mettant en évidence la tendance centrale et la dispersion.

Nous nous sommes attachés à caractériser la forme de la distribution et, ce faisant, à ouvrir la porte à une interprétation plus approfondie, en introduisant la distribution normale, démarche que nous compléterons avec d'autres lois de probabilité.

La mesure de la concentration est extrêmement importante pour faire ressortir des disparités sociales et économiques. Elle doit être aussi pour le lecteur l'occasion de s'assurer de la bonne maîtrise des fonctions cumulées, et des notions de masses et de médiane.

# Problèmes et exercices

Au-delà des caractéristiques de tendance centrale et de dispersion, une distribution statistique est également qualifiable par sa forme et sa concentration.

- Les exercices 1, 2 et 3 fournissent des exemples de calculs de caractéristiques de forme.
- L'exercice 4 s'attache à la notion de concentration, indissociable de l'indice de Gini.



## EXERCICE 1 CARACTÉRISTIQUES D'ASYMÉTRIE

### Énoncé

Le tableau ci-après indique la répartition du PIB par habitants (notée PPA, en euros) des pays de l'Europe des 25, hors Luxembourg, en 2001 :

PPA	Nombre de pays
[0 ; 9 000[	3
[9 000 ; 18 000[	7
[18 000 ; 27 000[	11
[27 000 ; 36 000[	3

Source : PNUD, Rapport mondial sur le développement humain, 2003

1. Dessinez l'histogramme correspondant. À partir de cet histogramme, concluez sur l'asymétrie de la distribution.
2. Concluez sur l'asymétrie de la distribution à partir du calcul des trois indicateurs suivants :
  - a. le mode ;
  - b. la moyenne ;
  - c. la médiane.
3. Concluez sur l'asymétrie de la distribution à partir du calcul des deux indicateurs suivants :
  - a. le coefficient d'asymétrie de Yule ;
  - b. le S de Pearson.
4. Concluez sur l'asymétrie de la distribution à partir du calcul des deux indicateurs suivants :
  - a. le coefficient d'asymétrie  $\beta_1$  de Pearson ;
  - b. le coefficient d'asymétrie  $\gamma_1$  de Fisher.
5. Concluez sur l'asymétrie de la distribution à partir de la boîte à moustaches.

## Solution

1. Les amplitudes de classes ( $a_i$ ) sont calculées dans la colonne C de la figure 4.10 :  
 $a_i = \sup(x_i) - \inf(x_i)$ .

Figure 4.10

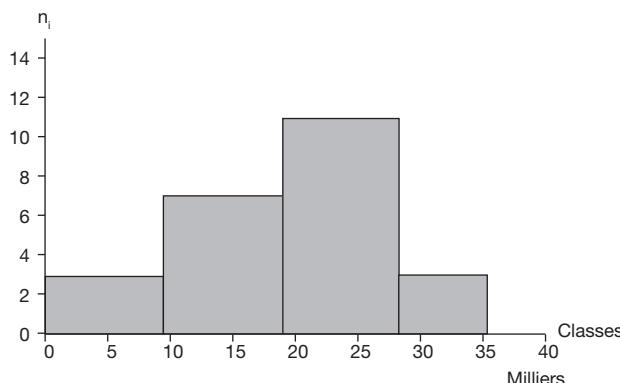
Résultats sous Excel.

	A PPA	B $n_i$	C $a_i$	D $x_i$	E $n_i x_i$	F $n_{cc}$	G $n_i x_i^2$	H $f_i(x_i - \bar{x})^2$
1								
2	[0 ; 9 000[	3	9 000	4 500	13 500	3	80 750 000	-361 705 078 125
3	[9 000 ; 18 000[	7	9 000	13 500	94 500	10	1 275 750 000	- 42 205 078 125
4	[18 000 ; 27 000[	11	9 000	22 500	247 500	21	5 568 750 000	24 169 921 875
5	[27 000 ; 36 000[	3	9 000	31 500	94 500	24	2 976 750 000	259 083 984 375
6	<b>Somme</b>	24			450 000		9 882 000 000	-120 666 250 000

Les amplitudes étant toutes égales, il n'est pas nécessaire d'utiliser les densités pour dessiner l'histogramme (voir figure 4.11), ces densités étant proportionnelles aux effectifs.

Figure 4.11.

Histogramme des PPA des pays de l'Europe des 25 (hors Luxembourg)



La réalisation de cet histogramme permet déjà de percevoir que **la distribution est asymétrique et étalée vers la gauche**.

2. a. Pour calculer le mode, nous vérifions en premier lieu que les amplitudes de classes sont égales, ici de valeur 9 000 €. La classe modale, celle qui a la plus grande densité, est donc celle qui a le plus grand effectif. Il s'agit de la classe [18 000 ; 27 000[, ce que montre bien l'histogramme.

Le mode est donc égal à  $Mo = \frac{k_2 x_1 + k_1 x_2}{k_1 + k_2} = \frac{(11-3) \times 18 000 + (11-7) \times 27 000}{(11-7) + (11-3)}$ , soit

$$Mo = 21 000 \text{ €.}$$

b. Pour calculer la moyenne, à la suite du tableau précédent, nous calculons les centres de classes ( $x_i$ ) en colonne D et les masses ( $n_i x_i$ ) en colonne E puis leur somme en cellule E6, sous Excel (voir figure 4.10).

La moyenne est égale à  $\bar{x} = \frac{1}{24} \sum_{i=1}^4 n_i x_i = \frac{450 000}{24}$ , soit  $\bar{x} = 18 750 \text{ €.}$

c. La médiane correspond à un effectif cumulé croissant de  $24 / 2 = 12$ . Les effectifs cumulés croissants ( $n_{cc}$ ) sont calculés en colonne F, à la suite du tableau précédent (voir figure 4.10).

12 est compris entre 10 et 21, donc la médiane appartient à la classe [18 000 ; 27 000[.

Par interpolation linéaire,  $Me = \frac{12-10}{21-10} \times (27\ 000 - 18\ 000) + 18\ 000$  ; soit  $Me = 19\ 636,36 \text{ €}$ .

Finalement,  $Mo > Me > \bar{x}$ , donc la distribution est asymétrique et étalée vers la gauche.

**3. a.** Le calcul du coefficient de Yule nécessite de déterminer au préalable les trois quartiles,  $Q_1$ ,  $Me$  et  $Q_3$ . La médiane a été calculée précédemment.

Le quartile d'ordre 1,  $Q_1$  correspond à un effectif cumulé croissant de  $24 / 4 = 6$ . Donc  $Q_1$  appartient à la classe  $[9\ 000 ; 18\ 000[$ .

Par interpolation linéaire,  $Q_1 = \frac{6,25-3}{10-3} \times (18\ 000 - 9\ 000) + 9\ 000$  ; soit  $Q_1 = 12\ 857,14 \text{ €}$ .

Le quartile d'ordre 3,  $Q_3$  correspond à un effectif cumulé croissant de  $24 \times 3 / 4 = 18$ . Donc  $Q_3$  appartient à la classe  $[18\ 000 ; 27\ 000[$ .

Par interpolation linéaire,  $Q_3 = \frac{18,75-10}{21-10} \times (27\ 000 - 18\ 000) + 18\ 000$  ; soit  $Q_3 = 24\ 545,45 \text{ €}$ .

D'où le coefficient de Yule  $C_Y = \frac{Q_1 + Q_3 - 2 \times Me}{Q_3 - Q_1}$ , soit

$$C_Y = \frac{12\ 857,14 + 24\ 545,45 - 2 \times 19\ 636,36}{24\ 545,45 - 12\ 857,14}.$$

D'où  $C_Y = -0,160$ . La distribution est asymétrique et étalée vers la gauche.

**b.** Le calcul du S de Pearson nécessite de déterminer au préalable le mode, la moyenne et l'écart-type. Les deux premiers indicateurs sont déjà calculés.

Pour déterminer la valeur de l'écart-type, les  $(n_i x_i^2)$  sont calculés en colonne G, à la suite du tableau précédent, puis leur somme en cellule G8 (voir figure 4.10).

Par la formule développée, la variance est égale à

$$V(x) = \frac{1}{24} \sum_{i=1}^4 n_i x_i^2 - \bar{x}^2 = \frac{9\ 882\ 000\ 000}{24} - 18\ 750^2, \text{ soit } V(x) = 60\ 187\ 500.$$

L'écart-type est égal à  $\sigma_x = \sqrt{V(x)} = \sqrt{60\ 187\ 500}$ , soit  $\sigma_x = 7\ 758$ .

D'où  $S = \frac{\bar{x} - Mo}{\sigma} = \frac{18\ 750 - 21\ 000}{7758}$ , soit  $S = -0,290$ . La distribution est asymétrique et étalée vers la gauche.

**4. a.** Le calcul du  $\beta_1$  de Pearson et du  $\gamma_1$  de Fisher nécessite de connaître la valeur de  $\mu_3$ , le moment centré d'ordre 3 défini par  $\mu_3 = \frac{1}{n} \sum_{i=1}^k n_i (x_i - \bar{x})^3 = \sum_{i=1}^k f_i (x_i - \bar{x})^3$ . Les  $f_i (x_i - \bar{x})^3$  sont calculés en colonne H, à la suite du tableau précédent, puis leur somme en cellule H6 (voir figure 4.10).

De là,  $\mu_3 = \sum_{i=1}^4 f_i (x_i - \bar{x})^3$ , soit  $\mu_3 = -120\ 656\ 250\ 000$ .

b. Sachant que  $\mu_2 = V(x)$ ,  $\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(-120\ 656\ 250\ 000)^2}{60\ 187\ 500^3}$ , soit  $\beta_1 = 0,067$ .  $\beta_1$  positif

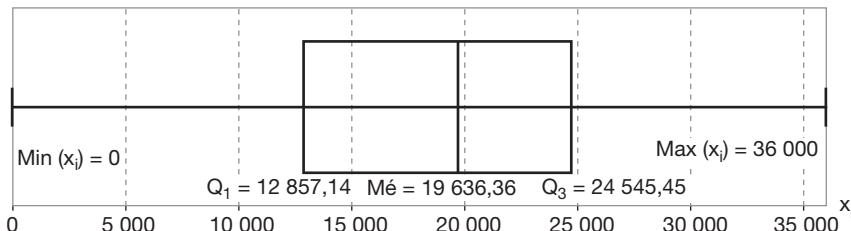
permet de conclure que la distribution est asymétrique et  $\mu_3$  négatif permet de conclure qu'elle est étalée vers la gauche.

De même,  $\gamma_1 = \frac{\mu_3}{\sigma^3} = \frac{-120\ 656\ 250\ 000}{7758^3}$  soit  $\gamma_1 = -0,258$ .  $\gamma_1$  permet de conclure que la distribution est asymétrique et étalée vers la gauche.

5.

**Figure 4.12**

**Boîte à moustaches.**



Ce diagramme permet de visualiser l'étalement vers la gauche de la distribution, la médiane étant plus proche de  $Q_3$  que de  $Q_1$ .



## EXERCICE 2 CARACTÉRISTIQUES D'APLATISSEMENT

### Énoncé

Le tableau ci-après indique la répartition du PIB par habitants (PPA) des pays de l'Europe des 25, en 2001 :

% de la population âgée de 65 ans ou plus	Nombre de pays
11	1
12	2
13	2
14	3
15	5
16	5
17	3
18	3
19	1

Source : PNUD, Rapport mondial sur le développement humain, 2003

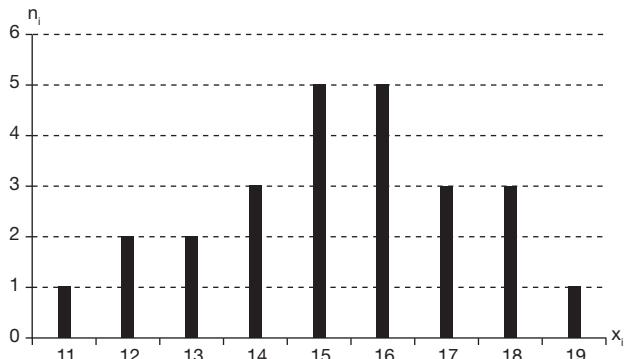
1. Dessinez le diagramme en bâtons correspondant.
2. Calculez le coefficient d'aplatissement de Pearson.
3. Calculez le coefficient d'aplatissement de Fisher.

**Solution**

1.

**Figure 4.13**

**Diagramme en bâtons du pourcentage de la population âgée de 65 ans ou plus des pays de l'Europe des 25.**



2. Le calcul du  $\beta_2$  de Pearson nécessite de connaître la valeur de la variance et de  $\mu_4$ , le moment centré d'ordre 4 défini par :  $\mu_4 = \frac{1}{n} \sum_{i=1}^k n_i (x_i - \bar{x})^4 = \sum_{i=1}^k f_i (x_i - \bar{x})^4$ .

Saisissez les modalités dans la colonne L1 et les effectifs dans la colonne L2 (voir figure 4.14).

**Figure 4.14**

**Saisie du tableau de données avec la calculatrice.**

L1	L2	L3	$\bar{x}$
11	1		
12	2		
13	2		
14	3		
15	5		
16	5		
17	3		
		L3(1)=	

Dans le menu CALC de STAT,appelez la fonction 1-Var Stats, puis indiquez dans l'ordre L1, L2 comme suit : 1-Var Stats L<sub>1</sub>,L<sub>2</sub>. Validez avec **ENTER**. Les résultats de la figure 4.15 s'affichent.

**Figure 4.15**

Résultats de l'analyse statistique effectuée avec la calculatrice.

```

1-Var Stats
x̄=15,28
Σx=382
Σx²=5940
Sx=2.072036036
σx=2.030172406
n=25

```

La moyenne est  $\bar{x} = 15,28$ .

La variance est égale à  $V(x) = \sigma_x^2 = 1,4447^2$ , soit  $V(x) = 2,096$ .

Pour calculer les  $f_i(x_i - \bar{x})^4$  dans la colonne L3, placez le curseur sur l'en-tête de colonne L3. Indiquez  $L3=L2÷25*(L1-15,28)^4$  puis appuyez sur **ENTER**.

Pour calculer leur somme, placez le curseur dans la cellule L3(10), et indiquez  $L3(10)=SUM(L3)$  en appelant la fonction SUM (voir annexe 1.2). Validez avec **ENTER**.

De là,  $\mu_4 = \sum_{i=1}^9 f_i(x_i - \bar{x})^4$ , soit  $\mu_4 = 40,5$  (voir figure 4.16).

**Figure 4.16**

Calcul de  $\mu_4$  avec la calculatrice.

L1	L2	L3	3
15	5	.00123	
16	5	.05375	
17	3	1.0503	
18	3	6.5684	
19	1	7.6601	
-----	-----	40,5	
L3(11) =			

D'où le coefficient d'aplatissement de Pearson  $\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{40,50}{4,12^2}$ , soit  $\beta_2 = 2,384$ .

**La distribution est platicurtique.**

3. Le coefficient d'aplatissement de Fisher  $\gamma_2 = \beta_2 - 3 = 2,384 - 3$ , soit  $\gamma_2 = -0,616$ .

**La distribution est platicurtique, c'est-à-dire plus « plate » que la distribution normale.**



### EXERCICE 3 CARACTÉRISTIQUES DE FORME ET MÉDIALE

#### Énoncé

Le tableau ci-après indique la répartition des salaires annuels bruts, par tranches, de l'entreprise Alpha :

Salaires (K€)	Effectifs
[25 ; 35[	22
[35 ; 45[	28

Salaires (K€)	Effectifs
[45 ; 55[	37
[55 ; 65[	51
[65 ; 80[	32
[80 ; 100[	12
[100 ; 120[	7

1. Dessinez l'histogramme correspondant.
2. Calculez la médiale. Interprétez.
3. Concluez sur la forme de la distribution à partir du calcul des deux coefficients suivants :
  - a. le coefficient d'asymétrie  $\beta_1$  de Pearson ;
  - b. le coefficient d'aplatissement  $\beta_2$  de Pearson.

**Solution**

1. Saisissez les centres de classes (modalités) dans la colonne L1, les effectifs dans la colonne L2 et les amplitudes de classes ( $a_i$ ) dans la colonne L3 (voir figure 4.17).

Comme les amplitudes de classes ne sont pas toutes égales, il est nécessaire de passer par les densités  $d_i$ . Pour calculer les densités, placez le curseur sur l'en-tête de colonne L4. Indiquez  $L4=L2/L3$  puis appuyez sur **ENTER** (voir figure 4.18).

**Figure 4.17 (gauche)**

Saisie du tableau de données avec la calculatrice.

L1	L2	L3	3
40	28	10	
50	37	10	
60	51	10	
72,5	32	15	
90	12	20	
110	7	20	
-----	-----	-----	-----

**Figure 4.18 (droite)**

Calcul des densités avec la calculatrice.

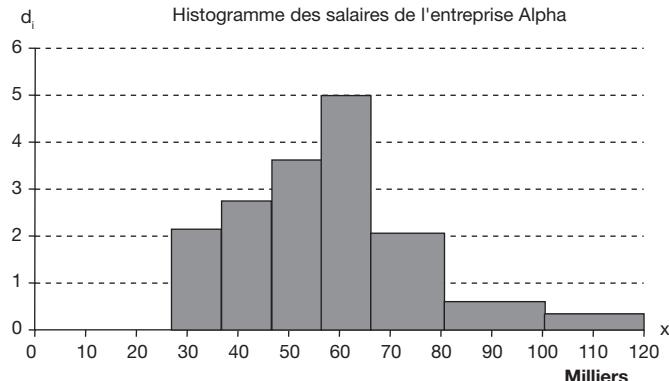
L2	L3	L4	4
28	10	2.8	
37	10	3.7	
51	10	5.1	
32	15	2.1333	
12	20	6	
7	20	35	
-----	-----	-----	-----

L3(B) = **L4(1)=2.2**

L'histogramme peut alors être dessiné d'après ces densités (voir figure 4.19).

**Figure 4.19**

**Histogramme des salaires de l'entreprise Alpha.**



2. La médiale est l'équivalent de la médiane sur la masse salariale (ici, la masse salariale est donnée par  $\sum_{i=1}^7 n_i x_i$ ), puisqu'elle partage la population en deux sous-populations de masses salariales égales.

Pour calculer les ( $n_i x_i$ ) dans la colonne L5, placez le curseur sur l'en-tête de colonne L5. Indiquez L5=L2\*L1, puis appuyez sur **ENTER**.

Pour obtenir les  $n_i x_i$  cumulés croissants ( $n_i x_{cc}$ ) dans la colonne L6, placez le curseur sur l'en-tête de colonne L6, puis entrez la formule L6=CumSum(L5), en appelant la fonction CUMSUM (voir annexe 1.2), puis appuyez sur **ENTER** (voir figure 4.20).

**Figure 4.20**

**Calcul des  $n_i x_i$  et des  $n_i x_i$  cumulés croissants avec la calculatrice.**

L4	L5	L6	
2.2	660	660	
2.8	1120	1780	
3.7	1850	3630	
5.1	3060	6690	
2.1333	2320	9010	
.6	1080	10090	
.35	770	10860	
<b>L6(1)=660</b>			

La médiale correspond à une masse relative cumulée croissante de :  $10\ 860 / 2 = 5\ 430$ , valeur comprise entre 3 630 et 6 690, donc la médiale appartient à la classe [55 ; 65[.

Par interpolation linéaire,  $Ml = \frac{5\ 430 - 3\ 630}{6\ 690 - 3\ 630} \times (65 - 55) + 55$  ; soit **Ml = 60,88 K€**. Les salariés qui perçoivent moins de 60 880 € de salaire annuel brut se partagent la moitié de la masse salariale.

3. a. Dans le menu CALC de STAT,appelez la fonction 1-Var Stats, puis indiquez dans l'ordre L1, L2 comme suit : 1-Var Stats L<sub>1</sub>,L<sub>2</sub>, Validez avec **ENTER**. Les résultats de la figure 4.21 s'affichent.

Figure 4.21

Résultats de l'analyse statistique effectuée avec la calculatrice.

1-Var Stats  
 $\bar{x}=57,46031746$   
 $\Sigma x=10860$   
 $\Sigma x^2=690800$   
 $Sx=18,84722345$   
 $\sigma x=18,79729694$   
 $\downarrow n=189$

La moyenne est  $\bar{x} = 57,46$ .

La variance est égale à  $V(x) = \sigma_x^2 = 18,7973^2$ , soit  $V(x) = 353,34$ .

Pour calculer les  $f_i(x_i - \bar{x})^3$  dans la colonne L7, placez le curseur sur l'en-tête de colonne L7 et nommez-la LA. Indiquez  $LA=L2\div189*(L1-57,46)^3$  puis appuyez sur **ENTER**. Pour calculer leur somme, placez le curseur dans la cellule LA(8), et indiquez  $LA(8)=SUM(LA)$  en appelant la fonction SUM (voir annexe 1.2) puis la ligne LA par le menu **LIST**, **NAMES**, 7:LA. Validez avec **ENTER**.

Pour calculer les  $f_i(x_i - \bar{x})^4$  dans la colonne L8, placez le curseur sur l'en-tête de colonne L8 et nommez-la LB. Indiquez  $LB=L2\div189*(L1-57,46)^4$  puis appuyez sur **ENTER**. Pour calculer leur somme, placez le curseur dans la cellule LB(8), et indiquez  $LB(8)=SUM(LB)$  en appelant la fonction SUM puis la ligne LB par le menu **LIST**, **NAMES**, 8:LB. Validez avec **ENTER**.

$$\mu_3 = \sum_{i=1}^7 f_i(x_i - \bar{x})^3, \text{ soit } \mu_3 = 4859,6.$$

Sachant que  $\mu_2 = V(x)$ ,  $\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{4859,6^2}{353,34^3}$ , soit  $\beta_1 = 0,535$ .  $\beta_1$  positif permet de conclure

que la distribution est asymétrique et  $\mu_3$  positif permet de conclure qu'elle est étalée vers la droite.

$$\mu_4 = \sum_{i=1}^7 f_i(x_i - \bar{x})^4, \text{ soit } \mu_4 = 442\,645 \text{ (voir figure 4.22).}$$

Figure 4.22

Calcul de  $\mu_3$  et de  $\mu_4$  avec la calculatrice.

L6	LA	LB	B
3630	-81,27	606,31	
6690	4,4219	11,232	
9010	576,01	8663,2	
10090	2187,6	71185	
10860	5371,6	282226	
-----	4859,6	442645	
	LA(8) =		

b. D'où le coefficient d'aplatissement de Pearson  $\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{442\,645}{353,34^2}$ , soit  $\beta_2 = 3,545$ .

La distribution est leptocurtique, c'est-à-dire plus pointue que la distribution normale.



## EXERCICE 4 CARACTÉRISTIQUE DE CONCENTRATION : L'INDICE DE GINI

### Énoncé

Le tableau ci-après indique la répartition des 22 régions françaises selon le nombre de lits dont elles disposent en maisons de retraite au 1<sup>er</sup> janvier 2005 :

NOMBRE DE LITS	NOMBRE DE RÉGIONS
[0 ; 12 250[	4
[12 250 ; 24 500[	12
[24 500 ; 36 750[	4
[36 750 ; 49 000[	2

Source : ministère de la Santé et des Solidarités, enquêtes EHPA, FINESS, SAE, 2005

1. Calculez la médiale.
2. Représentez la courbe de concentration.
3. Calculez l'indice de Gini. Interprétez.

### Solution

1. Les centres de classes sont calculés en colonne C, les fréquences ( $f_i$ ) en colonne D puis les fréquences cumulées croissantes ( $f_{i,cc}$ ) en colonne E (voir figure 4.23).

Les ( $n_i x_i$ ) sont calculés en colonne F. Leur somme représente la masse totale des lits disponibles en maisons de retraite dans les 22 régions françaises. La médiale partage la population en deux sous-populations de masses égales.

La quote-part  $q_i$  des masses dans la masse salariale ( $q_i$ ) est calculée en colonne G et leurs pourcentages cumulés croissants ( $q_{i,cc}$ ) sont calculés en colonne H.

Figure 4.23

Résultats sous Excel.

	A	B	C	D	E	F	G	H	I
1	Nombre de lits	$n_i$	$x_i$	$f_i$	$f_{i,cc}$	$n_i x_i$	$q_i$	$q_{i,cc}$	$S_i$
3	[0 ; 12 250[	4	6 125	18,18%	18,18%	24 500	5,41%	5,41%	0,0049
4	[12 250 ; 24 500[	12	18 375	54,65%	72,73%	220 500	48,65%	64,05%	0,1622
5	[24 500 ; 36 750[	4	30 625	18,18%	90,91%	122 500	27,03%	81,08%	0,1229
6	[36 750 ; 49 000[	2	42 875	9,09%	100,00%	85 750	18,92%	100,00%	0,0823
7	Somme	22		100,00%		453 250	100,00%		0,3722

La médiale se trouve dans l'intervalle où  $q_{i,cc}$  passe à 50 %, c'est-à-dire [12 250 ; 24 500[.

Par interpolation linéaire (voir chapitre 2),

$$Ml = \frac{0,5 - 0,0541}{0,5405 - 0,0541} \times (24\ 500 - 12\ 250) + 122\ 500, \text{ soit } Ml = 23\ 479,17 \text{ lits. } 50 \% \text{ des lits}$$

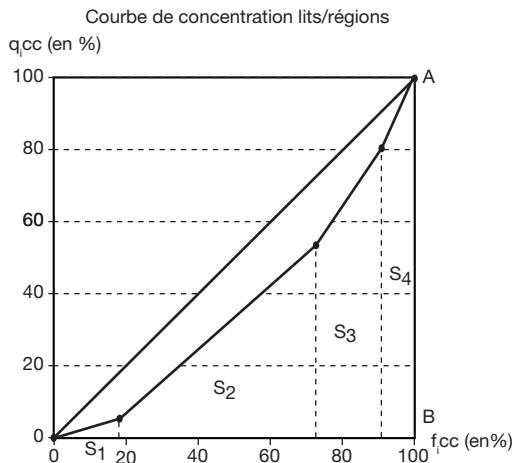
disponibles en maisons de retraite françaises proviennent de régions qui ont moins de 23 479 lits.

2. La courbe de concentration est obtenue en portant en abscisses les fréquences cumulées croissantes, notées  $f_{i,cc}$  (colonne E) et les  $q_{i,cc}$  (colonne H) en ordonnées. À la lecture de la ligne 4 du tableau Excel de la figure 4.23, il est possible de conclure que

72,73 % des régions détiennent 54,05 % des lits disponibles dans les maisons de retraite françaises (voir figure 4.24).

**Figure 4.24**

**Courbe de concentration des lits selon les régions.**



3. L'aire de la surface sous la courbe de concentration se calcule par la méthode des trapèzes. L'aire de chaque trapèze ( $S_i$ ) est calculée dans la colonne I, puis leur somme dans la cellule I6 (voir figure 4.23).

La première surface,  $S_1$ , est un triangle dont l'aire est égale à  $S_1 = \frac{f_1 \times q_1 cc}{2} = \frac{0,1818 \times 0,0541}{2} = 0,0049$ . La deuxième,  $S_2$ , est un trapèze d'aire

$$S_2 = \frac{f_2 \times (q_1 cc + q_2 cc)}{2} = \frac{0,5455 \times (0,0541 + 0,5405)}{2} = 0,1622.$$

$$\text{De même, } S_3 = \frac{f_3 \times (q_2 cc + q_3 cc)}{2} = \frac{0,1818 \times (0,5405 + 0,8108)}{2} = 0,1229.$$

$$\text{Et } S_4 = \frac{f_4 \times (q_3 cc + q_4 cc)}{2} = \frac{0,0909 \times (0,8108 + 1)}{2} = 0,0823.$$

L'aire de la surface située entre la courbe de concentration et l'axe des abscisses est la somme des aires des trapèzes.  $S = \sum_{i=1}^4 S_i = 0,3722$ . La surface de concentration, notée SC, est le domaine situé entre la diagonale du carré et la courbe de Lorentz. Son aire est égale à la différence entre l'aire du triangle rectangle OAB, soit  $\frac{1 \times 1}{2}$ , et la somme des aires des trapèzes calculée. D'où  $SC = 0,5 - 0,3722 = 0,1278$ .

D'où l'indice de Gini,  $I_G = \frac{0,1278}{0,5}$ , soit  $I_G = 0,2555$ . La concentration est faible, car

l'indice de Gini est plus proche de 0 que de 1. Autrement dit, les lits en maisons de retraite ne sont pas concentrés au sein de quelques régions françaises, mais sont relativement bien répartis sur ces régions.

# Bibliographie

- BAILLARGEON G., *Méthodes statistiques de l'ingénieur*, SMG, 1990.
- CALOT G., *Cours de statistique descriptive*, Dunod, 1969.
- CHAREILLE P. et PINAULT Y., *Statistique descriptive*, Collection AES, Montchrestien, Paris, 1996.
- DELMAS B., *Statistique descriptive*, Armand Colin, 2005.
- DELECROIX M., *Histogrammes et estimation de la densité*, Que sais-je ?, PUF, 1983.
- DODGE Y., *Statistique. Dictionnaire encyclopédique*, Springer, 2004.
- LIORZOU A., *Initiation à la pratique statistique*, Eyrolles, 1985.
- SAPORTA G., *Probabilités, analyse de données et statistique*, Technip, 1990.
- SCHLACHTHER D., *De l'analyse à la prévision*, Ellipses, 1986.
- TASSI Ph. et LEGAIT S., *Théorie des probabilités en vue des applications statistiques*, Technip, 1990.

# Les séries bivariées

1. Présentation des données....	108
2. Les caractéristiques des séries à deux caractères ....	113
3. Étude des liaisons entre deux variables .....	119
<b>Problèmes et exercices</b>	
1. Construction d'un tableau de contingence sur caractères discret et qualitatif .....	127
2. Construction d'un tableau de contingence sur caractères continus .....	131
3. Contenu d'un tableau de contingence.....	132
4. Indicateurs sur tableau de contingence.....	135
5. Dépendance entre deux variables .....	139

Dans de nombreuses sciences – démographie, médecine, économie –, le statisticien est amené à étudier plusieurs caractères sur une même population. L'évolution d'un caractère avec le temps est de la plus grande importance et donne lieu à l'étude des séries chronologiques, qui constituent un cas particulier des séries bivariées, c'est-à-dire des séries visant à étudier conjointement deux variables mesurées sur un même individu. Les modalités sont donc des couples et les données sont présentées dans des tableaux élémentaires ou dans des tableaux à double entrée, encore appelés tableaux de contingence. L'analyse de ces tableaux vise à mettre en évidence d'éventuelles relations ou corrélations entre les deux variables. Le concept de corrélation (« co-relation ») est né vers 1880, avec les travaux de Francis Galton. Karl Pearson a ensuite utilisé la notion de contingence dans le sens de mesure de la déviation par rapport à l'indépendance. Ce contexte sera l'occasion de s'initier à la théorie des tests statistiques, dont la paternité est attribuée à la collaboration (1925-1930) entre Jerzy Neyman et Egon Pearson, dénommé Pearson « deux », le fils de Karl.

Nous n'oublierons pas qu'à partir d'un tableau concernant deux variables nous pourrons toujours extraire les séries concernant chacun des caractères, encore appelées séries marginales.

Comme nous le verrons dans les différents exemples, les caractères étudiés peuvent être de même type, qualitatifs ou quantitatifs (discrets ou continus), ou de natures différentes, l'un qualitatif et l'autre quantitatif.

# 1 Présentation des données

Il existe deux façons de présenter une série bivariée :

- les tableaux simples, composés des observations en ligne et des variables en colonne ;
- les tableaux de contingence, qui croisent les modalités des deux variables.

## 1.1 DONNÉES EXHAUSTIVES : TABLEAUX SIMPLES

Les tableaux simples des séries bivariées sont constitués :

- des observations en ligne ;
- des deux variables en colonne.

Ainsi, chaque ligne comporte l'identifiant de l'observation dans la première colonne et les modalités observées pour chacune des deux variables dans les deux colonnes suivantes.

### Exemple 5.1

#### Série bivariée et tableau simple

Le tableau suivant indique, pour chacune des trois académies d'Île-de-France, le nombre de licenciés en 2005 et le nombre de licenciés poursuivant leurs études à l'université, en 2006. Il recense ainsi la poursuite des études à l'université après la licence.

Académie	Nombre de licenciés (2005)	Licenciés à l'université (2006)
Paris	14 150	11 271
Créteil	7 759	5 150
Versailles	7 254	5 107
<b>Total</b>	<b>29 163</b>	<b>21 528</b>

Source : ministère de l'Éducation nationale, 2006

Cette série double, ou bivariée, comporte trois modalités. Si l'on note X le nombre de licenciés en 2005 et Y le nombre de licenciés poursuivant leurs études à l'université en 2006, Créteil est représentée par la modalité  $(x_2; y_2) = (7\ 759; 5\ 150)$ . En exploitant chaque variable une par une, il est possible de calculer tous les indicateurs des séries univariées, comme les moyennes. Ainsi,  $\bar{x} = \frac{29\ 163}{3} = 9\ 721$  ; le nombre moyen de licenciés est de

9 721 étudiants par académie. De même,  $\bar{y} = \frac{21\ 528}{3} = 7\ 176$  ; le nombre moyen de licenciés poursuivant leurs études à l'université est de 7 176 étudiants par académie.

On représente cette série en plaçant dans un repère les trois points de coordonnées  $(x_i; y_i)$  pour  $i$  entier variant de 1 à 3 ; cette représentation s'appelle un nuage de points. Le point  $G$  de coordonnées respectives  $\bar{x}$  et  $\bar{y}$  est appelé point moyen du nuage.

Dans une série double de ce type, les effectifs de chaque modalité sont égaux à 1 et ne sont pas mentionnés. Il est possible de calculer sur les séries marginales les moyennes et tous les paramètres nécessaires à l'étude de la série bivariée, comme les variances.

Un nouveau paramètre, la covariance, sera introduit à la section 2.4.

## 1.2 LE TABLEAU CROISÉ OU TABLEAU DE CONTINGENCE

Les tableaux croisés ou tableaux de contingence sont les tableaux obtenus quand on étudie une population sous l'angle de deux caractères que l'on croise. Dans le cas particulier où ces caractères ont chacun deux modalités (cas binaire), on obtient le cas particulier des tableaux  $2 \times 2$ .

Côté « technique », les tableaux à double entrée ne sont pas différents de ce que les mathématiciens appellent matrices, tableaux de nombres à  $n$  lignes et  $p$  colonnes et dont nous noterons  $n_{ij}$  le terme situé à l'intersection de la ligne  $i$  et de la colonne  $j$ .

Une des caractéristiques des tableaux de contingence, qui sont très présents dans l'analyse de données, est d'attribuer un sens aux « marges », c'est-à-dire à une colonne supplémentaire à droite et à une ligne supplémentaire en bas, qui indiquent le nombre d'individus possédant une des modalités de l'un des deux caractères (ce que l'on appelle le tri à plat de ce caractère).

### Présentation des effectifs du tableau de contingence

Soit respectivement  $p$  et  $q$  les nombres de modalités des caractères X et Y.

- Les modalités du caractère X se notent  $x_i$  avec  $i = \{1, 2, \dots, p\}$ .
- Les modalités du caractère Y se notent  $y_j$  avec  $j = \{1, 2, \dots, q\}$ .

#### Définitions

**L'effectif partiel** de la modalité  $(x_i, y_j)$  est le nombre d'observations présentant simultanément les deux modalités  $x_i$  et  $y_j$ . Il se note  $n_{ij}$ .

**L'effectif marginal** de la modalité  $x_i$  se note  $n_{i+}$ , ou encore  $n_{+i}$ , tel que :  $n_{i+} = \sum_{j=1}^q n_{ij}$ . Cet effectif

désigne la somme des effectifs de la ligne  $i$ . La distribution des effectifs marginaux de X s'appelle distribution marginale de X.

De même,  $n_{+j} = \sum_{i=1}^p n_{ij}$ , ou encore  $n_{*j}$ , est l'effectif marginal de la modalité  $y_j$ . Il désigne la somme des effectifs de la colonne  $j$ . La distribution des effectifs marginaux de Y s'appelle distribution marginale de Y.

**L'effectif total** de la série double est la somme des effectifs marginaux de la série X (ou Y).

Il est noté  $n_{++}$ ,  $n_{..}$  ou simplement  $n$ , avec :  $n_{++} = \sum_{j=1}^q n_{+j} = \sum_{i=1}^p n_{i+} = \sum_{j=1}^q \sum_{i=1}^p n_{ij} = \sum_{i=1}^p \sum_{j=1}^q n_{ij}$ .

En adoptant l'ensemble de ces notations, le tableau de contingence contenant les effectifs se présente de la manière suivante :

- Les modalités  $x_i$  de X apparaissent dans la première colonne.
- Les modalités  $y_j$  de Y apparaissent sur la première ligne.
- L'effectif partiel  $n_{ij}$  de la modalité  $(x_i, y_j)$  est inscrit au croisement de la ligne  $i$  et de la colonne  $j$ .
- L'effectif marginal  $n_{i+}$  de X est reporté dans la dernière colonne du tableau. L'effectif marginal  $n_{+j}$  de Y est reporté sur la dernière ligne du tableau. La dernière ligne et la dernière colonne du tableau de contingence s'appellent les marges et contiennent la distribution marginale de X et de Y. Elles représentent les effectifs des séries simples X et Y.
- L'effectif total  $n_{++}$  est indiqué au croisement des deux distributions marginales de X et de Y.

D'où la présentation suivante du tableau de contingence :

<b>X \ Y</b>	<b><math>y_1</math></b>	<b><math>y_2</math></b>	<b>...</b>	<b><math>y_j</math></b>	<b>...</b>	<b><math>y_q</math></b>	<b><math>n_{i+}</math></b>
$x_1$	$n_{11}$	$n_{12}$	...	$n_{1j}$	...	$n_{1q}$	$n_{1+}$
$x_2$	$n_{21}$	$n_{22}$	...	$n_{2j}$	...	$n_{2q}$	$n_{2+}$
...	...	...	...	...	...	...	...
$x_i$	$n_{i1}$	$n_{i2}$	...	$n_{ij}$	...	$n_{iq}$	$n_{i+}$
...	...	...	...	...	...	...	...
$x_p$	$n_{p1}$	$n_{p2}$	...	$n_{pj}$	...	$n_{pq}$	$n_{p+}$
<b><math>n_{+j}</math></b>	<b><math>n_{+1}</math></b>	<b><math>n_{+2}</math></b>	...	<b><math>n_{+j}</math></b>	...	<b><math>n_{+q}</math></b>	<b><math>n_{++}</math></b>

### Exemple 5.2

#### Un tableau de contingence $2 \times 2$

Certaines entreprises mettent en avant auprès des consommateurs des engagements de « citoyenneté » : par exemple, fabriquer sans générer de pollution, ne pas avoir recours au travail des enfants, etc. Le tableau ci-après donne le résultat sur un échantillon constitué sur la base des résultats d'une enquête du Credoc de l'année 2006, en réponse à la question : « D'une façon générale, tenez-vous compte de ces éléments lorsque vous achetez un produit ? »

<b>Sexe \ Réponse</b>	<b>Oui (O)</b>	<b>Non (N)</b>	<b>Total</b>
Masculin (M)	290	410	700
Féminin (F)	141	159	300
<b>Total</b>	<b>431</b>	<b>569</b>	<b>1 000</b>

Source : Credoc, 2006

Ce tableau comporte deux caractères qualitatifs :

- X, le sexe, avec les deux modalités  $x_1 = M$  et  $x_2 = F$  ;

- Y, la réponse à la question de citoyenneté, avec les deux modalités  $y_1 = O$  et  $y_2 = N$ . L'effectif total est de 1 000 et il y a quatre modalités. Par exemple, le couple  $(x_1 ; y_2) = (M ; N)$  a un effectif situé au croisement de la première ligne et de la deuxième colonne et noté  $n_{12} = 410$ .

Les sommes des effectifs en ligne sont indiqués dans la dernière colonne et les sommes des effectifs en colonne sur la deuxième ligne du tableau.

- Par exemple, la somme des effectifs de la deuxième ligne,  $n_{2+} = 300$  est indiquée dans la dernière colonne, sur la dernière ligne. Il s'agit du nombre total de femmes, également appelé effectif marginal de la modalité Féminin de la variable Sexe.
- De même, la somme des effectifs de la première colonne  $n_{+1} = 431$  est donnée sur la dernière ligne. Il s'agit de l'effectif de la modalité Oui, sans distinction de sexe, également appelé effectif marginal de la modalité Oui de la variable Réponse.

Le détail des effectifs de cette série est donné dans la présentation générale suivante :

<b>X \ Y</b>	<b><math>y_1 = O</math></b>	<b><math>y_2 = N</math></b>	<b><math>n_{i+}</math></b>
$x_1 = M$	$n_{11} = 290$	$n_{12} = 410$	$n_{1+} = 700$
$x_2 = F$	$n_{21} = 141$	$n_{22} = 159$	$n_{2+} = 300$
$n_{+j}$	$n_{+1} = 431$	$n_{+2} = 569$	$n_{++} = 1000$

### Les fréquences des tableaux de contingence

À partir du tableau de contingence composé des effectifs, il est possible de calculer les fréquences (fréquences relatives). Il existe trois types de fréquences :

- les fréquences partielles ;
- les fréquences marginales ;
- les fréquences conditionnelles.

#### Définitions

La **fréquence partielle** de la modalité  $(x_i, y_j)$  est notée  $f_{ij}$  et est définie par  $f_{ij} = \frac{n_{ij}}{n_{++}}$ . Il est clair que  $\sum_{j=1}^q \sum_{i=1}^p f_{ij} = 1$ . On retrouve le concept d'intersection, ces individus appartenant à la modalité  $x_i$  de X et à la modalité  $y_j$  de Y.

La **fréquence marginale** de la modalité  $x_i$  est notée  $f_{i+}$  et est définie par  $f_{i+} = \frac{n_{i+}}{n_{++}}$ . Il est clair que  $f_{i+} = \sum_{j=1}^q f_{ij}$ .

De même, la **fréquence marginale** de la modalité  $y_j$  est notée  $f_{+j}$  et est définie par  $f_{+j} = \frac{n_{+j}}{n_{++}} = \sum_{i=1}^p f_{ij}$ .

### Exemple 5.3

#### Calcul des fréquences partielles et marginales sur tableau de contingence

Reprendons le tableau de contingence de l'exemple 5.2 ci-avant.

Il est possible de déterminer les fréquences partielles  $f_{ij}$ . Par exemple,  $f_{12} = 410 / 1000 = 0,41$ , soit

it 41 % des individus de notre enquête sont des hommes et ont répondu non.

Il est également possible de déterminer les fréquences marginales  $f_{i+}$  ou  $f_{+j}$ . Par exemple,  $f_{+1} = 431 / 1000 = 0,431$ , soit 43,1 % des individus de l'enquête ont répondu oui.

Les séries marginales peuvent éventuellement être extraites. Par exemple, l'extraction de la série marginale du caractère X donne :

Sexe	$n_{i+}$
Masculin (M)	700
Féminin (F)	300
<b>Total</b>	1 000

Cette présentation pourra faciliter les calculs de fréquence, moyenne, variance et écart-type dans le cas des caractères quantitatifs. Par exemple, ici, les fréquences marginales du caractère Sexe sont aisément repérables :  $f_{1+} = 0,70$  et  $f_{2+} = 0,30$ , soit 70 % d'hommes et 30 % de femmes.

Les fréquences conditionnelles nous permettent d'aborder la distribution conditionnelle. Cette distribution est à relier à la notion de probabilité conditionnelle, qui consiste à effectuer un changement de l'univers ou de la population étudiés (voir P. Roger, page 17). Cela revient à effectuer les calculs sur une sous-population présentant une modalité choisie au lieu de s'intéresser à la population entière.

### Définitions

**Distributions conditionnelles** : Si le caractère Y possède q modalités, on peut définir q distributions conditionnelles de X sachant Y. Les effectifs de ces distributions sont représentés par chacune des colonnes du tableau de contingence. L'effectif total de la distribution conditionnelle de X sachant  $Y = y_i$  étant alors  $n_{+i}$ .

De même, si le caractère X possède p modalités, on peut définir p distributions conditionnelles de Y sachant X. Les effectifs de ces distributions sont représentés par chacune des lignes du tableau de contingence. L'effectif total de la distribution conditionnelle de  $Y = y_i$  sachant  $X = x_i$  étant alors  $n_{i+}$ .

**Fréquences conditionnelles de X sachant Y** : La fréquence conditionnelle de la modalité  $x_i$  sachant  $y_i$  est donnée par  $f_{X=x_i/Y=y_i} = \frac{n_{ij}}{n_{+i}}$ . Ainsi,  $\sum_{i=1}^p f_{X=x_i/Y=y_i} = 1$ . Elle est aussi notée  $f_{i/+i}$ .

**Fréquences conditionnelles de Y sachant X** : La fréquence conditionnelle de la modalité  $y_i$  sachant  $x_i$  est donnée par  $f_{Y=y_i/X=x_i} = \frac{n_{ij}}{n_{i+}}$ . Ainsi,  $\sum_{j=1}^q f_{Y=y_j/X=x_i} = 1$ . Elle est aussi notée  $f_{i+j+}$ .

Il existe une relation entre les fréquences conditionnelles et les fréquences partielles précédemment définies :  $f_{ij} = f_{i/+j} \times f_{+/j}$ . Cette relation est similaire au théorème des probabilités composées qui indique que :

$$P((X=x_i) \cap (Y=y_j)) = P_{(Y=y_j)}(X=x_i) \times P(Y=y_j).$$

#### Exemple 5.4

#### Calcul des fréquences conditionnelles sur tableau de contingence : sexe et citoyenneté

Reprendons le tableau de contingence de l'exemple 5.2 ci-avant.

Au lieu de s'intéresser à la population entière, il est possible de s'intéresser à l'univers des femmes. L'univers de travail est alors la sous-population notée  $\{X = x_2\}$ . Elle est constituée des individus présentant la modalité F de la variable X.

Cherchons alors la proportion de réponses Oui, soit d'individus appartenant à la modalité  $y_1$  de Y dans cette sous-population. Cette fréquence conditionnelle est notée indifféremment  $f_{Y=y_1/X=x_2}$ ,  $f_{j=1/i=2}$  ou  $f_{1/2+}$  (on lit « f indice j = 1 sachant i = 2 » si les indices i et j ont été respectivement affectés aux modalités de X et de Y) et définie par :  $f_{j=1/i=2} = \frac{n_{21}}{n_{2+}} = \frac{141}{300} = 0,47$  ; ainsi, 47 % des femmes ont répondu oui.

Il est ainsi possible de calculer toutes les fréquences conditionnelles de X sachant Y.

<b>X \ Y</b>	<b><math>y_1 = O</math></b>	<b><math>y_2 = N</math></b>
$x_1 = M$	$f_{i=1/j=1} = 0,6729$	$f_{i=1/j=2} = 0,7206$
$x_2 = F$	$f_{i=2/j=1} = 0,3271$	$f_{i=2/j=2} = 0,2794$
$f_{+j}$	1	1

De même, il est possible de calculer toutes les fréquences conditionnelles de Y sachant X.

<b>X \ Y</b>	<b><math>y_1 = O</math></b>	<b><math>y_2 = N</math></b>	<b><math>f_{i+}</math></b>
$x_1 = M$	$f_{j=1/i=1} = 0,4143$	$f_{j=2/i=1} = 0,5857$	1
$x_2 = F$	$f_{j=1/i=2} = 0,47$	$f_{j=2/i=2} = 0,53$	1

## 2 Les caractéristiques des séries à deux caractères

Les fréquences, indicateurs qui se calculent dans le cadre des séries univariées, se calculent également sur des séries bivariées. Il en va de même pour les autres caractéristiques des séries statistiques que sont la moyenne, la variance et l'écart-type.

Ces caractéristiques peuvent être calculées sur des variables quantitatives, à partir :

- des distributions marginales : il s'agit de caractéristiques marginales ;
- des distributions conditionnelles : il s'agit de caractéristiques conditionnelles.

## 2.1 LES CARACTÉRISTIQUES MARGINALES

Les séries marginales sont des séries univariées. Les calculs des moyennes, variances et écarts-types marginaux se font donc de la façon habituelle, après extraction de la série marginale.

### Définitions

**Moyennes marginales :**  $\bar{x} = \frac{1}{n_{++}} \sum_{i=1}^p n_{i+} \times x_i = \frac{1}{n_{++}} \sum_{i=1}^p x_i \sum_{j=1}^q n_{ij}$  et de même  
 $\bar{y} = \frac{1}{n_{++}} \sum_{j=1}^q n_{+j} \times y_j = \frac{1}{n_{++}} \sum_{j=1}^q y_j \sum_{i=1}^p n_{ij}$ .

Remarque : certains auteurs notent ces moyennes marginales respectivement :  $\bar{\bar{x}}$  et  $\bar{\bar{y}}$ .

**Variances marginales :**  $V(x) = \frac{1}{n_{++}} \sum_{i=1}^p n_{i+} (x_i - \bar{x})^2$ , de formule développée  
 $V(x) = \frac{1}{n_{++}} \sum_{i=1}^p n_{i+} x_i^2 - \bar{x}^2$ . De même,  $V(y) = \frac{1}{n_{++}} \sum_{j=1}^q n_{+j} (y_j - \bar{y})^2 = \frac{1}{n_{++}} \sum_{j=1}^q n_{+j} y_j^2 - \bar{y}^2$ .

**Écarts-types marginaux :** Les écarts-types marginaux sont déduits des variances marginales,  $\sigma(x) = \sqrt{V(x)}$  et  $\sigma(y) = \sqrt{V(y)}$ .

### Exemple 5.5

#### Calcul des caractéristiques marginales

Soit un échantillon d'entreprises sur lequel sont observées les variables X, investissement annuel en milliers d'euros, et Y, chiffre d'affaires annuel en millions d'euros :

X \ Y	[10 ; 30[	[30 ; 50[	[50 ; 70[	Somme
[10 ; 30[	300	80	0	380
[30 ; 40[	70	200	50	320
[40 ; 50[	20	30	250	300
<b>Somme</b>	<b>390</b>	<b>310</b>	<b>300</b>	<b>1 000</b>

On extrait les séries marginales en utilisant les centres de classes. Les moyennes, variances et écarts-types marginaux sont ensuite calculés sur ces séries, comme dans le cas d'une série univariée ; la figure 5.1 donne la distribution marginale de X.

**Figure 5.1**

**Calcul des caractéristiques marginales de X.**

	A	B	C	D
1	$x_i$	$n_{i+}$	$n_{i+}x_i$	$n_{i+}x_i^2$
2	20	380	7 600	152 000
3	35	320	11 200	392 000
4	45	300	13 500	607 500
5	<b>Somme</b>	1 000	32 300	1 151 500

Ce qui donne :  $\bar{x} = \frac{32300}{1000} = 32,3$  ;  $V(x) = \frac{1151500}{1000} - (32,3)^2 = 108,21$  et  $\sigma(x) = 10,4$ .

En faisant de même pour la distribution marginale de Y, on obtient  $\bar{y} = \frac{38200}{1000} = 38,2$  ;

$V(y) = \frac{1732000}{1000} - (38,2)^2 = 272,76$  et  $\sigma(y) = 16,52$ .

## 2.2 LES CARACTÉRISTIQUES CONDITIONNELLES

Comme les caractéristiques marginales, les calculs des moyennes, variances et écarts-types conditionnels se font donc de la façon habituelle, après extraction de la distribution conditionnelle concernée.

### Définitions

**Moyennes conditionnelles :** Les moyennes conditionnelles de X sont les moyennes des distributions conditionnelles de X sachant Y.  $\bar{x}_i = \sum_{j=1}^p (f_{ij|i+}) \times x_i = \frac{1}{n_{i+}} \sum_{j=1}^p n_{ij} x_i$  est la moyenne conditionnelle de X sachant  $Y = y_i$ .

De même, les moyennes conditionnelles de Y sont les moyennes des distributions conditionnelles de Y sachant X.  $\bar{y}_i = \sum_{j=1}^q (f_{ij|i+}) \times y_i = \frac{1}{n_{i+}} \sum_{j=1}^q n_{ij} y_i$  est la moyenne conditionnelle de Y sachant  $X = x_i$ .

**Variances conditionnelles :** Les variances conditionnelles de X sont les variances des distributions conditionnelles de X sachant Y. La variance conditionnelle de X sachant  $Y = y_i$  est notée  $V_i(x) = \frac{1}{n_{i+}} \sum_{j=1}^p n_{ij} (x_i - \bar{x}_i)^2 = \frac{1}{n_{i+}} \sum_{j=1}^p n_{ij} x_i^2 - \bar{x}_i^2$ .

De même, les variances conditionnelles de Y sont les variances des distributions conditionnelles de Y sachant X. La variance conditionnelle de Y sachant  $X = x_i$  est notée

$V_i(y) = \frac{1}{n_{i+}} \sum_{j=1}^q n_{ij} (y_i - \bar{y}_i)^2 = \frac{1}{n_{i+}} \sum_{j=1}^q n_{ij} y_i^2 - \bar{y}_i^2$ .

**Écarts-types conditionnels :** Les écarts-types conditionnels sont déduits des variances conditionnelles,  $\sigma(x) = \sqrt{V(x)}$  et  $\sigma(y) = \sqrt{V(y)}$ .

**Exemple 5.6****Calcul des caractéristiques conditionnelles**

Reprendons les données de l'exemple 5.5.

Extrayons la distribution conditionnelle de X sachant  $Y = 60$ . À partir de cette série extraite, assimilable à une série univariée, nous effectuons les étapes nécessaires aux calculs de la moyenne et de la variance (voir figure 5.2).

**Figure 5.2**

**Distribution conditionnelle de X sachant  $Y = 60$ .**

	A	B	C	D
1	$x_i$	$n_{i3}$	$n_{i3}x_i$	$n_{i3}x_i^2$
2	20	0	0	0
3	35	50	1 750	61 250
4	45	250	11 250	506 250
5	<b>Somme</b>	300	13 000	567 500

$$\text{D'où les paramètres conditionnels : } \bar{x}_3 = \frac{13 000}{300} = 43,33 ;$$

$$V_3(X) = \frac{567 500}{300} - 43,33^2 = 14,18 \text{ et } \sigma_3(X) = \sqrt{14,18} = 3,77.$$

## 2.3 RELATIONS ENTRE LES MOYENNES MARGINALES ET CONDITIONNELLES

Les moyennes conditionnelles et marginales sont liées par la relation suivante : la moyenne des moyennes conditionnelles de X est égale à la moyenne marginale de X.

Cette propriété est à relier à la notion d'espérance conditionnelle en probabilité.

Soit  $\bar{\bar{x}}$  la moyenne des moyennes conditionnelles. La démonstration suivante montre que  $\bar{\bar{x}}$  est égale à la moyenne marginale de X, c'est-à-dire  $\bar{x}$  :

$$\begin{aligned} \bar{\bar{x}} &= \frac{1}{n_{++}} \sum_{j=1}^q n_{+j} \bar{x}_j = \frac{1}{n_{++}} \sum_{j=1}^q n_{+j} \left( \frac{1}{n_{+j}} \sum_{i=1}^p n_{ij} x_i \right) = \frac{1}{n_{++}} \sum_{j=1}^q \sum_{i=1}^p n_{ij} x_i \\ &= \frac{1}{n_{++}} \sum_{i=1}^p x_i \left( \sum_{j=1}^q n_{ij} \right) = \frac{1}{n_{++}} \sum_{i=1}^p x_i n_{i+} = \bar{x} \end{aligned}$$

**Exemple 5.7****Vérification de la relation entre moyennes marginales et conditionnelles**

Reprendons les données de l'exemple 5.5.

Extrayons les distributions conditionnelles de X sachant  $Y = y_1$  (voir figure 5.3).

**Figure 5.3**

**Distribution conditionnelle de X sachant  $Y = y_1$ .**

	A	B	C	D
1	$x_i$	$n_{i1}$	$n_{i1}x_i$	$n_{i1}x_i^2$
2	20	300	6000	120000
3	35	70	2 450	85 750
4	45	20	900	40 500
5	<b>Somme</b>	390	9 350	246 250

$$\text{D'où } \bar{x}_1 = \frac{1}{n_{+j}} \sum_{i=1}^3 n_{ij} x_i = \frac{9\,350}{390} = 23,97.$$

En faisant de même pour les distributions conditionnelles de X sachant  $Y = y_2$  et de X sachant  $Y = y_3$ , on obtient :

$$\bar{x}_2 = \frac{1}{n_{+j}} \sum_{i=1}^3 n_{ij} x_i = \frac{9\,950}{310} = 32,10 ; \quad \bar{x}_3 = \frac{1}{n_{+j}} \sum_{i=1}^3 n_{ij} x_i = \frac{13\,000}{300} = 43,33.$$

La distribution des moyennes conditionnelles de X est proposée figure 5.4.

**Figure 5.4**

**Distribution des moyennes conditionnelles de X.**

	A	B	C	D
1	j	$\bar{x}_j$	$n_{+j}$	$n_{+j}\bar{x}_j$
2	1	23,97	390	9 350
3	2	32,10	310	9 950
4	3	43,33	300	13 000
5	Somme		1 000	32 300

$$\text{D'où la moyenne des moyennes conditionnelles de X : } \bar{\bar{x}} = \frac{1}{n_{++}} \sum_{j=1}^3 n_{+j} \bar{x}_j = \frac{32\,300}{1000} = 32,3.$$

Or,  $\bar{x} = 32,3$  (voir exemple 5.5). Donc, la relation entre  $\bar{\bar{x}}$  et  $\bar{x}$  est vérifiée.

## 2.4 LA COVARIANCE

Nous avons vu que la variabilité des caractères quantitatifs à une variable autour de leur moyenne pouvait être mesurée par la variance. Dans le cas des séries doubles, nous disposons d'un indicateur comparable, appelé covariance, qui permet de mesurer les fluctuations simultanées de chaque variable par rapport à sa moyenne. Il est important de noter que, contrairement à la variance (moyenne de carrés) qui est toujours positive ou nulle, la covariance peut être de signe quelconque.

### Définition

**La covariance :** Soit X et Y deux caractères quantitatifs. La covariance du couple (X ; Y) est

$$\text{définie par : } \text{Cov}(X;Y) = \frac{1}{n_{++}} \sum_{j=1}^q \sum_{i=1}^p n_{ij} (x_i - \bar{x})(y_j - \bar{y}) = \sum_{j=1}^q \sum_{i=1}^p f_{ij} (x_i - \bar{x})(y_j - \bar{y}).$$

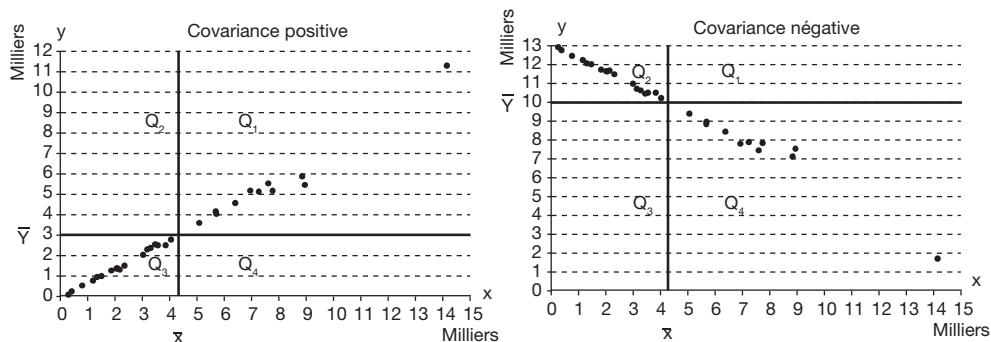
Graphiquement, cette définition revient à prendre un nouveau repère d'origine G ( $\bar{x}; \bar{y}$ ), le point moyen, et à diviser le plan en quatre quadrants, respectivement définis

par :  $Q_1 \begin{cases} x \geq \bar{x} \\ y \geq \bar{y} \end{cases}, Q_2 \begin{cases} x \leq \bar{x} \\ y \geq \bar{y} \end{cases}, Q_3 \begin{cases} x \leq \bar{x} \\ y \leq \bar{y} \end{cases}$  et  $Q_4 \begin{cases} x \geq \bar{x} \\ y \leq \bar{y} \end{cases}$ . On notera que les quadrants  $Q_1$  et  $Q_3$  sont associés, car les points M ( $x_i ; y_j$ ) du nuage situé dans le domaine  $Q_1 \cup Q_3$  sont caractérisés par  $(x_i - \bar{x})(y_j - \bar{y}) \geq 0$ , les quantités  $(x_i - \bar{x})$  et  $(y_j - \bar{y})$  étant de même signe. De même,  $Q_2 \cup Q_4$  est caractérisé par  $(x_i - \bar{x})(y_j - \bar{y}) \leq 0$ . Ainsi, le signe de la covariance nous indiquera si les points du nuage sont majoritairement dans  $Q_1 \cup Q_3$  ou

dans  $Q_2 \cup Q_4$  (voir figure 5.5) ; nous reviendrons sur cette remarque dans l'étude de la régression (voir chapitre 6).

**Figure 5.5**

**Nuages de points**  
 $(x_i ; y_i)$ .



Comme pour la variance, la covariance admet une formule développée. Cette formule est issue du théorème de Koenig.

### Définition

$$\text{Formule développée de la covariance : } \text{Cov}(X; Y) = \frac{1}{n} \sum_{j=1}^q \sum_{i=1}^p n_{ij} x_i y_j - \bar{x} \times \bar{y}.$$

Par cette formule, la covariance est la « moyenne des produits moins le produit des moyennes ».

De même que la variance, la covariance possède des propriétés très importantes :

### Propriétés

- $\text{Cov}(X ; Y) = \text{Cov}(Y ; X)$  : la covariance est symétrique.
- $\text{Cov}(X ; X) = \text{Var}(X)$  : la covariance est obtenue en « dédoublant » la formule de la variance.
- $\text{Cov}(aX ; a'Y) = aa' \text{Cov}(X ; Y)$  : multiplier chacune des séries par un réel multiplie la covariance par le produit de ces nombres.
- $\text{Cov}(X+b ; Y) = \text{Cov}(X ; Y)$  : ajouter une constante ne change pas la covariance.

Le signe de la covariance possède une signification (voir figure 5.5) :

- Une covariance positive indique que les caractères X et Y varient globalement dans le même sens, une hausse de l'un étant associée à une hausse de l'autre, ou encore une baisse de l'un étant associée à une baisse de l'autre.
- Une covariance négative indique que les caractères X et Y varient globalement en sens contraires, une hausse de l'un étant associée à une baisse de l'autre.

### Exemple 5.8

#### Calcul de covariance dans le cas de données exhaustives

Reprendons les données de l'exemple 5.1 et calculons la covariance avec la formule développée. On rappelle que  $n = 3$  ;  $\bar{x} = 9721$  ;  $\bar{y} = 7176$ . On calcule chacun des  $x_i y_i$  et on en fait la somme (voir figure 5.6).

**Figure 5.6**

**Calcul des  $x_i y_j$ .**

	A	B	C	D
1	Académie	x	y	$x_i y_j$
2	Paris	14 150	11 271	159 484 650
3	Créteil	7 759	5 150	39 958 850
4	Versailles	7 254	5 107	37 046 178
5	<b>Somme</b>	29 163	21 528	236 489 678

Source : ministère de l'Éducation nationale, 2006

D'où  $\sum_{j=1}^q \sum_{i=1}^p n_{ij} x_i y_j = 236 489 678$ . D'où, en utilisant la formule développée :

$$\text{Cov}(X; Y) = \frac{1}{3} \sum_{j=1}^3 \sum_{i=1}^3 n_{ij} x_i y_j - \bar{x} \times \bar{y} = \frac{1}{3} 236 489 678 - 9 721 \times 7 176,$$

soit **Cov(X ; Y) = 9 071 996,76**. Le nombre de licenciés en 2005 varie dans le même sens que le nombre de licenciés poursuivant leurs études à l'université en 2006.

Dans le chapitre suivant nous affinerons l'étude de la relation entre deux caractères et nous verrons le rôle de la covariance dans le calcul du coefficient de corrélation linéaire.

## 3 Étude des liaisons entre deux variables

La notion de liaison entre deux variables est un premier stade incontournable vers une éventuelle imputation causale qu'il est fondamental de mettre en évidence dans de nombreux domaines, notamment en épidémiologie, justice, économie, sociologie, etc. Dans le cas particulier de deux caractères quantitatifs, le degré d'association peut varier entre deux extrêmes : d'un côté la liaison fonctionnelle et de l'autre l'indépendance.

### 3.1 LIAISON FONCTIONNELLE ET INDÉPENDANCE

Prenons l'exemple du jeu de la roulette. La roulette comporte 37 numéros (numérotés de 0 à 36), 18 rouges, 18 noirs, le zéro étant vert. À la suite de 100 parties, notons respectivement X et Y le nombre de numéros rouges et de numéros noirs sortis.

Si le zéro n'existe pas, nous aurions entre X et Y la relation fonctionnelle  $X + Y = 100$  ; avec la présence d'une case verte, nous avons un degré de liaison très fort entre X et Y, le zéro ayant une probabilité faible de sortie.

#### Définition

Un caractère X est **lié fonctionnellement** au caractère Y si à chaque modalité de Y correspond une seule modalité de X.

La liaison fonctionnelle n'est pas symétrique : si X est fonctionnellement lié à Y, cela n'implique pas que Y le soit fonctionnellement à X.

#### Exemple 5.9

#### Liaison fonctionnelle et absence de symétrie

Supposons que, suite à l'introduction sur le marché d'un nouveau produit, une enquête de satisfaction sur un échantillon de 200 consommateurs des deux sexes ait donné les résultats suivants, avec X le sexe et Y la satisfaction :

<b>X \ Y</b>	<b>Insatisfait</b>	<b>Ni satisfait, ni insatisfait</b>	<b>Satisfait</b>	<b>Somme</b>
Masculin (M)	20	70	0	90
Féminin (F)	0	0	110	110
<b>Somme</b>	20	70	110	200

X est fonctionnellement lié à Y, car pour chaque modalité de Y résulte une seule modalité de X ; ainsi, un consommateur satisfait est nécessairement un homme. Par contre, Y n'est pas fonctionnellement lié à X, car à la modalité Masculin de X correspondent deux modalités possibles de Y : Insatisfait ou Ni satisfait, ni insatisfait.

La liaison fonctionnelle n'est pas symétrique.

Reprendons l'exemple du jeu de la roulette et imaginons une roulette comportant un très grand nombre de cases vertes : les caractères X et Y seraient à peu près indépendants.

### Définition

Deux variables statistiques X et Y sont **indépendantes** si les distributions conditionnelles de X sachant Y sont identiques, ce qui équivaut à :  $f_{(X=x_i/Y=y_j)} = f_{i+}$ , quels que soient les indices i et j (i entier compris entre 1 et p et j entre 1 et q). Dans le cas où X et Y sont indépendants, les distributions conditionnelles de X selon Y sont identiques à la distribution marginale de X.

Le concept d'indépendance étant symétrique, l'indépendance se traduit également par la relation  $f_{(Y=y_j/X=x_i)} = f_{+j}$ .

Cette notion est similaire à la notion d'indépendance probabiliste :

$$P_B(A) = P(A \cap B) / P(B).$$

### Exemple 5.10

#### Étude de l'indépendance

Reprendons les données de l'exemple 5.2. Leur étude a mené au calcul des fréquences conditionnelles de X sachant Y, rappelées dans le tableau suivant :

<b>X \ Y</b>	<b>y<sub>1</sub> = O</b>	<b>y<sub>2</sub> = N</b>
x <sub>1</sub> = M	$f_{i=1/j=1} = 0,6729$	$f_{i=1/j=2} = 0,7206$
x <sub>2</sub> = F	$f_{i=2/j=1} = 0,3271$	$f_{i=2/j=2} = 0,2794$
<b>f<sub>ij</sub></b>	1	1

De même, les fréquences marginales de X avaient été calculées :  $f_{1+} = 0,70$  et  $f_{2+} = 0,30$ , soit 70 % d'hommes et 30 % de femmes.

Parmi les individus ayant répondu oui, il y a 67,29 % d'hommes et 32,71 % de femmes, ce qui est différent des proportions d'hommes et de femmes dans l'échantillon étudié, qui sont respectivement de 70 % et de 30 %. Ces résultats montrent que les caractères X et Y ne sont pas indépendants, car les distributions conditionnelles de X selon Y ne sont pas égales à la distribution marginale de X (voir définition de l'indépendance, ci-avant) : la réponse d'un individu n'est pas indépendante de son sexe.

## 3.2 INTRODUCTION AU TEST DU KHI-DEUX ( $\chi^2$ )

Le test d'indépendance du khi-deux ( $\chi^2$ ) permet de se prononcer sur l'indépendance de deux variables qualitatives, observées sur un échantillon. Il s'effectue en deux étapes :

1. La première consiste à comparer le tableau des effectifs observés et le tableau des effectifs théoriques calculés sous l'hypothèse d'indépendance, ou plutôt de mesurer leur distance afin de disposer d'un indicateur permettant d'accepter ou de refuser l'hypothèse d'indépendance entre ces variables :
  - Si la distance entre les tableaux est « petite », les effectifs observés sont proches des effectifs théoriques. Les effectifs observés s'assimilent aux effectifs théoriques sous hypothèse d'indépendance : on ne peut rejeter l'hypothèse d'indépendance.
  - Si la distance entre les tableaux est « grande », les effectifs observés sont différents des effectifs théoriques calculés sous l'hypothèse d'indépendance. Les effectifs observés ne s'assimilent pas aux effectifs théoriques sous l'hypothèse d'indépendance : les deux variables ne sont pas indépendantes.
2. La deuxième étape, présente dans tous les tests d'hypothèses (voir focus 5.1), consiste à déterminer la probabilité associée à la décision d'accepter ou de refuser l'hypothèse d'indépendance. Ne pouvant prétendre à une certitude, il apparaît raisonnable de « minimiser » le risque d'erreur.

### Focus 5.1

#### Principe des tests d'hypothèses

Une hypothèse statistique est une assertion concernant les caractéristiques (valeurs des paramètres, nature de la distribution, indépendance, etc.) d'une ou de plusieurs variables statistiques sur **une population**.

L'examen de la validité d'une hypothèse se fait sur la base d'observations recueillies sur **un échantillon** de la population étudiée. Le test statistique est une démarche qui vise à fournir une règle de décision permettant de faire un choix entre deux hypothèses statistiques.

Les deux hypothèses envisagées s'appellent l'hypothèse nulle ( $H_0$ ) et l'hypothèse alternative ( $H_1$ ). La terminologie hypothèse nulle est une hypothèse de « différence nulle » entre les données observées sur un échantillon et l'hypothèse  $H_0$  que l'on désire tester (valeur d'un paramètre, adéquation à une loi de probabilité théorique, indépendance, etc.).

La démarche du test s'effectue en considérant  $H_0$  vraie ; c'est cette hypothèse que nous allons soit accepter – on parle alors de région de « non-rejet de  $H_0$  », soit rejeter – on parle alors de « région critique » de  $H_0$ . Le rejet éventuel de l'hypothèse nulle conduit à l'acceptation de l'hypothèse alternative (« contre-hypothèse »)  $H_1$ .

La décision de favoriser telle hypothèse est basée sur les résultats d'un échantillon et donc, à partir d'une information très partielle, il est impossible d'être sûr de prendre la bonne décision : on devra se contenter de limiter la probabilité que notre décision soit erronée.

On distinguera deux types d'erreur :

- Erreur de première espèce : rejeter à tort  $H_0$ . Ce risque, consenti à l'avance, de rejeter à tort l'hypothèse nulle alors qu'elle est vraie s'appelle le seuil de signification et est

noté  $\alpha$ . Les seuils les plus utilisés sont  $\alpha = 0,05$  et  $\alpha = 0,01$ , soit respectivement 5 % et 1 %.

- Erreur de seconde espèce : accepter  $H_0$  alors que  $H_1$  est vraie. La probabilité de cette erreur est notée  $\beta$ .

Le risque de première espèce est regrettable, mais inévitable, comme le rappelle Daniel Schwartz. La seule façon de ne pas se tromper, et de ne prendre aucun risque de rejeter à tort  $H_0$ , est d'accepter  $H_0$  dans tous les cas, ce qui augmente le risque d'accepter  $H_0$  alors qu'elle est fausse. Autrement dit, pour diminuer  $\alpha$ , il faut augmenter  $\beta$ . Pour ne pas prendre le moindre risque de condamner un innocent – risque  $\alpha$ –, on doit accepter le risque de relaxer tous les coupables – risque  $\beta$ .

## Effectifs observés et effectifs théoriques calculés

La première étape passe par le calcul des effectifs théoriques notés  $c_{ij}$ .

### Définition

**Les effectifs calculés (ou théoriques) :** Les effectifs calculés sous l'hypothèse d'indépendance, encore appelés effectifs théoriques, sont notés  $c_{ij}$  et donnés par :  $c_{ij} = n_{+i} \times n_{+j} / n_{++}$ .

Après détermination des effectifs calculés  $c_{ij}$ , il est possible de déterminer un indicateur de distance entre le tableau observé, composé des  $n_{ij}$ , et le tableau théorique, composé des  $c_{ij}$ . Cette « distance » est appelée distance du khi-deux.

### Définition

**Distance du khi-deux :** La distance entre les tableaux observé et théorique est appelée khi-deux calculé, notée  $\chi_c^2$ , et définie par  $\chi_c^2 = \sum_{i=1}^q \sum_{j=1}^p \frac{(n_{ij} - c_{ij})^2}{c_{ij}}$ , les coefficients  $c_{ij}$  désignant les effectifs théoriques ou calculés et les  $n_{ij}$  les effectifs observés.

Pour appliquer un calcul de distance du khi-deux entre deux tableaux, les deux conditions suivantes doivent être vérifiées :

- la taille de l'échantillon doit être supérieure ou égale à 30 ;
- tous les effectifs calculés doivent être supérieurs ou égaux à 5 (dans le cas contraire, on regroupe les classes adjacentes).

Karl Pearson a démontré que ce khi-deux calculé suit approximativement la distribution du khi-deux (voir focus 5.2), loi de probabilité continue, caractérisée par un paramètre  $v$  (nu), le degré de liberté.

### Définition

**Degré de liberté d'un tableau de contingence :** Soit un tableau de contingence formé de  $n$  lignes et de  $p$  colonnes. Son degré de liberté, noté  $ddl$ , est donné par :  $ddl = (n-1)(p-1)$ , ou encore  $ddl = (\text{nombre de lignes} - 1) \times (\text{nombre de colonnes} - 1)$ .

Pour comprendre la signification de la notion de degré de liberté, il convient d'observer que l'on peut remplir librement les  $(n-1)$  premières lignes et les  $(p-1)$  premières colonnes et qu'alors les effectifs marginaux imposent les valeurs restantes.

**Focus 5.2****La loi du khi-deux**

La loi du  $\chi^2$  finalisée par Karl Pearson au début du XX<sup>e</sup> siècle est une loi de probabilité continue représentant la distribution de la somme des carrés de n variables aléatoires indépendantes, chacune étant normale centrée réduite. Cette somme est appelée variable du  $\chi^2$  à n degrés de liberté ; on note v le degré de liberté (ddl). Les valeurs de  $\chi^2$  dépendent du degré de liberté v et du seuil de signification  $\alpha$ . Elles sont notées  $\chi^2_{(\alpha;v)}$  et sont tabulées sur la table du  $\chi^2$ , avec  $P(\chi^2 \geq \chi^2_{(\alpha;v)}) = \alpha$ .

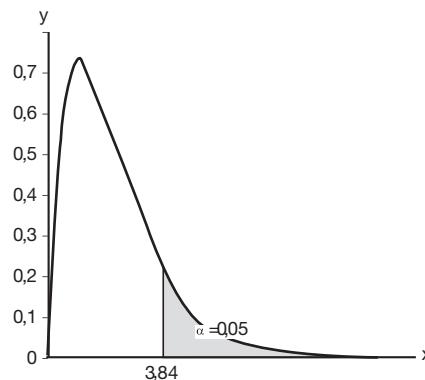
Prenons un exemple : pour un seuil de signification de 5 % et un ddl v = 1, on trouve :  $\chi^2_{(0,05;1)} = 3,84$  ; pour un seuil de signification de 1 % et un ddl de 1,  $\chi^2_{(0,01;1)} = 6,63$ .

Pour un ddl de 1, il y a une chance sur 100 pour que la variable aléatoire du  $\chi^2$  à 1 degré de liberté dépasse 6,63 (voir figure 5.7).

Autre démarche : on peut, à partir du khi-deux calculé et du ddl, déterminer le degré de signification correspondant. Par exemple, pour un khi-deux calculé de 2,8 et un ddl de 1, le degré de signification est de 9,43 % (ce degré de signification peut être déterminé en utilisant Excel ; voir exercice 5) ; si le seuil de 5 % a été assigné au test, alors on ne pourra pas rejeter l'hypothèse nulle, car notre seuil de signification est supérieur à 5 % (voir l'exercice 5 et la notion de p-valeur).

**Figure 5.7**

**Distribution du khi-deux à 1 degré de liberté.**



### Le test de l'hypothèse d'indépendance

La deuxième étape consiste à tester l'hypothèse d'indépendance, en respectant les quatre phases suivantes :

1. Formuler les hypothèses :
  - $H_0$  : les deux caractères sont indépendants.
  - $H_1$  : les deux caractères ne sont pas indépendants.
2. Choisir le seuil de signification, noté  $\alpha$ .

3. Déterminer le degré de liberté.
4. Définir la règle de décision à partir de  $\chi_c^2$  le khi-deux calculé et  $\chi_{(\alpha;v)}^2$  le khi-deux critique, dépendant du seuil de signification  $\alpha$  et du degré de liberté  $v$ .
  - Si  $\chi_c^2 \geq \chi_{(\alpha;v)}^2$ , l'hypothèse  $H_0$  d'indépendance entre les deux variables est rejetée et l'hypothèse  $H_1$  est acceptée : les deux caractères seront considérés comme statistiquement associés.
  - Si  $\chi_c^2 \leq \chi_{(\alpha;v)}^2$ , l'hypothèse  $H_0$  d'indépendance entre les deux variables n'est pas rejetée : il est impossible de conclure de façon significative à l'existence d'un lien statistique entre les variables.

### Exemple 5.11

#### Test du khi-deux

Reprendons l'exemple 5.10. Le tableau des effectifs observés est le suivant :

<b>X \ Y</b>	<b>y<sub>1</sub> = O</b>	<b>y<sub>2</sub> = N</b>
x <sub>1</sub> = M	n <sub>11</sub> = 290	n <sub>12</sub> = 410
x <sub>2</sub> = F	n <sub>21</sub> = 141	n <sub>22</sub> = 159

Les deux variables sont dépendantes (voir exemple 5.10). Il est possible de s'interroger sur les conditions qui auraient permis de conclure à l'indépendance. Pour cela, calculons les effectifs sous l'hypothèse d'indépendance, notés  $c_{ij}$ .

L'indépendance se traduit par :  $f_{(X=x_i/Y=y_j)} = f_{1+}$ , soit par le fait que la proportion d'individus de sexe masculin parmi « les oui » est égale à la proportion d'individus de sexe masculin dans la population étudiée, soit 70 %, ce qui donne :  $\frac{c_{11}}{431} = \frac{700}{1000}$ , soit

$$c_{11} = \frac{700}{1000} \times 431 = 302. \text{ Remarquons que } c_{11} = \frac{n_{1+}}{n_{++}} \times n_{1+}.$$

Ce problème compte apparemment quatre inconnues, mais en vérité elles sont liées : la donnée d'une de ces inconnues, par exemple  $c_{11}$ , fixe les valeurs des autres. Le tableau a un degré de liberté égal à 1.

Ainsi, à partir de  $c_{11}$ , il est possible de trouver toutes les autres valeurs du tableau :  $c_{12} = 700 - c_{11} = 398$ ;  $c_{21} = 431 - c_{11} = 129$  et  $c_{22} = 300 - c_{21} = 171$ .

D'où le tableau suivant, qui indique les effectifs calculés,  $c_{ij}$ , en supposant l'indépendance des caractères X et Y.

<b>X \ Y</b>	<b>y<sub>1</sub> = O</b>	<b>y<sub>2</sub> = N</b>	<b>Somme</b>
x <sub>1</sub> = M	302	398	700
x <sub>2</sub> = F	129	171	300
<b>Somme</b>	431	569	1 000

Calculons le  $\chi^2$  à l'aide de la formule  $\chi_C^2 = \sum_{j=1}^q \sum_{i=1}^p \frac{(n_{ij} - c_{ij})^2}{c_{ij}}$ . Pour cela, il est nécessaire

de calculer chacun des  $\frac{(n_{ij} - c_{ij})^2}{c_{ij}}$ , avant d'en faire la somme. Ainsi,

$$\frac{(n_{11} - c_{11})^2}{c_{11}} = \frac{(290 - 302)^2}{302} = 0,48 ; \quad \frac{(n_{12} - c_{12})^2}{c_{12}} = \frac{(410 - 398)^2}{398} = 0,36 ;$$

$$\frac{(n_{21} - c_{21})^2}{c_{21}} = \frac{(141 - 129)^2}{129} = 1,12 \text{ et } \frac{(n_{22} - c_{22})^2}{c_{22}} = \frac{(159 - 171)^2}{171} = 0,84 . \text{ Ces valeurs sont}$$

reportées dans le tableau suivant :

<b>X \ Y</b>	<b>y<sub>1</sub> = O</b>	<b>y<sub>2</sub> = N</b>	<b>Somme</b>
x <sub>1</sub> = M	0,48	0,36	0,84
x <sub>2</sub> = F	1,12	0,84	1,96
<b>Somme</b>	<b>1,60</b>	<b>1,20</b>	<b>2,80</b>

Ainsi,  $\chi_c^2 = 0,48 + 0,36 + 1,12 + 0,84 = 2,80$ , avec un ddl de 1 qui donne au seuil de 5 %  $\chi^2_{(0,05;1)} = 3,84$ .

$\chi_c^2 \leq \chi^2_{(\alpha;\nu)}$ , l'hypothèse  $H_0$  d'indépendance entre les deux variables n'est pas rejetée : il est impossible de conclure de façon significative à l'existence d'un lien statistique entre le sexe et le type de réponse.

### Focus 5.3

### Test du khi-deux sous Excel

Excel propose de réaliser un test du khi-deux uniquement à partir des tableaux de données observées et théoriques, sans avoir à calculer les distances du khi-deux. Pour cela, sélectionnez la cellule dans laquelle vous souhaitez faire apparaître le résultat, puis, dans la barre de menus, cliquez sur Insertion/Fonction. Dans la boîte de dialogue, sélectionnez la catégorie Statistiques, puis sélectionnez la fonction TEST.KHIDEUX. Cliquez sur **OK**. Dans la boîte de dialogue Arguments de la fonction (voir figure 5.8), dans le champ Plage\_réelle, indiquez la plage dans laquelle se trouve le tableau de données observées, soit B2:D4, et dans le champ Plage\_attendue, indiquez la plage dans laquelle se trouve le tableau de données théoriques, soit B22:D24 pour notre exemple. Cliquez sur **OK** pour faire apparaître le résultat.

**Figure 5.8**

**Réalisation du test du khi-deux sous Excel.**



La probabilité affichée, égale à 0,0000, est le degré de signification, c'est-à-dire le plus petit risque d'erreur pour lequel la différence entre le modèle observé et le modèle d'indépendance est significative. Si cette probabilité est supérieure au seuil de signification, alors  $H_0$  ne peut être rejetée. Dans notre exemple, avec un seuil de signification de 5 % et un degré de signification d'environ 0, on doit rejeter  $H_0$ .

## Conclusion

Ce chapitre est un chapitre clef à double titre : tout d'abord il a introduit les outils de base des séries bivariées, qui seront nécessaires pour aborder, au chapitre 6, la régression ; ensuite il a introduit le concept fondamental d'indépendance. Cette notion a été l'occasion de présenter une initiation aux tests statistiques, qui constituent un aspect fondamental de l'inférence statistique.

Le lecteur doit maîtriser les concepts d'effectifs (et de fréquences) conditionnels et marginaux, ainsi que les éléments ayant trait aux tableaux de contingence : utilisation rigoureuse des indices, notion de degré de liberté. La covariance, son calcul sous les deux formes et l'interprétation de son signe doivent être bien connus.

Enfin, le lecteur doit s'attacher à une rédaction rigoureuse et systématique dans l'élaboration d'un test d'hypothèse. Les calculs intervenant dans le test du khi-deux exigent une démarche, des notations et une présentation claires. Par ailleurs, indépendamment de l'utilisation du tableur, il est fondamental d'être familiarisé avec la table de la distribution du khi-deux.

# Problèmes et exercices

Par l'intermédiaire du tableau de contingence, ce chapitre présente une première approche des séries bivariées.

- Les exercices 1 et 2 initient à la construction du tableau de contingence selon la nature des variables étudiées.
- L'exercice 3 détaille les éléments constitutifs du contenu d'un tableau de contingence.
- L'exercice 4 applique aux séries bivariées le calcul des indicateurs précédemment mis en œuvre dans l'étude des séries univariées.
- L'exercice 5 introduit la notion de dépendance entre deux séries, à l'aide de la covariance et du test du khi-deux.



## EXERCICE 1 CONSTRUCTION D'UN TABLEAU DE CONTINGENCE SUR CARACTÈRES DISCRET ET QUALITATIF

### Énoncé

À l'occasion d'une enquête statistique, un enseignant demande à ses 28 étudiants d'indiquer sur un papier leur genre, masculin ou féminin, et le nombre de films qu'ils ont vus au cinéma au cours des deux derniers mois. Les résultats de l'enquête sont reportés dans le tableau suivant :

Étudiant	Nombre de films	Genre
Étudiant 1	1	Féminin
Étudiant 2	3	Masculin
Étudiant 3	4	Masculin
Étudiant 4	2	Féminin
Étudiant 5	1	Féminin
Étudiant 6	4	Féminin
Étudiant 7	0	Masculin
Étudiant 8	2	Masculin
Étudiant 9	2	Féminin
Étudiant 10	0	Féminin
Étudiant 11	4	Féminin
Étudiant 12	5	Masculin
Étudiant 13	2	Féminin
Étudiant 14	1	Masculin

Étudiant	Nombre de films	Genre
Étudiant 15	2	Féminin
Étudiant 16	2	Masculin
Étudiant 17	2	Féminin
Étudiant 18	3	Féminin
Étudiant 19	3	Masculin
Étudiant 20	3	Masculin
Étudiant 21	6	Masculin
Étudiant 22	3	Masculin
Étudiant 23	0	Féminin
Étudiant 24	0	Féminin
Étudiant 25	0	Masculin
Étudiant 26	2	Masculin
Étudiant 27	3	Masculin
Étudiant 28	2	Féminin

1. Précisez la nature des caractères étudiés.
2. Dressez le tableau de contingence présentant les deux distributions marginales.

### Solution

1. La variable « nombre de films » est une variable quantitative discrète.

La variable « genre » est une variable qualitative nominale.

2. Afin d'établir le tableau de contingence décrivant la série bivariée, nous pouvons soit faire un recensement manuel, soit utiliser le tableau croisé dynamique d'Excel.

Pour le recensement manuel, il convient de compter combien de femmes ont vu 0 film, 1 film, 2 films, etc., et de faire de même pour les hommes. Ce comptage aboutit au tableau de contingence de la figure 5.9, qui indique par exemple que 6 femmes ont vu 2 films ou encore que 5 hommes ont vu 3 films.

**Figure 5.9**

Réalisation manuelle  
d'un tableau de  
contingence.

Films \ Genre	Féminin	Masculin	Somme
0	3	2	5
1	2	1	3
2	6	3	9
3	1	5	6
4	2	1	3
5	0	1	1
6	0	1	1
<b>Somme</b>	<b>14</b>	<b>14</b>	<b>28</b>

La réalisation manuelle d'un tel tableau est souvent longue et fastidieuse. Excel permet de réaliser ce type de tableau automatiquement, à l'aide du tableau croisé dynamique. Ce tableau est dit dynamique, car une fois qu'il est réalisé à partir des données brutes, il est possible de le modifier à tout moment en faisant glisser les variables à l'aide de la souris.

Pour réaliser un tableau croisé dynamique sous Excel, ouvrez Excel sur la feuille contenant les données à traiter. Cliquez sur Données/Rapport de tableau croisé dynamique dans la barre de menus.

L'assistant tableau croisé dynamique apparaît (voir figure 5.10). Par défaut les données à analyser sont supposées être dans Excel. Il suffit donc de cliquer sur le bouton Suivant.

**Figure 5.10**

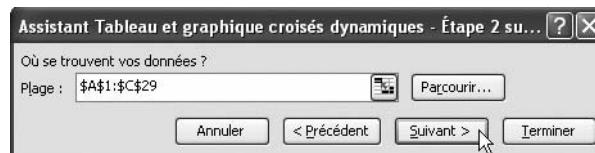
**Création d'un tableau croisé dynamique à l'aide de l'assistant.**



L'assistant tableau croisé dynamique demande alors d'indiquer la plage où se trouvent les données. Il convient donc de sélectionner à l'aide de la souris le tableau Excel, c'est-à-dire, ici, la plage A1:C29, comme indiqué sur la figure 5.11. Puis cliquez sur le bouton Suivant.

**Figure 5.11**

**Sélection des données à croiser dans l'assistant tableau croisé dynamique.**



Dernière étape : il convient d'indiquer l'endroit où vous souhaitez que le tableau croisé dynamique soit réalisé : soit sur une nouvelle feuille, soit sur la feuille existante. Nous choisissons ici de faire apparaître le tableau croisé dynamique sur une nouvelle feuille (voir figure 5.12) avant de cliquer sur le bouton Terminer.

**Figure 5.12**

**Sélection du lieu d'affichage des résultats dans l'assistant tableau croisé dynamique.**



L'assistant tableau croisé dynamique se ferme et le tableau croisé dynamique apparaît, vide, avec la liste de champs qui reprend les trois colonnes du tableau brut (voir figure 5.13).

**Figure 5.13**

**Tableau croisé dynamique à renseigner.**

Pour remplir le tableau croisé dynamique, il suffit de cliquer sur un des éléments de la liste de champs et de le faire glisser, à l'aide de la souris, à l'endroit souhaité du tableau de contingence. Dans notre cas :

- L'élément « Nombre de films » est déplacé à l'emplacement indiqué « Déposer champs de lignes Ici ».
- L'élément « Genre » est déplacé à l'emplacement indiqué « Déposer champs de colonnes Ici ».
- L'élément « Étudiant » est déplacé à l'emplacement indiqué « Déposer données Ici ».

Le tableau croisé dynamique construit fait ainsi apparaître le nombre de films en ligne, le sexe en colonne et compte le nombre d'étudiants présentant chaque modalité de l'une et de l'autre de ces deux variables (voir figure 5.14).

**Figure 5.14**

**Dénombrement par tableau croisé dynamique.**

	A	B	C	D	E	F	G	H
1	Déposer champs de page Ici							
2								
3	Nombre de Etudiant	Genre						
4	Nombre de films	(vide)	Féminin	Masculin	(vide)	Total		
5		0	3	2		5		
6		1	2	1		3		
7		2	6	3		9		
8		3	1	5		6		
9		4	2	1		3		
10		5		1		1		
11		6		1		1		
12	(vide)							
13	Total		14	14		28		
14								
15								
16								
17								
18								

 The 'Liste de champs de tableau croisé' pane on the right shows the same three fields: 'Etudiant', 'Nombre de films', and 'Genre'. The 'Genre' field is currently selected."/>

Un simple clic sur les cellules dynamiques (A3, A4 et B3) permet de modifier les options du tableau, notamment de faire disparaître la modalité indiquée (vide) pour chaque variable.

Ce tableau de contingence correspond à celui obtenu manuellement (voir figure 5.9).

## EXERCICE 2 CONSTRUCTION D'UN TABLEAU DE CONTINGENCE SUR CARACTÈRES CONTINUS

### Énoncé

Dans le cadre d'une étude sur l'aménagement touristique du territoire, des données relatives aux 22 régions françaises vous sont fournies :

- le nombre de chambres classées, qui reflète la capacité d'accueil des hôtels de la région ;
- le nombre de nuitées, qui correspond à la fréquentation de la région.

Région	Nombre de chambres classées (milliers)	Nombre de nuitées (milliers)
Alsace	18,874	5 783,190
Aquitaine	29,367	8 249,402
Auvergne	16,488	3 585,167
Basse-Normandie	13,916	4 717,249
Bourgogne	14,673	4 780,127
Bretagne	23,815	6 942,431
Centre	19,713	5 879,467
Champagne-Ardenne	8,119	2 747,915
Corse	11,288	2 720,622
Franche-Comté	7,807	2 080,166
Haute-Normandie	9,119	3 048,212
Île-de-France	146,247	61 479,881
Languedoc-Roussillon	25,981	7 834,973
Limousin	5,198	1 279,118
Lorraine	13,713	3 657,955
Midi-Pyrénées	40,124	9 602,892
Nord-Pas-de-Calais	16,901	5 819,472
Pays de la Loire	20,162	5 711,902
Picardie	7,833	2 486,715
Poitou-Charentes	15,965	4 499,656
Provence-Alpes-Côte-d'Azur	69,120	21 442,215
Rhône-Alpes	69,812	18 311,960

Source : Insee, direction du Tourisme, partenaires régionaux, 2007

- Précisez la nature des caractères étudiés.
- Dressez le tableau de contingence présentant les deux distributions marginales.  
Utilisez les classes  $[0 ; 15[$  ;  $[15 ; 30[$  et  $[30 ; 150[$  pour X.  
Utilisez les classes  $[0 ; 5\ 000[$  ;  $[5\ 000 ; 10\ 000[$  ;  $[10\ 000 ; 65\ 000[$  pour Y.

### Solution

- Les deux variables  $X = \text{« nombre de chambres classées »}$  et  $Y = \text{« nombre de nuitées »}$  sont des variables quantitatives continues.
- Afin de pouvoir réaliser un tableau de contingence, il est indispensable de discréteriser ces variables afin de les regrouper en classes (voir chapitre 1), sans quoi chacune d'entre elles aura 22 modalités et le tableau de contingence sera composé pour chaque ligne et pour chaque colonne d'une unique région dans les marges. En effet, aucune région n'a le même nombre de chambres ni le même nombre de nuitées qu'une autre.

Pour la variable  $X = \text{« nombre de chambres classées »}$ , nous choisissons les classes suivantes :  $[0 ; 15[$ ,  $[15 ; 30[$  et  $[30 ; 150[$  (en milliers).

Pour la variable  $Y = \text{« nombre de nuitées »}$ , nous choisissons les classes suivantes :  $[0 ; 5\ 000[$ ,  $[5\ 000 ; 10\ 000[$  et  $[10\ 000 ; 65\ 000[$  (en milliers).

Le recensement manuel permet d'obtenir le tableau suivant :

<b>X \ Y</b>	<b>[0 ; 5 000[</b>	<b>[5 000 ; 10 000[</b>	<b>[10 000 ; 65 000[</b>	<b>Somme</b>
$[0 ; 15[$	9	0	0	9
$[15 ; 30[$	2	7	0	9
$[30 ; 150[$	0	1	3	4
<b>Somme</b>	11	8	3	22



### EXERCICE 3 CONTENU D'UN TABLEAU DE CONTINGENCE

#### Énoncé

Le tableau suivant recense les pays de l'Europe des 25 selon :

- la taille de leur population (en millions d'habitants), notée X et indiquée en ligne ;
- le nombre de voix dont ils disposent au conseil de l'Union européenne, noté Y et indiqué en colonne.

<b>Pop (X) \ Voix (Y)</b>	<b>[0 ; 5[</b>	<b>[5 ; 10[</b>	<b>[10 ; 15[</b>	<b>[15 ; 30[</b>
$[0 ; 5[$	6	2	0	0
$[5 ; 10[$	0	3	2	0
$[10 ; 50[$	0	0	6	2
$[50 ; 100[$	0	0	0	4

Source : PNUD, Rapport mondial sur le développement humain, 2003

1. Dressez le tableau contenant les effectifs partiels et marginaux.
2. Dressez le tableau des fréquences partielles et marginales.
3. Dressez le tableau des fréquences conditionnelles de X selon Y.
4. Dressez le tableau des fréquences conditionnelles de Y selon X.
5. À partir des questions précédentes, concluez sur la dépendance entre X et Y.

**Solution**

1. Les **effectifs partiels** des caractères X et Y sont notés  $n_{ij}$  et sont indiqués dans le **corps du tableau de contingence** (voir figure 5.15). Ils correspondent aux effectifs donnés dans l'énoncé. Ainsi, par exemple,  $n_{23} = 2$ , soit 2 pays de l'Europe des 25 ont une population comprise entre 5 et 10 millions d'habitants et ont entre 10 et 15 voix au conseil de l'Union européenne.

Les **effectifs marginaux** du caractère X se notent  $n_{i+}$  et sont indiqués dans la **dernière colonne** du tableau de contingence, appelée marge (voir figure 5.15). Ainsi, par exemple,  $n_{2+} = \sum_{j=1}^4 n_{2j} = 5$ , soit 5 pays de l'Europe des 25 ont une population comprise entre 5 et 10 millions d'habitants.

Les **effectifs marginaux** du caractère Y se notent  $n_{+j}$  et sont indiqués dans la **dernière ligne** du tableau de contingence, appelée marge (voir figure 5.15). Ainsi, par exemple,  $n_{+3} = \sum_{i=1}^4 n_{i3} = 8$ , soit 8 pays de l'Europe des 25 ont entre 10 et 15 voix au conseil de l'Union européenne.

**Figure 5.15**

**Les effectifs partiels ( $n_{ij}$ ) et marginaux ( $n_{i+}$ ;  $n_{+j}$ ).**

10	Pop (X) \ Voix (Y)	$y_1=[0 ; 5]$	$y_2=[5 ; 10]$	$y_3=[10 ; 15]$	$y_4=[15 ; 30]$	$n_{i+}$
11	$x_1=[0 ; 5[$	6	2	0	0	8
12	$x_2=[5 ; 10[$	0	3	2	0	5
13	$x_3=[10 ; 50[$	0	0	6	2	8
14	$x_4=[50 ; 100[$	0	0	0	4	4
15	$n_{ij}$	6	5	8	6	25

2. Les **fréquences partielles** des caractères X et Y se notent  $f_{ij}$  et sont indiquées dans le **corps du tableau de contingence** (voir figure 5.16). Ainsi, par exemple,  $f_{23} = \frac{n_{23}}{n_{++}} = \frac{2}{25} = 8\%$ , soit 8 % des pays de l'Europe des 25 ont une population comprise entre 5 et 10 millions d'habitants et ont entre 10 et 15 voix au conseil de l'Union européenne.

Les **fréquences marginales** du caractère X se notent  $f_{i+}$  et sont indiquées dans la **dernière colonne** du tableau de contingence, appelée marge (voir figure 5.16). Ainsi, par exemple,  $f_{2+} = \frac{n_{2+}}{n_{++}} = \frac{5}{25} = 20\%$ , soit 20 % des pays de l'Europe des 25 ont une population comprise entre 5 et 10 millions d'habitants.

Les **fréquences marginales** du caractère Y se notent  $f_{+j}$  et sont indiquées dans la **dernière ligne** du tableau de contingence, appelée marge (voir figure 5.16). Ainsi, par exemple,  $f_{+3} = \frac{n_{+3}}{n_{++}} = \frac{8}{25} = 32\%$ , soit 32 % des pays de l'Europe des 25 ont entre 10 et 15 voix au conseil de l'Union européenne.

**Figure 5.16**

**Les fréquences partielles ( $f_{ij}$ ) et marginales ( $f_{i+}$  ;  $f_{+j}$ ).**

19	Pop (X) \ Voix (Y)	$y_1=[0 ; 5[$	$y_2=[5 ; 10[$	$y_3=[10 ; 15[$	$y_4=[15 ; 30[$	$f_{+}$
20	$x_1=[0 ; 5[$	24%	8%	0%	0%	32%
21	$x_2=[5 ; 10[$	0%	12%	8%	0%	20%
22	$x_3=[10 ; 50[$	0%	0%	24%	8%	32%
23	$x_4=[50 ; 100[$	0%	0%	0%	16%	16%
24	<b><math>f_{+j}</math></b>	24%	20%	32%	24%	100%

3. Les **fréquences conditionnelles de X selon Y** se notent  $f_{i+j}$  et sont indiquées dans le **corps du tableau de contingence** (voir figure 5.17). Ainsi, par exemple,

$$f_{i=2/j=3} = \frac{n_{23}}{n_{+3}} = \frac{2}{8} = 25\% .$$

Parmi les pays de l'Europe des 25 qui disposent de 10 à 15 voix au conseil de l'Union européenne, 25 % ont une population comprise entre 5 et 10 millions d'habitants.

La somme en colonne des fréquences conditionnelles de X selon Y fait 100 %. Ces fréquences correspondent donc aux **pourcentages en colonne** : la somme des pourcentages de chacune des colonnes est égale à 100 %.

**Figure 5.17**

**Les fréquences conditionnelles de X selon Y :  $f_{i+j}$**

28	Pop (X) \ Voix (Y)	$y_1=[0 ; 5[$	$y_2=[5 ; 10[$	$y_3=[10 ; 15[$	$y_4=[15 ; 30[$
29	$x_1=[0 ; 5[$	100%	40%	0%	0%
30	$x_2=[5 ; 10[$	0%	60%	25%	0%
31	$x_3=[10 ; 50[$	0%	0%	75%	33%
32	$x_4=[50 ; 100[$	0%	0%	0%	67%
33	<b>Somme</b>	100%	100%	100%	100%

4. Les **fréquences conditionnelles de Y selon X** se notent  $f_{j+i}$  et sont indiquées dans le **corps du tableau de contingence** (voir figure 5.18). Ainsi, par exemple,

$$f_{j=3/i=2} = \frac{n_{23}}{n_{2+}} = \frac{2}{5} = 40\% .$$

Parmi les pays de l'Europe des 25 qui ont une population comprise entre 5 et 10 millions d'habitants, 40 % disposent de 10 à 15 voix au conseil de l'Union européenne.

La somme en ligne des fréquences conditionnelles de Y selon X fait 100 %. Ces fréquences correspondent donc aux **pourcentages en ligne** : la somme des pourcentages de chacune des lignes est égale à 100 %.

**Figure 5.18**

**Les fréquences conditionnelles de Y selon X :  $f_{j+i}$**

37	Pop (X) \ Voix (Y)	$y_1=[0 ; 5[$	$y_2=[5 ; 10[$	$y_3=[10 ; 15[$	$y_4=[15 ; 30[$	<b>Somme</b>
38	$x_1=[0 ; 5[$	75%	25%	0%	0%	100%
39	$x_2=[5 ; 10[$	0%	60%	40%	0%	100%
40	$x_3=[10 ; 50[$	0%	0%	75%	25%	100%
41	$x_4=[50 ; 100[$	0%	0%	0%	100%	100%

**5. X n'est pas fonctionnellement lié à Y**, car à la modalité de  $y_2$  correspondent deux modalités possibles de X,  $x_1$  et  $x_2$ ; de même, **Y n'est pas fonctionnellement lié à X**, car à la modalité de  $x_2$  correspondent deux modalités possibles de Y,  $y_2$  et  $y_3$ .

Ainsi, par exemple, les pays de l'Europe des 25 dont la taille de la population est comprise entre 10 et 50 millions d'habitants peuvent disposer de 10 à 15 voix ou de 15 à 30 voix au conseil de l'Union européenne. Inversement, les pays de l'Europe des 25 qui ont entre 10 et 15 voix au conseil de l'Union européenne peuvent avoir une population comprise entre 5 et 10 millions ou entre 10 et 50 millions d'habitants.

**X et Y ne sont pas indépendants**, car les distributions conditionnelles ne sont pas égales aux distributions marginales. En effet, par exemple,  $f_{j=3/i=2} = 40\%$  est différent de  $f_{+3} = 32\%$ .

Puisque X et Y ne sont ni dans une relation de liaison fonctionnelle, ni dans une relation d'indépendance, on se trouve entre ces deux cas extrêmes et il est simplement possible de conclure qu'il existe **une liaison entre X et Y**.



## EXERCICE 4 INDICATEURS SUR TABLEAU DE CONTINGENCE

### Énoncé

Le tableau suivant recense le nombre de personnes tuées dans un accident de la route en 2005 (millions d'individus de la classe d'âge), en fonction de l'âge (X) et du sexe (Y) :

Age (X) \ Sexe (Y)	Homme	Femme
[0 ; 15[	15	10
[15 ; 20[	241	70
[20 ; 25[	362	77
[25 ; 45[	161	36
[45 ; 65[	102	35
[65 ; 95[	145	67

Source : ONISR, 2006

1. Pour la variable « âge des tués par accidents de la route », calculez :
  - a. la moyenne marginale  $\bar{x}$  ;
  - b. la variance marginale  $V(x)$ .
2. Pour la variable « âge des tués par accidents de la route » conditionnée par la modalité « homme » de la variable « sexe », calculez :
  - a. la moyenne conditionnelle, soit  $\bar{x}_1$  ;
  - b. la variance conditionnelle, soit  $V_1(x)$ .
3. Effectuez un test du khi-deux au seuil de signification de 5 %. Concluez sur la dépendance entre l'âge et le sexe des personnes tuées dans un accident de la route.

**Solution**

1. Saisissez les centres de classes de X dans la colonne L1, les effectifs partiels pour les hommes dans la colonne L2 et les effectifs partiels pour les femmes dans la colonne L3, comme indiqué figure 5.19.

**Figure 5.19**

**Saisie du tableau de contingence avec la calculatrice.**

L1	L2	L3	3
7,5	15	10	
17,5	241	70	
22,5	362	77	
35	161	36	
55	102	35	
80	145	62	
-----	-----	-----	-----
L3(7) =			

Pour calculer les effectifs marginaux ( $n_{i+}$ ) de X dans la colonne L4, placez le curseur sur l'en-tête de colonne L4. Indiquez L4=L2+L3. Puis appuyez sur **ENTER**.

Pour obtenir les ( $n_{i+}x_i$ ) dans la colonne L5, placez le curseur sur l'en-tête de colonne L5, puis indiquez L5=L4×L1. Puis appuyez sur **ENTER**.

Pour obtenir les ( $n_{i+}x_i^2$ ) dans la colonne L6, placez le curseur sur l'en-tête de colonne L6, puis indiquez L6=L5×L1. Puis appuyez sur **ENTER**.

Pour effectuer la somme des ( $n_{i+}$ ), placez le curseur sur la cellule L4(7), et indiquez L4(7)=sum(L4), en appelant la fonction SUM (voir annexe 1.2). Puis appuyez sur **ENTER**.

Pour effectuer la somme des ( $n_{i+}x_i$ ), placez le curseur sur la cellule L5(7), et indiquez L5(7)=sum(L5), en appelant la fonction SUM. Puis appuyez sur **ENTER**.

Pour effectuer la somme des ( $n_{i+}x_i^2$ ), placez le curseur sur la cellule L6(7), et indiquez L6(7)=sum(L6), en appelant la fonction SUM. Puis appuyez sur **ENTER** (voir figure 5.20).

**Figure 5.20**

**Calcul des  $n_{i+}x_i^2$  et de la somme des colonnes avec la calculatrice.**

L4	L5	L6	6
311	5442,5	95244	
439	9877,5	222244	
197	6895	241325	
132	7535	414425	
212	16960	13666	
1321	46898	2331443,75	
-----	-----	-----	-----
L6(7) =			

- a. La moyenne marginale de X est donc égale à  $\bar{x} = \frac{1}{1321} \sum_{i=1}^6 n_{i+}x_i = \frac{46897,5}{1321}$ , soit  $\bar{x} = 35,5$ . L'âge moyen des personnes tuées dans un accident de la route est de 35,5 ans.

b. La variance marginale de X est donc égale à :

$$V(x) = \frac{1}{1321} \sum_{i=1}^6 n_{ii} x_i^2 - \bar{x}^2 = \frac{2331443,8}{1321} - 35,5^2, \text{ soit } V(x) = 504,55.$$

La variance de l'âge des personnes tuées dans un accident de la route est de 504,55.

2. Effacez le contenu des colonnes L4 et L5 en plaçant le curseur sur chacun des en-têtes de colonnes et en appuyant sur **CLEAR** et **ENTER**.

Pour calculer les  $n_{ii}x_i$  dans la colonne L4, placez le curseur sur l'en-tête de colonne L4. Indiquez  $L4=L1\times L2$ . Puis appuyez sur **ENTER**.

Pour obtenir les  $n_{ii}x_i^2$  dans la colonne L5, placez le curseur sur l'en-tête de colonne L5, puis indiquez  $L5=L4\times L1$ . Puis appuyez sur **ENTER**.

Pour faire la somme des  $n_{ii}$ , placez le curseur sur la cellule L2(7), et indiquez  $L2(7)=\text{sum}(L2)$ , en appelant la fonction SUM (voir annexe 1.2). Puis appuyez sur **ENTER**.

Pour faire la somme des  $n_{ii}x_i$ , placez le curseur sur la cellule L4(7), et indiquez  $L4(7)=\text{sum}(L4)$ , en appelant la fonction SUM. Puis appuyez sur **ENTER**.

Pour faire la somme des  $n_{ii}x_i^2$ , placez le curseur sur la cellule L5(7), et indiquez  $L5(7)=\text{sum}(L5)$ , en appelant la fonction SUM. Puis appuyez sur **ENTER** (voir figure 5.21).

Figure 5.21

**Calcul des  $n_{ii}x_i^2$  et de la somme des colonnes avec la calculatrice.**

L3	L4	L5	5
70	4217,5	73806	
77	8145	183263	
36	5635	197225	
35	5610	308550	
67	11600	928000	
-----	35320	1691687,5	
<b>L5(7) = 1691687,5</b>			

a. La moyenne conditionnelle cherchée est donc :  $\bar{x}_1 = \frac{1}{1026} \sum_{i=1}^6 n_{ii} x_i = \frac{35320}{1026}$ , soit

$\bar{x}_1 = 34,42$ . L'âge moyen des hommes tués dans un accident de la route est de 34,42 ans.

b. La variance conditionnelle cherchée est :

$$V_1(x) = \frac{1}{1026} \sum_{i=1}^6 n_{ii} x_i^2 - \bar{x}_1^2 = \frac{1691687,5}{1026} - 34,42^2, \text{ soit } V(x) = 463,74.$$

La variance de l'âge des hommes tués dans un accident de la route est de 463,74.

3. Pour effectuer un test du khi-deux, il convient de saisir le tableau de données observées dans une matrice. Pour cela, appuyez sur la touche **MATRIX**, choisissez le menu EDIT. Tapez 1 pour éditer la matrice [A]. Saisissez le nombre de lignes, soit 6, et appuyez sur **ENTER**. Saisissez le nombre de colonnes, soit 2, et appuyez sur **ENTER**. Enfin, saisissez les valeurs en validant chacune d'entre elles par appui sur **ENTER**.

La matrice [A] de la calculatrice contient ainsi les données observées (voir figure 5.22).

Le test du khi-deux compare cette matrice observée avec la matrice théorique, construite sous l'hypothèse d'indépendance entre X et Y. Pour effectuer ce test à l'aide de la calculatrice, appuyez sur la touche **STAT**, choisissez le menu TESTS et tapez C pour appeler le test du khi-deux. Par défaut, la matrice de données observées est la matrice [A]. Tapez sur **ENTER** pour valider. Par défaut, la matrice où seront stockés les résultats de la matrice théorique est la matrice [B]. Tapez sur **ENTER** pour valider. Puis tapez une nouvelle fois sur **ENTER** pour lancer le test du khi-deux. Les résultats s'affichent à l'écran (voir figure 5.23).

Figure 5.22 (gauche)

Saisie de la matrice [A]  
des effectifs observés  
avec la calculatrice.

L1	L2	L3	3
4	5	1	
6	7	1	
7	11	16	
11	14	9	
2	1	1	
8	8	0	
9	4	25	

Figure 5.23 (droite)

Résultats du test du khi-deux avec la calculatrice.

X<sup>2</sup>-Test  
X<sup>2</sup>=23.51174785  
P=2.6939811e-4  
df=5

La probabilité 0,000269, soit environ 0,03 %, donnée ici est celle que l'on obtiendrait sous Excel avec la fonction LOI.KHIDEUX. La valeur du khi-deux de 23,51, avec un degré de liberté de 5, a une probabilité d'environ 0,03 % d'être dépassée ou correspond à un seuil de signification de 0,03 %. Ce seuil de signification est inférieur à 5 %, et induit donc le rejet de l'hypothèse nulle au seuil fixé de 5 % et l'acceptation de l'hypothèse alternative. Si H<sub>0</sub> est vraie, il y a 99,97 % de chances d'obtenir un échantillon correspondant à un khi-deux inférieur à 23,51 ; en rejetant H<sub>0</sub>, on prend ici un risque négligeable. Il existe donc un grand écart entre les données observées et les données théoriques sous hypothèse d'indépendance. Les données observées reflètent un degré de **dépendance statistique entre X et Y**. Autrement dit, il existe un lien entre l'âge et le genre des personnes tuées dans un accident de la route.

Au seuil de 5 %, avec un ddl de 5, la table ou la fonction statistique Excel KHIDEUX.INVERSE nous donne un khi-deux de 11,05, qui est largement dépassé ici par le khi-deux calculé.

La matrice [B] des données théoriques peut être visualisée en appuyant sur la touche **MATRIX**. Dans le menu EDIT, tapez 2 pour éditer la matrice [B] (voir figure 5.24).

Figure 5.24

Visualisation de la matrice [B] des effectifs calculés avec la calculatrice.

MATRIX[B] 6 ×2

[ 19.417	5.5829
[ 241.55	69.451
[ 340.96	98.036
[ 153.01	43.993
[ 106.41	30.594
[ 164.66	47.343

On vérifie que, pour chaque élément de la matrice,  $\chi^2_i = \frac{(Obs_i - Thq_i)^2}{Thq_i}$ . Par exemple, pour l'élément situé à l'intersection de la 2<sup>e</sup> ligne et de la 1<sup>re</sup> colonne,

$$241,55 = \frac{311 / 1321 \times 1026 / 1321}{1026 / 1321} \times 1321.$$

## EXERCICE 5 DÉPENDANCE ENTRE DEUX VARIABLES

### Énoncé

318 étudiants ont été interrogés sur leurs achats de jeux vidéo neufs et d'occasion au cours de la dernière année. Le tableau suivant croise le nombre de jeux achetés neufs (X) avec le nombre de jeux achetés d'occasion (Y).

Neuf (X) \ Occasion (Y)	0	1	[2 ; 4[
0	157	8	5
1	55	8	8
[2 ; 4[	49	9	19

- Calculez la moyenne marginale  $\bar{x}$  et la variance  $V(x)$ .
- Calculez la moyenne marginale  $\bar{y}$  et la variance  $V(y)$ .
- Calculez la covariance entre X et Y. Concluez sur la dépendance entre X et Y.
- Effectuez un test du khi-deux au seuil de signification de 5 %. Concluez sur la dépendance entre X et Y.

### Solution

1. Afin d'obtenir la valeur de la moyenne marginale de X, il convient de calculer :

- les effectifs marginaux ( $n_{i+}$ ) de X dans la colonne E, ainsi que leur somme dans la cellule E5 ;
- les centres de classes  $x_i$  dans la colonne F ;
- les ( $n_{i+}x_i$ ) dans la colonne G, ainsi que leur somme dans la cellule G5.

Pour le calcul de la variance marginale de X, les ( $n_{i+}x_i^2$ ) sont calculés dans la colonne H, et leur somme dans la cellule H5 (voir figure 5.25).

Figure 5.25

Résultats sous Excel.

A	B	C	D	E	F	G	H
1 X \ Y	y=0	y2=1	y3=[2 ; 4[	n <sub>i+</sub>	x <sub>i</sub>	n <sub>i+xi</sub>	n <sub>i+xi</sub> <sup>2</sup>
2 x <sub>i</sub> =0	157	8	5	170	0	0	0
3 x <sub>2</sub> =1	55	8	8	71	1	71	71
4 x <sub>3</sub> =[2 ; 4[	49	9	19	77	3	231	693
5 n <sub>i+</sub>	261	25	32	318		302	764
6 y <sub>i</sub>	0	1	3				
7 n <sub>i+yi</sub>	0	25	96	121			
8 n <sub>i+yi</sub> <sup>2</sup>	0	25	288	313			
9 Σ <sub>i</sub> n <sub>i+xi</sub> y <sub>i</sub>	0	35	195	230			

La moyenne marginale de X est donc égale à  $\bar{x} = \frac{1}{318} \sum_{i=1}^3 n_{i+} x_i = \frac{302}{318}$ , soit  $\bar{x} = 0,95$ .

Le nombre moyen de jeux vidéo achetés neufs lors de la dernière année est de 0,95.

La variance marginale de X est donc égale à  $V(x) = \frac{1}{318} \sum_{i=1}^3 n_{i+} x_i^2 - \bar{x}^2 = \frac{764}{318} - 0,95^2$ , soit  $V(x) = 1,50$ . La variance des jeux vidéo achetés neufs lors de la dernière année est de 1,50.

**2.** Afin d'obtenir la valeur de la moyenne marginale de Y, il convient de calculer à la suite du tableau précédent (voir figure 5.25) :

- les effectifs marginaux ( $n_{+j}$ ) de Y sur la ligne 5, ainsi que leur somme dans la cellule E5 ;
- les centres de classes  $y_j$  sur la ligne 6 ;
- les ( $n_{+j} y_j$ ) sur la ligne 7, ainsi que leur somme dans la cellule E7.

Pour le calcul de la variance marginale de X, les ( $n_{i+} x_i^2$ ) sont calculés sur la ligne 8, et leur somme dans la cellule E8.

La moyenne marginale de Y est donc égale à  $\bar{y} = \frac{1}{318} \sum_{j=1}^3 n_{+j} y_j = \frac{121}{318}$ , soit  $\bar{y} = 0,38$ .

Le nombre moyen de jeux vidéo achetés d'occasion lors de la dernière année est de 0,38.

La variance marginale de Y est donc égale à  $V(y) = \frac{1}{318} \sum_{i=1}^3 n_{+j} y_j^2 - \bar{y}^2 = \frac{313}{318} - 0,38^2$ , soit  $V(y) = 0,84$ . La variance des jeux vidéo achetés d'occasion lors de la dernière année est de 0,84.

**3.** Pour obtenir la valeur de la covariance entre X et Y, nous calculons à la suite du tableau précédent (voir figure 5.25) les  $\sum_{i=1}^p n_{ij} x_i y_j$  pour chaque colonne j, sur la ligne 9, dans les cellules B9, C9 et D9. Puis nous en effectuons la somme en faisant varier j dans la cellule E9, afin d'obtenir la somme :  $\sum_{j=1}^q \sum_{i=1}^p n_{ij} x_i y_j$ .

La covariance de (X ; Y) est donc égale à

$$COV(X;Y) = \frac{1}{318} \sum_{j=1}^3 \sum_{i=1}^3 n_{ij} x_i y_j - \bar{x} \times \bar{y} = \frac{230}{318} - 0,95 \times 0,38, \text{ soit } Cov(X;Y) = 0,36.$$

X et Y sont positivement liés. Le nombre de jeux vidéo achetés neufs est positivement lié au nombre de jeux vidéo achetés d'occasion.

**4.** Pour effectuer un test du khi-deux, il convient de calculer les effectifs théoriques (ou calculés, notés  $c_{ij}$ ) sous l'hypothèse d'indépendance entre X et Y. Les calculs sont présentés à la figure 5.26. Par exemple, pour l'effectif théorique  $c_{21}$ :  $c_{21} = \frac{n_{2+} \times n_{+1}}{n_{++}}$ , donc

$$c_{21} = \frac{71 \times 261}{318} = 58. \text{ Autre exemple : } c_{32} = \frac{77}{318} \times 25 = 6.$$

**Figure 5.26**

**Données théoriques sous hypothèse d'indépendance sous Excel.**

21	X \ Y	0 (y <sub>1</sub> )	1 (y <sub>2</sub> )	[2 ; 4[ (y <sub>3</sub> )	n <sub>i+</sub>
22	0 (x <sub>1</sub> )	139,53	13,36	17,11	170
23	1 (x <sub>2</sub> )	58,27	5,58	7,14	71
24	[2 ; 4[ (x <sub>3</sub> )	63,20	6,05	7,75	77
25	n <sub>+j</sub>	261	25	32	318

Le test du khi-deux compare cette matrice observée avec la matrice théorique, construite sous hypothèse d'indépendance entre X et Y. Pour cela, il convient de calculer chacune des distances du khi-deux par case tel que  $\chi_{ij}^2 = \frac{(n_{ij} - c_{ij})^2}{c_{ij}}$  (voir figure 5.27). Par exemple,  $\chi_{21}^2 = \frac{(55 - 58,27)^2}{58,27} = 0,18$ . Autre exemple :  $\chi_{32}^2 = \frac{(9 - 6,05)^2}{6,05} = 1,43$ .

**Figure 5.27**

**Distances du khi-deux sous Excel.**

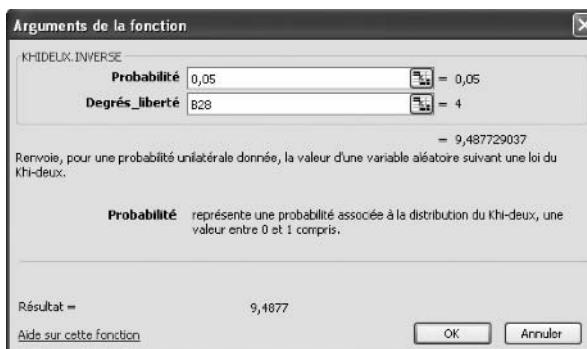
29	X \ Y	0 (y <sub>1</sub> )	1 (y <sub>2</sub> )	[2 ; 4[ (y <sub>3</sub> )	n <sub>i+</sub>
30	0 (x <sub>1</sub> )	2,19	2,15	8,57	12,91
31	1 (x <sub>2</sub> )	0,18	1,05	0,10	1,33
32	[2 ; 4[ (x <sub>3</sub> )	3,19	1,43	16,34	20,96
33	n <sub>+j</sub>	5,56	4,64	25,01	35,21

La somme des distances du khi-deux est de 35,21, soit  $\chi_c^2 = 35,21$ . Or, ce tableau a :  $(3 - 1) \times (3 - 1) = 4$  degrés de liberté. Pour définir la règle de décision, nous devons déterminer la valeur critique, c'est-à-dire  $\chi^2_{(0,05;4)}$ .

Pour effectuer une lecture de table du khi-deux sous Excel, sélectionnez la cellule dans laquelle vous souhaitez faire apparaître le résultat, puis, dans la barre de menus, cliquez sur Insertion/Fonction. Dans la boîte de dialogue, sélectionnez la catégorie Statistiques, puis sélectionnez la fonction KHIDEUX.INVERSE. Cliquez sur OK. Dans la boîte de dialogue Arguments de la fonction (voir figure 5.28), dans le champ Probabilité, indiquez le niveau de signification fixé, ici 0,05, puis, dans le champ Degrés\_liberté, indiquez la cellule dans laquelle vous aurez préalablement saisi le degré de liberté du tableau, soit 4, en cellule B28 pour notre exemple. Cliquez sur OK pour faire apparaître le résultat, soit un khi-deux d'environ 9,49.

**Figure 5.28**

**Lecture du khi-deux de la table sous Excel.**



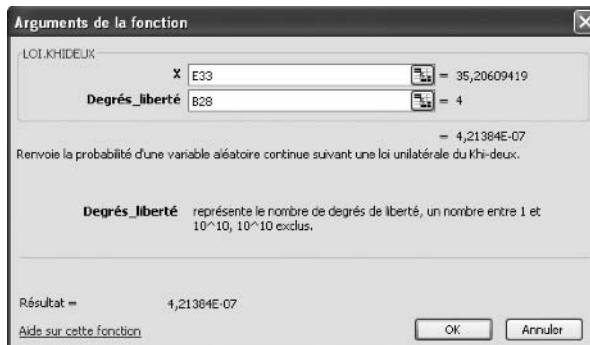
Il reste à prendre la décision : le khi-deux calculé est d'environ 35,21 ; il est supérieur au khi-deux de la table, on doit rejeter l'hypothèse d'indépendance et accepter l'hypothèse alternative de dépendance entre le nombre de jeux vidéo achetés neufs et le nombre de jeux vidéo achetés d'occasion.

Une démarche complémentaire consiste à chiffrer le degré de signification, ou la valeur p (p-value) en utilisant la fonction Excel LOI.KHIDEUX. Ce degré de signification est la probabilité d'avoir un khi-deux supérieur ou égal à 35,21 sous l'hypothèse nulle.

Pour déterminer le degré de signification, sélectionnez la cellule dans laquelle vous souhaitez faire apparaître le résultat, puis, dans la barre de menus, cliquez sur Insertion/Fonction. Dans la boîte de dialogue, sélectionnez la catégorie Statistiques, puis sélectionnez la fonction LOI.KHIDEUX. Cliquez sur OK. Dans la boîte de dialogue Arguments de la fonction (voir figure 5.29), dans le champ « x », indiquez la cellule dans laquelle se trouve la valeur du khi-deux, soit E33, et dans le champ Degrés\_Liberté, indiquez la cellule dans laquelle vous aurez préalablement saisi le degré de liberté du tableau, soit 4, en cellule B28 pour notre exemple. Cliquez sur OK pour faire apparaître le résultat.

Figure 5.29

Détermination du degré de signification pour un khi-deux sous Excel.



Pour une valeur du khi-deux de 35,21 et avec un degré de liberté de 4, la probabilité associée est de 4,2138E-07, soit 0,0000. Cette valeur du khi-deux a une probabilité pratiquement nulle d'être dépassée. Le degré de signification est inférieur au seuil de 5 % assigné au test, on doit donc rejeter l'hypothèse nulle d'indépendance entre les variables, le risque de prendre une mauvaise décision étant ici quasiment nul. Il existe donc un grand écart entre les données observées et les données théoriques sous hypothèse d'indépendance. Les données observées reflètent une **dépendance entre X et Y**. Autrement dit, il existe un lien entre le nombre de jeux vidéo achetés neufs et le nombre de jeux vidéo achetés d'occasion.

## Bibliographie

- BAILLARGEON G., *Méthodes statistiques de l'ingénieur*, SMG, 1990.
- BOUROCHE J.-M. et SAPORTA G., *L'analyse des données*, Que sais-je ?, PUF, 1990.
- CALOT G., *Cours de statistique descriptive*, Dunod, 1969.
- CHAREILLE P. et PINAULT Y., *Statistique descriptive*, Collection AES, Montchrestien, Paris, 1996.
- DODGE Y., *Statistique. Dictionnaire encyclopédique*, Springer, 2004.
- DODGE Y., *Premiers pas en statistique*, Springer, 2006.
- DROESBEKE J.-J. et TASSI Ph., *Histoire de la statistique*, Que sais-je ?, PUF, 1990.
- GRENON G. et VIAU S., *Méthodes quantitatives en sciences humaines*, Gaëtan Morin, 1999.
- MARTIN O., *L'analyse de données quantitatives. L'enquête et ses méthodes*, Armand Colin, 2005.
- WONNACOTT T.H. et R.J., *Statistique*, Economica, 1984.



# La régression

1. Les fondements de la régression..... 146
2. L'ajustement linéaire..... 150
3. Ajustements et absence de linéarité ..... 162

## Problèmes et exercices

1. Régression linéaire et indicateurs de qualité..... 164
2. Régression linéaire et prévisions..... 170
3. Régression sur tableau de contingence..... 174
4. Ajustement exponentiel et papier semi-logarithmique.. 175
5. Corrélation des rangs..... 179

Dans le chapitre précédent, nous avons vu que le degré d'association de deux caractères quantitatifs peut varier entre deux extrêmes, d'un côté la liaison fonctionnelle et de l'autre l'indépendance. La notion de corrélation consiste à préciser la dépendance mutuelle de deux variables statistiques.

Cette notion de corrélation a été esquissée pour la première fois par Francis Galton (1822-1911), dans ses travaux sur l'hérédité : il utilisait alors le terme « co-relation<sup>1</sup> ». Galton a montré que la taille moyenne des descendants était liée par une relation linéaire à la taille des parents. Les concepts introduits par Galton ont ensuite été développés par Karl Pearson (1857-1936).

Dans ce chapitre, nous étudierons essentiellement la corrélation linéaire, c'est-à-dire les situations où les variations relatives de deux caractères quantitatifs sont approximativement proportionnelles ; ce cas est fondamental, car il se produit quand le couple ( $X, Y$ ) suit une loi normale.

Ensuite, nous mesurerons l'intensité de cette corrélation à l'aide du coefficient de corrélation linéaire.

1. Formé de *cum*, avec, et de *relatio*, le mot latin *correlatio* signifie « relation mutuelle » (voir B. Hauchecorne).

L'analyse linéaire de la régression a un double objectif : d'une part expliciter le modèle décrivant les relations entre une variable privilégiée, appelée variable expliquée (dépendante ou endogène), et une variable appelée variable explicative (indépendante ou exogène), et d'autre part effectuer des prévisions de la variable expliquée en fonction de la variable explicative.

Dans ce cas, l'ajustement analytique sera effectué à l'aide de la méthode des moindres carrés, que nous devons à Carl Friedrich Gauss (1777-1855) et Adrien-Marie Legendre (1752-1833), et qui nous permettra de déterminer les équations des droites de régression. Nous envisagerons également des liaisons plus complexes (exponentielles), en utilisant une représentation graphique (nuage de points) comme outil de conjecture.

Enfin, une fois les calculs menés sur un échantillon, il importera d'utiliser un test statistique permettant de valider ou de rejeter l'existence d'un lien linéaire entre les variables sur la population.

# 1 Les fondements de la régression

## 1.1 TERMINOLOGIE

Il importe avant tout de préciser certains termes : régression, corrélation, indépendance.

Nous avons vu, au chapitre 5, un exemple (voir exemple 5.9) de liaison fonctionnelle non symétrique. De même, les notions de régression et de corrélation ne donnent pas un rôle symétrique aux deux variables. Quand deux variables ne sont pas liées par une relation fonctionnelle pure, on devra se contenter de regarder comment, en moyenne, se font les variations respectives de ces variables. On associera ainsi à chaque modalité  $x_i$  de X la moyenne conditionnelle  $\bar{y}_i$ .

### Définitions

**Liaison fonctionnelle** : On dit que la variable Y est fonctionnellement liée à X si à chaque modalité de X correspond une seule modalité de Y.

De même X est liée fonctionnellement à Y si à chaque modalité de Y correspond une seule modalité de X.

Si X est liée fonctionnellement à Y et Y est liée fonctionnellement à X, on parle de liaison fonctionnelle réciproque.

**Courbes de régression** : On appelle courbe de régression de Y selon x la courbe représentative des moyennes conditionnelles  $\bar{y}_i$  en fonction des valeurs  $x_i$  de X. On remarquera que si X est une variable discrète on aura en fait une suite de points appelée nuage de points.

On définit de même la courbe de régression de X selon y.

**Point moyen** : On appelle point moyen du nuage le point G de coordonnées respectives  $\bar{x}$  et  $\bar{y}$ .

Dans le cas particulier où les variables X et Y sont indépendantes, les distributions conditionnelles sont identiques entre elles (et confondues avec la distribution marginale correspondante). On a donc dans ce cas des moyennes conditionnelles constantes et donc des droites de régression parallèles aux axes et d'équations respectives  $x = \bar{x}$  et  $y = \bar{y}$ . On notera que la réciproque est fausse : des droites de régression parallèles aux axes n'impliquent pas l'indépendance.

Étudier la corrélation d'une variable Y avec une variable X consiste à étudier la dépendance des moyennes conditionnelles de Y en fonction des valeurs de X. L'étude de la corrélation de Y avec X se base sur la courbe de régression de Y selon X et sur la mesure de l'intensité de cette corrélation.

### Définition

**Corrélation :** Une variable Y est dite corrélée avec X si la courbe de régression de Y selon X n'est pas une droite parallèle à l'axe des abscisses.

On notera que :

- l'absence de corrélation n'est en général pas symétrique : X peut être corrélée avec Y sans que Y soit corrélée avec X ;
- si X et Y sont des variables indépendantes, X n'est pas corrélée à Y et Y n'est pas corrélée à X, mais l'indépendance n'est qu'un cas particulier d'absence de corrélation.

## 1.2 LES DIFFÉRENTS AJUSTEMENTS STATISTIQUES

Nous supposons que nous disposons d'un tableau simple donnant les modalités ( $x_i ; y_i$ ), pour  $i$  variant de 1 à  $n$ , d'un couple de variables quantitatives, pour un échantillon aléatoire, de taille  $n$ , prélevé dans la population. Avant toute étude, la série sera représentée par un diagramme de corrélation (ou de dispersion) afin d'apprecier le type d'ajustement adapté. Ce diagramme, appelé nuage de points, est obtenu en plaçant dans un repère les  $n$  points de coordonnées ( $x_i ; y_i$ ). La forme de ce nuage permettra de mettre au jour une éventuelle corrélation entre les variables. Réaliser un ajustement consiste à rechercher la meilleure relation possible entre les variables, donc à rechercher la courbe la plus « proche » de l'ensemble des points du nuage.

### Les liaisons fonctionnelles (rigides)

Ce type de liaison, que l'on rencontre par exemple dans de nombreuses lois physiques, a été défini au chapitre 5 et constitue un modèle déterministe. Une liaison fonctionnelle peut être linéaire ou non, conformément aux cas décrits dans les exemples 6.1 et 6.2.

### Exemple 6.1

#### Liaison fonctionnelle linéaire

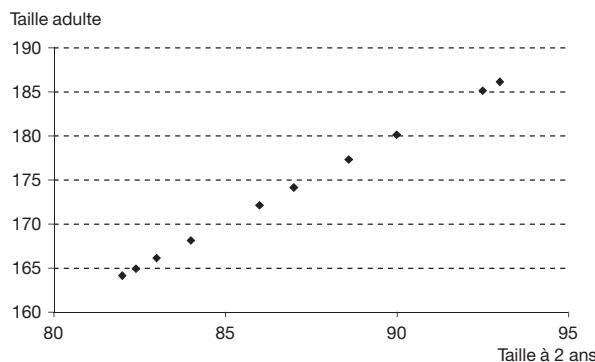
Le tableau suivant donne, pour un échantillon de 10 garçons de 18 ans prélevé dans la population d'un lycée, les tailles respectives (exprimées en centimètres),  $x_i$  et  $y_i$ , à 2 ans et à 20 ans :

X	Y
82	164,1
82,4	164,9
83	166,1
84	168,1
86	172,1
87	174,1
88,6	177,3

X	Y
90	180,1
92,5	185,1
93	186,1

**Figure 6.1**

**Exemple de relation fonctionnelle linéaire.**



Sur la figure 6.1, l’alignement des points met en évidence une relation fonctionnelle linéaire entre les deux variables. On peut vérifier que, sur cet échantillon, y est une fonction affine<sup>1</sup> de x :  $y_i = 2x_i + 0,1$ .

On notera que, si le modèle linéaire est fondamental, on ne peut négliger les autres ajustements : ajustement logarithmique, exponentiel, polynomial, puissance. Le lecteur pourra se familiariser avec ces différents modèles grâce à l’exemple 6.2 ci-après. Il pourra utiliser, dans l’assistant graphique d’Excel, le sous-menu « Ajouter une courbe de tendance », ou se reporter au corrigé de l’exercice 1, figures 6.6 et 6.7.

**Exemple 6.2**

**Liaison fonctionnelle non linéaire**

L’exemple qui suit est une illustration de « l’étonnante loi de Benford » qui modélise la fréquence d’apparition du premier chiffre significatif de données statistiques (voir J.-P. Delahaye).

On considère un échantillon de 300 pays. On note X le premier chiffre du nombre représentant la population de chaque pays (les modalités étant notées  $x_i$ ) et Y la variable dont les modalités notées  $y_i$  sont les fréquences des  $x_i$  :

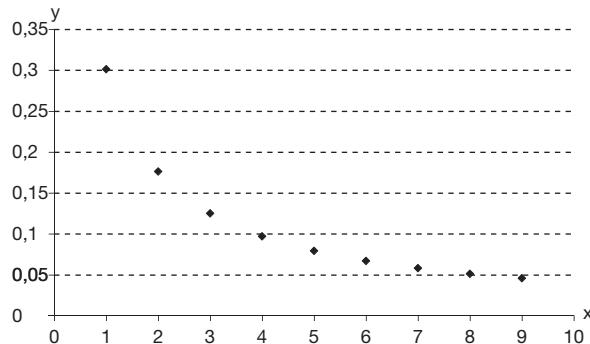
X	Y
1	0,3010
2	0,1760
3	0,1249
4	0,09691
5	0,0792

1. Une fonction affine est une fonction définie de R dans R par  $f(x) = ax + b$ , dont la représentation graphique est une droite non verticale.

X	Y
6	0,0669
7	0,0580
8	0,0511
9	0,0458

**Figure 6.2**

**Exemple de relation fonctionnelle non linéaire.**



Les points de la figure 6.2 ne sont pas alignés, mais le nuage montre l'existence d'une liaison non linéaire. Y et X sont liées par la relation logarithme décimal :  $Y = \log(1 + 1/x)$ .

### L'absence de liaison

Dans le cas d'un nuage de points diffus et répartis au hasard, il est possible de conclure à l'absence de liaison entre les variables X et Y, comme le montre l'exemple 6.3.

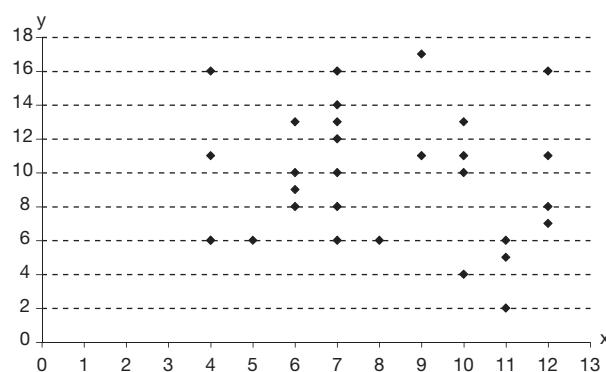
#### Exemple 6.3

#### Absence de liaison

Soit un échantillon de 31 étudiants ayant obtenu les notes X et Y dans deux matières. Le diagramme de dispersion correspondant est proposé figure 6.3.

**Figure 6.3**

**Absence de corrélation.**



Ce nuage de points sans liens apparents permet de conjecturer une absence de liaison entre les variables. Il n'y a pas corrélation entre X et Y.

## Les liaisons statistiques

Dans de nombreuses sciences, nous cherchons à mettre en évidence une liaison entre deux variables X et Y. Le plus souvent, la liaison cherchée n'est pas purement fonctionnelle et l'on parle de liaison stochastique pour exprimer qu'à une valeur de X correspond un ensemble de valeurs possibles de Y, distribuées suivant une loi de probabilité. Dans ce cas, les points ne sont plus alignés, mais le nuage de points a une forme allongée qui évoque une droite.

Cette droite constitue une liaison statistique entre les deux variables ; il nous reste à préciser en quoi cette droite est « la plus proche du nuage » et à exposer la méthode permettant de déterminer une équation de cette droite : la méthode des moindres carrés ordinaires (MCO).

# 2 L'ajustement linéaire

En cherchant à mettre en évidence une fonction  $f$  qui représente la liaison statistique entre deux variables X et Y, on se trouve face au problème général de l'interpolation. La détermination analytique de  $f$  aurait *a priori* comme seule contrainte de vérifier  $y_i = f(x_i)$ , avec Y la variable expliquée et en faisant abstraction des erreurs dues à l'échantillon.

Dans le cas où le nuage de points a une forme allongée, on présume un ajustement linéaire. La fonction cherchée est une fonction affine. Le but est de trouver la meilleure droite qui résume le nuage de points, ce qui nous amène à résoudre un problème d'interpolation linéaire. Pour cela, nous utilisons une propriété importante de la moyenne arithmétique : la moyenne arithmétique d'une série est le nombre le plus proche de cette série au sens des moindres carrés.

## 2.1 DROITES DE RÉGRESSION PAR LA MÉTHODE MCO

La loi normale ou de Laplace-Gauss est encore appelée loi des erreurs ou des écarts, car c'est ainsi qu'elle a été introduite. Le principe de la méthode des moindres carrés ordinaires (MCO) consiste à s'intéresser à la série statistique des erreurs ou résidus ( $e_i$ ).

On notera que l'on peut émettre des hypothèses sur le choix de la variable expliquée, mais que le statisticien doit également mener les calculs dans le cas où X est la variable expliquée. Il appartiendra au spécialiste concerné – économiste, médecin, etc. – de décider éventuellement d'écartier un des cas sur la base d'une analyse propre à sa spécialité.

### Définitions

On appelle **droite de régression de Y selon x**, notée  $D_{Y/x}$ , déterminée par la méthode des moindres carrés, la droite d'équation  $y = ax + b$ , pour laquelle la somme des carrés des résidus est minimale.

On note  $\hat{y}_i = ax_i + b$  la valeur de  $y_i$  estimée par la droite de régression de Y selon x.

De même, la **droite de régression de X selon y**, notée  $D_{x/y}$ , est la droite d'équation  $x = a'y + b'$  avec  $\hat{x}_i = a'y_i + b'$ .

Notons que, graphiquement, la somme des carrés des résidus représente la somme des carrés des écarts entre les points du nuage et la droite, écarts calculés parallèlement à l'axe des ordonnées dans le cas de la droite de régression de Y selon x.

À partir du modèle linéaire construit, il est possible d'effectuer des prévisions. Dans le cas d'une liaison linéaire avérée, une fois déterminée la droite de régression de Y selon x, on peut l'utiliser pour estimer la valeur de y associée à une valeur de x appartenant à l'étendue des valeurs de x retenues dans l'échantillon. Dans ce cas, il n'y a pas de raison statistique de supposer que le modèle linéaire puisse se prolonger au-delà de l'intervalle étudié. Si l'on effectue des prévisions en dehors de l'intervalle défini par les valeurs extrêmes de x, on peut obtenir des valeurs aberrantes.

On pourra sortir de cet intervalle, notamment dans les séries chronologiques, à condition d'avoir des informations sur la stabilité de la liaison linéaire.

### Détermination des droites de régression

Remarque préalable : nous cherchons à déterminer les paramètres a et b traduisant une éventuelle liaison linéaire du type  $Y = aX + b$  (dans le cas de la droite de régression de Y selon x) entre les variables X et Y ; pour cela, nous devons déterminer les paramètres a et b de la droite qui « s'éloigne » le moins du nuage de points constitué par un échantillon de taille n de la population. En conséquence, nous allons déterminer des estimateurs de a et b, c'est-à-dire des fonctions des n observations de l'échantillon, notées  $\hat{a}$  et  $\hat{b}$ , qui permettent d'obtenir les meilleures estimations possibles des paramètres a et b.

Dans les calculs, nous garderons les notations a et b de la statistique descriptive.

Posons :  $\hat{y} = ax + b$ , a désignant le coefficient directeur de la droite  $D_{Y/x}$  et b l'ordonnée à l'origine (on notera que certains auteurs prennent la notation :  $\hat{y} = a + bx$ ).

Nous devons déterminer les estimateurs de a et b qui minimisent

$$S = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - ax_i - b)^2 = S(a; b).$$

S est une fonction de deux variables et les mathématiques nous enseignent que les conditions nécessaires du premier ordre pour avoir un extremum (minimum ou maximum) sont :

$$\begin{cases} \frac{\partial S}{\partial a} = 0 \\ \frac{\partial S}{\partial b} = 0 \end{cases}, \text{c'est-à-dire la nullité des dérivées partielles premières.}$$

$$\frac{\partial S}{\partial a} = -2 \sum_{i=1}^n x_i (y_i - ax_i - b) \text{ et } \frac{\partial S}{\partial b} = -2 \sum_{i=1}^n (y_i - ax_i - b); \text{ on doit résoudre le système :}$$

$$\begin{cases} \sum_{i=1}^n x_i (y_i - ax_i - b) = 0 \\ \sum_{i=1}^n (y_i - ax_i - b) = 0 \end{cases}. \text{En utilisant les relations } \sum_{i=1}^n x_i = n\bar{x} \text{ et } \sum_{i=1}^n y_i = n\bar{y}, \text{ on obtient :}$$

$$\begin{cases} \sum_{i=1}^n (x_i y_i - ax_i^2 - bx_i) = 0 \\ n\bar{y} - an\bar{x} - nb = 0 \end{cases}. \text{ La deuxième équation du système s'écrit } b = \bar{y} - a\bar{x}, \text{ ce qui}$$

permet de remplacer  $b$  par sa valeur dans la première équation du système, ce qui donne :  $\sum_{i=1}^n x_i y_i - a \sum_{i=1}^n x_i^2 - (\bar{y} - a\bar{x}) \sum_{i=1}^n x_i = 0$  soit  $\sum_{i=1}^n x_i y_i - a \sum_{i=1}^n x_i^2 - n\bar{x}(\bar{y} - a\bar{x}) = 0$  soit  $a \left( \sum_{i=1}^n x_i^2 - n\bar{x}^2 \right) = \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}$ , qui donne :  $a = \frac{nCov(X;Y)}{nV(X)} = \frac{Cov(X;Y)}{V(X)}$ . Nous admettrons que ces valeurs correspondent bien à un minimum.

Les calculs sont similaires pour la droite de régression de  $X$  selon  $y$  ; on retiendra donc les résultats suivants pour les estimateurs de  $a$  et  $b$ , notés  $\hat{a}$  et  $\hat{b}$  :

$$D_{Y/x} : y = ax + b, \text{ avec } \begin{cases} \hat{a} = \frac{Cov(X;Y)}{V(X)} \\ \hat{b} = \bar{y} - \hat{a}\bar{x} \end{cases} \quad \text{et} \quad D_{X/y} : x = a'y + b', \text{ avec } \begin{cases} \hat{a}' = \frac{Cov(X;Y)}{V(Y)} \\ \hat{b}' = \bar{x} - \hat{a}'\bar{y} \end{cases}.$$

Ces deux droites se coupent au point moyen  $G$ . La droite de régression de  $X$  selon  $y$  peut être mise sous forme affine :  $y = (1/a')x - (b'/a')$ , de façon à faire apparaître son coefficient directeur :  $1/a'$ .

#### Exemple 6.4

#### Calculs de droites de régression

Le tableau suivant donne les indices du pouvoir d'achat (base 100 en 1951) du salaire minimum net, noté  $X$ , et du salaire moyen, noté  $Y$ , pour les salariés français des secteurs privé et semi-public.

Année	X	Y
1994	293	329
1995	296	336
1996	296	334,35
1997	302,15	337,33
1998	311,45	340,34
1999	313,93	345,76
2000	315,47	347,46
2001	321,99	349,17
2002	326,41	352,25
2003	330,57	350,87

Source : Insee, 2006

Pour calculer les coefficients des droites de régression, il est nécessaire de calculer les moyennes, les écarts-types et la covariance de  $X$  et  $Y$ . La figure 6.4 propose les calculs intermédiaires nécessaires, réalisés sous Excel.

**Figure 6.4****Calculs préalables sous Excel.**

	A	B	C	D	E	F
1	Année	$x_i$	y	$x_i^2$	$y^2$	$x_i y$
2	1994	293,00	329,00	85 849,00	108 241,00	96 397,00
3	1995	296,00	336,00	87 616,00	112 896,00	98 448,00
4	1996	296,00	334,35	87 616,00	111 789,92	97 964,55
5	1997	302,15	337,33	91 294,62	113 791,53	98 837,69
6	1998	311,45	340,34	97 001,10	115 631,32	99 719,62
7	1999	313,93	345,76	98 552,04	119 549,98	101 307,68
8	2000	315,47	347,46	99 521,32	120 728,45	101 805,78
9	2001	321,99	349,17	103 677,56	121 919,69	102 306,81
10	2002	326,41	362,25	106 643,49	124 080,06	103 209,25
11	2003	330,57	350,87	109 276,52	123 109,76	102 804,91
12	Somme	3 106,97	3 422,53	966 947,66	1 171 937,70	1 002 801,29

$$\text{De là, } \bar{x} = \frac{1}{10} \sum_{i=1}^{10} x_i = \frac{3 106,97}{11} = 310,7, \bar{y} = \frac{1}{10} \sum_{i=1}^{10} y_i = \frac{3 422,53}{10} = 342,25,$$

$$V(x) = \frac{1}{10} \sum_{i=1}^{10} x_i^2 - \bar{x}^2 = \frac{966 944,70}{10} - (310,7)^2 = 162,09 \text{ et}$$

$$V(y) = \frac{1}{10} \sum_{i=1}^{10} y_i^2 - \bar{y}^2 = \frac{1 171 938,41}{10} - (342,25)^2 = 56,66.$$

$$Cov(x; y) = \frac{1}{10} \sum_{i=1}^{10} x_i y_i - \bar{x} \times \bar{y} = \frac{1 064 294,54}{10} - 310,7 \times 342,25 = 92,57.$$

On dispose donc de tous les éléments pour calculer des estimations des paramètres a, b, a' et b' :

$$\hat{a} = \text{Cov}(x; y) / V(x) = 92,57 / 162,09 = 0,5711 \text{ et } \hat{b} = \bar{y} - \hat{a}\bar{x} = 342,25 - 0,5711 \times 310,7 = 164,80.$$

$$\hat{a}' = \text{Cov}(x; y) / V(y) = 92,57 / 56,66 = 1,6340 \text{ et } \hat{b}' = \bar{x} - \hat{a}'\bar{y} = 310,7 - 1,6340 \times 342,25 = -248,54.$$

On obtient les droites de régression :  $\begin{cases} D_{Y/x} : y = 0,5711x + 164,80 \\ D_{X/y} : x = 1,6340y - 248,54 \end{cases}$ .

On peut vérifier que ces deux droites sont sécantes au point moyen G.

Si nous validons provisoirement l'existence d'un lien linéaire entre X et Y, les valeurs de x varient dans l'intervalle [293 ; 330,57] et cet intervalle est en toute rigueur l'intervalle de validité du modèle. Si nous relevons une valeur de l'indice du pouvoir d'achat du salaire minimum x = 305, on peut faire une prévision pour y :  $\hat{y} = 0,5711x + 164,80$  soit  $\hat{y} = 0,5711 \times 305 + 164,80 = 338,99$ .

De même, sachant que l'indice du pouvoir d'achat du salaire minimum en 2005 est x = 341,90, il est possible d'utiliser  $D_{Y/x}$  pour faire une « prévision » de l'indice du pouvoir d'achat du salaire moyen en 2005, soit  $\hat{y} = 0,5711 \times 341,9 + 164,8 = 360,06$ . Cependant, la valeur x = 341,90 est hors de l'intervalle de construction du modèle défini par [293 ; 330,57]. C'est pourquoi nous n'avons pas d'information sur la fiabilité de cette prévision (en réalité, la vraie valeur est 351,56).

Les droites de régression peuvent également être construites dans le cas de données contenues dans un tableau de contingence. La détermination de  $a$ ,  $b$ ,  $a'$  et  $b'$ , coefficients des droites de  $Y$  selon  $x$  et de  $X$  selon  $y$ , nécessite les calculs de moyennes, de variances et de la covariance, qui peuvent être effectués à partir des valeurs du tableau de contingence (voir exercice 3).

Deux annexes, proposées en fin de chapitre, sont consacrées à la réalisation d'une droite de régression sous Excel (annexe 6.1), ou tout autre tableur équivalent, et avec une calculatrice graphique, Texas Instrument (annexe 6.2), ou toute autre calculatrice approchante.

## 2.2 LE COEFFICIENT DE CORRÉLATION LINÉAIRE

---

L'intensité de la corrélation est d'autant plus grande que les points du nuage sont plus concentrés au voisinage de la courbe de régression. On voit ainsi l'importance de s'intéresser à la dispersion et à la composition de la variance de  $Y$  dans l'étude de la corrélation.

### Décomposition de la variance

La droite de régression  $D_{Y/x}$  donne pour estimation de  $y$  :  $\hat{y} = ax + b$ .

En remplaçant  $b$  par sa valeur  $b = \bar{y} - a\bar{x}$ ,  $\hat{y} = ax + \bar{y} - a\bar{x}$  soit  $\hat{y} - \bar{y} = a(x - \bar{x})$ .

Calculons la variance des deux membres :  $V(\hat{y} - \bar{y}) = V(a(x - \bar{x}))$ , soit, en utilisant les propriétés de la variance :  $V(\hat{y}) = a^2 V(x - \bar{x}) = a^2 V(x)$ . Or,  $a = \frac{\text{Cov}(x; y)}{V(x)}$ , donc

$$V(\hat{y}) = \left( \frac{\text{Cov}(x; y)}{V(x)} \right)^2 V(x) = \frac{\text{Cov}(x; y)^2}{V(x)}.$$

Reprenons la somme des carrés des erreurs et calculons sa valeur minimale  $S_m$ , en remplaçant  $a$  et  $b$  par leurs valeurs :

$$S_m = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \bar{y} - a(x_i - \bar{x}))^2 = \sum_{i=1}^n (y_i - \bar{y})^2 + a^2 \sum_{i=1}^n (x_i - \bar{x})^2 - 2a \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})$$

soit, en divisant par  $n$  et en remplaçant  $a$  par sa valeur :

$$\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = V(y) + \frac{\text{Cov}(x; y)^2}{V(x)} - 2 \frac{\text{Cov}(x; y)^2}{V(x)} = V(y) - \frac{\text{Cov}(x; y)^2}{V(x)} = V(y) - V(\hat{y})$$

$$\text{soit } V(y) = V(\hat{y}) + \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

La variance totale de  $Y$ ,  $V(y)$ , est la somme de deux termes :

- $V(\hat{y})$ , appelée variance expliquée par la droite de régression. Elle mesure la dispersion de  $y$  quand on résume le nuage à la droite de régression  $D_{Y/x}$  et représente la dispersion le long de la droite de régression ;

- $\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$ , qui est la moyenne des carrés des écarts (comptés parallèlement à l'axe des ordonnées) entre les points du nuage et la droite  $D_{Y/x}$ . Elle représente la variance résiduelle, notée  $V_r(y)$ .

Ainsi, **Variance totale = Variance expliquée + Variance résiduelle**, soit  $V(y) = V(\hat{y}) + V_r(y)$ . Si la variance résiduelle est nulle, cela signifie que tous les points du nuage sont sur la droite de régression, et la variance est entièrement expliquée par la droite de régression.

On pourra utiliser les notations suivantes :

$$\begin{aligned} \text{somme des carrés totaux : } SCT &= \sum_{i=1}^n (y_i - \bar{y})^2, \text{ somme des carrés expliqués :} \\ SCE &= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \quad \text{et somme des carrés résiduels : } SCR = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad \text{avec :} \\ SCT &= SCE + SCR. \end{aligned}$$

### Exemple 6.5

#### L'équation de l'analyse de la variance

Reprendons les données de l'exemple 6.4 et calculons les variances expliquée et résiduelle à l'aide d'Excel (voir figure 6.5).

**Figure 6.5**

**Calcul des variances expliquée et résiduelle.**

	A	B	C	D	E	F
1 Année		$x_i$	$y_i$	$\hat{y}_i$	$(y_i - \hat{y}_i)^2$	$(\hat{y}_i - \bar{y})^2$
2	1994	293,00	329,00	332,15	9,90	102,10
3	1995	296,00	336,00	333,86	4,57	70,36
4	1996	296,00	334,36	333,86	0,24	70,36
5	1997	302,15	337,33	337,37	0,00	23,81
6	1998	311,45	340,34	342,68	5,48	0,19
7	1999	313,93	345,76	344,10	2,76	3,42
8	2000	315,47	347,46	344,98	6,15	7,45
9	2001	321,99	349,17	348,70	0,22	41,61
10	2002	326,41	362,26	361,23	1,05	80,59
11	2003	330,57	360,87	363,60	7,45	128,85
12	<b>Somme</b>	3 106,97	3 422,53	3 422,53	37,81	528,74

On a :  $SCE = 528,74$ ,  $SCR = 37,81$  et  $SCT = SCE + SCR = 566,55$ . La variance résiduelle est  $V_r(y) = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{37,81}{10} = 3,78$  et la variance expliquée par la droite de régression est  $V(\hat{y}) = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \frac{528,74}{10} = 52,87$ , notée aussi  $V_e$ . La variance totale de  $y$  est

$$V(y) = 56,65.$$

De là, l'équation de l'analyse de variance  $V(y) = V(\hat{y}) + V_r(y)$  est vérifiée :  $56,65 = 52,87 + 3,78$ .

On constate avec cet exemple que la variance expliquée représente  $52,87 / 56,65$ , soit environ 93,32 % de la variance totale. Autrement dit : 93,32 % de la variation de  $Y$  est expliquée par la variation de  $X$ . Ce résultat est un bon indicateur de la qualité de la liaison linéaire ; nous y reviendrons à la section suivante.

## Les coefficients de corrélation linéaire et de détermination

La variance expliquée est donnée par la relation  $V(\hat{y}) = \frac{\text{Cov}(x; y)^2}{V(x)}$ . Nous avons vu dans

l'exemple 6.5 qu'il est intéressant d'apprécier la part de la variance expliquée dans la variance totale. Nous allons donc transformer cette relation pour exprimer la variance expliquée en fonction de la variance totale. En multipliant numérateur et dénominateur

de la relation précédente par  $V(y)$ , il vient  $V(\hat{y}) = \frac{\text{Cov}(x; y)^2}{V(x)V(y)}V(y)$ , soit

$\frac{V(\hat{y})}{V(y)} = \frac{\text{Cov}(x; y)^2}{V(x)V(y)}$ . Ce rapport représente la part de variance expliquée sur la variance

totale. Il est appelé coefficient de détermination (noté  $R^2$ ) et amène les définitions ci-après.

### Définition

On appelle **coefficient de corrélation linéaire**, noté  $r$ , entre les variables quantitatives  $X$  et  $Y$ , le nombre sans dimension, défini par :  $r = \frac{\text{Cov}(x; y)}{\sigma(x)\sigma(y)}$ .

$r$  est symétrique par rapport à  $X$  et  $Y$  et est de même signe que la covariance : un coefficient positif (respectivement négatif) indique que  $X$  et  $Y$  varient dans le même sens (respectivement en sens contraire).

La relation entre les variances expliquée et totale s'écrit :  $V(\hat{y}) = r^2V(y)$ , et l'équation de l'analyse de variance  $V(y) = V(\hat{y}) + V_r(y)$  s'écrit alors :  $V(y) = r^2V(y) + V_r(y)$  soit  $V_r(y) = (1 - r^2)V(y)$ . Les variances étant positives, cette relation prouve que la quantité  $1 - r^2$  reste positive ou nulle, c'est-à-dire que :  $-1 \leq r \leq 1$ .

Les coefficients directeurs des droites de régression sont respectivement  $a$  ( $D_{Y/x}$ ) et  $1/a'$  ( $D_{x/y}$ ) et nous pouvons écrire :  $a = \frac{\text{Cov}(x; y)}{V(x)} = r \frac{\sigma(y)}{\sigma(x)}$  et  $\frac{1}{a'} = \frac{1}{r} \times \frac{\sigma(y)}{\sigma(x)}$ . On vérifie que  $a$ ,  $a'$  et  $r$  sont de même signe. Par ailleurs, les droites de régression sont confondues si et seulement si :  $a = 1/a'$  soit  $r^2 = 1$  soit  $r = -1$  ou  $r = 1$ .

### Définition

On appelle **coefficient de détermination**, noté  $R^2$ , le quotient entre la variance expliquée et la variance totale. On a :  $R^2 = \text{SCE} / \text{SCT}$ .

### Propriété

Le coefficient de détermination est le carré du coefficient de corrélation.

Quelques considérations importantes :

- Cet indice est compris entre 0 et 1 et mesure la qualité de l'ajustement de la droite de régression aux points du nuage.
- $R^2$  mesure la part de la variance expliquée par les droites de régression,  $V(\hat{y})$ , rapportée à la variance totale,  $V(y)$  ; ce coefficient de détermination s'exprime souvent en pourcentage.

- Plus  $R^2$  est grand (proche de 1), plus la variance résiduelle (inexpliquée par la droite de régression) est petite ; cela explique qu'il est souhaitable d'avoir un coefficient de détermination proche de 1 si l'on désire utiliser la régression pour faire des prévisions.

- On vérifie par un calcul immédiat :  $R^2 = \frac{Cov^2(x; y)}{V(x)V(y)} = \frac{V(\hat{y})}{V(y)} = \frac{V(\hat{x})}{V(x)}$ , ou encore

$R^2 = r^2 = a \times a'$ . Cette dernière expression permet de retrouver  $r$ , en étant vigilant sur son signe :  $r$  est du signe commun aux deux nombres  $a$  et  $a'$  et on aura donc :  $r = \sqrt{a \times a'}$  si  $a$  et  $a'$  sont positifs et  $r = -\sqrt{a \times a'}$  si  $a$  et  $a'$  sont négatifs.

- En valeur absolue, le coefficient de corrélation est supérieur ou égal au coefficient de détermination. En effet,  $-1 \leq r \leq 1$ , et  $0 \leq r^2 \leq 1$  ; or, la racine carrée d'un nombre compris entre 0 et 1 est supérieure ou égale à ce nombre. On en déduit que  $|r| \geq R^2$ .

### Exemple 6.6

#### Calcul des coefficients de corrélation linéaire et de détermination

Prolongeons l'exemple 6.5 en conclusion duquel nous avions montré que la variance expliquée représente  $52,87 / 56,65$  soit environ 93,32 % de la variance totale, autrement dit que 93,32 % de la variation de  $Y$  est expliquée par la variation de  $X$ . Ce résultat est retrouvé en calculant  $R^2 = \frac{V(\hat{y})}{V(y)} = \frac{52,87}{56,65} = 93,32\%$ .

Ou encore, à partir des résultats de l'exemple 6.4 :

$$R^2 = \frac{Cov^2(x; y)}{V(x) \times V(y)} = \frac{92,57^2}{162,09 \times 56,65} = 93,32\%, \text{ ou } R^2 = a \times a' = 0,5711 \times 1,6340 = 0,9332.$$

Puisque la corrélation est positive,  $a$  et  $a'$  sont positifs et  $r = \sqrt{a \times a'} = \sqrt{0,9332} = 0,9660$ .

#### Interprétation du coefficient de corrélation

Le coefficient de corrélation est toujours compris entre  $-1$  et  $1$  et *a priori* :

- si  $r$  est proche de  $1$  (droites de régression très voisines), la corrélation linéaire entre  $X$  et  $Y$  est positive et forte ;
- si  $r$  est proche de  $-1$  (droites de régression très voisines), la corrélation linéaire entre  $X$  et  $Y$  est négative et forte ;
- si  $r$  est voisin de  $0$  (droites de régression proches de l'orthogonalité), la corrélation linéaire entre  $X$  et  $Y$  est faible.

Quelques mises en garde dans l'interprétation du coefficient de corrélation linéaire doivent être effectuées :

- La corrélation n'est pas une relation de causalité. On a pu mettre en évidence une forte corrélation entre la vente de glaces et la vente de crèmes à bronzer, entre l'augmentation des salaires des enseignants et la consommation d'alcool. Il appartient au spécialiste du domaine d'étude de s'interroger sur un éventuel lien de causalité, à partir de connaissances extérieures au domaine statistique.

- L'absence de corrélation linéaire ne signifie pas l'absence de lien. Il peut exister une liaison fonctionnelle autre que linéaire (parabolique, exponentielle...).
- Le nombre d'observations utilisées pour déterminer le coefficient de corrélation est très important. Le coefficient de corrélation est généralement calculé à partir d'un échantillon de taille  $n$  extrait de la population totale et ne donne qu'une estimation ponctuelle du coefficient de corrélation inconnu, noté  $\rho$ , de la population totale.

## 2.3 TESTS SUR LES ÉLÉMENTS DE LA RÉGRESSION

### Corrélation significativement différente de zéro

Le problème a été évoqué à la fin de la section précédente : le coefficient de corrélation calculé sur un échantillon n'est jamais nul. Nous ne pouvons pourtant pas conclure à l'existence d'un lien linéaire dans tous les cas.

Nous allons donc vérifier l'hypothèse d'un lien linéaire entre les variables à l'aide d'un test statistique après avoir formulé le cadre théorique : supposons que les variables  $x$  et  $y$  suivent une loi normale. En cas d'absence de corrélation linéaire entre ces variables, la

variable  $T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$  est distribuée suivant la loi de Student (voir focus 6.1),  $T_{n-2}$ ,

à  $n - 2$  degrés de liberté,  $n$  désignant le nombre d'observations ; le nombre de degrés de liberté, noté  $v$ , est  $(n - 2)$ , car on a dû estimer les paramètres  $a$  et  $b$  de la droite de régression, leur calcul utilisant deux degrés de liberté.

#### Focus 6.1

#### La loi de Student

La loi de Student est due à William Sealy Gosset (1876-1937), statisticien, employé de la célèbre brasserie Guinness. Student était son pseudonyme. Si  $Z$  et  $X$  désignent deux variables aléatoires indépendantes suivant respectivement la loi normale centrée réduite

et la loi du khi-deux à  $n$  degrés de liberté, la variable aléatoire  $T_n = \sqrt{\frac{Z}{X/n}}$ , appelée le  $t$

de Student, suit la loi de Student à  $n$  degrés de liberté. La courbe représentative de sa densité est symétrique par rapport à l'axe des  $y$  et en forme de cloche comme celle de la loi normale. Cette loi est tabulée en fonction du nombre de degrés de liberté, noté en général  $v$ , et de la probabilité  $\alpha$  ; on note  $t_{\alpha, v}$  la valeur de  $t$  ayant la probabilité  $\alpha$  d'être dépassée. On notera que, dans le cas d'un test bilatéral, pour un seuil de signification de 5 %, on devra prendre  $\alpha / 2 = 2,5\%$ , de façon à avoir :  $P(-t_{\alpha/2; n-2} \leq T_n \leq t_{\alpha/2; n-2}) = 0,95$ .

La loi de Student est très utile pour caractériser la loi de la moyenne empirique d'une distribution normale de variance inconnue. Quand le nombre de degrés de liberté augmente,  $T$  se rapproche de la loi normale centrée réduite.

Posons  $t_{\alpha/2; n-2}$  la valeur de  $T$  donnée par la table de Student telle que  $P(-t_{\alpha/2; n-2} \leq T \leq t_{\alpha/2; n-2}) = 1 - \alpha$  et  $\rho$  le coefficient de corrélation linéaire de la population totale.

Tester l'existence éventuelle d'une corrélation linéaire entre X et Y au sein de la population nécessite de passer par les étapes suivantes :

1. Formuler les hypothèses à tester :
  - $H_0 : \rho = 0$  (absence de corrélation linéaire) ;
  - $H_1 : \rho \neq 0$  (présence de corrélation linéaire).
2. Déterminer le degré de liberté :  $n - 2$ .
3. Définir la règle de décision du test à partir de la valeur  $t_{\alpha/2; n-2}$  dépendant du seuil de signification  $\alpha$  et du degré de liberté :
  - Si  $T \geq t_{\alpha/2; n-2}$  ou si  $T \leq -t_{\alpha/2; n-2}$ , l'hypothèse  $H_0$  est rejetée et l'hypothèse  $H_1$  est acceptée : il y a une corrélation linéaire significative entre les variables.
  - Si  $-t_{\alpha/2; n-2} \leq T \leq t_{\alpha/2; n-2}$ , l'hypothèse  $H_0$  n'est pas rejetée : il est impossible de conclure de façon significative à l'existence d'une corrélation linéaire entre les variables.

### Exemple 6.7

#### Test du coefficient de corrélation linéaire

Reprendons les données de l'exemple 6.4.

$n = 10$  et la droite de régression nécessite d'estimer deux paramètres. Donc le degré de liberté est  $10 - 2 = 8$ .

Par ailleurs, à partir de ces mêmes données, nous avons calculé  $r = 0,9660$  (voir exemple 6.6). Nous noterons  $t_c$  la valeur de  $t$  calculée sur l'échantillon. On a

$$t_c = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0,9660\sqrt{8}}{\sqrt{1-0,9660^2}} = 10,57$$

et la table de Student donne  $t_{0,025; 8} = 2,3060$ .

Puisque  $10,57 \geq 2,3060$ , soit  $t_c \geq t_{\alpha/2; n-2}$ , il faut rejeter l'hypothèse  $H_0$ . Il y a donc une corrélation linéaire significative entre x et y.

#### Test de Student sur la pente $a$ de la droite de régression

Quittons l'approche descriptive pour adopter le point de vue de la statistique inférentielle : le problème est similaire à celui évoqué pour le coefficient de corrélation linéaire. Nous supposons que nous avons déterminé l'équation de la droite de régression de Y selon x et nous noterons cette équation :  $\hat{y} = \hat{a}x + \hat{b}$ , pour ne pas oublier que les coefficients de cette droite sont des coefficients empiriques calculés sur notre échantillon et qu'ils constituent des estimations ponctuelles des coefficients  $a$  et  $b$  inconnus dans la population.

On se place dans l'hypothèse où la distribution des  $y$  est normale et où la variance de Y est constante pour toute valeur de X. On démontre que l'écart-type de  $\hat{a}$ , noté  $\sigma(\hat{a})$ , est

estimé par :  $\sigma^2(\hat{a}) = \frac{SCR}{(n-2) \sum_{i=1}^n (x_i - \bar{x})^2}$  où  $SCR = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ ; le nombre noté

$S^2_{Y/x} = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2$  représente un estimateur de la variance résiduelle.

L'intervalle de confiance de  $a$  est donné par :  $a \pm t_{\frac{\alpha}{2}; n-2} \sigma(\hat{a})$ .

Tester l'hypothèse  $H_0 : a = 0$  revient à tester le parallélisme de la droite de régression de  $Y$  selon  $x$  avec l'axe des  $x$  et donc à tester la nullité du coefficient de corrélation.

1. Les hypothèses à tester :

- $H_0 : a = 0$  (absence de corrélation linéaire);
- $H_1 : a \neq 0$  (présence de corrélation linéaire).

2. Déterminer le degré de liberté :  $n - 2$ .

3. Définir la règle de décision du test à partir de la valeur  $t_{\alpha/2; n-2}$  dépendant du seuil de signification  $\alpha$  et du degré de liberté.

- Si  $t \geq t_{\alpha/2; n-2}$  ou si  $t \leq -t_{\alpha/2; n-2}$ , l'hypothèse  $H_0$  est rejetée en faveur de l'hypothèse alternative  $H_1 : a \neq 0$ .
- Si  $-t_{\alpha/2; n-2} \leq t \leq t_{\alpha/2; n-2}$ , l'hypothèse  $H_0$  n'est pas rejetée.

### Exemple 6.8

#### Test de student sur le paramètre $a$ (pente de $D_{Y/x}$ )

Si l'on reprend l'exemple 6.5, on a :  $S^2_{Y/x} = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{37,81}{8} = 4,7263$  et

$$S_{Y/x} = \sqrt{\frac{37,81}{8}} = 2,174, \text{ ce qui donne : } \sigma^2(\hat{a}) = \frac{S^2_{Y/x}}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{4,7263}{1620,9} \text{ et}$$

$$\sigma(\hat{a}) = \sqrt{\frac{4,7263}{1620,9}} = 0,0029. \text{ On calcule alors } t_c = \frac{\hat{a}}{\sigma(\hat{a})} = \frac{0,5711}{0,0029} = 196,93 \text{ et}$$

$t_{0,025; 8} = 2,3060$ , ce qui donne pour intervalle de confiance pour  $a$ , au seuil de signification de 5 % :  $0,5711 \pm 2,3060 \times 0,0029$ , soit  $[0,5644 ; 0,5778]$ .

$t_c > t_{0,025; 8}$ , donc on doit rejeter l'hypothèse  $H_0$  et conclure à l'existence d'une relation linéaire entre  $X$  et  $Y$ . Si on utilise l'intervalle de confiance, on aura la même conclusion, car il ne recouvre pas la valeur 0, ce qui signifie qu'au niveau de confiance 95 %  $a$  est différent de 0.

## Test de Student sur l'ordonnée à l'origine b de la droite de régression

On peut effectuer la même démarche pour le coefficient b et déterminer un intervalle de confiance pour ce paramètre, et tester l'hypothèse d'une droite de régression passant par l'origine ( $b = 0$ ).

$$\text{Avec les mêmes notations que précédemment, on obtient : } \sigma^2(\hat{b}) = \frac{\sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2} S_{Y/x}^2.$$

## Test de Fisher sur la pente a de la droite de régression

La seconde approche pour tester une régression linéaire passe par l'étude de la part de la variance expliquée dans la variance totale<sup>1</sup>. On démontre que la variable aléatoire

$$F = \frac{\frac{SCE}{1}}{\frac{SCR}{n-2}} \text{ suit la loi de Fisher avec 1 et } (n-2) \text{ degrés de liberté, notée } F(1; n-2).$$

Le nombre de degrés de liberté de la variance expliquée est de 1 et celui de la variance résiduelle de  $(n-2)$ , celui de la variance totale de  $(n-1)$ .

Les hypothèses à tester sont :

- $H_0 : SCE = SCR / (n-2)$ ;
- $H_1 : SCE > SCR / (n-2)$ .

On rejette  $H_0$  au seuil de signification  $\alpha$  si  $F_c > F_{(\alpha; 1, n-2)}$ ,  $F_c$  étant le F calculé et  $F_{(\alpha; 1, n-2)}$  le F théorique (lu dans la table ; voir focus 6.2).

On notera que

$$F = \frac{(n-2)SCE}{SCR} = (n-2) \times \frac{SCE}{SCT - SCE} = (n-2) \times \frac{SCE/SCT}{1 - SCE/SCT} = (n-2) \times \frac{r^2}{1 - r^2} = t_c^2.$$

### Focus 6.2

### La loi de Fisher

La comparaison de deux populations normales peut porter sur leurs variances. Pour tester l'hypothèse d'égalité de deux variances, on utilise la distribution du quotient de deux variances, appelée distribution de Fisher ou de Fisher-Snedecor.

Si  $\chi_1^2$  et  $\chi_2^2$  sont deux variables aléatoires indépendantes, suivant chacune la loi du Khi-deux avec respectivement  $v_1$  et  $v_2$  pour degrés de liberté, la variable aléatoire  $F = (\chi_1^2 / v_1) / (\chi_2^2 / v_2)$  suit la loi de Fisher à  $v_1$  et  $v_2$  degrés de liberté. Cette loi est dissymétrique et tend vers la loi normale à mesure que les degrés de liberté augmentent. Cette loi est tabulée, ses valeurs dépendant du seuil de signification  $\alpha$  et des degrés de liberté, et on a :  $P(F > F_{(\alpha; v_1, v_2)}) = \alpha$ .

1. Voir P. Roger.

**Exemple 6.9****Test de Fisher**

Reprendons l'exemple 6.5 et calculons  $F_c$ , le F calculé :  $F_c = \frac{SCE/1}{SCR/(n-2)} = \frac{528,74}{37,81/8} = 111,87$  ;

par ailleurs, le F de la table est :  $F_{(0,05; 1, 8)} = 5,32$ .  $F_c > F_{(0,05; 1, 8)}$ , donc  $H_0$  est rejetée et on conclut à l'existence d'une relation linéaire (tester  $H_0$  revient à tester  $a = 0$ ).

## 3 Ajustements et absence de linéarité

### 3.1 AJUSTEMENT LINÉAIRE PAR CHANGEMENT DE VARIABLE

Dans certains cas où, clairement, les points ne sont pas alignés, le graphique représentant le nuage de points permet de rejeter directement l'hypothèse d'une corrélation linéaire. Il est alors possible de revenir à la théorie de la corrélation linéaire en utilisant un changement de variable, afin de déterminer la relation fonctionnelle qui lie les deux variables.

Par exemple :

- Soit la relation non linéaire  $y = a \ln x + b$ . En posant  $X = \ln(x)$ , cette relation non linéaire est équivalente à la relation linéaire  $y = aX + b$ .
- Soit la relation non linéaire  $y = a \exp x + b$ . En posant  $X = \exp x$ , cette relation non linéaire est équivalente à la relation linéaire  $y = aX + b$ .
- Soit la relation non linéaire  $y = b \times a^x$ . En prenant le logarithme de cette expression,  $\ln y = \ln(b \times a^x) = \ln b + x \ln a$ , soit avec  $B = \ln b$  ;  $A = \ln a$  ;  $Y = \ln y$ , cette relation non linéaire est équivalente à la relation linéaire  $Y = Ax + B$ .
- Soit la relation non linéaire  $y = b \times x^a$ . En prenant le logarithme de cette expression,  $\ln y = \ln(b \times x^a) = \ln b + a \ln x$ , soit avec  $B = \ln b$  ;  $Y = \ln y$  ;  $X = \ln x$ , cette relation non linéaire est équivalente à la relation linéaire  $Y = aX + B$ .
- Modèle logistique : ce modèle est défini par  $y = k/(1 + a \exp(-bx))$  et peut être ramené à un modèle linéaire. Ce modèle a été introduit par Pierre François Verhulst (1804-1849), élève de Quetelet, lors de l'étude de l'évolution d'une population qui croît exponentiellement au début puis se stabilise, freinée par un phénomène de surpopulation (saturation), pour tendre vers sa capacité maximale. Ce modèle est utilisé notamment pour le traitement des séries chronologiques (voir chapitre 7).

## 3.2 COEFFICIENT DE CORRÉLATION DES RANGS

Certaines grandeurs ne sont pas mesurables, ou n'ont pu être mesurées, mais peuvent être classées. Il s'agit de variables ordinaires. Dans ce cas, le calcul du coefficient de corrélation linéaire, réservé aux variables quantitatives, est alors inapplicable. Pour autant, il peut être intéressant de calculer la corrélation entre deux variables ordinaires. Il convient alors de trouver un coefficient de corrélation, non pas entre les valeurs prises par les deux variables, mais entre les rangs de ces valeurs.

On doit à Charles Spearman, psychologue anglais (1863-1945), le coefficient de corrélation des rangs, qui permet de comparer la concordance du classement de deux variables et de mesurer leur degré de dépendance.

### Définition

Soit deux caractères X et Y. Soit  $d_i$  la différence des rangs de l'observation i pour les deux variables. On appelle **coefficient de corrélation des rangs** (coefficient de Spearman), noté  $r_s$ ,

$$\text{entre les variables } X \text{ et } Y, \text{ le nombre défini par : } r_s = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}.$$

Soit  $R(x_i)$  le rang de la modalité  $x_i$  et  $R(y_i)$  le rang de la modalité  $y_i$ .  $d_i = R(x_i) - R(y_i)$ . Le coefficient de Spearman est le coefficient de corrélation linéaire de la série bivariée ( $R(x_i)$  ;  $R(y_i)$ ). La simplicité de la formule donnée dans la définition vient du fait que  $R(x_i)$  et  $R(y_i)$  prennent les valeurs entières de 1 à n. Par définition, ce coefficient est compris entre -1 et 1 et constitue un outil précieux pour détecter une liaison. Il a l'avantage de ne pas être influencé par des valeurs aberrantes et de ne pas être tributaire de l'allure de la liaison éventuelle (linéaire, exponentielle, etc.).

## Résumé

Lors de l'étude du lien entre deux variables, la notion de corrélation est extrêmement importante. Il importe de dominer la technique de la méthode MCO, de connaître les formules, de savoir utiliser efficacement une calculatrice statistique et de rester prudent dans les interprétations.

Le lecteur doit, à l'issue de ce chapitre, pouvoir mener à bien les calculs de l'analyse de la variance.

Par ailleurs, il doit maîtriser les différents tests et la lecture des tables.

Dans le chapitre suivant nous aborderons les séries chronologiques, qui sont des séries bivariées dont une des variables est le temps. Pour analyser la tendance de ces séries, nous utiliserons les résultats incontournables de ce chapitre.

# Problèmes et exercices

L'analyse de régression fournit une seconde approche des séries bivariées, qui autorise l'approfondissement des liaisons étudiées au sein des tableaux de contingence.

- Les exercices 1 et 2 proposent l'application des calculs indispensables à la détermination d'une équation de régression linéaire incluant l'étude de la qualité de la régression et la réalisation de prévisions.
- L'exercice 3 met en œuvre ces mêmes calculs à partir de données présentées sous la forme d'un tableau de contingence.
- Les exercices 4 et 5 abordent respectivement les analyses de régression et de corrélation dans le cas de séries liées par une relation non linéaire.



## EXERCICE 1 RÉGRESSION LINÉAIRE ET INDICATEURS DE QUALITÉ

### Énoncé

Les données régionales de l'accidentologie 2005, transmises par la Sécurité routière, sont les suivantes (hors régions PACA et Île-de-France) :

Région	Nombre d'accidents corporels	Nombre de tués
Alsace	2 085	114
Aquitaine	4 523	333
Auvergne	1 817	141
Basse-Normandie	1 518	144
Bourgogne	2 065	208
Bretagne	2949	252
Centre	2 859	307
Champagne-Ardenne	1 512	168
Corse	845	35
Franche-Comté	1 224	147
Haute-Normandie	1 754	154
Languedoc-Roussillon	3 305	319
Limousin	1 124	82
Lorraine	2 672	213
Midi-Pyrénées	3 610	330
Nord-Pas-de-Calais	3 817	255
Pays de la Loire	3 778	314

Région	Nombre d'accidents corporels	Nombre de tués
Picardie	1 919	194
Poitou-Charentes	1 984	221
Rhône-Alpes	6 957	469

Source : ONISR, 2006

On note respectivement X et Y les variables « nombre d'accidents corporels » et « nombre de tués ».

1. Dessinez le nuage de points représentant cette série.
2. Établissez l'équation de la droite de régression de Y selon x, qui permet d'expliquer le nombre de tués par le nombre d'accidents corporels.
3. Donnez l'équation de l'analyse de la variance.
4. Calculez :
  - a. le coefficient de corrélation linéaire ;
  - b. le coefficient de détermination.
5. Calculez :
  - a. l'écart-type du coefficient a ;
  - b. l'écart-type du coefficient b.
6. Effectuez les tests :
  - a. de signification du coefficient de corrélation linéaire ;
  - b. de Student sur les coefficients a et b ;
  - c. de Fisher.

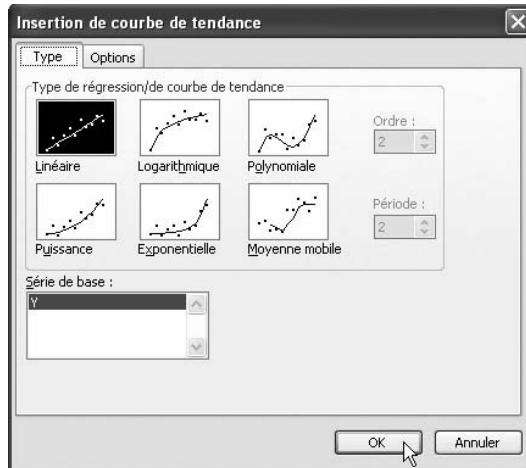
### Solution

1. Pour représenter le nuage de points sous Excel, cliquez sur Insertion/Graphique dans la barre de menus, puis, dans l'assistant graphique, choisissez le type de graphique Nuage de points, puis, dans Sous-type de graphique, sélectionnez l'image « Nuage de points. Compare des paires de valeurs ». Cliquez sur Suivant et indiquez dans le champ correspondant la plage où se trouvent les données (voir chapitre 1, exercice 5).

La droite de régression de Y selon X peut être ajoutée au nuage de points. Pour cela, une fois le nuage de points effectué, sélectionnez tous les points du graphique en cliquant sur l'un d'entre eux, puis cliquez sur le bouton droit de la souris et sélectionnez « Ajouter une courbe de tendance... ». La boîte de dialogue de la figure 6.6 apparaît :

**Figure 6.6**

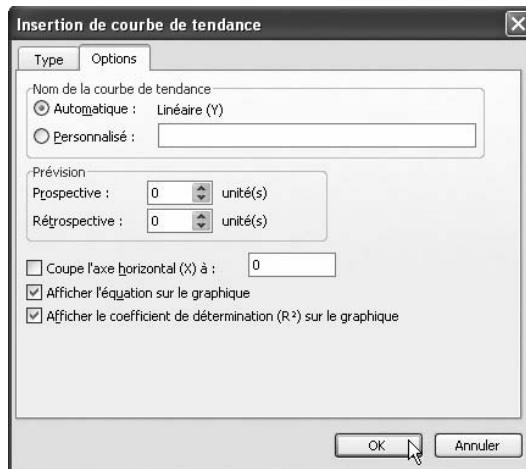
Ajout d'une courbe de tendance à un nuage de points.



Sélectionnez Linéaire, puis cliquez sur l'onglet Option (voir figure 6.7).

**Figure 6.7**

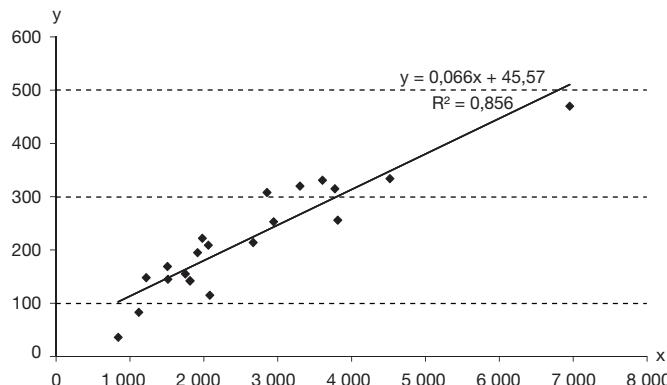
Affichage de l'équation de régression et du  $R^2$  sur un nuage de points.



Cochez les cases Afficher l'équation sur le graphique et Afficher le coefficient de détermination ( $R^2$ ) sur le graphique, puis cliquez sur OK (voir figure 6.8).

**Figure 6.8**

Nuage de points entre X et Y, avec courbe de tendance sous Excel.



L'équation de la droite de régression de Y selon x est indiquée, ainsi que la valeur du R<sup>2</sup>. Nous allons retrouver ces résultats en répondant aux questions suivantes.

**2.** La droite de régression qui permet d'expliquer le nombre de tués par le nombre d'accidents corporels correspond à la droite de régression de Y selon x. Pour établir l'équation de cette droite de régression, il convient de déterminer les valeurs de a et b dans l'équation  $y = ax + b$ .

Pour cela, il est nécessaire de calculer les valeurs de  $\bar{x}$ ,  $\bar{y}$ , V(x) et  $\sum_{i=1}^n x_i y_i$ .

Les moyennes de X et de Y ainsi que la variance de X peuvent être calculées en utilisant les fonctions d'Excel correspondantes, puisque les données sont des données brutes, avec  $n_i = 1$  quel que soit i. Pour cela, il convient d'appeler les fonctions MOYENNE et VAR.P d'Excel (voir annexe 1.1), ou bien d'effectuer les calculs comme exposé précédemment (voir chapitres 2 et 3). Les résultats de ces calculs sont indiqués figure 6.9.

Figure 6.9

Résultats sous Excel.

	A Région	B X	C Y	D XY
1				
2	ALSACE	2 085	114	237 690
3	AQUITAINE	4 523	333	1 506 159
4	AUVERGNE	1 817	141	256 197
5	BASSE-NORMANDIE	1 518	144	218 592
6	BOURGOGNE	2 065	208	429 520
7	BRETAGNE	2 949	252	743 148
8	CENTRE	2 859	307	877 713
9	CHAMPAGNE-ARDENNE	1 512	168	254 016
10	CORSE	845	35	29 575
11	FRANCHE-COMTE	1 224	147	179 928
12	HAUTE-NORMANDIE	1 754	154	270 116
13	LANGUEDOC-ROUSSILLON	3 305	319	1 064 295
14	LIMOUSIN	1 124	82	92 168
15	LORRAINE	2 672	213	569 136
16	MIDI-PYRENEES	3 610	300	1 191 300
17	NORD-PAS DE CALAIS	3 817	255	973 335
18	PAYS-DE-LA-LOIRE	3 778	314	1 186 292
19	PICARDIE	1 919	194	372 286
20	POITOU-CHARENTES	1 984	221	438 464
21	RHONE-ALPES	6 957	469	3 262 833
22	Somme	52 317,00	4 400,00	14 142 763
23	Moyenne	2 615,85	220,00	
24	Variance	1 974 311,73		

$$\text{De là, } \hat{a} = \frac{\sum_{i=1}^{20} x_i y_i - n\bar{y} \cdot \bar{x}}{nV(x)} = \frac{14\,142\,763 - 20 \times 2\,615,85 \times 220}{20 \times 1\,974\,311,73} = 0,0667 \text{ et}$$

$$\hat{b} = \bar{y} - \hat{a}\bar{x} = 220 - 0,0667 \times 2\,615,85 = 45,57.$$

L'équation de la droite de régression de Y selon x est donc :  $y = 0,0667x + 45,57$ . Ce résultat est conforme à l'équation de la courbe de tendance linéaire proposée par l'assistant graphique d'Excel (voir question 1).

**3.** Afin de donner l'équation de l'analyse de la variance, il convient de calculer la somme des carrés totaux (SCT), la somme des carrés expliqués (SCE) et la somme des carrés résiduels (SCR).

Le calcul de la somme des carrés expliqués (SCE) nécessite au préalable le calcul de la valeur de Y estimée par la droite de régression, telle que  $\hat{y}_i = 0,0667x_i + 45,57$ . Ces

calculs sont effectués à la suite du tableau précédent (voir figure 6.9) et les résultats de ces calculs sont indiqués figure 6.10.

**Figure 6.10**

Résultats sous Excel.

	E	F	G	H	I
1	$\hat{Y}$	$(Y - \bar{Y})^2$	$(\hat{Y} - \bar{Y})^2$	$(Y - \hat{Y})^2$	$(X - \bar{X})^2$
2	184,60	11 236,00	1 253,03	4 984,62	281 801,72
3	347,17	12 769,00	16 172,89	200,86	3 637 221,12
4	166,73	6 241,00	2 837,58	662,09	638 161,32
5	146,79	5 776,00	5 359,25	7,80	1 205 274,62
6	183,27	144,00	1 349,23	611,66	303 435,72
7	242,22	1 024,00	493,51	95,74	110 988,92
8	236,21	7 569,00	262,89	5 010,69	59 121,92
9	146,39	2 704,00	5 417,99	466,86	1 218 484,82
10	101,92	34 225,00	13 943,81	4 477,76	3 135 909,72
11	127,19	5 329,00	8 613,96	392,49	1 937 246,42
12	162,53	4 356,00	3 302,79	72,76	742 785,42
13	265,95	9 801,00	2 111,76	2 813,89	474 927,72
14	120,52	19 044,00	9 896,19	1 483,82	2 225 616,42
15	223,74	49,00	14,02	115,44	3 152,82
16	286,29	12 100,00	4 394,62	1 910,39	988 334,22
17	300,10	1 225,00	6 415,23	2 033,57	1 442 761,32
18	297,49	8 836,00	6 005,40	272,43	1 350 592,62
19	173,53	676,00	2 159,22	418,91	485 599,92
20	177,87	1,00	1 775,19	1 860,46	399 234,42
21	509,48	62 001,00	83 796,80	1 638,37	18 845 583,32
22	4 400,00	205 106,00	175 575,37	29 530,63	39 486 234,55

De là,  $SCT = 205 106$  ;  $SCE = 175 575$  et  $SCR = 29 531$ . L'équation de l'analyse de variance  $SCT = SCE + SCR$  est vérifiée, puisque  $205 106 = 175 575 + 29 531$ .

**a.** Le calcul du coefficient de corrélation linéaire nécessite de calculer la covariance entre X et Y et les écarts-types de X et de Y.

$$COV(x; y) = \frac{14 142 763}{20} - 2 615,85 \times 220, \text{ soit } Cov(x; y) = 131 651,15.$$

L'écart-type de X est la racine de  $V(X)$ , calculée précédemment, soit  $\sqrt{1 974 311,73} = 1 405,1$ . D'où  $\sigma_x = 1 405,1$ .

De même, l'écart-type de Y est la racine de  $V(Y)$ .  $V(Y)$  est calculé en utilisant la fonction VAR.P d'Excel (voir annexe 1.1) ou la méthode exposée précédemment (voir chapitres 2 et 3). On trouve :  $V(Y) = 10 255,30$ , soit  $\sqrt{10 255,30} = 101,27$ . D'où  $\sigma_y = 101,27$ .

On obtient alors :  $r = \frac{Cov(x; y)}{\sigma_x \sigma_y} = \frac{131 651,15}{1 405,1 \times 101,27}$ , soit  $r = 0,925$ . Il existe *a priori* une

forte corrélation linéaire positive entre X et Y, la droite de régression calculée est une bonne représentation du nuage de points.

**b.** Le coefficient de détermination est le carré du coefficient de corrélation linéaire, donc  $R^2 = 0,925^2$ , soit  $R^2 = 0,856$ .

$R^2$  représente la part de variabilité expliquée sur la variabilité totale, on vérifie que :

$$\frac{SCE}{SCT} = \frac{175 575}{205 106} = 0,856 = R^2.$$

**5. a.** Le calcul de l'écart-type de  $\hat{a}$ ,  $\sigma_a$ , nécessite le calcul de  $\sum_{i=1}^n (x_i - \bar{x})^2$ . Ces calculs sont effectués à la suite du tableau précédent (voir figure 6.10).

De là,  $\sigma_a^2 = \frac{1}{18} \times \frac{29\,530,63}{39\,486\,235,55}$ , soit  $\sigma_a^2 = 0,0000415$ ; d'où  $\sigma_a = 0,00645$ .

**b.** À partir des calculs précédents,  $\sigma_b^2 = 0,0000415 \times \left( \frac{39\,486\,235,55}{20} + 2\,615,85^2 \right)$ , soit  $\sigma_b^2 = 366,33$ ; d'où  $\sigma_b = 19,14$ .

**6. a.** À la suite des calculs précédents,  $T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0,925\sqrt{20-2}}{\sqrt{1-0,856^2}} = 7,593$  et la table de

Student donne  $t_{0,025; 18} = 2,445$ . Puisque  $7,593 \geq 2,445$ , soit  $T \geq t_{\alpha/2; n-2}$ , il faut rejeter l'hypothèse  $H_0$ . Il y a donc une corrélation linéaire significative entre le nombre d'accidents corporels et le nombre de tués.

Le  $t_{\text{théorique}}$  peut s'obtenir à l'aide d'Excel en appelant la fonction statistique LOI.STUDENT.INVERSE et en saisissant les arguments suivants : Probabilité = 0,025 et Degrés\_liberté = 18. Cette fonction est similaire dans son utilisation à celle rencontrée pour la lecture de la table de la loi du khi-deux dans l'exercice 5 du chapitre 5.

La probabilité de Student associée peut s'obtenir à l'aide d'Excel en appelant la fonction statistique LOI.STUDENT et en saisissant les arguments suivants : X = 7,593 (le Student calculé), Degrés\_liberté = 18 et Uni/bilatéral = 1.

**b.** À la suite des calculs précédents,  $t_a = \frac{\hat{a}}{\sigma_{\hat{a}}} = \frac{0,0667}{0,00645}$ , soit  $t_a = 10,345$  et

$$t_b = \frac{\hat{b}}{\sigma_{\hat{b}}} = \frac{45,57}{19,14}, \text{ soit } t_b = 2,381.$$

$t_a$  et  $t_b$  sont tous deux supérieurs au  $t_{\text{théorique}} = t_{(0,025; 8)} = 2,101$  obtenu par lecture de la table de Student, avec une probabilité de 0,05 ( $\alpha = 5\%$ ) et  $n - 2 = 18$  degrés de liberté. De plus, toujours par lecture de la table statistique, la probabilité associée à  $t_a$  ( $p = 0,000$ ) et celle associée à  $t_b$  ( $p = 0,029$ ) sont toutes deux inférieures à 5 %. (Pour un rappel sur les tests d'hypothèses, voir focus 5.1.)

Le test de Student pour le coefficient a de la régression permet de conclure que la valeur de a est significativement différente de 0. De même, le test de Student pour le coefficient b de la régression permet de conclure que la valeur de b est significativement différente de 0.

La probabilité de Student associée peut s'obtenir à l'aide d'Excel en appelant la fonction statistique LOI.STUDENT et en saisissant les arguments suivants : X = 10,345 pour a et X = 2,381 pour b (le Student calculé), Degrés\_liberté = 18 et Uni/bilatéral = 2.

**c.** À la suite des calculs précédents,  $F_c = \frac{SCE}{SCR} = \frac{175\,575,37}{29\,230,63} = 107,02$ , soit  $F_c = 107,02$ .

$F_c$  est supérieur au  $F_{(0,05; 1, 18)} = 4,414$  obtenu par lecture de la table de Fisher, avec une probabilité de 0,05 ( $\alpha = 5\%$ ),  $ddl_1 = 1$  et  $ddl_2 = n - 2 = 18$  degrés de liberté. On trouve donc  $F_c > F_{(0,05; 1, 18)}$ . On rejette donc  $H_0$  au seuil de signification 5 % et l'on conclut à l'existence d'une relation linéaire entre X et Y.

Le  $F(1; 18)_{\text{théorique}}$  est disponible sous Excel en appelant la fonction statistique INVERSE.LOI.F et en saisissant les arguments suivants : Probabilité = 0,05, Degrés.liberté1 = 1 et Degrés.liberté2 = 18.

La probabilité de Fisher associée peut s'obtenir à l'aide d'Excel en appelant la fonction statistique LOI.F et en saisissant les arguments suivants : X = 107,02 (le Fisher calculé), Degrés.liberté1 = 1 et Degrés.liberté2 = 18.



## EXERCICE 2 RÉGRESSION LINÉAIRE ET PRÉVISIONS

### Énoncé

Au cours des années 2000, le produit intérieur brut (PIB) et la consommation effective des ménages ont été les suivants (en milliards d'euros) :

Année	Consommation	PIB
2000	1 009,6	1 441,4
2001	1 053,9	1 497,2
2002	1 098,2	1 548,6
2003	1 145,5	1 594,8
2004	1 194,9	1 660,2
2005	1 243,6	1 717,9
2006	1 292,5	1 792,0

Source : Comptes nationaux - Base 2000, Insee

1. En utilisant la méthode des moindres carrés ordinaires, établissez l'équation de la droite de régression  $y = ax + b$  qui permet d'expliquer le PIB en fonction de la consommation.
2. Calculez les indicateurs de qualité de la régression :
  - a. le coefficient de détermination et le test associé ;
  - b. les tests de Student ;
  - c. le test de Fisher.
3. En stimulant la consommation pour lui permettre d'atteindre 1 400 milliards d'euros, à quel niveau de PIB peut s'attendre le gouvernement ?
4. En utilisant la méthode des moindres carrés ordinaires, établissez l'équation de la droite de régression  $x = a'y + b'$  qui permet d'expliquer la consommation en fonction du PIB.
5. Estimez la consommation correspondant à un PIB de 1 600 milliards d'euros.

**Solution**

1. Expliquer le PIB en fonction de la consommation des ménages selon la droite de régression  $y = ax + b$  nécessite de poser X = « consommation » et Y = « PIB ».

Pour établir l'équation de la droite de régression  $y = ax + b$ , il convient de déterminer les valeurs de a et b dans l'équation. Pour cela, il est nécessaire de calculer les valeurs de  $\bar{x}$ ,

$$\bar{y}, V(x) \text{ et } \sum_{i=1}^n x_i y_i.$$

Saisissez les valeurs de X, la consommation, dans la colonne L1 et celles de Y, le PIB, dans la colonne L2, comme indiqué figure 6.11.

Pour obtenir les calculs intermédiaires nécessaires, appuyez sur la touche **STAT**, puis choisissez le menu **CALC** et sélectionnez la fonction **2-Var Stats**. Puis appuyez sur **ENTER**. Tapez **2-Var Stats L1,L2** puis appuyez à nouveau sur **ENTER**. Les résultats de statistiques sur les variables X et Y, respectivement contenues dans L1 et L2, s'inscrivent (voir figure 6.12).

**Figure 6.11 (gauche)**

Saisie du tableau de données avec la calculatrice.

L1	L2	L3	z
1009,6	1441,4	-----	
1053,9	1497,2		
1098,2	1548,6		
1145,5	1594,8		
1194,9	1660,2		
1243,6	1712,9		
1292,5	1792,2		

**Figure 6.12 (droite)**

Statistiques sur L2(Y).

2-Var Stats  
 $\bar{x}=1607,442857$   
 $\Sigma y=11252,1$   
 $\Sigma y^2=18179499,3$   
 $S_y=124,0910131$   
 $\sigma_y=114,8859541$   
 $\downarrow \Sigma xy=12996965,9$

$$L2(7) = 1792$$

$$\text{De là, } \hat{a} = \frac{\sum_{i=1}^7 x_i y_i - n\bar{y}\bar{x}}{nV(x)} = \frac{12996965,9 - 7 \times 1148,31 \times 1607,44}{7 \times 94,653^2} = 1,212 \text{ et}$$

$$\hat{b} = \bar{y} - \hat{a}\bar{x} = 1607,44 - 1,212 \times 1148,31 = 215,52.$$

D'où l'équation de régression de Y selon x :  $y = 1,212 X + 215,52$ .

2. a. Le calcul du coefficient de corrélation linéaire nécessite de calculer la covariance entre X et Y en plus des écarts-types de x et de y, déjà connus.

$$COV(x; y) = \frac{12996965,5}{7} - 1148,31 \times 1607,44, \text{ soit } COV(x; y) = 10859,81.$$

De là,  $r = \frac{10859,81}{94,653 \times 114,886}$ , soit  $r = 0,999$ , soit  $R^2 = 0,997$ . Il existe une forte corrélation linéaire positive entre X et Y.

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0,999\sqrt{7-2}}{\sqrt{1-0,997^2}} = 30,633 \text{ et la table de Student donne } t_{0,025; 5} = 3,163. \text{ Puisque}$$

$30,633 \geq 3,163$ , soit  $t \geq t_{\alpha/2; n-2}$ , il faut rejeter l'hypothèse  $H_0$ . Il y a donc une corrélation linéaire hautement significative entre X et Y.

**b.** Afin de réaliser les tests de Student, les variables suivantes sont calculées (voir figure 6.13) :

- En L4 sont calculés les  $y$  estimés, notés  $\hat{y}_i$ . Pour cela, placez le curseur sur l'en-tête de colonne L4, indiquez L4=1,212\*L1+215,52, puis appuyez sur **ENTER**.
- En L5 sont calculés les  $(y - \hat{y}_i)^2$ . Pour cela, placez le curseur sur l'en-tête de colonne L5, indiquez L5=(L4-L2)^2, puis appuyez sur **ENTER**.
- En L6 sont calculés les  $(x_i - \bar{x})^2$ . Pour cela, placez le curseur sur l'en-tête de colonne L6, indiquez L6=(L1-1148,31)^2, puis appuyez sur **ENTER**.

Pour obtenir les calculs intermédiaires nécessaires à partir des variables nouvellement créées, appuyez sur la touche **STAT**, puis choisissez le menu CALC et sélectionnez la fonction 2:2-Var Stats. Puis appuyez sur **ENTER**. Tapez 2-Var Stats L5,L6 puis appuyez à nouveau sur **ENTER**. Les statistiques sur les variables  $(y - \hat{y}_i)^2$  et  $(x_i - \bar{x})^2$ , respectivement contenues dans L5 et L6, s'inscrivent (voir figure 6.14).

Figure 6.13 (gauche)

Calculs dans L4, L5 et L6 avec la calculatrice.

L4	L5	L6	
1439,2	5,0391	48240	
1492,8	18,95	8913,2	
1546,5	4,2502	2511	
1603,9	82,192	7,8961	
1663,7	12,523	2170,6	
1722,8	23,651	9080,2	
1782	99,401	20791	

Figure 6.14 (droite)

Statistiques sur L6,  
 $(x_i - \bar{x})^2$ .

2-Var Stats
$\bar{x}=8959,169814$
$\sum y=62714,1887$
$\sum y^2=975363603$
$S_y=8301,571509$
$\sigma_y=7685,761763$
$\sum xy=2585741,32$

$$L6(1)=19240,4641$$

De là,  $\sigma_a^2 = \frac{1}{5} \times \frac{246,006}{62\,714,189}$ , soit  $\sigma_a^2 = 0,0007839$ ; d'où  $\sigma_a = 0,028$ .

$$\sigma_b^2 = 0,0007839 \times \left( \frac{62\,714,189}{7} + 1148,31^2 \right), \text{ soit } \sigma_b^2 = 1040,71; \text{ d'où } \sigma_b = 32,26.$$

$$t_a = \frac{\hat{a}}{\sigma_a} = \frac{1,212}{0,028}, \text{ soit } t_a = 43,293 \text{ et } t_b = \frac{\hat{b}}{\sigma_b} = \frac{215,52}{32,26}, \text{ soit } t_b = 6,681.$$

$t_a$  et  $t_b$  sont tous deux supérieurs au  $t_{\text{théorique}} = 2,571$  obtenu par lecture de la table de Student, avec une probabilité de 0,05 ( $\alpha = 5\%$ ) et  $n - 2 = 5$  degrés de liberté. De plus, toujours par lecture de la table statistique, la probabilité associée à  $t_a$  ( $p = 0,000$ ) et celle associée à  $t_b$  ( $p = 0,001$ ) sont toutes deux inférieures à 5 %. (Pour un rappel sur les tests d'hypothèses, voir focus 5.1.)

Le test de Student pour le coefficient  $a$  de la régression linéaire permet de conclure que la valeur de  $a$  est significativement différente de 0. De même, le test de Student pour le coefficient  $b$  de la régression linéaire permet de conclure que la valeur de  $b$  est significativement différente de 0.

c. Afin de réaliser le test de Fisher, les  $(\hat{y}_i - \bar{y})^2$  sont calculées en L7 (voir figure 6.15). Pour cela, placez le curseur sur l'en-tête de la septième colonne, et, après l'avoir nommée L7, indiquez  $L7=(L4-1607,44)^2$ , puis appuyez sur **ENTER**.

Pour obtenir la somme des  $(\hat{y}_i - \bar{y})^2$ , appuyez sur la touche **STAT**, puis choisissez le menu CALC et sélectionnez la fonction 1:1-Var Stats. Puis appuyez sur **ENTER**. Tapez 1-Var Stats **L7** (*ne pas taper L7, mais l'appeler dans la liste de noms des variables : 2ND LIST*, menu NAMES, sélectionner 7:L7) puis appuyez à nouveau sur **ENTER**. Les statistiques sur la variable  $(\hat{y}_i - \bar{y})^2$ , contenues dans L7, s'inscrivent (voir figure 6.16).

Figure 6.15 (gauche)

**Calculs dans L7 avec la calculatrice.**

Figure 6.16 (droite)

**Statistiques sur L7,  $(\hat{y}_i - \bar{y})^2$ .**

L5	L6	L7	7
5.0391	19240	28319	
18,95	8913,2	77391	
4,2502	2511	13132	
82,192	7,8961	3709	
12,523	2170,6	12,773	
23,651	9080,2	3169,6	
99,401	20791	13299	
		30482	
<b>L7(1)=28319,77391...</b>			
<b>1-Var Stats</b>			
<b><math>\bar{x}=13160,54531</math></b>			
<b><math>\sum x=92123,8172</math></b>			
<b><math>\sum x^2=2104258717</math></b>			
<b><math>S_x=12191,93071</math></b>			
<b><math>\sigma_x=11287,53451</math></b>			
<b><math>\downarrow n=7</math></b>			

$$F_c = \frac{SCE/\frac{1}{1}}{SCR/\frac{5}{5}} = \frac{92123,82/\frac{1}{1}}{246,006/\frac{5}{5}}, \text{ soit } F_c = 1874,307.$$

$F_c$  est supérieur au  $F_{(0,05;1,5)} = 6,608$  obtenu par lecture de la table de Fisher, avec une probabilité de 0,05 ( $\alpha = 5\%$ ),  $ddl_1 = 1$  et  $ddl_2 = n - 2 = 5$  degrés de liberté. On rejette donc  $H_0$  au seuil de signification 5 % et l'on conclut à l'existence d'une relation linéaire entre X et Y.

3. En appliquant l'équation  $y = 1,212 x + 215,52$  pour une consommation  $x = 1\ 400$ ,  $y = 1,212 \times 1\ 400 + 215,52$ , soit  $y = 1\ 912,32$ . Pour une consommation de 1 400 milliards d'euros, le gouvernement peut s'attendre à un PIB de 1 912,32 milliards d'euros.

4. La droite de régression qui permet d'expliquer la consommation en fonction du PIB est telle que  $x = a'y + b'$ .

À partir de l'ensemble des calculs déjà effectués :

$$a' = \frac{\sum_{i=1}^7 x_i y_i - n\bar{y} \cdot \bar{x}}{nV(y)} = \frac{12\ 996\ 965,9 - 7 \times 1148,31 \times 1607,44}{7 \times 114,89^2} = 0,823 \text{ et}$$

$$b' = \bar{y} - a' \bar{x} = 1148,31 - 0,823 \times 1607,44 = -174,27.$$

D'où l'équation de régression de X selon y :  $x = 0,823 y - 174,27$ .

5. En appliquant l'équation  $x = 0,823 y - 174,27$  pour un PIB  $y = 1\ 600$ ,  $x = 0,823 \times 1\ 600 - 174,27$ , soit  $x = 1\ 142,5$ . Pour un PIB de 1 600 milliards d'euros, la consommation correspondante est de 1 142,5 milliards d'euros.



## EXERCICE 3 RÉGRESSION SUR TABLEAU DE CONTINGENCE

### Énoncé

Soit X l'espérance de vie des hommes et Y l'espérance de vie des femmes, relevées en 2004 dans 21 pays :

X \ Y	[75 ; 80[	[85 ; 85[
[65 ; 70[	4	0
[70 ; 75[	0	3
[75 ; 80[	2	12

Sources : Eurostat et instituts nationaux de statistique, 2004

En utilisant la méthode des moindres carrés ordinaires, établissez la droite de régression  $y = ax + b$ .

### Solution

Pour établir l'équation de la droite de régression  $y = ax + b$ , il convient de déterminer les valeurs de a et b dans cette équation. Pour cela, il est nécessaire de calculer les valeurs de  $\bar{x}$ ,  $\bar{y}$ ,  $V(x)$  et  $Cov(x; y)$ .

Ces valeurs sont calculées selon les étapes détaillées au chapitre 5 (voir figure 6.17).

Figure 6.17

Résultats sous Excel.

A	B	C	D	E	F	G
1 X \ Y	[75;80[	[80;85[	n <sub>i+</sub>	x <sub>j</sub>	n <sub>i+j</sub>	n <sub>i+j</sub> x <sub>j</sub> <sup>2</sup>
2 [65;70[	4	0	4	67,5	270,00	18 225,00
3 70;75[	0	3	3	72,5	217,50	15 768,75
4 [75;80[	2	12	14	77,5	1 085,00	84 087,50
5 n <sub>ij</sub>	6	15	21		1 572,50	118 081,25
6 y <sub>j</sub>	77,5	82,5				
7 n <sub>i+j</sub> y <sub>j</sub>	465,00	1 237,50	1 702,50			
8 n <sub>ij</sub> y <sub>j</sub> <sup>2</sup>	36 037,50	102 093,75	138 131,25			
9 Σn <sub>ij</sub> x <sub>i+j</sub> y <sub>j</sub>	32 937,50	94 668,75	127 606,25			

$$\text{De là, } \bar{x} = \frac{1}{n_{++}} \sum_{i=1}^3 n_{i+} \times x_i = \frac{1572,5}{21} = 74,88 \text{ et, de même,}$$

$$\bar{y} = \frac{1}{n_{++}} \sum_{i=1}^2 n_{+j} \times y_j = \frac{1702,5}{21} = 81,07 .$$

$$V(x) = \frac{1}{n_{++}} \sum_{i=1}^3 n_{i+} x_i^2 - \bar{x}^2 = \frac{118 081,25}{21} - \left( \frac{1572,5}{21} \right)^2 = 15,76 \text{ et}$$

$$Cov(x; y) = \frac{1}{n_{++}} \sum_{j=1}^2 \sum_{i=1}^3 n_{ij} x_i y_j - \bar{x} \times \bar{y} = \frac{127 606,25}{21} - (81,07 \times 74,88) = 5,78 .$$

À partir des formules  $\begin{cases} \hat{a} = \frac{Cov(X; Y)}{V(X)} \\ \hat{b} = \bar{y} - \hat{a}\bar{x} \end{cases}$ , il vient :  $\begin{cases} \hat{a} = \frac{5,78}{15,76} = 0,3669 \\ \hat{b} = 81,07 - 0,3669 \times 74,88 = 53,60 \end{cases}$

D'où  $D_{Y/x} : y = 0,3669 x + 53,60$ .



## EXERCICE 4 AJUSTEMENT EXPONENTIEL ET PAPIER SEMI-LOGARITHMIQUE

## Énoncé

Les données suivantes sont extraites d'une table de mortalité et de survie (1959-1963), ajustée par une loi de Makeham, actuaire anglais (décédé en 1892) :

Âge (X)	Taux instantané de mortalité (Y)
50	0,008541
51	0,009287
52	0,010103
53	0,010998
54	0,011978
55	0,013051
56	0,014228
57	0,015516
58	0,016928
59	0,018474
60	0,020169
61	0,022025
62	0,024059
63	0,026287
64	0,028728
65	0,031402
66	0,034332

1. Représentez le nuage de points entre X et Y.
2. Effectuez l'ajustement qui permet d'expliquer Y selon x, par la relation :  $Y = B \times A^x$ .
3. Calculez les indicateurs de qualité de la régression :
  - a. le coefficient de détermination et le test associé ;
  - b. les tests de Student ;
  - c. le test de Fisher.
4. Quel est le taux instantané de mortalité d'un individu de 70 ans ?

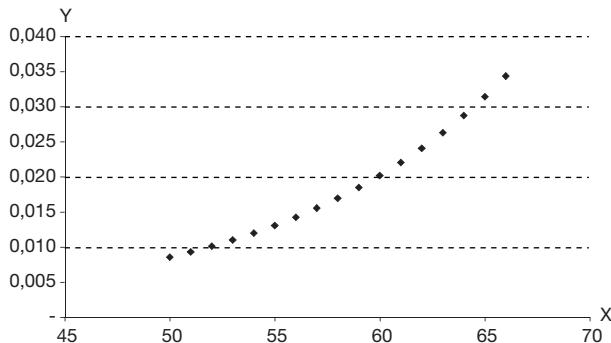
## Solution

1. Soit X : « âge » et Y : « taux instantané de mortalité ». Pour représenter le nuage de points sous Excel, cliquez sur Insertion/Graphique dans la barre de menus, puis, dans l'assistant graphique, choisissez le type de graphique Nuage de points, puis, dans Sous-

type de graphique, sélectionnez l'image « Nuage de points. Compare des paires de valeurs ». Cliquez sur Suivant et indiquez dans le champ correspondant la plage où se trouvent les données (voir chapitre 1, exercice 5).

**Figure 6.18**

**Nuage de points entre X et Y sous Excel.**

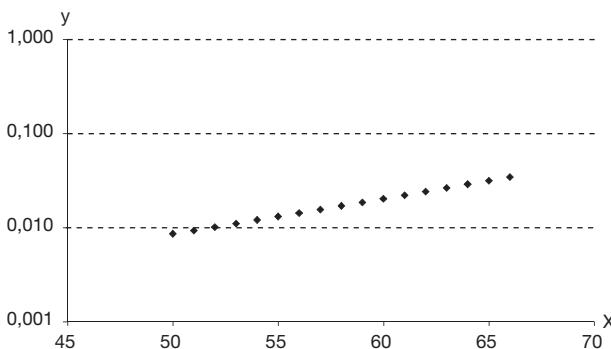


Le graphique de la figure 6.18 évoque une croissance de type exponentiel. Une croissance exponentielle se traduit par une équation du type  $Y = B \times A^X$ , soit, en passant aux logarithmes népériens :  $\ln Y = \ln B + X \ln A$ , en posant  $y = \ln Y$ ,  $b = \ln B$  et  $a = \ln A$  :  $y = ax + b$ , ce qui équivaut à une liaison linéaire entre  $x$  et  $y$ .

On peut tester graphiquement cette hypothèse, en représentant le nuage dans un graphique semi-logarithmique (l'échelle des ordonnées est logarithmique, l'échelle des abscisses reste identique). L'alignement des points valide l'hypothèse de liaison linéaire entre  $x$  et  $y$  (voir figure 6.19).

**Figure 6.19**

**Graphique semi-logarithmique.**



2. Afin de rapporter la relation  $Y = B \times A^X$  à une équation de droite, il est nécessaire de procéder au changement de variables en passant aux logarithmes népériens, comme indiqué dans la question 1 :  $\ln Y = \ln B + X \ln A$ , en posant  $y = \ln Y$ ,  $b = \ln B$  et  $a = \ln A$  :  $y = ax + b$ . La relation est linéaire, il est donc possible de procéder à l'estimation de la droite de régression par la méthode des moindres carrés ordinaires.

L'application du changement de variables sur les valeurs de  $Y$  est réalisée dans la colonne D. Puis l'ensemble des calculs nécessaires à l'estimation de la droite de régression est effectué à partir des valeurs calculées de  $X$  et de  $y$  (voir figure 6.20).

Figure 6.20

Résultats sous Excel.

	A	B	C	D	E
1	Observation	X	Y	y	Xy
2	Observation 1	50	0,009	- 4,763	- 238,14
3	Observation 2	51	0,009	- 4,679	- 238,64
4	Observation 3	52	0,010	- 4,595	- 238,94
5	Observation 4	53	0,011	- 4,510	- 239,03
6	Observation 5	54	0,012	- 4,425	- 238,93
7	Observation 6	55	0,013	- 4,339	- 238,64
8	Observation 7	56	0,014	- 4,253	- 238,14
9	Observation 8	57	0,016	- 4,166	- 237,46
10	Observation 9	58	0,017	- 4,079	- 236,57
11	Observation 10	59	0,018	- 3,991	- 235,49
12	Observation 11	60	0,020	- 3,904	- 234,22
13	Observation 12	61	0,022	- 3,816	- 232,75
14	Observation 13	62	0,024	- 3,727	- 231,09
15	Observation 14	63	0,026	- 3,639	- 229,24
16	Observation 15	64	0,029	- 3,550	- 227,19
17	Observation 16	65	0,031	- 3,461	- 224,96
18	Observation 17	66	0,034	- 3,372	- 222,53
19	Somme	986		-69,267	-3 981,95
20	Moyenne	58		- 4,075	
21	Variance	24		0,182	
22	Covariance (X;Y)			2,089	
23	Ecart type	4,899		0,426	

$$\text{De là, } a = \frac{\sum_{i=1}^{17} X_i y_i - n \bar{X} \bar{y}}{n V(X)} = \frac{-3981,95 - 17 \times 58 \times (-4,075)}{9 \times 24} = 0,0871 \text{ et}$$

$$b = \bar{y} - a \bar{X} = -4,075 - 0,0871 \times 58 = -9,12.$$

D'où l'équation de régression de y selon X :  $y = 0,0871 \times X - 9,12$ .

En effectuant le changement de variables qui permet de revenir à la relation initiale :  $b = \ln B \Leftrightarrow B = e^b$  et  $a = \ln A \Leftrightarrow A = e^a$ , d'où  $Y = e^{-9,12} \times e^{0,0871X}$ , soit  $Y = 0,000109 \times e^{0,0871X}$ .

3. Les indicateurs de qualité de la droite de régression sont calculés pour l'équation de la droite de régression  $y = 0,0871 \times X - 9,12$ . La qualité de cette droite conditionne la qualité de l'estimation non linéaire  $Y = 0,000109 \times e^{0,0871X}$ .

a. Le calcul du coefficient de corrélation linéaire nécessite de calculer la covariance entre x et y et les écarts-types de x et de y.

$$COV(X; y) = \frac{-3 981,95}{17} - 58 \times (-4,075), \text{ soit } COV(X; y) = 2,089.$$

L'écart-type de X est la racine de  $V(X)$ , calculée précédemment, soit  $\sqrt{24} = 4,899$ . D'où  $\sigma_x = 4,899$ .

De même, l'écart-type de y est la racine de  $V(y)$ .  $V(y)$  est calculée en utilisant la fonction VAR.P d'Excel (voir annexe 1.1). Elle peut également l'être selon la méthode exposée précédemment (voir chapitres 2 et 3).  $V(y) = 0,182$ ; soit  $\sqrt{0,182} = 0,426$ . D'où  $\sigma_y = 0,426$ .

De là,  $r = \frac{2,089}{4,899 \times 0,426}$ , soit  $r = 0,99996$ , soit  $R^2 = 0,99992$ . Il existe une forte corrélation linéaire positive entre X et y.

D'où  $T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0,99996\sqrt{17-2}}{\sqrt{1-0,99992^2}} = 302,08$  et la table de Student donne  $t_{0,025; 15} = 2,49$ .

Puisque  $302,08 \geq 2,49$ , soit  $t \geq t_{\alpha/2; n-2}$ , il faut rejeter l'hypothèse  $H_0$ . Il y a donc une corrélation linéaire significative entre X et y.

Le  $t_{\text{théorique}}$  est disponible sous Excel en appelant la fonction statistique LOI.STUDENT.INVERSE et en saisissant les arguments suivants : Probabilité = 0,025 et Degrés\_liberté = 15.

La probabilité de Student associée peut s'obtenir à l'aide d'Excel en appelant la fonction statistique LOI.STUDENT et en saisissant les arguments suivants : X = 302,08 (le Student calculé), Degrés\_liberté = 15 et Uni/bilatéral = 1.

**b.** Afin de réaliser les tests de Student, les calculs intermédiaires suivant sont réalisés.

**Figure 6.21**

**Résultats sous Excel.**

	A	F	G	H	I
1	Observation	$\hat{y}$	$(\hat{y} - \bar{y})^2$	$(\hat{y} - y)^2$	$(X - \bar{X})^2$
2	Observation 1	- 4,77	0,48	0,0001	64,00
3	Observation 2	- 4,68	0,37	0,0000	49,00
4	Observation 3	- 4,60	0,27	0,0000	36,00
5	Observation 4	- 4,51	0,19	0,0000	25,00
6	Observation 5	- 4,42	0,12	0,0000	16,00
7	Observation 6	- 4,34	0,07	0,0000	9,00
8	Observation 7	- 4,25	0,03	0,0000	4,00
9	Observation 8	- 4,16	0,01	0,0000	1,00
10	Observation 9	- 4,07	0,00	0,0000	-
11	Observation 10	- 3,99	0,01	0,0000	1,00
12	Observation 11	- 3,90	0,03	0,0000	4,00
13	Observation 12	- 3,81	0,07	0,0000	9,00
14	Observation 13	- 3,73	0,12	0,0000	16,00
15	Observation 14	- 3,64	0,19	0,0000	25,00
16	Observation 15	- 3,55	0,27	0,0000	36,00
17	Observation 16	- 3,47	0,37	0,0000	49,00
18	Observation 17	- 3,38	0,48	0,0000	64,00
19	Somme	- 69,27	3,09	0,0003	408,00

La détermination de l'écart-type de  $\hat{a}$ ,  $\sigma_{\hat{a}}$ , nécessite le calcul de  $SCR = \sum_{i=1}^n (\hat{y}_i - y_i)^2$  et de

$SCT = \sum_{i=1}^n (x_i - \bar{x})^2$ , effectué respectivement dans les cellules H19 et I19 (voir figure 6.21).

De là,  $\sigma_{\hat{a}}^2 = \frac{1}{15} \times \frac{0,0003}{408}$ , soit  $\sigma_{\hat{a}}^2 = 4,15E-08$ ; d'où  $\sigma_{\hat{a}} = 0,00020$ .

$\sigma_{\hat{b}}^2 = 4,15E-08 \times \left( \frac{408}{17} + 58^2 \right)$ , soit  $\sigma_{\hat{b}}^2 = 0,000141$ ; d'où  $\sigma_{\hat{b}} = 0,01186$ .

$t_a = \frac{\hat{a}}{\sigma_{\hat{a}}} = \frac{0,0871}{0,00020}$ , soit  $t_a = 427,195$  et  $t_b = \frac{\hat{b}}{\sigma_{\hat{b}}} = \frac{-9,12}{0,01186}$ , soit  $t_b = -769,205$ .

$t_a$  et  $t_b$  sont tous deux supérieurs au  $t_{\text{théorique}} = 2,131$  obtenu par lecture de la table de Student, avec une probabilité de 0,05 ( $\alpha = 5\%$ ) et  $n - 2 = 15$  degrés de liberté. De plus, toujours par lecture de la table statistique, la probabilité associée à  $t_a$  ( $p = 0,000$ ) et celle associée à  $t_b$  ( $p = 0,000$ ) sont toutes deux inférieures à 5 %.

Le test de Student pour le coefficient  $a$  de la régression linéaire permet de conclure que la valeur de  $a$  est significativement différente de 0. De même, le test de Student pour le coefficient  $b$  de la régression linéaire permet de conclure que la valeur de  $b$  est significativement différente de 0.

Le  $t_{\text{théorique}}$  est disponible sous Excel en appelant la fonction statistique LOI.STUDENT.INVERSE et en saisissant les arguments suivants : Probabilité = 0,05 et Degrés\_liberté = 15.

La probabilité de Student associée peut s'obtenir à l'aide d'Excel en appelant la fonction statistique LOI.STUDENT et en saisissant les arguments suivants :  $X = 427,195$  pour  $a$  et  $X = -769,205$  pour  $b$  (le Student calculé), Degrés\_liberté = 15 et Uni/bilatéral = 2.

c. La détermination du Fisher nécessite le calcul de  $SCE = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ , effectué dans la cellule G11 (voir figure 6.21).

$$F_c = \frac{\frac{SCE}{1}}{\frac{SCR}{15}} = \frac{\frac{3,09}{1}}{\frac{0,0003}{15}}, \text{ soit } F_c = 182\,495,41.$$

$F_c$  est supérieur au  $F_{(0,05; 1, 15)} = 4,543$  obtenu par lecture de la table de Fisher, avec une probabilité de 0,05 ( $\alpha = 5\%$ ),  $ddl_1 = 1$  et  $ddl_2 = n - 2 = 15$  degrés de liberté. On rejette donc  $H_0$  au seuil de signification 5 % et l'on conclut à l'existence d'une relation linéaire entre  $X$  et  $Y$ .

Le  $F_{(0,05; 1, 15)}$  est disponible sous Excel en appelant la fonction statistique INVERSE.LOI.F et en saisissant les arguments suivants : Probabilité = 0,05, Degrés\_liberté1 = 1 et Degrés\_liberté2 = 15.

La probabilité de Fisher associée peut s'obtenir à l'aide d'Excel en appelant la fonction statistique LOI.F et en saisissant les arguments suivants :  $X = 182\,495,41$  (le Fisher calculé), Degrés\_liberté1 = 1 et Degrés\_liberté2 = 15.

4. Nous utilisons l'équation initiale  $Y = 0,000109 \times e^{0,0871X}$  afin de réaliser une prévision à partir de la valeur  $X = 70$ . Ainsi,  $Y = 0,000109 \times e^{0,0871 \times 70} = 0,0483$ .

Le taux instantané de mortalité d'un individu de 70 ans est de 0,0483.



## EXERCICE 5 CORRÉLATION DES RANGS

### Énoncé

Le tableau suivant indique pour les 15 étudiants d'un TD de statistiques leur rang au partiel et leur rang à l'examen :

Observation	Partiel	Examen
Étudiant 1	4	5
Étudiant 2	6	7
Étudiant 3	7	11

Observation	Partiel	Examen
Étudiant 5	2	1
Étudiant 6	8	8
Étudiant 7	9	4
Étudiant 8	3	2
Étudiant 9	15	15
Étudiant 10	13	6
Étudiant 11	1	12
Étudiant 12	10	13
Étudiant 13	14	9
Étudiant 14	12	10
Étudiant 15	5	3

Calculez le coefficient de corrélation de rang de Spearman.

### Solution

Posons X : « rang au partiel » et Y : « rang à l'examen ». Calculons chacune des distances entre le rang d'un étudiant au partiel et son rang à l'examen :  $d_i = x_i - y_i$ .

Saisissez les valeurs de X, le rang au partiel, dans la colonne L1 et celles de Y, le rang à l'examen, dans la colonne L2. Pour obtenir les  $(x_i - y_i)^2$  dans la colonne L3, placez le curseur sur l'en-tête de colonne L3, puis indiquez  $L3=(L1-L2)^2$ . Puis appuyez sur **ENTER**. Le résultat de ces opérations est proposé figure 6.22.

Figure 6.22 (gauche)

Saisie des données et calcul des distances avec la calculatrice.

Figure 6.23 (droite)

Statistiques sur les  $d_i^2$ .

L1	L2	L3	3
4	5	1	
6	7	1	
7	11	16	
11	14	9	
2	1	1	
8	8	0	
9	4	25	

1-Var Stats  
 $\bar{x}=17.73333333$   
 $\Sigma x=266$   
 $\Sigma x^2=18746$   
 $Sx=31.65543661$   
 $\sigma x=30.58205719$   
 $\downarrow n=15$

Appuyez sur la touche **STAT**, puis choisissez le menu CALC et sélectionnez la fonction 1:1-Var Stats. Puis appuyez sur **ENTER**. Tapez 1-Var Stats L3 puis appuyez à nouveau sur **ENTER**. Les statistiques sur la variable  $d_i^2$ , contenue dans L3, s'inscrivent (voir figure 6.23).

$$\sum_{i=1}^{15} d_i^2 = 266, \text{ donc } r_s = 1 - \frac{6 \times 266}{15 \times (15^2 - 1)}, \text{ soit } r_s = 0,525. \text{ Il existe un lien entre le rang d'un étudiant au partiel et son rang à l'examen, mais ce lien n'est pas très fort.}$$

## Bibliographie

- BAILLARGEON G., *Méthodes statistiques de l'ingénieur*, SMG, 1990.
- BLUMENTHAL S., *Statistiques appliquées*, Éditions d'Organisation, 1989.
- BOWKER A.H. et LIEBERMAN G.J., *Méthodes statistiques de l'ingénieur*, Dunod, 1965.
- BOREL E., DELTHEIL R. et HURON R., *Probabilités. Erreurs*, Armand Colin, 1960.
- CALOT G., *Cours de statistique descriptive*, Dunod, 1969.
- CHAREILLE P. et PINAULT Y., *Statistique descriptive*, Collection AES, Montchrestien, Paris, 1996.
- DARMOIS G., *Statistiques et applications*, Armand Colin, 1952.
- DELAHAYE J.-P., « L'étonnante loi de Benford », *Pour la science*, janvier 2007
- DODGE Y., *Statistique. Dictionnaire encyclopédique*, Springer, 2004.
- DODGE Y., *Premiers pas en statistique*, Springer, 2006.
- DROESBEKE J.-J. et TASSI Ph., *Histoire de la statistique*, Que sais-je ?, PUF, 1990.
- GELLER S., *Abrégé de statistique*, Éditions Masson, 1979.
- GRENON G. et VIAU S., *Méthodes quantitatives en sciences humaines*, Gaëtan Morin, 1999.
- HAUCHECORNE B., *Les mots et les maths*, Ellipses, 2003.
- LIORZOU A., *Initiation à la pratique statistique*, Eyrolles, 1979.
- LEGRIS G., *Statistiques pour économistes*, Economica, 1987.
- ROGER P., *Probabilités, statistique et processus stochastiques*, Collection Synthex, Pearson Education, 2004.
- SCHLACTHER D., *De l'analyse à la prévision*, Ellipses, 1986.
- TINTNER G., *Mathématiques et statistiques pour les économistes*, Dunod, 1962.

# Annexe 6.1

## La fonction DROITEREG d'Excel

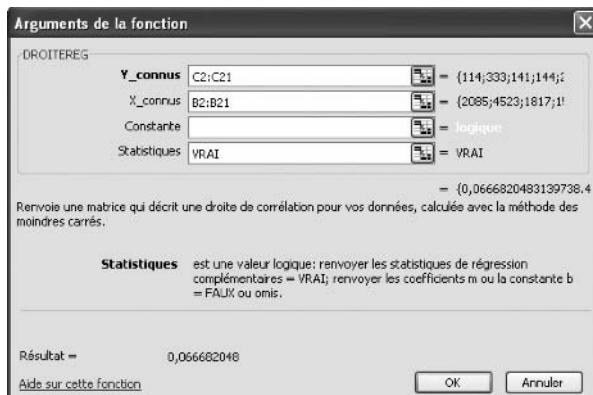
La droite de régression et plusieurs de ses indicateurs peuvent être obtenus en utilisant la fonction statistique DROITEREG d'Excel.

Pour cela, sélectionnez une plage de 2 lignes et 5 colonnes, soit 10 cellules, qui représentent la matrice dans laquelle les résultats seront affichés. Appelez la fonction statistique DROITEREG. Le masque de dialogue suivant s'affiche (voir figure 6.24) :

- Dans le champ Y\_connus, sélectionnez la colonne dans laquelle se trouvent les valeurs de Y.
- Dans le champ X\_connus, sélectionnez la colonne dans laquelle se trouvent les valeurs de X.
- Le champ Constante est laissé vide.
- Dans le champ Statistiques, saisissez VRAI.

Figure 6.24

Masque de dialogue de la fonction DROITEREG sous Excel.



Une fois le masque de dialogue rempli, ne cliquez pas sur **OK** : tenez enfoncées en même temps les touches Ctrl et Shift tout en appuyant sur **ENTRÉE**. Cette procédure permet l'affichage matriciel des résultats dans les 10 cellules sélectionnées précédemment (voir figure 6.25).

Figure 6.25

Résultat de la fonction DROITEREG sous Excel.

69	0,0667	45,5698
70	0,0064	19,1398
71	0,8560	40,5042
72	107,0196	18,0000
73	175,575,37	29,530,63

Ces résultats numériques correspondent aux indicateurs suivants, en respectant l'ordre des lignes et des colonnes de la figure 6.25 :

a	b
$\sigma_a$	$\sigma_b$
$r^2$	$\sigma_{\hat{y}}$
F	ddl
SCT	SCR

## Annexe 6.2

### La fonction LinReg(ax + b) de la calculatrice

La droite de régression et le  $r^2$  peuvent être obtenus en utilisant la fonction LinReg de la calculatrice.

Pour cela, commencez par activer le DiagnosticOn en appuyant sur les touches **2ND** et **CATALOG** et en sélectionnant la fonction DiagnosticOn.

Pour effectuer la régression, saisissez les valeurs des X en L1 et les valeurs de Y en L2, comme dans l'exercice 2. Appuyez sur la touche **STAT**, puis choisissez le menu CALC et sélectionnez la fonction 4:LinReg(ax + b). Puis appuyez sur **ENTER**.

Les résultats de la régression s'affichent (voir figure 6.26).

Figure 6.26

Résultat de la fonction LinReg(ax + b) de la calculatrice.

```
LinReg
y=ax+b
a=1.21214493
b=215.5195177
r^2=.9973394443
r=.9986688362
```



# Les séries chronologiques

1. Présentation de la série chronologique.....	186
2. Agrégation des composantes.....	197

## Problèmes et exercices

1. Méthode empirique et modèle additif .....	204
2. Méthode empirique et modèle multiplicatif.....	207
3. Méthode analytique et modèle additif .....	210
4. Méthode analytique et modèle multiplicatif.....	214

Parmi les séries doubles, certaines méritent d'être traitées à part : celles qui décrivent l'évolution d'un phénomène par rapport au temps, et que l'on nomme séries temporelles, chronologiques ou encore chroniques. Nous traiterons ici des séries doubles dont le premier caractère est le temps et dont le deuxième caractère est quantitatif. L'analyse des séries chronologiques est fondée sur l'existence d'une corrélation entre le caractère étudié et le temps. Ces séries interviennent dans des domaines aussi variés que l'astronomie, la démographie, l'économie, l'histoire, etc.

Ainsi que l'indique Jean-Marie Dufour dans son article intitulé « Histoire de l'analyse des séries chronologiques »<sup>1</sup>, c'est en astronomie que sont apparues les premières séries chronologiques.

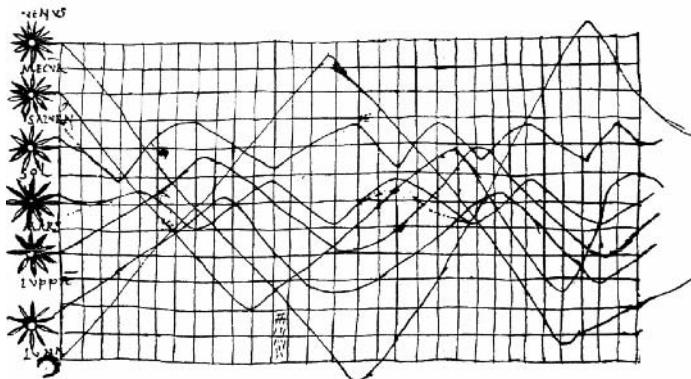
1. <http://www.fas.umontreal.ca/SCECO/Dufour>. Jean-Marie Dufour est titulaire de la chaire de recherche en économétrie à l'université de Montréal au Canada.

D'après Kendall, le plus ancien graphique connu d'une série chronologique se trouve dans un manuscrit du X<sup>e</sup> ou du XI<sup>e</sup> siècle et représente l'inclinaison des orbites de sept planètes en fonction du temps ; il est reproduit figure 7.1.

Figure 7.1

**Graphique chronologique.**

Source : Funkhauser (1936)



L'objectif de l'étude d'une série chronologique est de mettre en évidence l'évolution passée d'une variable statistique et sous certaines conditions d'extrapoler cette évolution afin d'effectuer des prévisions à court terme.

L'analyse des séries chronologiques consistera à mettre en évidence leurs quatre composantes : une composante tendancielle, une composante cyclique, une composante saisonnière et une composante accidentelle (bruit). Cette décomposition a été proposée en 1919 par le statisticien Warren Persons<sup>1</sup>.

Nous mettrons en évidence l'existence de deux modèles de composition de ces composantes : le modèle additif et le modèle multiplicatif.

Pour faire apparaître la composante tendancielle (appelée le trend), nous utiliserons la méthode MCO ou les moyennes mobiles.

## 1 Présentation de la série chronologique

La variable dont on suit l'évolution au cours du temps peut être un niveau (on parle aussi de stock), comme la température, le nombre de chômeurs, etc., ou un flux, c'est-à-dire un nombre d'événements observés au cours d'une période, comme le nombre mensuel de naissances, la consommation des ménages, etc. Dans les deux cas, le temps qui représente les dates ou les périodes d'observation sera repéré par l'indice  $t$  et numéroté de 1 à  $n$ .

### Définition

On appelle **série chronologique**, ou série temporelle, une suite d'observations chiffrées d'un caractère quantitatif  $Y$ , ordonnées dans le temps. La valeur prise par la variable  $Y$  à la date  $t$  est notée  $y_t$ .

1. Warren Persons (1878-1937) a développé un indicateur de la conjoncture économique, connu sous le nom de baromètre de Harvard.

Avant toute analyse, nous représenterons les données par une courbe exprimant la continuité de l'évolution de la variable étudiée. Nous supposerons que les dates d'observation sont équidistantes (mois, trimestres, années...) et nous les représenterons par les entiers naturels non nuls : 1, 2, 3...

## 1.1 LES REPRÉSENTATIONS GRAPHIQUES

L'analyse des séries chronologiques se fonde sur la décomposition de l'évolution d'un caractère en plusieurs composantes et, comme nous l'avons indiqué précédemment, il est nécessaire de réaliser une représentation graphique afin de guider la réflexion. La représentation graphique classique est calquée sur le nuage de points, mais les points seront reliés par des segments de droite pour traduire la chronologie. Le temps sera noté  $t$  et on lui donnera les valeurs 1, 2, ...,  $n$  si l'on a  $n$  périodes, les modalités du caractère étudié étant notées  $y_t$ .

### Exemple 7.1

#### La première série chronologique

Le tableau suivant donne les indices trimestriels de stocks de matières en valeur des industries agricoles et alimentaires (IAA) :

	1 <sup>e</sup> trimestre	2 <sup>e</sup> trimestre	3 <sup>e</sup> trimestre	4 <sup>e</sup> trimestre
2004	108,2	104,5	102,8	107,8
2005	107,9	106,2	104,5	112,3
2006	110,8	110,7	108	115,2

Source : Insee, 2007

On associera à cette série le tableau statistique de la figure 7.2.

Figure 7.2

Tableau statistique  
d'une série  
chronologique.

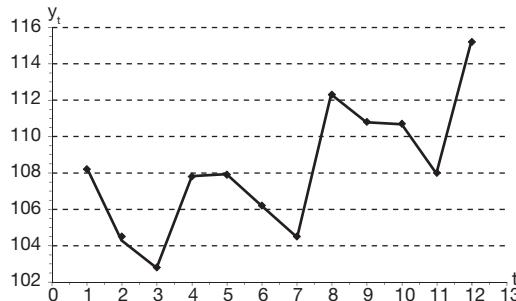
	A	B
1	t	$y_t$
2	1	108,2
3	2	104,5
4	3	102,8
5	4	107,8
6	5	107,9
7	6	106,2
8	7	104,5
9	8	112,3
10	9	110,8
11	10	110,7
12	11	108
13	12	115,2

La série sera ainsi représentée par le graphique de la figure 7.3.

Pour mettre en évidence une éventuelle variation périodique, ou une saisonnalité de la série, on réalise une représentation superposée des données, qui permet, dans notre exemple, de mettre en évidence le caractère propre de chaque trimestre (voir figure 7.4).

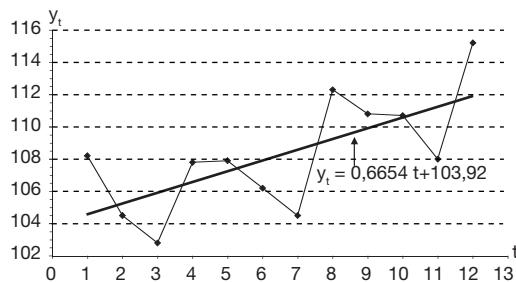
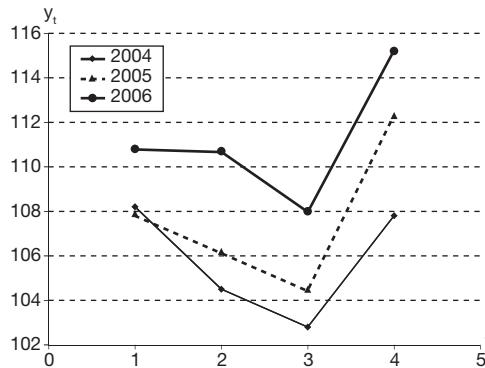
**Figure 7.3**

**Représentation graphique de la série chronologique des indices trimestriels IAA.**



**Figure 7.4**

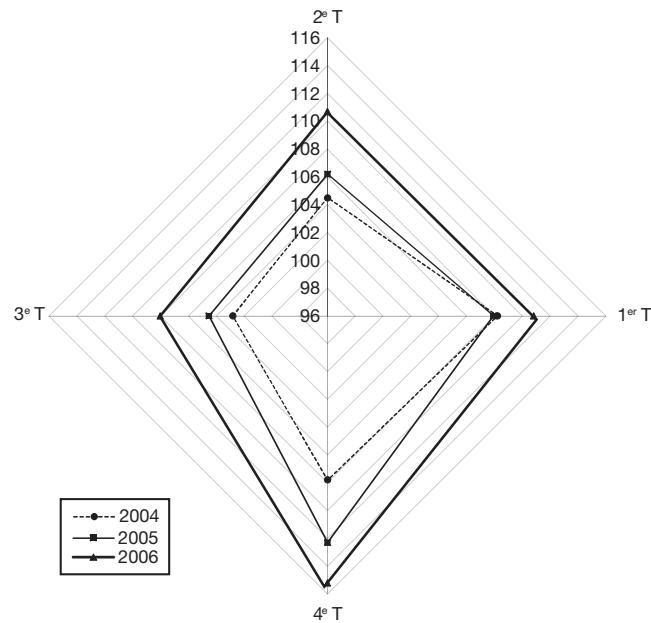
**Représentation superposée des données d'indices trimestriels IAA.**



On représente souvent les séries chronologiques par un graphique polaire s'inspirant de certains thermomètres enregistreurs, qui utilisent une feuille enroulée sur un cylindre permettant de visualiser rapidement la température tous les jours d'une semaine à la même heure. Excel ne permet pas de réaliser un graphique polaire, mais propose un graphique approchant, nommé « Radar », dont la figure 7.5 donne la représentation.

**Figure 7.5**

**Graphique « Radar » des indices trimestriels IAA.**

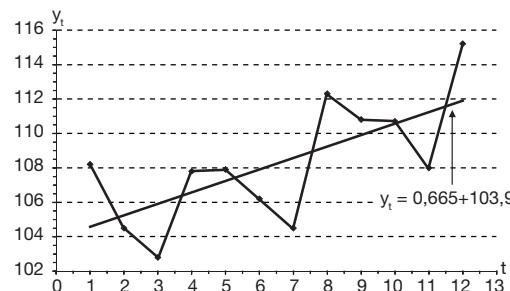


## 1.2 LES COMPOSANTES

Les fluctuations d'une série chronologique sont le fruit de la composition de plusieurs composantes. Nous avons repris ici l'exemple 7.1 auquel nous avons ajouté la droite de tendance calculée par la méthode MCO sous Excel.

**Figure 7.6**

**Série chronologique et trend.**



La droite de régression de  $Y$  en  $t$  représente la composante tendancielle de cette série chronologique. Elle exprime son mouvement de longue durée. La série est le résultat de la superposition de deux autres composantes à cette composante fondamentale.

### Définitions

On appelle **tendance** ou composante générale ou composante extra-saisonnière d'une série chronologique sa tendance générale. Cette tendance générale (dite séculaire) exprime une tendance durable à la croissance (mouvement de longue durée ascendant) ou à la décroissance (mouvement de longue durée descendant).

On décompose parfois cette composante tendancielle en deux éléments : la tendance à long terme et une composante périodique appelée **cycle**. Le mouvement cyclique résulte de la succession de périodes d'expansion et de dépression. La reprise est le passage de la

dépression à l'expansion et la crise le passage de l'expansion à la dépression. Ces deux composantes ne sont pas toujours distinguables et on ne cherchera pas à les distinguer ; on notera  $f_t$  cette composante tendancielle, que l'on identifiera à la tendance durable et que l'on appellera trend.

La **composante saisonnière** de la série est sa composante périodique dans le cadre de l'année (elle peut être due aux saisons, comme pour l'IAA, ou résulter des usages (fêtes, vacances, etc.) ; elle sera notée  $S_t$ .

On appelle **composante résiduelle** (bruit, aléa) ou accidentelle les fluctuations irrégulières et imprévisibles de la série ; elle sera notée  $\varepsilon_t$  (erreur).

## 1.3 DÉTERMINATION DE LA TENDANCE

---

Nous aborderons trois méthodes pour déterminer le trend :

- une méthode purement graphique : la méthode des points moyens (voir sur le site [www.pearsoned.fr](http://www.pearsoned.fr)) ;
- une méthode analytique : la méthode MCO (nous n'envisagerons que le cas du trend linéaire) ;
- des méthodes empiriques :
  - la méthode des moyennes échelonnées ;
  - la méthode des moyennes mobiles non centrées ;
  - la méthode des moyennes mobiles centrées.

### La méthode analytique : MCO

Dans le cas d'une série chronologique, la variable explicative est le temps ( $T$ ) et on ajustera une droite à l'ensemble des observations, par la méthode des moindres carrés, en cherchant la droite de régression de  $Y$  selon  $t$ , pour obtenir une équation du type :

$$y = at + b, \text{ avec : } \begin{cases} a = \frac{\text{Cov}(T; Y)}{V(T)} = \frac{\text{Cov}(T; Y)}{\sigma^2(T)} \\ b = \bar{y} - a\bar{t} \end{cases} .$$

On supposera que  $T$  prend les  $n$  valeurs : 1 ; 2 ; ... ;  $n$ .

Dans le cas de séries chronologiques, on peut alléger les calculs en utilisant les formules

suivantes : 
$$\begin{cases} \bar{t} = \frac{1+2+\dots+n}{n} = \frac{n(n+1)}{2n} = \frac{n+1}{2} \\ \sum_{i=1}^n t^2 = \frac{n(n+1)(2n+1)}{6} \end{cases} .$$

Le premier résultat vient de la formule exprimant la somme des termes d'une suite arithmétique et le second peut facilement être démontré par récurrence. Le second

résultat donnera pour la variance :  $V(T) = \sigma^2(T) = \frac{1}{n} \sum_{i=1}^n t^2 - \bar{t}^2 = \frac{(n+1)(2n+1)}{6} - \left(\frac{n+1}{2}\right)^2$ .

**Exemple 7.2****Le trend par la méthode MCO**

Considérons la série suivante donnant le taux mensuel de nuptialité (nombre de mariages pour 1 000 habitants) en France métropolitaine :

<b>Mois</b>	<b>2003</b>	<b>2004</b>	<b>2005</b>	<b>2006</b>
Janvier	1,3	1,40	1,40	1,30
Février	1,9	2,00	1,80	1,60
Mars	2,2	1,70	1,80	1,60
Avril	3,3	3,60	3,60	3,60
Mai	5,5	5,30	4,80	4,70
Juin	10,30	9,40	9,80	9,50
Juillet	8,40	10,10	10,70	10,10
Août	8,50	6,80	7,10	6,60
Septembre	6,30	6,30	6,50	7,10
Octobre	3,10	3,20	3,10	2,40
Novembre	1,80	1,70	1,80	1,60
Décembre	2,20	2,00	2,10	2,00

Source : Insee, département de la Démographie, 2006

À partir du tableau statistique de cette série, on obtient les résultats suivants :

$$\begin{cases} \bar{t} = \frac{n+1}{2} = \frac{49}{2} = 24,5 \\ \bar{y} = \frac{214,9}{48} = 4,4771 \\ \sum_{t=1}^n t^2 = \frac{n(n+1)(2n+1)}{6} = \frac{48 \times 49 \times 97}{6} = 38\,024 \end{cases}$$

D'où :

$$\begin{cases} Cov(T; Y) = \frac{1}{n} \sum_{t=1}^n t y_t - \bar{t} \times \bar{y} = \frac{5\,295,8}{48} - 24,5 \times 4,4771 = 0,6406 \\ V(T) = \frac{\sum_{t=1}^n t^2}{n} - (\bar{t})^2 = \frac{38\,024}{48} - (24,5)^2 = 191,92 \end{cases}$$

Il reste à calculer a et b :

$$\begin{cases} a = \frac{Cov(T; Y)}{V(T)} = \frac{0,6406}{191,92} = 0,0033 \\ b = \bar{y} - a\bar{t} = 4,4771 - 0,0033 \times 24,5 = 4,3953 \end{cases}$$

On obtient finalement la tendance donnée par l'équation :  $y = 0,0033 \times t + 4,3953$ .

Il est important de signaler que si la droite occupe une place privilégiée dans l'ajustement analytique, d'autres modèles sont incontournables, notamment la courbe de Gompertz<sup>1</sup>, utilisée entre autres pour les tables de mortalité (voir chapitre 6, exercice 4), et la courbe logistique<sup>2</sup>, utilisée pour modéliser l'évolution de certaines populations (voir chapitre 6, section 3.1).

Si les fluctuations de la série sont trop importantes, on pourra au préalable les atténuer en utilisant des moyennes adaptées, que nous allons aborder maintenant.

### La méthode des moyennes échelonnées

Afin de lisser les fluctuations, on peut remplacer les données périodiques par leurs moyennes sur plusieurs périodes – par exemple, des moyennes annuelles de données mensuelles. Ces moyennes ne subissent pas l'influence des variations saisonnières et ont l'avantage de minimiser les extrema. La méthode des moyennes échelonnées consiste à remplacer un certain nombre de données consécutives par leur moyenne.

#### Exemple 7.3

#### Le trend par la méthode des moyennes échelonnées

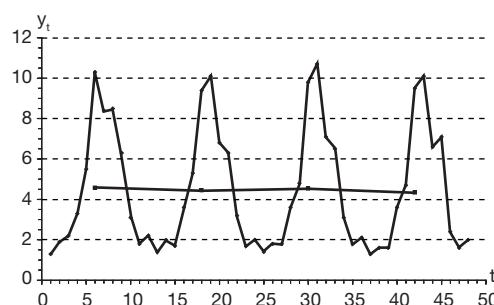
Reprendons la série de l'exemple 7.2. La méthode des moyennes échelonnées consiste à remplacer les données mensuelles par leur moyenne annuelle :

Année	Moyenne échelonnée
2003	4,57
2004	4,46
2005	4,54
2006	4,34

Ces moyennes échelonnées ont été affectées aux dates correspondant au milieu de chaque année, et les quatre points obtenus sont joints à la règle sur la figure 7.7 et donnent un ajustement de la tendance.

**Figure 7.7**

**Moyennes échelonnées (nuptialité).**



La série passe ainsi de 48 données mensuelles, qui varient selon les influences saisonnières, à 4 données annuelles indépendantes de ces variations.

1. Benjamin Gompertz, mathématicien anglais (1779-1865).

2. Découverte par le mathématicien belge Pierre François Verhulst (1804-1849), élève de Quetelet.

Cette méthode fait perdre trop de données, aussi utilisera-t-on plus généralement les moyennes mobiles, qui sont la méthode la plus utilisée dans le lissage des séries chronologiques. Elles permettent de suivre progressivement le phénomène par un système de chevauchement. On distingue en général deux types de moyennes mobiles :

- les moyennes mobiles non centrées ;
- les moyennes mobiles centrées.

### La méthode des moyennes mobiles non centrées

Dans le cas des moyennes mobiles non centrées d'ordre  $p$ , il convient de remplacer une valeur observée,  $y_t$ , par la moyenne arithmétique des  $p$  valeurs antérieures ( $t \geq p$ ), soit

$$\frac{1}{p} \sum_{i=0}^{p-1} y_{t-i}. \text{ On remplace donc } y_p \text{ par } \frac{1}{p} \sum_{i=1}^p y_t, \text{ puis } y_{p+1} \text{ par } \frac{1}{p} \sum_{i=2}^{p+1} y_t, \text{ etc.}$$

#### Définition

On appelle **moyenne mobile non centrée** d'ordre  $p$  à la date  $t$  le nombre noté  $MM_p(t)$  nc et défini par :

$$MM_p(t) \text{ nc} = \frac{1}{p} \sum_{i=1}^p y_t.$$

Les moyennes mobiles non centrées permettent d'exploiter les données récentes.

On notera que les moyennes mobiles non centrées « raccourcissent » la série, car aucune moyenne mobile n'est affectée aux  $(p - 1)$  premières dates.

#### Exemple 7.4

#### Moyennes mobiles non centrées

Prenons comme exemple le cours d'une action (en euros) en Bourse et la recherche d'une stratégie (simple) de décision : acheter en phase de hausse, quand le cours traverse la moyenne mobile de bas en haut, et vendre en phase de baisse, quand le cours traverse la moyenne mobile de haut en bas.

Jour	Cours	Jour	Cours
1	812,5	13	825
2	812,25	14	868,75
3	810	15	881,25
4	806,25	16	868,75
5	793,75	17	862,5
6	787,5	18	875
7	793,75	19	875
8	812,5	20	887,5
9	831,25	21	900
10	837,5	22	910
11	843,75	23	912,5
12	843,75	24	912

La moyenne non centrée d'ordre 4 est la moyenne des quatre valeurs qui précédent la période de calcul. Par exemple, pour le quatrième jour, la moyenne non centrée d'ordre 4 est  $MM4(4)_{nc} = \frac{812,5 + 812,25 + 810 + 806,25}{4} = 810,25$ .

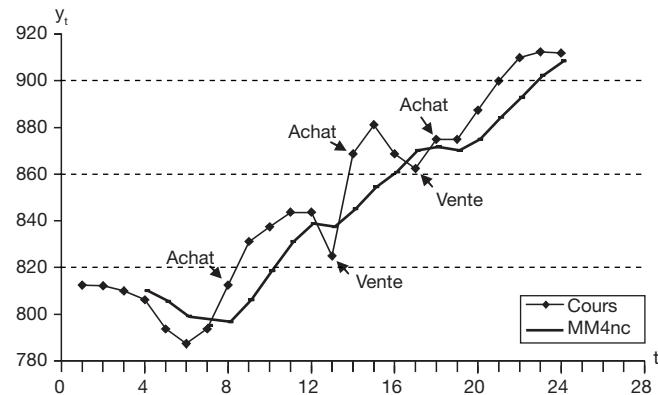
Le tableau de la figure 7.8 donne les moyennes mobiles non centrées d'ordre 4.

**Figure 7.8**  
**Moyennes mobiles non centrées.**

	A	B	C
1	Jour	Cours	MM4 nc
2	1	812,5	
3	2	812,25	
4	3	810	
5	4	806,25	810,25
6	5	793,75	805,56
7	6	787,5	799,38
8	7	793,75	795,31
9	8	812,5	796,88
10	9	831,25	806,25
11	10	837,5	818,75
12	11	843,75	831,25
13	12	843,75	839,06
14	13	825	837,50
15	14	868,75	845,31
16	15	881,25	854,69
17	16	868,75	860,94
18	17	862,5	870,31
19	18	875	871,88
20	19	875	870,31
21	20	887,5	875,00
22	21	900	884,38
23	22	910	893,13
24	23	912,5	902,50
25	24	912	908,63

La figure 7.9 est la traduction graphique de ce tableau qui permet de visualiser l'application de la décision d'achat et de vente des actions.

**Figure 7.9**  
**Moyennes mobiles non centrées du cours de Bourse.**



### La méthode des moyennes mobiles centrées

Dans le cas des moyennes mobiles centrées d'ordre  $p$ , il s'agit de remplacer une valeur observée,  $y_t$ , par la moyenne arithmétique de  $p$  valeurs centrées autour de  $y_t$ .

**Définition**

On appelle **moyenne mobile centrée** d'ordre  $p$  à la date  $t$  le nombre noté  $MMp(t)$  et défini par :

- si  $p$  est impair, soit  $p = 2k + 1$  :

$$MMp(t) = \frac{1}{p} (y_{t-k} + y_{t-k+1} + \dots + y_{t-1} + y_t + y_{t+1} + \dots + y_{t+k}),$$

$$\text{soit } MMp(t) = \frac{1}{p} \sum_{i=-k}^k y_{t+i};$$

- si  $p$  est pair, soit  $p = 2k$  :

$$MMp(t) = \frac{1}{p} (0,5y_{t-k} + y_{t-k+1} + \dots + y_{t-1} + y_t + y_{t+1} + \dots + 0,5y_{t+k}),$$

$$\text{soit } MMp(t) = \frac{1}{p} \left( 0,5y_{t-k} + \sum_{i=-k+1}^{k-1} y_{t+i} + 0,5y_{t+k} \right).$$

**Le cas des moyennes mobiles d'ordre impair** : posons  $p = 2k + 1$  ; dans ce cas tout indice  $t$  ( $t \geq (p + 1) / 2$ ) est la médiane d'une série de  $p$  dates et l'on remplace  $y_t$  par :

$\frac{1}{p} \sum_{i=-k}^k y_{t+i}$ , en prenant la moyenne arithmétique des  $p$  observations obtenues en réunissant les  $k$  observations immédiatement antérieures à  $y_t$ ,  $y_t$  et les  $k$  observations qui succèdent à  $y_t$ .

On notera que les moyennes mobiles centrées « raccourcissent » la série, car aucune moyenne mobile n'est affectée ni aux  $(p - 1)$  premières dates ni aux  $(p - 1)$  dernières dates.

**Exemple 7.5****Moyennes mobiles centrées d'ordre 3 (MM3)**

Considérons le tableau suivant donnant le cours journalier du baril de pétrole sur une période de 14 jours et recherchons le trend par la méthode des moyennes mobiles centrées d'ordre 3 (MM3).

Date	Cours (en US dollars)
29/10/2007	86,05
30/10/2007	85,69
31/10/2007	84,84
01/11/2007	87,61
02/11/2007	87,57
05/11/2007	88,13
06/11/2007	89,13
08/11/2007	90,71
09/11/2007	89,71
12/11/2007	88,8

Date	Cours (en US dollars)
14/11/2007	86,57
15/11/2007	87,01

Source : OPEC, novembre 2007

La moyenne centrée d'ordre 3 est la moyenne des trois valeurs qui entourent la valeur de la période de calcul, y compris elle-même. Par exemple, pour la deuxième date, la moyenne centrée d'ordre 3 est  $MM3(2) = \frac{86,05 + 85,69 + 84,84}{3} = 85,53$ .

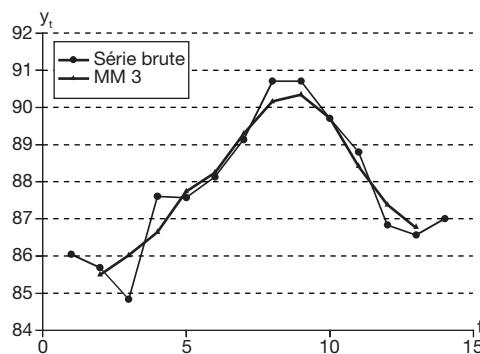
Le tableau de la figure 7.10 donne les moyennes mobiles centrées d'ordre 3.

**Figure 7.10**  
**Moyennes mobiles centrées d'ordre 3.**

	A	B	C
1	t	$y_t$	MM3
2	1	86,05	
3	2	85,69	85,53
4	3	84,84	86,05
5	4	87,61	86,67
6	5	87,57	87,77
7	6	88,13	88,28
8	7	89,13	89,32
9	8	90,71	90,18
10	9	90,71	90,38
11	10	89,71	89,74
12	11	88,8	88,45
13	12	86,84	87,40
14	13	86,57	86,81
15	14	87,01	

La figure 7.11 représente la série brute et la série lissée par les MM3.

**Figure 7.11**  
**Série brute et MM3.**



**Moyennes mobiles d'ordre p pair :** posons  $p = 2k$ . Dans ce cas une série de p dates n'admet pas de médiane, mais un intervalle médian. La règle adoptée consiste à prendre arbitrairement pour médiane la moyenne arithmétique des bornes de l'intervalle médian.

Prenons par exemple  $p = 4$ . Si l'on remplace  $y_1, y_2, y_3$  et  $y_4$  par leur moyenne arithmétique, on devra affecter cette valeur à la date 2,5 (pour centrer), ce qui n'est pas satisfaisant ; de même,  $y_2, y_3, y_4$  et  $y_5$  seraient remplacées par leur moyenne arithmétique affectée à la date 3,5.



Pour éviter cela, la méthode de calcul consiste à affecter à la date 3 la moyenne arithmétique des deux moyennes centrées qui l'encadrent :  $y_{2,5} = \frac{y_1 + y_2 + y_3 + y_4}{4}$  et

$y_{3,5} = \frac{y_2 + y_3 + y_4 + y_5}{4}$ . Ce qui donne :

$$\frac{y_{2,5} + y_{3,5}}{2} = \frac{y_1 + y_2 + y_3 + y_4 + y_2 + y_3 + y_4 + y_5}{8} = \frac{0,5 y_1 + y_2 + y_3 + y_4 + 0,5 y_5}{4}.$$

Finalement, pour former la première moyenne mobile centrée d'ordre 4, on utilise les 5 premières observations et l'on affecte à la date 3 leur moyenne arithmétique pondérée, en affectant aux valeurs extrêmes (la première et la cinquième) le coefficient 0,5 et aux trois valeurs centrales le coefficient 1.

On notera que les moyennes mobiles centrées n'autorisent pas d'estimation d'une valeur théorique, car elles sont subordonnées à la connaissance d'observations postérieures.

La série des moyennes mobiles comporte moins de termes que la série brute.

La série des moyennes mobiles est très inerte du fait qu'une brusque variation n'est retenue que pour 1 / p<sup>ème</sup> de sa valeur brute, les oscillations étant étalées sur les dates antérieures et postérieures.

En général, on choisira l'ordre des moyennes mobiles suivant la périodicité des données : MM7 pour des données journalières (7 jours de la semaine), MM4 pour des données trimestrielles (4 trimestres dans l'année), etc.

## 2 Agrégation des composantes

Nous avons défini précédemment les différentes composantes d'une série chronologique, nous devons maintenant nous intéresser à leur mode de composition et présenter les deux hypothèses que l'on fait habituellement : le schéma additif et le schéma multiplicatif.

### 2.1 PRÉSENTATION DES MODÈLES

---

Deux types de situations coexistent dans le cadre des séries temporelles :

- le modèle additif ;
- le modèle multiplicatif.

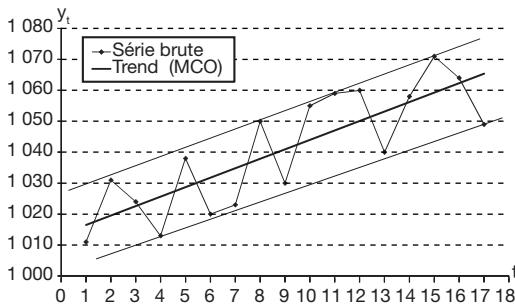
#### Modèle additif et modèle multiplicatif

Nous avons souligné dès le départ l'importance d'une représentation graphique dans l'analyse des séries chronologiques. Ces graphiques permettent de visualiser les deux types de situations.

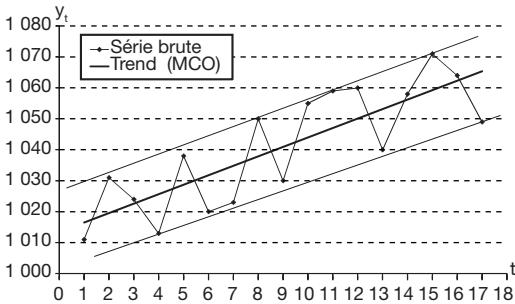
- Dans le cas du modèle additif, les fluctuations sont d'amplitude constante autour du trend, ce qui se traduit par un nuage de points limité par deux parallèles à la droite de tendance (voir figure 7.12).
- Dans le cas du modèle multiplicatif, les fluctuations sont d'amplitudes liées à la valeur du trend, ce qui se traduit par un nuage de points situés entre deux droites

concourantes (entonnoir). Les rapports entre les valeurs observées et les valeurs du trend sont pratiquement identiques d'une période à l'autre, ce qui représente des écarts égaux en pourcentage (voir figure 7.13).

**Figure 7.12**  
**Schéma additif**  
**(aspect d'un tube).**



**Figure 7.13**  
**Schéma multiplicatif**  
**(aspect conique).**



Il arrive que les choix ne soient pas aussi clairs et que l'on hésite entre les deux modèles qui pourront dans ce cas donner des valeurs proches.

Les deux modèles supposent que la composante saisonnière est parfaitement périodique<sup>1</sup>, qu'à l'intérieur d'une année le phénomène saisonnier est neutre, les variations saisonnières se compensant :

- dans le schéma additif, la moyenne des coefficients saisonniers est nulle sur une année ;
- dans le schéma multiplicatif, le produit des coefficients saisonniers est égal à 1 sur une année.

Cette convention est appelée principe de conservation des aires, les aires représentant les fluctuations saisonnières autour du mouvement général, qui se compensent.

Par ailleurs, le mouvement accidentel est supposé faible et de moyenne nulle sur quelques mois.

Pour mettre en évidence les composantes saisonnières et accidentelles, nous devrons distinguer les deux modèles.

### Composante saisonnière

La composante saisonnière est une fonction périodique, de période  $p$ , déterminée par la donnée de  $p$  coefficients saisonniers que nous noterons  $S_1, S_2, \dots, S_p$  et qui vérifient

1. La décomposition d'une fonction en sommes de termes périodiques à l'aide de fonctions sinusoïdales a été établie par le mathématicien Jean-Baptiste Fourier (1768-1813) dans ses travaux sur la chaleur.

$S_i = S_{i+p}$ . L'entier  $p$  détermine la période et on aura  $p = 12$  pour des données mensuelles,  $p = 4$  pour des données trimestrielles, etc.

### Définition

Soit  $p$  la période, les entiers  $i$ , pour  $i \in \{1; 2; \dots; p\}$ , définissent les **saisons** de la série. Les dates relatives à la saison  $i$  sont alors les dates définies par  $t = i + np$  ( $n$  entier naturel).

Si par exemple les données sont trimestrielles, on a quatre saisons que l'on nommera  $T_1$ ,  $T_2$ ,  $T_3$  et  $T_4$ . Les dates relatives à  $T_1$  sont les dates du type  $t = 1 + 4n$ , soit  $1; 5; 9$ ; etc.

La série CVS (corrigée des variations saisonnières), encore appelée série désaisonnalisée, est obtenue en éliminant les influences saisonnières. Cette série est fondamentale et utilisée constamment par l'Insee, qui donne par exemple les chiffres du chômage en « données CVS en fin de mois ». La série corrigée des variations saisonnières peut révéler des résultats paradoxaux, le chômage pouvant diminuer en données brutes un certain mois, et en fait augmenter en données corrigées des variations saisonnières.

## 2.2 SÉRIE CORRIGÉE DES VARIATIONS SAISONNIÈRES DU MODÈLE ADDITIF

Ce modèle se traduit par :  $y_t = f_t + S_t + \varepsilon_t$ . On va donc définir la série CVS en négligeant dans un premier temps la composante accidentelle supposée faible et par définition non prévisible.

Le principe de la neutralité « additive » de la composante périodique sur une période se

traduira par la propriété :  $\sum_{i=1}^p S_i = 0$ .

### Mise en évidence de la composante saisonnière

La prise en compte de la composante saisonnière passe par quatre étapes :

1. On calcule pour chaque date le coefficient  $s_t = y_t - f_t$ , appelé écart saisonnier , qui représente la différence entre la donnée brute  $y_t$  et la tendance déterminée soit par la méthode MCO soit par les moyennes mobiles.
2. On estime les coefficients saisonniers,  $S_i$ , par la moyenne arithmétique des écarts saisonniers  $s_t$  correspondant à la même saison ; si on dispose de données sur  $n$  périodes, donc de  $np$  dates, le coefficient saisonnier  $S_i$  correspondant à la saison  $i$  ( $i \in \{1; 2; \dots; p\}$ ) sera donné par :  $S_i = \frac{1}{n} \sum_{k=0}^{n-1} s_{i+kp}$ .

Si l'on dispose, par exemple, de données mensuelles sur 3 années, on obtiendra le

coefficient saisonnier de janvier par :  $S_1 = \frac{1}{3} \sum_{k=0}^2 s_{1+12k} = \frac{1}{3} (s_1 + s_{13} + s_{25})$ .

3. On contrôle que  $\sum_{i=1}^p S_i = 0$ . Si cette somme est significativement différente de zéro, on introduit des coefficients saisonniers corrigés selon l'étape 4.
4. On note  $m$  la moyenne arithmétique des  $S_i$ , soit  $m = \frac{1}{p} \sum_{i=1}^p S_i$ , et on introduit les coefficients saisonniers corrigés définis par :  $S'_i = S_i - m$  ; on aura alors  $\sum_{i=1}^p S'_i = 0$ .

## Série corrigée des variations saisonnières

La série CVS contient la composante tendancielle et la composante accidentelle.

### Définitions

**La série corrigée des variations saisonnières** est la série obtenue à partir de la série brute en éliminant la composante saisonnière. Dans le schéma **additif**, on aura donc :  $Y_{cvs} = Y - S$ , soit pour toute date  $t$ ,  $i$  désignant la saison relative à la date  $t$  :  $y_{cvs}(t) = y_t - S_i$ , dans le cas où  $\sum_{i=1}^P S_i = 0$ .

$y_{cvs}(t) = y_t - S'_i$  en utilisant les coefficients saisonniers corrigés dans le cas où  $\sum_{i=1}^P S_i \neq 0$ .

On peut alors isoler la **composante accidentelle** en calculant les termes  $\varepsilon_t$ , en éliminant la tendance de la série CVS :  $\varepsilon_t = y_{cvs}(t) - f_t$ .

### Exemple 7.6

#### Série corrigée des variations saisonnières (schéma additif)

Reprendons la série trimestrielle de l'exemple 7.1. Le graphique permet de conjecturer l'hypothèse d'un modèle additif. Déterminons la série CVS (voir figure 7.14) en utilisant le trend déterminé par la méthode MCO, c'est-à-dire  $f_t = 0,6654t + 103,92$  (voir figure 7.6).

Après avoir déterminé les valeurs du trend par la formule  $f_t = 0,6654t + 103,92$  dans la colonne C, on a calculé les coefficients  $s_t$  (colonne D), puis les coefficients saisonniers  $S_p$ , avec par exemple  $S_1 = \frac{s_1 + s_5 + s_9}{3} = 1,72$ .

$S_1$	1,72
$S_2$	-0,78
$S_3$	-3,48
$S_4$	2,52
<b>Total</b>	-0,01

La somme des coefficients est très proche de zéro, il est donc inutile de les corriger. On a fait figurer la série CVS en colonne E. Enfin, on a calculé la composante accidentelle en colonne F.

Figure 7.14

**Détermination de la série CVS et de la composante accidentelle.**

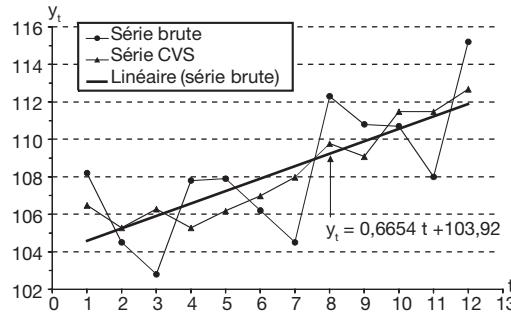
	A	B	C	D	E	F
1	t	$y_t$	$f_t$	$s_t - y_t - f_t$	$y_t - S_i - y_{cvs}(t)$	$\varepsilon_t = y_{cvs} - f_t$
2	1	108,20	104,59	3,61	106,48	1,89
3	2	104,50	106,25	-0,75	106,28	0,03
4	3	102,80	105,92	-3,12	106,28	0,36
5	4	107,90	106,58	1,22	105,28	-1,31
6	5	107,90	107,25	0,65	106,18	-1,07
7	6	106,20	107,91	-1,71	106,98	-0,93
8	7	104,50	108,58	-4,08	107,98	-0,60
9	8	112,30	109,24	3,06	109,78	0,53
10	9	110,80	109,91	0,89	109,08	-0,83
11	10	110,70	110,57	0,13	111,48	0,91
12	11	108,00	111,24	-3,24	111,48	0,24
13	12	115,20	111,90	3,30	112,68	0,77



On a représenté figure 7.15 la série des indices IAA, avec la série CVS et le trend linéaire.

**Figure 7.15**

**Série brute, série CVS et trend des indices IAA.**



## 2.3 SÉRIE CORRIGÉE DES VARIATIONS SAISONNIÈRES DU MODÈLE MULTIPLICATIF

Ce modèle se traduit par :  $y_t = f_t \times S_t \times \varepsilon_t$ . On définit la série CVS selon la même procédure que dans le cas additif. La composante saisonnière est une fonction périodique, de période  $p$ , déterminée par la donnée de  $p$  coefficients saisonniers que nous noterons  $S_1$ ,

$$\begin{cases} \frac{1}{p} \sum_{i=1}^p S_i = 1 \\ S_i = S_{i+p} \end{cases}, \text{ la première propriété traduisant le principe de la neutralité « multiplicative » (moyenne arithmétique des coefficients égale à 1) de cette composante sur une période et la seconde le fait que ces coefficients sont périodiques.}$$

On remarque que le modèle multiplicatif peut se ramener au modèle additif en passant aux logarithmes décimaux :  $\log y_t = \log f_t + \log S_t + \log \varepsilon_t$ .

### Mise en évidence de la composante saisonnière

La prise en compte de la composante saisonnière passe par quatre étapes :

1. On calcule pour chaque date le coefficient  $s_t = y_t / f_t$ , appelé rapport saisonnier, qui représente le rapport entre la donnée brute  $y_t$  et la tendance déterminée soit par la méthode MCO soit par les moyennes mobiles. Ce rapport saisonnier est le coefficient multiplicateur qui permet à la date  $t$  de passer de la tendance à la série brute.
2. On estime les coefficients saisonniers,  $S_p$ , par la moyenne arithmétique des rapports saisonniers  $s_t$  correspondant à la même saison ; si on dispose de données sur  $n$  périodes, donc de  $np$  dates, le coefficient saisonnier  $S_i$  correspondant à la saison  $i$  ( $i \in \{1; 2; \dots; p\}$ ) sera donné par :  $S_i = \frac{1}{n} \sum_{k=0}^{n-1} s_{i+kp}$ .

3. On contrôle que la contrainte de neutralité multiplicative  $\frac{1}{p} \sum_{i=1}^p S_i = 1$  est vérifiée. Si

cette somme est significativement différente de 1, on introduit des coefficients saisonniers corrigés selon l'étape 4.

4. Soit  $m$  la moyenne arithmétique des  $S_i$ ,  $m = \frac{1}{p} \sum_{i=1}^p S_i$ , les coefficients saisonniers corrigés sont  $S'_i = S_i / m$ ; on aura alors  $\frac{1}{p} \sum_{i=1}^p S'_i = 1$ .

### Série corrigée des variations saisonnières

La série CVS contient la composante tendancielle et la composante accidentelle.

#### Définitions

**La série corrigée des variations saisonnières** est la série obtenue à partir de la série brute en éliminant la composante saisonnière. Dans le schéma **multiplicatif**, on aura donc :  $Y_{\text{cvs}} = Y / S$ , soit pour toute date  $t$ ,  $i$  désignant la saison relative à la date  $t$  :

- $y_{\text{cvs}}(t) = y_t / S_i$ , dans le cas où  $\frac{1}{p} \sum_{i=1}^p S_i = 1$  ;
- $y_{\text{cvs}}(t) = y_t / S'_i$  en utilisant les coefficients saisonniers corrigés dans le cas où  $\frac{1}{p} \sum_{i=1}^p S'_i \neq 1$ .

On peut alors isoler la **composante accidentelle** en calculant les termes  $\varepsilon_t$  en éliminant la tendance de la série CVS :  $\varepsilon_t = y_{\text{cvs}}(t) / f_t$ .

## 2.4 PRÉVISIONS

---

### Modèles de prévision

« L'étude du passé sert à anticiper le futur et la prévision économique n'est pas autre chose, en grande partie du moins, que ce qu'on appelle en langage mathématique l'extrapolation des événements passés, des conjonctions passées<sup>1</sup>. »

À partir des méthodes exposées précédemment, l'analyste se situe à la période  $T$  et souhaite effectuer une prévision à l'horizon  $h$ . On envisagera uniquement le cas d'une prévision ponctuelle, c'est-à-dire de la recherche d'une valeur unique qui représente la meilleure estimation possible de la valeur future inconnue  $y_{T+h}$  à partir de la donnée ( $y_1$ ;  $y_2$ ; ...;  $y_T$ ). Cette estimation est notée  $\hat{y}_T(h)$ , ou encore  $\hat{y}_t$  avec  $t = T + h$ ,  $T$  représentant l'origine de la prévision. On supposera que l'on dispose d'une tendance linéaire, alors la prévision ponctuelle pourra être faite en utilisant les coefficients saisonniers en addition dans le modèle additif et en multiplication dans le modèle multiplicatif, ce qui donnera, à partir du trend linéaire noté  $f(t) = at + b$ ,  $S'_i$  désignant le coefficient saisonnier corrigé relatif à la date  $t = T + h$  :

- schéma additif:  $\hat{y}_T(h) = a(T+h) + b + S'_i$ , ou encore  $\hat{y}_t = at + b + S'_i$  ;
- schéma multiplicatif:  $\hat{y}_T(h) = (a(T+h) + b) \times S'_i$ , ou encore  $\hat{y}_t = (at + b) \times S'_i$ .

---

1. H. Guitton, *Statistique et économétrie*, Dalloz, 1959.

## Série ajustée

On définit la série ajustée sur le modèle de la série prévisionnelle exposée ci-avant. On notera alors pour les dates  $t$ , pour lesquelles on connaît la série brute ( $S'_i$  désignant le coefficient saisonnier corrigé relatif à la date  $t$ ) :

- schéma additif :  $\hat{y}_t = at + b + S'_i$  ;
- schéma multiplicatif :  $\hat{y}_t = (at + b) \times S'_i$ .

### Exemple 7.7

#### Prévision (schéma additif)

Reprendons les données de l'exemple 7.1.

On a déterminé le trend par la méthode MCO, et on a trouvé  $f_t = 0,6654 t + 103,92$  (voir figure 7.6) ; la prévision ponctuelle sera donnée par :  $\hat{y}_t = at + b + S'_i$ ,  $S_i$  désignant le coefficient saisonnier relatif à la date  $t = T + h$ . On aura donc :

$$\hat{y}_t = 0,6654t + 103,92 + \begin{pmatrix} 1,72 \\ -0,78 \\ -3,48 \\ 2,52 \end{pmatrix}, \text{ en choisissant le coefficient saisonnier relatif à la}$$

date  $t = T + h$  ; recherchons par exemple une prévision ponctuelle pour le deuxième trimestre 2007, soit  $T = 12$  et  $h = 2$ , d'où  $t = T + h = 12 + 2 = 14$ .

Dans ce cas,  $\hat{y}_{12}(2) = \hat{y}_{14} = 0,6654 \times 14 + 103,92 - 0,78 = 112,46$ .

## Conclusion

On notera que l'on devra rester très prudent pour les extrapolations, car on peut se retrouver face à un retournement de tendance ou à des changements dans les fluctuations périodiques. Si  $h > 1$ , on pourra tester la qualité du modèle, en utilisant les premières observations de la période  $T + 1$  devenues disponibles et en les comparant aux prévisions qu'elles n'ont pas contribué à déterminer. Cette confrontation de prévisions fondées sur le passé et de valeurs actuelles est très précieuse pour valider l'estimation.

Pour conclure cette introduction aux séries chronologiques, nous devons signaler que nous n'avons abordé que l'aspect déterministe et que nous avons laissé de côté l'aspect aléatoire, que nous avons simplement notifié à l'occasion de la composante accidentelle. Nous n'avons pas abordé les modèles autorégressifs<sup>1</sup>, qui traduisent une caractéristique particulière des séries chronologiques, la corrélation entre les termes, c'est-à-dire la dépendance statistique du présent et du passé, et le lecteur pourra consulter de nombreux ouvrages complémentaires (notamment l'ouvrage d'économétrie d'Éric Dor).

En résumé, à l'issue de ce chapitre, le lecteur doit connaître les deux modèles de décomposition d'une série chronologique, savoir utiliser la méthode MCO et les différentes moyennes mobiles pour mettre en évidence le trend et les différentes composantes. Ces techniques doivent permettre d'expliciter la série corrigée des variations saisonnières et d'aborder l'aspect prévisionnel.

1. L'article de référence en la matière est dû au statisticien George Udny Yule (1871-1951).

# Problèmes et exercices

L'analyse des séries temporelles est un prolongement de l'analyse de régression puisqu'il s'agit d'expliquer un phénomène selon le temps. Pour cela, quatre modes d'application des séries temporelles coexistent selon les combinaisons effectuées entre méthodes empirique et analytique et modèles additif et multiplicatif :

- l'exercice 1 combine la méthode empirique avec le modèle additif ;
- l'exercice 2 associe la méthode empirique et le modèle multiplicatif ;
- l'exercice 3 allie méthode analytique et modèle additif ;
- l'exercice 4 met en œuvre la méthode analytique avec le modèle multiplicatif.

## EXERCICE 1 MÉTHODE EMPIRIQUE ET MODÈLE ADDITIF

### Énoncé

Le tableau ci-après indique les entrées par quadrimestres (durée de quatre mois), en millions, dans les salles de cinéma en France :

Quadrimestre	2003	2004	2005	2006*
1	61,33	67,86	61,04	72,58
2	48,16	65,3	53,06	55,21
3	63,97	62,17	61,23	60,66

\* Données provisoires - Source : [www.cnc.fr](http://www.cnc.fr), 2007

1. Représentez graphiquement cette série chronologique et déterminez sa saisonnalité.
2. En utilisant le modèle empirique additif :
  - a. Calculez les coefficients saisonniers.
  - b. Déterminez la série ajustée.
  - c. Déterminez la série CVS.
3. Représentez sur un même graphique la série brute, la tendance et la série CVS.

### Solution

1. La première étape consiste à présenter le tableau de données sous la forme d'un tableau statistique indiquant les valeurs de  $t$ , le temps, et de  $Y_t$ , valeur des entrées en période  $t$  (voir figure 7.16).

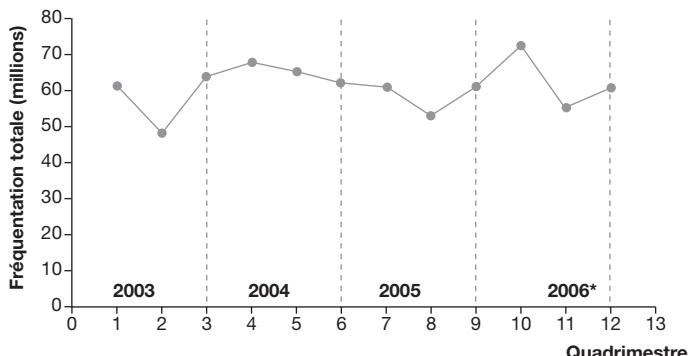
Afin de représenter graphiquement cette série chronologique, il convient de tracer la courbe avec le temps,  $t$ , en abscisses, et la valeur des entrées,  $Y_t$ , en ordonnées.

**Figure 7.16**  
Résultats sous Excel.

	A	B	C	D	E	F	G	H	I	J	K
1	Année	Quadrimestre	t	$Y_t$	$MM_3(t)$	$t$	s <sub>t</sub>	$s_{1t}$	$s_{12t}$	$\hat{Y}_t$	$Y_{cvs(t)}$
2	2003	1	1	61,33				4,67	4,73	66,60	
3	2003	2	2	48,16	57,82	57,82	-9,66	-5,62	-5,55	52,27	63,71
4	2003	3	3	63,97	60,00	60,00	3,97	0,75	0,81	60,81	63,16
5	2004	1	4	67,86	66,71	65,71	2,15	4,67	4,73	70,44	63,13
6	2004	2	5	65,3	65,11	65,11	0,19	-5,62	-5,55	59,56	70,85
7	2004	3	6	62,17	62,84	62,84	-0,67	0,75	0,81	63,65	61,36
8	2005	1	7	61,04	58,76	58,76	2,28	4,67	4,73	63,49	56,31
9	2005	2	8	53,06	58,44	58,44	-5,38	-5,62	-5,55	52,89	58,61
10	2005	3	9	61,23	62,29	62,29	-1,06	0,75	0,81	63,10	60,42
11	2006*	1	10	72,58	63,01	63,01	9,57	4,67	4,73	67,74	67,85
12	2006*	2	11	55,21	62,82	62,82	-7,61	-5,62	-5,55	57,27	60,76
13	2006*	3	12	60,66				0,75	0,81		59,85

Pour représenter une courbe sous Excel, cliquez sur Insertion/Graphique dans la barre de menus, puis, dans l'assistant graphique, choisissez le type de graphique Nuage de points, puis, dans Sous-type de graphique, sélectionnez l'image représentant le nuage de points reliés par une courbe. Cliquez sur Suivant et indiquez dans le champ correspondant la plage où se trouvent les données (voir figure 7.17).

**Figure 7.17**  
Fréquentation des salles de cinéma – France.



La saisonnalité des entrées cinématographiques en France est annuelle. La structure des entrées subit un creux au deuxième quadrimestre, pour remonter au troisième quadrimestre, à l'exception de l'année 2004, pour laquelle les ventes continuent de chuter.

**2. a.** Pour déterminer les coefficients saisonniers, il est nécessaire de calculer la tendance. Dans le cadre de la méthode empirique, la tendance est déterminée par des moyennes mobiles. Puisque la saisonnalité est annuelle, composée de trois quadrimestres, les moyennes mobiles adaptées sont les moyennes mobiles d'ordre 3.

La première moyenne mobile calculable est  $MM_3(2)$ . Explicitons les premiers calculs :

$$MM_3(2) = \frac{Y_1 + Y_2 + Y_3}{3} = \frac{61,33 + 48,16 + 63,97}{3}, \quad \text{soit} \quad MM_3(2) = 57,82 ;$$

$$MM_3(3) = \frac{Y_2 + Y_3 + Y_4}{3} = \frac{48,16 + 63,97 + 67,86}{3}, \quad \text{soit} \quad MM_3(3) = 60,00 .$$

La dernière moyenne mobile calculable est  $MM_3(11)$ . Les moyennes mobiles figurent dans la colonne E du tableau de la figure 7.16.

À la suite de ces calculs, les écarts saisonniers peuvent être calculés, selon le modèle additif.  $s_2 = Y_2 - MM_3(2)_2 = 48,16 - 57,82$ , soit  $s_2 = -9,66$ . Les écarts  $s_1$  et  $s_{12}$  ne sont pas calculables. On trouvera dans la colonne G du tableau de la figure 7.16 les écarts saisonniers.

Les coefficients saisonniers sont ensuite calculés en effectuant pour chaque saison (quadrimestre) la moyenne arithmétique des écarts saisonniers disponibles :

$$S_1 = \frac{s_4 + s_7 + s_{10}}{3} = \frac{2,15 + 2,28 + 9,57}{3}, \text{ soit } S_1 = 4,67 ;$$

$$S_2 = \frac{s_2 + s_5 + s_8 + s_{11}}{4} = \frac{-9,66 + 0,19 - 5,38 - 7,61}{4}, \text{ soit } S_2 = -5,62 ;$$

$$S_3 = \frac{s_3 + s_6 + s_9}{3} = \frac{3,97 - 0,67 - 1,06}{3}, \text{ soit } S_3 = 0,75.$$

On rappelle que les coefficients saisonniers sont périodiques et que, dans cet exercice, la période est de 3 : on a donc calculé  $S_1$ ,  $S_2$  et  $S_3$ .

On calcule ensuite la moyenne  $m$  des coefficients saisonniers pour effectuer, si leur moyenne n'est pas nulle, la correction nécessaire au respect de la compensation :

$$m = \frac{S_1 + S_2 + S_3}{3} = \frac{4,67 - 5,62 + 0,75}{3}, \text{ soit } m = -0,07.$$

D'où les coefficients saisonniers corrigés,  $S'_1 = S_1 - m = 4,67 + 0,07$ , soit  $S'_1 = 4,73$ . De même,  $S'_2 = -5,55$  et  $S'_3 = 0,81$ .

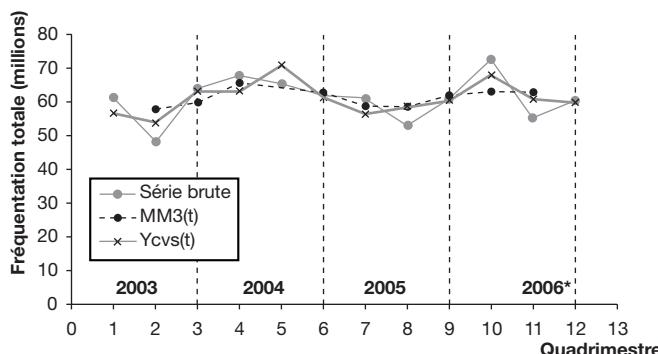
Les calculs sont détaillés dans les colonnes H et I de la figure 7.16.

**b.** Pour le modèle additif, la série ajustée est  $\hat{Y}_t = f_t + S'_t$ , d'où  $\hat{Y}_2 = MM_3(2) + S'_2 = 57,82 - 5,55$ , soit  $\hat{Y}_2 = 52,27$  ;  $\hat{Y}_3 = T_3 + S'_3 = 60,00 + 0,81$ , soit  $\hat{Y}_3 = 60,81$ . De même,  $\hat{Y}_4 = 70,44$  ;  $\hat{Y}_5 = 59,56$  ;  $\hat{Y}_6 = 63,65$  ;  $\hat{Y}_7 = 63,49$  ;  $\hat{Y}_8 = 52,89$  ;  $\hat{Y}_9 = 63,10$  ;  $\hat{Y}_{10} = 67,74$  ;  $\hat{Y}_{11} = 57,27$ .  $\hat{Y}_{12}$  est indéterminé, pour la même raison que  $\hat{Y}_1$ . Ces calculs sont détaillés à la suite des calculs précédents, dans la figure 7.16.

**c.** La série CVS est différente de la série ajustée, car elle inclut les aléas. Pour le modèle additif, la série CVS est  $Y_{CVS}(t) = Y_t - S'_t$ , d'où  $Y_{CVS}(1) = Y_1 - S'_1 = 61,33 - 4,73$ , soit  $Y_{CVS}(1) = 56,60$ . Ces calculs sont détaillés dans la figure 7.16.

3. Les trois courbes sont tracées sur le même graphique, à partir des données de la figure 7.16, avec le temps, t, en abscisses, et les différentes séries en ordonnées (voir figure 7.18).

**Figure 7.18**  
Fréquentation des salles de cinéma, tendance et série CVS – France.





## EXERCICE 2 MÉTHODE EMPIRIQUE ET MODÈLE MULTIPLICATIF

### Énoncé

Une entreprise de location et vente de matériel de montagne réalise l'essentiel de son chiffre d'affaires sur deux saisons :

- l'hiver, avec le matériel de ski ;
- l'été, avec le matériel de randonnée.

Son chiffre d'affaires (en milliers d'euros) des trois dernières années est indiqué dans le tableau suivant :

Saison	2005	2006	2007
Automne	4,86	4,33	3,11
Hiver	6,52	6,73	7,61
Printemps	5,16	4,41	2,83
Été	6,75	7,01	7,51

1. Représentez graphiquement cette série chronologique et justifiez l'utilisation du modèle multiplicatif.
2. En estimant la tendance de cette série par les moyennes mobiles d'ordre 4 et à l'aide du modèle multiplicatif :
  - a. Calculez les coefficients saisonniers.
  - b. Déterminez la série ajustée.
  - c. Déterminez la série CVS.
3. Représentez sur un même graphique la série brute, la tendance et la série ajustée.

### Solution

1. La première étape consiste à présenter le tableau de données sous la forme d'un tableau statistique indiquant les valeurs de  $t$ , le temps, et de  $Y_t$ , valeur des entrées en période  $t$  :

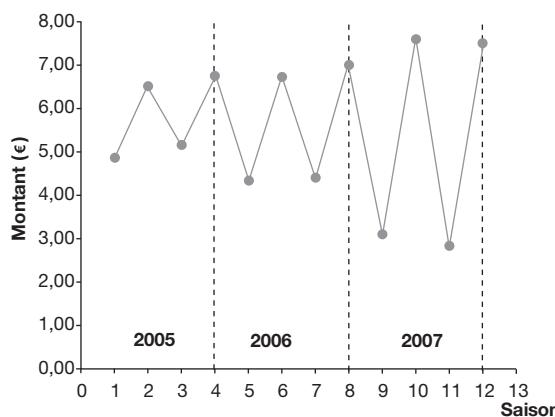
Année	Saison	t	$Y_t$
2005	Automne	1	4,86
2005	Hiver	2	6,52
2005	Printemps	3	5,16
2005	Été	4	6,75
2006	Automne	5	4,33
2006	Hiver	6	6,73
2006	Printemps	7	4,41
2006	Été	8	7,01

Année	Saison	t	$Y_t$
2007	Automne	9	3,11
2007	Hiver	10	7,61
2007	Printemps	11	2,83
2007	Été	12	7,51

Afin de représenter graphiquement cette série chronologique, il convient de tracer la courbe avec le temps, t, en abscisses, et la valeur des entrées,  $Y_t$ , en ordonnées (voir figure 7.19).

Figure 7.19

Ventes par saisons.



Les variations des ventes sont d'amplitudes de plus en plus grandes, le schéma ayant un aspect « conique », ce qui justifie de recourir au modèle multiplicatif.

**2. a.** Pour déterminer les coefficients saisonniers, il est nécessaire de calculer la tendance. Dans le cadre de la méthode empirique, la tendance sera déterminée par des moyennes mobiles centrées d'ordre 4 (une saisonnalité annuelle). La première moyenne mobile calculable est  $MM_4(3)$ , que nous calculons selon la méthode vue dans la partie théorique de ce chapitre sur 5 termes :

$$MM_4(3) = \frac{0,5 \times Y_1 + Y_2 + Y_3 + Y_4 + 0,5 \times Y_5}{4} = \frac{0,5 \times 4,86 + 6,52 + 5,16 + 6,75 + 0,5 \times 4,33}{4},$$

soit  $MM_4(3) = 5,76$ . De même,

$$MM_4(4) = \frac{0,5 \times Y_2 + Y_3 + Y_4 + Y_5 + 0,5 \times Y_6}{4} = \frac{0,5 \times 6,52 + 5,16 + 6,75 + 4,33 + 0,5 \times 6,73}{4},$$

soit  $MM_4(4) = 5,72$ . De même,  $MM_4(5) = 5,65$ ;  $MM_4(6) = 5,59$ ;  $MM_4(7) = 5,47$ ;  $MM_4(8) = 5,43$ ;  $MM_4(9) = 5,34$ ;  $MM_4(10) = 5,20$ . La tendance est ainsi déterminée par les valeurs des moyennes mobiles.

À la suite de ces calculs, les variations saisonnières par période peuvent être calculées selon le modèle multiplicatif. Les rapports saisonniers  $s_1$  et  $s_2$  sont indéterminés;  $s_3 = Y_3 / MM_4(3) = 5,16 / 5,76$ , soit  $s_3 = 0,90$ ;  $s_4 = 6,75 / 5,72$ , soit  $s_4 = 1,18$ . De même,  $s_5 = 0,77$ ;  $s_6 = 1,20$ ;  $s_7 = 0,81$ ;  $s_8 = 1,29$ ;  $s_9 = 0,58$  et  $s_{10} = 1,46$ .

Les coefficients saisonniers sont ensuite calculés :

$$S_1 = \frac{s_5 + s_9}{2}, \text{ car } s_1 \text{ est inconnu, donc } S_1 = \frac{0,77 + 0,58}{2}, \text{ soit } S_1 = 0,68 ;$$

$$S_2 = \frac{s_6 + s_{10}}{2}, \text{ donc } S_2 = \frac{1,20 + 1,46}{2}, \text{ soit } S_2 = 1,33 ;$$

$$S_3 = \frac{s_3 + s_7}{2}, \text{ donc } S_3 = \frac{s_3 + s_7}{2} = \frac{0,90 + 0,81}{2}, \text{ soit } S_3 = 0,86 ;$$

$$S_4 = \frac{s_4 + s_8}{2}, \text{ donc } S_4 = \frac{s_4 + s_8}{2} = \frac{1,18 + 1,29}{2}, \text{ soit } S_4 = 1,24.$$

Notons que le coefficient saisonnier d'un trimestre est le même pour chaque année, d'où  $S_1 = S_5 = S_9$ ;  $S_2 = S_6 = S_{10}$ ;  $S_3 = S_7 = S_{11}$  et  $S_4 = S_8 = S_{12}$ .

En appliquant la correction nécessaire au respect de la compensation entre coefficients saisonniers,  $m = \frac{S_1 + S_2 + S_3 + S_4}{4} = \frac{0,68 + 1,33 + 0,86 + 1,24}{4}$ , soit  $m = 1,03$ .

D'où les coefficients saisonniers corrigés :  $S'_1 = S_1 / m = 0,67 / 1,02$ , soit  $S'_1 = 0,66$ . De même,  $S'_2 = 1,29$ ;  $S'_3 = 0,83$  et  $S'_4 = 1,20$ .

Comme pour les coefficients saisonniers,  $S'_1 = S'_5 = S'_9$ ;  $S'_2 = S'_6 = S'_{10}$ ;  $S'_3 = S'_7 = S'_{11}$  et  $S'_4 = S'_8 = S'_{12}$ .

**b.** Pour le modèle multiplicatif, la série ajustée est  $\hat{Y}_t = MM_4(t) \times S'_t$ , pour  $t$  entier variant de 3 à 10. On a :  $\hat{Y}_3 = MM_4(3) \times S'_3$ , soit  $\hat{Y}_3 = 5,76 \times 0,83$ , soit  $\hat{Y}_3 = 4,78$  ;

$\hat{Y}_4 = MM_4(4) \times S'_4 = 5,72 \times 1,20$ , soit  $\hat{Y}_4 = 6,86$ . De même,  $\hat{Y}_5 = 3,73$ ;  $\hat{Y}_6 = 7,21$ ;  $\hat{Y}_7 = 4,54$ ;  $\hat{Y}_8 = 6,52$ ;  $\hat{Y}_9 = 3,52$ ;  $\hat{Y}_{10} = 6,71$ . Pour effectuer ces calculs à l'aide de la calculatrice, saisissez  $MM_4(t)$  dans la colonne L1, en saisissant la valeur 0 pour les dates 1, 2, 11, 12, et  $S'_j$  dans la colonne L2; placez le curseur sur l'en-tête de colonne L3. Indiquez  $L3=L1 \times L2$ . Puis appuyez sur **ENTER**. La colonne L3 fait alors apparaître la série ajustée (voir figure 7.20).

**c.** La série CVS est différente de la série ajustée car elle inclut les aléas. Pour le modèle multiplicatif, la série CVS est  $Y_{CVS}(t) = Y_t / S'_t$ , d'où  $Y_{CVS}(1) = Y_1 / S'_1 = 4,86 / 0,66$ , soit  $Y_{CVS}(1) = 7,36$ .  $Y_{CVS}(2) = Y_2 / S'_2 = 6,52 / 1,29$ , soit  $Y_{CVS}(2) = 5,05$ . De même,  $Y_{CVS}(3) = 6,22$ ;  $Y_{CVS}(4) = 5,63$ ;  $Y_{CVS}(5) = 6,56$ ;  $Y_{CVS}(6) = 5,22$ ;  $Y_{CVS}(7) = 5,31$ ;  $Y_{CVS}(8) = 5,84$ ;  $Y_{CVS}(9) = 4,71$ ;  $Y_{CVS}(10) = 6,56$ ;  $Y_{CVS}(11) = 3,41$ ;  $Y_{CVS}(12) = 6,26$ .

Pour effectuer ces calculs, à l'aide de la calculatrice, à la suite du tableau précédent, saisissez  $Y_t$  dans la colonne L4, placez le curseur sur l'en-tête de colonne L5. Indiquez  $L5=L4/L2$ . Puis appuyez sur **ENTER**. La colonne L5 fait alors apparaître la série CVS (voir figure 7.21).

**Figure 7.20 (gauche)**

**Calculs de la série ajustée avec la calculatrice.**

L1	L2	L3	3
0	.66	0	
0	1.29	0	
5.76	.83	4.7808	
5.72	1.2	6.864	
5.65	.66	3.729	
5.59	1.29	7.2111	
5.47	.83	4.5401	

$$L3(1)=0$$

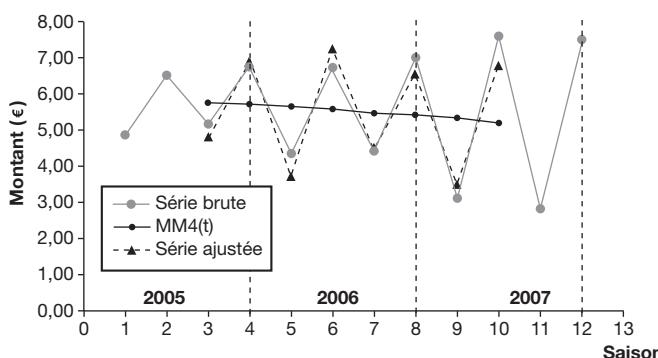
L3	L4	L5	5
0	4.86	5.0543	
0	6.52	6.2169	
4.7808	5.16	5.625	
6.864	6.75	6.5606	
3.729	4.33	5.2171	
7.2111	6.73	5.3133	
4.5401	4.41		

$$L5(1)=7.363636363\dots$$

3. Les trois courbes sont tracées sur le même graphique, à partir du graphique précédemment présenté (voir figure 7.19), avec le temps, t, en abscisses, et les valeurs du chiffre d'affaires,  $Y_t$ ,  $T_t$  et  $\hat{Y}_t$ , en ordonnées (voir figure 7.22).

**Figure 7.22**

**Chiffre d'affaires, tendance et série ajustée.**



### EXERCICE 3 MÉTHODE ANALYTIQUE ET MODÈLE ADDITIF

#### Énoncé

Le tableau ci-après indique les entrées trimestrielles, en millions, dans les salles de cinéma en France :

Trimestre	2004	2005	2006*
1	50,46	45,34	51,63
2	51,46	41,86	51,06
3	41,07	35,14	35
4	52,34	52,99	50,76

\* Données provisoires - Source : [www.cnc.fr](http://www.cnc.fr), 2007

- Déterminez la droite de régression de  $Y_t$  selon le temps.
- À partir de la droite de régression de  $Y_t$  selon le temps et en utilisant le modèle additif :
  - Calculez les coefficients saisonniers.
  - Déterminez la série ajustée.

3. Représentez sur un même graphique la série brute, la tendance obtenue par la droite de régression et la série ajustée.
4. À la suite des calculs précédents, calculez la série CVS.
5. Proposez des prévisions de fréquentations trimestrielles pour l'année 2007.

**Solution**

**1.** La première étape consiste à présenter le tableau de données sous la forme d'un tableau statistique indiquant les valeurs de  $t$ , le temps, et de  $Y_t$ , valeur des entrées en période  $t$ . La droite de régression  $f_t = a \times t + b$  est déterminée par la méthode des MCO vue au chapitre 6. Il convient de déterminer les valeurs de  $a$  et  $b$  dans l'équation  $f_t = a \times t + b$ .

Pour cela, il est nécessaire de calculer les valeurs de  $\bar{t}$ ,  $\bar{y}$ ,  $V(t)$  et  $\sum_{i=1}^n t_i y_i$ .

Les moyennes de  $t$  et de  $Y$  ainsi que la variance de  $t$  peuvent être calculées en utilisant les fonctions d'Excel correspondantes, puisque les données sont des données uniques (avec  $n_i = 1$  quel que soit  $i$ ). Pour cela, il convient d'appeler les fonctions MOYENNE et VAR.P d'Excel (voir annexe 1.1), ou bien de les calculer comme exposé précédemment (voir chapitres 2 et 3). On peut aussi utiliser pour  $t$  les formules spécifiques (voir chapitre 7, section 1.3, la méthode MCO). Ces calculs sont détaillés figure 7.23.

**Figure 7.23****Résultats sous Excel.**

	A	B	C	D	E
1	Année	Trimestre	$t$	$Y_t$	$t \cdot Y_t$
2	2004	1	1	50,46	50,46
3	2004	2	2	51,46	102,92
4	2004	3	3	41,07	123,21
5	2004	4	4	52,34	209,36
6	2005	1	5	45,34	226,70
7	2005	2	6	41,86	251,16
8	2005	3	7	35,14	245,98
9	2005	4	8	52,99	423,92
10	2006*	1	9	51,63	464,67
11	2006*	2	10	51,06	510,60
12	2006*	3	11	35,00	385,00
13	2006*	4	12	50,76	609,12
14		Somme	78,00	559,11	3 603,10
15		Moyenne	6,50	46,59	
16		Variance	11,92		

$$\text{De là, } a = \frac{3603,10 - 12 \times 6,5 \times 46,59}{12 \times 11,92} = -0,218 \text{ et } b = 46,59 + 0,218 \times 6,5 = 48,007,$$

d'où :  $f_t = -0,218 \times t + 48,007$ .

**2. a.** Pour déterminer les coefficients saisonniers, il est nécessaire de calculer la tendance. Dans le cadre de la méthode analytique, ces tendances sont calculées en utilisant l'équation de la droite de régression. Pour  $t = 1$ ,  $f_1 = -0,218 \times 1 + 48,007$ , soit  $f_1 = 47,79$  ;  $f_2 = -0,218 \times 2 + 48,007$ , soit  $f_2 = 47,57$  ; de même,

$f_3 = 47,35$  ;  $f_4 = 47,14$  ;  $f_5 = 46,92$  ;  $f_6 = 46,70$  ;  $f_7 = 46,48$  ;  $f_8 = 46,27$  ;  $f_9 = 46,05$  ;  $f_{10} = 45,83$  ;  $f_{11} = 45,61$  et  $f_{12} = 45,40$ .

À la suite de ces calculs, les écarts saisonniers par période sont, selon le modèle additif :  $s_1 = Y_1 - f_1 = 50,46 - 47,79$ , soit  $s_1 = 2,67$  ;  $s_2 = Y_2 - f_2 = 51,46 - 47,57$ , soit  $s_2 = 3,89$ . De même,  $s_3 = -6,28$  ;  $s_4 = 5,20$  ;  $s_5 = -1,58$  ;  $s_6 = -4,84$  ;  $s_7 = -11,34$  ;  $s_8 = 6,72$  ;  $s_9 = 5,58$  ;  $s_{10} = 5,23$  ;  $s_{11} = -10,61$  ;  $s_{12} = 5,36$ .

Les coefficients saisonniers sont donc :

$$S_1 = \frac{s_1 + s_5 + s_9}{3} = \frac{2,67 - 1,58 + 5,58}{3}, \text{ soit } S_1 = 2,22 ;$$

$$S_2 = \frac{s_2 + s_6 + s_{10}}{3} = \frac{3,89 - 4,84 + 5,23}{3}, \text{ soit } S_2 = 1,43 ;$$

$$S_3 = \frac{s_3 + s_7 + s_{11}}{3} = \frac{-6,28 - 11,34 - 10,61}{3}, \text{ soit } S_3 = -9,41 ;$$

$$S_4 = \frac{s_4 + s_8 + s_{12}}{3} = \frac{5,20 + 6,72 + 5,36}{3}, \text{ soit } S_4 = 5,76.$$

Rappelons que les coefficients saisonniers sont périodiques (période 4, ici), d'où  $S_1 = S_5 = S_9$  ;  $S_2 = S_6 = S_{10}$  ;  $S_3 = S_7 = S_{11}$  et  $S_4 = S_8 = S_{12}$ .

La compensation entre coefficients saisonniers est respectée, puisque  $S_1 + S_2 + S_3 + S_4 = 0$ .

Il est donc inutile de corriger les coefficients saisonniers. Les calculs sont détaillés figure 7.24.

**Figure 7.24**

**Résultats sous Excel.**

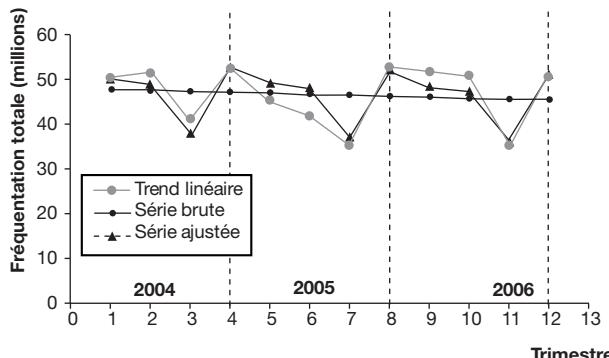
	C	D	E	F	G	H	I	J	K
1	t	$Y_t$	$t \cdot Y_t$	f <sub>t</sub>	s <sub>t</sub>	S <sub>j</sub>	S <sub>j</sub> '	$\hat{Y}_t$	$Y_{CVS(t)}$
2	1	50,46	50,46	47,79	2,67	2,22	2,22	50,01	48,24
3	2	51,46	102,92	47,57	3,89	1,43	1,43	49,00	50,03
4	3	41,07	123,21	47,35	-6,28	-9,41	-9,41	37,94	50,40
5	4	52,34	209,36	47,14	5,20	5,76	5,76	52,90	46,56
6	5	46,34	226,70	46,92	-1,58	2,22	2,22	49,14	43,12
7	6	41,86	251,16	46,70	-4,84	1,43	1,43	48,13	40,43
8	7	35,14	245,98	46,48	-11,34	-9,41	-9,41	37,07	44,55
9	8	52,99	423,92	46,27	6,72	5,76	5,76	52,03	47,23
10	9	51,63	464,67	46,05	5,58	2,22	2,22	48,27	49,41
11	10	51,06	510,60	45,83	5,23	1,43	1,43	47,26	49,63
12	11	35,00	385,00	45,61	-10,61	-9,41	-9,41	36,20	44,41
13	12	50,76	609,12	45,40	5,36	5,76	5,76	51,16	45,00

b. Pour le modèle additif, la série ajustée est donnée par :  $\hat{Y}_t = f_t + S'_t$ , d'où  $\hat{Y}_1 = f_1 + S'_1 = 49,79 + 2,22$ , soit  $\hat{Y}_1 = 50,01$  ;  $\hat{Y}_2 = f_2 + S'_2 = 47,55 - 1,43$ , soit  $\hat{Y}_2 = 49,00$ . De même,  $\hat{Y}_3 = 37,94$  ;  $\hat{Y}_4 = 52,90$  ;  $\hat{Y}_5 = 49,14$  ;  $\hat{Y}_6 = 48,13$  ;  $\hat{Y}_7 = 37,07$  ;  $\hat{Y}_8 = 52,03$  ;  $\hat{Y}_9 = 48,27$  ;  $\hat{Y}_{10} = 47,26$  ;  $\hat{Y}_{11} = 36,20$  ;  $\hat{Y}_{12} = 51,16$ . Ces calculs sont détaillés à la suite du tableau précédent (voir figure 7.24).

3. Les trois courbes sont représentées sur le même graphique, avec le temps, t, en abscisses, et les valeurs des entrées,  $Y_t$ ,  $f_t$  et  $\hat{Y}_t$ , en ordonnées.

**Figure 7.25**

**Fréquentation des salles de cinéma, tendance et série ajustée – France.**



4. Pour le modèle additif, la série CVS est donnée par :  $Y_{CVS}(t) = Y_t - S'_t$ , d'où  $Y_{CVS}(1) = Y_1 - S'_1 = 50,46 - 2,22$ , soit  $Y_{CVS}(1) = 48,24$  ;  $Y_{CVS}(2) = Y_2 - S'_2 = 51,46 + 1,43$ , soit  $Y_{CVS}(2) = 50,03$ . De même,  $Y_{CVS}(3) = 50,48$  ;  $Y_{CVS}(4) = 46,58$  ;  $Y_{CVS}(5) = 43,12$  ;  $Y_{CVS}(6) = 40,43$  ;  $Y_{CVS}(7) = 44,55$  ;  $Y_{CVS}(8) = 47,23$  ;  $Y_{CVS}(9) = 49,41$  ;  $Y_{CVS}(10) = 49,63$  ;  $Y_{CVS}(11) = 44,41$  ;  $Y_{CVS}(12) = 45,00$ . Ces calculs sont détaillés à la suite des calculs précédents (voir figure 7.24).

5. L'utilisation de l'équation de la droite de régression permet d'obtenir des prévisions de fréquentations trimestrielles pour l'année 2007. En appliquant le coefficient saisonnier  $S'_j$ , nous obtenons la série ajustée qui donne les prévisions de fréquentations trimestrielles pour l'année 2007. Ces prévisions sont à manier avec précaution, puisque le modèle de régression est estimé sur la période 2004-2006 (voir chapitre 6).

Ainsi, au premier trimestre 2007,  $t = T + h = 12 + 1 = 13$ , donc :

$f_{13} = -0,218 \times 13 + 48,007$ , soit  $f_{13} = 45,18$  ; au deuxième trimestre 2007,  $t = 14$ , donc :

$f_{14} = -0,218 \times 14 + 48,007$ , soit  $f_{14} = 44,96$ . De même,  $f_{15} = 44,74$  et  $f_{16} = 44,53$ .

D'où  $\hat{Y}_{13} = f_{13} + S'_{13} = 45,18 + 2,22$ , soit  $\hat{Y}_{13} = 47,40$ . La fréquentation prévisionnelle pour le premier trimestre de 2007 est de 47,40 millions d'entrées.

$\hat{Y}_{14} = f_{14} + S'_{14} = 44,96 - 1,43$ , soit  $\hat{Y}_{14} = 46,39$ . La fréquentation prévisionnelle pour le deuxième trimestre de 2007 est de 46,39 millions d'entrées.

De même,  $\hat{Y}_{15} = 35,33$  ;  $\hat{Y}_{16} = 50,29$ . Les fréquentations prévisionnelles pour les troisième et quatrième trimestres de 2007 sont respectivement de 35,33 et 50,29 millions d'entrées.

Ces calculs sont détaillés figure 7.26.

**Figure 7.26**

**Résultats sous Excel.**

	A	B	C	D	E	F
21	Trimestre	Année	$t=T+h$	$f_t$	$S'_j$	$\hat{Y}_t$
22	1	2007	13	45,18	2,22	47,40
23	2	2007	14	44,96	1,43	46,39
24	3	2007	15	44,74	-9,41	35,33
25	4	2007	16	44,53	5,76	50,29

Notons que ce modèle permet d'estimer la fréquentation totale de 2007 à 179,41 millions d'entrées. En réalité, le nombre total d'entrées sur 2007 a été de 178,14 millions d'entrées (en données provisoires au 4 janvier 2008, selon le CNC).



## EXERCICE 4 MÉTHODE ANALYTIQUE ET MODÈLE MULTIPLICATIF

### Énoncé

À partir des données de l'exercice 2 :

1. Déterminez la droite de régression de  $Y_t$  selon le temps.
2. À partir de la droite de régression de  $Y_t$  selon le temps et en utilisant le modèle multiplicatif :
  - a. Calculez les coefficients saisonniers.
  - b. Déterminez la série ajustée.
  - c. Déterminez la série CVS.
3. Proposez des prévisions de chiffre d'affaires pour l'année 2008.
4. Représentez sur un même graphique la tendance obtenue par la droite de régression entre 2005 et 2008 et la série brute de 2005 à 2007, prolongée de la série ajustée en 2008.

### Solution

1. La première étape consiste à présenter le tableau de données sous la forme d'un tableau statistique indiquant les valeurs de  $t$ , le temps, et de  $Y_t$ , valeur des entrées en période  $t$ . La droite de régression  $f_t = a \times t + b$  est déterminée par la méthode des MCO vue au chapitre 6. Il convient de déterminer les valeurs de  $a$  et  $b$  dans l'équation  $f_t = a \times t + b$ .

Pour cela, il est nécessaire de calculer les valeurs de  $\bar{t}$ ,  $\bar{y}$ ,  $V(t)$  et  $\sum_{i=1}^n t_i y_i$ .

Saisissez les valeurs de  $t$  dans la colonne L1 et celles de  $Y$  dans la colonne L2, comme indiqué figure 7.27.

Figure 7.27

Saisie du tableau de données avec la calculatrice.

L1	L2	L3	3
1	4.86		
2	6.52		
3	5.16		
4	6.75		
5	4.33		
6	6.73		
7	4.41		

L3(1)=

Pour obtenir les calculs intermédiaires nécessaires, appuyez sur la touche **STAT**, puis choisissez le menu **CALC** et sélectionnez la fonction **2:2-Var Stats**. Puis appuyez sur **ENTER**. Tapez **2-Var Stats L1,L2** puis appuyez à nouveau sur **ENTER**. Les résultats de



statistiques sur les variables t, notée X par la calculatrice, et Y, respectivement contenues dans L1 et L2, s'inscrivent (voir figures 7.28a et b).

**Figure 7.28a (gauche)**

**Statistiques sur L1 (t).**

**2-Var Stats**

$\bar{x}=6,5$

$\sum x=78$

$\sum x^2=650$

$Sx=3,605551275$

$\sigma x=3,45205253$

$\downarrow n=12$

**Figure 7.28b (droite)**

**Statistiques sur L2 (Y).**

**2-Var Stats**

$\bar{y}=5,569166667$

$\sum y=66,83$

$\sum y^2=402,9413$

$Sy=1,672067084$

$\sigma y=1,600882352$

$\downarrow \sum xy=434,7$

$$\text{De là, } a = \frac{434,70 - 12 \times 6,5 \times 5,57}{12 \times 11,92} = 0,002 \quad \text{et} \quad b = 5,57 - 0,002 \times 6,5 = 5,55, \quad \text{d'où :}$$

$$Y_t = 0,002 \times t + 5,555.$$

**2. a.** Pour déterminer les coefficients saisonniers, il est nécessaire de calculer la tendance. Dans le cadre de la méthode analytique, la tendance est calculée en utilisant l'équation de la droite de régression.

Pour  $t = 1$ ,  $f_1 = 0,002 \times 1 + 5,555$ , soit  $f_1 = 5,557$ ;  $f_2 = 0,002 \times 2 + 5,555$ , soit  $f_2 = 5,559$ .

De même,  $f_3 = 5,561$ ;  $f_4 = 5,563$ ;  $f_5 = 5,565$ ;  $f_6 = 5,567$ ;  $f_7 = 5,569$ ;  $f_8 = 5,571$ ;  $f_9 = 5,573$ ;  $f_{10} = 5,575$ ;  $f_{11} = 5,577$  et  $f_{12} = 5,579$ .

Pour calculer les valeurs de la tendance par période à l'aide de la calculatrice à la suite du tableau précédent, placez le curseur sur l'en-tête de colonne L3. Indiquez  $L3=0,002\times L1+5,555$ . Puis appuyez sur **ENTER**. La colonne L3 fait alors apparaître les valeurs de la tendance par période (voir figure 7.29).

**Figure 7.29**

**Calcul des valeurs de la tendance avec la calculatrice.**

L1	L2	L3	3
1	4,86	5,557	
2	6,52	5,559	
3	5,16	5,561	
4	6,75	5,563	
5	4,33	5,565	
6	6,73	5,567	
7	4,41	5,569	
<b>L3(1)=5,557</b>			

À la suite de ces calculs, les rapports saisonniers par période peuvent être calculés, selon le modèle multiplicatif.  $s_1 = Y_1 / f_1 = 4,86 / 5,557$ , soit  $s_1 = 0,875$ ;  $s_2 = Y_2 / f_2 = 6,52 / 5,559$ , soit  $s_2 = 1,173$ .

De même,  $s_3 = 0,928$ ;  $s_4 = 1,213$ ;  $s_5 = 0,778$ ;  $s_6 = 1,209$ ;  $s_7 = 0,792$ ;  $s_8 = 1,258$ ;  $s_9 = 0,558$ ;  $s_{10} = 1,365$ ;  $s_{11} = 0,507$ ;  $s_{12} = 1,346$ .

Pour calculer les variations saisonnières par période à l'aide de la calculatrice à la suite du tableau précédent, placez le curseur sur l'en-tête de colonne L4. Indiquez  $L4=L2/L3$ . Puis appuyez sur **ENTER**. La colonne L4 fait alors apparaître les valeurs des rapports saisonniers (voir figure 7.30).

**Figure 7.30**

**Calcul des valeurs des rapports saisonniers avec la calculatrice.**

L2	L3	L4	4
4,86	5,557	5,737	
6,52	5,559	1,1729	
5,16	5,561	.92789	
6,75	5,563	1,2134	
4,33	5,565	.77808	
6,73	5,567	1,2089	
4,41	5,569	.79188	
<b>L4(1)=.8745726111...</b>			

Les coefficients saisonniers sont ensuite calculés :

$$S_1 = \frac{s_1 + s_5 + s_9}{3} = \frac{0,875 + 0,778 + 0,558}{3}, \text{ soit } S_1 = 0,747 ;$$

$$S_2 = \frac{s_2 + s_6 + s_{10}}{3} = \frac{1,173 + 1,209 + 1,365}{3}, \text{ soit } S_2 = 1,249 ;$$

$$S_3 = \frac{s_3 + s_7 + s_{11}}{3} = \frac{0,928 + 0,792 + 0,507}{3}, \text{ soit } S_3 = 0,742 ;$$

$$S_4 = \frac{s_4 + s_8 + s_{12}}{3} = \frac{1,213 + 1,258 + 1,346}{3}, \text{ soit } S_4 = 1,272 .$$

Notons que le coefficient saisonnier d'un trimestre est le même pour chaque année, d'où  $S_1 = S_5 = S_9 ; S_2 = S_6 = S_{10} ; S_3 = S_7 = S_{11}$  et  $S_4 = S_8 = S_{12}$ .

La compensation entre coefficients saisonniers est respectée, donc les coefficients saisonniers corrigés sont identiques aux coefficients saisonniers.

**b.** Pour le modèle multiplicatif, la série ajustée est  $\hat{Y}_t = f_t \times S'_t$ , d'où  $\hat{Y}_1 = f_1 \times S'_1 = 5,557 \times 0,737$ , soit  $\hat{Y}_1 = 4,09$  ;  $\hat{Y}_2 = f_2 \times S'_2 = 5,559 \times 1,249$ , soit  $\hat{Y}_2 = 6,94$ . De même,  $\hat{Y}_3 = 4,13$  ;  $\hat{Y}_4 = 7,08$  ;  $\hat{Y}_5 = 4,10$  ;  $\hat{Y}_6 = 6,95$  ;  $\hat{Y}_7 = 4,13$  ;  $\hat{Y}_8 = 7,09$  ;  $\hat{Y}_9 = 4,11$  ;  $\hat{Y}_{10} = 6,96$  ;  $\hat{Y}_{11} = 4,14$  ;  $\hat{Y}_{12} = 7,10$ .

**c.** Pour le modèle multiplicatif, la série CVS est  $Y_{CVS}(t) = Y_t / S'_t$ , d'où  $Y_{CVS}(1) = Y_1 / S'_1 = 4,86 / 0,737$ , soit  $Y_{CVS}(1) = 6,60$  ;  $Y_{CVS}(2) = Y_2 / S'_2 = 6,52 / 1,249$ , soit  $Y_{CVS}(2) = 5,22$ . De même,  $Y_{CVS}(3) = 6,95$  ;  $Y_{CVS}(4) = 5,31$  ;  $Y_{CVS}(5) = 5,88$  ;  $Y_{CVS}(6) = 5,39$  ;  $Y_{CVS}(7) = 5,94$  ;  $Y_{CVS}(8) = 5,51$  ;  $Y_{CVS}(9) = 4,22$  ;  $Y_{CVS}(10) = 6,09$  ;  $Y_{CVS}(11) = 3,81$  ;  $Y_{CVS}(12) = 5,90$ .

**3.** L'utilisation de l'équation de la droite de régression permet d'obtenir des prévisions de chiffre d'affaires pour l'année 2008. En appliquant le coefficient saisonnier  $S'_p$ , nous obtenons la série ajustée qui donne les prévisions de chiffres d'affaires trimestriels pour l'année 2008. Ces prévisions sont à manier avec précaution, puisque le modèle de régression est estimé sur la période 2005-2007 (voir chapitre 6).

Ainsi, au premier trimestre 2008,  $t = T + H = 12 + 1 = 13$ , donc  $f_{13} = 0,002 \times 13 + 5,555$ , soit  $f_{13} = 5,583$ . Au deuxième trimestre 2008,  $t = 14$ , donc  $f_{14} = 0,002 \times 14 + 5,555$ , soit  $f_{14} = 5,585$ . De même,  $f_{15} = 5,587$  et  $f_{16} = 5,589$ .

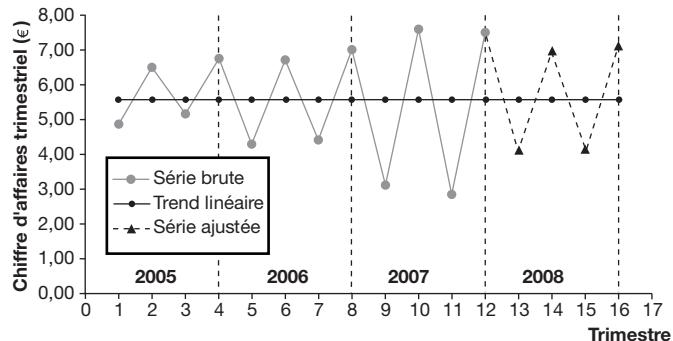
D'où  $\hat{Y}_{13} = f_{13} \times S'_{13} = 5,583 \times 0,737$ , soit  $\hat{Y}_{13} = 4,11$ . Le chiffre d'affaires prévisionnel pour le premier trimestre de 2007 est de 4,11 milliers d'euros.

$\hat{Y}_{14} = f_{14} \times S'_{14} = 5,585 \times 1,249$ , soit  $\hat{Y}_{14} = 6,97$ . Le chiffre d'affaires prévisionnel pour le deuxième trimestre de 2007 est de 6,97 milliers d'euros.

De même,  $\hat{Y}_{15} = 4,15$  ;  $\hat{Y}_{16} = 7,11$ . Les chiffres d'affaires prévisionnels pour les troisième et quatrième trimestres de 2007 sont respectivement de 4,15 et 7,11 milliers d'euros.

4. Les deux courbes sont représentées sur le même graphique (voir figure 7.31), avec le temps, t, en abscisses, et la tendance  $f_t$  et les valeurs du chiffre d'affaires  $Y_t$  prolongé de  $\hat{Y}_t$  en ordonnées.

**Figure 7.31**  
Chiffre d'affaires,  
tendance et  
prévisions.



# Bibliographie

- CALOT G., *Cours de statistique descriptive*, Dunod, 1969.
- CHAREILLE P. et PINAULT Y., *Statistique descriptive*, Collection AES, Montchrestien, Paris, 1996.
- DARMOIS G., *Statistiques et applications*, Armand Colin, 1952.
- DODGE Y., *Statistique. Dictionnaire encyclopédique*, Springer, 2004.
- DOR E., *Économétrie*, Collection Synthex, Pearson Education, 2004.
- DROESBEKE J.-J. et TASSI Ph., *Histoire de la statistique*, Que sais-je ?, PUF, 1990.
- LIORZOU A., *Initiation à la pratique statistique*, Eyrolles, 1979.
- GUERBER L et HENNEQUIN P.-L., *Initiation à la statistique*, Bibliothèque d'enseignement mathématique A.P.M.E.P., 1967.
- SCHLACTHER D., *De l'analyse à la prévision*, Ellipses, 1986.
- WONNACOTT T. et R., *Statistiques*, Economica, 1984.

# Les indices

- 1. Les indices élémentaires ..... 220
- 2. Les indices synthétiques ..... 226

## Problèmes et exercices

- 1. Indices élémentaires ..... 236
- 2. Indices synthétiques ..... 238
- 3. Coefficients budgétaires et relation entre indices ..... 240

Dans de nombreux domaines, notamment dans le domaine économique, nous devons savoir décrire et analyser l'évolution temporelle ou spatiale de différentes grandeurs. Les pourcentages ne disposent pas des qualités propres à décrire simplement ces variations<sup>1</sup>. L'indicateur fondamental de l'évolution des variables économiques et sociales est l'indice.

On distingue deux types d'indices : les indices portant sur une seule grandeur, appelés indices élémentaires, et les indices portant sur des grandeurs complexes (agrégation de plusieurs grandeurs), nommés indices synthétiques dans le cas où les grandeurs sont de même nature (indice des prix regroupant un panier de biens) ou indices composites quand il s'agit de grandeurs de natures différentes (l'indice boursier de Shanghai, qui comprend à la fois les actions A libellées en yuans et les actions B libellées en devises, est un indice composite). Il est vivement conseillé au lecteur d'aller explorer le site de l'Insee ([www.insee.fr](http://www.insee.fr)), qui offre une grande richesse d'information sur les différents indices.

---

1. Les pourcentages, par exemple, ne s'ajoutent pas : une hausse de 10 % suivie d'une hausse de 20 % correspond à une hausse globale de 32 % (coefficent multiplicateur).

Nous verrons que les indices synthétiques apparaissent comme des moyennes pondérées (arithmétiques, géométriques ou harmoniques) des indices élémentaires et nous définirons les coefficients budgétaires qui constituent les pondérations.

# 1 Les indices élémentaires

Nous commencerons par un petit rappel sur les calculs de variations, avant d'exposer les indices élémentaires et leurs propriétés.

## 1.1 VOCABULAIRE DES VARIATIONS, COEFFICIENT MULTIPLICATEUR

Avant de définir les indices, il est important de dire ici qu'un indice évalue une variation et non un niveau et qu'il mesure cette variation en valeur relative et non absolue. Ainsi, dire qu'en 2007 l'indice base 100 en 2000 du prix du pain (baguette) est de 123,72 et celui du café moulu de 103,8 n'indique évidemment pas que le prix de la baguette est supérieur à celui du café, mais que la baguette a augmenté de 23,72 % de 2000 à 2007 et le café de 3,8 % dans la même période.

Nous commencerons donc par clarifier le vocabulaire des outils permettant de mesurer les variations d'une grandeur (économique, sociale, etc.) et par définir le coefficient multiplicateur.

### Définitions

La **variation absolue** d'une grandeur  $G$  de la date 0 à la date  $t$  est la différence entre la valeur finale (à la date  $t$ ) et la valeur initiale (à la date 0) de cette grandeur. Cette variation absolue est notée :  $\Delta G = G_t - G_0$ .

Une variation absolue positive traduit une augmentation et une variation négative une baisse.

La **variation relative** d'une grandeur  $G$  de la date 0 à la date  $t$  est le rapport entre la variation absolue et la valeur **initiale** de cette grandeur. Cette variation relative est notée :  $\Delta G / G = (G_t - G_0) / G_0$ .

Une variation relative s'exprime souvent en pourcentage de la valeur initiale, ce pourcentage étant donné par :  $(G_t - G_0) \times 100 / G_0$ .

Quand une grandeur passe de la valeur  $G_0$  à la valeur  $G_t$ , on note  $a$  le **coefficient multiplicateur** défini par :  $a = G_t / G_0$ .

Un coefficient plus grand que 1 traduit une hausse et un coefficient inférieur à 1, une baisse. On notera que le coefficient multiplicateur ne possède pas d'unité.

### Exemple 8.1

#### Coefficient multiplicateur

Le tableau suivant donne la population de la France (France métropolitaine et DOM) :

Année	Population (en milliers)
2003	62 042
2004	62 445

Source : Insee, Tableaux de l'économie française, 2007

Nous pouvons calculer la variation absolue, la variation relative et le coefficient multiplicateur de 2003 à 2004. Nous noterons respectivement  $P_0$  et  $P_1$  les populations en 2003 et 2004.

La variation absolue est :  $\Delta P = P_1 - P_0 = 62\ 445 - 62\ 042 = 403$  milliers d'habitants.

La variation relative est :  $\Delta P / P = (P_1 - P_0) / P_0 = 403 / 62\ 042 = 0,0065$ , soit une augmentation de 0,65 %.

Le coefficient multiplicateur est :  $a = P_1 / P_0 = 62\ 445 / 62\ 042 = 1,0065$  ; il est supérieur à 1 et traduit une hausse dont le taux est :  $t = a - 1 = 0,0065$ .

On rappelle que, pour mesurer l'effet global de plusieurs variations successives, on doit employer les coefficients multiplicateurs, comme le montre l'exemple 8.2.

### Exemple 8.2

#### Coefficient multiplicateur et pourcentages

Supposons qu'une grandeur subisse une augmentation de 30 % suivie d'une baisse de 10 % et mesurons l'effet global de ces variations en pourcentage : nous utiliserons les coefficients multiplicateurs successifs  $a_1 = 1,30$  et  $a_2 = 0,90$ , ce qui donne un coefficient multiplicateur global :  $a = a_1 \times a_2 = 1,30 \times 0,90 = 1,17$ , soit une hausse de 17 %. On constate que les pourcentages ne s'ajoutent pas.

Notons  $p_0$  le prix hors taxe et  $p_1$  le prix TTC, après application de la TVA à 19,6 %. Déterminons la variation en pourcentage, permettant de revenir du prix TTC au prix HT. On a :  $p_1 = 1,196 p_0$ , soit  $p_0 = p_1 / 1,196$ , ce qui donne un coefficient multiplicateur  $a = 1 / 1,196 = 0,8361$  quand on passe de  $p_1$  à  $p_0$ , soit une baisse de taux :  $t = 1 - 0,8361 = 0,1639$ , soit 16,39 %. La TVA représente 16,39 % du prix TTC affiché en magasin.

On constate que les pourcentages ne sont pas réversibles, c'est-à-dire qu'une hausse de 19,6 % n'est pas neutralisée par une baisse de 19,6 %.

L'exemple 8.2 nous a montré les « défauts » des pourcentages et la nécessité d'utiliser un outil plus adapté à la mesure des variations : l'indice.

## 1.2 INDICES ÉLÉMENTAIRES BASE 1 ET BASE 100

Pour décrire les variations de grandeurs simples telles que le prix du baril de pétrole, le smic, le taux de fécondité, on compare leurs valeurs dans le temps ou dans l'espace en effectuant le rapport des valeurs de la grandeur considérée à deux dates différentes (indice chronologique), ou en deux lieux distincts (indice spatial).

### Définitions

**Indice base 1** : on appelle indice élémentaire de la grandeur simple  $G$ , à la date  $t$ , base 1 à la date 0, le rapport noté  $I_{t/0}(G) = G_t / G_0$ . La date 0 est appelée la date de référence, et la date  $t$  la date courante.

On reconnaît le coefficient multiplicateur. On notera que  $I_{0/0}(G) = 1$ .

**Indice base 100** : on appelle indice élémentaire de la grandeur simple  $G$ , à la date  $t$ , base 100 à la date 0, le rapport noté  $I_{t/0}(G)$  et défini par :  $I_{t/0}(G) = (G_t / G_0) \times 100$ .

On notera que  $I_{0/0}(G) = 100$ .

Un indice ne possède pas d'unité. Un indice supérieur à 100 représente une hausse et un indice inférieur à 100 une baisse. On parlera souvent d'année de base ou d'année de référence pour dénommer la date 0.

Les indices base 100 sont les plus courants, car bien adaptés aux pourcentages. On notera que les bases 1 ou 100 n'apparaissent pas dans la notation, mais qu'on indique au départ le type d'indice utilisé.

### Exemple 8.3

#### Indices base 1 et base 100

Reprendons l'exemple 8.1. Nous pouvons écrire, en notant  $P$  la population de la France :  $I_{2004/2003}(P) = 1,0065$  en utilisant un indice base 1, ce qui signifie que la population a augmenté de 2003 à 2004 comme une grandeur qui valait 1 en 2003 et qui vaut 1,0065 en 2004.

Si l'on utilise un indice base 100, on notera :  $I_{2004/2003}(P) = 100,65$ , ce qui donne la même variation qu'une grandeur qui valait 100 en 2003 et 100,65 en 2004.

Il est possible de calculer le pourcentage de variation entre deux périodes grâce aux deux indices relatifs à ces périodes. À partir de deux indices base 100 année 0, d'une même grandeur, aux dates respectives  $t_1$  et  $t_2$ , la variation en pourcentage de la grandeur de l'année  $t_1$  à l'année  $t_2$  est donnée par la variation relative de l'indice :

$$\frac{I_{t_2/0}(G) - I_{t_1/0}(G)}{I_{t_1/0}(G)} \times 100$$
. Au numérateur, la variation absolue  $I_{t_2/0}(G) - I_{t_1/0}(G)$  se mesure en **points d'indice**.

### Exemple 8.4

#### Points d'indice et variation en pourcentage

Le tableau suivant donne la population de la France (en milliers, source Insee 2007) et les indices base 100 en 1990 :

Année	$P_t$	$I_{t/1990}(P)$
1990	58 171	100,00
2000	60 751	104,44
2005	62 818	107,99

Utilisons les indices  $I_{t/1990}(P)$  pour déterminer la variation en pourcentage de la population de 2000 à 2005.

De 2000 à 2005 la variation absolue de l'indice a été de :  $I_{2005/1990}(P) - I_{2000/1990}(P) = 107,99 - 104,44 = 3,55$  ; l'indice a augmenté de 3,55 points d'indice de 2000 à 2005 ; on dit aussi que cet indice a pris 3,55 points d'indice.

On peut évaluer la variation en pourcentage de la population de 2000 à 2005 en évaluant la variation relative de l'indice, c'est-à-dire :  $(I_{2005/1990}(P) - I_{2000/1990}(P)) / I_{2000/1990}(P) = 3,55 / 104,44 = 0,034$ , soit une hausse de 3,4 %.

## 1.3 PROPRIÉTÉS DES INDICES ÉLÉMENTAIRES

Les indices élémentaires possèdent des propriétés qui manquent aux pourcentages et que nous allons exposer ici. Ces propriétés sont détaillées dans le focus 8.1. On notera au préalable que les formules sur les indices élémentaires sont données sous forme duale : en base 1 pour la compréhension et en base 100 pour l'usage.

### La circularité, ou transférabilité

C'est la propriété fondamentale des indices, qui permet de « voyager » dans le temps et qui se traduit par une relation multiplicative, de type relation de Chasles<sup>1</sup>. On rappelle que la relation de Chasles est la relation vectorielle  $\vec{MP} + \vec{PS} = \vec{MS}$ , qui lie trois points quelconques de l'espace. C'est une relation basée sur la correspondance (type SNCF) : pour aller de Marseille à Strasbourg, allez de Marseille à Paris et prenez la correspondance à Paris pour Strasbourg.

#### Définition

Un indice est **transférable** si et seulement si il vérifie la relation :

- pour les indices base 1 :  $I_{t_2/t_0}(G) = I_{t_2/t_1}(G) \times I_{t_1/t_0}(G)$  ;
- pour les indices base 100 :  $100I_{t_2/t_0}(G) = I_{t_2/t_1}(G) \times I_{t_1/t_0}(G)$ .

On devra contrôler dans les formules base 100 l'homogénéité. Dans la formule multiplicative précédente il y a deux indices dans le membre de droite et un seul dans celui de gauche, il y a donc un facteur 100 pour « équilibrer » la relation.

#### Propriété

Les indices élémentaires sont transférables.

### La réversibilité

La réversibilité consiste à permute l'année courante et l'année de référence.

#### Définition

Un indice est **réversible** si et seulement si il vérifie la relation :

- pour les indices base 1 :  $I_{t_1/t_0}(G) = 1 / I_{t_0/t_1}(G)$  ;
- pour les indices base 100 :  $I_{t_1/t_0}(G) = 10000 / I_{t_0/t_1}(G)$ .

On notera que ces formules découlent de la circularité.

En base 1,  $I_{t_1/t_0}(G) \times I_{t_0/t_1}(G) = I_{t_1/t_1}(G) = 1$  (base 1). On retrouve une relation de Chasles avec un « aller-retour ».

$$I_{t_1/t_0}(G) \times I_{t_0/t_1}(G) = 100I_{t_1/t_1}(G) = 100^2 \text{ (base 100).}$$

#### Propriété

Les indices élémentaires sont réversibles.

### L'enchaînement

Dans de nombreuses situations, on doit suivre l'évolution d'une grandeur d'une année sur l'autre et on utilise alors des indices chaînes, en prenant pour année de référence l'année qui précède l'année courante.

1. Michel Chasles, mathématicien français (1793-1830) dont le nom est lié à la relation du même nom.

**Définition**

Les **indices chaînes** sont des indices pour lesquels l'année de référence est l'année qui précède l'année courante. Ils sont notés :  $I_{t/t-1}(G)$ .

La généralisation de la transférabilité donne :

- pour les indices base 1 :  $I_{t/t-1}(G) \times I_{t-1/t-2}(G) \times \dots \times I_{1/0}(G) = I_{t/0}(G)$  ;
- pour les indices base 100 :  $I_{t/t-1}(G) \times I_{t-1/t-2}(G) \times \dots \times I_{1/0}(G) = 100^{t-1} \times I_{t/0}(G)$  (car il y a  $t$  indices dans le membre de gauche et un seul à droite).

**Propriété**

Les indices élémentaires sont enchaînables.

**Focus 8.1****Propriétés des indices élémentaires**

Le tableau suivant donne le prix moyen TTC de l'eau à la consommation en métropole, en janvier de chacune des années. Ces prix sont suivis des indices du prix de l'eau base 100 en 2002 et des indices enchaînés ( $I_{2002/2001}$  n'étant pas calculable, puisque 2001 n'est pas communiqué).

<b>Année</b>	<b>Prix</b>	$I_{t/2002}(P)$	$I_{t/t-1}(P)$
2002	165,65	100	—
2003	170,45	102,90	102,90
2004	172,19	103,95	101,02
2005	178,93	108,02	103,91
2006	187,19	113,00	104,62

Source : Insee, 2007

On vérifie l'ensemble des propriétés des indices élémentaires :

- Circularité : on a (base 100) :  $I_{2005/2003}(P) = (178,93 / 170,45) \times 100$ ,  $I_{2003/2002}(P) = (170,45 / 165,65) \times 100$  et  $I_{2005/2002}(P) = (178,93 / 165,65) \times 100$ ; on vérifie sans effectuer les calculs la circularité :  $I_{2005/2003}(P) \times I_{2003/2002}(P) = 100 I_{2005/2002}(P)$ , le facteur 170,45 (prix intermédiaire de 2003) s'éliminant.
- Réversibilité : on a (base 100)  $I_{2005/2002}(P) = (178,93 / 165,65) \times 100 = 108,02$  et  $I_{2002/2005}(P) = (165,65 / 178,93) \times 100$ , et on établit :  $I_{2005/2002}(P) \times I_{2002/2005}(P) = 10\,000$  soit la formule de réversibilité, ce qui donne :  $I_{2002/2005}(P) = 10\,000 / 108,2 = 92,42$ . Interprétation : de 2002 à 2005 le prix de l'eau a augmenté de 8,02 %. La réversibilité permet de conclure qu'en 2002 le prix de l'eau était 7,58 % ( $100 - 92,42$ ) moins élevé qu'en 2005.
- Indices enchaînés : on peut vérifier que  $I_{2006/2005}(P) \times I_{2005/2004}(P) \times I_{2004/2003}(P) \times I_{2003/2002}(P) = 113\,003\,320$  soit environ (approximations) :  $100^3 \times I_{2006/2002}(P) = 1\,000\,000 \times (187,19 / 165,5) \times 100$ .

## Opérations

Les indices élémentaires possèdent des propriétés précieuses relatives au produit et au quotient.

### Propriétés

- **Produit**

En base 1, l'indice élémentaire d'un produit de deux grandeurs est le produit des indices.

En base 100, on a :  $I_{t/0}(A \times B) = I_{t/0}(A) \times I_{t/0}(B) / 100$ .

- **Quotient**

En base 1, l'indice élémentaire d'un quotient de deux grandeurs est le quotient des indices.

En base 100, on a :  $I_{t/0}(A / B) = (I_{t/0}(A) / I_{t/0}(B)) \times 100$ .

On citera notamment l'indice de pouvoir d'achat, qui s'obtient par la formule :  $I_{t/0}(\text{Pouvoir achat}) = (I_{t/0}(S) / I_{t/0}(P)) \times 100$ , S désignant le salaire et P les prix. Il s'agit donc du quotient de l'indice des salaires nominaux par l'indice des prix.

### Exemple 8.5

#### Indices élémentaires et opérations

D'après une étude de l'Insee, de 1986 à 1998, le nombre d'entrées au cinéma est passé de 170 millions à 160 millions alors que le prix de la place de cinéma passait de 4 € à 5,90 €. Dans le tableau suivant, on note P le prix d'une place (en euros), Q la quantité de places vendues (en millions) et V la valeur globale (qui correspond ici à la recette :  $V = P \times Q$ ).

Année	P	Q	V
1986	4	170	680
1998	5,9	160	944

Source : Insee, 2002

On peut calculer les indices élémentaires de quantité et de prix en 1998, base 100 en 1986.

On a :  $I_{1998/1986}(P) = (5,90 / 4) \times 100 = 147,5$  ;  $I_{1998/1986}(Q) = (160 / 170) \times 100 = 94,12$  et  $I_{1998/1986}(V) = (944 / 680) \times 100 = 138,82$ .

On vérifie que  $I_{1998/1986}(V) = I_{1998/1986}(P) \times I_{1998/1986}(Q) / 100 = 147,5 \times 94,12 / 100$ .

Ainsi, la hausse de 38,82 % de la recette est due à l'effet conjugué d'une baisse de la quantité et d'une augmentation du prix.

## 1.4 L'INDEXATION

La publication des grands indicateurs fait régulièrement la une des journaux, et l'indice des prix tient régulièrement la vedette, du fait qu'il joue un rôle central dans l'appréciation de la situation économique du pays, mais aussi de par les répercussions importantes qu'il entraîne par le biais des indexations<sup>1</sup>.

Le smic est revalorisé au 1<sup>er</sup> juillet de chaque année, notamment en fonction de l'évolution de l'indice des prix à la consommation (indice pour les « ménages urbains dont le chef est ouvrier ou employé, hors tabac »). L'indexation a pour but d'assurer un maintien du pouvoir d'achat ; elle nécessite une durée ou périodicité (l'année, dans le cas du smic), une date (1<sup>er</sup> juillet, pour le smic) et un indice de référence. L'exemple 8.6 donne un exemple pour un loyer indexé sur l'indice du coût de la construction (ICC, indice trimestriel).

### Exemple 8.6

#### Indexation

Supposons qu'un locataire ait signé le 15 janvier 2007 un bail avec un loyer mensuel de 750 euros, ce loyer étant réévalué chaque année à la date anniversaire du bail, l'indice de référence étant l'indice du coût de la construction (ICC) du 2<sup>e</sup> trimestre 2006. L'indice du coût de la construction du 2<sup>e</sup> trimestre 2006, base 100 au 4<sup>e</sup> trimestre 1953, vaut 1 366 et celui du 2<sup>e</sup> trimestre 2007, 1 435. Calculons le loyer de ce locataire au 15 janvier 2008.

Ce loyer va suivre la progression de l'indice sur un an, ce qui donne un coefficient multiplicateur  $a = 1435 / 1366 = 1,0505$ , ce qui donnera un nouveau loyer de :  $750 \times 1,0505 = 787,88$  euros.

## 2 Les indices synthétiques

À sa création en 1946, l'Insee a repris l'indice des 34 articles établi base 100 en 1914 et base 100 en 1938, calculé par la Statistique générale de la France, et qui faisait suite à un indice de 13 articles publié depuis 1916. La liste des 34 articles comprenait 29 denrées alimentaires, 4 articles de chauffage et éclairage, un seul article (le savon) pour l'entretien ménager ; la plupart des produits manufacturés, dont l'habillement, n'étaient pas représentés, les services étant complètement absents.

L'indice a beaucoup évolué, et l'IPC (indice des prix à la consommation) base 1998 est la septième génération d'indice. Il couvre l'ensemble de la population et du territoire (métropole et DOM) et se décompose aujourd'hui en 305 postes, chacun d'eux étant représenté par un indice (« œufs », « pantalons pour enfants », « coiffeurs pour femme », « maisons de retraite »...). Il exclut le tabac et les alcools.

On comprend que le problème pour composer un « bon » indice des prix vient de la difficulté à prendre en compte l'importance de chacun des postes dans la constitution d'un indice synthétique et à tenir compte des évolutions des modes de consommation.

1. *Index* désignait, chez les Romains, « celui qui montre ».

**Focus 8.2****Comment construire un indice synthétique ?**

Le tableau suivant donne pour les années 2001 et 2007 les valeurs du smic horaire brut en euros (heures légales). On a supposé une majoration de 25 % pour les heures supplémentaires en 2001 et de 40 % en 2007 ; les durées légales du travail mensuel sont celles qui ont prévalu dans les entreprises dans la période du passage aux 35 heures et on a supposé que l'employé moyen assurait en 2001 en moyenne 2 heures supplémentaires par mois et en 2007 4 heures supplémentaires par mois.

Comment définir un « bon » indice de salaire en 2007, base 100 en 2001 ?

<b>Heures</b>	<b>Smic 2001</b>	<b>Quantité 2001</b>	<b>Smic 2007</b>	<b>Quantité 2007</b>
Légales	6,67	169	8,44	151,67
Supplémentaires	8,3375	2	11,816	4

On peut calculer pour chacune des années un salaire global, noté  $S$ , et en déduire ainsi un indice :  $S_{2001} = 169 \times 6,67 + 2 \times 8,3375 = 1\,143,90$  € et

$S_{2007} = 151,67 \times 8,44 + 4 \times 11,816 = 1\,327,36$  €, ce qui donnerait pour l'indice de salaire global :  $I_{2007/2001}(S) = (1\,327,36 / 1\,143,9) \times 100 = 116,03$ , soit une augmentation de 16,03 %. Cependant, cet indice est « brouillé », dans la mesure où sa signification traduit simultanément une évolution de la quantité d'heures de travail et une évolution du salaire horaire, sans que l'on puisse isoler l'impact de ces évolutions. Pour résumer les indices élémentaires de salaire, on va donc introduire un indice synthétique de salaire horaire, de façon à gommer l'influence due à la variation des quantités, en les considérant comme constantes. On peut alors opter pour deux possibilités :

- Fixer les quantités à leur niveau pris l'année de base, c'est-à-dire privilégier le mode de travail du salarié de 2001. On forme alors l'indice de Laspeyres des salaires horaires, noté :  $L_{2007/2001}(s) = (169 \times 8,44 + 4 \times 11,816) / (169 \times 6,67 + 2 \times 8,3375) \times 100 = (1\,449,99 / 1\,143,91) \times 100 = 126,76$ .
- Fixer les quantités à leur niveau pris l'année courante, c'est-à-dire privilégier le mode de travail du salarié de 2007. On forme l'indice de Paasche des salaires horaires, noté :  $P_{2007/2001}(s) = (151,67 \times 8,44 + 4 \times 11,816) / (151,67 \times 6,67 + 4 \times 8,3375) \times 100 = (1\,327,36 / 1\,044,99) \times 100 = 127,02$ .
- Le choix entre ces deux indices présente un certain arbitraire, et nous verrons plus loin que le statisticien américain Fisher<sup>1</sup> a proposé dans les années 1920 un indice « idéal », qui est la moyenne géométrique des deux indices précédents.

Nous allons maintenant définir les indices synthétiques de Laspeyres<sup>2</sup> et de Paasche<sup>3</sup>, indices de prix et de quantités. Ces indices vont respecter le principe évoqué dans le focus précédent : dans un indice de prix, seuls les prix varient, les quantités restant constantes,

1. Irving Fisher, économiste, mathématicien américain (1867-1947).

2. Étienne Laspeyres, économiste, statisticien allemand (1834-1913).

3. Hermann Paasche, statisticien, économiste allemand (1851-1925).

et, dans un indice de quantité, seules les quantités varient, les prix restant fixes. Auparavant nous allons introduire les notations et définir les coefficients budgétaires.

## 2.1 LES COEFFICIENTS BUDGÉTAIRES

Soit un panier de consommation, composé de  $n$  produits, le produit  $i$  ( $i$  entier variant de 1 à  $n$ ) ayant pour prix unitaires respectifs  $P_i^0$  et  $P_i^t$  aux années de base (année 0) et courante ( $t$ ), les quantités consommées étant respectivement notées  $Q_i^0$  et  $Q_i^t$ .

On notera respectivement  $V_i^0$  et  $V_i^t$  les valeurs globales du bien  $i$  aux dates 0 et  $t$  et  $V^0$  et  $V^t$  les valeurs globales de ce panier aux années de base et courante, avec  $V_i^0 = P_i^0 Q_i^0$ ,

$$V_i^t = P_i^t Q_i^t, \quad V^0 = \sum_{i=1}^n P_i^0 Q_i^0 \quad \text{et} \quad V^t = \sum_{i=1}^n P_i^t Q_i^t.$$

### Définition

Étant donné un panier de consommation, on appelle **coefficients budgétaires** d'un bien  $j$  de ce panier, l'année 0 (respectivement l'année  $t$ ), la part du budget total de l'année 0 (respectivement l'année  $t$ ) affectée au bien  $j$ ; ce coefficient sera noté  $C_j^0$  (respectivement  $C_j^t$ ) et défini par :  $C_j^0 = \frac{P_j^0 Q_j^0}{\sum_{i=1}^n P_i^0 Q_i^0} = \frac{P_j^0 Q_j^0}{V^0}$  (respectivement  $C_j^t = \frac{P_j^t Q_j^t}{\sum_{i=1}^n P_i^t Q_i^t} = \frac{P_j^t Q_j^t}{V^t}$ ).

On a :  $\sum_{i=1}^n C_i^0 = \sum_{i=1}^n C_i^t = 1$  ou 100 % s'ils sont exprimés en pourcentage (voir les masses relatives, chapitre 4, section 3 sur la concentration).

### Exemple 8.7

#### Coefficients budgétaires

Considérons le panier de consommation suivant composé de deux denrées, la baguette de pain et la viande de bœuf, l'année de référence étant l'année 1980 et l'année courante, l'année 2003.

Les quantités de consommation  $Q$  sont données pour un mois, en nombre de baguettes et en kilos de viande. Les prix sont notés  $P$  et les valeurs globales  $V$ .

Bien	$Q_i^0$ 1980	$P_i^0$ 1980	$V_i^0$ 1980	$Q_i^t$ 2003	$P_i^t$ 2003	$V_i^t$ 2003
Baguette	21,00	0,15	3,20	18,00	0,75	13,50
Viande de bœuf	1,23	6,74	8,30	2,10	16,50	34,58

$$V^0 = \sum_{i=1}^2 P_i^0 Q_i^0 = 3,20 + 8,30, \quad \text{soit} \quad V^0 = 11,50 \quad \text{et} \quad V^t = \sum_{i=1}^2 P_i^t Q_i^t = 13,50 + 34,58, \quad \text{soit} \\ V^t = 48,08.$$

Calculons les coefficients budgétaires de chacun des biens l'année de base et l'année courante. On a pour le bien 1 (pain) :  $C_1^0 = \frac{V_1^0}{V^0} = \frac{3,20}{11,51} = 0,2783$ , soit 27,83 % du budget du consommateur de 1980 consacré au pain. On trouve de même :  $C_2^0 = \frac{V_2^0}{V^0} = \frac{8,30}{11,51} = 0,7217$  ;  $C_1^t = \frac{13,50}{48,08} = 0,2808$  et  $C_2^t = \frac{13,50}{48,08} = 0,7192$ .

## 2.2 LES INDICES DE LASPEYRES

Nous allons définir deux indices de Laspeyres, l'un relatif au prix, l'autre aux quantités.

### Indice des prix de Laspeyres

#### Définition

On appelle **indice des prix de Laspeyres**, année  $t$ , base 100 l'année 0, l'indice noté  $L_{t/0}(P)$  obtenu en fixant les quantités à l'année de base. Il est défini par :  $L_{t/0}(P) = \frac{\sum_{i=1}^n Q_i^0 P_i^t}{\sum_{i=1}^n Q_i^0 P_i^0} \times 100$ .

On a :  $L_{t/0}(P) = \frac{\sum_{i=1}^n Q_i^0 P_i^t}{\sum_{i=1}^n Q_i^0 P_i^0} \times 100 = \frac{\sum_{i=1}^n Q_i^0 P_i^t}{V^0} \times 100 = \sum_{i=1}^n \frac{Q_i^0 P_i^0}{V^0} \left( \frac{P_i^t}{P_i^0} \times 100 \right)$  ; on reconnaît dans la parenthèse l'indice élémentaire du bien  $i$  et le coefficient  $\frac{Q_i^0 P_i^0}{V^0}$  est  $C_i^0$  le coefficient budgétaire du bien  $i$ , l'année de base. On rappelle que les coefficients budgétaires de l'année 0 ont pour somme 1. D'où la propriété suivante.

#### Propriété

L'indice des prix de Laspeyres est la moyenne arithmétique pondérée des indices élémentaires de prix des biens composant le panier. Les coefficients de pondération sont les coefficients budgétaires de l'année de base.

#### Exemple 8.8

### Indice des prix de Laspeyres

Reprendons l'exemple 8.7 et calculons l'indice des prix de Laspeyres en 2003, base 100 en 1980.

$$L_{2003/1980}(P) = \frac{\sum_{i=1}^2 Q_i^0 P_i^t}{\sum_{i=1}^2 Q_i^0 P_i^0} \times 100 = \frac{21 \times 0,75 + 1,23 \times 16,5}{21 \times 0,1525 + 1,23 \times 6,74} \times 100 = 313,66, \text{ soit une augmentation de } 213,66\%.$$

Laspeyres s'intéresse au mode de consommation du consommateur de 1980 : si ce dernier consomme en 2003 de la même façon qu'en 1980, cela lui coûtera 213,66 % plus cher.

En utilisant la propriété de l'indice des prix de Laspeyres, on trouve effectivement que la valeur de cet indice est la moyenne arithmétique des indices élémentaires de prix pondérée par les coefficients budgétaires de l'année de base : pour la baguette,  $I_{2003/1980}(P_1) = \frac{0,75}{0,1525} \times 100 = 491,80$  et  $C_1^0 = 0,2783$  ; pour la viande de bœuf,  $I_{2003/1980}(P_2) = \frac{16,5}{6,74} \times 100 = 244,96$  et  $C_2^0 = 0,7217$ , ce qui donne pour l'indice des prix de Laspeyres :  $L_{2003/1980}(P) = 0,2783 \times 491,8 + 0,7217 \times 244,96 = 313,66$ .

### Indice des quantités de Laspeyres

#### Définition

On appelle **indice des quantités de Laspeyres**, année t, base 100 l'année 0, l'indice noté  $L_{t/0}(Q)$  obtenu en fixant les prix à l'année de base. Il est défini par :

$$L_{t/0}(Q) = \frac{\sum_{i=1}^n P_i^0 Q_i^t}{\sum_{i=1}^n P_i^0 Q_i^0} \times 100$$

#### Propriété

L'indice des quantités de Laspeyres est la moyenne arithmétique pondérée des indices élémentaires de quantités des biens composant le panier. Les coefficients de pondération sont les coefficients budgétaires de l'année de base.

#### Exemple 8.9

##### Indice des quantités de Laspeyres

Reprendons l'exemple 8.7 et calculons l'indice des quantités de Laspeyres en 2003, base 100 en 1980 :

$$L_{2003/1980}(Q) = \frac{\sum_{i=1}^2 P_i^0 Q_i^t}{\sum_{i=1}^2 P_i^0 Q_i^0} \times 100 = \frac{0,1525 \times 18 + 6,74 \times 2,1}{0,1525 \times 21 + 6,74 \times 1,23} \times 100 = 146,54, \text{ soit}$$

une augmentation de 46,54 % des quantités.

### Indice de Laspeyres chaîné

En pratique, l'IPC (indice des prix à la consommation) est un indice de Laspeyres et pose donc la question fondamentale : combien de temps garder le même panier ?

En France, le panier est mis à jour chaque année et l'indice est calculé sous la forme d'un indice de Laspeyres chaîné annuellement. Les pondérations utilisées pour agréger les 21 000 indices élémentaires sont mises à jour chaque année.

On adopte en général pour les séries mensuelles le mois de décembre précédent comme base intermédiaire.

Par exemple :  $I_{déc2007/98}(P) = \frac{I_{déc2007/déc2006} \times I_{déc2006/98}}{100}$ , les indices étant calculés avec la formule de Laspeyres.

Nous rappelons ici que « mathématiquement » l'indice de Laspeyres n'est pas transférable, même si, dans la pratique, sur des périodes courtes, on obtient des approximations acceptables.

## 2.3 LES INDICES DE PAASCHE

Nous allons définir deux indices de Paasche, l'un relatif au prix, l'autre aux quantités.

### Indice des prix de Paasche

#### Définition

On appelle **indice des prix de Paasche**, année  $t$ , base 100 l'année 0, l'indice noté  $P_{t/0}(P)$  obtenu en fixant les quantités à l'année courante. Il est défini par :  $P_{t/0}(P) = \frac{\sum_{i=1}^n Q_i^t P_i^t}{\sum_{i=1}^n Q_i^0 P_i^0} \times 100$ .

On a :

$$P_{t/0}\{P\} = \frac{\sum_{i=1}^n Q_i^t P_i^t}{\sum_{i=1}^n Q_i^0 P_i^0} \times 100 = \frac{V^t}{\sum_{i=1}^n Q_i^t P_i^0} \times 100 = \frac{V^t}{\sum_{i=1}^n Q_i^t P_i^t \left( \frac{P_i^0}{P_i^t} \right)} \times 100 = \frac{1}{\sum_{i=1}^n \frac{Q_i^t P_i^t}{V^t} \left( \frac{P_i^0}{100 P_i^t} \right)} = \frac{1}{\sum_{i=1}^n I_{t/0}(P_i)}$$

on reconnaît dans la parenthèse l'inverse de l'indice élémentaire du prix du bien  $i$  et le coefficient  $\frac{Q_i^t P_i^t}{V^t}$  est  $C_i^t$ , le coefficient budgétaire du bien  $i$ , l'année courante. D'où la propriété suivante.

#### Propriété

L'indice des prix de Paasche est la moyenne harmonique pondérée des indices élémentaires de prix des biens composant le panier. Les coefficients de pondération sont les coefficients budgétaires de l'année courante.

#### Exemple 8.10

### Indice des prix de Paasche

Reprends l'exemple 8.7 et calculons l'indice de Paasche des prix en 2003, base 100 en 1980, de deux façons : à partir de la définition et comme moyenne harmonique des indices élémentaires de prix.

$$L_{2003/1980}(P) = \frac{\sum_{i=1}^2 Q_i^t P_i^t}{\sum_{i=1}^2 Q_i^0 P_i^0} \times 100 = \frac{18 \times 0,75 + 2,10 \times 16,5}{18 \times 0,1525 + 2,10 \times 6,74} \times 100 = 285,14, \text{ soit une augmentation de } 185,14\%.$$

En utilisant la propriété de l'indice des prix de Paasche, on vérifie que la valeur de cet indice est la moyenne harmonique des indices élémentaires de prix pon-

dérée par les coefficients budgétaires de l'année courante : pour la baguette,  $I_{2003/1980}(P_1) = \frac{0,75}{0,1525} \times 100 = 491,80$  et  $C_1^0 = 0,2808$  ; pour la viande de bœuf,  $I_{2003/1980}(P_2) = \frac{16,5}{6,74} \times 100 = 244,96$  et  $C_2^t = 0,7192$ , ce qui donne pour l'indice des prix de Laspeyres :  $P_{t/0}(P) = \frac{1}{\sum_{i=1}^2 \frac{C_i^t}{I_{t/0}(P_i)}} = \frac{1}{\frac{0,2808}{491,8} + \frac{0,7192}{244,96}} = 285,14$ .

On note que l'indice des prix de Paasche est inférieur à l'indice des prix de Laspeyres, ce qui n'est pas un hasard ; nous reviendrons plus loin sur la comparaison entre ces indices (voir section 2.4).

### Indice des quantités de Paasche

#### Définition

On appelle **indice des quantités de Paasche**, année  $t$ , base 100 l'année 0, l'indice noté  $P_{t/0}(Q)$  obtenu en fixant les prix à l'année courante. Il est défini par :

$$P_{t/0}(Q) = \frac{\sum_{i=1}^n P_i^t Q_i^t}{\sum_{i=1}^n P_i^0 Q_i^0} \times 100.$$

#### Propriété

L'indice des quantités de Paasche est la moyenne harmonique pondérée des indices élémentaires de quantités des biens composant le panier, les coefficients de pondération étant les coefficients budgétaires de l'année courante.

#### Exemple 8.11

##### Indice des quantités de Paasche

Reprendons l'exemple 8.7 et calculons l'indice de Paasche des quantités en 2003, base 100

$$\text{en } 1980 : P_{2003/1980}(Q) = \frac{\sum_{i=1}^2 P_i^t Q_i^t}{\sum_{i=1}^2 P_i^0 Q_i^0} = \frac{0,75 \times 18 + 16,50 \times 2,1}{0,75 \times 21 + 16,50 \times 1,23} \times 100 = 133,22, \text{ soit une}$$

augmentation de 33,22 % des quantités.

On note que l'indice des quantités de Paasche est inférieur à l'indice des quantités de Laspeyres, ce qui n'est pas un hasard ; nous reviendrons plus loin sur la comparaison entre ces indices (voir section 2.4).

## 2.4 LIENS ET COMPARAISONS ENTRE LES INDICES DE LASPEYRES ET DE PAASCHE

Les indices de Paasche et de Laspeyres ne possèdent pas les propriétés de circularité et de réversibilité des indices élémentaires. Ils ne vérifient pas non plus la propriété relative au produit mais sont liés par une relation faisant intervenir l'indice de valeur globale.

### Indice de valeur globale

L'indice de valeur globale est donné par :  $I_{t/0}(V) = \frac{V^t}{V^0} \times 100 = \frac{\sum_{i=1}^n P_i^t Q_i^t}{\sum_{i=1}^n P_i^0 Q_i^0} \times 100$ . On

rappelle (voir section 2.1) que  $V^0$  et  $V^t$  sont les valeurs globales d'un panier aux années de base et courante, telles que  $V^0 = \sum_{i=1}^n P_i^0 Q_i^0$  et  $V^t = \sum_{i=1}^n P_i^t Q_i^t$ .

#### Propriété

L'indice de valeur globale est lié aux indices de Laspeyres et de Paasche par la relation suivante :  $100I_{t/0}(V) = L_{t/0}(P) \times P_{t/0}(Q) = L_{t/0}(Q) \times P_{t/0}(P)$ .

La preuve est immédiate, elle s'obtient en utilisant les définitions des indices de Laspeyres et de Paasche.

### Comparaison et utilisation des indices de Laspeyres et de Paasche

L'indice de Laspeyres est le plus couramment utilisé, car il permet de conserver la même pondération pour toutes les années : celle de l'année de base. Cet avantage du point de vue des calculs devient vite un inconvénient, car le panier « figé » s'éloigne de plus en plus de la réalité économique.

Pour un indice de prix, par exemple, l'indice de Laspeyres pondère les différents articles proportionnellement aux habitudes de consommation du passé, alors que celui de Paasche prend en compte les habitudes de consommation actuelles.

L'incorporation inévitable dans les indices de prix d'articles dont la quantité produite a nettement augmenté et dont le prix relatif a de ce fait souvent diminué, introduit des disparités dans les résultats obtenus pour les indices de Paasche et de Laspeyres.

Mathématiquement, on démontre que la moyenne harmonique est inférieure ou égale à la moyenne arithmétique. L'indice de Laspeyres étant une moyenne arithmétique des indices élémentaires et l'indice de Paasche une moyenne harmonique, « en général », l'indice de Paasche sera inférieur ou égal à l'indice de Laspeyres. Cependant, il faut prendre en compte que la situation est plus complexe. Les coefficients de pondération étant différents, ils peuvent influer sur la tendance de l'indice de Laspeyres à surestimer les variations et celle de l'indice de Paasche à la sous-estimer.

Dans le cas d'un indice de prix, le jeu des substitutions renforce le phénomène : on cesse en général d'acheter un produit dont le prix augmente pour le remplacer par un produit substituable, au prix plus avantageux, et l'indice de Laspeyres, qui utilise les quantités de la période de départ, donne un poids trop grand aux produits dont les

prix augmentent beaucoup, alors que la part de ces produits va diminuer dans le panier du consommateur.

### Propriété

**Pseudo-réversibilité** : si l'on inverse le temps dans un indice de Laspeyres, on obtient une relation qui s'apparente à la réversibilité, mais avec un indice de Paasche,  $L_{0/t}(P) \times P_{t/0}(P) = 10^4$ , soit  $L_{0/t}(P) = \frac{10^4}{P_{t/0}(P)}$ .

## 2.5 L'INDICE IDÉAL DE FISHER

En 1922, l'économiste américain Irving Fisher propose un indice synthétique qu'il qualifie d'idéal, dans la mesure où il est réversible.

### Définition

**L'indice synthétique de Fisher** est défini comme étant la moyenne géométrique des indices de Laspeyres et de Paasche.

- Pour les prix :  $F_{t/0}(P) = \sqrt{L_{t/0}(P) \times P_{t/0}(P)}$ .
- Pour les quantités :  $F_{t/0}(Q) = \sqrt{L_{t/0}(Q) \times P_{t/0}(Q)}$ .

### Propriété

**Réversibilité** de l'indice de Fisher :

$$F_{0/t}(P) = \sqrt{L_{0/t}(P) \times P_{0/t}(P)} = \sqrt{\frac{10^4}{P_{t/0}(P)} \times \frac{10^4}{L_{t/0}(P)}} = \frac{10^4}{\sqrt{L_{t/0}(P) \times P_{t/0}(P)}} = \frac{10^4}{F_{t/0}(P)}.$$

### Exemple 8.12

#### Indice des prix de Fisher

Reprendons l'exemple 8.7 et calculons l'indice des prix de Fisher, en 2003, base 100 en 1980 :

$$F_{2003/1980}(P) = \sqrt{L_{2003/1980}(P) \times P_{2003/1980}(P)} = \sqrt{313,66 \times 285,14} = 299,06, \text{ soit une augmentation de } 199,06\%.$$

On notera que l'indice de Fisher est toujours compris entre l'indice de Paasche et celui de Laspeyres puisqu'il est défini comme leur moyenne géométrique.

# Conclusion

À l'issue de ce chapitre, le lecteur doit connaître les différents indices, élémentaires et synthétiques, ainsi que leurs propriétés qui sont résumées dans le tableau ci-après. Mais il est évidemment extrêmement important à ce stade de se familiariser avec les grands indices économiques, boursiers, et de donner un sens à ces formules.

Indice	Laspeyres (1864)	Paasche (1874)	Fisher (1922)
Notation	L	P	F
Référence	Année de base	Année courante	
Moyenne	Arithmétique	Harmonique	Géométrique
Pondérations	Coefficients budgétaires année de base	Coefficients budgétaires année courante	
Réversibilité	Non	Non	Oui
Circularité	Non	Non	Non
Aggrégation	Oui	Oui	Non
Effet	Surévalue la hausse	Sous-évalue la hausse	

On note que :

- La moyenne géométrique de deux nombres est comprise entre ces deux nombres, on a donc en général :  $P_{t/0} \leq F_{t/0} \leq L_{t/0}$ .
- Les trois indices synthétiques sont liés par la relation :  $F_{t/0}(P) \times F_{t/0}(Q) = L_{t/0}(P) \times P_{t/0}(Q) = L_{t/0}(Q) \times P_{t/0}(P) = 100I_{t/0}(V)$ . Cette relation se démontre facilement à partir de la définition de l'indice de Fisher et de la relation liant les indices de Laspeyres, Paasche et l'indice de valeur globale (section 2.4).
- L'indice de Fisher n'a pas une structure de moyenne comme les indices de Paasche et de Laspeyres ; il ne satisfait pas à la propriété d'agrégation. En effet, les indices de Laspeyres et de Paasche ont des structures de moyennes, ce qui permet d'utiliser des moyennes partielles, c'est-à-dire de scinder l'ensemble considéré en plusieurs sous-ensembles ; ces indices possèdent la **propriété d'agrégation**. Par exemple, pour calculer l'indice des prix à la consommation, qui regroupe 305 postes de dépenses, on utilise la formule de Laspeyres, mais, au préalable, on procède à des regroupements par grandes fonctions : alimentation, produits manufacturés, services, etc., on calcule les indices partiels de Laspeyres de chacun de ces regroupements, puis on effectue la moyenne arithmétique des indices partiels en prenant pour coefficients de pondération les parts de chacun de ces regroupements dans la valeur de la consommation totale. On a alors agrégé les produits en groupes, et on peut publier des indices partiels.

# Problèmes et exercices

Les indices autorisent les comparaisons de données longitudinales, en figeant un point de comparaison selon la base annuelle retenue.

- L'exercice 1 expose le calcul des indices élémentaires et leurs propriétés.
- L'exercice 2 s'intéresse aux indices particuliers que sont les indices synthétiques.
- L'exercice 3 propose une lecture de ces indices par les coefficients budgétaires et montre que ces indices sont liés entre eux.



## EXERCICE 1 INDICES ÉLÉMENTAIRES

### Énoncé

Les séries suivantes indiquent l'évolution du revenu moyen disponible par ménage et celle du nombre de ménages (France). Par ailleurs, on définit le revenu disponible des Français par la multiplication du revenu moyen disponible par ménage avec le nombre de ménages.

Année	Revenu moyen disponible par ménage (€)	Nombre de ménages (milliers)
1975	23 016	17 745
1990	26 529	21 542
1999	26 612	23 808

Source : Insee, recensement de la population, 1999

1. Calculez les indices relatifs au revenu moyen disponible par ménage, notés IRM :
  - a.  $IRM_{1999/1990}$  ;
  - b.  $IRM_{1990/1975}$  ;
  - c.  $IRM_{1999/1975}$  à l'aide de la propriété de circularité ;
  - d.  $IRM_{1975/1999}$  à l'aide de la propriété de réversibilité.
2. Calculez les indices relatifs au nombre de ménages, notés INM :
  - a.  $INM_{1999/1990}$  ;
  - b.  $INM_{1990/1975}$  ;
  - c.  $INM_{1999/1975}$  à l'aide de la propriété de circularité ;
  - d.  $INM_{1975/1999}$  à l'aide de la propriété de réversibilité.
3. En utilisant la propriété liée à la multiplication, calculez les indices relatifs au revenu disponible des Français, notés IRF :
  - a.  $IRF_{1999/1990}$  ;
  - b.  $IRF_{1990/1975}$  ;
  - c.  $IRF_{1999/1975}$  ;
  - d.  $IRF_{1975/1999}$ .

**Solution**

**1. a.**  $\text{IRM}_{1999/1990} = \frac{V_{1999}}{V_{1990}} \times 100 = \frac{26\,612}{26\,529} \times 100$ , soit  $\text{IRM}_{1999/1990} = 100,31$ . Le revenu moyen disponible par ménage a augmenté de 0,31 % entre 1990 et 1999.

**b.**  $\text{IRM}_{1990/1975} = \frac{V_{1990}}{V_{1975}} \times 100 = \frac{26\,529}{23\,016} \times 100$ , soit  $\text{IRM}_{1990/1975} = 115,26$ . Le revenu moyen disponible par ménage a augmenté de 15,26 % entre 1975 et 1990.

**c.** En s'appuyant sur la propriété de circularité,

$$\text{IRM}_{1999/1975} = \text{IRM}_{1999/1990} \times \text{IRM}_{1990/1975} / 100 = 100,31 \times 115,26 / 100, \text{ soit } \text{IRM}_{1999/1975} = 115,62.$$

Le revenu moyen disponible par ménage a augmenté de 15,62 % entre 1975 et 1999.

**d.** En s'appuyant sur la propriété de réversibilité,  $\text{IRM}_{1975/1999} = \frac{10\,000}{\text{IRM}_{1999/1975}} = \frac{10\,000}{115,62}$ , soit  $\text{IRM}_{1975/1999} = 86,49$ . Le revenu moyen disponible par ménage en 1975 représente 86,49 % du revenu disponible par ménage en 1999.

**2. a.**  $\text{INM}_{1999/1990} = \frac{V_{1999}}{V_{1990}} \times 100 = \frac{23\,808}{21\,542} \times 100$ , soit  $\text{INM}_{1999/1990} = 110,52$ . Le nombre de ménages a augmenté de 10,52 % entre 1990 et 1999.

**b.**  $\text{INM}_{1990/1975} = \frac{V_{1990}}{V_{1975}} \times 100 = \frac{21\,542}{17\,745} \times 100$ , soit  $\text{INM}_{1990/1975} = 121,40$ . Le nombre de ménages a augmenté de 21,40 % entre 1975 et 1990.

**c.** En s'appuyant sur la propriété de circularité,

$$\text{INM}_{1999/1975} = \text{INM}_{1999/1990} \times \text{INM}_{1990/1975} / 100 = 110,52 \times 121,40 / 100 = \text{INM}_{1999/1975} = 134,17.$$

Le nombre de ménages a augmenté de 34,17 % entre 1975 et 1999.

**d.** En s'appuyant sur la propriété de réversibilité,  $\text{INM}_{1975/1999} = \frac{10\,000}{\text{INM}_{1999/1975}} = \frac{10\,000}{134,17}$ , soit  $\text{INM}_{1975/1999} = 74,53$ . Le nombre de ménages en 1975 représente 74,53 % du nombre de ménages en 1999.

**3. a.** En s'appuyant sur la propriété des indices relative à la multiplication, on obtient :  $\text{IRF}_{1999/1990} = \text{IRM}_{1999/1990} \times \text{INM}_{1999/1990} / 100 = 100,31 \times 110,52 / 100$ , soit  $\text{IRF}_{1999/1990} = 110,86$ .

Le revenu disponible des Français a augmenté de 10,86 % entre 1990 et 1999.

**b.** En s'appuyant sur la propriété des indices relative à la multiplication, on obtient :  $\text{IRF}_{1990/1975} = \text{IRM}_{1990/1975} \times \text{INM}_{1990/1975} / 100 = 115,26 \times 121,40 / 100$ , soit  $\text{IRF}_{1990/1975} = 139,93$ .

Le revenu disponible des Français a augmenté de 39,93 % entre 1975 et 1990.

**c.** De même, on obtient :  $\text{IRF}_{1999/1975} = \text{IRM}_{1999/1975} \times \text{INM}_{1999/1975} / 100 = 115,62 \times 134,17 / 100$ ,

soit  $\text{IRF}_{1999/1975} = 155,13$ . Le revenu disponible des Français a augmenté de 55,13 % entre 1975 et 1999.

**d.** De même, on obtient :  $\text{IRF}_{1975/1999} = \text{IRM}_{1975/1999} \times \text{INM}_{1975/1999} / 100 = 86,49 \times 74,53 / 100$ , soit

$\text{IRF}_{1975/1999} = 64,46$ . Le revenu disponible des Français en 1975 représente 64,46 % du revenu disponible des Français en 1999.



## EXERCICE 2 INDICES SYNTHÉTIQUES

### Énoncé

Le tableau suivant recense les prix moyens des chambres d'hôtel en 2006 et 2007, selon leur catégorie et le nombre de nuitées annuelles.

Catégorie	Prix 2006 (€)	Prix 2007 (€)	Nuitées 2006 (milliers)	Nuitées 2007 (milliers)
0 & 1 étoile	33	35	1 676	1 909
2 étoiles	57	59	3 631	3 813
3 étoiles	86	88	3 475	3 850
4 étoiles & luxe	175	187	2 371	2 229

Sources : Insee, 2007, et KPMG, 2007

- Calculez l'indice des prix de Laspeyres en 2007 base 100 en 2006. Interprétez.
- À prix constants (base 2006), quelle est l'augmentation des nuitées entre 2006 et 2007 ? Quel indice connu avez-vous calculé ?
- Calculez l'indice des quantités de Paasche en 2007 base 100 en 2007. Interprétez.
- À nuitées constantes (base 2007), quelle est l'augmentation du prix des chambres entre 2006 et 2007 ? Quel indice connu avez-vous calculé ?
- Calculez les indices de Fisher en 2007, base 100 en 2006 :
  - des prix ;
  - des quantités.

### Solution

1. Afin de pouvoir calculer l'indice des prix de Laspeyres en 2007 (base 2006), il est nécessaire de connaître les sommes des produits des prix 2007 par les quantités 2006 et des prix 2006 par les quantités 2006.

Les produits et leurs sommes sont calculés dans les colonnes F et G de la figure 8.1.

Figure 8.1

### Résultats sous Excel.

	A	B	C	D	E	F	G	H	I
1	Catégorie	Prix 2006	Prix 2007	Nuitées 2006	Nuitées 2007	p <sub>2007</sub> q <sub>2006</sub>	p <sub>2006</sub> q <sub>2006</sub>	p <sub>2006</sub> q <sub>2007</sub>	p <sub>2007</sub> q <sub>2007</sub>
2	0 & 1 étoile	33,00	35,00	1 676	1 909	58 660	55 308	62 997	66 815
3	2 étoiles	57,00	59,00	3 631	3 813	214 229	206 967	217 341	224 967
4	3 étoiles	86,00	88,00	3 475	3 869	306 800	298 950	331 100	339 900
5	4 étoiles & lu	175,00	187,00	2 371	2 229	443 377	414 925	390 075	416 823
6	Somme					1 022 066	976 050	1 001 513	1 047 406

$$\text{D'où : } L_{2007/2006}(P) = 100 \times \frac{\sum_{i=1}^4 p_{2007}^i \times q_{2006}^i}{\sum_{i=1}^4 p_{2006}^i \times q_{2006}^i} = 100 \times \frac{1 022 066}{976 050}, \text{ soit } L_{2007/2006}(P) = 104,71.$$

À quantités constantes (base 2006), les prix des chambres d'hôtel, toutes catégories confondues, ont augmenté de 4,71 % entre 2006 et 2007.

**2.** Afin de pouvoir calculer l'augmentation des nuitées entre 2006 et 2007 à prix constant (base 2006), il est nécessaire de connaître les sommes des produits des prix 2006 par les quantités 2007 et des prix 2006 par les quantités 2006. Il s'agit de calculer l'**indice des quantités de Laspeyres** entre 2006 et 2007 (base 2006).

Les produits des prix 2006 par les quantités 2007 et leur somme sont présentés à la suite des précédents calculs, dans la colonne H de la figure 8.1.

$$\text{D'où : } L_{2007/2006}(Q) = 100 \times \frac{\sum_{i=1}^4 p_{2006}^i \times q_{2007}^i}{\sum_{i=1}^4 p_{2006}^i \times q_{2006}^i} = 100 \times \frac{1001513}{976050}, \text{ soit } L_{2007/2006}(Q) = 102,61.$$

À prix constants (base 2006), le nombre de nuitées, toutes catégories d'hôtel confondues, a augmenté de 2,61 % entre 2006 et 2007.

**3.** Afin de pouvoir calculer l'indice de Paasche des quantités entre 2006 et 2007 (base 2007), il est nécessaire de connaître les sommes des produits des prix 2007 par les quantités 2007 et des prix 2007 par les quantités 2006.

Les produits des prix 2007 par les quantités 2007 et leur somme sont présentés à la suite des précédents calculs, dans la colonne I de la figure 8.1.

$$\text{D'où : } P_{2007/2006}(Q) = 100 \times \frac{\sum_{i=1}^4 p_{2007}^i \times q_{2007}^i}{\sum_{i=1}^4 p_{2007}^i \times q_{2006}^i} = 100 \times \frac{1047405}{1022066}, \text{ soit } P_{2007/2006}(Q) = 102,48.$$

À prix constants (base 2007), le nombre de nuitées, toutes catégories d'hôtel confondues, a augmenté de 2,48 % entre 2006 et 2007.

**4.** Afin de pouvoir calculer l'augmentation des prix des chambres entre 2006 et 2007 à nuitées constantes (base 2007), il est nécessaire de connaître les sommes des produits des prix 2007 par les quantités 2007 et des prix 2006 par les quantités 2007. Il s'agit de calculer l'**indice de Paasche des prix** entre 2006 et 2007 (base 2007).

$$\text{D'où : } P_{2007/2006}(P) = 100 \times \frac{\sum_{i=1}^4 p_{2007}^i \times q_{2007}^i}{\sum_{i=1}^4 p_{2006}^i \times q_{2007}^i} = 100 \times \frac{1047405}{1001513}, \text{ soit } P_{2007/2006}(P) = 104,58.$$

À quantités constantes (base 2007), les prix des chambres d'hôtel, toutes catégories confondues, ont augmenté de 4,58 % entre 2006 et 2007.

**5. a.**  $F_{2007/2006}(P) = \sqrt{L_{2007/2006}(P) \times P_{2007/2006}(P)} = \sqrt{104,71 \times 104,58},$

soit  $F_{2007/2006}(P) = 104,65.$

**b.**  $F_{2007/2006}(Q) = \sqrt{L_{2007/2006}(Q) \times P_{2007/2006}(Q)} = \sqrt{102,61 \times 102,48},$

soit  $F_{2007/2006}(Q) = 102,54.$

Ces indices de Fisher sont dans chaque cas compris entre les indices de Laspeyres et de Paasche, ce qui est une « obligation mathématique » due à leur statut de moyenne. Pour les prix, par exemple, l'indice de Laspeyres a tendance à surestimer les augmentations, l'indice de Paasche à les sous-estimer, l'indice « idéal » de Fisher se voulant un juste compromis entre ces deux tendances.



### EXERCICE 3 COEFFICIENTS BUDGÉTAIRES ET RELATION ENTRE INDICES

#### Énoncé

Le tableau suivant indique le montant de la consommation effective, par fonctions, des ménages (France entière) entre 2003 et 2006, en milliards d'euros courants :

Désignation du poste	2003	2004	2005	2006
Prod. alimentaires et boissons non alcoolisées	128,305	130,626	132,517	136,163
Boissons alcoolisées et tabac	29,378	29,877	29,684	30,266
Articles d'habillement et chaussures	45,472	46,182	46,521	46,923
Logement, eau, gaz, électricité et autres combustibles	209,182	220,424	234,899	250,150
Meubles, articles de ménage et entretien courant de l'habitation	53,331	55,753	57,379	58,870
Santé	29,154	30,995	32,583	33,936
Transport	127,489	134,619	142,175	146,247
Communications	24,380	25,447	26,868	27,970
Loisirs et culture	82,862	87,084	89,380	92,637
Éducation	5,730	6,202	6,729	7,385
Hôtels, cafés et restaurants	56,086	57,971	59,682	61,970
Autres biens et services	98,530	102,350	105,460	110,851

Source : Insee, 2007

Les indices chaînés des prix à la consommation entre ces deux mêmes années vous sont également communiqués (base 100 l'année précédente) :

Désignation du poste	2003	2004	2005	2006
Prod. alimentaires et boissons non alcoolisées	103,666	101,809	101,448	102,751
Boissons alcoolisées et tabac	99,660	101,699	99,354	101,961
Articles d'habillement et chaussures	102,888	101,563	100,734	100,863
Logement, eau, gaz, électricité et autres combustibles	105,764	105,375	106,567	106,493
Meubles, articles de ménage et entretien courant de l'habitation	103,505	104,543	102,916	102,598
Santé	104,392	106,316	105,123	104,152

Désignation du poste	2003	2004	2005	2006
Transport	101,151	105,593	105,612	102,865
Communications	107,375	104,377	105,584	104,102
Loisirs et culture	103,480	105,096	102,637	103,643
Éducation	106,470	108,237	108,497	109,749
Hôtels, cafés et restaurants	103,958	103,361	102,951	103,834
Autres biens et services	103,554	103,877	103,039	105,112

Source : Insee, 2007

- Calculez le coefficient budgétaire de chaque fonction de consommation pour chacune des années de 2003 à 2006.
- Proposez le tableau des indices des prix à la consommation, base 100 en 2003, pour chacune des années 2003, 2004, 2005 et 2006.
- Calculez l'indice des prix à la consommation en 2006, base 100 en 2003, selon la méthode de Laspeyres.
- Calculez, selon la méthode de Paasche, l'indice des prix en 2006, base 100 l'année 2003.
- De combien a augmenté la consommation des ménages en volume entre l'année 2003 et l'année 2006 ?

## Solution

### 1. Le coefficient budgétaire représente le poids de la fonction de consommation dans l'ensemble des dépenses du ménage.

Il convient dans un premier temps de calculer la somme des dépenses totales des ménages. Par exemple, la dépense des ménages en 2003 est de  $128,305 + 29,378 + \dots + 98,53 = 889,897$  milliards d'euros.

Ensuite, il suffit de calculer la part de chaque poste dans le montant de ces dépenses. Par exemple, les produits alimentaires et boissons non alcoolisées représentent 128,305 milliards d'euros sur les 889,897 milliards d'euros de dépense des ménages en 2003, soit 14,42 %.

Ces calculs sont détaillés dans la figure 8.2.

Figure 8.2

### Résultats sous Excel.

A	B	C	D	E
32 Désignation du poste	2003	2004	2005	2006
33 Prod. alimentaires et bois	14,42%	14,08%	13,75%	13,57%
34 Boissons alcoolisées et t	3,30%	3,22%	3,08%	3,02%
35 Articles d'habillement et c	5,11%	4,98%	4,83%	4,68%
36 Logement, eau, gaz, élec	23,51%	23,76%	24,37%	24,93%
37 Meubles, articles de mén	5,99%	6,01%	5,95%	5,87%
38 Santé	3,28%	3,34%	3,38%	3,38%
39 Transport	14,33%	14,51%	14,75%	14,58%
40 Communications	2,74%	2,74%	2,79%	2,79%
41 Loisirs et culture	9,31%	9,39%	9,27%	9,23%
42 Éducation	0,64%	0,67%	0,70%	0,74%
43 Hôtels, cafés et restaurar	6,30%	6,25%	6,19%	6,18%
44 Autres biens et services	11,07%	11,03%	10,94%	11,05%
45 Somme	100%	100%	100%	100%

**2.** Les indices en 2003 valent tous 100, puisqu'il s'agit de l'année de référence.

Les indices en 2004 conservent leur valeur puisqu'il était en base 100 l'année précédente, c'est-à-dire 2003.

Pour calculer les indices élémentaires en 2005 et 2006, base 100 l'année 2003, on utilise la propriété de circularité (transférabilité) des indices élémentaires :

$$I_{2005/2003} = I_{2005/2004} \times I_{2004/2003} / 100 .$$

Par exemple, pour les produits alimentaires et boissons non alcoolisées :

$$I_{2005/2003} = I_{2005/2004} \times I_{2004/2003} / 100 = 101,45 \times 101,81 / 100 , \text{ soit } I_{2005/2003} = 103,28 .$$

Ces calculs sont détaillés dans la figure 8.3.

**Figure 8.3**

**Résultats sous Excel.**

A	B	C	D	E
49 Désignation du poste	2003	2004	2005	2006
50 Produits alimentaires et boissons	100	101,81	103,28	104,24
51 Boissons alcoolisées et tisanes	100	101,70	101,04	101,30
52 Articles d'habillement et chaussures	100	101,56	102,31	101,60
53 Logement, eau, gaz, électricité	100	105,37	112,29	113,49
54 Meubles, articles de ménage	100	104,54	107,59	105,59
55 Santé	100	106,32	111,76	109,49
56 Transport	100	105,59	111,52	108,64
57 Communications	100	104,38	110,21	109,91
58 Loisirs et culture	100	105,10	107,87	106,38
59 Education	100	108,24	117,43	119,07
60 Hôtels, cafés et restaurants	100	103,36	106,41	106,90
61 Autres biens et services	100	103,88	107,03	108,31

**3.** L'indice de Laspeyres est la moyenne arithmétique des indices élémentaires pondérés par les coefficients budgétaires de l'année de base.

Ainsi,  $L_{2006/2003}(P) = 104,24 \times 0,1375 + 101,30 \times 0,0308 + \dots + 108,31 \times 0,1094 ,$  soit

$$L_{2006/2003}(P) = 108,25 .$$

**4.** L'indice de Paasche est la moyenne harmonique des indices élémentaires pondérés par les coefficients budgétaires de l'année de base.

$$\text{Ainsi, } P_{2006/2003}(P) = \frac{1}{\frac{1}{104,24/0,1357} + \frac{1}{101,30/0,0302} + \dots + \frac{1}{108,31/0,1105}} ,$$

$$\text{soit } P_{2006/2003}(P) = 108,18 .$$

**5.** On sait qu'un indice de valeur globale est le produit d'un indice de volume par un indice de prix ; plus précisément, selon l'indice des prix que nous retenons, l'indice de volume de la consommation des ménages varie.

Si nous retenons l'indice des prix de Laspeyres, l'indice de volume est un indice de Paasche, et symétriquement, en utilisant l'indice des prix de Paasche, nous obtenons un indice de volume de Laspeyres, selon la formule :

$$L_{2006/2003}(Q) \times P_{2006/2003}(P) = 100 \times I_{2006/2003}(V) .$$

Nous allons calculer les indices des quantités de Laspeyres et de Paasche. Commençons par l'indice de Laspeyres.

L'indice des prix de Paasche est :  $P_{2006/2003}(P)=108,18$ . Calculons l'indice de valeur globale de la consommation des ménages en 2006, base 100 l'année 2003 :

$$I_{2006/2003}(V) = \frac{1003,368}{889,897} \times 100, \text{ soit } I_{2006/2003}(V) = 112,75. \text{ On obtient alors :}$$

$$L_{2006/2003}(Q) = \frac{I_{2006/2003}(V)}{P_{2006/2003}(P)} \times 100 = \frac{112,75}{108,18} \times 100, \text{ soit } L_{2006/2003}(Q) = 104,23. \text{ Entre l'année 2003 et l'année 2006, la consommation des ménages a augmenté selon la méthode de Laspeyres de 4,23 \% en volume.}$$

En utilisant l'indice des prix de Laspeyres, on obtient :

$$P_{2006/2003}(Q) = \frac{I_{2006/2003}(V)}{L_{2006/2003}(P)} \times 100 = \frac{112,75}{104,23} \times 100, \text{ soit } P_{2006/2003}(Q) = 108,25. \text{ Entre l'année 2003 et l'année 2006, la consommation des ménages a augmenté selon la méthode de Paasche de 4,16 \% en volume.}$$

## Bibliographie

CHAREILLE P. et PINAULT Y., *Statistique descriptive*, Collection AES, Montchrestien, Paris, 1996.

DAMON J.-P., *La méthode statistique en économie*, Éditions Paris-8 Vincennes, 1976.

DUPONT-KIEFFER A., *Ragnar Frisch et l'économétrie : l'invention de modèles et d'instruments à des fins normatives*, Thèse pour le doctorat en science économique (arrêté du 30 mars 1992), université Paris-1 Sorbonne, 2003.

DODGE Y., *Statistique. Dictionnaire encyclopédique*, Springer, 2004.

DROESBEKE J.-J. et TASSI Ph., *Histoire de la statistique*, Que sais-je ?, PUF, 1990.

FERREOL G. et SCHLACTHER D., *Dictionnaire des techniques quantitatives appliquées aux sciences économiques et sociales*, Armand Colin, 1995.

LIORZOU A., *Initiation à la pratique statistique*, Eyrolles, 1979.

GUERBER L et HENNEQUIN P.-L., *Initiation à la statistique*, Bibliothèque d'enseignement mathématique A.P.M.E.P., 1967.

INSEE METHODES, *Pour comprendre l'indice des prix*, Édition 1998.

SCHLACTHER D., *De l'analyse à la prévision*, Ellipses, 1986.

# Annexes

Loi de Student  $P(X > t) = p$ ; on lit  $t$  dans la table et  $p$  figure sur la première ligne

ddl\proba	0,1	0,05	0,025	0,01	0,005	0,001
1	3,078	6,314	12,706	31,821	63,656	318,289
2	1,886	2,92	4,303	6,965	9,925	22,328
3	1,638	2,353	3,182	4,541	5,841	10,214
4	1,533	2,132	2,776	3,747	4,604	7,173
5	1,476	2,015	2,571	3,365	4,032	5,894
6	1,44	1,943	2,447	3,143	3,707	5,208
7	1,415	1,895	2,365	2,998	3,499	4,785
8	1,397	1,86	2,306	2,896	3,355	4,501
9	1,383	1,833	2,252	2,821	3,25	4,297
10	1,372	1,812	2,228	2,764	3,169	4,144
11	1,363	1,796	2,201	2,718	3,106	4,025
12	1,356	1,782	2,179	2,681	3,055	3,93
13	1,35	1,771	2,16	2,65	3,012	3,852
14	1,345	1,761	2,145	2,624	2,977	3,787
15	1,341	1,753	2,131	2,602	2,947	3,733
16	1,337	1,746	2,12	2,583	2,921	3,686
17	1,333	1,74	2,11	2,567	2,898	3,646
18	1,33	1,734	2,101	2,552	2,878	3,61
19	1,328	1,729	2,093	2,539	2,861	3,579
20	1,325	1,725	2,086	2,528	2,845	3,552
21	1,323	1,721	2,08	2,518	2,831	3,527
22	1,321	1,717	2,074	2,508	2,819	3,505
23	1,319	1,714	2,069	2,5	2,807	3,485
24	1,318	1,711	2,064	2,492	2,797	3,467
25	1,316	1,708	2,06	2,485	2,787	3,45
26	1,315	1,706	2,056	2,479	2,779	3,435
27	1,314	1,703	2,052	2,473	2,771	3,421
28	1,313	1,701	2,048	2,467	2,763	3,408
29	1,311	1,699	2,045	2,462	2,756	3,396
30	1,31	1,697	2,042	2,457	2,75	3,385
60	1,296	1,671	2	2,39	2,66	3,232
90	1,291	1,662	1,987	2,368	2,632	3,183
120	1,289	1,658	1,98	2,358	2,617	3,16

Loi de Khi-2  $P(X > t)$

ddl \ t	0,001	0,005	0,01	0,025	0,05	0,1	0,9	0,95	0,975	0,99	0,995	0,999
1	10,83	7,88	6,64	5,02	3,84	2,71	0,02	0,00	0,00	0,00	0,00	0,00
2	13,82	10,60	9,21	7,38	5,99	4,61	0,21	0,10	0,05	0,02	0,01	0,00
3	16,27	12,84	11,35	9,35	7,82	6,25	0,58	0,35	0,22	0,12	0,07	0,02
4	18,47	14,86	13,28	11,14	9,49	7,78	1,06	0,71	0,48	0,30	0,21	0,09
5	20,52	16,75	15,09	12,83	11,07	9,24	1,61	1,15	0,83	0,55	0,41	0,21
6	22,46	18,55	16,81	14,45	12,59	10,65	2,20	1,64	1,24	0,87	0,68	0,38
7	24,32	20,28	18,48	16,01	14,07	12,02	2,83	2,17	1,69	1,24	0,99	0,60
8	26,12	21,96	20,09	17,54	15,51	13,36	3,49	2,73	2,18	1,65	1,34	0,86
9	27,88	23,59	21,67	19,02	16,92	14,68	4,17	3,33	2,70	2,09	1,74	1,15
10	29,59	25,19	23,21	20,48	18,31	15,99	4,87	3,94	3,25	2,56	2,16	1,48
11	31,26	26,76	24,73	21,92	19,68	17,28	5,58	4,58	3,82	3,05	2,60	1,83
12	32,91	28,30	26,22	23,34	21,03	18,55	6,30	5,23	4,40	3,57	3,07	2,21
13	34,53	29,82	27,69	24,74	22,36	19,81	7,04	5,89	5,01	4,11	3,57	2,62
14	36,12	31,32	29,14	26,12	23,69	21,06	7,79	6,57	5,63	4,66	4,08	3,04
15	37,70	32,80	30,58	27,49	25,00	22,31	8,55	7,26	6,26	5,23	4,60	3,48
16	39,25	34,27	32,00	28,85	26,30	23,54	9,31	7,96	6,91	5,81	5,14	3,94
17	40,79	35,72	33,41	30,19	27,59	24,77	10,09	8,67	7,56	6,41	5,70	4,42
18	42,31	37,16	34,81	31,53	28,87	25,99	10,87	9,39	8,23	7,02	6,27	4,91
19	43,82	38,58	36,19	32,85	30,14	27,20	11,65	10,12	8,91	7,63	6,84	5,41
20	45,31	40,00	37,57	34,17	31,41	28,41	12,44	10,85	9,59	8,26	7,43	5,92
21	46,80	41,40	38,93	35,48	32,67	29,62	13,24	11,59	10,28	8,90	8,03	6,45
22	48,27	42,80	40,29	36,78	33,92	30,81	14,04	12,34	10,98	9,54	8,64	6,98
23	49,73	44,18	41,64	38,08	35,17	32,01	14,85	13,09	11,69	10,20	9,26	7,53
24	51,18	45,56	42,98	39,36	36,42	33,20	15,66	13,85	12,40	10,86	9,89	8,09
25	52,62	46,93	44,31	40,65	37,65	34,38	16,47	14,61	13,12	11,52	10,52	8,65
26	54,05	48,29	45,64	41,92	38,89	35,56	17,29	15,38	13,84	12,20	11,16	9,22
27	55,48	49,65	46,96	43,20	40,11	36,74	18,11	16,15	14,57	12,88	11,81	9,80
28	56,89	50,99	48,28	44,46	41,34	37,92	18,94	16,93	15,31	13,57	12,46	10,39
29	58,30	52,34	49,59	45,72	42,56	39,09	19,77	17,71	16,05	14,26	13,12	10,99
30	59,70	53,67	50,89	46,98	43,77	40,26	20,60	18,49	16,79	14,95	13,79	11,59
31	61,10	55,00	52,19	48,23	44,99	41,42	21,43	19,28	17,54	15,66	14,46	12,20
32	62,49	56,33	53,49	49,48	46,19	42,59	22,27	20,07	18,29	16,36	15,13	12,81
33	63,87	57,65	54,78	50,73	47,40	43,75	23,11	20,87	19,05	17,07	15,82	13,43
34	65,25	58,96	56,06	51,97	48,60	44,90	23,95	21,66	19,81	17,79	16,50	14,06
35	66,62	60,28	57,34	53,20	49,80	46,06	24,80	22,47	20,57	18,51	17,19	14,69

**Loi de Fisher-Snedecor ( $n_1, n_2$ )  $P(X > t) = 1\%$**

$n_1 \setminus n_2$	1	2	3	4	5	6	7	8	9	10	15	20	25	30	60	120
1	4052,185	98,502	34,116	21,198	16,258	13,745	12,246	11,259	10,562	10,044	8,683	8,096	7,770	7,562	7,077	6,851
2	4999,340	99,000	30,816	18,000	13,274	10,925	9,547	8,649	8,022	7,559	6,359	5,849	5,568	5,390	4,977	4,787
3	5403,534	99,164	29,457	16,694	12,060	9,780	8,451	7,591	6,992	6,552	5,417	4,938	4,675	4,510	4,126	3,949
4	5624,257	99,251	28,710	15,977	11,392	9,148	7,847	7,006	6,422	5,994	4,893	4,431	4,177	4,018	3,649	3,480
5	5763,955	99,302	28,237	15,522	10,967	8,746	7,460	6,632	6,057	5,636	4,556	4,103	3,855	3,699	3,339	3,174
6	5858,950	99,331	27,911	15,207	10,672	8,466	7,191	6,371	5,802	5,386	4,318	3,871	3,627	3,473	3,119	2,956
7	5928,334	99,357	27,671	14,976	10,456	8,260	6,993	6,178	5,613	5,200	4,142	3,699	3,457	3,305	2,953	2,792
8	5980,954	99,375	27,489	14,799	10,289	8,102	6,840	6,029	5,467	5,057	4,404	3,564	3,324	3,173	2,823	2,663
9	6022,397	99,390	27,345	14,659	10,158	7,976	6,719	5,911	5,351	4,942	3,895	3,457	3,217	3,067	2,718	2,559
10	6055,925	99,397	27,223	14,546	10,051	7,874	6,620	5,814	5,257	4,848	3,805	3,368	3,129	2,979	2,632	2,472
11	6083,399	99,408	27,132	14,452	9,963	7,790	6,538	5,734	5,178	4,772	3,730	3,294	3,056	2,906	2,555	2,399
12	6106,682	99,419	27,052	14,374	9,888	7,718	6,469	5,667	5,111	4,706	3,666	3,231	2,993	2,843	2,496	2,336
13	6125,774	99,422	26,983	14,306	9,825	7,657	6,410	5,609	5,055	4,650	3,612	3,177	2,939	2,789	2,442	2,282
14	6143,004	99,426	26,924	14,249	9,770	7,605	6,359	5,559	5,005	4,601	3,564	3,130	2,892	2,742	2,394	2,234
15	6156,974	99,433	26,872	14,198	9,722	7,559	6,314	5,515	4,962	4,558	3,522	3,088	2,850	2,700	2,352	2,191
16	6170,012	99,437	26,826	14,154	9,680	7,519	6,275	5,477	4,924	4,520	3,485	3,051	2,813	2,663	2,315	2,154
17	6181,188	99,441	26,786	14,114	9,643	7,483	6,240	5,442	4,890	4,487	4,352	3,018	2,780	2,630	2,281	2,119
18	6191,432	99,444	26,751	14,079	9,609	7,451	6,209	5,412	4,860	4,457	4,243	2,989	2,751	2,600	2,251	2,089
19	6200,746	99,448	26,719	14,048	9,580	7,422	6,181	5,384	4,833	4,430	3,396	2,962	2,724	2,573	2,223	2,060
20	6208,662	99,448	26,690	14,019	9,553	7,396	6,155	5,359	4,808	4,405	3,372	2,938	2,699	2,549	2,198	2,035
21	6216,113	99,451	26,664	13,994	9,528	7,372	6,132	5,336	4,786	4,383	3,350	2,916	2,677	2,526	2,175	2,011
22	6223,097	99,455	26,639	13,970	9,506	7,351	6,111	5,316	4,765	4,363	3,330	2,895	2,657	2,506	2,155	1,989
23	6228,685	99,455	26,617	13,949	9,485	7,331	6,092	5,297	4,746	4,344	3,311	2,877	2,638	2,487	2,134	1,969
24	6234,273	99,455	26,597	13,929	9,466	7,313	6,074	5,279	4,729	4,327	3,294	2,859	2,620	2,469	2,115	1,950
25	6239,861	99,459	26,579	13,911	9,449	7,296	6,058	5,263	4,713	4,313	3,278	2,843	2,604	2,453	2,098	1,932
26	6244,518	99,462	26,562	13,894	9,433	7,281	6,043	5,248	4,698	4,296	3,264	2,829	2,589	2,437	2,083	1,916
27	6249,174	99,462	26,546	13,878	9,418	7,266	6,029	5,234	4,684	4,283	3,250	2,815	2,575	2,423	2,068	1,901
28	6252,900	99,462	26,531	13,864	9,404	7,253	6,016	5,221	4,672	4,270	3,237	2,802	2,562	2,410	2,054	1,886
29	6257,091	99,462	26,517	13,850	9,391	7,240	6,003	5,209	4,660	4,258	3,225	2,790	2,550	2,398	2,041	1,873
30	6260,350	99,466	26,504	13,838	9,379	7,229	5,992	5,198	4,649	4,247	3,214	2,778	2,538	2,386	2,028	1,860
60	6312,970	99,484	26,316	13,652	9,202	7,057	5,824	5,032	4,483	4,082	3,047	2,608	2,364	2,208	1,836	1,656
120	6339,513	99,491	26,221	13,558	9,112	6,969	5,737	4,946	4,398	3,996	2,959	2,517	2,270	2,111	1,726	1,533

**Loi de Fisher-Snedecor  $P(X > t) = 5\%$**

$n_1 \setminus n_2$	1	2	3	4	5	6	7	8	9	10	15	20	25	30	60	120
1	161,446	18,513	10,128	7,709	6,608	5,987	5,591	5,318	5,117	4,965	4,543	4,351	4,242	4,171	4,001	3,920
2	199,499	19,000	9,552	6,944	5,786	5,143	4,737	4,459	4,256	4,103	3,682	3,493	3,385	3,316	3,150	3,072
3	215,707	19,164	9,277	6,591	5,409	4,757	4,347	4,066	3,863	3,708	3,287	3,098	2,991	2,922	2,758	2,680
4	224,583	19,247	9,117	6,388	5,192	4,534	4,120	3,838	3,633	3,478	3,056	2,866	2,759	2,690	2,525	2,447
5	230,160	19,296	9,013	6,256	5,050	4,387	3,972	3,688	3,482	3,326	2,901	2,711	2,603	2,534	2,368	2,290
6	233,988	19,329	8,941	6,163	4,950	4,284	3,866	3,581	3,374	3,217	2,790	2,599	2,490	2,421	2,254	2,175
7	236,767	19,352	8,887	6,094	4,876	4,207	3,787	3,500	3,293	3,135	2,707	2,514	2,405	2,334	2,167	2,087
8	238,884	19,371	8,845	6,041	4,818	4,147	3,726	3,438	3,230	3,072	2,641	2,447	2,337	2,266	2,097	2,016
9	240,543	19,385	8,812	5,999	4,772	4,099	3,677	3,388	3,179	3,020	2,588	2,393	2,282	2,211	2,040	1,959
10	241,882	19,396	8,785	5,964	4,735	4,060	3,630	3,347	3,137	2,978	2,544	2,348	2,236	2,165	1,993	1,910
11	242,981	19,405	8,763	5,936	4,704	4,027	3,603	3,313	3,102	2,943	2,507	2,310	2,198	2,126	1,952	1,869
12	243,905	19,412	8,745	5,912	4,678	4,000	3,575	3,284	3,073	2,913	2,475	2,278	2,165	2,092	1,917	1,834
13	244,690	19,419	8,729	5,891	4,655	3,976	3,550	3,259	3,048	2,887	2,448	2,250	2,136	2,063	1,887	1,803
14	245,363	19,424	8,715	5,873	4,636	3,956	3,529	3,237	3,025	2,865	2,424	2,225	2,111	2,037	1,860	1,775
15	245,949	19,429	8,703	5,858	4,619	3,938	3,511	3,218	3,006	2,845	2,403	2,203	2,089	2,015	1,836	1,750
16	246,466	19,433	8,692	5,844	4,604	3,922	3,494	3,202	2,989	2,828	2,385	2,184	2,069	1,995	1,815	1,728
17	246,917	19,437	8,683	5,832	4,590	3,908	3,480	3,187	2,974	2,812	2,368	2,167	2,051	1,976	1,796	1,709
18	247,324	19,440	8,675	5,821	4,579	3,896	3,467	3,173	2,960	2,798	2,353	2,151	2,035	1,960	1,778	1,690
19	247,688	19,443	8,667	5,811	4,568	3,884	3,455	3,161	2,948	2,785	2,340	2,137	2,021	1,945	1,763	1,674
20	248,016	19,446	8,660	5,803	4,558	3,874	3,445	3,156	2,936	2,774	2,328	2,124	2,007	1,932	1,748	1,659
21	248,307	19,448	8,654	5,795	4,549	3,865	3,435	3,140	2,926	2,764	2,316	2,112	1,995	1,919	1,735	1,645
22	248,579	19,450	8,648	5,787	4,541	3,856	3,426	3,131	2,917	2,754	2,306	2,102	1,984	1,908	1,722	1,632
23	248,823	19,452	8,643	5,781	4,534	3,849	3,418	3,123	2,908	2,745	2,297	2,092	1,974	1,897	1,711	1,620
24	249,052	19,454	8,638	5,774	4,527	3,841	3,410	3,115	2,900	2,737	2,288	2,082	1,964	1,887	1,700	1,608
25	249,260	19,456	8,634	5,769	4,521	3,835	3,404	3,108	2,893	2,730	2,280	2,074	1,955	1,878	1,690	1,598
26	249,453	19,457	8,630	5,763	4,515	3,829	3,397	3,102	2,886	2,723	2,272	2,066	1,947	1,870	1,681	1,588
27	249,631	19,459	8,626	5,759	4,510	3,823	3,391	3,095	2,880	2,716	2,265	2,059	1,939	1,862	1,672	1,579
28	249,798	19,460	8,623	5,754	4,505	3,818	3,386	3,090	2,874	2,710	2,259	2,052	1,932	1,854	1,664	1,570
29	249,951	19,461	8,620	5,750	4,500	3,813	3,381	3,084	2,869	2,705	2,253	2,045	1,926	1,847	1,656	1,562
30	250,096	19,463	8,617	5,746												

# Index

## A

Ajustement  
linéaire *Voir* Droite de régression  
non linéaire, 162, 175  
Amplitude de classe, 6, 13, 25, 28, 30  
Analyse de variance, 154, 164  
Aplatissement, 88  
Asymétrie, 85, 95

## B

Boîte à moustaches, 65, 75, 85, 95  
Box plot *Voir* Boîte à moustaches

## C

Caractère, 18, 22, 25, 27, *Voir* Variable  
Centile, 51, 56  
Centre de classe, 6  
Classe, 6  
Coefficient  
budgétaire, 228, 240  
d'aplatissement  
de Fisher, 89, 98  
de Pearson, 89, 98, 100  
d'asymétrie  
de Fisher, 87, 95  
de Pearson, 86, 95, 100  
de corrélation

de rang, 163, 179  
linéaire, 156, 164  
de détermination, 164, 170, 175  
de Kendall, 86  
de Spearman *Voir* Coefficient de corrélation de rang  
de variation, 71, 76, 77  
de Yule, 86, 95  
saisonnier, 199, 201, 204, 207, 210

Composante  
extra-saisonnière, 190, 200, 202  
générale, 190, 200, 202  
résiduelle, 190, 200, 202  
saisonnière, 190, 198, 200, 202

Corrélation, 147

Courbe  
de concentration, 104  
de régression, 146  
Covariance, 117, 139  
formule développée, 118  
propriétés, 118  
Cycle, 190, 200, 202

## D

Décile, 50  
Degré de liberté, 122  
Densité, 12, 25, 28, 36  
Diagramme  
circulaire, 11, 22

cumulatif *Voir* Fonction de répartition  
de Tukey *Voir* Boîte à moustaches  
en barres *Voir* Diagramme en tuyaux d'orgue  
en bâtons, 12, 18, 98  
en tuyaux d'orgue, 11  
Discrétisation, 8, 27  
Distribution, 8, 11, 12, 13, 22, 25, 26, 29, 51, 65, 77  
Donnée brute, 8  
Droite de régression, 150, 164, 170, 174, 210, 214

## E

Écart  
absolu moyen, 66, 73  
intercentile, 64, 74  
interdécile, 64, 74  
interquantile, 64, 73  
interquartile, 64, 74  
saisonnier, 199, 201  
type, 66, 76, 77, 79  
    conditionnel, 115  
    marginal, 114

Échantillon, 2

Effectif, 3, 8  
    corrigé, 13, 25, 28  
    cumulé, 8, 22  
        croissant, 9, 18  
        décroissant, 9, 18  
    marginal, 109, 132  
    partiel, 109, 132

Étendue, 64, 73

## F

Fonction  
    affine, 150  
    de répartition, 15, 22

Fréquence, 8, 21  
    absolue, 3  
    conditionnelle, 112, 132  
    cumulée, 8, 29  
        croissante, 9, 18  
        décroissante, 9, 18  
    marginale, 111, 132  
    partielle, 111, 132  
    propriétés, 4  
    relative, 3

## G-I

Graphique semi-logarithmique, 175  
Histogramme, 12, 25, 26, 29, 59, 100  
Indépendance, 120  
Indice  
    de Fisher, 234, 238  
    de Gini, 93, 104  
    de Laspeyres, 229, 238, 240  
    de Paasche, 231, 238, 240

de valeur globale, 233  
des prix, 229, 231, 238, 240  
des quantités, 230, 232, 238  
élémentaire, 220, 236  
propriétés, 223, 229, 231, 234, 236  
synthétique, 226, 238

Individu, 2

Intervalle interquantile *Voir* Écart interquantile

## K-L

Kurtosis, 88  
Leptocurtique, 89  
Loi  
    de Fisher, 161  
    de Student, 158  
    normale, 84

## M

Médiale, 91, 104  
Médiane, 45, 54, 55, 56, 59, 95  
Méthode  
    analytique, 190, 210, 214  
    empirique, 193, 194, 204, 207  
Modalité, 3, 18, 22, 25, 27  
Mode, 36, 38, 54, 55, 56, 59, 95  
Modèle  
    additif, 197, 204, 207, 210  
    multiplicatif, 197, 214  
Moindres carrés ordinaires, 150, 170, 174  
Moyenne, 39, 55, 56, 59, 95  
    arithmétique, 39  
    conditionnelle, 115, 135  
    échelonnée, 192  
    géométrique, 42, 60  
    harmonique, 43, 61  
    marginale, 114, 135, 139  
    mobile, 207  
        centrée, 194  
        non centrée, 193  
    propriétés, 41, 79  
    quadratique, 44

## N-P

Nature, 4, 18, 22, 25, 27, 127, 131  
Platocurtique, 89  
Polygone  
    des effectifs, 16, 59  
    des fréquences, 29  
Population, 2, 18, 22, 25, 27  
Pyramide *Voir* Diagramme en tuyaux d'orgue

## Q

Quantile, 44  
Quartile, 50, 55, 56, 59

## R

Régression  
courbe, 146  
droite, 150, 164, 170, 174, 210, 214

## S

Série  
ajustée, 203, 204, 207, 210, 214  
brute, 204, 207, 210, 214  
chronologique, 187  
CVS, 202, 204, 207, 210, 214  
temporelle, 187

## T

Tableau  
croisé *Voir* Tableau de contingence  
de contingence, 109, 127, 131  
élémentaire, 8  
simple, 108  
statistique, 8, 18, 27  
Tendance, 190, 200, 202, 204, 207, 210, 214  
Test, 121  
de corrélation, 159, 164, 170, 175

de Fisher, 170, 175  
de Student, 159, 164, 170, 175  
du khi-deux, 121, 135, 139

Tri  
à plat, 8  
croisé *Voir* Tableau de contingence

## V

Variable  
qualitative, 4  
nominale, 4  
ordinale, 5  
quantitative, 6, 8  
continue, 6, 25, 27  
discrète, 6, 19, 22  
statistique, 3  
variance  
décomposition, 164  
Variance, 66, 76  
conditionnelle, 115, 135  
décomposition, 154  
formule développée, 68, 77  
marginale, 114, 135, 139  
propriétés, 69, 79

# Synthèse de cours & exercices corrigés

## Les auteurs :

**Étienne Bressoud** est maître de conférences à l'université Paris 8 Vincennes-Saint-Denis et à l'*European Business School* (EBS) Paris. Il enseigne la statistique descriptive et les études quantitatives appliquées au marketing.

**Jean-Claude Kahané** est enseignant à l'université Paris 8 Vincennes-Saint-Denis et à l'*École nationale d'assurance* (ENASS). Il enseigne les statistiques, les probabilités et les mathématiques. Il est également membre du jury de CAPES externe de sciences économiques et sociales.

## Direction de collection :

**Roland Gillet**, professeur à l'université Paris 1 Panthéon-Sorbonne

## Dans la même collection :

- **Analyse de données avec SPSS**, M. Carricano et F. Poujal
- **Analyse financière et évaluation d'entreprise**, S. Parienté
- **Performance de portefeuille**, P. Grandin *et al.*
- **Création de valeur et capital-investissement**, M. Cherif et S. Dubreuil
- **Contrôle de gestion**, 2<sup>e</sup> ed., Y. de Rongé et K. Cerrada
- **Économétrie**, É. Dor
- **Finance**, A. Farber *et al.*
- **Les enquêtes par questionnaire avec Sphinx**, S. Ganassali
- **Marketing, une approche quantitative**, une approche quantitative, A. Steyer *et al.*
- **Mathématiques appliquées à la gestion**, A. Szafarz *et al.*
- **Probabilités, statistique et processus stochastiques**, P. Roger
- **Stratégie**, A. Desreumaux *et al.*

# Statistique descriptive avec Excel et la calculatrice

Ce livre est une introduction complète à la statistique descriptive. À la fois accessible à tous et d'une grande rigueur mathématique et statistique, il présente d'abord les notions fondamentales (variables statistiques et graphiques), pour détailler ensuite les caractéristiques de tendance centrale (moyenne, médiane, etc.), de dispersion (écart-type, variance...), de forme et de concentration, les tableaux croisés, la régression linéaire et non linéaire, les séries chronologiques et les indices. Il aborde également les tests statistiques (notamment le test du Khi-deux) et permet d'approfondir vers la statistique inférentielle et l'économétrie.

Toutes les notions sont illustrées à partir de données réelles issues des observatoires statistiques (INSEE, Médiamétrie...). Les exercices occupent une part importante de l'ouvrage et sont appliqués à la gestion, à l'économie et aux sciences humaines. Les corrections détaillent tous les calculs et sont présentées soit à l'aide du tableur Excel soit de la calculatrice (graphique ou scientifique). Ce double choix donne au livre une dimension pratique précieuse et en fait un véritable outil de travail.

L'ouvrage s'adresse aux étudiants de licence en sciences de gestion, en économie, en AES et en sciences humaines, ainsi qu'aux étudiants en IUT et en écoles de management.

**Toutes les données des exercices au format Excel, ainsi que des exemples supplémentaires, sont disponibles sur le site [www.pearson.fr](http://www.pearson.fr).**

**La collection Synthex propose aux gestionnaires et aux économistes de découvrir ou de réviser une discipline et de se familiariser avec ses outils au travers d'exercices résolus.**

Chaque ouvrage présente une synthèse pédagogique et rigoureuse des techniques et fondements théoriques, qu'une sélection d'exercices aux corrigés détaillés permet d'assimiler progressivement.

Le lecteur, étudiant ou professionnel, est ainsi conduit au cœur de la discipline considérée, et, via la résolution de nombreux problèmes, acquiert une compréhension rapide et un raisonnement solide.

ISBN : 978-2-7440-4052-8

