LES CARRÉS

Elisabeth Olivier

L'essentiel de Statistique descriptive

Gualino

lextenso éditions

Sommaire

INTRODUCTION

Cet ouvrage développe l'ensemble des ***concepts et méthodes nécessaires à la compréhension et à l'utilisation de la statistique descriptive***.

Il présente successivement :

– les ***notions fondamentales*** (population, variables, modalités…), puis les ***caractéristiques des distributions à une variable*** : caractéristiques de position (mode, quantiles, moyennes), de dispersion (étendue, variance, écart-type…) et de concentration (courbe de Lorenz, indice de Gini…) ;

– les ***indices élémentaires*** et les ***indices synthétiques*** ;

– les ***distributions à deux variables*** : les tableaux de contingence, la corrélation linéaire et l'ajustement affine par les moindres carrés, ainsi que les séries chronologiques.

Chaque concept et méthode est illustré par une ou plusieurs applications avec son corrigé détaillé. Les données analysées portent sur l'économie et la gestion.

Cet ouvrage s'adresse à tous les étudiants de licence en économie, en AES ou gestion, aux étudiants des écoles de management, des IUT ou des sections BTS qui ont à leur programme un cours de statistique descriptive.

Les concepts de base de la statistique descriptive

CHAPITRE 1

Pour analyser les données fournies par un tableau statistique ou un graphique, il faut au préalable en avoir effectué une lecture complète et correcte.

Cette lecture repose sur la connaissance des concepts de base présentés dans ce chapitre.

Cette première phase est essentielle : elle permet de mettre en évidence toutes les informations contenues dans le tableau ou le graphique et détermine la qualité du travail réalisé ultérieurement sur les données.

1 Population, individus, variable statistique (ou caractère) et modalités

Les premières informations à identifier dans un document statistique sont la ou les population(s) étudiée(s), la (ou les) variables(s) observée(s) ainsi que les modalités.

■ *Population et individus*

La ***population*** est l'ensemble des éléments observés. Ces éléments portent le nom d'***individus*** ou unités statistiques.

Ces individus peuvent être des êtres humains : l'ensemble des Français, les salariés d'une entreprise, les étudiants d'une université, les condamnés à une peine de justice… Mais le terme d'individu a une signification beaucoup plus large en statistique.

Un individu peut être aussi un ensemble d'objets (les véhicules achetés par les ménages français, les produits fabriqués par une entreprise), un ensemble d'entités géographiques (les départements français, les pays de l'Union européenne), un ensemble non concret (un ensemble de séjours de vacances, un ensemble d'accidents de la route) ou encore un ensemble d'entreprises, un ensemble de logements, etc.

Les individus doivent appartenir à un ensemble bien délimité : il ne doit y avoir aucune ambiguïté sur les unités à observer, leur définition doit être parfaitement claire ainsi que le moment où elles sont observées.

■ *Variable statistique et modalités*

a) Variable statistique (ou caractère)

Le terme variable n'a pas la même signification en mathématique et en statistique :

– en mathématique, une ***variable*** est l'argument d'une fonction : à la variable x, la fonction f associe f(x) ;

– en statistique, une variable est un aspect particulier des individus auxquels on s'intéresse, une caractéristique qui peut varier d'un individu à l'autre. Elle porte aussi le nom de caractère. Il s'agit par exemple de l'âge, le sexe, le nombre d'enfants d'un salarié, la marque ou le prix d'un véhicule, le mode d'hébergement ou la durée d'un séjour de vacances, le lieu d'un accident de la route ou la période à laquelle il s'est produit.

L'ensemble des observations élémentaires d'une variable statistique forme l'ensemble des ***modalités*** de ce caractère.

Pour des êtres humains, si la variable est l'âge, les modalités sont l'ensemble des âges des individus observés ; pour la variable « sexe », il y a deux modalités : homme ou femme ; les modalités de la variable « nombre d'enfants » peuvent être tous les nombres entiers allant de 0 à 5 ou 6, voire plus éventuellement.

Pour un ensemble de véhicules, si la variable est la marque, les modalités sont alors constituées de la liste de toutes les marques observées dans l'ensemble des véhicules étudiés. Si c'est le prix des véhicules qui les caractérisent, les modalités sont l'ensemble de tous les prix constatés.

Pour les accidents de la route, la Sécurité routière les classe, dans son bilan annuel, notamment selon le lieu de l'accident en distinguant quatre modalités (autoroutes, routes nationales, routes départementales et autres routes), mais aussi selon le mois durant lequel l'accident s'est produit (les modalités sont alors les douze mois de l'année).

À chaque individu doit être associée une modalité unique de la variable, c'est-à-dire « au plus une » et « au moins une » :

– « ***Au plus une*** » : il ne doit pas être possible d'associer à un même individu deux modalités ou plus. Cela suppose que les ***modalités*** soient ***incompatibles***, c'est-à-dire parfaitement distinctes les unes des autres. Si un être humain est de sexe masculin, il ne peut être aussi de sexe féminin. Si la variable est l'âge, le même âge ne doit pas figurer dans deux modalités différentes ;

– « ***Au moins une*** » : il ne doit pas être possible qu'un individu ne soit associé à aucune modalité. Il faut pour cela que les ***modalités*** soient ***exhaustives*** : absolument tous les aspects particuliers de

la variable doivent être présents dans la liste des modalités. C'est pourquoi la modalité « autres » ou « indéterminé » est parfois nécessaire. Par exemple, dans l'ensemble des lieux d'accidents de la route répertoriés par la Sécurité routière, la dernière modalité « autres routes » permet de classer, sans avoir à en préciser le détail et risquer d'en oublier, toutes les sortes de routes autres que autoroutes, routes nationales et routes départementales.

D'un point de vue mathématique, une variable peut donc être définie comme une application entre la population (l'ensemble de départ) et l'ensemble des modalités (l'ensemble d'arrivée). Cette application est souvent notée X. Elle associe à chaque individu ω_i de l'ensemble de départ (la population Ω) une unique modalité de l'ensemble d'arrivée notée $X(\omega_i)$ ou m_i.

On distingue deux types de variables : les variables qualitatives et les variables quantitatives.

b) Variable qualitative

Une ***variable*** est ***qualitative*** lorsque l'ensemble des modalités n'est pas un ensemble de nombres.

Dans les exemples ci-dessus, sont des variables qualitatives : le « sexe » d'un être humain, la « marque » d'un véhicule, le « lieu » d'un accident de la route, le « mois » pendant lequel il a eu lieu.

Les différentes modalités m_1, m_2…, m_i…, m_k d'une variable qualitative constituent les rubriques d'une nomenclature. Le nombre de rubriques est au minimum égal à deux ; c'est la cas pour la variable « sexe ». Pour la variable « mois », il y a en nécessairement douze si les observations sont annuelles. Mais une nomenclature peut en compter beaucoup plus. Par exemple, la nomenclature des professions et catégories socioprofessionnelles (dite PCS) de l'Insee est constituée de 8 postes au niveau le plus agrégé (niveau 1), mais 497 au niveau le plus désagrégé (niveau 4).

c) Variable quantitative

Une ***variable*** est ***quantitative*** lorsque l'ensemble de ses modalités est un ensemble de nombres. Elle peut être discrète ou continue.

Une ***variable*** est ***quantitative discrète*** si ses modalités sont des nombres isolés les uns des autres. Il s'agit souvent de nombres entiers, par exemple le « nombre d'enfants » d'un ensemble de ménages ou le « nombre de pièces » d'un ensemble de logements. Les modalités ou valeurs de la variable sont notées x_1, x_2…, x_i…, x_k s'il y a k modalités.

Une ***variable*** est ***quantitative continue*** lorsque ses modalités peuvent prendre toutes les valeurs d'un intervalle réel. Ces valeurs sont regroupées dans des intervalles de valeurs numériques appelés classes.

Dans un tableau, on reconnaît donc une variable quantitative continue au fait que les valeurs de la variable ont été regroupées en classes. Il s'agit de « l'âge » ou du « revenu » d'individus, du

« nombre de salariés » ou du « chiffre d'affaires » des entreprises d'un secteur d'activité, du « prix » de véhicules. En réalité, ces variables, lorsqu'elles ont été mesurées, l'ont souvent été en nombres entiers et elles ne prennent pas toutes les valeurs d'un intervalle réel. C'est évidemment le cas pour le nombre de salariés, mais aussi pour l'âge (même si dans ce cas une plus grande précision serait possible en mois, et jours !). Le revenu, le chiffre d'affaires ou le prix peuvent être évalués au dixième ou au centième d'euros (ou de toute autre unité monétaire), mais, lorsque les montants sont élevés, ils sont généralement arrondis à l'unité.

Le regroupement en classes se justifie par l'existence d'un grand nombre de modalités ; la présentation des données s'en trouve simplifiée.

Les classes sont notées $[e_i ; e_{i+1}[$. L'intervalle est fermé à gauche et ouvert à droite : il inclut toutes les valeurs de la variable supérieures ou égales à la borne inférieure e_i et strictement inférieures à la borne supérieure e_{i+1}.

La différence $e_{i+1} - e_i$ s'appelle l'amplitude de la classe ; elle est notée a_i.

$$a_i = e_{i+1} - e_i$$

La moyenne des extrémités de classe $\frac{e_i + e_{i+1}}{2}$ est appelée centre de la classe et notée x_i (comme les valeurs de la variable lorsqu'elle est discrète).

$$x_i = \frac{e_i + e_{i+1}}{2}$$

2 Effectif et fréquence

■ *Effectif*

Le nombre d'individus présentant la modalité m_i (variable qualitative) ou x_i (variable quantitative discrète) ou une modalité incluse dans $[e_i ; e_{i+1}[$ (variable quantitative continue) s'appelle l'***effectif***. Il est noté n_i. S'il y a k modalités de la variable, les effectifs sont donc notés n_1, n_2…, n_k.

Les modalités de la variable étant à la fois incompatibles et exhaustives, chaque individu est associé à une et une seule d'entre elles. La somme des effectifs, ou effectif total, est donc égale au nombre total d'individus de la population. Ce nombre est noté n.

$$n = n_1 + n_2 + \ldots + n_k = \sum_{i=1}^{k} n_i$$

Attention ! Ne pas confondre population et effectif : une population est un ensemble d'individus ; l'effectif total est le nombre d'individus appartenant à cet ensemble.

■ *Fréquence*

La ***fréquence*** associée à une modalité, ou à un ensemble de modalités regroupées en classes, indique la proportion d'individus présentant cette modalité, ou cet ensemble de modalités, par rapport à l'ensemble des individus. La fréquence associée à la i^e modalité, ou i^e classe, est notée f_i.

Par définition :

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

C'est un nombre compris entre 0 et 1, ou 0 % et 100 %.

La somme des fréquences est égale à 1 ou 100 % :

$$\sum_{i=1}^{k} f_i = f_1 + f_2 + \ldots + f_k = \frac{n_1}{n} + \frac{n_2}{n} + \ldots + \frac{n_k}{n} = \frac{n_1 + n_2 + \ldots + n_k}{n} = \frac{n}{n} = 1.$$

Le plus souvent, les données à analyser concernent une seule population et une seule variable. L'***analyse statistique*** est alors dite ***univariée***. L'ensemble des couples $(m_i ; n_i)$ ou $(m_i ; f_i)$ constituent la distribution statistique de la variable. Les premiers chapitres de cet ouvrage sont consacrés aux outils d'analyse de ces distributions.

Lorsqu'une population est caractérisée à l'aide de deux variables, l'***analyse statistique*** est dite ***bivariée***. Si l'une des variables est le temps, la distribution statistique est appelée série chronologique. Les trois derniers chapitres de cet ouvrage sont consacrés aux séries bivariées.

3 Lecture des données et représentations graphiques des distributions à une variable

Les données peuvent être présentées dans un tableau et/ou un graphique. Il est nécessaire de savoir construire et lire l'un et l'autre.

Si l'on réalise un tableau ou un graphique, il faut veiller à ce que toutes les informations nécessaires pour le comprendre soient présentes. En particulier, il ne faut pas oublier de mettre le titre qui contient une partie de ces informations. La population et la variable doivent être clairement définies, le lieu et la date d'observation précisés. La source du document doit être fournie ; elle permettra,

si besoin est, de trouver des informations complémentaires. Il faut penser à indiquer l'unité dans laquelle sont exprimés les effectifs.

Si l'on utilise un tableau ou un graphique, il faut bien vérifier que le concepteur n'a omis aucune de ces informations.

Nous allons voir, à partir d'exemples, comment lire un tableau statistique et réaliser un graphique lorsque la variable est qualitative puis lorsqu'elle est quantitative.

■ *Variable qualitative*

a) Lecture du tableau de données

La répartition des logements selon la catégorie était la suivante en France, en 2007 (source : Insee, www.insee.fr) :

	Nombre de logements (en millions)	Part des logements (%)
Résidences principales	27,161	84,2
Résidences secondaires	3,235	10,0
Logements vacants	1,864	5,8
Ensemble	32,260	100,0

La population est l'ensemble des logements en France en 2007. Chaque logement est un individu de cette population. Il est caractérisé par la catégorie à laquelle il appartient : résidences principales, résidences secondaires ou logements vacants. La variable étudiée dans ce tableau est donc la catégorie de logement. Elle a trois modalités : résidence principale, résidence secondaire, logement vacant.

L'effectif total, n, est le nombre total de logements : 32,26 millions.

Le nombre de résidences principales, n_1, est 27,161 millions ; leur part dans l'ensemble des logements f_1 est n_1/n, soit 84,2 %. Le nombre de résidences secondaires n_2 est 3,235 millions, le nombre de logements vacants n_3 est 1,864 million ; leur part dans l'ensemble des logements est respectivement 10 % ($f_2 = n_2/n$) et 5,8 % ($f_3 = n_3/n$).

b) Représentation graphique

Un graphique fournit rarement autant d'informations qu'un tableau. Il permet, en revanche, de mieux mettre en évidence certaines informations données par le tableau. Les graphiques les plus utilisés sont

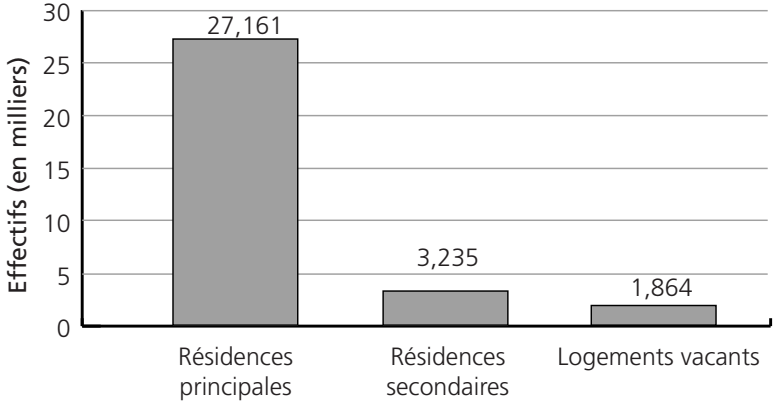
le diagramme à bandes (ou diagramme en tuyaux d'orgues) et le diagramme à secteurs circulaires (ou « camembert »), réalisables par exemple à l'aide d'un tableur (MS Excel, Open Office Calc, etc.).

1) Diagramme à bandes

Dans un diagramme à bandes, une bande verticale (ou horizontale) est associée à chaque modalité. La distance entre chaque bande est constante. La largeur de chacune des bandes est la même, et sa hauteur (ou longueur) est égale à l'effectif ou à la fréquence de la modalité correspondante. Par conséquent, la surface d'une bande est proportionnelle à l'effectif (et donc à la fréquence) de la modalité associée.

Le diagramme à bandes verticales ci-dessous met en évidence les effectifs. Les étiquettes au-dessus des bandes permettent même d'afficher l'effectif exact pour chaque modalité. Ce graphe matérialise aussi, sans la chiffrer, l'importance respective des différentes catégories de logements ; la part prépondérante des résidences principales est évidente ; cependant, sa valeur n'apparaît pas dans le graphique.

Répartition des logements selon la catégorie, en France en 2007 (source : Insee)



Si l'on souhaite mettre l'accent sur la part de chacune des catégories, il est possible de réaliser le même type de graphique mais en portant les fréquences en ordonnée, et éventuellement en étiquettes au-dessus des bandes.

2) Diagramme à secteurs circulaire

Un diagramme à secteurs circulaire est un graphique qui divise un disque en secteurs angulaires dont les angles aux centres sont proportionnels aux effectifs (ou aux fréquences) de chaque modalité.

L'angle au centre α_i, en degrés, associé à la modalité m_i d'effectif n_i, est égal à :

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

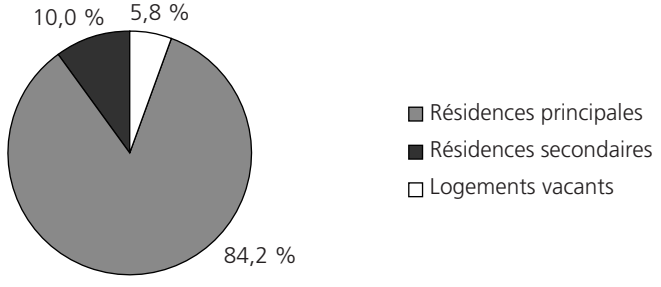
Dans ce graphique, l'aire du disque matérialise l'effectif total (ou la somme des fréquences).

Dans le diagramme circulaire relatif à la répartition des logements selon la catégorie, l'angle au centre associé à chaque modalité est évalué ci-dessous :

Modalité	Effectif (en millions)	Fréquence (%)	Angle au centre (en degrés)
Résidences principales	27,161	84,19	303,1
Résidences secondaires	3,235	10,03	36,1
Logements vacants	1,864	5,78	20,8
Ensemble	32,260	100,00	360,0

Ce calcul n'est évidemment pas nécessaire (sauf si le graphique est réalisé avec un compas et un rapporteur, ce qui est rarement le cas aujourd'hui). Tous les tableurs réalisent les diagrammes circulaires directement à partir des données.

Répartition des logements selon la catégorie, en France en 2007 (source : Insee)



Dans le diagramme ci-dessus, il est possible de remplacer les fréquences par les effectifs.

■ *Variable quantitative discrète*

a) Lecture du tableau de données

Les notes (sur 10) obtenues par 40 étudiants à une épreuve d'économie se répartissent ainsi :

Note : x_i	1	2	3	4	5	6	7	8	9	10
Fréquence : f_i (%)	5,0	5,0	12,5	15,0	22,5	12,5	10,0	10,0	5,0	2,5
Effectif : n_i	2	2	5	6	9	5	4	4	2	1

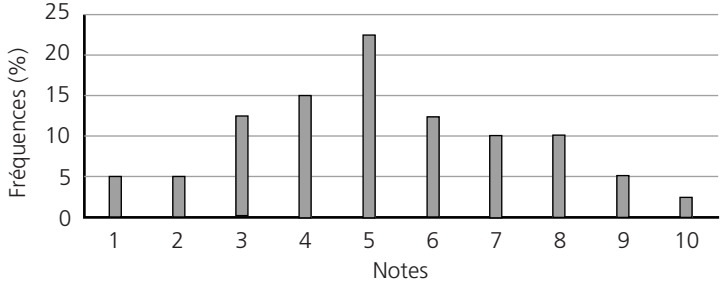
La population est l'ensemble des étudiants qui ont participé à l'épreuve d'économie. L'effectif total n est le nombre total d'étudiants : 40. La variable est la note. Elle a 10 modalités : $x_1 = 1$, $x_2 = 2$…, $x_{10} = 10$.

La part f_1 des étudiants qui ont obtenu 1 est 5 %, leur nombre est $n_1 = f_1.n = 0{,}05.40 = 2$…, la part f_{10} des étudiants qui ont obtenu 10 est 2,5 %, leur nombre est $n_{10} = f_{10}.n = 0{,}025.40 = 1$.

b) Représentation graphique

La représentation graphique d'une distribution à variable quantitative discrète est un diagramme en bâtons. Pour le réaliser sous Excel, on utilise un diagramme à bandes en rétrécissant au minimum la largeur des bandes. Les valeurs de la variable sont portées en abscisse ; en ordonnée, figurent les effectifs ou les fréquences.

Répartition des notes



■ *Variable quantitative continue*

a) Lecture du tableau de données

En France, en 2005, les exploitations agricoles se répartissaient comme suit selon la surface agricole utilisée (SAU) (source : Ministère de l'Agriculture et de la Pêche, Scees, www.agriculture.gouv.fr) :

SAU (en hectares)	[0 ; 5[	[5 ; 20[	[20 ; 50[	[50 ; 100[	[100 ; 200[	200 ou plus
Effectif (en milliers)	132*	105	109	113	70	17
Fréquence (en %)	24,2 %*	19,2 %	20,0 %	20,7 %	12,8 %	3,1 %

* Y compris les exploitations sans superficie agricole utilisée.

La population est l'ensemble des exploitations agricoles, en France, en 2005. La variable est la SAU. Les modalités sont les nombres réels positifs et inférieurs à la borne supérieure de la dernière classe qui n'est pas donnée par le tableau. Attention ! Il est faux de dire qu'il y a six modalités : ce sont les classes qui sont au nombre de six tandis que les modalités sont toutes les valeurs des SAU de l'ensemble des exploitations !

Le nombre total d'exploitations (l'effectif total n) est la somme des effectifs :

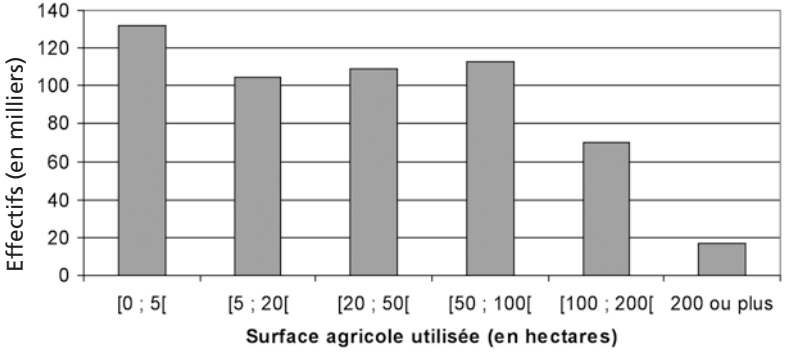
$n = 132 + 105 + \ldots + 17 = 546$ (en milliers).

Il y a 132 000 exploitations de moins de 5 hectares, leur part dans l'ensemble des exploitations est 24,2 % (n_1/n soit 132/546)… ; il y a 17 000 exploitations de 200 hectares ou plus, leur part dans l'ensemble des exploitations est 3,1 % (n_6/n soit 17/546).

b) Représentation graphique

Une représentation graphique simple proposée par les tableurs est un diagramme à bandes. En abscisse sont portées les classes, en ordonnée les effectifs ou les fréquences.

Répartition des exploitations agricoles selon la SAU, en France en 2005 (source : Scees)



Cette représentation est généralement nommée histogramme par les tableurs, mais pas dans la plupart des manuels de statistique.

On nomme généralement histogramme un diagramme formé d'un ensemble de rectangles contigus dont la base est déterminée par les extrémités de classe et dont la surface doit être proportionnelle à l'effectif (ou à la fréquence) de la classe. La surface totale de l'ensemble des rectangles est donc égale à l'effectif total ou à la somme des fréquences, soit 100 %.

Si les classes sont toutes de même amplitude, il suffit pour réaliser l'histogramme de porter en ordonnée les effectifs ou les fréquences.

En revanche, lorsqu'elles ne le sont pas, pour le réaliser sans risque d'erreur, on porte en abscisse les extrémités de classe et en ordonnée les effectifs par unité d'amplitude n_i/a_i, appelés densités d'effectif et notés d_i, ou les fréquences par unité d'amplitude f_i/a_i, nommées densités de fréquence et notées d'_i.

Ainsi la surface de chaque rectangle (hauteur x base) $\frac{n_i}{a_i} a_i$ ou $\frac{f_i}{a_i} a_i$ est bien égale à n_i ou f_i.

Le problème, c'est qu'aucun tableur ne permet de réaliser de manière simple ce graphique. En pratique, c'est donc un histogramme tel que proposé par les tableurs qui est représenté dans le cas où la variable est quantitative continue.

4 Effectifs cumulés et fréquences cumulées

Lorsque la variable est quantitative, discrète ou continue, le tableau des données initiales peut être complété par le calcul des effectifs cumulés croissants et/ou des effectifs cumulés décroissants, des fréquences cumulées croissantes et/ou des fréquences cumulées décroissantes.

■ *Variable quantitative discrète*

a) Définitions

Lorsque la variable est discrète, un effectif ou une fréquence cumulé(e) est associé(e) à une modalité x_i de la variable. Les valeurs de la variable étant toujours classées par ordre croissant, x_i est supérieure à x_{i-1} et inférieure à x_{i+1}.

L'*effectif cumulé croissant associé à la modalité* x_i, noté N_i, est égal au nombre d'individus dont la valeur de la variable est strictement inférieure à x_i, donc au plus égale à x_{i-1}.

La *fréquence cumulée croissante associée à la modalité* x_i, notée F_i, est égale à la proportion d'individus dont la valeur de la variable est strictement inférieure à x_i. D'où :

$$F_i = \frac{N_i}{n}$$

Dans le tableau de calculs, N_i et F_i sont positionnés sur la même ligne que x_i parce que leur valeur s'interprète par rapport à x_i.

Modalité x_i	Effectif n_i	Fréquence f_i	Effectif cumulé croissant N_i	Fréquence cumulée croissante F_i
x_1	n_1	f_1	$N_1 = 0$	$F_1 = 0$
x_2	n_2	f_2	$N_2 = n_1$	$F_2 = f_1$
x_3	n_3	f_3	$N_3 = n_1 + n_2$	$F_3 = f_1 + f_2$
. . .	. . .	. . .	. . .	. . .
x_k	n_k	f_k	$N_k = n_1 + n_2 + \ldots + n_{k-1}$	$F_k = f_1 + f_2 + \ldots + f_{k-1}$
			$N_{k+1} = n$	$F_{k+1} = 1$
Ensemble	n	1		

N_{k+1} est égal à l'effectif total, de même F_{k+1} est égal à la somme des fréquences 1 ou 100 % : c'est respectivement le nombre et la part des individus dont la valeur de la variable est inférieure ou égale à x_k.

Les effectifs cumulés décroissants et les fréquences cumulées décroissantes se déduisent respectivement des effectifs cumulés croissants et des fréquences cumulées croissantes. L'***effectif cumulé décroissant*** $(n - N_i)$ est le nombre d'individus dont la valeur de la variable est supérieure ou égale à x_i, donc au moins égale à x_i.

La ***fréquence cumulée décroissante*** $(1 - F_i)$ est la proportion d'individus dont la valeur de la variable est supérieure ou égale à x_i.

Ce sont les effectifs et les fréquences cumulés croissants qui sont le plus souvent calculés. Ils sont généralement nommés simplement « effectifs cumulés » et « fréquences cumulées ».

b) Application

La répartition des notes de statistique obtenues à une épreuve par 40 étudiants est donnée par les trois premières colonnes du tableau ci-dessous. Des fréquences sont déduites les fréquences cumulées croissantes, des effectifs sont déduits les effectifs cumulés croissants. Dans les deux dernières colonnes, figurent les fréquences cumulées décroissantes et les effectifs cumulés décroissants.

Note x_i	Fréq. f_i	Effectif n_i	Fréquence cumulée croissante F_i	Effectif cumulé croissant Ni	Fréquence cumulée décroissante $(100\ \% - F_i)$	Effectif cumulé décroissant $n - Ni$
1	5 %	2	0 %	0	100 %	40
2	5 %	2	5 %	2	95 %	38
3	12,5 %	5	10 %	4	90 %	36
4	15 %	6	22,5 %	9	77,5 %	31
5	22,5 %	9	37,5 %	15	62,5 %	25
6	12,5 %	5	60 %	24	40 %	16
7	10 %	4	72,5 %	29	27,5 %	11
8	10 %	4	82,5 %	33	17,5 %	7
9	5 %	2	92,5 %	37	7,5 %	3
10	2,5 %	1	97,5 %	39	2,5 %	1
			100 %	40	0 %	0

Lecture des lignes du tableau :

– *ligne 1* : aucun étudiant n'a une note strictement inférieure à 1 (aucun étudiant n'a eu 0), donc $F_1 = 0$ % et $N_1 = 0$. D'où 100 % des étudiants (les 40 étudiants) ont une note supérieure ou égale à 1 ;

– *ligne 2* : 5 % des étudiants, soit deux étudiants, ont eu strictement moins de 2, c'est-à-dire une note inférieure ou égale à 1, donc $F_2 = 5$ % et $N_2 = 2$. D'où 95 % des étudiants ont eu 2 ou plus, soit 38 étudiants ;

– *ligne 3* : 10 % des étudiants, soit quatre étudiants, ont eu strictement moins de 3, c'est-à-dire une note inférieure ou égale à 2, donc $F_3 = 10$ % et $N_3 = 4$. D'où 90 % des étudiants ont eu 3 ou plus, soit 36 étudiants ;

…

– *ligne 10* : 97,5 % des étudiants, soit 39 étudiants, ont eu strictement moins de 10, c'est-à-dire une note inférieure ou égale à 9, donc $F_{10} = 97{,}5$ % et $N_{10} = 39$. D'où 2,5 % des étudiants ont eu 10, soit 1 étudiant.

– *dernière ligne* :100 % des étudiants, soit 40 étudiants, ont eu 10 ou moins et évidemment aucun n'a eu plus de 10.

■ *Variable quantitative continue*

a) Définitions

Lorsque la variable est continue, un effectif ou une fréquence cumulé(e) est associé(e) à une extrémité de classe.

L'***effectif cumulé croissant associé à l'extrémité de classe*** e_i, noté N_i, est égal au nombre d'individus dont la valeur de la variable est strictement inférieure à e_i.

La ***fréquence cumulée croissante associée à l'extrémité de classe*** e_i, notée F_i, est égale à la proportion d'individus dont la valeur de la variable est strictement inférieure à e_i. Donc, par définition : $F_i = \frac{N_i}{n}$.

Pour être sûr de donner la signification exacte d'un effectif cumulé ou d'une fréquence cumulée, il est préférable de construire le tableau de calculs comme suit :

– les extrémités des classes figurent dans la 1re colonne les unes en dessous des autres et une ligne sur deux ;

– dans la 2e colonne, on porte les centres des classes, une ligne sur deux également, en les intercalant entre les extrémités des classes correspondantes : x_1 est entre e_1 et e_2, x_2 est entre e_2 et e_3, etc. ;

– sur la même ligne que x_i sont portés dans les colonnes suivantes l'effectif n_i et la fréquence f_i car leur valeur s'interprète en référence à x_i : n_i individus, soit une part f_i de la population, ont une valeur moyenne de la variable égale à x_i ;

– sur la même ligne que e_i (et non pas x_i) sont ensuite portés les effectifs cumulés N_i et les fréquences cumulées F_i parce que leur valeur est liée à e_i : N_i individus, soit une part F_i de la population, ont une valeur de la variable strictement inférieure à e_i.

Extrémité de classe e_i	Centre de classe x_i	Effectif n_i	Fréquence f_i	Effectif cumulé N_i	Fréquence cumulée F_i
e_1				$N_1 = 0$	$F_1 = 0$
	x_1	n_1	f_1		
e_2				$N_2 = n_1$	$F_2 = f_1$
	x_2	n_2	f_2		
e_3				$N_3 = n_1 + n_2$	$F_3 = f_1 + f_2$
. . .	. . .	. . .	. . .	. . .	. . .
e_k				$N_k = n_1 + n_2 + \ldots + n_{k-1}$	$F_k = f_1 + f_2 + \ldots + f_{k-1}$
	x_k	n_k	f_k		
e_{k+1}				$N_{k+1} = n$	$F_{k+1} = 1$
Ensemble		n	1		

Les effectifs cumulés décroissants et les fréquences cumulées décroissantes se déduisent respectivement des effectifs cumulés croissants et des fréquences cumulées croissantes.

L'***effectif cumulé décroissant ($n - N_i$)*** est le nombre d'individus dont la valeur de la variable est supérieure ou égale à e_i.

La ***fréquence cumulée décroissante (1 – F_i)*** est la proportion d'individus dont la valeur de la variable est supérieure ou égale à e_i.

Comme lorsque la variable est quantitative discrète, ce sont les effectifs et fréquences cumulés croissants qui sont le plus souvent calculés et nommés simplement « effectifs cumulés » et « fréquences cumulées ».

b) Application

La répartition des exploitations agricoles selon la SAU, en France, en 2005, est donnée par les deux premières colonnes du tableau ci-dessous. Les effectifs permettent de calculer les effectifs cumulés, croissants et décroissants (3e et 4e colonne), puis les fréquences (5e colonne), dont sont déduites les fréquences cumulées croissantes et décroissantes (6e et 7e colonne).

SAU (ha) e_i	Effectif (milliers) n_i	Effectif cumulé croissant (milliers) N_i	Effectif cumulé décroissant (milliers) $N - N_i$	Fréquence f_i	Fréquence cumulée croissante F_i	Fréquence cumulée décroissante $100\ \% - F_i$
$e_1 = 0$		$N_{1 =} 0$	546		$F_1 = 0{,}0\ \%$	100,0 %
	132			24,2 %		
$e_2 = 5$		$N_2 = 132$	414		$F_2 = 24{,}2\ \%$	75,8 %
	105			19,2 %		
$e_3 = 20$		$N_3 = 237$	309		$F_3 = 43{,}4\ \%$	56,6 %
	109			20,0 %		
$e_4 = 50$		$N_4 = 346$	200		$F_4 = 63{,}4\ \%$	36,6 %
	113			20,7 %		
$e_5 = 100$		$N_5 = 459$	87		$F_5 = 84{,}1\ \%$	15,9 %
	70			12,8 %		
$e_6 = 200$		$N_6 = 529$	17		$F_6 = 96{,}9\ \%$	3,1 %
	17			3,1 %		
$e_7 = ?$		$N_7 = 546$	0		$F_7 = 100{,}0\ \%$	0,0 %

Lecture des lignes du tableau :

– *ligne 1* : aucune exploitation n'a une SAU inférieure à 0, donc les 546 000 exploitations, soit 100 % des exploitations, ont une surface supérieure ou égal à 0 ;

– *ligne 3* : 132 000 exploitations ont une SAU inférieure à 5 hectares, soit 24,2 % des exploitations ; donc 414 000 exploitations, soit 75,8 % des exploitations, ont une SAU supérieure ou égale à 5 hectares ;

…

– *dernière ligne* : les 546 000 exploitations ont toutes une SAU inférieure à e_7 (dont la valeur n'est pas fournie par le tableau initial), soit 100 % des exploitations, et aucune exploitation n'a une SAU supérieure ou égale à e_7.

Caractéristiques de position : le mode et les quantiles

CHAPITRE 2

Lorsque la variable est quantitative, pour poursuivre l'analyse des données, on les « résume » par des valeurs numériques qui portent le nom de caractéristiques. Dans ce chapitre et le suivant, nous étudions des caractéristiques de position : le mode et les quantiles (chapitre 2) puis les moyennes (chapitre 3). Le mode et la médiane, qui est l'un des quantiles les plus usités, sont des caractéristiques de tendance centrale, c'est-à-dire des valeurs situées « au centre » de la distribution.

1 Mode et classe modale

Lorsque la variable est quantitative discrète, on détermine le mode. Lorsqu'elle est quantitative continue, on détermine la classe modale.

■ *Mode*

a) Définition

La variable est quantitative discrète. Les valeurs de la variable sont donc des nombres bien isolés les uns des autres.

Le mode est la valeur de la variable la plus fréquemment observée. C'est donc la valeur pour laquelle l'effectif est le plus élevé ou la fréquence la plus élevée. C'est cette signification qui lui confère le statut de caractéristique de valeur centrale : c'est la valeur qui domine les autres. Sur le diagramme en bâtons, c'est la valeur qui correspond au bâton le plus haut.

Si une série présente, pour plusieurs valeurs de la variable, un effectif ou une fréquence maximum identique, on dit qu'elle est plurimodale. Les différents modes sont les différentes valeurs de la variable correspondant à cet effectif ou à cette fréquence maximale.

b) Application

Une entreprise de services a relevé au cours des cent derniers jours ouvrés le nombre de plaintes de clients par jour. Le tableau suivant synthétise les résultats :

Nombre de plaintes	0	1	2	3	4	5	6	7
Nombre de jours	11	22	19	16	11	9	7	5

La population étudiée est l'ensemble des cent derniers jours ouvrés. Chaque jour est caractérisé par le nombre de plaintes de clients : la variable (ou caractère) est donc le nombre de plaintes. Les valeurs de la variable étant des nombres bien distincts les uns des autres, la variable est quantitative discrète.

L'effectif le plus élevé est 22, la valeur correspondante de la variable est 1 : le mode est égal à 1.

■ *Classe modale*

a) Définition

La variable est quantitative continue. Les valeurs de la variable sont donc regroupées en classes notées $[e_i ; e_{i+1}[$; la différence entre les extrémités de classe est l'amplitude a_i.

Si les classes sont toutes de même amplitude, la ***classe modale*** est la classe d'effectif le plus élevé ou de fréquence la plus élevée.

Si les classes ne sont pas toutes de même amplitude, pour déterminer la classe modale, c'est-à-dire la classe la plus « fréquemment » observée, il faut comparer les effectifs (ou les fréquences) à amplitude constante. L'amplitude choisie est généralement l'amplitude d'une unité. On calcule alors, pour chaque classe, soit la densité d'effectif, soit la densité de fréquence.

La densité d'effectif, notée d_i, est l'effectif pour une amplitude d'une unité :

$$d_i = \frac{n_i}{a_i}$$

La densité de fréquence, notée d'_i, est la fréquence pour une amplitude d'une unité :

$$d'_i = \frac{f_i}{a_i}$$

La classe de densité d'effectif (ou de fréquence) la plus élevée est la classe modale. C'est celle qui, pour chaque unité d'amplitude, présente le plus grand nombre d'individus.

Dans le cas particulier où une classe est caractérisée à la fois par l'effectif le plus élevé et l'amplitude la plus faible, elle a nécessairement la densité la plus forte. Dans ce cas, le calcul des densités n'est pas nécessaire pour connaître la classe modale.

b) Application

Le tableau suivant indique la répartition des entreprises industrielles (hors IAA, bâtiment, génie civile et agricole) de 20 salariés ou plus selon le nombre de salariés, en France, en 2005 (source : Ministère de l'Économie, de l'Industrie et de l'Emploi, Services des études et des statistiques industrielles, *Enquête annuelle d'entreprises*, www.industrie.gouv.fr/sessi/).

Nombre de salariés	[20 ; 50[	[50 ; 100[	[100 ; 250[	[250 ; 500[	500 ou plus	Ensemble
Nombre d'entreprises	10 409	4 155	2 872	1 031	870	19 337

Les classes n'étant pas toutes de même amplitude, la classe modale est la classe de densité d'effectif la plus grande. Pour la déterminer, il faut comparer les densités d'effectifs des différentes classes.

On calcule d'abord l'amplitude de chaque classe. Le calcul exact de l'amplitude de la dernière classe ne peut pas être effectué, mais on peut, en revanche, affirmer qu'elle n'est pas inférieure à 250 (l'amplitude la plus élevée des autres classes), car il existe en France des entreprises industrielles de plus de 750 salariés. La densité de la dernière classe ne peut donc pas être supérieure à 870/250, soit 3,48.

Nombre de salariés	[20 ; 50[	[50 ; 100[	[100 ; 250[	[250 ; 500[	500 ou plus
Effectif	10 409	4 155	2 872	1 031	870
Amplitude	30	50	150	250	au moins 250
Densité d'effectif	347,0	83,1	19,1	4,1	au plus 3,48

La classe modale est la classe [20 ; 50[car elle a la densité la plus élevée : 347. Quelle est la signification concrète de ce nombre ? Il indique que si, en 2005, les entreprises industrielles étaient réparties de manière uniforme (« régulière ») à l'intérieur de la classe [20 ; 50[, 347 entreprises avaient 20 salariés, 347 en avaient 21…, 347 en avaient 48 et enfin, 347 en avaient 49.

On peut remarquer que dans cet exemple, le calcul des densités n'était pas nécessaire pour déterminer la classe modale. En effet, on constate ici que la 1re classe a l'effectif n_1 le plus élevé mais aussi l'amplitude a_1 la plus faible ; la division de cet effectif par cette amplitude n_1/a_1 donne donc nécessairement la valeur de densité la plus forte : la 1re classe est la classe modale.

2 Quantiles : définitions

Soit une variable quantitative X, discrète ou continue, dont les modalités sont classées par valeurs croissantes.

■ *Quantile d'ordre α %*

Le quantile d'ordre α %, noté q_α, est la valeur de la variable telle que α % des valeurs observées sont inférieures à q_α. Donc, par définition, la fréquence cumulée croissante associée au quantile d'ordre α % est α %, soit :

$$F(q_\alpha) = \alpha\ \%$$

Le quantile d'ordre α % peut aussi être défini comme la valeur de la variable qui partage la population en deux sous-populations, telles que dans la première on a (α %)n individus et dans la seconde (100 – α %)n individus. L'effectif cumulé croissant associé au quantile d'ordre α % est donc égal à (α %)(n), soit :

$$N(q_\alpha) = (\alpha\ \%)(n)$$

■ *Principaux quantiles*

Les quantiles les plus couramment utilisés sont les quartiles, les déciles et les centiles :

– les ***quartiles*** sont les ***trois valeurs de la variable notées q_{25}, q_{50}*** et ***q_{75}*** qui partagent les observations en quatre groupes d'effectifs égaux : 25 % (resp. 50 % et 75 %) des individus de la population ont une valeur de la variable inférieure à q_{25} (resp. q_{50} et q_{75}) ;

– les ***déciles*** sont les ***neuf valeurs de la variable notées q_{10}, q_{20}…, q_{90}*** qui partagent les observations en dix groupes d'effectifs égaux : 10 % des individus de la population ont une valeur de la variable inférieure à q_{10}…, 90 % des individus de la population ont une valeur de la variable inférieure à q_{90} ;

– les ***centiles*** sont les ***99 valeurs de la variable notées q_1, q_2…, q_{99}*** qui partagent les observations en 100 groupes d'effectifs égaux : 1 % des individus de la population ont une valeur de la variable inférieure à q_1…, 99 % des individus de la population ont une valeur de la variable inférieure à q_{99}.

La notation q_α a le mérite d'être identique pour tous les quantiles et d'indiquer, par le α présent en indice, la part de la population qui présente une valeur de la variable inférieure à ce quantile. Il existe cependant d'autres notations fréquemment adoptées : Q_1, Q_2, Q_3 pour les quartiles, D_1, D_2…, D_9, pour les déciles et C_1, C_2…, C_{99}. pour les centiles. Elles ont l'avantage de mettre en évidence par le Q, le D ou le C le type de quantile évalué, mais l'inconvénient de noter de manière différente des

quantiles qui sont identiques : le quantile d'ordre 10 % est D_1 mais aussi C_{10}, le quantile d'ordre 25 % est Q_1 mais aussi C_{25}, le quantile d'ordre 50 % est Q_2, mais aussi D_5 et C_{50}. Dans cet ouvrage, nous utilisons principalement la notation q_α.

Parmi tous les quantiles, il en est un beaucoup plus usité que les autres, c'est *le quantile* q_{50} appelé médiane et aussi noté Mé (ou Q_2, ou D_5, ou C_{50}). La médiane est la valeur qui partage les observations en deux groupes de même effectif. C'est une caractéristique de valeur centrale au sens usuel du terme : sa valeur est au milieu de toutes les autres. Il y a autant d'individus de la population qui présentent une valeur de la variable inférieure (ou éventuellement égale) à la médiane que d'individus qui présentent une valeur de la variable supérieure (ou éventuellement égale) à la médiane.

3 Quantiles : évaluation

■ *Cas d'une variable quantitative discrète*

La méthode d'évaluation diffère selon que les données sont ou non groupées dans un tableau statistique.

a) Cas où les données ne sont pas groupées dans un tableau statistique

1) Méthode

Les données sont classées par ordre croissant ou décroissant. Comme elles sont généralement peu nombreuses, c'est principalement la médiane qui est évaluée.

Si le nombre n des données est impair, la médiane est la valeur de la variable qui occupe la position centrale, c'est-à-dire le rang (n + 1)/2. Si leur nombre est pair, la médiane est égale à la demi-somme des deux valeurs centrales qui occupent respectivement le rang n/2 et le rang (n/2) + 1.

Pour les autres quantiles, il n'est pas toujours possible de leur donner une valeur exactement conforme à la définition.

2) Applications

Considérons le taux de chômage des jeunes (en %) dans les 27 pays de l'UE en 2006 (source : Insee, *Tableaux de l'économie française*, édition 2007, www.insee.fr).

Pays–Bas	6,6	Lettonie	12,2	République tchèque	17,5	Roumanie	21,4
Danemark	7,7	Slovénie	13,9	Espagne	17,9	Italie	21,6
Irlande	8,6	Royaume–Uni	14,1	Finlande	18,7	France	23,1
Autriche	9,2	Allemagne	14,2	Hongrie	19,1	Grèce	25,2
Lituanie	9,8	Luxembourg	16,2	Bulgarie	19,5	Slovaquie	26,6
Chypre	10,5	Portugal	16,3	Belgique	20,5	Pologne	29,8
Estonie	12,0	Malte	16,4	Suède	20,7		

Le taux de chômage médian est celui qui occupe le rang (27 + 1)/2, soit le 14e rang. Il est égal à 16,4 %. 13 pays ont un taux de chômage des jeunes inférieur à celui de Malte et 13 pays un taux de chômage des jeunes supérieur.

Observons maintenant les 20 plus fortes capitalisations boursières en actions françaises (en milliards d'euros) d'Euronext Paris au 31 mai 2007 (source : Insee, *Tableaux de l'économie française*, édition 2007, www.insee.fr). Quelle est la capitalisation boursière médiane ?

Total	134,0	Société Générale	66,8	Crédit Agricole	50,5	Danone	30,4
EDF	125,6	France Télécom	59,5	LVMH	43,0	Renault	30,3
Sanofi–Aventis	97,4	Arcelor Mittal	58,9	Carrefour	38,2	Saint–Gobain	30,1
BNP Paribas	84,1	L'Oréal	55,3	Vivendi	37,4	Vinci	28,1
Axa	68,1	Suez	54,7	Gaz de France	36,7	Dexia	27,7

Le nombre de données 20 est pair. La médiane est égale à la demi-somme des deux valeurs centrales qui occupent respectivement le rang 10 et la rang 11. La capitalisation boursière médiane est (54,7 + 50,5)/2 = 52,6 (en milliards d'euros).

b) Cas où les données sont groupées dans un tableau statistique

1) Méthode

S'il existe une valeur du caractère x_i telle que $N(x_i) = (\alpha\ \%)(n)$ ou $F(x_i) = \alpha\ \%$, alors cette valeur x_i est le quantile d'ordre $\alpha\ \%$: $\alpha\ \%$ des valeurs observées sont inférieures à x_i et $(100 - \alpha)\ \%$ des valeurs observées lui sont supérieures ou égales.

Sinon, on nomme quantile d'ordre α % la valeur x_i de la variable telle que α % des valeurs observées lui sont inférieures ou égales et (100 – α) % lui sont supérieures ou égales. C'est la valeur associée à l'effectif cumulé directement inférieur à (α %)(n) et par conséquent aussi à la fréquence cumulée directement inférieure à α %.

2) Application

Illustrons en reprenant la répartition des notes des 40 étudiants à une épreuve d'économie du chapitre 1.

Note x_i	Fréquence f_i (%)	Effectif n_i	Fréquence cumulée croissante F_i (%)	Effectif cumulé croissant Ni
1	5	2	0	0
2	5	2	5	2
3	12,5	5	10	4
4	15	6	22,5	9
5	22,5	9	37,5	15
6	12,5	5	60	24
7	10	4	72,5	29
8	10	4	82,5	33
9	5	2	92,5	37
10	2,5	1	97,5	39
			100	40

On peut lire directement dans le tableau le 5e centile, le 1er et le 6e décile : 5 % des notes sont inférieures à 2, donc $q_5 = 2$; 10 % sont inférieures à 3, donc $q_{10} = 3$; et 60 % sont inférieures à 6, d'où $q_{60} = 6$.

Pour déterminer les autres quantiles q_α, il faut chercher dans le tableau la fréquence cumulée directement inférieure à α %.

Si l'on cherche la médiane q_{50}, on constate dans le tableau que 37,5 % des étudiants (les 15 étudiants ayant les notes les plus faibles) ont une note inférieure à 5 et 60 % (les 24 ayant les notes les plus

faibles) ont une note inférieure à 6, donc inférieure ou égale à 5. Il n'existe donc pas de note telle que exactement 50 % des étudiants aient une note inférieure à celle-ci et 50 % une note supérieure ou égale. En revanche, si l'on écrit toutes les notes les unes à la suite des autres, par ordre croissant, on peut affirmer que la 20e note est 5 et la 21e est 5 également, donc 50 % des étudiants ont une note inférieure ou égale à 5 et 50 % une note supérieure ou égale à 5. La note médiane est 5.

Néanmoins, on peut, dans ce cas, se poser la question de la pertinence de l'information ainsi fournie et donc de la pertinence de l'utilisation du terme médiane. Certes, la valeur 5 occupe bien la position centrale, mais elle en occupe aussi d'autres puisque, sur l'ensemble des 40 notes, la note 5 occupe les rangs 16 à 24. Dans un cas comme celui-ci, plutôt que d'affirmer que la médiane est 5, il est préférable, pour transmettre une information juste, de dire simplement que 37,5 % des étudiants ont une note inférieure à 5 et 60 % une note inférieure à 6 car cette information là est beaucoup plus précise.

En appliquant le même raisonnement, on trouve $q_{25} = 4$, car 25 % des étudiants ont une note inférieure ou égale à 4 et 75 % ont une note supérieure ou égale à 4, mais il est plus précis de dire que 22,5 % des notes sont inférieures à 4 et 37,5 % sont inférieures à 5, donc inférieures ou égales à 4.

■ *Cas d'une variable quantitative continue*

Lorsque la variable est quantitative continue, un quantile peut être déterminé graphiquement ou par le calcul.

a) Détermination graphique

1) Méthode

Pour déterminer graphiquement un quantile, on peut se servir soit de la courbe cumulative des fréquences, soit de la courbe cumulative des effectifs :

– la ***courbe cumulative des fréquences*** (ou courbe des fréquences cumulées, ou encore polygone des fréquences cumulées) est la ligne polygonale qui joint, dans un système d'axes orthogonaux, les points d'abscisse e_i et d'ordonnée F_i ;

– la ***courbe cumulative des effectifs*** (ou courbe des effectifs cumulés, ou encore polygone des effectifs cumulés) est la ligne polygonale qui joint, dans un système d'axes orthogonaux, les points d'abscisse e_i et d'ordonnée N_i.

Pour chacune de ces courbes, en joignant les points par des segments de droite, on « fait comme si » les observations étaient réparties de manière uniforme, c'est-à-dire régulière, à l'intérieure de chaque classe, ce qui dans la réalité n'est pas nécessairement vrai.

On cherche graphiquement la solution de l'équation $F(q_\alpha) = \alpha$ % sur la courbe cumulative des fréquences ou la solution de l'équation $N(q_\alpha) = (\alpha\ \%)(n)$ sur la courbe cumulative des effectifs.

En pratique, on utilise plutôt la courbe cumulative des fréquences pour deux raisons. D'une part, les fréquences sont des nombres compris entre 0 et 1 (ou 0 % et 100 %) donc des valeurs plus faciles à manipuler que les effectifs qui peuvent atteindre des valeurs importantes. D'autre part, la fréquence cumulée α % est repérable immédiatement sur l'axe des ordonnées, tandis que l'effectif cumulé $(\alpha\ \%)(n)$ doit d'abord être calculé avant d'être positionné sur l'axe des ordonnées.

2) Application

Le tableau suivant indique la répartition des films long métrage d'initiative française, en 2006, en fonction du montant du devis, en millions d'euros (source : Centre national de la cinématographie, Bilan 2006, mai 2007, www.cnc.fr) :

Montant du devis (millions d'euros)	Moins de 1	[1 ; 4[	[4 ; 7[	Plus de 7	Ensemble
Nombre de films	28	72	19	45	164

Nous allons déterminer graphiquement la médiane et le 7e décile à partir de la courbe des fréquences cumulées.

Pour tracer la courbe cumulative des fréquences, il faut connaître toutes les extrémités de classe. Or la borne inférieure de la première classe et la borne supérieure de la dernière ne sont pas données. Comment les choisir ? Pour la première, le plus simple est de choisir la valeur 0, même si évidemment la vraie borne est un nombre positif. Pour la dernière classe, en l'absence de toute information supplémentaire sur les valeurs de la variable, on choisit souvent de lui donner une amplitude double de la précédente. Ici, la classe [4 ; 7[a pour amplitude 3. Une amplitude double pour la dernière classe implique une borne supérieure égale à 7 + 6, soit 13.

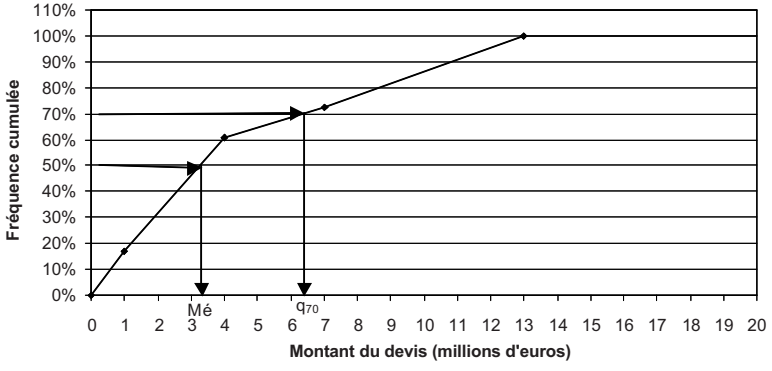
Dans le tableau suivant figurent les données, les fréquences et les fréquences cumulées.

e_i	n_i	f_i	F_i
0			0,0 %
	28	17,1 %	
1			17,1 %
	72	43,9 %	
4			61,0 %
	19	11,6 %	
7			72,6 %
	45	27,4 %	
13			100,0 %
Ensemble	164	100,0 %	

Les points (e_i ; F_i) sont (0 ; 0 %), (1 ; 17,1 %)…, (13, 100 %). Ces points sont joints par des segments de droite. Au-delà de la valeur 13 de la variable, la fréquence cumulée est constante et toujours égale à 100 %.

Pour déterminer graphiquement la médiane, on part de la fréquence cumulée 50 % sur l'axe des ordonnées, on projette ce point parallèlement à l'axe des abscisses jusqu'à la courbe. Le point atteint sur la courbe est projeté à son tour sur l'axe des abscisses parallèlement à l'axe des ordonnées. La valeur atteinte sur l'axe des abscisses est la médiane. Les flèches sur le graphique matérialisent ce cheminement.

La médiane apparaît égale à environ 3,3 millions d'euros : en France, en 2006, le devis de la moitié des films long métrage d'initiative française était inférieur à 3,3 millions d'euros.



En procédant de la même manière, à partir de la valeur 70 % sur l'axe des ordonnées, on trouve q_{70} égal à environ 6,3 millions d'euros. Le devis des 70 % des films long métrage les moins chers était inférieur à 6,3 millions d'euros.

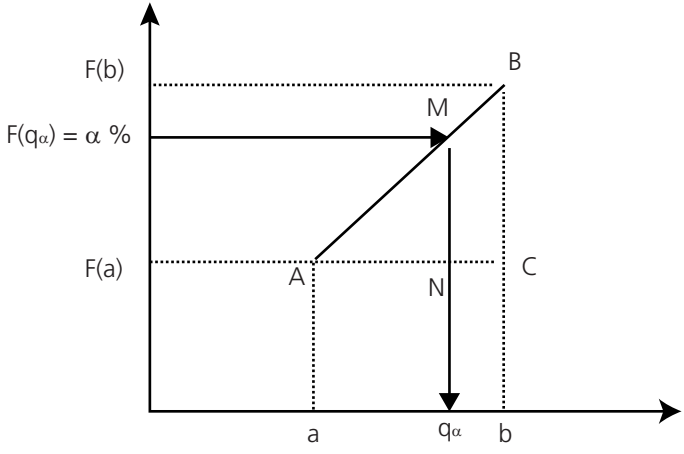
b) Détermination par le calcul

1) Méthode

S'il existe une extrémité de classe e_i telle que $N(e_i) = (\alpha\ \%)(n)$ ou $F(x_i) = \alpha\ \%$, alors cette valeur e_i est le quantile d'ordre α % : α % des valeurs observées sont inférieures à e_i et $(100 - \alpha)$ % des valeurs observées lui sont supérieures ou égales.

Sinon, on détermine le quantile d'ordre α % en effectuant une interpolation linéaire. Cette méthode repose sur l'hypothèse d'une répartition uniforme des valeurs observées à l'intérieur de la classe qui contient le quantile d'ordre α %. Si cette hypothèse est vérifiée, à l'intérieur de cette classe, la courbe représentative des fréquences cumulées est une droite.

Soient a et b les extrémités de classe telles que $F(a) < \alpha\ \% < F(b)$. La représentation graphique de la courbe des fréquences cumulées entre a et b est la suivante :



Le graphique permet de mettre en évidence deux triangles semblables : ANM et ACB. Ils sont semblables car les angles correspondants sont égaux et les côtés correspondants de longueurs proportionnelles.

En particulier : $\frac{\overline{AN}}{\overline{AC}} = \frac{\overline{NM}}{\overline{CB}}$. Or $\frac{\overline{AN}}{\overline{AC}} = \frac{q_\alpha - a}{b - a}$ et $\frac{\overline{NM}}{\overline{CB}} = \frac{F(q_\alpha) - F(a)}{F(b) - F(a)}$.

On a donc l'égalité : $\frac{q_\alpha - a}{b - a} = \frac{F(q_\alpha) - F(a)}{F(b) - F(a)}$. Elle implique :

$$q_\alpha = a + [(b - a)\frac{F(q_\alpha) - F(a)}{F(b) - F(a)}].$$

Le raisonnement est le même à partir de la courbe représentative des effectifs cumulés. Sur le graphe, en ordonnée, à la place de F(a), F(q_α) et F(b), figurent respectivement N(a), N(q_α) et N(b). Par conséquent $\frac{q_\alpha - a}{b - a} = \frac{N(q_\alpha) - N(a)}{N(b) - N(a)}$, d'où :

$$q_\alpha = a + [(b - a)\frac{N(q_\alpha) - N(a)}{N(b) - N(a)}]$$

Il ne faut pas oublier que le calcul ainsi effectué repose sur l'hypothèse d'une répartition uniforme des observations à l'intérieur de la classe qui contient q_α, ce qui n'est probablement pas tout à fait vérifié. Le résultat de ce calcul est donc un évaluation qui ne peut prétendre à l'exactitude.

2) Application

Nous allons calculer la médiane et le 7e décile des devis des films en effectuant une interpolation linéaire à partir successivement des effectifs cumulés et des fréquences cumulées, calculés ci-dessous.

e_i	n_i	N_i	f_i	F_i
0		0		0,0 %
	28		17,1 %	
1		28		17,1 %
	72		43,9 %	
4		100		61,0 %
	19		11,6 %	
7		119		72,6 %
	45		27,4 %	
13		164		100,0 %
Ensemble	164		100,0 %	

Calcul de Mé, la médiane :

Par définition, N(Mé) = n/2. L'effectif total est 164, d'où n/2 = 82.

On constate, dans la colonne des effectifs cumulés, que 28 < 82 < 100, soit N(1) < N(Mé) < N(4). La fonction N étant une fonction strictement croissante, on en déduit 1 < Mé < 4 et on calcule la médiane par interpolation linéaire :

$$\text{Mé} = 1 + [(4-1)\frac{82-28}{100-28}] = 1 + (3).(0,75) = 3,25 .$$

On obtient évidemment le même résultat en utilisant les fréquences cumulées.

Dans la colonne des fréquences cumulées 17,1 % < 50 % < 61 %, donc F(1) < 50 % < F(4).

La fonction F étant strictement croissante, on en déduit que la médiane est comprise entre 1 et 4 et on la calcule par interpolation linéaire :

$$\text{Mé} = 1 + [(4-1)\frac{50\,\% - 17,1\,\%}{61\,\% - 17,1\,\%}] = 1 + (3).(0,75) = 3,25 .$$

On retrouve le résultat de l'évaluation graphique : la médiane est égale à environ 3,3 millions d'euros.

Cependant, le bilan 2006 du CNC nous indique que le devis médian s'est élevé à 2,82 millions. Cette dernière valeur a été calculée à partir de la connaissance des montants précis des devis de tous les films. Elle est donc exacte. Comment s'explique alors la différence avec notre résultat ? Simplement par la fait qu'à l'intérieur de la classe [1 ; 4[les observations ne sont pas réparties uniformément. Si elles l'étaient, les deux résultats seraient absolument identiques. L'hypothèse sur laquelle repose notre calcul n'est pas exacte, mais nous n'avions pas d'autre solution pour réaliser cette évaluation.

Calcul de q_{70}, le 7e décile : $N(q_{70}) = 70\,\%.n = (0,7)(164) = 114,8$.

Dans la colonne des effectifs cumulés, 100 < 114,8 < 119, donc $N(4) < N(q_{70}) < N(7)$.

La fonction N étant une fonction strictement croissante, on en déduit $4 < q_{70} < 7$ et on calcule q_{70} par interpolation linéaire : $q_{70} = 4 + [(7-4)\frac{114,8-100}{119-100}] = 4 + (3).(0,78) = 6,34$.

En utilisant les fréquences cumulées :

On constate que 61 % < 70 % < 72,6 %, soit F(4) < 70 % < F(7). Le 7e décile est donc compris entre 4 et 7. On le calcule par interpolation linéaire :

$$q_{70} = 4 + [(7-4)\frac{70\,\% - 61\,\%}{72,6\,\% - 61\,\%}] = 4 + (3).(0,78) = 6,34 .$$

On retrouve bien sûr un résultat proche de l'évaluation graphique.

Caractéristiques de position : les moyennes

CHAPITRE 3

Une moyenne est une caractéristique de tendance centrale. La moyenne la plus connue et la plus souvent utilisée est la moyenne arithmétique. La moyenne géométrique sert principalement à effectuer des calculs de coefficients multiplicateurs moyens (annuels ou trimestriels par exemple) dont sont déduits des taux de variation moyens. La moyenne harmonique permet l'évaluation de moyennes de rapports. Quant à la moyenne quadratique, en économie, gestion et sciences sociales, elle sert essentiellement à calculer une caractéristique de… dispersion.

Une moyenne peut être simple ou pondérée.

1 Moyenne simple et moyenne pondérée

Soit X une variable quantitative définie sur une population composée de n individus.

Lorsque cette variable est discrète et que ses valeurs se présentent sous la forme d'une suite de n nombres x_1, x_2…, x_n, la moyenne de X est dite simple. Sinon, elle est dite pondérée.

Lorsque la variable X est discrète et que sa distribution se présente sous la forme d'un tableau (comme suit) associant les k valeurs de la variable aux effectifs ou aux fréquences, la moyenne de X est dite pondérée car, dans son calcul, chacune des valeurs observées x_i est affectée du coefficient correspondant au nombre n_i de fois où cette valeur a été observée ou à la fréquence f_i de cette observation ; ce coefficient est un coefficient de pondération.

Valeur de X	x_1	x_2	…	x_{k-1}	x_k
Effectif	n_1	n_2	...	n_{k-1}	n_k
Fréquence	f_1	f_2	…	f_{k-1}	f_k

Lorsque la variable X est continue, la valeur exacte de X n'est pas connue pour chaque individu puisque les valeurs de la variable sont regroupées en classes.

Classe	$[e_1 ; e_2[$	$[e_2 ; e_3[$	…	$[e_{k-1} ; e_k[$	$[e_k ; e_{k+1}[$
Effectif	n_1	n_2	…	n_{k-1}	n_k
Fréquence	f_1	f_2	…	f_{k-1}	f_k

Pour pouvoir calculer une moyenne, on va supposer qu'à l'intérieur de chaque classe ces valeurs sont réparties uniformément car, dans ce cas, la moyenne des valeurs d'une classe est égale à la moyenne des extrémités de classe $\frac{e_i + e_{i+1}}{2}$. Cette moyenne est appelée centre de la classe, et notée x_i, comme les valeurs de la variable lorsqu'elle est discrète. Pour effectuer les calculs de moyenne, on « fait comme si » les valeurs de la variable d'une même classe étaient toutes égales au centre de la classe.

La moyenne de X est alors dite pondérée car, dans son calcul, chacun des centres de classe x_i est pondéré du coefficient égal au nombre n_i d'individus dont la valeur de la variable est compris entre e_i et e_{i+1} ou à la fréquence f_i correspondante.

2 Moyenne arithmétique $\bar{x}$

La moyenne arithmétique est le plus souvent nommée simplement « moyenne ».

■ *Définition*

La moyenne arithmétique de la variable X est la somme des valeurs observées divisée par le nombre d'observations n.

Si chaque valeur de la variable est observée une seule fois, la ***moyenne arithmétique*** est dite ***simple*** et s'écrit $\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$, ou en utilisant le symbole de la somme Σ (lire « sigma ») :

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

Si x_1 a été observé n_1 fois, x_2 n_2 fois…, x_k n_k fois, par définition la moyenne arithmétique est égale à :

$$\frac{(x_1 + x_1 + \ldots + x_1) + (x_2 + x_2 + \ldots + x_2) + \ldots + (x_k + x_k + \ldots + x_k)}{n}.$$

Or $(x_1 + x_1 + \ldots + x_1) = n_1 x_1$, $(x_2 + x_2 + \ldots + x_2) = n_2 x_2 \ldots$, $(x_k + x_k + \ldots + x_k) = n_k x_k$.

La formule de la ***moyenne arithmétique pondérée*** s'écrit donc $\bar{x} = \frac{n_1 x_1 + n_2 x_2 + \ldots + n_k x_k}{n_1 + n_2 + \ldots + n_k}$, ou en utilisant le symbole Σ :

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

La moyenne arithmétique pondérée peut aussi être calculée en utilisant les fréquences. En effet : $\bar{x} = \frac{n_1}{n} x_1 + \frac{n_2}{n} x_2 + \ldots + \frac{n_k}{n} x_k = f_1 x_1 + f_2 x_2 + \ldots + f_k x_k$, soit :

$$\bar{x} = \sum_{i=1}^{k} f_i x_i$$

On voit donc que, contrairement au mode et à la médiane, les autres indicateurs de tendance centrale, la moyenne prend en compte toutes les observations. Elle a aussi l'avantage de se prêter aux calculs algébriques, ce qui n'est le cas ni du mode, ni de la médiane. En revanche, il n'est pas rare qu'aucun individu de la population ne soit caractérisé par la valeur moyenne, alors que le mode est, par définition, une valeur observée de la variable. Pour la médiane, on peut affirmer qu'il s'agit d'une valeur observée seulement lorsque la variable est quantitative discrète et l'effectif total un nombre impair.

■ *Position de la moyenne arithmétique par rapport au mode et à la médiane : asymétrie*

Selon les positions respectives du mode, de la médiane et de la moyenne arithmétique, la série étudiée est symétrique ou asymétrique. Dans ce dernier cas, elle est dite étalée à droite, ou étalée à gauche.

Si $M_o = M_e = \bar{x}$, la série est parfaitement symétrique.

La représentation graphique de la série présente un axe de symétrie : la verticale qui passe par le point d'abscisse correspondant à la valeur commune de ces trois caractéristiques. La valeur de la variable la plus fréquemment observée (le mode) est aussi celle qui sépare les observations en deux groupes d'effectifs égaux (la médiane) et la valeur moyenne de l'ensemble des valeurs observées.

Si $M_o < M_e < \bar{x}$, la ***série*** est ***étalée vers la droite***.

Pour un grand nombre d'individus, les valeurs de la variable sont faibles. En revanche, pour un petit nombre elles atteignent des valeurs élevées, voire très élevées. La moyenne est alors « tirée vers la haut » par ces valeurs.

C'est le cas des salaires en France : il y a beaucoup de salariés qui ont des bas salaires, et plus on monte dans l'échelle des salaires, plus le nombre de salariés est faible. En 2005, le salaire mensuel net médian était de 1 528 euros, il était nettement inférieur au salaire mensuel net moyen qui était de 1 904 euros, soit 25 % plus élevé (source : Insee, *Tableaux de l'économie française*, édition 2007, www.insee.fr).

C'est aussi le cas des patrimoines : en 2004, le patrimoine médian des ménages était de 98 000 euros, tandis que le patrimoine moyen s'élevait à 165 000 euros, soit 68 % de plus (source : Insee, *Les revenus et le patrimoine des ménages*, édition 2006, www.insee.fr).

Si $\bar{x} < M_e < M_o$, la ***série*** est ***étalée vers la gauche***.

Les valeurs faibles de la variable « tirent la moyenne vers le bas ».

■ *Applications*

a) Moyenne arithmétique simple

Soit la série brute des notes obtenues en statistique par 19 élèves :

2 3 4 5 6 6 7 8 9 9 10 11 12 13 14 15 17 17 19.

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{2 + 3 + \ldots + 17 + 19}{19} = \frac{187}{19} = 9,8.$$

La note moyenne est 9,8.

b) Moyenne arithmétique pondérée : cas d'une variable discrète

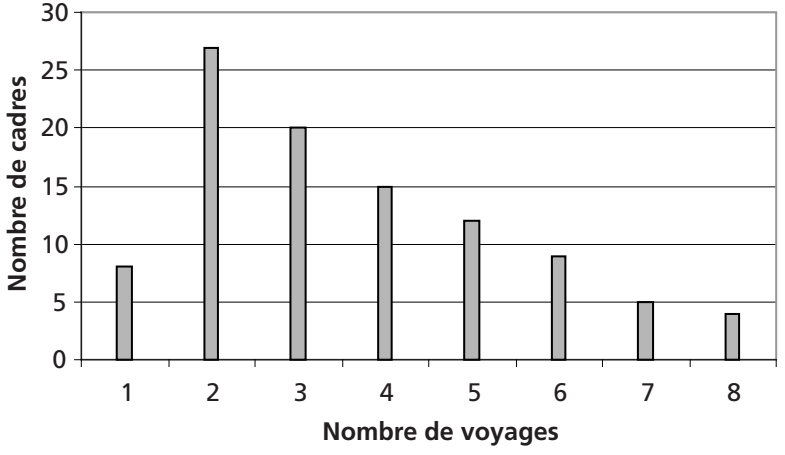
Le tableau suivant indique la répartition des cadres d'une entreprise selon le nombre de voyages en avion effectués en 2007 :

Nombre de voyages	0	1	2	3	4	5	6	7
Nombre de cadres	8	27	20	15	12	9	5	4

Le nombre moyen de voyages effectués par un cadre est :

$$\bar{x} = \frac{\sum_{i=1}^{k} n_i x_i}{n} = \frac{8.0 + 27.1 + \ldots + 5.6 + 4.7}{8 + 27 + \ldots + 5 + 4} = \frac{263}{100} = 2,63.$$

Le mode est égal à 1, la médiane est égale à 2 (le nombre de voyages a été inférieur ou égal à 2 pour 50 cadres et supérieur ou égal à 2 pour les 50 autres) et la moyenne arithmétique à 2,63 : $M_o < M_e < \bar{x}$. La série est étalée à droite, ce que confirme la graphe ci-dessous.



c) Moyenne arithmétique pondérée : cas d'une variable continue

L'Observatoire des entreprises a effectué une étude portant sur 2 255 entreprises industrielles ayant eu au moins un incident de paiement en 2004. La répartition de ces entreprises selon leur taux de défaut de paiement (mesuré par le rapport montant d'impayés de l'année/dettes fournisseurs) était la suivante (source : Banque de France, *Observatoire des entreprises*, www.banque-france.fr).

Taux de défaut de paiement (%)	Moins de 0,5	[0,5 ; 1[	[1 ; 2[	[2 ; 10[	10 ou plus
Nombre d'entreprises	651	154	202	654	594

La borne inférieure de la première classe sera assimilée à 0 et la borne de la dernière à 60.

x_i (%)	0,25	0,75	1,5	6	35	Ensemble
n_i	651	154	202	654	594	2 255
$n_i x_i$	162,75	115,5	303	3 924	20 790	25 295,25

Moyenne arithmétique des taux de défaut de paiement :

$$\bar{x} = \frac{\sum_{i=1}^{k} n_i x_i}{n} = \frac{(0,25)(651) + (0,75)(154) + \ldots + (35)(594)}{2\,255}$$

$$\bar{x} = \frac{162,75 + 115,5 + \ldots + 20\,790}{2\,255} = \frac{25\,295,25}{2\,255} = 11,2$$

Le taux moyen de défaut de paiement de ces entreprises est 11,2 %.

3 Moyenne géométrique G

■ *Définition*

La ***moyenne géométrique*** de la variable X est la ***racine n^e du produit des valeurs observées*** de cette variable, ces valeurs devant être toutes positives.

Si chaque valeur de la variable est observée une seule fois, la ***moyenne géométrique*** est dite ***simple*** et s'écrit :

$$G = \sqrt[n]{x_1 x_2 \ldots . x_n} = (x_1 x_2 \ldots . x_n)^{1/n}$$

ou, en utilisant le symbole du produit Π (lire « pi ») :

$$G = \left(\prod_{i=1}^{n} x_i \right)^{1/n}$$

Si x_1 a été observé n_1 fois, x_2 n_2 fois…, x_k n_k fois, par définition la moyenne géométrique est égale à :

$\sqrt[n]{(x_1 x_1 \ldots x_1)(x_2 x_2 \ldots x_2)\ldots(x_k x_k \ldots x_k)}$.

Or, $x_1 x_1 \ldots x_1 = x_1^{n1}$, $x_2 x_2 \ldots x_2 = x_2^{n2}$, …, $x_k x_k \ldots x_k = x_k^{nk}$.

La moyenne géométrique pondérée s'écrit donc :

$$G = \sqrt[n]{x_1^{n_1}.x_2^{n_2}.\ ....x_k^{n_k}} = (x_1^{n_1}.x_2^{n_2}.\ ....x_k^{n_k})^{1/n}$$

ou en utilisant le symbole Π :

$$G = \left(\prod_{i=1}^{k} x_i^{n_i}\right)^{1/n}$$

La moyenne géométrique pondérée peut aussi être calculée en utilisant les fréquences. En effet : $G = x_1^{n_1/n}.x_2^{n_2/n}.\ldots x_k^{n_k/n} = x_1^{f_1}.x_2^{f_2}.\ldots.x_k^{f_k}$:

$$G = x_1^{f_1}.x_2^{f_2}.\ldots.x_k^{f_k} = \prod_{i=1}^{k} x_i^{f_i}$$

La moyenne géométrique est principalement utilisée pour calculer les coefficients multiplicateurs moyens dont sont déduits les taux de variation moyens (annuels, semestriels, mensuels…). Les applications suivantes en donnent deux exemples.

■ *Applications*

a) Moyenne géométrique simple : multiplicateur annuel moyen du PIB

Le taux de variation annuel du produit intérieur brut (PIB) en volume a été le suivant en France de 2002 à 2006 (source : Insee, comptes nationaux, www.insee.fr) :

Année	2002	2003	2004	2005	2006
Taux de variation (%)	1,0	1,1	2,5	1,7	2,0

Notons PIB_{01} le PIB de l'année 2001, PIB_{02} le PIB de l'année 2002, etc.

D'après les données, le taux de variation du PIB entre 2001 et 2002 est égal à 1,0 % :

$$TV(PIB)_{02/01} = \frac{PIB_{02} - PIB_{01}}{PIB_{01}} = 1,0\ \%$$

Donc $PIB_{02} = PIB_{01} + (1,0\ \%)PIB_{01} = PIB_{01}(1+ 0,010) = PIB_{01}\ (1,010)$,

1,010 est le coefficient multiplicateur du PIB entre 2001 et 2002.

De même, on peut écrire :

$PIB_{03} = PIB_{02} + (1,1\ \%)PIB_{02} = PIB_{02}(1+ 0,011) = PIB_{02}\ (1,011) = [PIB_{01}\ (1,010)](1,011)$.

Et ainsi de suite. De proche en proche, on obtient :

$PIB_{06} = PIB_{01}(1,010)(1,011)(1,025)(1,017)(1,020) = PIB_{01}\ (1,0857)$.

En 5 ans, le PIB en volume a été multiplié par 1,0857, ou encore : le coefficient multiplicateur du PIB est 1,0857.

Soit m le coefficient multiplicateur annuel moyen. C'est le coefficient multiplicateur, ***identique chaque année***, qui aurait permis ***globalement*** au PIB de connaître, entre 2001 et 2006, la même croissance que les véritables coefficients multiplicateurs. Il vérifie donc l'égalité :

$PIB_{06} = PIB_{01}(m)^5 = PIB_{01}(1,010)(1,011)(1,025)(1,017)(1,020)$.

Donc $m^5 =(1,010)(1,011)(1,025)(1,017)(1,020)$.

D'où $m = [(1,010)(1,011)(1,025)(1,017)(1,020)]^{1/5} = (1,0857)^{1/5} = 1,0166$.

Ce multiplicateur est une moyenne géométrique simple des cinq multiplicateurs annuels :

$G = (X_1X_2\ldots X_n)^{1/n}$

En moyenne, chaque année de 2002 à 2006, le PIB en volume a été multiplié par 1,0166.

Le taux de variation du PIB en volume, ou taux de croissance de l'économie, a donc été, en moyenne, chaque année de 2002 à 2006, de 1,66 %.

b) Moyenne géométrique pondérée : multiplicateur mensuel moyen du chiffre d'affaires d'une entreprise

De janvier à décembre 2007, le taux de variation mensuel du chiffre d'affaires (CA) d'une entreprise a été de 3 % pendant les deux premiers mois, puis de 5 % pendant les trois mois suivants, puis de – 1 % pendant les deux mois suivants et enfin de 2 % pendant chacun des cinq derniers mois. Quel a été le taux de variation mensuel moyen du chiffre d'affaires ?

Soit CA_0 le chiffre d'affaires de l'entreprise en décembre 2006, CA_1 son chiffre d'affaires en janvier 2007…, CA_{12} son chiffre d'affaires en décembre 2007.

D'après les données :

$CA_1 = CA_0 + (3\%)CA_0 = CA_0(1 + 0,03) = CA_0\ (1,03)$

$CA_2 = CA_1 + (3\%)CA_1 = CA_1(1 + 0,03) = CA_1\ (1,03) = [CA_0(1,03)](1,03) = CA_0\ (1,03)^2$

et ainsi de suite.

On obtient finalement :

$CA_{12} = CA_0\ (1,03)^2(1,05)^3(0,99)^2(1,02)^5 = CA_0\ (1,329)$.

À noter que dans cette égalité le coefficient multiplicateur 0,99 est inférieur à 1 parce que le taux de variation correspondant est négatif (– 1 %, soit – 0 ,01) :

Coefficient multiplicateur = 1 + Taux de variation = 1 – 0,01 = 0,99.

En un an, le CA a été multiplié par 1,329 ; il a donc augmenté de 32,9 %.

Soit m le coefficient multiplicateur mensuel moyen. C'est le coefficient multiplicateur, identique chaque mois, qui aurait permis globalement au CA de connaître en un an une croissance de 32,9 %. Il vérifie donc l'égalité :

$CA_{10} = CA_0 (1,03)^2(1,05)^3(0,99)^2(1,02)^5 = CA_0 (m)^2(m)^3(m)^2(m)^5 = CA_0 (1,329)$

Donc $m^{12} = (1,03)^2(1,05)^3(0,99)^2(1,02)^5$

d'où $m = [(1,03)^2(1,05)^3(0,99)^2(1,02)^5]^{1/12} = (1,329)^{1/12} = 1,024$.

En moyenne chaque mois, le chiffre d'affaires a été multiplié par 1,024. Le taux de variation mensuel moyen du chiffre d'affaires est donc égal à 2,4 %.

Le coefficient multiplicateur mensuel moyen est une moyenne géométrique pondérée des quatre coefficients multiplicateurs :

$$G = (x_1^{n_1}.x_2^{n_2}.x_3^{n_3}.x_4^{n_4})^{1/n}$$

avec $x_1 = 1,03$ et $n_1 = 2$, $x_2 = 1,05$ et $n_2 = 3$, $x_3 = 0,99$ et $n_3 = 2$, $x_4 = 1,02$ et $n_4 = 5$.

4 Moyenne harmonique H

■ *Définition*

La ***moyenne harmonique*** d'une variable X est égale à l'***inverse de la moyenne arithmétique des inverses des valeurs observées*** de cette variable (aucune des valeurs observées ne devant être nulle).

Si chaque valeur de la variable est observée une seule fois, la ***moyenne harmonique est dite simple*** et s'écrit :

$$H = \frac{1}{\frac{1}{n}\left(\frac{1}{x_1} + \frac{1}{x_2} + ... + \frac{1}{x_n}\right)}$$

ou, en utilisant le symbole Σ :

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

Si x_1 a été observé n_1 fois, x_2 n_2 fois, …, x_k n_k fois, par définition la moyenne harmonique est égale à :

$$\frac{1}{\frac{1}{n}[(\frac{1}{x_1}+\frac{1}{x_1}+\ldots+\frac{1}{x_1})+(\frac{1}{x_2}+\frac{1}{x_2}+\ldots+\frac{1}{x_2})+\ldots+(\frac{1}{x_k}+\frac{1}{x_k}+\ldots+\frac{1}{x_k})]} = \frac{1}{\frac{1}{n}(\frac{n_1}{x_1}+\frac{n_2}{x_2}+\ldots+\frac{n_k}{x_k})}.$$

La moyenne harmonique pondérée s'écrit donc :

$$H = \frac{1}{\frac{1}{n}(\frac{n_1}{x_1}+\frac{n_2}{x_2}+\ldots.+\frac{n_k}{x_k})}$$

ou :

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

La moyenne harmonique pondérée peut aussi être calculée en utilisant les fréquences.

$$H = \frac{1}{\frac{1}{n}(\frac{n_1}{x_1}+\frac{n_2}{x_2}+\ldots.+\frac{n_k}{x_k})} = \frac{1}{\frac{n_1/n}{x_1}}+\frac{1}{\frac{n_2/n}{x_2}}+\ldots+\frac{1}{\frac{n_k/n}{x_k}} = \frac{1}{\frac{f_1}{x_1}}+\frac{1}{\frac{f_2}{x_2}}+\ldots+\frac{1}{\frac{f_k}{x_k}}, \text{ soit :}$$

$$H = \frac{1}{\frac{f_1}{x_1}}+\frac{1}{\frac{f_2}{x_2}}+\ldots+\frac{1}{\frac{f_k}{x_k}} = \frac{1}{\sum_{i=1}^{k} \frac{f_i}{x_i}}$$

La moyenne harmonique est employée dans le calcul des moyennes de rapports. Deux exemples sont donnés dans les applications suivantes.

■ *Applications*

a) Moyenne harmonique simple : vitesse moyenne

Un automobiliste effectue un aller–retour entre deux villes. À l'aller sa vitesse moyenne est de 130 km/h ; au retour, elle n'est que de 70 km/h. Quelle a été sa vitesse moyenne sur l'ensemble du parcours ?

La réponse n'est pas la moyenne arithmétique des vitesses : (130 + 70)/2 = 100 ! Pourquoi ? Parce que la vitesse moyenne est un rapport, le rapport de la distance totale parcourue au temps nécessaire pour la parcourir. Pour la calculer, notons d la distance entre les deux villes. Le numérateur du rapport, la distance totale parcourue, est égal à 2d.

D'après la définition de la vitesse moyenne, le temps nécessaire pour parcourir un trajet est le rapport de la distance à la vitesse moyenne, soit d/130 pour l'aller et d/70 pour le retour. Le dénominateur du rapport est donc égal à $\frac{d}{130}+\frac{d}{70}$.

Le calcul de la vitesse moyenne est le suivant :

$$\text{Vitesse moyenne} = \frac{\text{Distance}}{\text{Temps}} = \frac{2d}{\frac{d}{130}+\frac{d}{70}} = \frac{2}{\frac{1}{130}+\frac{1}{70}} = 91 \text{ km/h}.$$

La vitesse moyenne est une moyenne harmonique simple :

$H = \frac{1}{\frac{1}{n}(\frac{1}{x_1}+\frac{1}{x_2})} = \frac{n}{(\frac{1}{x_1}+\frac{1}{x_2})}$, avec n = 2, x_1 = 130 et x_2 = 70.

b) Moyenne harmonique pondérée : densité moyenne

En Limousin, au 1er janvier 2006, la part de la population des trois départements dans la population régionale et la densité moyenne étaient les suivantes (source : Insee, www.insee.fr) :

	Corrèze	Creuse	Haute-Vienne
Part de la population dans la population régionale	32,8 %	16,9 %	50,3 %
Densité (hab/km²)	41	22	66

Quelle était la densité moyenne de la population en Limousin en 2006 ?

La densité moyenne est le nombre moyen d'habitants au km².

Soit n le nombre d'habitants en Limousin en 2006.

D'après la définition de la densité, la surface de chaque département est le rapport du nombre d'habitants à la densité. En Corrèze, par exemple, la population représente 32,8 % de la population du Limousin, le nombre d'habitants est égal à 0,328 n, et la densité moyenne vaut 41 ; la surface de la Corrèze s'obtient alors en effectuant 0,328 n/41. On procède de même pour les deux autres départements. On obtient finalement :

$$\text{Densité} = \frac{\text{Nombre d'habitants}}{\text{Surface}} = \frac{n}{\frac{0,328n}{41}+\frac{0,169n}{22}+\frac{0,503n}{66}} = \frac{1}{\frac{0,328}{41}+\frac{0,169}{22}+\frac{0,503}{66}} = 42,9$$

Il y avait en moyenne, au 1er janvier 2006, en Limousin, 43 habitants au km².

Nous constatons que le calcul effectué est une moyenne harmonique pondérée :

$$H = \frac{1}{\frac{f_1}{x_1}+\frac{f_2}{x_2}+\ldots+\frac{f_k}{x_k}} = \frac{1}{\sum_{i=1}^{k}\frac{f_i}{x_i}}.$$

5 Moyenne quadratique Q

En sciences économiques, sciences sociales et gestion, la moyenne quadratique est utilisée essentiellement pour calculer un indicateur de dispersion autour de la moyenne, l'écart-type, que nous définirons au chapitre 4.

La moyenne quadratique d'une variable est la racine carrée de la moyenne arithmétique des carrées des valeurs observées de cette variable.

La moyenne quadratique simple s'écrit :

$$Q = \sqrt{\frac{x_1^2 + x_2^2 + \ldots + x_n^2}{n}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$$

La formule de la moyenne quadratique pondérée est :

$$Q = \sqrt{\frac{n_1x_1^2 + n_2x_2^2 + \ldots + n_kx_k^2}{n}} = \sqrt{\frac{1}{n}\sum_{i=1}^{k} n_ix_i^2}$$

En utilisant les fréquences, elle s'écrit :

$$Q = \sqrt{f_1 x_1^2 + f_2 x_2^2 + \ldots + f_k x_k^2} = \sqrt{\sum_{i=1}^{k} f_i x_i^2}$$

Remarque : lorsque, pour une série statistique, chacune des quatre moyennes existe, on peut démontrer que :

$$H \leq G \leq \bar{x} \leq Q$$

Conseils

1. Lorsqu'on calcule une moyenne, il faut d'abord réfléchir à la signification du calcul à réaliser et construire son calcul par rapport à ce raisonnement, plutôt que de chercher à utiliser *a priori* telle ou telle formule de moyenne.

Un taux de variation annuel moyen n'est pas la moyenne arithmétique des taux de variation annuels ! Il se déduit du coefficient multiplicateur annuel moyen qui est une moyenne géométrique des coefficients multiplicateurs annuels.

Un rapport moyen n'est pas une moyenne arithmétique de rapports, mais une moyenne harmonique.

2. Une caractéristique de position ne peut jamais être supérieure à la plus grande valeur de la variable ni être inférieure à la plus petite de ces valeurs. Un résultat hors de ces bornes est nécessairement faux ; il faut alors chercher où est l'erreur…

3. Pour résumer la tendance centrale d'une série, il ne faut pas hésiter à déterminer les trois caractéristiques – mode, médiane et moyenne arithmétique – car elles sont complémentaires. Dans le cadre d'un rapport ou d'un mémoire, un commentaire précisant la signification de chacun des résultats est indispensable pour éclairer le lecteur ; s'il s'agit d'un rapport ou mémoire d'études, il permet au rédacteur de mettre en évidence auprès du jury ses compétences en analyse statistique.

Les caractéristiques de dispersion

CHAPITRE 4

Deux séries peuvent avoir la même moyenne, voire la même médiane et le même mode, mais correspondre à des observations qui se distribuent très différemment : dans l'une, les valeurs du caractère peuvent s'écarter considérablement des valeurs centrales, et, dans l'autre en être très proches. On ne peut donc se limiter aux valeurs centrales pour caractériser une série.

L'objet de ce chapitre est de définir les principaux indicateurs qui permettent de mesurer la dispersion des observations autour des valeurs centrales. On les appelle caractéristiques ou paramètres de dispersion.

1 Étendue

Un enseignant a corrigé les copies des étudiants de deux groupes de travaux dirigés (Groupe 1 et Groupe 2) de 30 étudiants chacun. Les notes classées par valeurs croissantes sont les suivantes :

Groupe 1

1	1	2	2	4	4	6	7	7	8	8	8	9	9	11	11	11	11	12	12	13	14	14	14	15	15	16	17	18	20

Groupe 2

6	7	7	7	7	7	8	8	8	9	9	9	9	10	11	11	11	11	11	11	12	12	12	12	12	12	12	13	13	13

Une observation rapide montre que les notes de la première série vont de très faibles à très élevées tandis que, dans la deuxième, les notes ne descendent pas en dessous de 6, mais ne dépassent pas 13. Une première manière simple de traduire cette observation est de calculer l'étendue de chaque série.

L'étendue est égale à la différence entre la plus grande et la plus petite des valeurs observées.

Dans notre exemple, l'étendue de la 1re série est 19 alors que celle de la deuxième est 7 ; la dispersion des notes est donc nettement plus forte dans le premier groupe que dans le deuxième.

L'étendue est l'indicateur de dispersion le plus simple à évaluer. En outre, sa signification est aisée à comprendre. Néanmoins, l'étendue a l'inconvénient de ne pas se prêter au calcul car on ne peut l'intégrer dans une formule.

Les autres indicateurs de dispersion plus élaborés sont des indicateurs de dispersion autour de la médiane et autour de la moyenne.

Chacune des deux séries de trente notes ci-dessus a la même médiane, c'est la note 11 : dans chaque groupe de travaux dirigés, 15 étudiants ont une note inférieure ou égale à 11 et 15 une note supérieure ou égale à 11. Elles ont aussi la même moyenne : dans les deux cas, la somme des notes est égal à 300, la moyenne est donc égale à 10. On voit bien néanmoins que les notes s'éloignent beaucoup plus de ces deux valeurs centrales dans la 1re série que dans la 2e. Nous allons voir comment mesurer ces dispersions.

2 Indicateurs de dispersion autour de la médiane

Les principaux indicateurs de dispersion autour de la médiane sont les intervalles, écarts et rapports interquantiles. Ils peuvent être complétés par un diagramme appelé boîte à moustaches.

■ *Intervalles, écarts et rapports interquantiles*

Il existe trois intervalles interquantiles tous centrés sur la médiane :

– l'intervalle interquartile, noté I_{50}, est $[q_{25} ; q_{75}]$. Il contient 50 % des observations centrales ;

– l'intervalle interdécile, noté I_{80}, est $[q_{10} ; q_{90}]$. Il contient 80 % des observations centrales ;

– l'intervalle intercentile, noté I_{98}, est $[q_1 ; q_{99}]$. Il contient 98 % des observations centrales.

Plus l'écart entre les bornes de chacun de ces intervalles est grand, plus la dispersion de la distribution autour de la médiane est grande. Pour apprécier cet écart, on peut calculer :

soit la différence entre les bornes : c'est l'écart interquantile ;

soit le rapport entre les bornes : c'est le rapport interquantile.

	interquartile	interdécile	intercentile
Intervalle	$[Q_{25} ; Q_{75}]$	$[Q_{10} ; Q_{90}]$	$[Q_1 ; Q_{99}]$
Écart	$Q_{75} - Q_{25}$	$Q_{90} - Q_{10}$	$Q_{99} - Q_1$
Rapport	Q_{75}/ Q_{25}	Q_{90}/Q_{10}	Q_{99}/Q_1

Ces écarts et rapports interquantiles sont essentiellement utilisés pour comparer la dispersion de plusieurs distributions statistiques à une même date ou la dispersion d'une même série à deux dates différentes.

■ *Boîte à moustaches*

Une boîte à moustaches est un indicateur graphique de dispersion qui permet de visualiser la plus ou moins grande dispersion d'une distribution en matérialisant l'écart interquartile par une « boîte » et l'étendue par des « moustaches ».

Ce graphe prend place dans un système d'axes orthonormés, l'axe des ordonnées étant gradué de la plus petite à la plus grande valeur du caractère.

La boîte matérialise l'écart interquartile : c'est un rectangle dont la base inférieure a pour ordonnée le premier quartile et la base supérieure le troisième quartile ; à l'intérieur de ce rectangle, un trait horizontal indique la position de la médiane (exactement au milieu de la boîte lorsque la distribution est parfaitement symétrique).

Les moustaches, de part et d'autre de la boîte, matérialisent l'étendue : la première est un trait vertical qui joint la plus petite valeur de la variable (le minimum) à la base inférieure du rectangle (le premier quartile), la deuxième est un trait vertical qui joint la base supérieure du rectangle (le troisième quartile) à la plus grande valeur de la variable (le maximum).

■ *Applications*

a) Intervalle, écart et rapport interdéciles des salaires en France

Le tableau suivant fournit la distribution des salaires annuels nets de tous prélèvements, en euros courants, en France, en 2005 (source : Insee, www.insee.fr). Les déciles sont ici notés D1, D2…, D9.

Décile	Femmes	Hommes
1er décile (D1)	11 853	12 983
2e décile (D2)	13 144	14 530
3e décile (D3)	14 238	15 948
4e décile (D4)	15 394	17 447
Médiane (D5)	**16 845**	**19 162**
6e décile (D6)	18 624	21 348
7e décile (D7)	20 948	24 433
8e décile (D8)	24 133	29 399
9e décile (D9)	30 324	39 760

Ce tableau nous permet de comparer la dispersion autour de la médiane des salaires annuels nets des hommes et des femmes en déterminant l'intervalle, l'écart et le rapport interdécile.

	Femmes	Hommes
Intervalle interdécile	[11 853 ; 30 324]	[12 983 ; 39 760]
Écart interdécile	18 471	26 777
Rapport interdécile	2,56	3,06

L'intervalle interdécile est plus large chez les hommes que chez les femmes, l'écart et le rapport interdécile sont donc plus grands :

– l'écart interdécile des salaires des hommes est 45 % plus élevé que celui des femmes ;

– pour les salariés hommes, le dernier décile est égal à 3,1 fois le premier alors que, chez les femmes, il ne représente que 2,6 fois le premier.

La dispersion des salaires est donc plus importante dans l'ensemble des hommes salariés que des femmes salariées.

Pour affiner l'analyse, on peut décomposer l'écart et le rapport en distinguant l'écart et le rapport d'une part entre la médiane et le premier décile et d'autre part entre le dernier décile et la médiane : D9 – D1 = (D9 – D5) + (D5 – D1) et D9/D1 = (D9/D5)(D5/D1).

	Femmes	Hommes
D9 – D5	13 479	20 598
D5 – D1	4 992	6 179
D9 – D1	18 471	26 777

	Femmes	Hommes
D9/D5	1,80	2,07
D5/D1	1,42	1,48
D9/D1	2,56	3,06

Ce tableau montre que, pour les femmes comme pour les hommes, l'écart entre la médiane et le dernier décile est nettement plus élevé qu'entre le premier décile et la médiane : la dispersion des salaires est plus importante au sein des salaires élevés (au-dessus de la médiane) qu'au sein des petits salaires (en dessous de la médiane).

b) Boîte à moustaches

Revenons sur l'exemple des notes des deux groupes d'étudiants.

Groupe 1

1	1	2	2	4	4	6	7	7	8	8	8	9	9	11	11	11	11	12	12	13	14	14	14	15	15	16	17	18	20

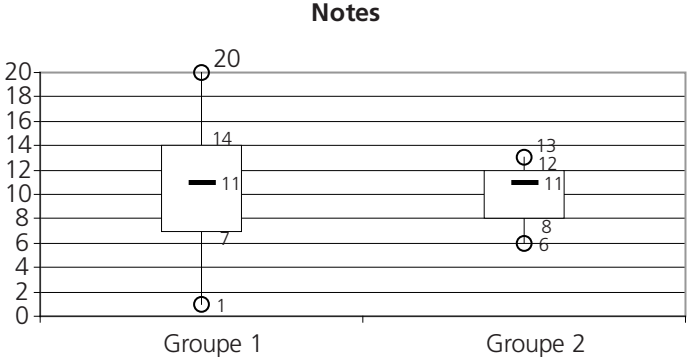
Groupe 2

6	7	7	7	7	7	8	8	8	9	9	9	9	10	11	11	11	11	11	11	12	12	12	12	12	12	12	13	13	13

Le tableau suivant indique les quartiles, le minimum et le maximum de chaque groupe :

Note	Groupe 1	Groupe 2
Minimum	1	6
Premier quartile	7	8
Médiane	11	11
Troisième quartile	14	12
Maximum	20	13

On peut alors tracer les deux boîtes à moustaches :



Ce graphe met clairement en évidence :

– une même note médiane dans les deux groupes : 11 ;

– une dispersion des notes du groupe 1 nettement plus forte que celle du groupe 2 :

- du point de vue de l'étendue : 20 – 1 pour G1, 13 – 6 pour G2,
- du point de vue de l'écart interquartile : 14 – 7 pour G1, 12 – 8 pour G2.

3 Indicateurs de dispersion autour de la moyenne

Les indicateurs de dispersion autour de la moyenne les plus usités sont la variance, l'écart-type et le coefficient de variation.

Écart-type et variance reposent sur le calcul de la différence moyenne (ou écart moyen) entre chaque valeur de la variable et la moyenne arithmétique de ces valeurs. La moyenne étant une valeur centrale, certains de ces écarts sont positifs mais d'autres négatifs. Si l'on effectue directement la somme de ces écarts pour en calculer ensuite la moyenne, les écarts positifs seront exactement compensés par les écarts négatifs (de par la définition même de la moyenne) et la moyenne des écarts sera nulle. La variance fournit une solution en effectuant le calcul de la moyenne non pas des écarts eux-mêmes, mais du carré de chacun de ces écarts, les carrés étant nécessairement positifs ou nuls.

■ *Variance*

La variance d'une variable statistique X est égale à la moyenne des carrés des écarts entre les valeurs de la variable et la moyenne. Plus la valeur de la variance est élevée, plus les écarts entre les valeurs de la variable et la moyenne sont grands, donc plus la dispersion autour de la moyenne est grande.

a) Cas où les données ne sont pas groupées dans un tableau de distribution

Les n valeurs de la variable sont notées x_1, x_2…, x_n. On effectue, pour chaque valeur de la variable, le calcul de l'écart entre cette valeur et la moyenne arithmétique $(x_i - \overline{x})$, puis on élève au carré cet écart $(x_i - \overline{x})^2$. Ces écarts au carré sont ensuite additionnés et la somme ainsi obtenue $\sum_{i=1}^{n}(x_i - \overline{x})^2$ est divisée par le nombre d'écarts (donc le nombre de valeurs observées n) pour obtenir la moyenne :

$$V(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

En développant et simplifiant cette formule, on obtient une autre formulation de la variance :

$$V(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i}(x_i^2 - 2x_i\overline{x} + \overline{x}^2)$$

$$= \frac{1}{n}\sum_{i}x_i^2 - 2\overline{x}\left(\frac{1}{n}\sum_{i}x_i\right) + \frac{1}{n}\sum_{i}\overline{x}^2 = \frac{1}{n}\sum_{i}x_i^2 - 2\overline{x}^2 + \frac{1}{n}n\overline{x}^2 = \frac{1}{n}\sum_{i}x_i^2 - \overline{x}^2$$

La formule obtenue est dite formule de Koenig :

$$V(X) = \frac{1}{n}\sum_i x_i^2 - \bar{x}^2$$

Cette formule a le mérite d'être plus simple et rapide à mettre en œuvre que la formule de définition : on calcule la somme des x_i^2, on divise cette somme par n, puis on soustrait au résultat le carré de la moyenne arithmétique. Autrement dit la variance est la moyenne des carrés des valeurs observées moins le carré de la moyenne.

Attention de ne pas soustraire le carré de la moyenne arithmétique à la somme des x_i^2 pour ensuite diviser le résultat par n ! C'est un erreur commise couramment par les étudiants ! Pour éviter de se tromper, il suffit d'écrire $V(X) = (\frac{1}{n}\sum_i x_i^2) - \bar{x}^2$. Les parenthèses introduites ici ne sont pas nécessaires, mais peuvent rendre service aux étudiants étourdis.

b) Cas de données groupées dans un tableau de distribution

Soit une variable quantitative X définie sur une population composée de n individus. Les valeurs de cette variable ou les centres de classes sont x_1, x_2…, x_k. Les effectifs sont n_1, n_2…, n_k.

Le premier écart $(x_1 - \bar{x})$ ayant été observé n_1 fois, le deuxième écart $(x_2 - \bar{x})$ n_2 fois…, le k^e n_k fois, la variance s'écrit :

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

C'est la moyenne pondérée des carrés des écarts à la moyenne.

Comme précédemment, si l'on développe puis simplifie cette formule, on obtient la formule de Koenig :

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

La variance peut aussi s'exprimer en fonction des fréquences :

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

$$V(X) = \sum_{i=1}^{k} f_i (x_i - \bar{x})^2$$

La formule de Koenig s'écrit alors :

$$V(X) = \sum_{i=1}^{k} f_i x_i^2 - \bar{x}^2$$

■ *Écart-type*

a) Définition

La variance étant une moyenne d'écarts élevés au carré, l'unité dans laquelle elle s'exprime est le carré de l'unité dans laquelle est évaluée la variable (des « euros au carré » par exemple si la variable est le salaire en euros). Sa valeur n'est donc pas directement interprétable.

En revanche, en calculant la racine carrée de ce nombre, on obtient une valeur dont la signification concrète est simple : c'est l'écart moyen entre une valeur quelconque de la variable et la valeur moyenne de la variable, exprimé dans la même unité que la variable. Cet écart moyen se nomme écart-type.

Lorsque les données ne sont pas groupées dans un tableau de distribution, il s'écrit :

$$\sigma(X) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Lorsqu'elles sont groupées dans un tableau de distribution, il s'écrit :

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

L'écart-type est une moyenne quadratique. Nous avons vu dans le chapitre 3 que la moyenne quadratique d'une variable est la racine carrée de la moyenne arithmétique des carrés des valeurs observées de cette variable :

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

Dans la formule de l'écart-type, ce n'est pas la moyenne arithmétique des carrés des valeurs observés dont on prend la racine carrée, mais la moyenne arithmétique des écarts entre les valeurs de la variable et la moyenne.

L'écart-type est donc la moyenne quadratique des écarts à la moyenne.

b) Position de l'écart-type par rapport à la moyenne

Plus l'écart-type est faible, plus les valeurs de la variable sont proches de $\bar{x}$, plus la moyenne arithmétique est représentative de l'ensemble des valeurs de la variable. Plus il est élevé, moins la

moyenne arithmétique est pertinente pour résumer l'ensemble des données. ***Il est donc indispensable de compléter le calcul de la moyenne par celui de l'écart-type.***

L'écart-type peut être supérieur à la moyenne. Plus la distribution est asymétrique, plus il est probable qu'il en sera ainsi.

En ajoutant l'écart-type à la moyenne, on peut obtenir une valeur supérieure à la plus grande valeur de la variable. En soustrayant l'écart-type à la moyenne, on peut obtenir une valeur inférieure à la plus petite valeur de la variable. En revanche, si $\bar{x} + \sigma$ est supérieur à la plus grande valeur de la variable, alors $\bar{x} - \sigma$ ne peut pas être inférieur à la plus petite valeur de la variable. De même, si $\bar{x} - \sigma$ est inférieur à la plus petite valeur de la variable, alors $\bar{x} + \sigma$ ne peut pas être supérieur à la plus grande valeur de la variable.

■ *Coefficient de variation*

Le coefficient de variation est un indicateur de dispersion relative utilisé à la place de l'écart-type dans deux cas :

– ***cas 1*** : lorsqu'on veut comparer la dispersion de deux séries qui ne sont pas exprimées dans la même unité : on ne peut pas utiliser l'écart-type puisqu'il est exprimé dans la même unité que la variable ;

– ***cas 2*** : lorsqu'on veut comparer la dispersion de deux séries exprimées dans la même unité mais dont les ordres de grandeurs sont très différents dans cette même unité : l'écart-type ne permet pas une comparaison pertinente parce qu'il est fonction de la moyenne ; plus les valeurs de la variable sont élevées, plus l'écart-type a des chances de l'être.

Pour supprimer à la fois l'effet d'unité et l'effet d'ordre de grandeur, on divise l'écart-type par la moyenne. On obtient ainsi un coefficient sans dimension, appelé coefficient de variation :

$$CV = \frac{\sigma(X)}{\bar{x}}$$

Plus le coefficient de variation est faible, plus les valeurs de la variable sont proches de $\bar{x}$, donc plus la moyenne est bien représentative de l'ensemble des valeurs de la variable.

■ *Applications*

a) Variable quantitative discrète

1) Données non groupées dans un tableau : la dispersion de notes autour de la moyenne

Revenons sur l'exemple des notes des deux groupes d'étudiants.

Groupe 1

1	1	2	2	4	4	6	7	7	8	8	8	9	9	11	11	11	11	12	12	13	14	14	14	15	15	16	17	18	20

Groupe 2

6	7	7	7	7	7	8	8	8	9	9	9	9	10	11	11	11	11	11	11	12	12	12	12	12	12	12	13	13	13

La note moyenne du groupe 1 est $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n} = \dfrac{300}{30} = 10$.

La variance est $V(X) = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \dfrac{1}{30}[(1-10)^2 + (1-10)^2 + \ldots + (20-10)^2] = 25,93$.

En utilisant la formule de Koenig : $V(X) = (\dfrac{1}{n}\sum_{i=1}^{n} x_i^2) - \bar{x}^2 = \dfrac{3778}{30} - 10^2 = 25,93$.

D'où $\sigma(x) = \sqrt{25,93} = 5,1$: l'écart moyen entre une note du groupe 1 et la note moyenne de ce groupe est de 5,1 points.

La note moyenne du groupe 2 est $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n} = \dfrac{300}{30} = 10$.

La variance est $V(X) = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \dfrac{1}{30}[(6-10)^2 + (7-10)^2 + \ldots + (13-10)^2] = 4,60$.

En utilisant la formule de Koenig : $V(X) = (\dfrac{1}{n}\sum_{i=1}^{n} x_i^2) - \bar{x}^2 = \dfrac{3138}{30} - 10^2 = 4,60$.

D'où $\sigma(x) = \sqrt{4,60} = 2,1$: l'écart moyen entre une note du groupe 2 et la note moyenne de ce groupe est de 2,1 points.

La dispersion autour de la moyenne du groupe 1 est quasiment le double de celle du groupe 2. La note moyenne de 10 est donc beaucoup plus représentative de l'ensemble des notes dans le groupe 2 que dans le groupe 1.

Ici, il est inutile de calculer le coefficient de variation pour comparer la dispersion autour de la moyenne des deux groupes : la moyenne des deux groupes étant la même, les coefficients de variation sont proportionnels aux écart-types.

2) Données non groupées dans un tableau : la volatilité de 2 actions

Le calcul de l'écart-type est utilisé sur les marchés financiers pour évaluer ce que les cambistes appellent la « volatilité » du rendement d'un titre, une action par exemple.

Le cours d'une action varie chaque jour. Le rendement d'une action sur une période est mesuré par le taux de variation de son cours pendant cette période.

La volatilité du rendement peut être calculée sur un mois par exemple. Le rendement moyen est la moyenne arithmétique des rendements journaliers et la volatilité l'écart moyen autour de ce rendement moyen.

Plus la volatilité du rendement d'une action est élevée, plus elle est risquée.

Les cours des actions ***Groupe Danone*** et ***Club Méditerranée*** ont été les suivants, durant le mois de janvier 2008, à la Bourse de Paris (source : boursorama.fr) :

Date	Groupe Danone	Club Méditerranée	Date	Groupe Danone	Club Méditerranée
2/1	61,24	43,24	17/1	58,45	31,45
3/1	61,04	42,26	18/1	58,6	29,9
4/1	59,98	40,15	21/1	54,6	27,05
7/1	61,8	38,95	22/1	55,73	28
8/1	63,71	38,59	23/1	52,79	27,99
9/1	61,8	36,8	24/1	53,08	32,2
10/1	60,19	34,53	25/1	53,32	32,25
11/1	58,26	34,46	28/1	52,86	31,77
14/1	59	34,39	29/1	53,8	32,75
15/1	57,34	32,82	30/1	53	32,67
16/1	58,46	30,75	31/1	54,03	32,01

Pour évaluer la volatilité de chacune de ces actions pendant le mois de janvier, il faut d'abord calculer leur rendement journalier, ou taux de variation du cours entre deux journées successives.

Par exemple, le 3 janvier 2008, le cours de clôture de l'action Groupe Danone était 61,04 €, la veille il était de 61,24 €. Le 3/1/08, le rendement journalier de cette action valait :

$$\frac{61{,}04 - 61{,}24}{61{,}24} = -0{,}0033 = -0{,}33\ \%$$

L'ensemble des rendements journaliers figurent dans le tableau ci-dessous :

Date	Groupe Danone	Club Méditerranée	Date	Groupe Danone	Club Méditerranée
2/1			17/1	– 0,02 %	2,28 %
3/1	– 0,33 %	– 2,27 %	18/1	0,26 %	– 4,93 %
4/1	– 1,74 %	– 4,99 %	21/1	– 6,83 %	– 9,53 %
7/1	3,03 %	– 2,99 %	22/1	2,07 %	3,51 %
8/1	3,09 %	– 0,92 %	23/1	– 5,28 %	– 0,04 %
9/1	– 3,00 %	– 4,64 %	24/1	0,55 %	15,04 %
10/1	– 2,61 %	– 6,17 %	25/1	0,45 %	0,16 %
11/1	– 3,21 %	– 0,20 %	28/1	– 0,86 %	– 1,49 %
14/1	1,27 %	– 0,20 %	29/1	1,78 %	3,08 %
15/1	– 2,81 %	– 4,57 %	30/1	– 1,49 %	– 0,24 %
16/1	1,95 %	– 6,31 %	31/1	1,94 %	– 2,02 %

Calculons la moyenne arithmétique et l'écart-type de chacune de ces séries :

– action *Groupe Danone* :

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{-11{,}76\ \%}{21} = -0{,}56\ \% = -0{,}0056$$

$$V(X) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - \bar{x}^2 = \frac{0{,}0150}{21} - (-0{,}0056)^2 = 0{,}0007 \text{ et } \sigma(X) = \sqrt{V(X)} = 2{,}6\ \%$$

– action *Club Méditerranée* :

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{-27,44\ \%}{21} = -1,31\ \% = -0,0131$$

$$V(X) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - \bar{x}^2 = \frac{0,0535}{21} - (-0,0131)^2 = 0,0024 \text{ et } \sigma(X) = \sqrt{V(X)} = 4,9\ \%$$

Conclusion : en janvier 2008, la volatilité de l'action *Club Méditerranée* a représenté presque le double de celle de l'action *Groupe Danone*.

3) Données groupées dans un tableau

Le propriétaire d'une importante concession-automobile a fait le bilan du nombre de voitures vendues par jour pendant 60 jours.

Nombre de voitures vendues	0	1	2	3	4	5	6	7	8	9	10
Nombre de jours	8	7	3	6	3	4	5	4	5	7	8

Calcul de la moyenne arithmétique : $\bar{x} = \frac{\sum_{i=1}^{k} n_i x_i}{n} = \frac{(8)(0) + (7)(1) + \ldots + (8)(10)}{60} = \frac{304}{60} = 5,07$.

x_i	0	1	2	3	4	5	6	7	8	9	10	Ensemble
n_i	8	7	3	6	3	4	5	4	5	7	8	50
$n_i x_i$	0	7	6	18	12	20	30	28	40	63	80	304
$n_i x_i^2$	0	7	12	54	48	100	180	196	320	567	800	2 284

Le concessionnaire a vendu en moyenne 5 voitures par jour.

Calcul de la variance en utilisant la formule de Koenig :

$$V(X) = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^2 - \bar{x}^2 = \frac{(8)(0^2) + (7)(1^2) \ldots + (8)(10^2)}{60} - 5,07^2 = \frac{2\,284}{60} - 5,07^2 = 12,396$$

Écart-type de X : $\sigma(x) = \sqrt{12,396} = 3,52$.

L'écart moyen entre le nombre de voitures vendues en une journée et le nombre moyen de voitures vendues est égal à 3,5. Cet écart est élevé par rapport à l'étendue de la distribution :

$\bar{x} - \sigma = 1,5$: $\bar{x} - \sigma$ est proche de 0, la plus petite valeur de la variable ;
$\bar{x} + \sigma = 8,5$: $\bar{x} + \sigma$ est proche de 10, la plus grande valeur de la variable.
La dispersion autour de la moyenne est forte. La moyenne ne résume pas bien les données.

b) Variable quantitative continue

Une enquête sur les revenus mensuels des ménages d'une commune a permis d'établir le tableau suivant :

Revenu (milliers d'euros)	[1 ; 2[	[2 ; 3[	[3 ; 4[	[4 ; 5[	[5 ; 6[	[6 ; 7[
Nombre de ménages	150	180	220	300	100	50

Le calcul de la moyenne arithmétique, nécessaire au calcul de la variance, exige le calcul préalable des centres de classe.

$[e_i ; e_{i+1}[$	x_i	n_i	$n_i x_i$	$n_i x_i^2$
[1 ; 2[	1,5	150	225	337,5
[2 ; 3[	2,5	180	450	1125
[3 ; 4[	3,5	220	770	2 695
[4 ; 5[	4,5	300	1 350	6 075
[5 ; 6[	5,5	100	550	3 025
[6 ; 7[	6,5	50	325	2 112,5
Ensemble		1 000	3 670	15 370

$$\bar{x} = \frac{\sum_{i=1}^{k} n_i x_i}{n} = \frac{(150)(1,5) + (180)(2,5) + \ldots + (50)(6,5)}{1\ 000} = \frac{225 + 450 + \ldots + 325}{1\ 000} = \frac{3\ 670}{1\ 000} = 3,67.$$

Le revenu mensuel moyen d'un ménage est égal à 3 670 euros.

Calcul de la variance (formule de Koenig) :

$$V(X) = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^2 - \bar{x}^2 = \frac{(150)(1,5^2) + (180)(2,5^2) + \ldots + (50)(6,5^2)}{1\ 000} - 3,67^2$$

$$V(X) = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^2 - \bar{x} = \frac{15\ 370}{1\,000} - 3,67^2 = 1,9011$$

Écart-type de X : $\sigma(X) = \sqrt{1,9011} = 1,38$

L'écart moyen à la moyenne, c'est-à-dire l'écart moyen entre le revenu mensuel d'un ménage et le revenu mensuel moyen, est égal à 1 380 euros. C'est une dispersion autour de la moyenne qui n'est pas très forte, étant donné que l'étendue est égale à 6 000 euros. La moyenne est donc ici un indicateur plutôt pertinent pour résumer la tendance centrale.

Les caractéristiques de concentration

CHAPITRE 5

L'étude de la concentration d'une distribution permet de mesurer les inégalités de répartition. Elle s'applique particulièrement aux salaires, aux revenus, aux patrimoines, mais aussi aux chiffres d'affaires ou à l'emploi. Elle concerne principalement des variables quantitatives continues dont les valeurs sont toutes positives.

Les outils d'analyse de la concentration sont la médiale (à ne pas confondre avec la médiane !), la courbe de concentration ou courbe de Lorenz, et l'indice de Gini.

1 Introduction

Si, dans une distribution, la population est constituée de l'ensemble des employés d'une entreprise et la variable étudiée le salaire, l'analyse de la concentration de cette distribution va permettre de savoir si, dans cette entreprise, la masse salariale (c'est-à-dire le montant total des salaires) est répartie de manière faiblement ou fortement inégalitaire entre les employés.

Si tous les salariés perçoivent le même salaire et donc la même part de la masse salariale, la répartition est parfaitement égalitaire. On dit alors que la concentration est nulle.

Plus on s'éloigne de cette situation, plus une faible part des salariés (ceux dont les salaires sont les plus élevés) perçoit une forte part de la masse salariale, plus la masse salariale est concentrée entre un petit nombre d'individus. À la limite, si un seul salarié percevait la totalité de la masse salariale et les autres une rémunération nulle, l'inégalité serait à son maximum et la concentration également.

La masse salariale est le montant global des salaires ou « masse des valeurs globales » de la variable ; l'étude de la concentration est l'analyse de la répartition de cette masse entre les salariés.

De manière générale, pour évaluer les caractéristiques de concentration d'une distribution, il est nécessaire de calculer d'abord les valeurs globales et la masse de ces valeurs, pour ensuite s'intéresser à la manière dont cette masse se répartit entre les individus de la population.

2 Masse des valeurs globales et répartition de cette masse

La variable étudiée est quantitative. Elle peut être discrète ou continue, mais en pratique c'est essentiellement aux variables quantitatives continues que s'applique l'étude de la concentration. Nous nous limiterons donc à ces variables.

Soit X une variable statistique quantitative continue dont les valeurs sont toutes positives. Les extrémités de classe sont notées $[e_1 , e_2[$, $[e_2 ; e_3[$…, $[e_i ; e_{i+1}[$…, $[e_k ; e_{k+1}[$, les centres de classes x_1, x_2…, x_i…, x_k, et les effectifs correspondants sont n_1, n_2…, n_i…, n_k.

Classe	$[e_1 ; e_2[$	$[e_2 ; e_3[$	…	$[e_i ; e_{i+1}[$	…	$[e_k ; e_{k+1}[$
Centre de classe	x_1	x_2		x_i		x_k
Effectif	n_1	n_2	…	n_i	…	n_k

■ *Masse des valeurs globales*

La ***valeur globale*** associée au couple $(x_i ; n_i)$ est la somme des valeurs observées de la variable pour ces n_i individus. Il arrive que cette valeur globale soit donnée par le tableau des données, dans une ligne (ou colonne) supplémentaire. Si ce n'est pas le cas, la valeur exacte de la variable n'étant pas connue pour chaque individu, elle est supposée en moyenne égale au centre x_i de la classe et la valeur globale est évaluée par le produit $n_i x_i$.

La ***masse des valeurs globales*** est la somme des valeurs globales. Elle est calculée soit en additionnant les valeurs globales fournies par le tableau, soit en effectuant $\sum_{i=1}^{k} n_i x_i$.

Que représentent concrètement les valeurs globales et la masse des valeurs globales ? Le tableau suivant donne trois exemples pour mieux comprendre.

Population	Variable	Valeur globale associée au couple $(x_i ; n_i)$	Masse des valeurs globales
Un ensemble de ménages	Le revenu	Revenu global perçu par les n_i ménages	Revenu global perçu par l'ensemble de tous les ménages

Un ensemble d'entreprises	Le chiffre d'affaires	Chiffre d'affaires global réalisé par les n_i entreprises	Chiffre d'affaires global réalisé par l'ensemble de toutes les entreprises
Un ensemble d'entreprises	Le nombre de salariés	Nombre total de salariés employés par les n_i entreprises	Nombre total de salariés employés par l'ensemble de toutes les entreprises

■ *Valeur globale relative*

La ***valeur globale relative*** associée au couple $(x_i ; n_i)$ mesure la proportion que représente la valeur globale correspondant à la modalité x_i par rapport à la masse des valeurs globales. Cette proportion sera notée g_i. Si chaque valeur globale a été évaluée par le produit $n_i x_i$, la valeur globale relative est le rapport $\frac{n_i x_i}{\sum_{i=1}^{k} n_i x_i}$.

Le tableau ci-dessous précise la signification concrète des valeurs globales relatives dans les trois exemples précédents :

Population	Variable	Valeur globale relative associée au couple $(x_i ; n_i)$
Un ensemble de ménages	Le revenu	Part du revenu global perçue par les n_i ménages dont le revenu est compris entre e_i et e_{i+1}
Un ensemble d'entreprises	Le chiffre d'affaires	Part du chiffre d'affaires global réalisé par les n_i entreprises dont le chiffre d'affaires est compris entre e_i et e_{i+1}
Un ensemble d'entreprises	Le nombre de salariés	Part de l'ensemble des salariés employés par les n_i entreprises dont le nombre de salariés est compris entre e_i et e_{i+1}

Dans le cas où le tableau de distribution ne fournit pas les effectifs mais seulement les fréquences, il n'est pas possible d'évaluer la masse des valeurs globales, mais cela n'empêche pas de calculer les valeurs globales relatives. En effet :

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

■ *Valeur globale relative cumulée*

Chaque valeur globale et chaque valeur globale relative sont associées à une classe, comme le sont chaque effectif et fréquence. En revanche, une valeur globale relative cumulée est, comme une fréquence cumulée, associée à une extrémité de classe.

La valeur globale relative cumulée associée à l'extrémité de classe e_i indique la proportion de la masse totale des valeurs globales que se partagent les individus dont la valeur de la variable est strictement inférieure à e_i. Elle est la somme des valeurs globales relatives associées à la classe $[e_{i-1} ; e_i[$ et aux classes précédentes. Elle sera notée G_i.

Par définition : $G_1 = 0$, $G_2 = g_1 \ldots$, $G_i = g_1 + g_2 + \ldots + g_{i-1} \ldots$, $G_{k+1} = 1$.

G_i indique la proportion de la masse totale des valeurs globales que se partagent les individus dont la valeur de la variable est inférieure à e_i.

La signification concrète des valeurs globales relatives cumulées dans les trois exemples précédents est précisée ci-dessous :

Population	Variable	Valeur globale relative cumulée associée à e_i
Un ensemble de ménages	Le revenu	Part du revenu global perçu par les ménages dont le revenu est inférieur à e_i
Un ensemble d'entreprises	Le chiffre d'affaires	Part du chiffre d'affaires global réalisé par les entreprises dont le chiffre d'affaires est inférieur à e_i
Un ensemble d'entreprises	Le nombre de salariés	Part de l'ensemble des salariés employés par les entreprises dont le nombre de salariés est inférieur à e_i

■ *Applications*

a) Répartition de la masse salariale d'une entreprise

Les salariés d'une entreprise sont répartis comme suit en fonction de leur salaire brut mensuel (en milliers d'euros), en France, en 2008 :

Salaires (10^3 euros)	[3 ; 4[	[4 ; 5[	[5 ; 6[	[6 ; 7[	[7 ; 8[
Effectifs	26	33	64	7	10

La population est l'ensemble des salariés de l'entreprise. La variable est le salaire brut mensuel ; c'est une variable quantitative continue dont les valeurs sont comprises entre 3 000 et 8 000 euros.

Les valeurs globales $n_i x_i$, les valeurs globales relatives g_i et les valeurs globales relatives cumulées G_i figurent dans le tableau ci-dessous. Chaque valeur globale et valeur globale relative est sur la même ligne que le centre de classe et l'effectif qui lui sont associés. Chaque valeur globale relative cumulée est sur la même ligne que l'extrémité de classe correspondante.

e_i	x_i	n_i	$n_i x_i$	g_i	G_i
3					0
	3,5	26	91	0,128	
4					0,128
	4,5	33	148,5	0,209	
5					0,337
	5,5	64	352	0,494	
6					0,831
	6,5	7	45,5	0,064	
7					0,895
	7,5	10	75	0,105	
8					1
Ensemble		140	712	1	

Les 26 salariés dont la rémunération est comprise entre 3 000 et 4 000 euros se partagent une masse salariale égale à 91 000 euros. La masse salariale globale étant égale à 712 000 euros, ils perçoivent une part de cette masse égale à 91/712, soit 0,128 ou 12,8 %.

Les 33 salariés dont la rémunération est comprise entre 4 000 et 5 000 euros se partagent une masse salariale égale à 148 500 euros. Ils perçoivent une part de la masse salariale globale égale à 148,5/712, soit 0,209 ou 20,9 %.

…

Les 10 salariés dont la rémunération est comprise entre 7 000 et 8 000 euros se partagent une masse salariale égale à 75 000 euros. Ils perçoivent une part de la masse salariale globale égale à 75/712, soit 0,105 ou 10,5 % de la masse salariale.

Les valeurs globales relatives cumulées G_i indiquent successivement la part des salariés dont le salaire est inférieur à e_i :

$e_1 = 3$, $G_1 = 0$: aucun salarié n'a une rémunération inférieure à 3 000 euros.

$e_2 = 4$, $G_2 = 0{,}128$: la part des salariés dont la rémunération est inférieure à 4 000 euros est 12,8 %.

...

$e_5 = 7$, $G_5 = 0{,}895$: la part des salariés dont la rémunération est inférieure à 7 000 euros est 89,5 %.

$e_6 = 8$, $G_6 = 1$: 100 % des salariés ont une rémunération inférieure à 8 000 euros.

b) Répartition de l'impôt de solidarité sur la fortune (ISF) en France, en 2005

Le tableau suivant indique la répartition des redevables de l'impôt de solidarité sur la fortune (ISF) selon le montant de l'impôt payé, en France, en 2005, ainsi que, pour chaque tranche, le montant total de l'impôt payé. La valeur du patrimoine à partir de laquelle l'ISF était perçu en 2005 était 732 000 euros et il existait six tranches de barème. Le nombre de contribuables à l'ISF s'est élevé à 394 500 (source : Commission des finances du Sénat, www.senat.fr).

Impôt (en euros)	Part des contribuables (%)	Montant global de l'impôt (millions d'euros)
[0 ; 2 464[	49,2	251,1
[2 464 ; 11 156,5[	38,9	908,3
[11 156,5 ; 24 376,5[	7,4	514,6
[24 376,5 ; 68 004,5[	3,3	542,5
[68 004,5 ; 203 436,5[	0,9	384,4
203 436,5 et plus	0,3	499,1
Ensemble	100,0	3 100,0

La population est l'ensemble des redevables de l'impôt de solidarité sur la fortune. La variable est l'impôt de solidarité sur la fortune. C'est une variable quantitative continue.

La masse des valeurs globales est le montant global de l'impôt payé par l'ensemble des redevables. Elle figure en bas de la dernière colonne du tableau ; elle est égale à 3,1 milliards d'euros. Elle est la somme des valeurs globales qui figurent dans cette colonne. Ce sont les valeurs globales exactes. Il n'est donc pas nécessaire de les estimer pour chaque classe en utilisant les centres de classe.

Les valeurs globales relatives g_i et les valeurs globales relatives cumulées G_i figurent dans le tableau ci-dessous.

e_i	f_i	Montant global de l'impôt	g_i	G_i
0				0,0 %
	49,2	251,1	8,1 %	
2 464				8,1 %
	38,9	908,3	29,3 %	
11 156,5				37,4 %
	7,4	514,6	16,6 %	
24 376,5				54,0 %
	3,3	542,5	17,5 %	
68 004,5				71,5 %
	0,9	384,4	12,4 %	
203 436,5				83,9 %
	0,3	499,1	16,1 %	
?				100,0 %
Ensemble	100	3 100,00	100,0 %	

Les redevables dont l'impôt est compris entre 0 et 2 464 euros paient ensemble un montant d'impôt égal à 251,1 millions d'euros, soit 8,1 % du montant total de l'ISF.

Les redevables dont l'impôt est compris entre 2 464 et 11 156,5 euros paient ensemble un montant d'impôt égal à 908,3 millions d'euros, soit 29,3 % du montant total de l'ISF.

…

Les redevables dont l'impôt est supérieur à 203 436,5 euros paient ensemble un montant d'impôt égal à 499,1 millions d'euros, soit 16,1 % du montant total de l'ISF.

Les valeurs globales relatives cumulées G_i indiquent successivement la part des redevables dont l'impôt est inférieur à e_i :

$e_1 = 0$, $G_1 = 0$: aucun redevable ne paie un impôt inférieur à 0 euros.

$e_2 = 2\,464$, $G_2 = 8{,}1\ \%$: la part des redevables dont l'impôt est inférieur à 2 464 euros est 8,1 %.

…

$e_6 = 203\,436{,}5$, $G_6 = 83{,}9\ \%$: la part des redevables dont l'impôt est inférieur à 203 436,5 euros est 83,9 %.

e_7 = ?, G_7 = 100 % : 100 % des redevables paient un impôt inférieur à un montant qui n'est pas donné par le tableau. C'est le montant payé par le contribuable dont l'impôt est le plus élevé.

3 Médiale

■ *Définition*

La médiale Ml d'une distribution statistique est la valeur de la variable qui partage la masse des valeurs globales en deux parties égales. C'est donc la valeur de la variable associée à la valeur globale relative cumulée 50 %.

Soit G la fonction qui à chaque valeur de la variable associe la valeur globale relative cumulée correspondante. Par définition : G(Ml) = 50 %.

La signification concrète de la médiale est moins simple à exprimer que celle de la médiane ; il faut veiller à ne pas faire de confusion entre les deux. La moitié de la masse des valeurs globales est répartie entre les individus dont la valeur de la variable est inférieure à la médiale, alors que la moitié des individus présentent une valeur de la variable inférieure à la médiane. Par exemple :

– si la variable est le revenu d'un ensemble de ménages :
  • la moitié de la masse globale des revenus est répartie entre les ménages dont le revenu est inférieur à la médiale,
  • tandis que la moitié des ménages ont un revenu inférieur à la médiane ;
– si la variable est le chiffre d'affaires d'un groupe d'entreprises :
  • la moitié du chiffre d'affaires global est réalisée par les entreprises dont le chiffre d'affaires est inférieur à la médiale,
  • tandis que la moitié des entreprises ont un chiffre d'affaires inférieur à la médiane ;
– si la variable est le nombre de salariés des entreprises d'un secteur :
  • la moitié des salariés de ce secteur est employée par les entreprises dont le nombre de salariés est inférieur à la médiale,
  • tandis que la moitié des entreprises a un nombre de salariés inférieur à la médiane.

■ *Détermination de la médiale*

S'il existe une extrémité de classe e_i telle que $G_i = G(e_i) = 50$ %, alors e_i est la médiale. Sinon, pour déterminer la médiale, on procède comme pour la médiane, par interpolation linéaire, à partir de la courbe représentative non pas des fréquences cumulées, mais des valeurs globales relatives cumulées.

Dans un repère portant en abscisse les extrémités de classes et en ordonnée les valeurs globales relatives cumulées, cette courbe (qui est en réalité une ligne polygonale) joint par des segments de droite les points de coordonnées $(e_i ; G_i)$. La médiale est l'abscisse du point de cette courbe dont l'ordonnée est 0,5 ou 50 %. Soient G(a) et G(b) les valeurs globales relatives cumulées qui encadrent 0,5. La fonction G étant une fonction strictement croissante, la médiale est comprise entre a et b.

Par interpolation linéaire, on obtient : $\frac{Ml - a}{b - a} = \frac{G(Ml) - G(a)}{G(b) - G(a)}$

donc : $Ml = a + [(b - a)\frac{0,5 - G(a)}{G(b) - G(a)}]$.

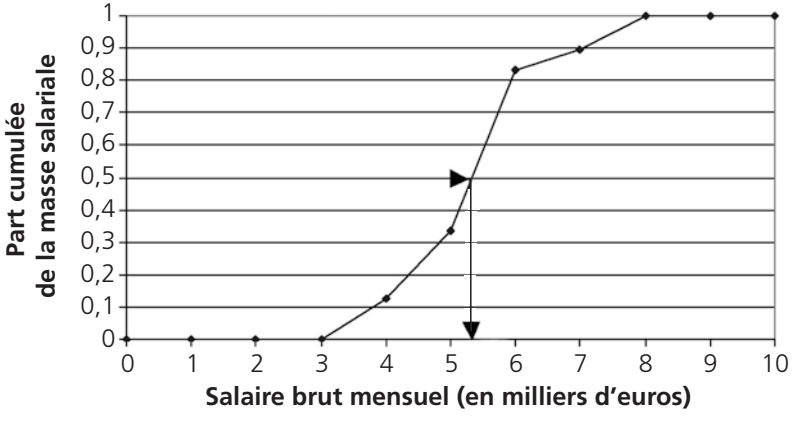
■ *Applications*

a) Médiale de la distribution des salaires dans une entreprise

Les couples $(e_i ; G_i)$ relatifs à la distribution des salaires sont repris dans le tableau ci-dessous :

e_i	3	4	5	6	7	8
Gi	0,000	0,128	0,337	0,831	0,895	1,000

La courbe représentative des valeurs globales relatives cumulées (ici parts cumulées de la masse salariale) permet une évaluation de la médiale.



La médiale est la valeur de la variable dont l'ordonnée sur ce graphe est 0,5. Elle est donc comprise entre 5 et 5,5 (milliers d'euros).

Pour une évaluation plus précise, il faut réaliser une interpolation linéaire.

La valeur globale relative cumulée 0,5 est comprise entre 0,337 et 0,895, c'est-à-dire G(5) et G(6). La fonction G étant une fonction strictement croissante, on en déduit que la médiale est comprise entre 5 et 6 :

$0{,}337 < 0{,}5 < 0{,}836 \Rightarrow G(5) < G(Ml) < G(6) \Rightarrow 5 < Ml < 6$

La médiale est alors calculée par interpolation linéaire :

$$\frac{Ml-5}{6-5} = \frac{0{,}5-0{,}337}{0{,}836-0{,}337} \text{ d'où : } Ml = 5+[(6-5)\frac{0{,}5-0{,}337}{0{,}836-0{,}337}] = 5{,}33.$$

La médiale est égale à 5,33 milliers d'euros : 50 % de la masse salariale est perçue par les salariés dont le salaire est inférieur à 5 330 euros, ou encore les salariés dont le salaire mensuel est inférieur à 5 330 euros se partagent la moitié de la masse salariale.

b) Médiale de la distribution de l'impôt de solidarité sur la fortune

Les couples $(e_i ; G_i)$ relatifs à la distribution de l'ISF figurent ci-dessous.

e_i	0	2 464	11 156,5	24 376,5	68 004,5	203 436,5	?
G_i	0,0 %	8,1 %	37,4 %	54,0 %	71,5 %	83,9 %	100,0 %

Il n'existe pas d'extrémité de classe e_i vérifiant $G(e_i) = 50$ %. La valeur globale relative cumulée 50 % est comprise entre 37,4 % et 54,0 %, c'est-à-dire G(11 156,5) et G(24 376,5). La médiale est donc comprise entre 11 156,5 et 24 376,5 :

$37{,}4\ \% < 50\ \% < 54\ \% \Rightarrow G(11\,156{,}5) < G(Ml) < G(24\,376{,}5) \Rightarrow 11\,156{,}5 < Ml < 24\,376{,}5.$

La médiale est alors calculée par interpolation linéaire :

$$Ml = 11\,156{,}5+[(24\,376{,}5-11\,156{,}5)\frac{50\ \%-37{,}4\ \%}{54\ \%-37{,}4\ \%}] = 11\,156{,}5+10\,034{,}5 = 21\,191{,}0.$$

La médiale est égale à 21 191 euros : la moitié du montant total de l'ISF est payée par les redevables dont l'impôt de solidarité sur la fortune est inférieur à 21 191 euros.

4 Écart médiale – médiane

■ *Définition*

Soit Δ_M l'écart entre la médiale et la médiane : $\Delta_M = Ml - Mé$

La médiale étant toujours supérieure ou égale à la médiane (on peut le démontrer mathématiquement), cet écart est toujours positif.

La médiale est d'autant plus proche de la médiane que la répartition des valeurs globales est proche d'une répartition égalitaire (ou équirépartition). En effet, intuitivement, dans ce cas, 10 % des individus se répartissent environ 10 % de la masse des valeurs globales, 20 % des individus se répartissent environ 20 % de la masse des valeurs globales, etc. et notamment 50 % des individus, ceux dont la valeur de la variable est inférieure à la médiane, se répartissent environ 50 % de la masse des valeurs globales. La valeur de la variable qui sépare les individus en deux groupes de même effectif (la médiane) est donc quasiment la même que celle qui partage la masse des valeurs globales en deux masses égales (la médiale).

Plus la répartition de la masse des valeurs globales devient inégalitaire, plus l'écart entre la médiale et la médiane augmente, et elle risque d'augmenter d'autant plus que la différence entre les valeurs extrêmes de la variable (ou étendue de la distribution) est elle-même grande. C'est pourquoi, pour avoir un indicateur de concentration facilement interprétable, il est nécessaire de rapporter l'écart médiale – médiane à l'étendue de la distribution :

$$\frac{\text{Médiale – Médiane}}{\text{Étendue}}$$

Plus la valeur de ce ratio est grande, plus un petit nombre d'individus concentre une part importante de la masse des valeurs globales : on dit que la concentration est forte.

■ *Applications*

a) Écart médiale – médiane de la distribution des salaires

Pour évaluer la médiane, les fréquences cumulées sont calculées ci-dessous.

e_i	3		4		5		6		7		8	
n_i		26		33		64		7		10		140
f_i		0,186		0,236		0,457		0,050		0,071		1,000
F_i	0,000		0,186		0,422		0,879		0,929		1,000	

Par définition F(Mé) = 0,5. Dans le tableau ci-dessus, il n'existe aucune extrémité de classe telle que $F(e_i) = 0,5$. La médiane doit donc être évaluée par interpolation linéaire.

On constate que 0,422 < 0,5 < 0,879, soit F(5) < F(Mé) < F(6). La médiane est donc comprise entre 5 et 6.

$$\text{Mé} = 5 + [(6-5)\frac{0,5-0,422}{0,879-0,422}] = 5 + 0,17 = 5,17\,.$$

La médiane est égale à 5,17 milliers d'euros : 50 % des salariés ont un salaire mensuel inférieur à 5 170 euros.

D'où : $\frac{\text{Médiale} - \text{Médiane}}{\text{Étendue}} = \frac{5,33-5,17}{8-3} = \frac{0,16}{5} = 0,032 = 3,2\,\%$.

L'écart médiale – médiane représente une très faible part de l'étendue : la concentration des salaires est faible, leur répartition est très peu inégalitaire.

b) Écart médiale – médiane de la distribution de l'impôt de solidarité sur la fortune

Le calcul des fréquences cumulées permet de déterminer la médiane.

ei	fi	Fi
0		0,0 %
	49,2 %	
2 464,0		49,2 %
	38,9 %	
11 156,5		88,1 %
	7,4 %	
24 376,5		95,5 %
	3,3 %	
68 004,5		98,8 %
	0,9 %	
203 436,5		99,7 %
	0,3 %	
?		100,0 %
Ensemble	100,0 %	

Dans le tableau ci-dessus, on constate que F(2 464) est 49,2 %, soit presque 50 %. Le médiane est donc légèrement supérieure à 2 464. Elle doit être évaluée par interpolation linéaire :

49,2 % < 50 % < 88,1 %, soit F(2 464) < F(Mé) < F(11 156,50) : la médiane est comprise entre 2 464 et 11 156,50.

$$\text{Mé} = 2\,464 + [(11\,156,5 - 2\,464)\,\frac{50\,\% - 49,2\,\%}{88,1\,\% - 49,2\,\%}] = 2\,464 + 179 = 2\,643.$$

La médiane est égale à 2 643 euros : 50 % des redevables de l'ISF paient un impôt inférieur à 2 643 euros.

La médiale étant égale à 21 191 euros, l'écart médiale – médiane est égal à :

$$\Delta_M = \text{Ml} - \text{Mé} = 21\,191 - 2\,643 = 18\,548.$$

La dernière extrémité de classe n'étant pas connue, il n'est pas possible de diviser cet écart par l'étendue. Néanmoins, la médiale étant nettement supérieure à la médiane, l'écart entre la médiale et la médiane est élevé. On peut donc en conclure que la répartition de l'ISF est loin d'être égalitaire.

C'est le propre d'un impôt par tranches d'être réparti de manière inégalitaire. Il augmente plus que proportionnellement à la base de calcul, ici le patrimoine ; il est dit progressif.

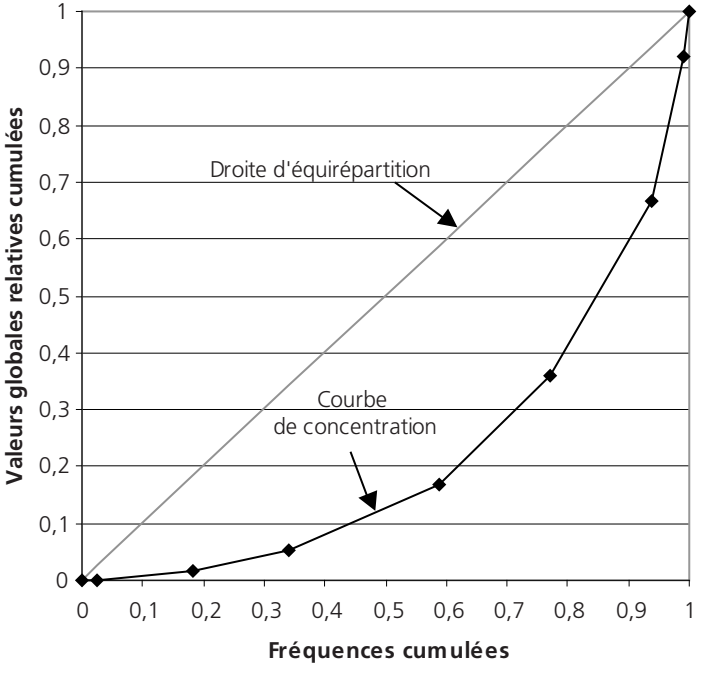
5 Courbe de concentration

■ *Définition*

La courbe de concentration (ou courbe de Lorenz) est un outil précieux de l'étude de la concentration. Elle permet d'apprécier rapidement, à travers un graphique, la plus ou moins grande concentration de la distribution. Elle sert aussi à définir et calculer un indicateur de la concentration, appelé indice de Gini, qui va préciser le message visuel donné par la courbe.

La courbe de concentration est obtenue en portant, dans un repère orthonormé, les fréquences cumulées sur l'axe des abscisses et les valeurs globales relatives cumulées sur l'axe des ordonnées. À chaque fréquence cumulée F_i est associée une valeur globale relative cumulée G_i qui détermine un point de la courbe de concentration. La courbe de concentration est obtenue en joignant ces points de coordonnées (F_i ; G_i) à main levée ou par des segments de droite.

Cette courbe est inscrite dans un carré dont chaque côté a pour longueur 1 (ou 100 %). Elle se situe sous la diagonale du carré qui joint les extrémités de la courbe car on a toujours $F_i \geq G_i$. La surface comprise entre la diagonale et la courbe de concentration s'appelle la surface de concentration.



Si la courbe de concentration est confondue avec la diagonale, cela signifie que, quel que soit x compris entre 0 et 100, x % des individus « se partagent » x % de la masse des valeurs globales : la concentration est nulle, la répartition est parfaitement égalitaire. Pour cette raison, la diagonale est souvent nommée droite d'équirépartition.

Plus la courbe de concentration s'éloigne de la droite d'équirépartition, plus on s'éloigne d'une répartition égalitaire et plus la concentration est forte. La concentration est à son maximum si la courbe est confondue avec les côtes du carré : un individu perçoit la totalité de la masse des valeurs globales et les autres rien.

■ *Applications*

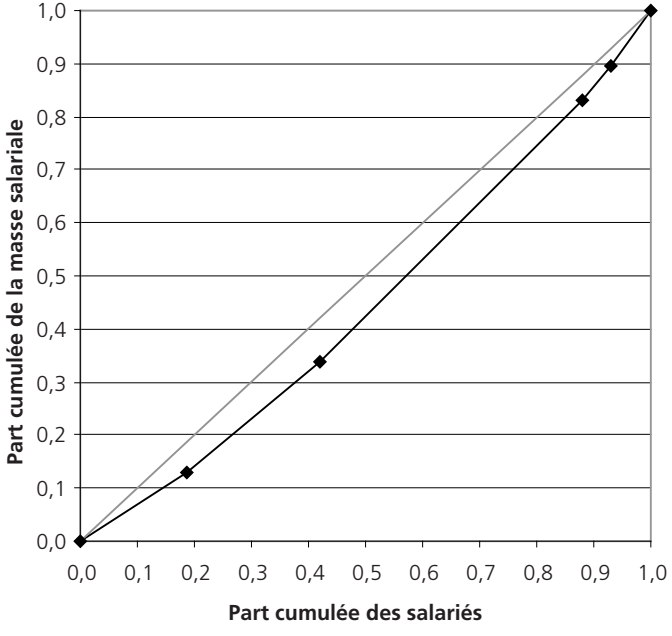
a) Courbe de concentration des salaires

Les coordonnées (F_i ; G_i) des points de la courbe de concentration des salaires ont été reportés dans le tableau suivant.

F_i	0,000	0,186	0,422	0,879	0,929	1,000
G_i	0,000	0,128	0,337	0,831	0,895	1,000

Ils nous indiquent que les 10,8 % des salariés qui perçoivent les salaires les plus faibles se partagent 12,8 % de la masse salariale, les 42,1 % des salariés qui perçoivent les salaires les plus faibles se partagent 33,6 % de la masse salariale, etc. ; 100 % des salariés se partagent 100 % de la masse salariale.

La courbe de concentration des salaires figure ci-dessous :



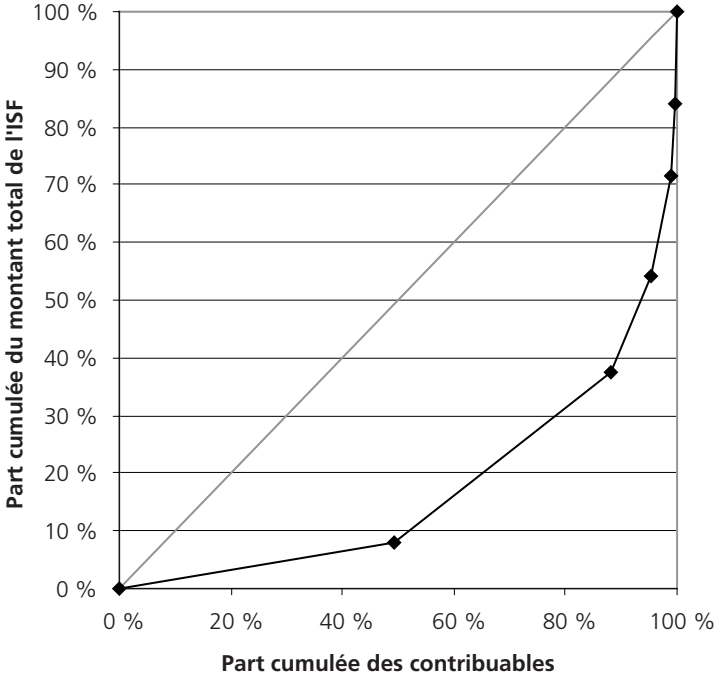
La courbe de concentration est proche de la droite d'équirépartition : la concentration est faible.

b) Courbe de concentration de l'impôt de solidarité sur la fortune

Les fréquences cumulées et valeurs globales relatives cumulées de la distribution de l'ISF, ci-dessous dans le tableau, permettent de tracer le courbe de concentration.

F_i	0,0 %	49,2 %	88,1 %	95,5 %	98,8 %	99,7 %	100,0 %
G_i	0,0 %	8,1 %	37,4 %	54,0 %	71,5 %	83,9 %	100,0 %

Les fréquences cumulées F_i (parts cumulées des contribuables) sont portées en abscisse, les valeurs globales relatives cumulées (parts cumulées du montant total de l'ISF) sont portées en ordonnée.



La courbe de concentration est nettement éloignée de la droite d'équirépartition : l'impôt de solidarité sur la fortune est assez fortement concentré. On peut illustrer cette concentration par les données : la moitié des redevables paient moins de 10 % de l'impôt et les 88 % qui en paient le moins paient environ 37 % du montant global, donc les 12 % qui paient le plus en paient 63 %.

6 Indice de concentration (ou indice de Gini)

■ *Définition*

Nous venons d'observer que la concentration est d'autant plus forte que la courbe de concentration est éloignée de la droite d'équirépartition. Elle est donc d'autant plus forte que l'aire de la surface de concentration est grande. Cette aire est au minimum nulle et au maximum égale à 0,5. En multipliant par 2 cette aire, on obtient l'indice de concentration. C'est un nombre sans dimension, compris entre 0 et 1.

Cet indice porte aussi le nom d'indice de Gini, du nom de son auteur, Corrado Gini, un statisticien italien qui l'a défini au cours de ses travaux sur les disparités de revenus, en 1912.

■ *Calcul de l'indice de Gini*

Il existe deux méthodes de calcul de l'indice de Gini : la méthode des triangles et la méthode des trapèzes.

a) Méthode des triangles

Cette méthode repose sur le découpage de la surface de concentration en un ensemble de k triangles contigus. La formule qui résulte de ce calcul est la suivante :

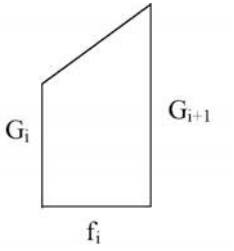
$$i_G = \sum_{i=1}^{k} (F_i G_{i+1} - F_{i+1} G_i) \text{ ou } i_G = \sum_{i=1}^{k} F_i G_{i+1} - \sum_{i=1}^{k} F_{i+1} G_i$$

b) Méthode des trapèzes

Cette méthode détermine l'aire de la surface de concentration par différence entre l'aire du triangle dans lequel s'inscrit la courbe de concentration et l'aire située sous la surface de concentration.

L'aire du triangle dans lequel s'inscrit la courbe de concentration est égale à la moitié de l'aire du carré, donc 0,5. Comme la courbe de concentration est une ligne polygonale, la surface située sous cette ligne peut être découpée en un ensemble de k trapèzes rectangles contigus dont l'aire est aisée à déterminer.

Chaque polygone est ainsi constitué :



Sa surface est donc égale à : $S = f_i(\frac{G_i + G_{i+1}}{2})$.

D'où : $i_G = 2[0,5 - \sum_{i=1}^{k} f_i(\frac{G_i + G_{i+1}}{2})] = 1 - \sum_{i=1}^{k} f_i(G_i + G_{i+1})$

$$i_G = 1 - \sum_{i=1}^{k} f_i(G_i + G_{i+1})$$

■ *Applications*

a) Indice de Gini de la distribution des salaires dans une entreprise

Calcul par la méthode des triangles : $i_G = \sum_{i=1}^{k}(F_iG_{i+1} - F_{i+1}G_i) = \sum_{i=1}^{k} F_iG_{i+1} - \sum_{i=1}^{k} F_{i+1}G_i$.

e_i	F_i	G_i	F_iG_{i+1}	$F_{i+1}G_i$
3	0,000	0,000		
4	0,186	0,128	0,000	0,000
5	0,422	0,337	0,063	0,054
6	0,879	0,831	0,351	0,296
7	0,929	0,895	0,787	0,772
8	1,000	1,000	0,929	0,895
			2,130	2,017

$$\sum_{i=1}^{k} F_i G_{i+1} = (0)(0,128) + (0,186)(0,337) + \ldots + (0,929)(1,000) = 2,130.$$

$$\sum_{i=1}^{k} F_{i+1} G_i = (0,186)(0) + (0,422)(0,128) + \ldots + (1)(0,895) = 2,017.$$

$i_G = 2,13 - 2,02 = 0,11.$

La valeur de l'indice de concentration proche de 0 confirme la faible concentration des salaires dans l'entreprise.

Calcul par la méthode des trapèzes : $i_G = 1 - \sum_{i=1}^{k} f_i (G_i + G_{i+1})$.

e_i	f_i	G_i	$f_i(G_i + G_{i+1})$
3		0,000	
	0,186		0,024
4		0,128	
	0,236		0,110
5		0,337	
	0,457		0,534
6		0,831	
	0,050		0,086
7		0,895	
	0,071		0,135
8		1,000	
			0,889

$$\sum_{i=1}^{k} f_i (G_i + G_{i+1}) = 0,186(0 + 0,128) + 0,236(0,128 + 0,337) + \ldots + 0,071(0,895 + 1).$$

$$\sum_{i=1}^{k} f_i (G_i + G_{i+1}) = 0,024 + 0,110 + \ldots + 0,135 = 0,889.$$

$$i_G = 1 - \sum_{i=1}^{k} f_i (G_i + G_{i+1}) = 1 - 0,89 = 0,11.$$

b) Indice de Gini de la distribution de l'impôt de solidarité sur la fortune

Calcul par la méthode des triangles :

$$i_G = \sum_{i=1}^{k} (F_i G_{i+1} - F_{i+1} G_i) = \sum_{i=1}^{k} F_i G_{i+1} - \sum_{i=1}^{k} F_{i+1} G_i = 3{,}17 - 2{,}51 = 0{,}66.$$

F_i	G_i	$F_i G_{i+1}$	$F_{i+1} G_i$
0,000	0,000		
0,492	0,081	0,000	0 ,000
0,881	0,374	0,184	0,071
0,955	0,540	0,476	0,357
0,988	0,715	0,683	0,534
0,997	0,839	0,829	0,713
1,000	1,000	0,997	0,839
		3,169	2,514

Calcul par la méthode des trapèzes : $i_G = 1 - \sum_{i=1}^{k} f_i (G_i + G_{i+1}) = 1 - 0{,}35 = 0{,}65.$

f_i	G_i	$f_i(G_i + G_{i+1})$
	0,000	
0,492		0,040
	0,081	
0,389		0,177
	0,374	
0,074		0,068
	0,540	
0,033		0,041
	0,715	
0,009		0,014
	0,839	
0,003		0,006
	1,000	
		0,345

La valeur de l'indice de Gini est beaucoup plus proche de 1 que de 0. Ce résultat confirme les résultats antérieurs : la concentration de l'impôt de solidarité sur la fortune est plutôt forte.

Les indices élémentaires

CHAPITRE 6

Les économistes, les sociologues et les gestionnaires étudient l'évolution dans le temps de grandeurs représentatives de phénomènes économiques et sociaux tels que les prix, la production, le chômage, les salaires. Ils comparent également ces grandeurs dans différents lieux ou pour différents groupes (entreprises, branches, départements, régions, pays,…). Pour évaluer ces évolutions ou effectuer ces comparaisons, ils utilisent principalement des indices.

Il existe deux catégories d'indices : les indices élémentaires et les indices synthétiques. Les indices élémentaires font l'objet de ce chapitre ; les indices synthétiques sont traités dans le chapitre suivant.

1 Introduction

Les indices sont des ***nombres sans unité*** qui traduisent les ***variations relatives*** de grandeurs dans le temps ou dans l'espace. Ces grandeurs peuvent être « simples » ou « complexes ».

Une grandeur est dite simple lorsque, dans chacune des situations dans lesquelles elle est observée, sa valeur est exprimée par un nombre unique. Le prix ou la quantité d'un produit sont des grandeurs simples, mais aussi la valeur d'un produit ou d'un ensemble de produits, tels que la dépense générée par l'achat d'un ou de plusieurs biens, le chiffre d'affaires ou les coûts d'une entreprise. Un nombre de personnes observées à un moment donné ou dans un lieu donné (bacheliers, personnes pauvres ou salariés par exemple), un nombre d'objets, une superficie d'un territoire sont aussi des grandeurs simples.

Un indice calculé sur une grandeur simple est appelé indice élémentaire.

Un indice calculé sur une grandeur complexe est appelé indice synthétique.

2 Variations absolues, variations relatives

Pour étudier l'évolution dans le temps de grandeurs simples ou comparer ces grandeurs dans des lieux différents (ou pour des groupes différents), deux types d'approches sont possibles : l'évaluation

de la variation absolue et l'évaluation de la variation relative. Ces évaluations sont complémentaires et donc nécessaires l'une et l'autre.

■ *Définitions*

Soit Z une grandeur économique simple observée dans plusieurs situations, une situation pouvant être une date précise, une période ou un lieu.

Soit z_t la valeur de Z à la situation t et z_0 la valeur de Z à la situation 0.

La variation absolue de Z entre la situation 0 et la situation t est égale à la différence entre z_t et z_0. Elle est notée ΔZ.

$$\Delta Z = z_t - z_0$$

Elle s'exprime dans la même unité que la grandeur Z.

La variation relative de Z entre la situation 0 et la situation t est couramment évaluée par le taux de variation de Z entre ces deux situations. Il est égal au rapport $\frac{z_t - z_0}{z_0}$ et noté $TV(Z)_{t/0}$.

$$TV(Z)_{t/0} = \frac{z_t - z_0}{z_0}$$

Un taux de variation est généralement exprimé en pourcentage.

■ *Application*

Le nombre de demandeurs d'emploi de catégorie 1, fin janvier 2007 et fin janvier 2008, dans trois régions françaises, Languedoc-Roussillon, Nord-Pas-de-Calais et Poitou-Charentes, est donné par le tableau suivant (données brutes, source : ministère du Travail, des Relations sociales et de la Solidarité, www.travail-solidarite.gouv.fr).

	Janvier 2007	Janvier 2008
Languedoc-Roussillon	109 173	104 935
Nord-Pas-de-Calais	190 803	177 320
Poitou-Charentes	56 475	52 480

La variation absolue du nombre de demandeurs d'emploi entre janvier 2007 et janvier 2008 est la différence entre le nombre de demandeurs d'emploi en janvier 2008, noté N_{08}, et le nombre de

demandeurs d'emploi en janvier 2007, noté N_{07}. L'unité dans laquelle cette variation est exprimée est « demandeur d'emploi ».

$$\text{Variation absolue} = N_{08} - N_{07}$$

Le taux de variation du nombre de demandeurs d'emploi $\frac{N_{08} - N_{07}}{N_{07}}$ évalue la variation relative du nombre de demandeurs d'emploi entre janvier 2007 et janvier 2008.

Les résultats de ces calculs sont les suivants pour chacune des trois régions :

	Variation absolue	Variation relative
Languedoc-Roussillon	– 4 238	– 3,9 %
Nord-Pas-de-Calais	– 13 483	– 7,1 %
Poitou-Charentes	– 3 995	– 7,1 %

Analysons ces variations : alors qu'en Languedoc-Roussillon et en Poitou-Charentes, la diminution du nombre de demandeurs d'emploi (variation absolue) est proche, un peu plus de 4 200 pour l'une et presque 4 000 pour l'autre, les baisses relatives (variations relatives) sont très différentes : – 3,9 % en Languedoc-Roussillon contre – 7,1 % en Poitou-Charentes. Proportionnellement, la performance du Poitou-Charentes est donc meilleure que celle du Languedoc-Roussillon.

Alors que la diminution du nombre de chômeurs en Nord-Pas-de-Calais est égale à environ 3,4 fois la diminution du nombre de chômeurs en Poitou-Charentes, la variation relative est identique (une baisse de 7,1 %) parce que, dans la première, le nombre de chômeurs en 2007 et 2008 est égal à environ 3,4 fois le nombre de chômeurs dans la seconde. Proportionnellement, la performance des deux régions est la même.

Deux variations absolues identiques peuvent donc correspondre à des variations relatives très éloignées l'une de l'autre ; inversement, deux variations relatives égales peuvent être associées à des variations absolues très différentes l'une de l'autre. Les deux formes de calcul sont donc bien nécessaires pour analyser la variation d'une grandeur.

Le calcul de la variation absolue d'une grandeur simple est unique.

L'évaluation de sa variation relative peut s'effectuer de trois manières différentes, en calculant un taux de variation, comme nous venons de le faire, mais aussi un coefficient multiplicateur ou un indice élémentaire. Ces trois indicateurs fournissent exactement la même information, sous des formes différentes. Cependant, les indices élémentaires possèdent des propriétés qui rendent leur utilisation attractive.

3 Définitions

■ *Indice élémentaire base 1, indice élémentaire base 100*

Soit Z une grandeur simple observée dans plusieurs situations, une situation pouvant être une date précise, une période ou un lieu.

Soit z_t la valeur de Z à la situation t et z_0 la valeur de Z à la situation 0.

L'*indice élémentaire de Z à la situation t, base 1 à la situation 0*, noté $i(Z)_{t/0}$, est le rapport de ces deux nombres z_t et z_0. Il est égal au coefficient multiplicateur, nommé aussi plus simplement multiplicateur.

$$i(Z)_{t/0} = \frac{z_t}{z_0}$$

L'*indice élémentaire de Z à la situation t, base 100 à la situation 0*, noté $I(Z)_{t/0}$, est le rapport des deux nombres z_t et z_0 multiplié par 100.

$$I(Z)_{t/0} = \frac{z_t}{z_0}.100$$

La situation t est appelée situation courante, la situation 0 situation de base ou situation de référence.

« Base 1 à la situation 0 » signifie que la valeur 1 de l'indice de la variable est associée à la situation 0. La valeur de l'indice de Z à la situation t, base 1 à la situation 0, traduit donc l'évolution de la valeur de Z par rapport à la valeur 1.

« Base 100 à la situation 0 » signifie que la valeur 100 de l'indice de la variable est associée à la situation 0. Donc, la valeur de l'indice de Z à la situation t, base 100 à la situation 0, traduit l'évolution de la valeur de Z par rapport à la valeur 100.

■ *Relation entre taux de variation et indice élémentaire*

Il existe une relation simple entre le taux de variation d'une grandeur et l'indice de cette grandeur, base 1 ou base 100.

$$TV(Z)_{t/0} = \frac{\text{Valeur courante} - \text{Valeur de référence}}{\text{Valeur de référence}} = \frac{z_t - z_0}{z_0} = \frac{z_t}{z_0} - 1 \text{ et } i(Z)_{t/0} = \frac{z_t}{z_0} \text{ donc :}$$

$$TV(Z)_{t/0} = i(Z)_{t/0} - 1 \text{ et } i(Z)_{t/0} = 1 + TV(Z)_{t/0}$$

Par définition, $I(Z)_{t/0} = i(z)_{t/0}.100$, donc :

$$TV(Z)_{t/0} = \frac{I(Z)_{t/0}}{100} - 1 \quad \text{et} \quad I(Z)_{t/0} = (1 + TV(Z)_{t/0}).100$$

Lorsque les situations 0 et t sont des dates, une augmentation de la grandeur Z se traduit par un indice à la date t, base 1 à la situation 0, supérieur à 1 et un indice à la date t, base 100 à la situation 0, supérieur à 100.

■ *Applications*

a) La situation est une date

Considérons le montant horaire du SMIC brut en euros courants le 1er juillet de chaque année de 2003 à 2007 (source : Insee, www.insee.fr) :

Année	2003	2004	2005	2006	2007
SMIC horaire brut	7,19	7,61	8,03	8,27	8,44

Calculons l'indice du SMIC chacune des années 2004 à 2007 :

– d'abord en prenant pour base 1 l'année 2003 (1re ligne du tableau $i(S)_{t/03}$) ;

– ensuite en prenant pour base 100 l'année qui précède l'année courante (2e ligne du tableau $I(S)_{t/t-1}$).

	2003	2004	2005	2006	2007
$i(S)_{t/03}$	1	$\frac{7,61}{7,19} = 1,0584$	$\frac{8,03}{7,19} = 1,1168$	$\frac{8,27}{7,19} = 1,1502$	$\frac{8,44}{7,19} = 1,1739$
$I(S)_{t/t-1}$	–	$\frac{7,61}{7,19}.100 = 105,84$	$\frac{8,03}{7,61}.100 = 105,52$	$\frac{8,27}{8,03}.100 = 102,99$	$\frac{8,44}{8,27}.100 = 102,06$

Ces indices sont calculés avec quatre chiffres derrière la virgule (à 10^{-4} près) pour les premiers et deux chiffres derrière la virgule (à 10^{-2} près) pour les suivants, car ils seront utilisés ensuite pour calculer d'autres indices ; les erreurs d'approximation sont ainsi limitées.

$i(S)_{t/03}$ traduit l'évolution du SMIC entre l'année 2003 et l'année t :

$i(S)_{04/03}$ = 1,0584 signifie que, si la valeur 1 est associée au SMIC en 2003, c'est la valeur 1,0584 qui doit être associée au SMIC en 2004. Le SMIC a donc augmenté de 0,0584 par rapport à 1, donc de 5,84 par rapport à 100, soit 5,84 %, entre 2003 et 2004.

$i(S)_{05/03} = 1{,}1168$ traduit donc une augmentation du SMIC de 11,68 % entre 2003 et 2005, $i(S)_{06/03} = 1{,}1502$ une augmentation de 15,02 % entre 2003 et 2006, et $i(S)_{07/03} = 1{,}1739$ une augmentation de 17,39 % de 2003 à 2007.

$I(S)_{t/t-1}$ traduit l'évolution du SMIC entre l'année t – 1 et l'année t : la valeur du SMIC en 2002 ne figurant pas dans le tableau des données, l'indice du SMIC en 2003, base 100 en 2002, ne peut être évalué.

$I(S)_{04/03} = 105{,}84$ signifie que, si la valeur 100 est associée au SMIC en 2003, c'est la valeur 105,84 qui doit être associée au SMIC en 2004. Le SMIC a donc augmenté de 5,84 par rapport à 100, donc de 5,84 %, entre 2003 et 2004.

$I(S)_{05/04} = 105{,}52$ traduit une hausse du SMIC de 5,52 % entre 2004 et 2005, $I(S)_{06/05} = 102{,}99$ une hausse de 2,99 % entre 2005 et 2006, et $I(S)_{07/06} = 102{,}06$ une augmentation de 2,06 % entre 2006 et 2007.

b) La situation est un lieu

Le tableau suivant indique la densité moyenne (nombre moyen d'habitants par km²) dans chacun des continents du monde, en 2005. La densité moyenne dans le monde entier est égale à 48 (source : Insee, www.insee.fr) :

Continent	Europe	Afrique	Amérique	Asie	Océanie
Densité moyenne	32	30	22	123	4

Calculons l'indice de densité moyenne I(D) de chacun des continents en prenant pour base 100 la densité moyenne dans le monde.

Continent	Europe	Afrique	Amérique	Asie	Océanie
I(D)	$\frac{32}{48}.100 = 66{,}7$	$\frac{30}{48}.100 = 62{,}5$	$\frac{22}{48}.100 = 45{,}8$	$\frac{123}{48}.100 = 256{,}2$	$\frac{4}{48}.100 = 8{,}3$

En Europe, la densité moyenne est inférieure à la densité moyenne mondiale de 33,3 % ; en Afrique, elle lui est inférieure de 37,5 %, en Amérique, de 54,2 %, en Océanie de 91,7 %. En revanche, en Asie, la densité moyenne représente plus de 2,5 fois la densité moyenne mondiale : elle lui est supérieure de 156,2 %.

4 Propriétés

Les indices élémentaires vérifient les propriétés de transitivité (ou circularité) et de réversibilité. Ces propriétés permettent, à partir de la connaissance de certains indices, d'en calculer d'autres.

■ *Transitivité*

Un indice élémentaire est transitif (ou circulaire) car, quelles que soient les situations r, s et t :

$$i(Z)_{t/s} \cdot i(Z)_{s/r} = i(Z)_{t/r} \text{ et } I(Z)_{t/s} \cdot I(Z)_{s/r} = I(Z)_{t/r} \cdot 100$$

En effet :

$$i(Z)_{t/s} \cdot i(Z)_{s/r} = \frac{Z_t}{Z_s} \cdot \frac{Z_s}{Z_r} = \frac{Z_t}{Z_r} = i(Z)_{t/r}$$

$$I(Z)_{t/s} \cdot I(Z)_{s/r} = \frac{Z_t}{Z_s} \cdot 100 \cdot \frac{Z_s}{Z_r} \cdot 100 = \left(\frac{Z_t}{Z_r} \cdot 100\right) \cdot 100 = I(Z)_{t/r} \cdot 100$$

Cette propriété peut également s'écrire :

$$i(Z)_{t/s} = \frac{i(Z)_{t/r}}{i(Z)_{s/r}} \text{ et } I(Z)_{t/s} = \frac{I(Z)_{t/r}}{I(Z)_{s/r}} \cdot 100$$

Elle permet alors d'effectuer des changements de base : à partir de deux indices de Z exprimés dans la même situation de base, ici la situation r, on obtient un troisième indice exprimé dans une autre situation de base, ici la situation s. C'est le grand intérêt de la propriété de transitivité : elle permet de déterminer $i(Z)_{t/s}$, ou $I(Z)_{t/s}$, sans connaître les valeurs de la grandeur Z aux situations t et s.

La propriété de transitivité permet aussi de réaliser l'enchaînement d'indices, fréquemment utilisé lorsque les situations sont des périodes ou des dates.

Lorsque les indices sont en base 1 :

$$i(Z)_{t/0} = i(Z)_{t/t-1} \cdot i(Z)_{t-1/t-2} \cdot i(Z)_{t-2/t-3} \cdot \ \ldots \ \cdot i(Z)_{2/1} \cdot i(Z)_{1/0}$$

Lorsque les indices sont en base 100 :

$$I(Z)_{t/0} = I(Z)_{t/t-1} \cdot I(Z)_{t-1/t-2} \cdot I(Z)_{t-2/t-3} \cdot \ \ldots \ \cdot I(Z)_{2/1} \cdot I(Z)_{1/0} \cdot 100^{-(t-1)}$$

L'inconvénient de cette dernière formule est qu'elle oblige à multiplier le produit des indices par $100^{-(t-1)}$ pour obtenir un indice final en base 100. On peut éviter cette complication en évaluant $i(Z)_{t/0}$ en utilisant la formule précédente ; on multiplie ensuite $i(Z)_{t/0}$ par 100 pour obtenir $I(Z)_{t/0}$.

■ *Réversibilité*

Un indice élémentaire est réversible car, quelles que soient les situations t et 0, il vérifie :

$$i(Z)_{t/0} \cdot i(Z)_{0/t} = 1 \quad \text{et} \quad I(Z)_{t/0} \cdot I(Z)_{0/t} = 100^2$$

En effet :

$$i(Z)_{t/0} \cdot i(Z)_{0/t} = \frac{z_t}{z_0} \cdot \frac{z_0}{z_t} = 1$$

$$I(Z)_{t/0} \cdot I(Z)_{0/t} = \frac{z_t}{z_0} \cdot 100 \cdot \frac{z_0}{z_t} \cdot 100 = 100 \cdot 100 = 100^2$$

Cette propriété permet également d'effectuer des changements de base : de la connaissance d'un indice de Z à la situation t, base 100 à la situation 0, on déduit l'indice de Z en 0, base 100 à la situation t, ou inversement.

$$i(Z)_{0/t} = \frac{1}{i(Z)_{t/0}} \text{ et } I(Z)_{0/t} = \frac{100^2}{I(Z)_{t/0}}$$

■ *Applications*

a) Transitivité

Les indices du SMIC, base 100 l'année précédente, sont les suivants de 2004 à 2007 :

	2004	2005	2006	2007
$I(S)_{t/t-1}$	105,84	105,52	102,99	102,06

La propriété de transitivité des indices permet de calculer l'indice du SMIC en 2007, base 100 en 2003. D'après cette propriété :

$$I(S)_{07/03} = I(S)_{07/06} \cdot I(S)_{06/05} \cdot I(S)_{05/04} \cdot I(S)_{04/03} \cdot 100^{-3}$$

$$I(S)_{07/03} = 102,06 \cdot 102,99 \cdot 105,52 \cdot 105,84 \cdot 100^{-3} = 117,39$$

En utilisant les indices base 1, on écrit :

$$i(S)_{07/03} = i(S)_{07/06} \cdot i(S)_{06/05} \cdot i(S)_{05/04} \cdot i(S)_{04/03} = 1,0206 \cdot 1,0299 \cdot 1,0552 \cdot 1,0584 = 1,1739$$

$$I(S)_{07/03} = i(S)_{07/03} \cdot 100 = 1,1739 \cdot 100 = 117,39$$

Le résultat est évidemment identique à celui obtenu à partir des valeurs respectives du SMIC en 2007 (8,44 €) et 2003 (7,19 €) : $\frac{8,44}{7,19} \cdot 100 = 117,39$.

b) Transitivité et changement de base

Les indices du SMIC, base 1 en 2003, pour les années 2004 à 2007, figurent dans le tableau suivant :

	2004	2005	2006	2007
$i(S)_{t/03}$	1,0584	1,1168	1,1502	1,1739

Calculons l'indice du SMIC en 2006, base 1 en 2004.

D'après la propriété de transitivité :

$$i(S)_{06/04} = \frac{i(S)_{06/03}}{i(S)_{04/03}} = \frac{(1,1502)}{(1,0584)} = 1,0867$$

De juillet 2004 à juillet 2006, le SMIC horaire a augmenté de 8,67 %

c) Propriété de réversibilité et « non-symétrie » des variations relatives

En utilisant la propriété de réversibilité des indices élémentaires et le calcul d'indice effectué en b), calculons l'indice du SMIC en 2004, base 100 en 2006.

$$I(S)_{04/06} = \frac{100^2}{I(S)_{06/04}} = \frac{100^2}{100.i(S)_{06/04}} = \frac{100^2}{108,67} = 92,02$$

Si la valeur 100 est associée au SMIC en 2006, c'est la valeur 92,02 qui doit être associée au SMIC en 2004. En 2004, le SMIC ne représentait que 92,02 % du SMIC de 2006. Il lui était inférieur de 7,98 par rapport à 100, donc de 8 % environ.

Le SMIC en 2004 était inférieur au SMIC en 2006 de 8 %, alors que le SMIC de 2006 était supérieur à celui de 2004 de 8,7 %. Autrement dit, la variation relative à appliquer au SMIC en 2006, si l'on avait souhaité qu'il revienne à sa valeur de 2004, était –8 % et non pas –8,7 %. Ce résultat illustre la « non-symétrie » des variations relatives.

$$TV(Z)_{t/0} = \frac{z_t - z_0}{z_0} \text{ et } TV(Z)_{0/t} = \frac{z_0 - z_t}{z_t}$$

Le numérateur des ces deux rapports est le même en valeur absolu ; en revanche, si z_0 est inférieur à z_t, le premier rapport est plus grand que le deuxième, et inversement. On dit que les variations relatives ne sont pas symétriques.

Si une grandeur a augmenté de 100 % entre 0 et t, c'est-à-dire a été multipliée par 2, cela signifie qu'en 0 sa valeur était inférieure de 50 % à sa valeur en t ; elle devra diminuer de 50 % (et non pas de 100 % car elle deviendrait nulle !) pour retrouver sa valeur initiale.

$$\frac{2z - z}{z} = 1 = 100\ \%\ \text{et}\ \frac{z - 2z}{2z} = -\frac{1}{2} = -0,5 = -50\ \%$$

Si elle a augmenté de 200 %, elle a été multipliée par 3. Sa valeur initiale était inférieure de 2/3 à sa valeur en t ; elle devra diminuer de 2/3, soit 66,7 % pour retrouver cette valeur.

$$\frac{3z - z}{z} = 2 = 200\ \%\ \text{et}\ \frac{z - 3z}{3z} = -\frac{2}{3} = -0,667 = -66,7\ \%$$

Les écarts entre les variations relatives « à la hausse » et « à la baisse » sont d'autant plus élevés que les variations relatives sont elles-mêmes élevées.

Les indices synthétiques

CHAPITRE 7

Un indice synthétique est un indice calculé sur une grandeur « complexe ». Une grandeur complexe est soit un ensemble de prix, soit un ensemble de quantités de deux ou plusieurs produits.

L'indice synthétique le plus connu est l'indice des prix à la consommation de l'Insee, mais l'Insee calcule de nombreux autres indices synthétiques, notamment dans les comptes nationaux.

1 Introduction

Les économistes et les gestionnaires s'intéressent principalement aux variations des prix, des quantités ou des valeurs monétaires globales. Les indices calculés sont donc essentiellement des indices de prix, des indices de quantités et des indices de valeurs monétaires.

Les situations d'observations sont presque toujours des périodes ; les indices seront désormais considérés à une date ou une période courante par rapport à une date ou une période de référence.

L'indice de valeur d'un ensemble de n produits est un indice élémentaire puisque la valeur des n produits est unique à chacune des dates où elle est évaluée.

Soient p_0^i le prix du bien i à la date 0 et p_t^i le prix du bien i à la date t.

Soient q_0^i la quantité du bien i à la date 0 et q_t^i la quantité du bien i à la date t.

La valeur du produit i à la date 0, $i \in \{1, 2\ldots, n\}$, est $p_0^i q_0^i$ et $\sum_{i=1}^{n} p_o^i . q_o^i$ la valeur globale des n produits à la date 0.

$$I(V)_{t/0} = \frac{\sum_{i=1}^{n} p_t^i . q_t^i}{\sum_{i=1}^{n} p_0^i . q_0^i} 100$$

Cet indice permet de connaître le taux de variation de la valeur de cet ensemble de n produits entre l'année de base (date 0) et l'année courante (date t). C'est une information intéressante mais insuffisante, car si, par exemple, la valeur globale en t ($\sum_{i=1}^{n} p_t^i q_t^i$) augmente, ce seul indice ne nous permet pas de savoir si la hausse est due à une hausse des prix et des quantités, ou une hausse des prix et une baisse (ou une stagnation) des quantités, ou l'inverse.

Pour savoir s'il y a eu, entre les dates 0 et t, une hausse ou une baisse des prix, il faut éliminer l'influence des quantités dans le calcul de l'indice, c'est-à-dire considérer les quantités comme fixes.

De même, pour savoir s'il y a eu entre les dates 0 et t une hausse ou une baisse des quantités, il faut éliminer l'influence des prix dans le calcul de l'indice, c'est-à-dire considérer les prix comme fixes.

C'est en procédant ainsi que sont construits les indices de prix et les indices de quantités de Laspeyres et de Paasche. Ces indices servent eux-mêmes aux calculs d'autres indices synthétiques : les indices de Fisher.

2 Définitions

■ *Indices de Laspeyres*

a) Indice de Laspeyres des prix

Dans l'indice de Laspeyres des prix, les quantités valorisées en 0 et en t sont les quantités à la date 0 :

$$L(p)_{t/0} = \frac{\sum_{i=1}^{n} p_t^i . q_0^i}{\sum_{i=1}^{n} p_0^i . q_0^i} 100$$

Cet indice traduit l'évolution de l'ensemble des prix des n produits de la date 0 à la date t, puisque, entre le numérateur et le dénominateur, seuls les prix diffèrent : ce sont les prix à la date 0 au dénominateur, les prix à la date t au numérateur. Pour les quantités de chaque produit, on « fait comme si » elles étaient restées, à la date t, les mêmes qu'à la date 0.

L'indice de Laspeyres des prix se définit également comme une moyenne des indices élémentaires des prix.

Soit c_0^i la part de la valeur du bien i à la date 0 dans la valeur de l'ensemble des n biens à la date 0, appelée coefficient budgétaire du bien i à la date 0 :

$$c_0^i = \frac{p_0^i.q_0^i}{\sum_{i=1}^{n} p_0^i.q_0^i}$$

On démontre que l'indice de Laspeyres des prix est la moyenne arithmétique des indices élémentaires des prix pondérés par les coefficients budgétaires de la période de base.

$$L(p)_{t/0} = \sum_{i=1}^{n} c_0^i.I(p_i)_{t/0}$$

En effet : $L(p)_{t/0} = \frac{\sum_{i=1}^{n} p_t^i.q_0^i}{\sum_{i=1}^{n} p_0^i.q_0^i}100 = \sum_{i=1}^{n} (\frac{p_0^i.q_0^i}{\sum_{i=1}^{n} p_0^i.q_0^i})(\frac{p_t^i}{p_0^i})100 = \sum_{i=1}^{n} c_0^i.I(p_i)_{t/0}$

Un coefficient budgétaire est un nombre compris entre 0 et 1. La somme des n coefficients budgétaires est nécessairement égal à 1. Ces coefficients jouent donc le rôle des fréquences f_i dans la formule de la moyenne arithmétique ; les indices élémentaires correspondent aux valeurs du caractère x_i.

Les coefficients de pondération sont qualifiés de « budgétaires » parce que, les biens étudiés étant souvent des biens de consommation, chaque coefficient indique alors la part de la dépense en un bien particulier dans le budget consacré à la dépense totale. Lorsque les biens sont les produits vendus par une entreprise, chaque coefficient « budgétaire » représente une part du chiffre d'affaires de l'entreprise dans le chiffre d'affaires total ; le qualificatif « budgétaire » est néanmoins conservé.

b) Indice de Laspeyres des quantités

Dans l'indice de Laspeyres des quantités, les prix qui valorisent les quantités en 0 et en t sont les prix à la date 0 :

$$L(q)_{t/0} = \frac{\sum_{i=1}^{n} p_0^i.q_t^i}{\sum_{i=1}^{n} p_0^i.q_0^i}100$$

Cet indice traduit l'évolution de l'ensemble des quantités des n produits de la date 0 à la date t puisque, entre le numérateur et le dénominateur, seuls les prix diffèrent. Pour les prix de chaque produit, on « fait comme si » ils étaient restés, à la date t, les mêmes qu'à la date 0.

L'indice de Laspeyres des quantités se définit également en tant que moyenne : c'est la moyenne arithmétique des indices élémentaires des quantités pondérés par les coefficients budgétaires de la période de base :

$$L(q)_{t/0} = \sum_{i=1}^{n} c_0^i . I(q_i)_{t/0}$$

En effet : $L(q)_{t/0} = \frac{\sum_{i=1}^{n} p_0^i . q_t^i}{\sum_{i=1}^{n} p_0^i . q_0^i} 100 = \sum_{i=1}^{n} \left(\frac{p_0^i . q_0^i}{\sum_{i=1}^{n} p_0^i . q_0^i}\right)\left(\frac{q_t^i}{q_0^i}\right)100 = \sum_{i=1}^{n} c_0^i . I(q_i)_{t/0}$

■ *Indices de Paasche*

a) Indice de Paasche des prix

Dans l'indice de Paasche des prix, les quantités valorisées en 0 et en t sont les quantités à la date t :

$$P(p)_{t/0} = \frac{\sum_{i=1}^{n} p_t^i . q_t^i}{\sum_{i=1}^{n} p_0^i . q_t^i} 100$$

Cet indice traduit l'évolution de l'ensemble des prix des n produits de la date 0 à la date t puisque, entre le numérateur et le dénominateur, seuls les prix diffèrent.

L'indice de Paasche des prix se définit également comme une moyenne des indices élémentaires des prix mais une moyenne harmonique et non pas arithmétique.

L'indice de Paasche des prix est la moyenne harmonique des indices élémentaires des prix pondérés par les coefficients budgétaires de la période courante.

En effet :

$$P(p)_{t/0} = \frac{1}{\sum_{i=1}^{n} \frac{c_t^i}{I(p_i)_{t/0}}}$$

$$P(p)_{t/0} = \frac{\sum_{i=1}^{n} p_t^i . q_t^i}{\sum_{i=1}^{n} p_0^i . q_t^i} 100 = \frac{1}{\frac{\sum_{i=1}^{n} p_t^i . q_t^i \frac{p_0^i}{p_t^i}}{\sum_{i=1}^{n} p_t^i . q_t^i}} 100 = \frac{1}{\sum_{i=1}^{n} \frac{p_t^i . q_t^i}{\sum_{i=1}^{n} p_t^i . q_t^i} \frac{p_0^i}{p_t^i} \frac{1}{100}} = \frac{1}{\sum_{i=1}^{n} \frac{c_t^i}{\frac{p_t^i}{p_0^i} 100}} = \frac{1}{\sum_{i=1}^{n} \frac{c_t^i}{I(p_i)_{t/0}}}$$

b) Indice de Paasche des quantités

Dans l'indice de Paasche des quantités, les prix qui valorisent les quantités en 0 et en t sont les prix à la date t :

$$P(q)_{t/0} = \frac{\sum_{i=1}^{n} p_t^i . q_t^i}{\sum_{i=1}^{n} p_t^i . q_0^i} 100$$

Cet indice traduit l'évolution de l'ensemble des quantités des n produits de la date 0 à la date t puisque, entre le numérateur et le dénominateur, seules les quantités diffèrent.

L'indice de Paasche des quantités est également une moyenne harmonique.

L'indice de Paasche des quantités est la moyenne harmonique des indices élémentaires des quantités pondérés par les coefficients budgétaires de la période courante.

$$P(q)_{t/0} = \frac{1}{\sum_{i=1}^{n} \frac{c_t^i}{I(q_i)_{t/0}}}$$

En effet :

$$P(q)_{t/0} = \frac{\sum_{i=1}^{n} p_t^i . q_t^i}{\sum_{i=1}^{n} p_t^i . q_0^i} 100 = \frac{1}{\frac{\sum_{i=1}^{n} p_t^i . q_t^i \frac{q_0^i}{q_t^i}}{\sum_{i=1}^{n} p_t^i . q_t^i}} 100 = \frac{1}{\sum_{i=1}^{n} \frac{p_t^i . q_t^i}{\sum_{i=1}^{n} p_t^i . q_t^i} \frac{q_0^i}{q_t^i} \frac{1}{100}} = \frac{1}{\sum_{i=1}^{n} \frac{c_t^i}{\frac{q_t^i}{q_0^i} 100}} = \frac{1}{\sum_{i=1}^{n} \frac{c_t^i}{I(q_i)_{t/0}}}$$

■ *Indices de Fisher*

L'indice de Fisher des prix est la moyenne géométrique des indices des prix de Laspeyres et de Paasche :

$$F(p)_{t/0} = \sqrt{L(p)_{t/0} . P(p)_{t/0}}$$

L'indice de Fisher des quantités est la moyenne géométrique des indices des quantités de Laspeyres et de Paasche :

$$F(q)_{t/0} = \sqrt{L(q)_{t/0} . P(q)_{t/0}}$$

En tant que moyenne de deux indices de Laspeyres et de Fisher, un indice de Fisher a nécessairement une valeur comprise entre celles de ces deux indices : c'est un « compromis » entre les deux.

■ *Application*

Pour trois produits A, B et C, les prix en euros par kg et les quantités consommées en kg ont été les suivants aux dates 0 et t :

Produit	Prix (date 0)	Quantité (date 0)	Prix (date t)	Quantité (date t)
A	10	5	15	4
B	8	3	9	4
C	10	2	9	2

1°) Pour chaque produit, calcul de l'indice élémentaire du prix à la date t, base 100 à la date 0, puis de l'indice élémentaire de quantité et l'indice élémentaire de dépense :

$$I(p_A)_{t/0} = \frac{15}{10}100 = 150 \quad I(p_B)_{t/0} = \frac{9}{8}100 = 112,5 \quad I(p_C)_{t/0} = \frac{9}{10}100 = 90$$

$$I(q_A)_{t/0} = \frac{4}{5}100 = 80 \quad I(q_B)_{t/0} = \frac{4}{3}100 = 133,33 \quad I(q_C)_{t/0} = \frac{2}{2}100 = 100$$

$$I(d_A)_{t/0} = \frac{60}{50}100 = 120 \quad I(d_B)_{t/0} = \frac{36}{24}100 = 150 \quad I(d_C)_{t/0} = \frac{18}{20}100 = 90$$

Bien A : le prix a augmenté de 50 %, la quantité a diminué de 20 %, donc la dépense a augmenté de 20 %.

Bien B : le prix a augmenté de 12,5 %, la quantité a augmenté de 33,33 %, donc la dépense a augmenté de 50 %.

Bien C : le prix a diminué de 10 %, la quantité est restée constante, donc la dépense a diminué de 10 %.

2°) Calcul du coefficient budgétaire de chaque produit en 0, puis en t :

$$c_0^A = \frac{50}{50+24+20} = \frac{50}{94} = 0,5319 \quad c_0^B = \frac{24}{94} = 0,2553 \quad c_0^C = \frac{20}{94} = 0,2128$$

$$c_t^A = \frac{60}{60+36+18} = \frac{60}{114} = 0,5263 \quad c_t^B = \frac{36}{114} = 0,3158 \quad c_t^C = \frac{18}{114} = 0,1579$$

3°) Calcul de l'indice de la dépense globale à la date t, base 100 à la date 0 :

$$I(d)_{t/0} = \frac{\sum_{i=1}^{n} p_t^i . q_t^i}{\sum_{i=1}^{n} p_0^i . q_0^i} 100 = \frac{114}{94} 100 = 121,27$$

La dépense globale a donc augmenté de 21,27 %.

4°) ***Calcul des indices des prix et des quantités de Laspeyres et de Paasche***, en utilisant les deux formules pour chacun des indices :

$$L(p)_{t/0} = \frac{\sum_{i=1}^{n} p_t^i . q_0^i}{\sum_{i=1}^{n} p_0^i . q_0^i} 100 = \frac{120}{94} 100 = 127,66 \text{ et } P(p)_{t/0} = \frac{\sum_{i=1}^{n} p_t^i . q_t^i}{\sum_{i=1}^{n} p_0^i . q_t^i} 100 = \frac{114}{92} 100 = 123,91$$

$$L(p)_{t/0} = \sum_{i=1}^{n} c_0^i I(p_i)_{t/0} = 0,5319.150 + 0,2553.112,5 + 0,2128.90 = 127,66$$

$$P(p)_{t/0} = \frac{1}{\sum_{i=1}^{n} \frac{c_t^i}{I(p_i)_{t/0}}} = \frac{1}{\frac{0,5263}{150} + \frac{0,3158}{112,5} + \frac{0,1579}{90}} = 123,91$$

D'après l'indice de Laspeyres, les prix des trois produits ont augmenté de 27,67 %, tandis que d'après l'indice de Paasche, ils ont augmenté de 23,91 %. Cette différence dans les résultats est due, d'une part, à l'évolution des coefficients budgétaires entre les dates 0 et t et, d'autre part, à l'utilisation pour les calculs de deux types de moyennes distinctes. Il n'y a aucun critère objectif qui permette de prétendre qu'un des résultats est « meilleur » que l'autre.

$$L(q)_{t/0} = \frac{\sum_{i=1}^{n} p_0^i . q_t^i}{\sum_{i=1}^{n} p_0^i . q_0^i} 100 = \frac{92}{94} 100 = 97,87 \text{ et } P(q)_{t/0} = \frac{\sum_{i=1}^{n} p_t^i . q_t^i}{\sum_{i=1}^{n} p_t^i . q_0^i} 100 = \frac{114}{120} 100 = 95$$

$$L(q)_{t/0} = \sum_{i=1}^{n} c_0^i I(q_i)_{t/0} = 0,5319.80 + 0,2553.133,33 + 0,2128.100 = 97,87$$

$$P(q)_{t/0} = \frac{1}{\sum_{i=1}^{n} \frac{c_t^i}{I(q_i)_{t/0}}} = \frac{1}{\frac{0,5263}{80} + \frac{0,3158}{133,33} + \frac{0,1579}{100}} = 95$$

D'après l'indice de Laspeyres, les quantités consommées des trois produits ont diminué de 2,13 %, tandis que d'après l'indice de Paasche, elles ont diminué de 5 %. Ces deux résultats sont sensiblement différents, et, comme pour les indices de prix, il n'est pas possible de considérer que l'un est plus satisfaisant que l'autre.

5°) ***Calcul de l'indice des quantités et l'indice des prix de Fisher***, à la date t, base 100 à la date 0 :

$$F(p)_{t/0} = \sqrt{L(p)_{t/0}.P(p)_{t/0}} = \sqrt{127,66.123,91} = 125,77$$

$$F(q)_{t/0} = \sqrt{L(q)_{t/0}.P(q)_{t/0}} = \sqrt{97,87.95} = 96,42$$

D'après ce calcul, globalement, les prix des trois produits ont augmenté de 25,77 % et les quantités consommées ont diminué de 3,58 %.

Ces calculs résultant de moyennes, ils donnent des valeurs intermédiaires entre les valeurs des indices de Laspeyres et de Paasche : 25,77 % est compris entre 27,67 % et 23,91 %, 3,58 % est compris entre 2,13 % et 5 %.

3 Décomposition volume-prix de l'indice de la valeur

■ *Méthode*

Le calcul des indices synthétiques des prix et des quantités permet de décomposer la variation de la valeur en une variation liée au prix et une variation liée aux quantités (ou volumes).

L'indice de la valeur est égal :

– au produit de l'indice des quantités de Paasche par l'indice des prix de Laspeyres, divisé par 100 :

$$I(V)_{t/0} = P(q)_{t/0}.L(p)_{t/0.}\frac{1}{100}$$

– au produit de l'indice des prix de Paasche par l'indice des quantités de Laspeyres, divisé par 100 :

$$I(V)_{t/0} = P(p)_{t/0}.L(q)_{t/0.}\frac{1}{100}$$

En effet :

$$I(V)_{t/0} = \frac{\sum_{i=1}^{n} p^i_t.q^i_t}{\sum_{i=1}^{n} p^i_0.q^i_0}100 = \frac{\sum_{i=1}^{n} p^i_t.q^i_t}{\sum_{i=1}^{n} p^i_t.q^i_0}100.\frac{\sum_{i=1}^{n} p^i_t.q^i_0}{\sum_{i=1}^{n} p^i_0.q^i_0}100.\frac{1}{100} = P(q)_{t/0}.L(p)_{t/0}.\frac{1}{100}$$

$$I(V)_{t/0} = \frac{\sum_{i=1}^{n} p^i_t.q^i_t}{\sum_{i=1}^{n} p^i_0.q^i_0}100 = \frac{\sum_{i=1}^{n} p^i_t.q^i_t}{\sum_{i=1}^{n} p^i_0.q^i_t}100.\frac{\sum_{i=1}^{n} p^i_0.q^i_t}{\sum_{i=1}^{n} p^i_0.q^i_0}100.\frac{1}{100} = P(p)_{t/0}.L(q)_{t/0}.\frac{1}{100}$$

L'indice de la valeur est aussi égal au produit de l'indice des prix de Fisher par l'indice des quantités de Fisher, divisé par 100 :

$$I(V)_{t/0} = F(p)_{t/0}F(q)_{t/0}\frac{1}{100}$$

En effet :

$$F(p)_{t/0}F(q)_{t/0} = \sqrt{P(p)_{t/0}.L(p)_{t/0}}.\sqrt{P(q)_{t/0}.L(q)_{t/0}} = \sqrt{(P(p)_{t/0}L(q)_{t/0}).(P(q)_{t/0}L(p)_{t/0})}$$

d'où $F(p)_{t/0}\,F(q)_{t/0} = \sqrt{(100.I(V)_{t/0}).(100.I(V)_{t/0})} = 100\,I(V)_{t/0}$

■ *Application*

Dans l'application précédente, l'indice de la dépense a été évalué à 121,27 : $I(d)_{t/0} = 121,27$.

Les indices de Laspeyres, de Paasche et de Fisher des prix et des quantités étaient égaux à :

$L(p)_{t/0} = 127,66$ $P(p)_{t/0} = 123,91$ $F(p)_{t/0} = 125,77$;

$L(q)_{t/0} = 97,87$ $P(q)_{t/0} = 95$ $F(q)_{t/0} = 96,42$.

L'indice de la valeur est égal au produit de l'indice des quantités de Paasche par l'indice des prix de Laspeyres divisé par 100 : $I(V)_{t/0} = P(q)_{t/0}.L(p)_{t/0}.\frac{1}{100} = (95).(127,66)\frac{1}{100} = 121,27$.

D'après ce calcul, la hausse de la dépense est due à une baisse des quantités consommées de 5 % et une hausse des prix des produits consommés de 27,66 %.

L'indice de la valeur est aussi égal au produit de l'indice des prix de Paasche par l'indice des quantités de Laspeyres divisé par 100 : $I(V)_{t/0} = P(p)_{t/0}.L(q)_{t/0.} \frac{1}{100} = (123,91).(97,87)\frac{1}{100} = 121,27.$

Selon ce calcul, la hausse de la dépense est due à une baisse des quantités consommées de 2,13 % et une hausse des prix des produits consommés de 23,91 %.

Enfin, l'indice de la valeur est aussi le produit des indices de Fisher des prix et des quantités, divisé par 100 : $I(V)_{t/0} = F(p)_{t/0}F(q)_{t/0}\frac{1}{100} = (125,77).(96,42)\frac{1}{100} = 121,27.$

La hausse de la dépense s'explique ici par une augmentation des prix de 25,77 % et une baisse des quantités de 3,58 %.

Est-ce qu'une de ces décompositions est « la bonne » et les autres sont « mauvaises » ? Non. Toutes les trois sont correctes. La dernière a l'avantage de fournir une solution intermédiaire entre les deux autres et, à ce titre, peut être préférée.

4 Propriétés

■ *Transitivité*

Les indices de Laspeyres, de Paasche et de Fisher ne possèdent pas la propriété de transitivité (ou circularité). Considérons, par exemple, l'indice de Laspeyres des prix.

$$L(p)_{2/1}L(p)_{1/0} = \frac{\sum_{i=1}^{n} p^i_2.q^i_1}{\sum_{i=1}^{n} p^i_1 q^i_1}100.\frac{\sum_{i=1}^{n} p^i_1.q^i_0}{\sum_{i=1}^{n} p^i_0.q^i_0}100 \text{ et } L(p)_{2/0} = \frac{\sum_{i=1}^{n} p^i_2.q^i_0}{\sum_{i=1}^{n} p^i_0.q^i_0}100$$

Le produit des deux indices $L(p)_{2/1}L(p)_{1/0}$ n'est pas égal à $L(p)_{2/0}$. Ce n'est donc pas un indice de Laspeyres. En revanche, c'est un « indice chaîne de Laspeyres » ou encore un « Laspeyres chaîné », noté $Lch(p)_{2/0}$: $Lch(p)_{2/0} = L(p)_{2/1}L(p)_{1/0}$.

Plus généralement :

$$Lch(p)_{t/0} = L(p)_{t/t-1}L(p)_{t-1/t-2}\ L(p)_{t-2/t-3} \ldots\ L(p)_{2/1}L(p)_{1/0}$$

Le chaînage s'applique aussi aux indices de Paasche et aux indices de Fisher ; un indice de Paasche chaîné est noté Pch, un indice de Fisher chaîné Fch.

L'indice des prix à la consommation, les indices des prix à la production dans l'industrie, l'indice du coût du travail dans l'industrie, la construction et le tertiaire, et les indices des comptes nationaux

annuels français calculés par l'Insee sont des exemples d'indices de Laspeyres chaînés. Cette technique a l'avantage, en faisant intervenir les dates intermédiaires entre la date de base et la date courante, de prendre en compte, à travers les coefficients budgétaires de chacune des sous périodes, le lien qui existe entre prix et quantités.

Les indices chaînés se prêtent également à la décomposition prix-volume :

$$I(V) = Lch(p).Pch(q) = Lch(q).Pch(p) = Fch(p).Fch(q)$$

Dans la décomposition de la valeur utilisée par l'Insee dans les comptes nationaux annuels français, l'indice des prix est un indice de Paasche chaîné, et l'indice de volumes un indice de Laspeyres chaîné. En revanche, dans les comptes nationaux américains, la décomposition prix-volume se fait à l'aide d'indices de Fisher chaînés.

■ *Réversibilité*

Les indices de Laspeyres et de Paasche ne possèdent pas la propriété de réversibilité. En revanche, par définition de ces indices, ils vérifient les égalités :

$$L_{t/0} = \frac{100^2}{P_{0/t}} \text{ et } P_{t/0} = \frac{100^2}{L_{0/t}}$$

Un indice de Fisher est réversible. En effet :

$$F_{t/0}F_{0/t} = \sqrt{L_{t/0}.P_{t/0}}\sqrt{L_{0/t}.P_{0/t}} = \sqrt{(L_{t/0}.P_{0/t}).(L_{0/t}.P_{t/0})} = \sqrt{100^2.100^2} = 100^2$$

D'où :

$$F_{t/0} = \frac{100^2}{F_{0/t}}$$

■ *Agrégation*

Les indices de Laspeyres et de Paasche possèdent la propriété d'agrégation : si un ensemble de produits est constitué de plusieurs groupes :

– l'indice de Laspeyres de l'ensemble est égal à l'indice de Laspeyres des indices de Laspeyres de chaque groupe ;

– l'indice de Paasche de l'ensemble est égal à l'indice de Paasche des indices de Paasche de chaque groupe.

Ces indices possèdent cette propriété parce qu'ils ont une structure de moyenne : moyenne arithmétique pour Laspeyres, moyenne harmonique pour Paasche.

Un indice de Fisher n'a pas une structure de moyenne ; il ne possède pas la propriété d'agrégation.

La propriété d'agrégation est très intéressante : si l'ensemble des produits appartient à une nomenclature emboîtée, elle permet, au lieu de calculer directement un indice synthétique à partir de l'ensemble des postes de la nomenclature, de le calculer à partir des indices intermédiaires. C'est ainsi que procède l'Insee pour calculer les indices synthétiques de prix ou de quantités d'un ensemble de biens, notamment l'indice des prix à la consommation (IPC).

Pour déterminer l'IPC, l'Insee construit, d'abord pour chaque agglomération dans laquelle des relevés sont effectués (96 agglomérations de plus de 2 000 habitants dispersées sur le territoire métropolitain et de toute taille ainsi que 10 agglomérations dans les DOM), et ensuite pour l'ensemble du territoire, des indices par type de produit, et ce pour 305 produits. Par exemple, l'Insee calcule un indice de « l'huile de tournesol » à Toulouse, à Paris, etc. et par des agrégations successives, obtient l'indice national pour « l'huile de tournesol ». Puis d'autres agrégations sont faites pour obtenir les différents indices selon les niveaux de la nomenclature des produits de consommation, du plus fin jusqu'à l'indice d'ensemble : « l'huile de tournesol » fait partie des « huiles et margarines », incluses dans les « huiles et graisses », qui sont elles-mêmes des « produits alimentaires » appartenant au premier poste du niveau 1 de la nomenclature « produits alimentaires et boissons alcoolisées ». Ces agrégats sont pondérés suivant la structure de la consommation de l'ensemble des ménages, qui est mise à jour chaque année. Par exemple, en 2007, le coefficient de pondération était 18/10 000 pour les « huiles et margarines », 36/10 000 pour les « huiles et graisses », 1 360/10 000 pour les « produits alimentaires » et 1 488/10 000 pour les « produits alimentaires et boissons alcoolisées ».

Le niveau 1 de la nomenclature des produits de consommation utilisée par l'Insee pour construire l'IPC comprend douze postes, appelés fonctions de consommation, dont les intitulés et les coefficients de pondération figurent ci-dessous :

Fonction de consommation	Coefficient
Ensemble	10 000
Produits alimentaires et boissons non alcoolisées	1 488
Boissons alcoolisées et tabac	338
Habillement et chaussures	511
Logement, eau, gaz, électricité et autres combustibles	1 345
Ameublement, équipement ménager et entretien courant de la maison	603
Santé	1 004
Transport	1 650
Communications	323
Loisirs et culture	925
Éducation	26
Hôtellerie, cafés, restauration	653
Autres bien et services	1 134

Pour plus d'informations sur cet indice, on peut utilement consulter la rubrique « Comprendre l'indice des prix à la consommation » sur le site Internet de l'Insee (www.insee.fr).

5 Conclusion : quel type d'indice synthétique choisir ?

C'est l'indice de Laspeyres qui est le plus utilisé pour deux raisons :
– il utilise les coefficients budgétaires à la date de base, donc il permet de calculer une série temporelle d'indices en utilisant les mêmes coefficients de pondération pour chaque indice ; il n'est pas nécessaire de connaître, et donc d'avoir déterminé les coefficients budgétaires pour les différentes périodes courantes ;
– il possède la propriété d'agrégation.

Un indice de Paasche a l'inconvénient d'utiliser des coefficients de pondération calculés à la période courante qu'il faut donc déterminer pour chacune de ces périodes. Cette détermination peut prendre du temps. En outre, l'interprétation d'une série d'indices de Paasche peut poser des problèmes : les pondérations étant différentes entre deux périodes, un indice synthétique peut augmenter alors que chaque indice élémentaire baisse. En pratique, les indices de Paasche servent principalement dans le cadre de la décomposition volume-prix (cf. ci-dessus) : par exemple, l'indice

des prix du PIB est un indice de Paasche qui, multiplié par l'indice de Laspeyres du volume, permet d'obtenir l'indice du PIB nominal, ou PIB en euros courants.

Pour calculer un indice de Fisher, il faut déterminer l'indice de Paasche correspondant ; on retrouve alors l'inconvénient évoqué précédemment. Par ailleurs, un indice de Fisher ne possède pas la propriété d'agrégation. Pour ces deux raisons, l'utilisation d'un indice de Fisher est très limitée. Son utilisation est cependant recommandée lorsque les coefficients de pondération (coefficients budgétaires) varient beaucoup entre la période de base et la période courante car, dans ce cas, l'indice de Paasche et l'indice de Laspeyres donnent des résultats nettement différents. L'indice de Fisher, moyenne géométrique de ces deux indices, est alors un bon compromis.

Les distributions statistiques à deux variables : les tableaux de contingence

CHAPITRE 8

Une distribution à deux variables s'intéresse à une population caractérisée simultanément par deux variables, par exemple l'âge et le niveau de diplôme préparé par un ensemble d'étudiants, ou le nombre de salariés et le secteur d'activité d'un ensemble d'entreprises.

La présentation des données s'effectue alors souvent dans des tableaux à double entrée nommés tableaux de contingence. Ce chapitre est consacré à l'analyse élémentaire des données contenues dans ces tableaux.

1 Tableaux de présentation des données

Soient deux variables X et Y définies sur une même population composée de n individus, chacune des variables pouvant être qualitative, quantitative discrète ou quantitative continue.

La distribution statistique relative au couple de variables (X, Y) est toujours présentée dans un tableau qui est soit un tableau élémentaire, soit un tableau de contingence ou tableau à double entrée.

■ *Tableau élémentaire*

Un tableau élémentaire précise pour chaque individu i de la population la modalité x_i de la variable X et la modalité y_i de la variable Y qui lui sont associées.

Individu i	Modalité de la variable X : x_i	Modalité de la variable Y : y_i
1	x_1	y_1
2	x_2	y_2
…	…	…
n – 1	x_{n-1}	y_{n-1}
n	x_n	y_n

La série statistique est donc constituée de l'ensemble des couples $(x_i, y_i)_{i=1, 2..., n.}$

Cette présentation des données est adoptée dans un document papier (ouvrage, revue, rapport…) lorsque le nombre d'individus est faible, et/ou lorsque pas ou peu d'individus présentent le même couple de modalités pour X et Y. Elle est couramment utilisée dans un fichier informatique quel que soit le nombre d'observations. L'analyse de ces tableaux fera l'objet du chapitre 9.

■ *Tableau de contingence*

a) Définition

Un tableau de contingence définit la distribution statistique relative à un couple de variables (X, Y) par la donnée des modalités de X, des modalités de Y et des effectifs correspondants à chaque couple de modalités. Cette distribution est appelée distribution conjointe.

Un tableau de contingence se présente généralement comme ci-dessous :

Modalités de Y : y_j / Modalités de X : x_i	y_1	y_2	…	y_p
x_1	n_{11}	n_{12}	…	n_{1p}
x_2	n_{21}	n_{22}	…	n_{2p}
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
x_k	n_{k1}	n_{k2}	…	n_{kp}

1) Variables et modalités

La ***variable X*** a k modalités qui figurent dans ***la marge de gauche du tableau***. À chaque modalité de X correspond une ligne du tableau. Par convention, ***l'indice de la ligne est i et les modalités sont x_i,*** i variant de 1 à k.

La ***variable Y*** a p modalités qui figurent dans la ***marge supérieure du tableau***. À chaque modalité de Y correspond une colonne du tableau. Par convention, ***l'indice de la colonne est j et les modalités sont y_j,*** j variant de 1 à p.

Lorsque la variable X (ou Y) est qualitative ou quantitative discrète, les modalités x_i (ou y_j) sont les valeurs observées de la variable. Dans le cas où la variable X est quantitative continue, la i^e classe des valeurs de X est représentée par son centre x_i ; de même, si la variable Y est quantitative continue, la j^e classe des valeurs de Y est représentée par son centre y_j.

2) Distribution conjointe du couple (X, Y), effectifs et fréquences conjoints

Le tableau, formé par les marges gauche et supérieure, les k lignes et p colonnes, donne la distribution du couple de variables, appelée ***distribution conjointe du couple (X, Y)***. C'est l'ensemble des (x_i, y_j, n_{ij})***, avec i = {1, 2..., k} et j = {1, 2..., p}.***

L'effectif de la modalité (x_i, y_j), noté n_{ij}, est le nombre d'individus qui présentent à la fois la modalité x_i de la variable X et la modalité y_j de la variable Y. Il est nommé soit « effectif », soit « effectif conjoint ».

L'***effectif total*** de la distribution conjointe est la somme des effectifs conjoints :

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

Dans le tableau de contingence, les effectifs peuvent être remplacés par les fréquences correspondantes. La fréquence du couple (x_i, y_j), notée f_{ij}, est le rapport de l'effectif n_{ij} à l'effectif total :

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

Elle mesure la part dans l'ensemble de la population des individus présentant à la fois la modalité x_i de la variable X et la modalité y_j de la variable Y. Elle est nommée soit « fréquence », soit « fréquence conjointe ».

La somme des fréquences conjointes est égale à 1 : $\sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij} = \sum_{j=1}^{p}\sum_{i=1}^{k} f_{ij} = 1$

b) Application

Au 1er janvier 2006, la population des départements d'outre-mer (DOM) se répartissait comme suit en fonction du département et de l'âge. Les effectifs sont exprimés en milliers (source : Insee, www.insee.fr) :

	Moins de 20 ans	De 20 à 59 ans	60 ans ou plus
Guadeloupe	141,2	236,5	69,3
Guyane	91,3	99,4	11,3
Martinique	116,9	211,1	71
La Réunion	277,1	424,7	83,2

La population est l'ensemble des habitants des DOM au 1er janvier 2006. Ces habitants sont caractérisés selon deux variables :

– l'une, qui sera notée X, est qualitative : c'est le département d'habitation. Ses modalités figurent dans la marge de gauche ;

– l'autre, qui sera notée Y, est quantitative continue : c'est l'âge. Ses modalités figurent dans la marge supérieure.

Dans les lignes et les colonnes du tableau figurent les effectifs conjoints. Au 1er janvier 2006, en Guadeloupe, 141 200 habitants avaient moins de 20 ans, 236 500 avaient entre 20 et 59 ans et 69 300 avaient 60 ans ou plus ; en Guyane, 91 300 habitants avaient moins de 20 ans, etc.

L'effectif total, en milliers, est égal à :

$$141{,}2 + 236{,}5 + \ldots + 83{,}2 = 1\,833$$

Il y avait 1 833 000 habitants dans l'ensemble des DOM au 1er janvier 2006.

Le tableau de contingence incluant les fréquences conjointes est obtenu en divisant chaque effectif conjoint par l'effectif total :

	Moins de 20 ans	De 20 à 59 ans	60 ans ou plus
Guadeloupe	7,7 %	12,9 %	3,8 %
Guyane	5,0 %	5,4 %	0,6 %
Martinique	6,4 %	11,5 %	3,9 %
La Réunion	15,1 %	23,2 %	4,5 %

On lit : 7,7 % des habitants des DOM avaient moins de 20 ans et habitaient en Guadeloupe, 12,9 % avaient de 20 à 59 ans et habitaient en Guadeloupe…, 4,5 % avaient 60 ans ou plus et habitaient à La Réunion.

2 Distributions marginales

De la distribution du couple (X, Y), on peut déduire la distribution relative à la seule variable X, appelée distribution marginale de X, et la distribution relative à la seule variable Y, appelée distribution marginale de Y.

■ *Distribution marginale de X*

La distribution marginale de X est l'ensemble des k couples $(x_i , n_{i.})_{i = 1, 2, ..., k}$, où x_i est la modalité n° i de la variable X et $n_{i.}$ (lire « n i point ») l'effectif correspondant. Cet effectif, appelé ***effectif marginal de la modalité x_i***, est le nombre d'individus dont la modalité de la variable X est x_i et dont la modalité de la variable Y est y_1 ou y_2 ou… ou y_p. Il est égal au total des effectifs de la ligne i :

$$n_{i.} = n_{i1} + n_{i2} + ... + n_{ip} = \sum_{j=1}^{p} n_{ij}$$

Le « point » mis à la place de j dans $n_{i.}$ traduit le fait que l'indice j a varié de 1 à p.

La somme des effectifs marginaux est égal à l'effectif total n, noté aussi n.. dans certains manuels pour matérialiser le fait que l'indice i a varié de 1 à k :

$$n_{1.} + n_{2.} + ... + n_{k.} = \sum_{i=1}^{k} n_{i.} = n_{..} = n$$

La distribution marginale de X peut être présentée sous forme de tableau :

Modalité de la variable X : x_j	Effectif marginal : $n_{.j}$
x_1	$n_{1.}$
x_2	$n_{2.}$
. . .	. . .
x_k	$n_{k.}$
Ensemble	n

La fréquence marginale de la modalité x_i est notée $f_{i.}$; elle est égale à $\frac{n_{i.}}{n}$:

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

C'est la proportion des individus dont la modalité de la variable X est x_i.

La somme des fréquences marginales $f_{i.}$ est égale à 1 (ou 100 %) : $\sum_{i=1}^{k} f_{i.} = 1$.

■ *Distribution marginale de Y*

La distribution marginale de Y est l'ensemble des p couples $(y_j , n_{.j})_{j = 1, 2..., p}$, où y_j est la modalité n° j de la variable Y et $n_{.j}$ (lire « n point j ») l'effectif correspondant. Cet effectif, appelé ***effectif marginal de la modalité y_j***, est le nombre d'individus dont la modalité de la variable Y est y_j et dont la modalité de la variable X est x_1 ou x_2 ou… ou x_k. Il est égal au total des effectifs de la colonne j :

$$n_{.j} = n_{1j} + n_{2j} + ... + n_{kj} = \sum_{i=1}^{k} n_{ij}$$

Le « point » mis à la place de i dans $n_{.j}$ traduit le fait que l'indice i a varié de 1 à k.

La somme des effectifs marginaux est égal à l'effectif total :

$$n_{.1} + n_{.2} + ... + n_{.p.} = \sum_{j=1}^{p} n_{.j} = n_{..} = n$$

La distribution marginale de Y peut être également présentée sous forme de tableau :

Modalité de la variable Y : y_j	Effectif marginal : $n_{.j}$
y_1	$n_{.1}$
y_2	$n_{.2}$
. . .	. . .
y_p	$n_{.p}$
Ensemble	n

La fréquence marginale de la modalité y_j notée $f_{.j}$ est égale à $\frac{n_{.j}}{n}$:

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

C'est la proportion des individus dont la modalité de la variable Y est y_j.

La somme des fréquences marginales $f_{.j}$ est égale à 1 (ou 100 %) : $\sum_{j=1}^{p} f_{.j} = 1$.

■ *Distributions marginales et tableau de contingence*

Les effectifs marginaux $n_{i.}$ figurent souvent dans une colonne supplémentaire du tableau de la distribution conjointe du couple (X, Y), et les effectifs marginaux $n_{.j}$ dans une ligne supplémentaire, comme ci-dessous :

Modalités de Y : y_j / Modalités de X : x_i	y_1	y_2	…	y_p	Ensemble
x_1	n_{11}	n_{12}	…	n_{1p}	$n_{1.}$
x_2	n_{21}	n_{22}	…	n_{2p}	$n_{2.}$
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
x_k	n_{k1}	n_{k2}	…	n_{kp}	$n_{k.}$
Ensemble	$n_{.1}$	$n_{.2}$	…	$n_{.p}$	n

La distribution marginale de X est alors donnée par la marge de gauche et la dernière colonne du tableau, la distribution marginale de Y par la marge supérieure et la dernière ligne.

Dans ce tableau, les effectifs conjoints et effectifs marginaux peuvent être remplacés par les fréquences conjointes et fréquences marginales.

■ *Application*

Reprenons la répartition de la population des départements d'outre-mer en fonction du département et de l'âge. La distribution marginale de X indique la répartition des habitants des DOM selon le département d'habitation. Elle est définie par les couples $(x_i, n_{i.})$ ou $(x_i, f_{i.})$.

Modalités de X	Effectif (marginal) : $n_{i.}$ (en milliers)	Fréquence (marginale) : $f_{i.}$
Guadeloupe	447	24,4 %
Guyane	202	11,0 %
Martinique	399	21,8 %
La Réunion	785	42,8 %
Ensemble	1 833	100,0 %

Il y avait 447 000 habitants en Guadeloupe au 1er janvier 2006, soit 24,4 % des habitants des DOM, 202 000 habitants en Guyane, soit 11,0 % des habitants des DOM, etc.

La distribution marginale de Y donne la répartition des habitants des DOM selon l'âge. Elle est définie par les couples $(y_j, n_{.j})$ ou $(y_j, f_{.j})$.

Modalités de Y	Effectif (marginal) : $n_{.j}$ (en milliers)	Fréquence (marginale) : $f_{.j}$
Moins de 20 ans	626,5	34,2 %
De 20 à 59 ans	971,7	53,0 %
60 ans ou plus	234,8	12,8 %
Ensemble	1 833	100,0 %

Dans les DOM, il y avait, au 1er janvier 2006, 626 500 habitants de moins de 20 ans, 971 700 habitants de 20 à 59 ans, et 234 800 habitants de 60 ans ou plus, qui représentaient respectivement 34,2 %, 53 % et 12,8 % de la population.

3 Moyennes et variances marginales

Chacune des distributions marginales est une distribution à une variable. Lorsque cette variable est quantitative, elle se prête donc à tous les calculs de caractéristiques de position, dispersion ou concentration présentés dans les chapitres 2 à 5, en particulier les calculs de moyennes, variances et écarts-types, caractéristiques les plus couramment utilisées.

■ *Moyennes marginales*

La moyenne marginale de X est notée $\bar{x}$. Elle correspond à la moyenne de la variable X prise seule (telle que définie dans le chapitre 3).

C'est une moyenne arithmétique pondérée :

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

En utilisant les fréquences marginales $f_{i.}$, elle s'écrit : $\bar{x} = \sum_{i=1}^{k} f_{i.} x_i$.

De même, la moyenne marginale de Y, notée $\bar{y}$, correspond à la moyenne de la variable Y prise seule.

C'est une moyenne arithmétique pondérée définie par :

$$\bar{y} = \frac{1}{n}\sum_{j=1}^{p} n_{.j} y_j$$

En utilisant les fréquences marginales $f_{.j}$, elle s'écrit : $\bar{y} = \sum_{j=1}^{p} f_{.j} y_j$.

■ *Variances et écarts-types marginaux*

La variance marginale de X, notée V(X), correspond à la variance de la variable X prise seule (telle que définie dans le chapitre 4) :

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

En développant et simplifiant cette formule, on obtient la formule de Koenig :

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

En introduisant les fréquences marginales, on obtient :

$$V(X) = \sum_{i=1}^{k} f_{i.} (x_i - \bar{x})^2 \text{ et } V(X) = \sum_{i=1}^{k} f_{i.} x_i^2 - \bar{x}^2$$

De même, la variance marginale de Y, notée V(Y), correspond à la variance de la variable Y prise seule. Elle est définie par :

$$V(Y) = \frac{1}{n}\sum_{j=1}^{p} n_{.j} (y_j - \bar{y})^2$$

En développant et simplifiant cette formule, on obtient la formule de Koenig :

$$V(Y) = \frac{1}{n}\sum_{j=1}^{p} n_{.j} y_j^2 - \overline{y}^2$$

En introduisant les fréquences marginales :

$$V(Y) = \sum_{j=1}^{p} f_{.j}(y_j - \overline{y})^2 \text{ et } V(Y) = \sum_{j=1}^{p} f_{.j} y_i^2 - \overline{y}^2$$

L'écart-type marginal de X ou de Y est la racine carrée de la variance marginale :

$$\sigma(X) = \sqrt{V(X)} \text{ et } \sigma(Y) = \sqrt{V(Y)}$$

■ *Application*

Reprenons la répartition de la population des départements d'outre-mer en fonction du département et de l'âge. La variable X étant qualitative, on ne peut effectuer aucun calcul de caractéristique statistique sur cette variable.

Pour calculer la moyenne et l'écart-type de Y, il faut d'abord évaluer les centres de classe. L'extrémité de la dernière classe n'est pas donnée. Nous allons la supposer égale à 100 pour éviter de surestimer l'âge moyen des plus de 60 ans.

Modalités de Y	y_j	$n_{.j}$	$n_{.j} y_j$	$n_{.j} y_j^2$
Moins de 20 ans	10	626,5	6 265	62 650
De 20 à 59 ans	40	971,7	38 868	1 554 720
60 ans ou plus	80	234,8	18 784	1 502 720
Ensemble		1 833,0	63 917	3 120 090

$$\overline{y} = \frac{1}{n}\sum_{j=1}^{p} n_{.j} y_j = \frac{1}{1\,833}(63\,917) = 34,9$$

$$V(Y) = \frac{1}{n}\sum_{j=1}^{p} n_{.j} y_j^2 - \overline{y}^2 = \frac{1}{1\,833}(3\,120\,090) - 34,87^2 = 486,26$$

$$\sigma(X) = \sqrt{486,26} = 22,05$$

L'âge moyen d'un habitant des DOM est 35 ans environ. L'écart moyen entre l'âge d'un de ces habitants et l'âge moyen est de 22 années.

4 Distributions conditionnelles, moyennes et variances conditionnelles

On appelle distribution conditionnelle de X selon $Y = y_j$ (ou liée par $Y = y_j$) la distribution de X concernant uniquement les individus présentant la modalité y_j de Y.

De même, on appelle distribution conditionnelle de Y selon X (ou liée par $X = x_i$) la distribution de Y concernant uniquement les individus présentant la modalité x_i de X.

■ *Distributions conditionnelles de X selon Y*

a) Définition

Les modalités de Y étant au nombre de p, la population peut être divisée en p sous-populations distinctes : les individus pour lesquels $Y = y_1$, ceux pour lesquels $Y = y_2$…, ceux pour lesquels $Y = y_p$. À chacune de ces sous-populations, on associe la distribution de ces individus selon les modalités de la variable X, dite distribution conditionnelle.

On obtient ainsi p distributions conditionnelles de X selon Y (ou sachant Y) :

– la distribution conditionnelle de X liée par $Y = y_1$;

– la distribution conditionnelle de X liée par $Y = y_2$;

– …

– la distribution conditionnelle de X liée par $Y = y_p$.

Chacune de ces distributions est définie par un ensemble de couples (x_i, n_{ij}), i variant de 1 à k et j étant fixé. Elle peut être présentée dans un tableau : dans la 1re colonne figurent les modalités de X, dans la 2e les effectifs figurant dans la je colonne du tableau de contingence.

Modalité de la variable X : x_i	Effectif conditionnel : n_{ij}
x_1	n_{1j}
x_2	n_{2j}
. . .	. . .
x_k	n_{kj}
Ensemble	$n_{.j}$

L'effectif total est égal à $n_{.j}$:

$$n_{1j} + n_{2j} + \ldots + n_{kj} = \sum_{i=1}^{k} n_{ij} = n_{.j}$$

Ce tableau peut être complété par le calcul des fréquences conditionnelles f_{x_i/y_j} définies par :

$$f_{x_i/y_j} = \frac{n_{ij}}{n_{.j}}$$

f_{x_i/y_j} indique la part de la sous-population considérée (c'est-à-dire les individus pour lesquels $Y = y_j$) qui présente la modalité x_i de la variable X.

La somme des fréquences conditionnelles f_{x_i/y_j} est égale à 1 (ou 100 %) :

$$\sum_{i=1}^{k} f_{x_i/y_j} = \sum_{i=1}^{k} \frac{n_{ij}}{n_{.j}} = \frac{1}{n_{.j}} \sum_{i=1}^{k} n_{ij} = \frac{1}{n_{.j}} n_{.j} = 1$$

b) Tableau des profils-colonnes

Les distributions conditionnelles de X selon Y sont souvent regroupées dans un unique tableau contenant les modalités de X dans la marge de gauche et dans la 1re colonne, les fréquences conditionnelles de X selon $Y = y_1$, dans la 2e les fréquences conditionnelles de X selon $Y = y_2$…, dans la pe les fréquences conditionnelles de X selon $Y = y_p$. Ce tableau, souvent complété par une dernière colonne donnant les fréquences marginales de X, s'appelle le tableau des profils-colonnes.

Modalités de X : x_i	$Y = y_1$	$Y = y_2$	…	$Y = y_p$	Ensemble
x_1	f_{x_1/y_1}	f_{x_1/y_2}		f_{x_1/y_p}	$f_{1.}$
x_2	f_{x_2/y_1}	f_{x_2/y_2}		f_{x_2/y_p}	$f_{2.}$
. . .	. . .	. . .	. . .	. . .	. . .
x_k	f_{x_k/y_1}	f_{x_k/y_2}		f_{x_k/y_p}	$f_{k.}$
Ensemble	1	1	…	1	1

■ *Distributions conditionnelles de Y selon X*

a) Définition

Les modalités de X étant au nombre de k, la population peut être divisée en k sous-populations distinctes : les individus pour lesquels $X = x_1$, ceux pour lesquels $X = x_2$…, ceux pour lesquels $X = x_k$. À chacune de ces sous-populations, on associe la distribution de ces individus selon les modalités de la variable Y, dite distribution conditionnelle.

On obtient ainsi k distributions conditionnelles de Y selon X (ou sachant X) :

– la distribution conditionnelle de Y liée par $X = x_1$;

– la distribution conditionnelle de Y liée par $X = x_2$;

– …

– la distribution conditionnelle de Y liée par $X = x_k$.

Chacune de ces distributions est définie par un ensemble de couples (y_j, n_{ij}), j variant de 1 à p et i étant fixé. Elle peut être présentée sous forme de tableau : dans la 1re colonne figurent les modalités de Y, dans la 2e les effectifs figurant dans la ie ligne du tableau de contingence.

Modalité de la variable Y : y_j	Effectif conditionnel : n_{ij}
y_1	n_{i1}
y_2	n_{i2}
. . .	. . .
y_p	n_{ip}
Ensemble	$n_{i.}$

L'effectif total est égal à $n_{i.}$:

$$n_{i1} + n_{i2} + \ldots + n_{ip} = \sum_{j=1}^{p} n_{ij} = n_{i.}$$

Ce tableau peut être complété par le calcul des fréquences conditionnelles f_{y_j/x_i} définies par :

$$f_{y_j/x_i} = \frac{n_{ij}}{n_{i.}}$$

$f_{yj/xi}$ indique la part de la sous-population considérée (c'est-à-dire les individus pour lesquels $X = X_i$) qui présente la modalité y_j de la variable Y.

La somme des fréquences conditionnelles $f_{yj/xi}$ est égale à 1 (ou 100 %) :

$$\sum_{j=1}^{p} f_{y_j/x_i} = \sum_{j=1}^{p} \frac{n_{ij}}{n_{i.}} = \frac{1}{n_{i.}} \sum_{j=1}^{p} n_{ij} = \frac{1}{n_{i.}} n_{i.} = 1$$

b) Tableau des profils-lignes

Les distributions conditionnelles de Y selon X sont souvent regroupées dans un unique tableau dans lequel figurent dans la marge supérieure les modalités de Y et, dans la 1re ligne, les fréquences conditionnelles de Y selon $X = x_1$, dans la 2e les fréquences conditionnelles de Y selon $X = x_2$…, dans la ke les fréquences conditionnelles de Y selon $X = x_k$. Ce tableau, souvent complété par une dernière ligne donnant les fréquences marginales de Y, s'appelle le tableau des profils-lignes :

Modalités de Y : y_j	y_1	y_2	…	y_p	Ensemble
$X = x_1$	f_{y_1/x_1}	f_{y_2/x_1}	…	f_{y_p/x_1}	1
$X = x_2$	f_{y_1/x_2}	f_{y_2/x_2}	…	f_{y_p/x_2}	1
. . .	. . .	. . .	. . .	. . .	. . .
$X = x_k$	f_{y_1/x_k}	f_{y_2/x_k}	…	f_{y_p/x_k}	1
Ensemble	$f_{.1}$	$f_{.2}$	…	$f_{.p}$	1

■ *Moyennes et variances conditionnelles*

a) Moyennes conditionnelles

On calcule une ***moyenne conditionnelle de X*** pour chacune des p distributions conditionnelles de X. La moyenne conditionnelle de X liée par $Y = y_j$ est notée $\bar{x}_j$. C'est une moyenne arithmétique pondérée définie par :

$$\bar{x}_j = \frac{1}{n_{.j}} \sum_{i=1}^{k} n_{ij} x_i$$

En utilisant les fréquences conditionnelles, elle s'écrit : $\bar{x}_j = \sum_{i=1}^{k} f_{x_i/y_j} x_i$.

De même, on calcule une ***moyenne conditionnelle de Y*** pour chacune des k distributions conditionnelles de Y. La moyenne conditionnelle de Y liée par $X = x_i$ est notée $\bar{y}_i$. C'est une moyenne arithmétique pondérée définie par : $\bar{y}_i = \frac{1}{n_{i.}} \sum_{j=1}^{p} n_{ij} y_j$

ou en utilisant les fréquences conditionnelles : $\bar{y}_i = \sum_{j=1}^{p} f_{y_j/x_i} y_j$

Pour chacune des variables, on a :

Moyenne marginale	=	moyenne des moyennes conditionnelles

Pour la variable X : $\bar{x} = \frac{1}{n} \sum_{j=1}^{p} n_{.j} \bar{x}_j$ et pour la variable Y : $\bar{y} = \frac{1}{n} \sum_{i=1}^{k} n_{i.} \bar{y}_i$

b) Variances et écarts-types conditionnels

Il y a autant de variances et écart-types conditionnels de X que de distributions conditionnelles de X, soit p. La variance conditionnelle de X liée par $Y = y_j$ est notée $V_j(X)$. Elle est définie par :

$$V_j(X) = \frac{1}{n_{.j}} \sum_{i=1}^{k} n_{ij} (x_i - \bar{x}_j)^2$$

ou, en développant et simplifiant : $V_j(X) = \frac{1}{n_{.j}} \sum_{i=1}^{k} n_{ij} x_i^2 - \bar{x}_j^2$.

En introduisant les fréquences conditionnelles, on obtient :

$$V_j(X) = \sum_{i=1}^{k} f_{x_i/y_j} (x_i - \bar{x}_j)^2 \text{ et } V_j(X) = \sum_{i=1}^{k} f_{x_i/y_j} x_i^2 - \bar{x}_j^2 .$$

On note $\sigma_j(X)$ l'écart-type conditionnel associé : $\sigma_j(X) = \sqrt{V_j(X)}$.

De même, il y a autant de variances et écart-types conditionnels de Y que de distributions conditionnelles de Y, soit k. La variance conditionnelle de Y liée par $\mathbf{X = x_i}$ est notée $V_i(Y)$. Elle est définie par :

$$V_i(Y) = \frac{1}{n_{i.}} \sum_{j=1}^{p} n_{ij} (y_j - \bar{y}_i)^2$$

ou, en développant et simplifiant : $V_i(Y) = \frac{1}{n_{i.}} \sum_{j=1}^{p} n_{ij} y_j^2 - \bar{y}_i^2$. En utilisant les fréquences conditionnelles :

$V_i(Y) = \sum_{j=1}^{p} f_{y_j/x_i} (y_j - \bar{y}_i)^2$ et $V_i(Y) = \sum_{j=1}^{p} f_{y_j/x_i} y_j^2 - \bar{y}_i^2$.

L'écart-type conditionnel associé est noté $\sigma_i(Y)$: $\sigma_i(Y) = \sqrt{V_i(Y)}$.

Pour chacune des variables X et Y, on a la relation suivante :

Variance marginale	=	moyenne des variances conditionnelles	+	variance des moyennes conditionnelles

Cette formule appliquée à la variance de X s'écrit : $V(X) = \frac{1}{n} \sum_{j=1}^{p} n_{.j} V_j(X) + \frac{1}{n} \sum_{j=1}^{p} n_{.j} (\bar{x}_j - \bar{x})^2$.

Appliquée à la variance de Y, elle devient : $V(Y) = \frac{1}{n} \sum_{i=1}^{k} n_{i.} V_i(Y) + \frac{1}{n} \sum_{i=1}^{k} n_{i.} (\bar{y}_i - \bar{y})^2$.

La moyenne des variances conditionnelles mesure la moyenne des variances des différentes sous-populations étudiées dans chacune des distributions conditionnelles. C'est donc la moyenne des dispersions à l'intérieur de chacune de ces sous-populations. Elle est aussi appelée variance intrapopulation.

La variance des moyennes conditionnelles mesure la dispersion des moyennes conditionnelles des différentes sous-populations autour de la moyenne marginale, donc la dispersion entre les sous-populations. Elle est aussi appelée variance interpopulation.

■ *Application*

a) Énoncé

La répartition des entreprises (hors agriculture) selon le nombre de salariés et le secteur d'activité, en France, au 1er janvier 2006, est donnée par le tableau ci-dessous (source : Insee, Connaissance locale de l'appareil productif (Clap), www.insee.fr).

Nbr de salariés / Secteur	0	[1 ; 10[	[10 ; 50[	[50 ; 200[	[200 ; 500[	500 et plus
Industrie	102 588	101 133	32 204	7 135	1 645	854
Construction	174 854	161 930	24 866	1 920	246	120
Tertiaire	1 277 461	652 898	94 836	13 081	2 315	1 108

La population est l'ensemble des entreprises (hors agriculture), en France, au 1er janvier 2006. Ces entreprises sont caractérisées selon deux variables. L'une est qualitative : c'est le secteur d'activité ; elle sera notée X. L'autre est quantitative continue : c'est le nombre de salariés ; elle sera notée Y.

b) Distributions conditionnelles de X selon Y

Les modalités de Y sont réparties en six groupes ; il y a donc six distributions conditionnelles de X selon Y. La première est la distribution conditionnelle de X liée par Y = 0. Dans le tableau de contingence, elle est définie par la marge de gauche (les modalités de X) et la première colonne (les effectifs n_{i1}). Elle peut être présentée dans un tableau séparé dans lequel sont calculées les fréquences conditionnelles :

Modalité : x_i	Effectif : n_{i1}	Fréquence : f_{x_i/y_1}
Industrie	102 588	6,6 %
Construction	174 854	11,2 %
Tertiaire	1 277 461	82,2 %
Ensemble	1 554 903	100,0 %

Cette distribution indique la répartition des 1 554 903 entreprises qui n'ont aucun salarié en fonction de leur secteur d'activité. Plus de 80 % d'entre elles appartiennent au secteur tertiaire, 11,2 % à la construction et seulement 6,6 % à l'industrie.

La deuxième est la distribution conditionnelle de X liée par $Y \in [1 ; 10[$. Elle est définie par la marge de gauche et la deuxième colonne du tableau de contingence, ou par le tableau ci-dessous, qui précise les fréquences conditionnelles :

Modalité : x_i	Effectif : n_{i2}	Fréquence : f_{x_i/y_2}
Industrie	101 133	11,0 %
Construction	161 930	17,7 %
Tertiaire	652 898	71,3 %
Ensemble	915 961	100,0 %

Les entreprises qui ont entre 1 et 9 salariés sont au nombre de 915 961. Elles appartiennent au secteur tertiaire pour 71,3 %, à la construction pour 17,7 % et à l'industrie pour 11 %.

On procède successivement de la même manière pour dresser le tableau de distribution des quatre autres distributions conditionnelles.

Le caractère X étant qualitatif, il ne se prête pas aux calculs de moyennes et écarts-types.

c) Distributions conditionnelles de Y selon X, moyennes et variances conditionnelles

La variable X a trois modalités, il y a donc trois distributions conditionnelles de Y selon X.

La première est la distribution conditionnelle de Y sachant que X est « l'industrie ». Dans le tableau de contingence, elle est définie par la marge supérieure (les modalités de Y) et la première ligne (les effectifs n_{1j}). Elle peut être présentée dans un tableau séparé dans lequel sont calculées les fréquences conditionnelles :

	0	[1 ; 10[	[10 ; 50[	[50 ; 200[	[200 ; 500[	500 et plus	Ensemble
y_j	0	5,5	30	125	350	?	
n_{1j}	102 588	101 133	32 204	7 135	1 645	854	245 559
f_{y_j/x_1}	41,8 %	41,2 %	13,1 %	2,9 %	0,7 %	0,3 %	100,0 %

Les entreprises industrielles étaient au nombre de 245 559 en 2006. 41,8 % d'entre elles n'avaient pas de salarié, 41,2 % en avaient au moins 1 et au plus 9…, 0,3 % en avaient 500 ou plus.

Pour effectuer les calculs des moyennes et écarts-types, il faut choisir une valeur pour l'extrémité de la dernière classe. En l'absence de toute information sur cette valeur, on décide souvent de donner à la dernière classe une amplitude double de la précédente. Ici cela conduit à une extrémité égale à 1 100, donc un centre de classe égal à (500 + 1 100)/2, soit 800.

Lorsque X est « l'industrie » :

$$\bar{y}_1 = \frac{1}{n_{1.}} \sum_{j=1}^{p} n_{1j} y_j = \frac{(102\,588)(0) + (101\,133)(5,5) + \ldots + (854)(800)}{245\,559} = 14,958$$

$$V_1(Y) = \frac{1}{n_{1.}} \sum_{j=1}^{p} n_{1j} y_j^2 - \bar{y}_1^2 = \frac{(102\,588)(0^2) + (101\,133)(5,5^2) + \ldots + (854)(800^2)}{245\,559} - 14,958^2 = 3\,407,144$$

$$\sigma_1(Y) = 58,4$$

Il y avait en moyenne 15 salariés dans une entreprise industrielle, en France en 2006. L'écart moyen à la moyenne était de 58,4 salariés, un nombre très supérieur à la moyenne parce que la distribution est asymétrique, étalée à droite. Plus de 80 % des entreprises ont moins de 10 salariés, concentrés « à gauche » de la distribution. La distribution s'étale ensuite vers la droite jusqu'à des valeurs beaucoup plus élevées (plus de 500).

La deuxième distribution conditionnelle est celle de Y, sachant que X est « la construction ». Dans le tableau de contingence, elle est définie par la marge supérieure (les modalités de Y) et la deuxième ligne (les effectifs n_{2j}). Elle peut être présentée comme précédemment dans un tableau intégrant les fréquences conditionnelles :

	0	[1 ; 10[	[10 ; 50[	[50 ; 200[	[200 ; 500[	500 et plus	Ensemble
y_j	0	5,5	30	125	350	800	
n_{2j}	174 854	161 930	24 866	1 920	246	120	363 936
f_{y_j/x_2}	48,0 %	44,5 %	6,8 %	0,5 %	0,1 %	0,0 %	100,0 %

Le secteur d'activité « construction » comprenait 363 936 entreprises. Parmi elles, 48 % n'avaient pas de salarié, 44,5 % en avaient au moins 1 et au plus 9…, moins de 0,1 % (donc 0,0 % dans le tableau) en avaient 500 ou plus.

Calcul des moyennes et variances conditionnelles, sachant que X est « la construction » :

$$\bar{y}_2 = \frac{1}{n_{2.}}\sum_{j=1}^{p} n_{2j}y_j = 5,657\ ,\ V_2(Y) = \frac{1}{n_{2.}}\sum_{j=1}^{p} n_{2j}y_j^2 - \bar{y}_2^2 = 419,215 \text{ d'où } \sigma_2(Y) = 20,5.$$

Il y avait en moyenne un peu moins de 5,7 salariés dans une entreprise de la construction, en France en 2006. L'écart moyen à la moyenne était de 20,5 salariés, un nombre très supérieur à la moyenne pour la même raison que précédemment (distribution asymétrique, étalée à droite).

La dernière distribution conditionnelle de Y sachant que X est « le tertiaire » est définie de même par les modalités de Y et les effectifs n_{3j} dont on déduit les fréquences conditionnelles :

	0	[1 ; 10[	[10 ; 50[	[50 ; 200[	[200 ; 500[	500 et plus	Ensemble
y_j	0	5,5	30	125	350	800	
n_{3j}	1 277 461	652 898	94 836	13 081	2 315	1 108	2 041 699
f_{y_j/x_3}	62,6 %	32,0 %	4,6 %	0,6 %	0,1 %	0,1 %	100,0 %

En 2006, il y avait 2 041 699 entreprises dans le secteur tertiaire. Parmi elles, 62,6 % n'avaient pas de salarié, 32 % en avaient au moins 1 et au plus 9…, 0,1 % en avaient 500 ou plus.

Calcul des moyennes et variances conditionnelles, sachant que X est « le tertiaire » :

$$\bar{y}_3 = \frac{1}{n_{3.}}\sum_{j=1}^{p} n_{3j}y_j = 4,784\ ,\ V_3(Y) = \frac{1}{n_{3.}}\sum_{j=1}^{p} n_{3j}y_j^2 - \bar{y}_3^2 = 614,914 \text{ d'où } \sigma_3(Y) = 24,8.$$

Il y avait en moyenne 4,8 salariés dans une entreprise du tertiaire, en France en 2006. L'écart moyen à la moyenne était de 24,8 salariés, un nombre très supérieur à la moyenne, comme précédemment et pour la même raison (distribution asymétrique, étalée à droite).

d) Moyenne et variance marginales de Y

Les moyennes et variances conditionnelles permettent de calculer les moyenne et variance marginales (qui pourraient aussi être obtenues à partir de la distribution marginale de Y).

La moyenne des moyennes conditionnelles est égale à la moyenne marginale de Y :

$$\frac{1}{n}\sum_{i=1}^{k} n_{i.}\bar{y}_i = \frac{1}{2\ 651\ 194}(245\ 559)(14,\ 96) + (363\ 936)(5,\ 66) + (2\ 041\ 699)(4,\ 78) = 5,85 = \bar{y}$$

Il y avait en moyenne 5,9 salariés dans une entreprise, en France en 2006.

Variance de Y : $V(Y) = \frac{1}{n}\sum_{i=1}^{k} n_{i.} V_i(Y) + \frac{1}{n}\sum_{i=1}^{k} n_{i.} (\overline{y}_i - \overline{y})^2$

$$\frac{1}{n}\sum_{i=1}^{k} n_{i.} V_i(Y) = \frac{(245\,559)(3\,407,\ 144) + (363\,936)(419,\ 215) + (2\,041\,699)(614,\ 914)}{2\,651\,194} = 846,67$$

$$\frac{1}{n}\sum_{i=1}^{k} n_{i.} (\overline{y}_i - \overline{y})^2 = \frac{(245\,559)(14,\ 96 - 5,85)^2 + (363\,936)(5,\ 66 - 5,85)^2 + (2\,041\,699)(4,\ 78 - 5,85)^2}{2\,651\,194}$$

d'où $\frac{1}{n}\sum_{i=1}^{k} n_{i.} (\overline{y}_i - \overline{y})^2 = 8,57$

$V(Y) = 846,67 + 8,57 = 855,24$ et $\sigma(Y) = 29,2$.

La variance des moyennes conditionnelles est très faible car il y a très peu d'écart entre les moyennes conditionnelles et la moyenne marginale : les nombres moyens de salariés sont respectivement 15 pour l'industrie, 5,7 pour la construction, 4,8 pour le tertiaire, et 5,9 pour l'ensemble des secteurs. La dispersion autour de la moyenne marginale est due principalement à la dispersion autour de chaque moyenne conditionnelle, c'est-à-dire à la dispersion du nombre de salariés à l'intérieur de chacune des sous-populations.

L'écart moyen à la moyenne est égal à 29,2, valeur supérieure à la moyenne marginale puisque la distribution marginale est, comme chacune des distributions conditionnelles, asymétrique, étalée à droite.

5 Distribution conjointe : indépendance, liaison fonctionnelle et liaison relative

■ *Indépendance*

a) Définition

Deux variables X et Y sont indépendantes lorsque, dans le tableau des profils-lignes, les lignes sont toutes exactement semblables. On dit que les profils-lignes sont identiques.

Quel que soit x_i, pour une valeur de j donnée, les fréquences conditionnelles f_{y_j/x_i} sont égales et égales à la fréquence marginale $f_{.j}$: $f_{y_j/x_i} = f_{.j}$

Dans le tableau des profils-colonnes, les colonnes sont aussi toutes semblables. On dit que les profils-colonnes sont identiques.

Quel que soit y_j, pour une valeur de i donnée, les fréquences conditionnelles f_{x_i/y_j} sont égales et égales à la fréquence marginale $f_{i.}$: $f_{x_i/y_j} = f_{i.}$

Les lignes du tableau de contingence sont alors proportionnelles entre elles et les colonnes sont également proportionnelles entre elles : $n_{ij} = \frac{n_{i.} n_{.j}}{n}$ et $f_{ij} = f_{i.} f_{.j}$.

b) Application

Soient X et Y deux variables décrivant un ensemble de 105 salariés d'une entreprise. La variable X est le nombre d'enfants ; elle possède quatre modalités : 1, 2, 3 ou 4. La variable Y est le statut ; elle possède trois modalités : ouvrier, employé ou cadre.

X \ Y	Ouvrier	Employé	Cadre	Ensemble
1	2	4	8	14
2	6	12	24	42
3	3	6	12	21
4	4	8	16	28
	15	30	60	105

1) Tableau des profils-lignes

X \ Y	Ouvrier	Employé	Cadre	Ensemble
1	2/14 = 14,3 %	4/14 = 28,6 %	8/14 = 57,1 %	14/14 = 100,0 %
2	14,3 %	28,6 %	57,1 %	100,0 %
3	14,3 %	28,6 %	57,1 %	100,0 %
4	14,3 %	28,6 %	57,1 %	100,0 %
	15/105 = 14,3 %	30/105 = 28,6 %	60/105 = 57,1 %	105/105 = 100,0 %

Quel que soit le nombre d'enfants, la proportion d'ouvriers est la même (14,3 %), la proportion d'employés est la même (28,6 %) et la proportion de cadres est aussi la même (57,1 %).

2) Tableau des profils-colonnes

X \ Y	Ouvrier	Employé	Cadre	Ensemble
1	2/15 = 13,3 %	13,3 %	13,3 %	14/105 = 13,3 %
2	6/15 = 40,0 %	40,0 %	40,0 %	42/105 = 40,0 %
3	3/15 = 20,0 %	20,0 %	20,0 %	21/105 = 20,0 %
4	4/15 = 26,7 %	26,7 %	26,7 %	28/105 = 26,7 %
	15/15 = 100,0 %	100,0 %	100,0 %	105/105 = 100,0 %

Quel que soit le statut du salarié dans l'entreprise, la proportion de ceux qui ont un enfant est la même (13,3 %), la proportion de ceux qui en ont deux est la même (40 %), la proportion de ceux qui en ont trois est la même (20 %), enfin la proportion qui en ont quatre est la même (26,7 %).

Dans cette population, il n'y a aucune influence du statut sur le nombre d'enfants et du nombre d'enfants sur le statut. Les deux variables sont indépendantes.

■ *Liaison fonctionnelle*

a) Définitions

La variable ***Y est liée fonctionnellement à X*** lorsqu'à chaque modalité de X (ou classe si X est quantitative continue) correspond une modalité (ou classe) unique de la variable Y. Dans ce cas, il n'y a dans le tableau de la distribution conjointe qu'un seul effectif non nul (ou une seule fréquence non nulle) par ligne.

La variable ***X est liée fonctionnellement à Y*** lorsqu'à chaque modalité (ou classe) de Y correspond une modalité (ou classe) unique de la variable X. Dans ce cas, il n'y a dans le tableau de la distribution conjointe qu'un seul effectif non nul (ou une seule fréquence non nulle) par colonne.

Deux variables ***X et Y sont en liaison fonctionnelle réciproque*** (ou réciproquement dépendantes) lorsqu'à chaque modalité (ou classe) de la variable X correspond une modalité (ou classe) unique de la variable Y et réciproquement. Cela n'est possible que si les deux variables ont le même nombre

de modalités. Le tableau de la distribution conjointe a autant de lignes que de colonnes et il y a un seul effectif non nul (ou une seule fréquence non nulle) par ligne et par colonne.

b) Application

1) Variable Y liée fonctionnellement à X

Soient X et Y deux variables décrivant un ensemble de 100 salariés d'une entreprise. La variable X est le salaire mensuel (en euros) ; ses modalités sont regroupées en quatre classes : [1 000 ; 1 500[, [1 500 ; 2 000[, [2 000 ; 2 500[et [2 500 ; 3 000[. La variable Y est le statut avec trois modalités : ouvrier, employé ou cadre.

Y / X	Ouvrier	Employé	Cadre	Ensemble
[1 000 ; 1 500[	20	0	0	20
[1 500 ; 2 000[	0	20	0	20
[2 000 ; 2 500[	0	10	0	10
[2 500 ; 3 000[	0	0	50	50
Ensemble	20	30	50	100

À chaque classe de salaire de X correspond une seule modalité de Y. Tous les salariés dont le salaire est compris entre 1 000 et 1 500 euros sont ouvriers ; ceux dont le salaire est compris entre 1 500 et 2 000 euros sont employés, il en est de même pour ceux dont le salaire est compris entre 2 000 et 2 500 euros ; ceux dont le salaire est compris entre 2 500 et 3 000 euros sont tous cadres.

La variable Y est donc liée fonctionnellement à X : le salaire détermine le statut. Sur chaque ligne, un seul effectif est non nul.

2) Variable X liée fonctionnellement à Y

Soient X et Y deux variables décrivant un autre ensemble de 100 salariés d'une entreprise. La variable X est le salaire mensuel (en euros) ; ses modalités sont regroupées en trois classes : [1 000 ; 1 500[, [1 500 ; 2 000[, [2 000 ; 2 500[. La variable Y est le statut avec quatre modalités : ouvrier, technicien, technicien supérieur ou cadre.

Y / X	Ouvrier	Technicien	Technicien supérieur	Cadre	Ensemble
[1 000 ; 1 500[	20	0	0	0	20
[1 500 ; 2 000[	0	20	15	0	35
[2 000 ; 2 500[	0	0	0	45	45
Ensemble	20	20	15	45	100

À chaque statut correspond une seule classe de X. Tous les ouvriers ont un salaire compris entre 1 000 et 1 500 euros ; tous les techniciens ont un salaire compris entre 1 500 et 2 000 euros ; tous les techniciens supérieurs ont un salaire compris entre 1 500 et 2 000 euros ; tous les cadres ont un salaire compris entre 2 000 et 2 500 euros.

La variable X est donc liée fonctionnellement à Y : le statut détermine le salaire. Sur chaque colonne, un seul effectif est non nul.

3) X et Y en liaison fonctionnelle réciproque

Soient X et Y deux variables décrivant un troisième ensemble de 200 salariés d'une entreprise. La variable X est le salaire mensuel (en euros) ; ses modalités sont regroupées en quatre classes : [1 000 ; 1 500[, [1 500 ; 2 000[, [2 000 ; 2 500[, [2 500 ; 3 000[. La variable Y est le statut avec quatre modalités : ouvrier, technicien, technicien supérieur ou cadre.

Y / X	Ouvrier	Technicien	Technicien supérieur	Cadre	Ensemble
[1 000 ; 1 500[	65	0	0	0	65
[1 500 ; 2 000[	0	24	0	0	24
[2 000 ; 2 500[	0	0	38	0	38
[2 500 ; 3 000[	0	0	0	73	73
Ensemble	65	24	38	73	200

Tous les ouvriers gagnent entre 1 000 et 1 500 euros et les salariés qui gagnent entre 1 000 et 1 500 euros sont tous ouvriers, tous les techniciens gagnent entre 1 500 et 2 000 euros et les

salariés qui gagnent entre 1 000 et 2 000 euros sont tous ouvriers, etc. Le statut détermine le salaire et réciproquement. Il n'y a qu'un seul effectif non nul par ligne et par colonne.

Les variables X et Y sont réciproquement dépendantes.

■ *Liaison relative*

Deux variables sont en liaison relative lorsqu'elles ne sont ni indépendantes ni en liaison fonctionnelle. C'est le cas le plus courant.

Dans ce cas, il existe des méthodes (ne relevant pas de ce manuel) pour quantifier l'intensité du lien entre les variables. Si l'intensité du lien le justifie, on cherche alors à déterminer si certaines modalités de l'une des variables correspondent à certaines modalités particulières de l'autre et à mesurer l'intensité de cette correspondance. Ce travail relève d'une technique d'analyse des données, nommée « analyse factorielle des correspondances », qui suppose l'utilisation de logiciels appropriés. Elle n'est donc pas traitée dans cet ouvrage.

Les distributions statistiques à deux variables quantitatives : corrélation et régression

CHAPITRE 9

Ce chapitre porte sur l'étude du lien qui peut exister entre deux variables quantitatives X et Y. L'analyse macro-économique s'intéresse par exemple au lien entre la consommation des ménages et leur revenu, entre l'investissement des entreprises et le taux d'intérêt ; l'analyse micro-économique étudie la liaison entre le prix d'un produit et la quantité vendue de ce produit ; un chef d'entreprise souhaite savoir si les dépenses en publicité de son entreprise sont en lien avec ses ventes…

L'étude de ce lien passe par la détermination, à l'aide de la méthode des moindres carrés, de l'équation de la courbe qui ajuste au mieux le nuage de points et par l'évaluation de l'intensité de la corrélation.

1 Liaisons entre variables quantitatives

■ *Corrélation*

Considérons une population de n individus pour laquelle on s'intéresse à deux variables quantitatives X et Y. À chaque individu i est associé un couple de valeurs $(x_i ; y_i)$. La distribution statistique est donc constituée de l'ensemble des couples $(x_i ; y_i)$, $i \in \{1, 2\ldots, n\}$. Ces données sont représentées graphiquement en portant X en abscisse et Y en ordonnée ; à chaque valeur de X est associée une valeur de Y. On obtient un nuage de n points $(x_1 ; y_1)$, $(x_2 ; y_2)$, …, $(x_n ; y_n)$, appelé diagramme de dispersion.

Dans le cas particulier où l'une des variables est le temps, elle est notée T et l'autre est généralement notée Y. La distribution est nommée série chronologique. Les couples de valeurs sont $(t_1 ; y_1)$, $(t_2 ; y_2)$, …, $(t_n ; y_n)$.

On dit qu'il existe une ***corrélation*** entre les variables X et Y si les variations de l'une entraînent des variations de l'autre. Soit les variations des deux variables ont tendance à se produire dans le même sens et la corrélation est dite ***positive*** ; soit les variations des deux variables ont tendance à se produire en sens inverse et la corrélation est dite ***négative***.

La représentation du diagramme de dispersion donne une idée de l'éventuelle corrélation entre les deux variables. Le diagramme 1 illustre une corrélation négative : en général, lorsque X augmente, Y diminue. Le diagramme 2 illustre une corrélation positive : lorsque X augmente, on voit que le plus souvent Y augmente aussi.

Diagramme 1 : corrélation négative

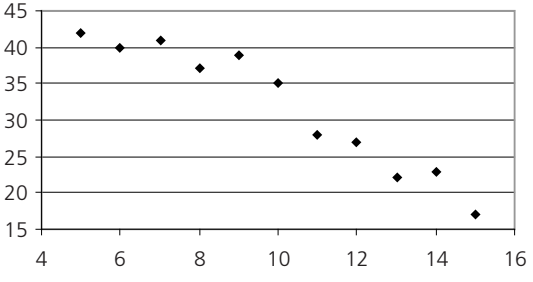
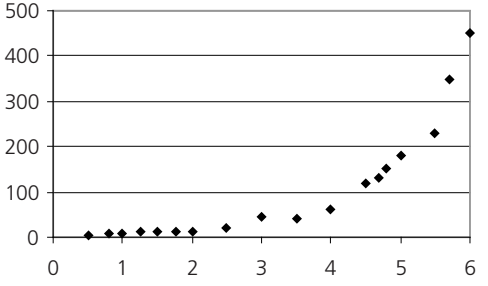


Diagramme 2 : corrélation positive



Il existe un indicateur statistique qui permet de dire si, globalement, deux variables varient ou non dans le même sens : c'est la covariance.

■ *Covariance*

a) Définition

La covariance du couple (X, Y) est la moyenne des produits des écarts aux moyennes $\bar{x}$ et $\bar{y}$:

$$\text{Cov}(X, Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

C'est un indicateur du sens de la variation simultanée. Sa valeur n'a pas de signification particulière, seul son signe en a une.

Si la covariance est positive, X et Y varient « globalement » dans le même sens. « Globalement » indique que, lorsque la variable X augmente, la variable Y tend également à augmenter, mais elle n'augmente pas nécessairement. Il peut y avoir des cas où une hausse de X s'accompagne d'une baisse de Y.

Si la covariance est négative, X et Y varient « globalement » en sens inverse. Une hausse de X se traduit le plus souvent, éventuellement tout le temps, par une baisse de Y.

En développant et simplifiant la formule ci-dessus, on obtient :

$$\text{Cov}(X, Y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}$$

Cette formule est plus simple à mettre en œuvre que la formule de la définition. C'est donc celle qui en pratique est utilisée pour calculer la covariance.

b) Application

Le tableau suivant indique la surface moyenne (en m²) des résidences principales selon le nombre de pièces, en France en 2002 (source : Insee, Enquête logement 2002, www.insee.fr) :

Nombre de pièces	1	2	3	4	5	6 et plus
Surface moyenne	29	48	70	90	109	147

La population est l'ensemble des résidences principales. Soit X la variable « nombre de pièces » et Y la variable « surface moyenne ». La dernière modalité de la variable X regroupe plusieurs valeurs ; pour tracer le diagramme et effectuer les calculs, nous assimilons « 6 et plus » à « 7 ».

Le tableau des données montre que chaque augmentation du nombre de pièces se traduit par une augmentation de la surface moyenne. La corrélation entre X et Y est donc positive.

Calculons la covariance : $\text{Cov}(X, Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = (\frac{1}{n}\sum_{i=1}^{n} x_i y_i) - \bar{x}\bar{y}$.

Pour calculer Cov(X,Y), il faut d'abord calculer $\bar{x}$ et $\bar{y}$:

x_i	y_i	$x_i.y_i$
1	29	29
2	48	96
3	70	210
4	90	360
5	109	545
7	147	1 029
22	493	2 269

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{6}(22) = 3,67$ et $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{6}(493) = 82,17.$

D'où : $Cov(X,Y) = (\frac{1}{6}2269) - (3,67)(82,17) = 76,60$.

La covariance est évidemment positive.

■ *Liaison linéaire et liaison non linéaire*

Le diagramme de dispersion donne aussi une idée de la forme de la liaison entre les variables.

Par exemple, le diagramme 1 ci-dessous a une forme allongée qui laisse penser à une liaison linéaire tandis que le diagramme 2 semble plutôt correspondre à une liaison exponentielle.

Diagramme 1 : liaison linéaire

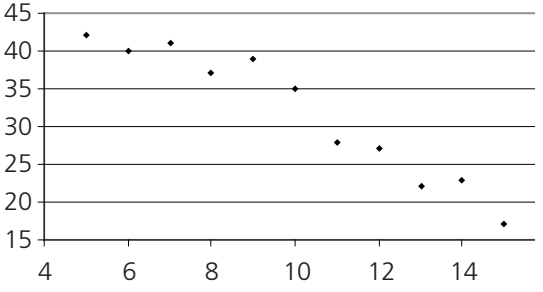
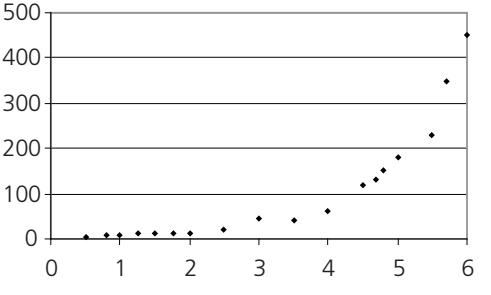


Diagramme 2 : liaison exponentielle



Pour analyser le lien qui existe entre les variables X et Y, on cherche à déterminer l'équation d'une courbe qui passe « le plus près possible » des points du nuage. Dans le cas du diagramme 1, ce sera l'équation d'une droite ; dans la cas du diagramme 2, l'équation d'une fonction exponentielle.

L'équation ainsi déterminée permet de préciser comment varie Y en fonction de X, ou X en fonction de Y, donc d'analyser le lien entre ces variables. Si, par exemple, l'équation est celle d'une droite et s'écrit $Y = -3,2X + 5$, la pente – 3,2 de cette droite nous indique que lorsque X augmente d'une unité, Y tend à diminuer de 3,2 unités. Cette équation va aussi permettre de prévoir la valeur que prendra Y pour une valeur prévue ou possible de X.

Lorsque la distribution est une série chronologique, l'étude du lien entre Y et T vise à déterminer l'évolution de la variable Y sur longue période appelée tendance de la série.

La méthode des moindres carrés permet de déterminer l'équation d'une courbe qui passe « le plus près possible » des points du nuage. Nous allons d'abord envisager le cas où cette courbe est une

droite, appelée droite d'ajustement par les moindres carrés ou droite de régression. Nous verrons ensuite comment procéder lorsque la courbe illustre une fonction exponentielle ou une fonction puissance.

2 Ajustement affine par la méthode des moindres carrés : régression linéaire

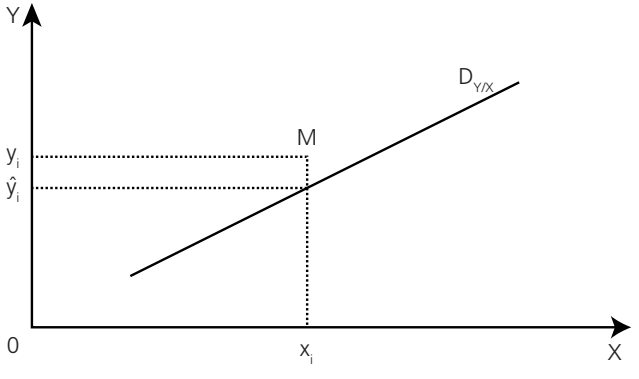
■ *Méthode des moindres carrés*

Lorsque les points du nuage paraissent relativement alignés, on va chercher à déterminer l'équation de la droite qui passe le plus prés possible de tous les points. La méthode dite des moindres carrés consiste à déterminer l'équation de la droite qui rend minimale la somme des carrés des écarts entre chaque point du nuage et la droite. Selon que les écarts sont mesurés parallèlement à l'axe des ordonnées ou à l'axe des abscisses, on obtient la droite de régression de Y en X d'équation $Y = aX + b$ ou la droite de régression de X en Y d'équation $X = a'Y + b'$.

■ *Droite de régression de Y en X : Y = aX + b (droite $D_{Y/X}$)*

La droite de régression de Y en X est la droite qui rend minimale la somme des carrés des distances entre chaque point du nuage et $D_{Y/X}$, les distances étant prises parallèlement à l'axe des ordonnées. L'objectif est de déterminer la valeur des coefficients a et b.

Soit M un point du nuage de points de coordonnées $(x_i ; y_i)$. Soit $\hat{y}_i$ l'ordonnée du point de la droite d'ajustement d'abscisse x_i : $\hat{y}_i = ax_i + b$.



Le carré de la distance entre M et la droite $D_{Y/X}$ est égal à $(y_i - \hat{y}_i)^2$.

La somme des carrés des distances entre les différents points du nuage et la droite $D_{Y/X}$ est égale à $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ avec $\hat{y}_i = ax_i + b$.

Pour déterminer l'équation de la droite de régression de Y en X, il faut donc minimiser $\sum_{i=1}^{n} (y_i - ax_i - b)^2$. Cette somme est une fonction à deux variables a et b.

Ce problème de minimisation a pour solution : $a = \frac{Cov(X, Y)}{V(X)}$ et $b = \bar{y} - a\bar{x}$.

L'équation de la droite de régression de Y en X est donc $Y = \frac{Cov(X, Y)}{V(X)} X + (\bar{y} - a\bar{x})$.

Quelles sont les caractéristiques de cette droite de régression ?

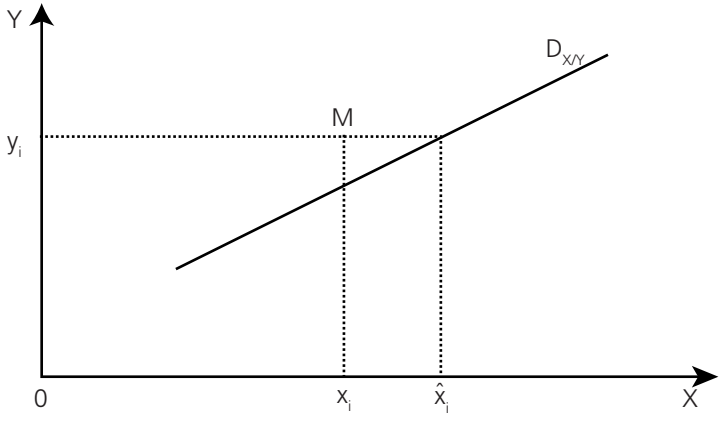
– sa pente $\frac{Cov(X, Y)}{V(X)}$ est de même signe que la covariance puisque la variance est toujours positive ;

– elle passe par le point moyen $(\bar{x}, \bar{y})$: $\bar{y} = a\bar{x} + b$.

■ *Droite de régression de X en Y : X = a'Y + b' (droite $D_{X/Y}$)*

La droite de régression de X en Y est la droite qui rend minimale la somme des carrés des distances entre chaque point du nuage et $D_{X/Y}$, les distances étant prises parallèlement à l'axe des abscisses. L'objectif est de déterminer a' et b'.

Soit M un point du nuage de points de coordonnées $(x_i ; y_i)$. Soit $\hat{x}_i$ l'abscisse du point de la droite d'ajustement d'ordonnée y_i : $\hat{x}_i = a'y_i + b'$.



La somme des carrés des distances entre les différents points du nuage et la droite $D_{X/Y}$ est égale à $\sum_{i=1}^{n}(x_i - \hat{x}_i)^2$, avec $\hat{x}_i = a'y_i + b'$. Pour déterminer l'équation de la droite de régression de X en Y, il faut donc minimiser $\sum_{i=1}^{n}$ pour obtenir $\sum_{i=1}^{n}(x_i - a'y_i - b')^2$. La solution de ce problème de minimisation est :

$$a' = \frac{Cov(X, Y)}{V(Y)} \quad \text{et } b' = \bar{x} - a'\bar{y}$$

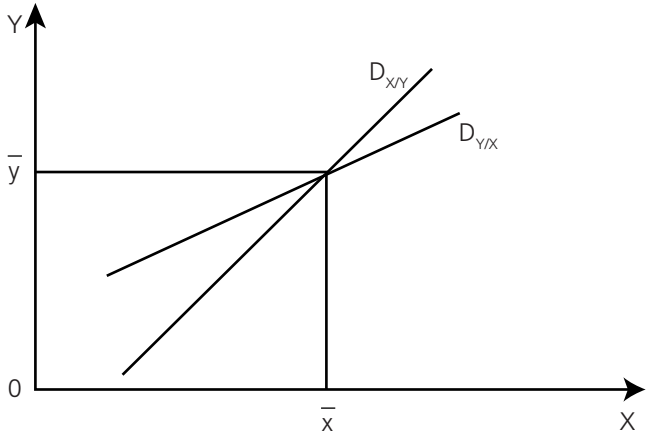
L'équation de la droite de régression de X en Y s'écrit : $X = \frac{Cov(X, Y)}{V(Y)}Y + (\bar{x} - a'\bar{y})$.

Quelles sont les caractéristiques de cette droite de régression ?

– sa pente $\frac{Cov(X, Y)}{V(Y)}$ est de même signe que la covariance ;

– elle passe par le point moyen $(\bar{x}, \bar{y})$: $\bar{x} = a'\bar{y} + b'$.

Les deux droites de régression $D_{Y/X}$ et $D_{X/Y}$ se coupent donc au point moyen $(\bar{x}, \bar{y})$.



La plupart des calculatrices incluent un programme statistique qui calcule, directement à partir des deux séries de données, la pente et l'ordonnée à l'origine de chacune des droites de régression. Il n'est pas nécessaire de calculer auparavant les moyennes, les variances et la covariance. C'est évidemment aussi le cas des logiciels statistiques (Excel, SPSS…). Dans les applications de ce chapitre, les calculs intermédiaires sont, le plus souvent, détaillés.

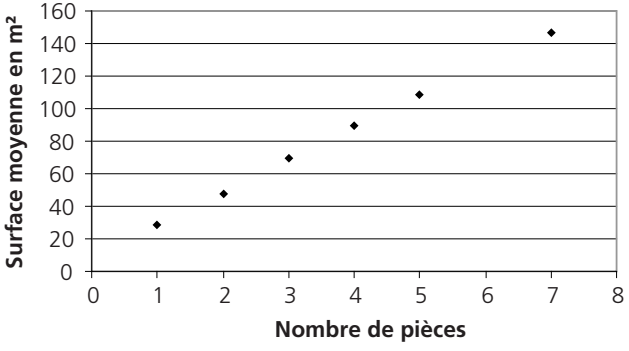
■ *Applications*

a) Droite de régression de Y en X

Nous avons vu que la surface moyenne des résidences principales et le nombre de pièces, en France en 2002, évoluaient comme suit :

Nombre de pièces (m²)	1	2	3	4	5	6 et plus
Surface moyenne	29	48	70	90	109	147

Les points du diagramme des données apparaissent presque alignés. C'est donc un ajustement affine qui s'impose.



L'équation de la droite de régression de Y en X est $Y = aX + b$, avec :

$$a = \frac{Cov(X, Y)}{V(X)} \quad \text{et} \quad b = \overline{y} - a\overline{x}$$

Les moyennes et la covariance ont été calculées antérieurement :

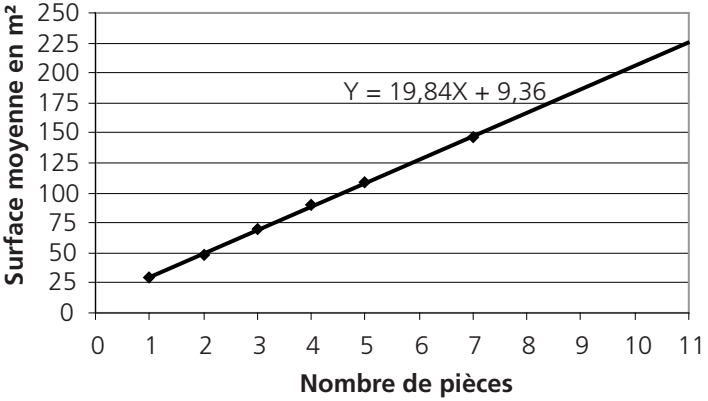
$\overline{x} = 3,67$ et $\overline{y} = 82,17$; $Cov(X,Y) = 76,60$

Calcul de la variance de X : $V(X) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2 = (\frac{1}{6}104) - 3,67^2 = 3,8644$

Pente de la droite de régression : $a = \frac{76,60}{3,86} = 19,84$

Ordonnée à l'origine : $b = 82,17 - (19,84)(3,67) = 9,36$

La droite de régression de Y en X est : Y = 19,84 X + 9,36. Elle est représentée ci-dessous.



L'équation obtenue est un « modèle » qui nous donne des informations que les données seules ne fournissaient pas.

La pente de la droite de régression représente la variation de Y induite par une variation de X d'une unité. En effet :

$Y(x + 1) – Y(x) = [19,84 (x + 1) + 9,36] – [19,84 x + 9,36] = 19,84 = a$

On peut donc dire qu'une pièce de plus à un logement augmente la surface moyenne d'environ 20 m².

L'équation de la droite de régression nous permet également de faire une évaluation de la surface moyenne d'un logement ayant par exemple 10 pièces :

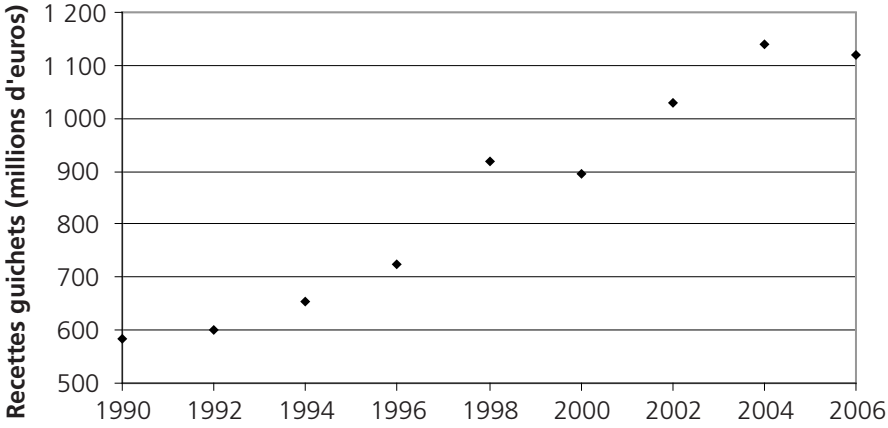
$$\hat{y}(10) = (19,84)(10) + 9,36 = 207,76$$

La surface moyenne d'un logement de 10 pièces est environ 208 m².

b) Droite de régression de Y en T : tendance d'une série chronologique

Le tableau suivant indique le montant de la recette totale (en millions d'euros) au guichet des salles de cinéma en France de 1990 à 2006 (source : Centre national de la cinématographie, www.cnc.fr) :

	1990	1992	1994	1996	1998	2000	2002	2004	2006
Recettes guichets	583,3	600,8	653,5	726,0	917,0	894,0	1 030,0	1 138,9	1 120,3



Le nuage de points suggère une liaison linéaire entre les recettes et le temps. Pour simplifier les calculs, on pose T = 0 en 1990, T = 2 en 1992…, T = 16 en 2006. Soit Y les recettes guichets. L'équation de la droite de régression de Y en T s'écrit : $Y = aT + b$.

Pour calculer a et b, il faut auparavant déterminer $\overline{t}$, $\overline{y}$, V(T) et Cov(T,Y).

Les sommes nécessaires pour effectuer ces calculs figurent en gras dans la dernière ligne du tableau ci-dessous.

t_i	y_i	t_i^2	y_i^2	$t_i y_i$
0	583,3	0	340 238,89	0,0
2	600,8	4	360 960,64	1 201,6
4	653,5	16	427 062,25	2 614,0
6	726,0	36	527 076,00	4 356,0
8	917,0	64	840 889,00	7 336,0
10	894,0	100	799 236,00	8 940,0
12	1 030,0	144	1 060 900,00	12 360,0
14	1 138,9	196	1 297 093,21	15 944,6
16	1 120,3	256	1 255 072,09	17 924,8
72	**7 663,8**	**816**	**6 908 528,08**	**70 677,0**

Moyennes de T et de Y : $\bar{t} = \frac{\sum_{i=1}^{n} t_i}{n} = \frac{72}{9} = 8$ et $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{7663,8}{9} = 851,5333$

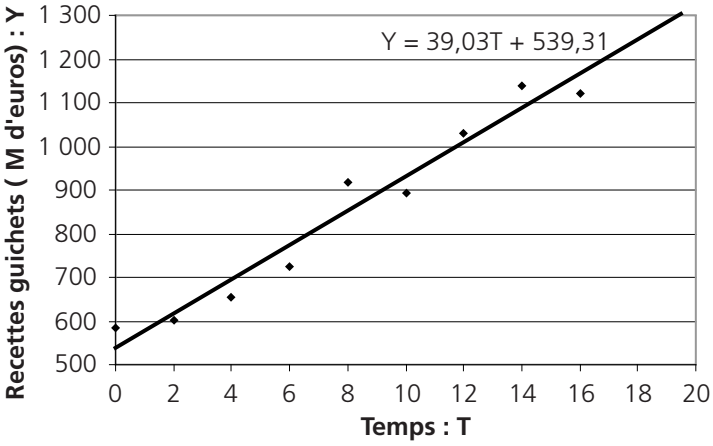
Variance de T : $V(T) = \frac{\sum_{i=1}^{n} t_i^2}{n} - \bar{t}^2 = \frac{816}{9} - 8^2 = 26,6667$

Covariance de T et Y : $Cov(T, Y) = \frac{\sum_{i=1}^{n} t_i y_i}{n} - \bar{t}\bar{y} = \frac{70677}{9} - (8)(851,5333) = 1040,7333$

D'où : $a = \frac{Cov(T, Y)}{V(T)} = \frac{1040,73333}{26,6667} = 39,0275$

$b = \bar{y} - a\bar{t} = 851,5333 - (39,0275)(8) = 539,3133$

En arrondissant a et b à 10^{-2} près : Y = 39,03 T + 539,31



La pente de cette droite traduit une tendance à la hausse des recettes guichets d'environ 39 millions d'euros chaque année entre 1990 et 2006.

Si cette tendance se poursuit, le montant prévu des recettes guichets pour 2008 serait :

$$\hat{y}\,(18) = (39{,}03)(18) + 539{,}31 = 1\,241{,}85$$

Il est cependant fort probable que le montant observé sera supérieur à cette prévision suite au succès exceptionnel du film *Bienvenue chez les Ch'tis* au premier semestre de l'année 2008.

3 Mesure de la qualité de la régression linéaire : le coefficient de corrélation linéaire

Plus les points du nuage sont proches de la droite de régression, plus cette droite résume bien le nuage de points et plus l'intensité de la liaison linéaire entre les variables X et Y est forte : on dit que la corrélation linéaire entre ces variables est forte. La dispersion des points du nuage autour de la droite de régression est mesurée par la variance dite résiduelle autour de la droite de régression.

■ *Variance résiduelle et variance expliquée par les droites de régression*

a) Droite de régression de Y en X ($D_{Y/X}$)

L'équation de la droite de régression de Y en X est Y = aX + b. Soit $\hat{y}_i$ l'ordonnée sur cette droite du point d'abscisse x_i :

$$\forall i \in \{1, 2..., n\} \quad \hat{y}_i = ax_i + b$$

On démontre que la variance de Y peut être décomposée en la somme de deux variances :

$$V(Y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = V_r(Y) + V_e(Y)$$

La variance résiduelle autour de $D_{Y/X}$, notée $V_r(Y)$, est la moyenne des carrés des écarts entre chaque point du nuage et la droite de régression, ces écarts étant pris parallèlement à l'axe des Y ; elle mesure donc la dispersion des points du nuage autour de la droite de régression $D_{Y/X}$.

$$V_r(Y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

La variance « expliquée » par la droite $D_{Y/X}$, notée $V_e(Y)$, est la moyenne des carrés des écarts sur la droite de régression entre chaque ordonnée $\hat{y}_i$ et la moyenne $\bar{y}$; elle mesure donc la dispersion des points sur la droite de régression.

$$V_e(Y) = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

b) Droite de régression de X en Y ($D_{X/Y}$)

L'équation de la droite de régression de X en Y est X = a'Y + b'. Soit $\hat{x}_i$ l'abscisse sur cette droite du point d'ordonnée y_i :

$$\forall i \in \{1, 2..., n\} \quad \hat{x}_i = a'y_i + b'$$

On démontre comme précédemment que la variance de X peut être décomposée en la somme de deux variances :

$$V(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2 + \frac{1}{n}\sum_{i=1}^{n}(\hat{x}_i - \bar{x})^2 = V_r(X) + V_e(X)$$

La variance résiduelle autour de $D_{X/Y}$, $V_r(X)$, mesure la dispersion des points du nuage autour de la droite de régression $D_{X/Y}$. La variance « expliquée » par la droite $D_{X/Y}$, $V_e(X)$, mesure la dispersion des points sur la droite de régression $D_{X/Y}$.

■ *Coefficient de détermination r²*

D'après la décomposition de la variance, plus la variance expliquée est forte, plus la variance résiduelle est faible et donc plus les points du nuage sont proches de la droite de régression. La rapport de la variance expliquée à la variance totale est donc un indicateur de la qualité de la régression linéaire.

Ce rapport est appelé ***coefficient de détermination*** ; il est noté r².

$$r^2 = \frac{V_e(Y)}{V(Y)} = \frac{V_e(X)}{V(X)}$$

r² mesure la part de la variance expliquée par les droites de régression. Il est toujours compris entre 0 et 1 puisque $V_e(Y) \leq V(Y)$ et $V_e(X) \leq V(X)$.

$$0 \leq r^2 \leq 1$$

Plus il est proche de 1, plus la corrélation linéaire entre les variables est forte.

Plus il s'approche de 0, plus elle est faible. Cela ne signifie pas pour autant que les variables ne sont pas corrélées. Elles peuvent être corrélées, mais pas linéairement.

Pour calculer le coefficient de détermination, on n'utilise généralement pas la formule ci-dessus. On démontre en effet que le coefficient de détermination est aussi égal au rapport $\frac{Cov^2(X,Y)}{V(X).V(Y)}$. Ce rapport est égal au produit des pentes des deux droites de régression a et a'. C'est un indicateur de la valeur de l'angle que forment les deux droites $D_{Y/X}$ et $D_{X/Y}$.

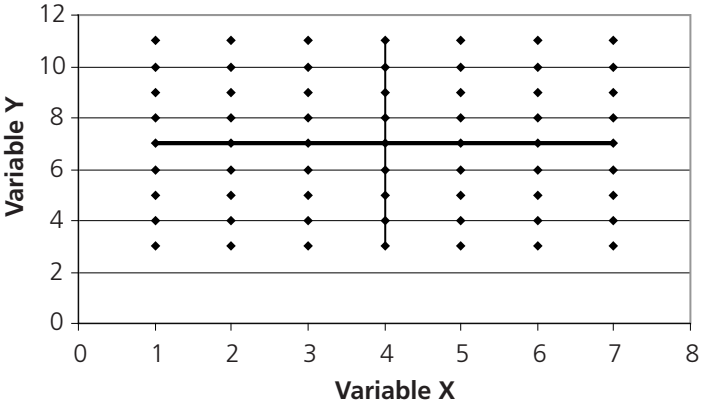
Si aa' = 1, les deux droites de régression $D_{Y/X}$ et $D_{X/Y}$ sont confondues, elles passent par tous les points du nuage qui sont parfaitement alignés.

La liaison linéaire entre les deux variables est dite « totale » ou « fonctionnelle » : l'équation de la droite qui passe en chacun des points est la fonction qui, connaissant une des valeurs de la variable X (respectivement Y), permet de déterminer exactement la valeur de la variable Y (respectivement X) qui lui est associée.

Si aa' = 0, les deux droites de régression sont perpendiculaires : les deux pentes a et a' sont nulles, la droite $D_{Y/X}$ est parallèle à l'axe OX et la droite $D_{X/Y}$ est parallèle à l'axe OY.

On dit que la liaison linéaire est « nulle » ou encore qu'il y a indépendance linéaire.

Le graphique ci-dessous en donne un exemple. La droite $D_{Y/X}$ a pour équation Y = 7 et la droite $D_{X/Y}$ a pour équation X = 4.



Lorsque X = 1, Y peut être égal à 3, 4, 5…, 10 ou 11 ; il en est de même lorsque X est égal à 2, 3, 4, 5, 6 ou 7. Quelle que soit la valeur prise par X, les valeurs possibles pour Y sont les mêmes.

Lorsque Y = 3, X peut être égal à 1, 2, 3, 4, 5, 6 ou 7 ; il en est de même lorsque Y est égal à 4, 5…, 10 ou 11. Quelle que soit la valeur prise par Y, les valeurs possibles pour X sont les mêmes.

Lorsque r^2 est compris entre 0 et 1, la liaison linéaire est dite « relative » ou « partielle ».

Plus r^2 se rapproche de 0, plus l'angle que forme $D_{Y/X}$ et $D_{X/Y}$ est ouvert, plus les deux droites de régression s'éloignent l'une de l'autre, plus la corrélation linéaire entre X et Y est faible.

En revanche, plus r^2 se rapproche de 1, plus l'angle que forme $D_{Y/X}$ et $D_{X/Y}$ est fermé, plus les deux droites de régression sont proches l'une de l'autre, et plus la corrélation linéaire entre X et Y est forte. On voit donc que ***lorsque le nuage de points est très bien ajusté par une droite, la droite de régression de Y en X et la droite de régression de X en Y sont quasiment les mêmes. Il est donc à peu près indifférent de déterminer l'équation de l'une ou de l'autre.***

■ *Coefficient de corrélation linéaire r*

Le ***coefficient de corrélation linéaire*** entre les variables X et Y est le nombre sans dimension, noté r, défini par le rapport $\frac{\text{Cov}(X, Y)}{\sigma(X).\sigma(Y)}$.

$$r = \frac{Cov(X, Y)}{\sigma(X).\sigma(Y)}$$

Le coefficient de corrélation linéaire r est égal à la racine carrée du coefficient de détermination r^2 et il a le signe de la covariance. r^2 étant compris entre 0 et 1, r est compris entre – 1 et + 1.

$$-1 \leq r \leq +1$$

r est positif si la covariance est positive, donc si globalement X et Y varient dans le même sens. r est négatif si la covariance est négative, donc si globalement X et Y varient en sens inverse.

On dit qu'il y a une forte corrélation linéaire entre X et Y (ou une forte dépendance linéaire) lorsque r est voisin de ± 1.

Les calculatrices incluant un programme statistique « régression linéaire » et bien sûr les logiciels statistiques (Excel, SPSS…) évaluent directement, à partir des deux séries de données, le coefficient de corrélation linéaire et le coefficient de détermination.

■ *Applications*

a) Qualité de la régression linéaire entre le nombre de pièces et la surface moyenne

Calculons le coefficient de corrélation linéaire : $r = \frac{Cov(X, Y)}{\sqrt{V(X).V(Y)}}$

Les calculs de Cov(X,Y) et V(X) ont été effectués ci-dessus :

Cov(X,Y) = 76,60 et V(X) = 3,8644

$$V(Y) = \frac{\sum_{i=1}^{n} y_i^2}{n} - \bar{y}^2 = \frac{49635}{6} - 82,17^2 = 1520,5911$$

$$\text{Donc } r = \frac{76,60}{\sqrt{3,8644}\sqrt{1520,5911}} = 0,9993$$

r est très proche de 1 : la liaison linéaire entre le nombre de pièces et la surface moyenne est forte.

Le coefficient de détermination r^2 est égal à 0,9985. Cela signifie que la variance expliquée par la droite de régression de Y en X représente 99,85 % de la variance de Y et la variance résiduelle en représente seulement 0,15 %. La dispersion des points du nuage autour de la droite est donc quasi nulle. L'ajustement de ces points par la droite déterminée par la méthode des moindres carrés est un excellent ajustement, ce que laissait présager la représentation graphique de la droite de régression sur le nuage de points : elle semblait passer très près de tous les points.

b) Qualité de la régression linéaire entre les recettes guichets des salles de cinéma et le temps

$$r = \frac{Cov(T, Y)}{\sqrt{V(T).V(Y)}}$$

Cov(T, Y) et V(T) ont été calculés antérieurement : Cov(T,Y) = 1 040,7333 ; V(T) = 26,6667

$$V(Y) = \frac{\sum_{i=1}^{n} y_i^2}{n} - \bar{y}^2 = \frac{6\,908\,528,\,08}{9} - 851,5333^2 = 42\,505,\,2701$$

$$r = \frac{Cov(T, Y)}{\sqrt{V(T).V(Y)}} = \frac{1\,040,\,7333}{\sqrt{(26,\,6667)(42\,505,\,2701)}} = 0,9775 \text{ et } r^2 = 0,9556.$$

La variance « expliquée » par la droite de régression de Y en T représente 95,56 % de la variance totale de Y et la variance « résiduelle » en représente 4,44 %. Les points du nuage sont donc très peu dispersés autour la droite de régression de Y en T ; elle « résume » bien le nuage de points.

4 Régressions non linéaires

En opérant les changements de variables appropriés, la méthode des moindres carrés permet d'ajuster des nuages de points à l'aide de fonctions non linéaires telles que les fonctions exponentielle et puissance. Ces changements de variables vont être présentés ci-dessous.

Néanmoins, les calculatrices et logiciels statistiques permettent de déterminer directement à partir de la série des données, donc sans effectuer de changements de variables, les fonctions exponentielle ou puissance résultant d'un ajustement par les moindres carrés.

■ *Régression exponentielle*

L'expression d'une fonction exponentielle diffère selon la base dans laquelle elle est exprimée. Les deux bases les plus couramment utilisées sont la base e et la base 10. En base e, une fonction exponentielle est du type $Y = e^{aX+b}$, en base 10, elle peut s'écrire $Y = 10^{aX+b}$, a et b étant des constantes.

En prenant le logarithme népérien des deux membres de l'équation $Y = e^{aX+b}$, on obtient une fonction linéaire ; il en est de même en prenant le logarithme en base 10 des deux membres de l'équation $Y = 10^{aX+b}$:

$\ln Y = \ln e^{aX+b} \Leftrightarrow \ln Y = (aX+ b) \ln e = aX + b$

$\log_{10} Y = \log_{10} 10^{aX+b} \Leftrightarrow \log_{10} Y = (aX+ b) \log_{10} 10 = aX + b$

Soit $Z = \ln Y$ ou $Z = \log_{10} Y$, alors $Z = aX + b$: les coefficients a et b peuvent être déterminés en effectuant une régression linéaire entre Z et X.

La fonction exponentielle est notamment utilisée pour modéliser les phénomènes qui varient dans le temps à un taux à peu près constant. La fonction qui ajuste au mieux ces phénomènes s'écrit :

$$Y = B\,(1+ r)^T$$

Le taux de variation constant est r. En effet, quel que soit T :

$$TV(Y)_{(T+1)/T} = \frac{Y(T+1)}{Y(T)} - 1 = \frac{B(1+r)^{T+1}}{B(1+r)^T} - 1 = (1+r) - 1 = r$$

Or $B(1+r)^T = e^{\ln B} e^{\ln(1+r)^T} = e^{T\ln(1+r)+\ln B}$.

La fonction ci-dessus s'écrit alors : $Y = e^{aT+b}$, où $a = \ln(1 + r)$ et $b = \ln B$. C'est une fonction exponentielle dont a et b vont pouvoir être déterminés par la méthode des moindres carrés.

■ *Régression puissance*

Une fonction puissance est du type $Y = bX^a$, où a et b sont des constantes.

En prenant le logarithme népérien des deux membres de cette égalité (mais l'utilisation du logarithme décimal est possible également), on obtient :

$$\ln Y = \ln bX^a \Leftrightarrow \ln Y = \ln b + a \ln X$$

Posons $\ln Y = Z$, $\ln X = U$ et $\ln b = \beta$; l'équation ci-dessus s'écrit alors :

$$Z = aU + \beta$$

La méthode des moindres carrés permet de déterminer β et a :

$$a = \frac{Cov(U, Z)}{V(U)} \text{ et } \beta = \bar{z} - a\bar{u}$$

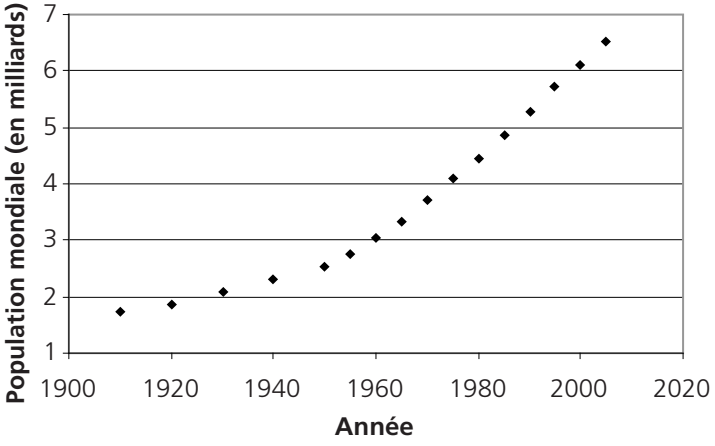
■ *Applications*

a) Régression exponentielle

La population mondiale, en milliards d'habitants, a évolué comme suit entre les années 1910 et 2005 (sources : Bureau du recensement des États-Unis, Historical Estimates of World Population, http://www.census.gov/ipc/www/worldhis.html, et ONU, United Nations Population Division, http://esa.un.org/unpp/):

Année	Population	Année	Population
1910	1,75	1970	3,70
1920	1,86	1975	4,08
1930	2,07	1980	4,45
1940	2,30	1985	4,86
1950	2,54	1990	5,29
1955	2,77	1995	5,72
1960	3,03	2000	6,12
1965	3,34	2005	6,52

La représentation graphique du nuage de points ci-dessous suggère un ajustement exponentiel.



Soient P la population et Y le logarithme (en base e) de P. Soit T la variable associée aux années prenant respectivement les valeurs 0 en 1910, 10 en 1920, 30 en 1940…, 90 en 2000 et 95 en 2005.

t_i	p_i	y_i	t_i^2	y_i^2	$t_i y_i$
0	1,75	0,55962	0	0,31317	0,00000
10	1,86	0,62058	100	0,38512	6,20576
20	2,07	0,72755	400	0,52933	14,55097
30	2,30	0,83291	900	0,69374	24,98727
40	2,54	0,93216	1 600	0,86893	37,28656
45	2,77	1,01885	2 025	1,03805	45,84813
50	3,03	1,10856	2 500	1,22891	55,42813
55	3,34	1,20597	3 025	1,45437	66,32839
60	3,70	1,30833	3 600	1,71173	78,49997
65	4,08	1,40610	4 225	1,97711	91,39630
70	4,45	1,49290	4 900	2,22876	104,50329
75	4,86	1,58104	5 625	2,49968	118,57788
80	5,29	1,66582	6 400	2,77495	133,26546
85	5,72	1,74397	7 225	3,04143	148,23735
90	6,12	1,81156	8 100	3,28176	163,04059
95	6,52	1,87487	9 025	3,51515	178,11307
870		19,89079	59 650	27,54218	1 266,26913

Pour apprécier la qualité de l'ajustement linéaire de Y en T, on évalue le coefficient de corrélation linéaire entre T et Y : $\frac{\text{Cov}(T, Y)}{\sigma(T).\sigma(Y)}$.

$$\text{Cov}(T, Y) = \frac{1}{n}\sum_{i=1}^{n}(t_i - \bar{t})(y_i - \bar{y}) = (\frac{1}{n}\sum_{i=1}^{n} t_i y_i) - \bar{t}\bar{y}$$

$$\bar{t} = \frac{1}{n}\sum_{i=1}^{n} t_i = \frac{1}{16}(870) = 54,375 \quad \text{et} \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{16}(19,89079) = 1,24317$$

D'où $\text{Cov}(T, Y) = (\frac{1}{16}1266,26913) - (54,375).(1,24317) = 11,54445$

En utilisant la formule de Koenig, on calcule l'écart-type de T et l'écart-type de Y :

$$\sigma(T) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} t_i^2 - \bar{t}^2} = \sqrt{(\frac{1}{16}59650) - 54,375^2} = \sqrt{771,48438} = 27,77561$$

$$\sigma(Y) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} y_i^2 - \bar{y}^2} = \sqrt{(\frac{1}{16}27,54218) - 1,24317^2} = \sqrt{0,17591} = 0,41942$$

Donc $r = \frac{11,54445}{(27,77561)(0,41942)} = 0,99097$ et $r^2 = 0,9820$. La variance « expliquée » par la droite de régression de Y en T représente 98,2 % de la variance de Y. La qualité de l'ajustement linéaire de lnP en T par les moindres carrés est excellente. Il en est de même pour l'ajustement exponentiel de P en T.

L'équation de la droite de régression de Y en T déterminée par les moindres carrés est :

$Y = aT + b$ avec $a = \frac{\text{Cov}(T, Y)}{V(T)}$ et $b = \bar{y} - a\bar{t}$.

Les coefficients a et b sont donc les suivants :

$a = \frac{11,54445}{771,48438} = 0,01496$ et $b = 1,24317 - (0,01496)(54,375) = 0,42972$, donc :

$$Y = 0,01496\,T + 0,42972$$

On peut alors déterminer la fonction exponentielle qui ajuste au moins le nuage de points :

$Y = 0,01496\,T + 0,42972 \Leftrightarrow \ln P = 0,01496\,T + 0,42972$

D'où : $P = e^{0,01496T + 0,42972} = e^{0,01496T}.e^{0,42972}$, soit $P = (1,0151)^T (1,5368)$

$P = (1,0151)^T (1,5368) \Leftrightarrow P = (1 + 0,0151)^T (1,5368)$

Cette fonction traduit une évolution de la population mondiale à un taux de variation annuel moyen de 1,5 %.

Remarque : on aurait également pu déterminer le taux de croissance annuel moyen r à partir uniquement des données extrêmes $P_{1910} = 1,75$ et $P_{2005} = 6,52$. Par définition, r vérifie l'égalité :

$$P_{2005} = P_{1910}\,(1 + r)^{95}$$

Soit $(1+r)^{95} = \frac{6,52}{1,75} = 3,7257$ d'où $1 + r = 3,7257^{1/95} = 1,01394$.

Le taux de croissance annuel moyen est ici évalué à environ 1,4 % et non pas 1,5 % comme précédemment. Laquelle des deux évaluations faut-il préférer ? Sans aucun doute la première, car elle tient compte de toutes les observations alors que la deuxième ne prend en compte que les valeurs extrêmes. Si l'une de ces valeurs extrêmes est atypique (sensiblement plus élevée ou plus faible), l'évaluation du taux de variation annuel moyen dans la seconde méthode va être faussée.

L'équation $P = (1{,}0151)^T (1{,}5368)$ permet d'effectuer des prévisions, de déterminer par exemple en quelle année la population mondiale atteindra 10 milliards d'habitants, si la tendance observée se prolonge.

$P = 10 = (1 + 0{,}0151)^T (1{,}5368)$ implique $\ln 10 = 0{,}01496\ T + 0{,}42972$.

D'où :

$$T = \frac{\ln 10 - 0,42972}{0,01496} = 125,19$$

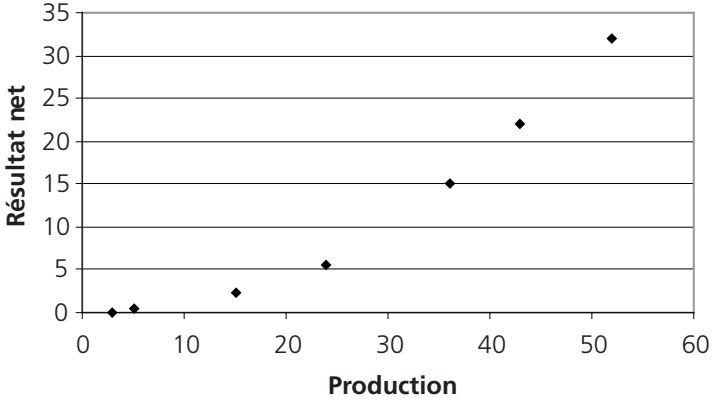
En 1910, T = 0, donc la valeur de T sera légèrement supérieure à 125 dans le courant de l'année 2036. Si la tendance observée se prolonge, la population mondiale atteindra 10 milliards d'habitants au cours de l'année 2036.

b) Régression puissance

Le tableau suivant donne la production annuelle X (en centaines de milliers) et le résultat net Y (en millions d'euros) d'une entreprise, de 2001, date de sa création, à 2007 :

Année	2001	2002	2003	2004	2005	2006	2007
Production	3	5	15	24	36	43	52
Résultat net	0,04	0,4	2,3	5,5	15	22	32

La représentation graphique du nuage de points suggère la possibilité d'un ajustement linéaire, mais aussi d'un ajustement exponentiel ou puissance.



Pour choisir le meilleur ajustement, il faut comparer les coefficients de corrélation linéaire des séries $(x_i ; y_i)$, $(x_i ; \ln y_i)$ et $(\ln x_i ; \ln y_i)$. Ces coefficients peuvent être calculés soit à l'aide du programme statistique approprié de la calculatrice, soit à l'aide d'un tableur.

Le coefficient de corrélation linéaire de la série $(x_i ; y_i)$ est égal à 0,9279, celui de la série $(x_i ; \ln y_i)$ à 0,8525 et celui de la série $(\ln x_i ; \ln y_i)$ à 0,9791. C'est donc à partir de cette dernière série qu'est obtenu le meilleur ajustement linéaire. Par conséquent, c'est une fonction puissance $y = \beta x^{\alpha}$ qui ajuste au mieux le nuage de points.

Soit $Z = \ln Y$ et $U = \ln X$. Déterminons la droite de régression de Z en U : $Z = aU + b$, avec $a = \dfrac{\text{Cov}(U, Z)}{V(U)}$ et $b = \bar{z} - a\bar{u}$.

x_i	y_i	$u_i = \ln x_i$	$z_i = \ln y_i$	$(\ln x_i)^2 = u_i^2$	$(\ln y_i)^2 = z_i^2$	$\ln x_i . \ln y_i = u_i . z_i$
3	0,04	1,09861	– 3,21888	1,20695	10,36116	– 3,53630
5	0,4	1,60944	– 0,91629	2,59029	0,83959	– 1,47471
15	2,3	2,70805	0,83291	7,33354	0,69374	2,25556
24	5,5	3,17805	1,70475	10,10003	2,90617	5,41778
36	15	3,58352	2,70805	12,84161	7,33354	9,70435
43	22	3,76120	3,09104	14,14663	9,55454	11,62603
52	32	3,95124	3,46574	15,61233	12,01133	13,69397
		19,89012	**7,66732**	**63,83136**	**43,70006**	**37,68668**

$$\bar{u} = \frac{1}{n}\sum_{i=1}^{n} u_i = \frac{19,89012}{7} = 2,84145 \quad \text{et} \quad \bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i = \frac{7,66732}{7} = 1,09533$$

$$V(U) = \frac{1}{n}\sum_{i=1}^{n} u_i^2 - \bar{u}^2 = \frac{1}{7}63,83136 - 2,84145^2 = 1,04493$$

$$\text{Cov}(U, Z) = \frac{1}{n}\sum_{i=1}^{n} u_i z_i - \bar{u}\bar{z} = \frac{1}{7}37,68668 - (2,84145)(1,09533) = 2,27149$$

Pente de la droite de régression de Z en U : $a = \dfrac{\text{Cov}(U, Z)}{V(U)} = \dfrac{2,27149}{1,04493} = 2,17382$

Ordonnée à l'origine : $b = \bar{z} - a\bar{u} = 1,09533 - (2,17382)(2,84145) = -5,08147$

Donc Z = 2,17382U –5,08147. On peut alors déterminer l'équation de la fonction puissance qui ajuste au mieux le nuage des points $(x_i ; y_i)$.

$\ln Y = 2,17382 \ln X - 5,08147 \Leftrightarrow e^{\ln y} = e^{\ln x^{2,17382}} e^{-5,08147} = X^{2,17382} 0,00621$

D'où, en arrondissant les coefficients à 10^{-4} près : $Y = X^{2,1738.} 0,0062$

C'est un fonction à élasticité constante : le rapport de la variation relative de Y à la variation relative de X est égal à 2,1738, quelle que soit la valeur de X :

$$e_{Y/X} = \frac{\dfrac{dy}{y}}{\dfrac{dx}{x}} = \frac{\dfrac{dy}{dx}}{\dfrac{y}{x}} = \frac{2,1738x^{1,1738}0,0062}{\dfrac{x^{2,1738}0,0062}{x}} = 2,1738$$

Une augmentation de la production de 1 % entraîne une augmentation du résultat net d'environ 2,17 %.

Les séries chronologiques

CHAPITRE 10

Une série chronologique est une distribution statistique qui décrit l'évolution d'une grandeur au cours du temps. Cette évolution peut être relativement régulière, mais elle peut aussi présenter des variations à la hausse ou à la baisse qui se reproduisent à certaines périodes, voire des variations exceptionnelles liées à des événements ponctuels.

L'étude d'une série chronologique a pour objectif l'analyse de cette évolution en mettant en évidence les différents mouvements qui l'affectent. Elle peut contribuer à élaborer, pour cette série, des prévisions conjoncturelles.

1 Définition d'une série chronologique

Une série chronologique ou chronique est une suite d'observations chiffrées, ordonnées dans le temps et portant sur une même grandeur. En économie et gestion, le temps est repéré le plus souvent en années, trimestres, mois, éventuellement jours.

Les observations elles-mêmes concernent soit un stock, soit un flux. Un flux est une quantité mesurée par unité de temps ; un stock est une quantité mesurée en un point donné du temps.

Les entreprises achètent chaque année des machines, des équipements, pour produire des biens et des services. Le montant de ces achats constitue un flux appelé investissement qui modifie le stock de capital fixe de l'entreprise.

Les naissances sont un flux qui modifient le nombre d'habitants qui, lui, est un stock.

Les demandeurs d'emploi qui s'inscrivent au cours d'un mois pour la première fois forment un flux qui fait varier le stock de demandeurs au 1er de chaque mois.

Une série chronologique peut également être définie comme une distribution à deux variables dont l'une est le temps. Le temps est généralement noté T et la grandeur étudiée Y. La série chronologique est alors formée de n observations :

– les n valeurs prises par le temps sont notées 1, 2…, n ou 0,1, 2…, n – 1 ;

– les n valeurs correspondantes prises par Y sont notées y_1, y_2…, y_n ou y_0, y_1…, y_{n-1}.

Elle est représentée graphiquement en portant en abscisse le temps et en ordonnée les valeurs de Y. Les points du graphe sont le plus souvent reliés par des segments de droite.

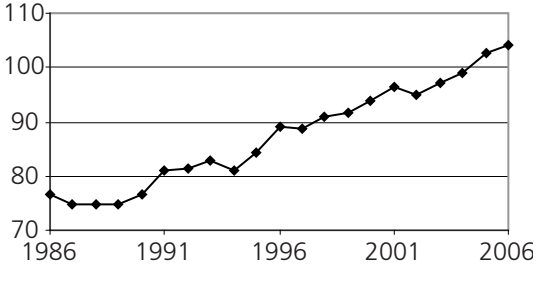
Nous allons voir que la représentation graphique d'une série chronologique permet de mettre en évidence ses composantes.

2 Composantes d'une série chronologique

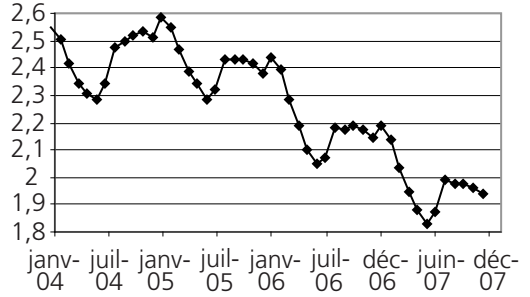
L'étude d'une série chronologique consiste à « décomposer » la série en « composantes » modélisant chacune une partie des variations de la série. Ces composantes sont la tendance et, lorsqu'elles existent, les variations cycliques, les variations saisonnières et les variations résiduelles.

Prenons l'exemple de la production d'énergie (prix chaînés, base 2000), évaluée en milliards d'euros, en France, de 1986 à 2006, représentée par le graphe 1 (source : Insee, comptes nationaux, www.insee.fr) et le nombre de demandes d'emploi en fin de mois de catégorie 1, en France, de 2004 à 2007, représenté par le graphe 2 (source : STMT-DARES, ANPE, www.travail-solidarité.gouv.fr).

Graphe 1

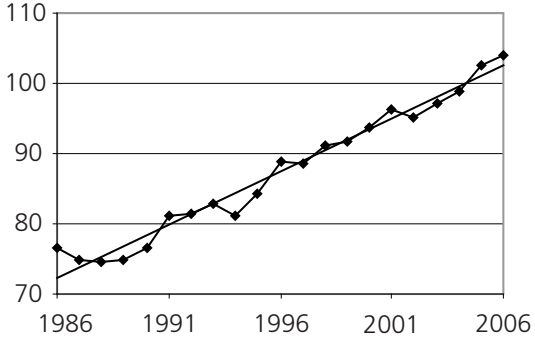


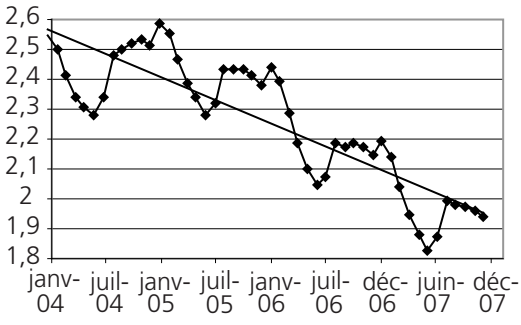
Graphe 2



Deux différences fondamentales apparaissent entre ceux deux courbes : la courbe 1 traduit globalement une augmentation de la quantité produite d'énergie tandis que le graphe 2 fait apparaître une baisse du nombre de demandeurs d'emploi. On dit que la tendance générale est à la hausse dans un cas, à la baisse dans l'autre.

Cette ***tendance*** (également nommée *trend*) est matérialisée par une courbe qui résume le phénomène, qui ajuste l'ensemble des points du nuage de points. Elle lisse la série. Cette courbe est souvent une droite, comme ci-dessous :





Les données du graphe 1 sont annuelles ; elles ne sont donc soumises à aucun phénomène de variations saisonnières. Celles du graphe 2, mensuelles, font apparaître des variations régulières à la hausse ou à la baisse : chaque année, le nombre de demandeurs d'emploi diminue de janvier à juin, il augmente ensuite, puis se stabilise, en décembre il baisse, puis connaît à nouveau une hausse en janvier.

On nomme ***variations saisonnières*** les variations qui résultent de répétitions d'événements plus ou moins réguliers dont les causes peuvent être diverses : fêtes religieuses, coutumes, faits météorologiques, etc. La hausse du nombre de demandeurs d'emploi à partir de juillet s'explique principalement par la fin de l'année scolaire et universitaire.

On appelle période un ensemble de « saisons », définies en fonction des variations observées. Lorsque les données sont mensuelles, chaque mois est une saison et une période peut être, selon le phénomène étudié, une année, un semestre ou un trimestre ; si elles sont trimestrielles, chaque trimestre est une saison et une période peut être une année ou un semestre. Pour des données quotidiennes, une période peut être une semaine.

Pour certaines séries chronologiques dont la période d'observation est longue, certains modèles économiques considèrent une autre composante appelée ***cycle***. Cette composante peut être intégrée à la tendance ; elle ne sera donc pas étudiée indépendamment du *trend*.

Les autres ***variations*** sont dites ***résiduelles***. Ce sont toutes celles qui ne peuvent être expliquées par la tendance et les variations saisonnières. Elles sont liées à des événements imprévisibles dont l'influence sur les valeurs observées peut être plus ou moins prononcée : accidents conjoncturels tels que grèves, incendies, inondations, épidémies, etc., événements qui peuvent être de grande ampleur, ou bien aléas dont l'influence est mineure. Une variation accidentelle de grande ampleur

se traduit graphiquement par un « pic » ou un « creux » à caractère exceptionnel qui rompt la régularité de l'évolution de l'ensemble. Aucune des deux courbes ci-dessus ne fait apparaître de variations accidentelles conjoncturelles.

Lorsqu'une série de données brutes présente une ou plusieurs variation(s) accidentelle(s) prononcée(s), avant d'analyser la série, il est nécessaire soit d'enlever chacune des données correspondantes, soit de les remplacer par des données vraisemblables. Sinon, la détermination de la tendance et celle du mouvement saisonnier seront faussées.

L'analyse d'une série chronologique repose sur l'étude de ses composantes. Pour effectuer cette analyse, on se réfère à un modèle théorique.

3 Modèles théoriques d'analyse des séries chronologiques

■ *Variables G, S et R*

Dans les modèles théoriques que nous allons définir, la variable associée aux données brutes Y est fonction de trois variables :

– une ***variable G associée à la composante tendancielle*** ;

– une ***variable S associée à la composante saisonnière*** ;

– une ***variable R associée à la composante résiduelle***.

Les résidus représentent les écarts entre la série reconstituée avec les composantes G et S et la série réellement observée.

Nous allons étudier deux modèles : le modèle additif et le modèle multiplicatif.

■ *Modèle additif*

Le modèle additif s'applique à une variable dont l'amplitude du mouvement saisonnier est constante ; les composantes G, S et R sont alors indépendantes les unes des autres.

Le modèle additif s'écrit : $Y = G + S + R$.

Il permet de décomposer chaque donnée brute y_t en une somme de trois valeurs :

– g_t, la valeur de la tendance générale à la date t ;

– s_t, la valeur du coefficient saisonnier à la date t ;

– r_t, la valeur résiduelle à la date t.

$$y_t = g_t + s_t + r_t$$

Modèle multiplicatif

Le modèle multiplicatif s'applique à une variable dont l'amplitude du mouvement saisonnier croît ou décroît avec la tendance. Dans ce cas, les composantes sont dépendantes, proportionnelles les unes par rapport aux autres.

Le modèle s'écrit : $Y = G \,.\, S \,.\, R$

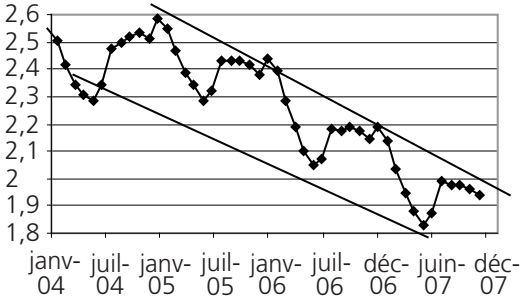
Il permet de décomposer chaque donnée brute en un produit de trois valeurs g_t, s_t et r_t.

$$y_t = g_t \,.\, s_t \,.\, r_t$$

Choix du modèle

Comment, en pratique, choisir le modèle théorique qui va être utilisé pour décomposer la série ?

Une méthode graphique simple peut permettre de faire ce choix. Elle consiste à tracer une droite qui passe le plus près possible des minima et une droite qui passe le plus près possible des maxima et à regarder si l'on obtient deux droites à peu près parallèles ou non. Si oui, comme ci-dessous, cela signifie que l'amplitude des variations saisonnières est à peu près constante ; on choisit alors le modèle additif. Sinon, on opte pour le modèle multiplicatif.



Il existe également une méthode analytique, la méthode de Buys et Ballot.

Cette méthode consiste, à partir de la série des données brutes, à calculer pour chaque période la moyenne et l'écart-type des données. Si les écart-types sont approximativement constants d'une période à l'autre, le modèle est additif, sinon il est multiplicatif.

■ *Applications*

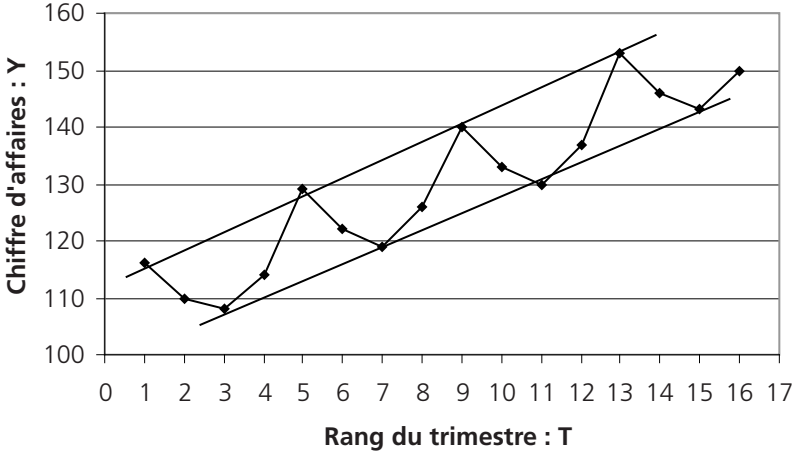
a) Modèle additif

Le chiffre d'affaires trimestriel d'une entreprise (en milliers d'euros) a évolué de la manière suivante de 2004 à 2007 :

Année	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4
2004	116	110	108	114
2005	129	122	119	126
2006	140	133	130	137
2007	153	146	143	150

Soit T le rang du trimestre. Les valeurs de T vont de 1 (1er trimestre de l'année 2004) à 16 (4e trimestre de l'année 2007).

Si l'on trace une droite qui passe le plus près possible des minima et une droite qui passe le plus près possible des maxima, on obtient deux droites qui semblent à peu près parallèles. La représentation graphique suggère donc un modèle additif.



Est-ce que la méthode de Buys et Ballot le confirme ? Calculons les moyennes et les écarts-types du chiffres d'affaires de chaque année. En 2004 :

Moyenne du chiffre d'affaires : $\bar{y} = \frac{\sum_{i=1}^{4} y_i}{4} = \frac{116 + 110 + 108 + 114}{4} = 112$

Variance de Y : $V(Y) = \frac{\sum_{i=1}^{4} y_i^2}{4} - \bar{y}^2 = \frac{116^2 + 110^2 + 108^2 + 114^2}{4} - 112^2 = 10$ d'où $\sigma(Y) = 3{,}162$.

On calcule de même les moyennes et les écarts-types des années suivantes. Les résultats figurent dans le tableau ci-dessous :

	2004	2005	2006	2007
Moyenne	112	124	135	148
Écart-type	3,162	3,808	3,808	3,808

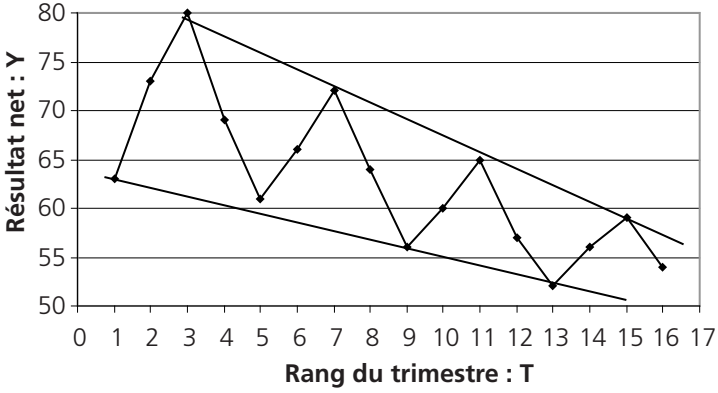
L'écart-type est quasiment stable d'une année à l'autre. C'est donc bien le modèle additif qu'il faut choisir.

b) Modèle multiplicatif

Le résultat net trimestriel (en milliers d'euros) d'un restaurant situé dans une station balnéaire a évolué comme suit au cours de la période 2004-2007 :

Année	1er trimestre	2e trimestre	3e trimestre	4e trimestre
2004	63	73	80	69
2005	61	66	72	64
2006	56	60	65	57
2007	52	56	59	54

En portant en abscisse le rang du trimestre, la représentation graphique de cette série est la suivante :



Les écarts entre les points les plus bas et les points les plus hauts diminuent : l'amplitude du mouvement saisonnier décroît. Le modèle est multiplicatif.

Calculons les écarts-types du résultat net :

	2004	2005	2006	2007
Moyenne	71,25	65,75	59,5	55,25
Écart-type	6,180	4,023	3,500	2,586

L'écart-type du résultat net décroît nettement chaque année : la méthode de Buys et Ballot confirme le modèle multiplicatif.

4 Méthodes de détermination de la tendance

La détermination de la tendance se fait par « lissage » des irrégularités. Trois méthodes peuvent être utilisées : la méthode des moyennes échelonnées, la méthode des moyennes mobiles et la méthode des moindres carrés. Les deux premières méthodes sont des méthodes empiriques et la troisième est une méthode analytique.

■ *Méthode des moyennes échelonnées*

Soient y_1, y_2…,y_n la série des données brutes et k le nombre de saisons de cette série. Par exemple, si les données sont mensuelles k = 12, si elles sont trimestrielles k = 4.

La 1re valeur de la tendance est la moyenne arithmétique des k premières données brutes :

$$g_1 = \frac{1}{k}(y_1 + y_2 + \ldots + y_k)$$

La 2e valeur est la moyenne arithmétique des k données brutes suivantes :

$$g_2 = \frac{1}{k}(y_{k+1} + y_{k+2} + \ldots + y_{2k})$$

La 3e valeur est la moyenne arithmétique des k données brutes suivantes :

$$g_3 = \frac{1}{k}(y_{2k+1} + y_{2k+2} + \ldots + y_{3k})$$

et ainsi de suite.

La valeur de k est choisie égale au nombre de saisons de chaque période parce que l'objectif est d'éliminer la composante saisonnière. La moyenne obtenue sur une période ne subit pas l'influence des variations saisonnières.

Cette méthode est simple à mettre en œuvre, mais elle a l'inconvénient de trop simplifier, trop « réduire » la réalité, et ce d'autant plus que la valeur de k est grande. Si les données sont trimestrielles, quatre données successives sont remplacées par une seule ; pour une série portant sur six années, la série des valeurs de G comprendra seulement six valeurs alors que la série des données brutes en contient 24. Pour des données mensuelles portant aussi sur six ans, on passe de 72 données à 6 moyennes annuelles.

En poussant la méthode à l'extrême, on peut diviser le nuage de points en deux sous-nuages dont on détermine la valeur moyenne de y. On détermine ensuite l'équation de la droite qui passe par ces deux points. Cette méthode est appelée méthode de Mayer.

■ *Méthode des moyennes mobiles centrées*

Soient $y_1, y_2 \ldots, y_n$ la série des données brutes et k le nombre de saisons de cette série.

La moyenne mobile centrée d'ordre k, à la date t, est le nombre g_t défini par :

– si k est impair (k = 2p + 1) : $g_t = \frac{1}{2p+1}(y_{t-p} + \ldots + y_{t-1} + y_t + y_{t+1} + \ldots + y_{t+p})$;

– si k est pair (k = 2p) : $g_t = \frac{1}{2p}(\frac{y_{t-p}}{2} + y_{t-p+1} + \ldots + y_{t-1} + y_t + y_{t+1} + \ldots + y_{t+p-1} + \frac{y_{t+p}}{2})$.

C'est donc la moyenne arithmétique de y_t et des valeurs qui l'encadrent, le nombre de ces valeurs étant déterminé par la valeur de k.

Selon la valeur de k, le calcul de la moyenne mobile centrée d'ordre k n'est pas possible pour certaines valeurs de t. Ainsi, si k = 3, il faut disposer d'une donnée avant y_t et d'une donnée après pour calculer la moyenne mobile ; ce calcul n'est donc pas possible pour t = 1 et t = n ; si k = 4, il faut disposer de deux données avant y_t et deux données après, le calcul de la moyenne mobile n'est donc pas possible pour t = 1, t = 2, t = n – 1 et t = n.

Cette méthode a un effet de lissage d'autant plus fort que la valeur de k est grande.

■ *Méthode des moindres carrés*

La tendance d'une série chronologique déterminée par un ajustement affine des données brutes s'écrit : G = aT + b, les coefficients a et b vérifiant les égalités suivantes :

$$a = \frac{Cov(T, Y)}{V(T)} \text{ et } b = \bar{y} - a\bar{t}$$

Cette équation permet d'associer à chaque valeur de t une valeur de la tendance, notée g_t, telle que : $g_t = at + b$.

Les trois méthodes de détermination de la tendance présentées ici sont toutes satisfaisantes pour assurer le lissage des données et donc déterminer la tendance. En revanche, lorsqu'on désire aussi déterminer la composante saisonnière, la méthode des moyennes échelonnées fournit trop peu de valeurs de G pour déterminer les coefficients saisonniers. Lorsqu'on souhaite aussi effectuer des prévisions à partir des composantes tendancielles et saisonnières déterminées, il faut que la tendance ait été évalué par une méthode analytique telle que la méthode des moindres carrés, sinon il n'est pas possible de prévoir une valeur pour la tendance.

■ *Application*

Le tableau suivant reprend la série des chiffres d'affaires trimestriels d'une entreprise (en milliers d'euros) de 2004 à 2007 :

Année	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4
2004	116	110	108	114
2005	129	122	119	126
2006	140	133	130	137
2007	153	146	143	150

a) Moyennes échelonnées d'ordre 4

Chaque moyenne échelonnée est la moyenne arithmétique des quatre données trimestrielles de l'année.

Année	Moyenne échelonnée
2004	$g_1 = (116 + 110 + 108 + 114)/4 = 112$
2005	$g_2 = (129 + 122 + 119 + 126)/4 = 124$
2006	$g_3 = (140 + 133 + 130 + 137)/4 = 135$
2007	$g_4 = (153 + 146 + 143 + 150)/4 = 148$

b) Moyennes mobiles centrées d'ordre 4

La moyenne mobile centrée d'ordre 4 associée à t = 3 est égale à :

$$\frac{\frac{y_1}{2} + y_2 + y_3 + y_4 + \frac{y_5}{2}}{4} = \frac{\frac{116}{2} + 110 + 108 + 114 + \frac{129}{2}}{4} = 113,625$$

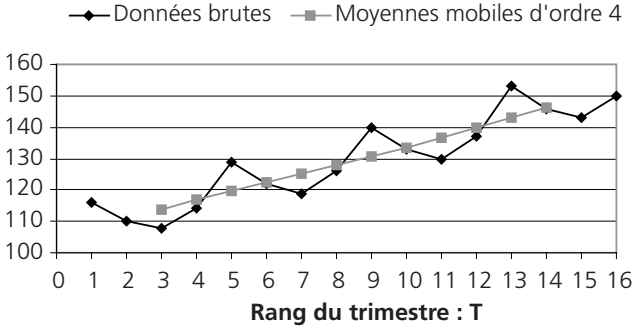
La moyenne mobile centrée d'ordre 4 associée à t = 4 est égale à :

$$\frac{\frac{y_2}{2} + y_3 + y_4 + y_5 + \frac{y_6}{2}}{4} = \frac{\frac{110}{2} + 108 + 114 + 129 + \frac{122}{2}}{4} = 116,75$$

On procède ainsi successivement jusqu'à t = 14.

t	Série brute	Moyenne mobile centrée d'ordre 4
1	116	
2	110	
3	108	113,625
4	114	116,75
5	129	119,625
6	122	122,5
7	119	125,375
8	126	128,125
9	140	130,875
10	133	133,625
11	130	136,625
12	137	139,875
13	153	143,125
14	146	146,375
15	143	
16	150	

La représentation graphique ci-dessous montre le lissage réalisé par la courbe des moyennes mobiles.



c) Méthode des moindres carrés

Soit G = aT + b l'équation de la tendance déterminée par un ajustement affine de Y en T.

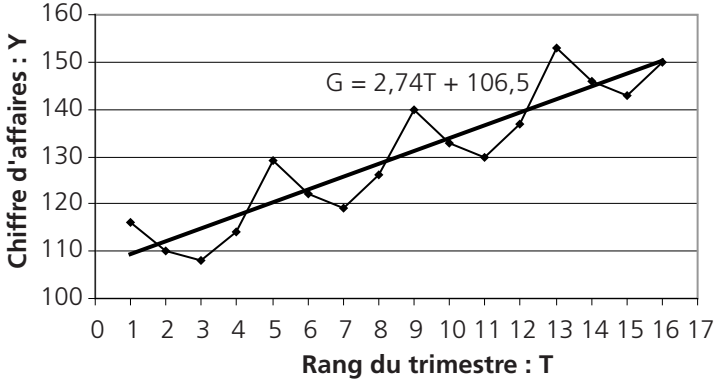
Les coefficients a et b vérifient : $a = \frac{Cov(T, Y)}{V(T)}$ et $b = \bar{y} - a\bar{t}$

$\bar{t} = \frac{1}{n}\sum_{i=1}^{n} t_i = \frac{1}{16}(136) = 8,5$ et $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{16}(2076) = 129,75$

$V(T) = \frac{\sum_{i=1}^{n} t_i^2}{n} - \bar{t}^2 = \frac{1496}{16} - 8,5^2 = 21,25$

$Cov(T, Y) = (\frac{1}{n}\sum_{i=1}^{n} t_i y_i) - \bar{t}\bar{y} = \frac{1}{16}18576 - (8,5)(129,75) = 58,125$

D'où : $a = \frac{58,125}{21,25} = 2,735$ et $b = 129,75 - (2,735)(8,5) = 106,50$. En arrondissant les valeurs de a et b à 10^{-2}, l'équation de la tendance s'écrit : G = 2,74 T + 106,50.



5 Méthodes de détermination de la composante saisonnière

■ *Principes fondamentaux*

L'objectif est de déterminer la composante saisonnière S qui correspond pour chaque « saison » à un coefficient qui mesure l'influence saisonnière correspondante. Ce coefficient est appelé coefficient saisonnier, et noté s_p, p étant le nombre de saisons. Deux principes fondamentaux sont à la base de la détermination des coefficients saisonniers dans un modèle théorique.

Le premier est la ***répétition à l'identique*** : on suppose que, dans un modèle théorique, toute variation saisonnière se répète identiquement à chaque période, une période étant un ensemble de saisons. Si le nombre de saisons d'une série chronologique est p, on déterminera donc seulement p coefficients saisonniers. Par exemple, pour des données trimestrielles, pendant une période, c'est-à-dire une année, il y a quatre saisons ; on déterminera donc quatre coefficients saisonniers ; pour des données mensuelles, la période est également l'année, il y a douze saisons donc on déterminera douze coefficients saisonniers.

Le deuxième principe est la ***neutralité de l'influence des variations saisonnières sur chaque période***.

Dans le cas du modèle additif, on a vu que chaque donnée brute y_t s'écrit comme la somme de trois composantes : $y_t = g_t + s_t + r_t$. Pour que l'influence des variations saisonnières soit neutre sur une période, il est donc nécessaire que, pour cette période, la valeur moyenne de s_t soit nulle. Il faut alors que la moyenne des coefficients $s_1, s_2 \ldots s_p$ soit nulle : $\frac{1}{p}\sum_{j=1}^{p} s_j = 0$.

Dans le cas du modèle multiplicatif, chaque donnée brute s'écrit comme le produit de trois composantes : $y_t = g_t \cdot s_t \cdot r_t$. Pour que l'influence des variations saisonnières soit neutre sur une période, il faut donc que pour cette période la valeur moyenne de s_t soit égale à 1. Par conséquent, la moyenne des coefficients $s_1, s_2 \ldots s_p$ doit être égale à 1 : $\frac{1}{p}\sum_{j=1}^{p} s_j = 1$.

■ *Calcul des coefficients saisonniers dans le cas du modèle additif*

Ce calcul s'effectue en trois étapes.

– ***1re étape*** : on calcule, pour chaque date t, l'***écart saisonnier*** es_t (ou écart à la tendance). Cet écart est égal à la différence entre la donnée brute y_t et la valeur de la tendance g_t déterminée par l'une des trois méthodes vues précédemment.

$$es_t = y_t - g_t$$

Ces écarts sont positifs si $y_t > g_t$, négatifs si $y_t < g_t$;

– ***2e étape*** : on calcule, pour chaque saison j, la ***moyenne arithmétique des écarts saisonniers*** correspondant à cette saison et on note s_j cette moyenne. C'est le ***coefficient saisonnier*** de la saison j ;

– ***3e étape*** : on s'assure que le principe de neutralité de l'influence des variations saisonnières sur chaque période est bien vérifié, c'est-à-dire que $\frac{1}{p}\sum_{j=1}^{p} s_j = 0$. En pratique, le plus souvent cette moyenne est approximativement nulle mais pas exactement nulle. Il faut alors calculer des coefficients corrigés notés s_j^*, tels que la somme de ces coefficients soit bien égale à 0. Pour cela, il suffit de soustraire à chaque coefficient s_j la valeur moyenne des s_j. Si $\frac{1}{p}\sum_{j=1}^{p} s_j = \lambda$, alors $s_j^* = s_j - \lambda$ et $\sum_{j=1}^{p} s_j^* = 0$.

■ *Calcul des coefficients saisonniers dans le cas du modèle multiplicatif*

Ce calcul s'effectue en trois étapes :

– ***1re étape*** : on calcule, pour chaque date t, le ***rapport saisonnier*** rs_t, ou rapport à la tendance. Ce rapport s'obtient en divisant la donnée brute y_t par la valeur de la tendance g_t déterminée par l'une des trois méthodes vues précédemment.

$$rs_t = y_t / g_t$$

Ces écarts sont supérieurs à 1 si $y_t > g_t$, inférieurs à 1si $y_t < g_t$.

Il est recommandé de les calculer avec une précision d'au moins trois chiffres après la virgule ;

– ***2e étape*** : on calcule, pour chaque saison j, la ***moyenne arithmétique des rapports saisonniers*** correspondant à cette saison et on note s_j cette moyenne. C'est le ***coefficient saisonnier*** de la saison j.

– ***3e étape*** : on s'assure que le principe de neutralité de l'influence des variations saisonnières sur chaque période est bien vérifié, c'est-à-dire que $\frac{1}{p}\sum_{j=1}^{p} s_j = 1$. En pratique, le plus souvent cette

moyenne est approximativement égale à 1 mais pas exactement égale à 1. Il faut alors calculer des coefficients corrigés notés s_j^*, tels que la somme de ces coefficients soit bien égale à p et donc leur moyenne égale à 1. Pour cela, il suffit de diviser chaque coefficient s_j par la valeur moyenne des s_j.

Si $\frac{1}{p}\sum_{j=1}^{p} s_j = \lambda$, alors $s_j^* = s_j / \lambda$ et $\frac{1}{p}\sum_{j=1}^{p} s_j^* = 1$.

■ *Applications*

a) Calcul des coefficients saisonniers, modèle additif

Les données sont celles de la série des chiffres d'affaires trimestriels de 2004 à 2007.

Le calcul des coefficients saisonniers nécessite d'abord l'évaluation, pour chacune des valeurs de T, des valeurs de la tendance. Ici nous choisissons d'utiliser la méthode des moindres carrés pour exprimer la tendance.

Le calcul effectué antérieurement a donné : $g_t = 2{,}74t + 106{,}50$.

On aurait aussi pu choisir la série des moyennes mobiles d'ordre 4. En revanche, la série des moyennes échelonnées fournit trop peu de valeurs de g_t pour être utilisée.

Pour chaque trimestre, le tableau ci-dessous reprend la donnée brute et donne la valeur de la tendance et de l'écart saisonnier es_t.

Année	t	y_t	$g_t = 2{,}74t + 106{,}50$	$es_t = y_t - g_t$
2004	1	116	109,24	6,76
	2	110	111,98	– 1,98
	3	108	114,72	– 6,72
	4	114	117,46	– 3,46
2005	5	129	120,20	8,80
	6	122	122,94	– 0,94
	7	119	125,68	– 6,68
	8	126	128,42	– 2,42
2006	9	140	131,16	8,84
	10	133	133,90	– 0,90
	11	130	136,64	– 6,64
	12	137	139,38	– 2,38
2007	13	153	142,12	10,88
	14	146	144,86	1,14
	15	143	147,60	– 4,60
	16	150	150,34	– 0,34

Chaque coefficient s_j est égal à la moyenne des écarts saisonniers du trimestre correspondant :

$s_1 = (es_1 + es_5 + es_9 + es_{13}) / 4 = (6,76 + 8,80 + 8,84 + 10,88) / 4 = 8,82$

$s_2 = (es_2 + es_6 + es_{10} + es_{14}) / 4 = (-1,98 - 0,94 - 0,90 + 1,14) / 4 = -0,67$

$s_3 = (es_3 + es_7 + es_{11} + es_{15}) / 4 = (-6,72 - 6,68 - 6,64 - 4,6) / 4 = -6,16$

$s_4 = (es_4 + es_8 + es_{12} + es_{16}) / 4 = (-3,46 - 2,42 - 2,38 - 0,34) / 4 = -2,15$

La somme de ces coefficients n'est pas nulle mais égale à – 0,16. Il faut donc les corriger en soustrayant à chacun d'eux la moyenne de leur somme, soit – 0,04. On obtient :

$s_1^* = 8,86 \quad s_2^* = -0,63 \quad s_3^* = -6,12 \quad s_4^* = -2,11$

La valeur positive ou négative des coefficients saisonniers nous renseigne sur la position des données brutes par rapport à la tendance. Le coefficient s_1^* est nettement supérieur à 0 : la série des données brutes est au dessus de la tendance au 1er trimestre ; le coefficient s_2^* est légèrement

inférieur à 0 : les données brutes sont plutôt au dessous de la tendance (sauf en 2007) et proches de la tendance ; etc.

b) Calcul des coefficients saisonniers, modèle multiplicatif

Les données sont les résultats nets trimestriels (en milliers d'euros) du restaurant situé dans une station balnéaire, de 2004 à 2007.

Pour calculer les coefficients saisonniers, il faut d'abord déterminer les valeurs de la tendance. Elles sont évaluées dans le tableau ci-dessous par des moyennes mobiles centrées d'ordre 4. Les rapports saisonniers rs_t sont calculés dans la dernière colonne.

t	y_t	g_t	$rs_t = y_t/g_t$
1	63		
2	73		
3	80	71,00	1,127
4	69	69,88	0,987
5	61	68,00	0,897
6	66	66,38	0,994
7	72	65,13	1,106
8	64	63,75	1,004
9	56	62,13	0,901
10	60	60,38	0,994
11	65	59,00	1,102
12	57	58,00	0,983
13	52	56,75	0,916
14	56	55,63	1,007
15	59		
16	54		

Chaque coefficient s_j est égal à la moyenne des écarts saisonniers du trimestre correspondant.

$s_1 = (0,897 + 0,901 + 0,916)/3 = 0,905$; $s_2 = (0,994 + 0,994 + 1,007)/3 = 0,998$

$s_3 = (1,127 + 1,106 + 1,102)/3 = 1,112$; $s_4 = (0,987 + 1,004 + 0,983)/3 = 0,991$

La somme de ces coefficients est égale à 4,006. Leur moyenne est égale à 1,0015, donc légèrement différente de 1. Il faut corriger ces coefficients en divisant chacun d'eux par cette moyenne. On obtient : $s_1^* = 0{,}904$ $s_2^* = 0{,}996$ $s_3^* = 1{,}110$ $s_4^* = 0{,}990$.

6 Série désaisonnalisée et série ajustée

■ *Définitions*

La ***série désaisonnalisée*** ou ***série corrigée des variations saisonnières (CVS)*** est la série obtenue à partir de la série brute après élimination de la composante saisonnière. Elle exprime ce qu'aurait été la réalité du phénomène étudié s'il n'y avait pas eu de variations saisonnières. En pratique, les économistes utilisent les séries CVS dans de nombreux domaines, notamment en analyse conjoncturelle : pour éviter des erreurs d'interprétation des évolutions d'un mois à un autre ou d'un trimestre à un autre, les séries sont toutes CVS. Il y a même des séries qui sont aussi corrigées des jours ouvrables (CVS-CJO) car il peut y avoir des variations importantes des valeurs de certaines variables dues au nombre de jours ouvrables du même nom (lundi, mardi…) dans un mois. Par exemple, le chiffre d'affaires des supermarchés est plus élevé dans les mois ayant cinq samedis que dans les mois en ayant quatre ; inversement, la production de l'industrie est plus faible les mois ayant cinq samedis que les mois en ayant quatre. Il faut donc corriger les séries correspondantes pour prendre en compte ces phénomènes.

La ***série ajustée*** est la série obtenue à partir de la tendance générale en intégrant la composante saisonnière. Elle exprime l'évolution qu'aurait connue la variable si le mouvement saisonnier avait été parfaitement régulier de période en période. Lorsque la tendance a été déterminée par la méthode des moindres carrés, la série ajustée est utilisée pour effectuer des prévisions car c'est elle qui ajuste le mieux le nuage de points en tenant compte des variations saisonnières.

■ *Cas du modèle additif $Y = G + S + R$*

Série CVS : dans ce modèle, on élimine les variations saisonnières en soustrayant S à Y. Soit ***Y^{cvs} la variable associée à la série CVS***. Alors, $Y^{cvs} = Y - S$.

En pratique, pour déterminer la série CVS, il faut calculer les y_t^{cvs} tels que :

$$y_t^{cvs} = y_t - s_j, \text{ j étant la saison associée à la date t}$$

Série ajustée : pour restituer au mieux les variations de la variable, il faut prendre en compte les variations saisonnières de cette variable, donc dans ce modèle additionner la composante saisonnière à la tendance.

Soit ***Y^{aj} la variable associée à la série ajustée***. Alors : $Y^{aj} = G + S$.

En pratique, pour déterminer la série ajustée, il faut calculer les y_t^{aj} tels que :

$$y_t^{aj} = g_t + s_j, \text{ j étant la saison associée à la date t}$$

Des valeurs de g_t et s_j se déduisent les valeurs résiduelles :

$$r_t = y_t - (g_t + s_j) = y_t - y_t^{aj}$$

Ce sont les écarts entre les valeurs observées et les valeurs de la série reconstituée avec G et S. Ils sont le plus souvent non nuls, mais proches de 0, positifs ou négatifs.

■ *Cas du modèle multiplicatif $Y = G \,.\, S \,.\, R$*

Série CVS : dans ce modèle, pour éliminer les variations saisonnières, il faut diviser Y par S. Soit ***Y^{cvs} la variable associée à la série CVS***. Alors $Y^{cvs} = Y / S$

En pratique, pour déterminer la série CVS, il faut calculer les y_t^{cvs} tels que :

$$y_t^{cvs} = y_t / s_j \text{ , j étant la saison associée à la date t}$$

Série ajustée : pour restituer au mieux les variations de la variable dans un modèle multiplicatif, il faut multiplier la composante saisonnière par la tendance.

Soit ***Y^{aj} la variable associée à la série ajustée***. Alors $Y^{aj} = G \,.\, S$.

En pratique, pour déterminer la série ajustée, il faut calculer les y_t^{aj} tels que :

$$y_t^{aj} = g_t \,.\, s_j \text{ , j étant la saison associée à la date t}$$

Des valeurs de g_t et s_j se déduisent les valeurs résiduelles :

$$r_t = (y_t / (g_t / s_j)) = y_t / y_t^{aj}$$

Elles sont le plus souvent légèrement supérieures ou légèrement inférieures à 1, éventuellement égales à 1.

■ *Applications*

a) Série CVS et série ajustée, modèle additif

Les données sont celles de la série des chiffres d'affaires trimestriels de 2004 à 2007 :

Année	t	y_t	g_t	s_j*	$y_t^{cvs} = y_t - s_j$*	$y_t^{aj} = g_t + s_j$*	$r_t = y_t - y_t^{aj}$
2004	1	116	109,24	8,86	107,14	118,10	– 2,10
	2	110	111,98	– 0,63	110,63	111,35	– 1,35
	3	108	114,72	– 6,12	114,12	108,60	– 0,60
	4	114	117,46	– 2,11	116,11	115,35	– 1,35
2005	5	129	120,20	8,86	120,14	129,06	– 0,06
	6	122	122,94	– 0,63	122,63	122,31	– 0,31
	7	119	125,68	– 6,12	125,12	119,56	– 0,56
	8	126	128,42	– 2,11	128,11	126,31	– 0,31
2006	9	140	131,16	8,86	131,14	140,02	– 0,02
	10	133	133,90	– 0,63	133,63	133,27	– 0,27
	11	130	136,64	– 6,12	136,12	130,52	– 0,52
	12	137	139,38	– 2,11	139,11	137,27	– 0,27
2007	13	153	142,12	8,86	144,14	150,98	2,02
	14	146	144,86	– 0,63	146,63	144,23	1,77
	15	143	147,60	– 6,12	149,12	141,48	1,52
	16	150	150,34	– 2,11	152,11	148,23	1,77

Série CVS : s'il n'y avait pas eu de variations saisonnières, le chiffre d'affaires aurait été égal à 107,14 au 1er trimestre 2004, 110,63 au 2e trimestre…, 152,11 au 4e trimestre 2007.

Série ajustée et résidus : la série reconstituée à l'aide de la tendance et des coefficients saisonniers donne pour le 1er trimestre 2004 un chiffre d'affaires de 118,10, donc supérieur de 2,10 au chiffre observé, pour le 2e trimestre un chiffre d'affaires de 111,35, donc supérieur de 1,35 au chiffre observé, etc. Le modèle ne traduit donc pas exactement l'évolution de la variable observée. C'est évidemment le cas le plus fréquent.

La tendance et les coefficients peuvent néanmoins être utilisés pour prévoir le chiffre d'affaires de l'entreprise. L'équation de la tendance déterminée précédemment par un ajustement affine est $g_t = 2{,}74t + 106{,}50$. On peut donc prévoir à la date t un chiffre d'affaires égal : $g_t + s_j^* = (2{,}74t + 106{,}50) + s_j^*$.

Par exemple, au troisième trimestre 2009, soit t = 23 :

$$g_{23} + s_3^* = [2{,}74(23) + 106{,}50] + (-6{,}12) = 163{,}40$$

Si la tendance observée antérieurement se poursuit, si les variations saisonnières continuent de se produire à l'identique, et bien sûr si aucun accident ne perturbe l'activité de l'entreprise, son chiffre d'affaires devrait être proche de 163,40 milliers d'euros au 3e trimestre 2009.

b) Série CVS et série ajustée, modèle multiplicatif

Les données sont les résultats nets trimestriels (en milliers d'euros) du restaurant situé dans une station balnéaire, de 2004 à 2007 :

Année	t	y_t	g_t	s_j^*	$y_t^{cvs} = y_t / s_j^*$	$y_t^{aj} = g_t \cdot s_j^*$	$r_t = y_t / y_t^{aj}$
2004	1	63		0,904	69,69		
	2	73		0,996	73,29		
	3	80	71,00	1,110	72,07	78,81	1,02
	4	69	69,88	0,990	69,70	69,18	1,00
2005	5	61	68,00	0,904	67,48	61,47	0,99
	6	66	66,38	0,996	66,27	66,11	1,00
	7	72	65,13	1,110	64,86	72,29	1,00
	8	64	63,75	0,990	64,65	63,11	1,01
2006	9	56	62,13	0,904	61,95	56,16	1,00
	10	60	60,38	0,996	60,24	60,13	1,00
	11	65	59,00	1,110	58,56	65,49	0,99
	12	57	58,00	0,990	57,58	57,42	0,99
2007	13	52	56,75	0,904	57,52	51,30	1,01
	14	56	55,63	0,996	56,22	55,40	1,01
	15	59		1,110	53,15		
	16	54		0,990	54,55		

Série CVS : s'il n'y avait pas eu de variations saisonnières, le résultat net du restaurant aurait été égal à 69,69 au 1er trimestre 2004, 73,29 au 2e trimestre…, 54,55 au 4e trimestre 2007.

Série ajustée et résidus : la série reconstituée à l'aide de la tendance et des coefficients saisonniers donne pour le 3e trimestre 2004 un résultat net de 78,81, égal à 1,02 fois le résultat observé, pour le 4e trimestre un résultat net 69,18 légèrement supérieur au résultat observé (69), etc. Comme dans la 1re application, le modèle ne restitue donc pas exactement l'évolution de la variable observée.

Dans cette application, la tendance n'a pas été déterminée par un méthode analytique mais empirique. Il n'est donc pas possible d'évaluer la tendance pour une date future, et d'effectuer des prévisions de résultat net.

Dans la collection Les Carrés

Droit constitutionnel et Institutions politiques

- L'essentiel du droit constitutionnel – T. 1 Théorie générale du droit constitutionnel (G. Champagne) – 7e édition – 2008
- L'essentiel du droit constitutionnel – T. 2 Les institutions de la Ve République (G. Champagne) – 8e édition – 2008
- L'essentiel de l'Histoire constitutionnelle et politique de la France (de 1789 à nos jours) (J.-C. Zarka) – 4e édition – 2008
- L'essentiel du droit des politiques sociales (E. Aubin) – 3e édition – 2008
- L'essentiel du régime juridique des droits et libertés en France (G. Armand) – 2007
- L'essentiel de l'introduction à la vie politique (E. Aubin) – 2e édition 2007
- L'essentiel de la sociologie politique (J.-P. Lecomte) – 2006
- L'essentiel des droits politiques, économiques et sociaux (C. Marliac-Négrier) – 2003

Droit civil et procédure civile

- L'essentiel du droit des biens (S. Druffin-Bricca) – 4e édition – 2008
- L'essentiel du droit de la famille (C. Renault-Brahinsky) – 7e édition – 2008
- L'essentiel du droit des personnes (C. Renault-Brahinsky) – 4e édition – 2008
- L'essentiel du droit des successions (C. Renault-Brahinsky) – 4e édition – 2008
- L'essentiel de la Bioéthique et du droit de la Biomédecine (E. Mondielli) – 2008
- L'essentiel des institutions judiciaires (N. Fricero) – 3e édition – 2008
- L'essentiel du droit des régimes matrimoniaux (C. Renault-Brahinsky) – 4e édition – 2008
- L'essentiel de l'introduction générale au droit (S. Druffin-Bricca) – 5e édition – 2007
- L'essentiel du droit des obligations (C. Renault-Brahinsky) – 4e édition – 2007
- L'essentiel de la procédure civile (N. Fricero) – 5e édition – 2007
- L'essentiel des règles de procédure civile (J.-P. Branlard) – 3e édition – 2006
- L'essentiel de l'introduction historique à l'étude du droit (A. Mory) – 2005
- L'essentiel de l'organisation judiciaire en France (J.-P. Branlard) – 2e édition – 2004

Droit administratif

- L'essentiel du droit de l'urbanisme (I. Savarit-Bourgeois) – 6e édition – 2008
- L'essentiel du droit de la construction (M. Faure-Abbad) – 2e édition – 2008

- L'essentiel du droit des étrangers (E. Aubin) – 2008
- L'essentiel du droit de l'environnement (C. Roche) – 3e édition – 2008
- L'essentiel des institutions politiques et administratives de la France (D. Grandguillot) – 6e édition – 2008
- L'essentiel du contentieux administratif (M.-Ch. Rouault) – 2008
- L'essentiel du droit de l'eau (B. Drobenko) – 2008
- L'essentiel des marchés publics (F. Allaire) – 2007
- L'essentiel du droit administratif des biens (F. Colin) – 2007
- L'essentiel du droit administratif général (M.-Ch. Rouault) – 6e édition – 2007
- L'essentiel du droit administratif des biens (P. Binczak, S. Nicinski) – 3e édition – 2007
- L'essentiel du droit de la Fonction publique (E. Aubin) – 3e édition – 2007
- L'essentiel de l'organisation administrative (M.-Ch. Rouault) – 1re édition – 2006
- L'essentiel du nouveau droit de la décentralisation (E. Aubin, C. Roche) – 2006
- L'essentiel du droit du service public (R. Le Mestre) – 2003

Finances publiques

- L'essentiel du Droit fiscal 2008 (B. et F. Grandguillot) – 9e édition – 2008
- L'essentiel du droit des marchés financiers (A.-D. Merville) – 2008
- L'essentiel des finances locales (P. Mouzet) – 4e édition – 2008
- L'essentiel des finances publiques 2008 (F. Chouvel) – 9e édition – 2008
- L'essentiel des finances publiques communautaires (M.-Ch. Steckel-Montes) – 2e édition – 2007
- L'essentiel de la nouvelle constitution financière en France (M. Paul) – 2e édition – 2007
- L'essentiel de la LOLF (M. Paul) – 2005

Relations internationales

- L'essentiel du droit international public et du droit des relations internationales (C. Roche) – 3e édition – 2008
- L'essentiel des relations internationales (A. Gazano) – 4e édition – 2007
- L'essentiel de la Justice pénale internationale (S. Maupas) – 2007
- Chronologies générale et thématiques des relations internationales (J.-J. Roche) – 2007
- L'essentiel de l'organisation mondiale du commerce (D. Colard-Fabregoule)
- L'essentiel des organisations européennes de coopération (C. Roche et D. Thierry) – 2002

Droit commercial et des affaires

- L'essentiel de la macroénomie (T. Tacheix) – 4e édition – 2008
- L'essentiel des mécanismes de l'économie (G. Leguirriec-Milner) – 2e édition – 2008

- L'essentiel de la statistique descriptive (E. Olivier) – 2008
- L'essentiel des marchés financiers (C. Karyotis) – 2008
- L'essentiel de la micro-économie (B. Gendron) – 2008
- L'essentiel du droit des entreprises en difficulté (L. Antonini-Cochin et L.-C. Henry) – 2008
- L'essentiel du droit des sociétés. Sociétés commerciales, autres sociétés (B. et F. Grandguillot) – 7e édition – 2008
- L'essentiel des Stratégies d'internationalisation de l'entreprise (C. Mercier-Suissa et C. Bouveret-Rivat) – 2007
- L'essentiel des marchés de capitaux français (C. Karyotis) – 2007
- L'essentiel du droit des garanties de paiement (J.-P. Branlard) – 2e édition – 2007
- L'essentiel du droit général des sociétés (J.-P. Branlard)
- L'essentiel du droit spécial des sociétés (J.-P. Branlard)
- L'essentiel du droit des effets de commerce (J.-P. Branlard)

Droit européen et international

- L'essentiel des institutions de l'Union européenne (J.-C. Zarka) – 10e édition – 2008
- L'essentiel de l'Union européenne et droit communautaire (J.-M. Favret) – 8e édition – 2008
- L'essentiel du Droit des institutions de l'Union européenne (S. Leclerc) – 2007
- L'essentiel du droit international privé (L.-C. Henry) – 2005
- L'essentiel du contentieux communautaire (J.-M. Favret) – 2001

Droit du travail

- L'essentiel du droit du travail (D. Grandguillot) – 9e édition – 2008
- L'essentiel du droit de la Sécurité sociale (D. Grandguillot) – 7e édition – 2008
- L'essentiel de la gestion des ressources humaines (L. Lethielleux) – 2e édition – 2008
- L'essentiel de l'Histoire du droit social (Y. Delbrel) – 2006

Droit pénal et procédure pénale

- L'essentiel du droit pénal général (P. Kolb et L. Leturmy) – 5e édition – 2008
- L'essentiel de la procédure pénale (C. Renault-Brahinsky) – 8e édition – 2007
- L'essentiel du droit des peines (C. Renault-Brahinsky) – 2001

Comptabilité et gestion de l'entreprise

- L'essentiel de la théorie des organisations (R. Aïm) – 2e édition – 2008
- L'essentiel des opérations courantes en comptabilité générale (B. et F. Grandguillot) – 2e édition – 2008

- L'essentiel des opérations de fin d'exercice en comptabilité générale (B. et F. Grandguillot) – 2e édition – 2008
- L'essentiel du Marketing (S. Soulez) – 2008
- L'essentiel du contrôle de gestion (B. et F. Grandguillot) – 3e édition – 2008
- L'essentiel de l'économie de l'entreprise (S. Josien et S. Landrieux-Kartochian) – 2008
- L'essentiel de l'analyse financière (B. et F. Grandguillot) – 7e édition – 2008
- L'essentiel des formules types du courrier d'entreprise (lettres et e-mails) (A. Nishimata) – 3e édition – 2008
- L'essentiel de la gestion de projet (R. Aïm) – 3e édition – 2007
- Les 150 lettres et e-mails du créateur (A. Nishimata) – 2006
- L'essentiel des normes comptables internationales IAS/IFRS (S. Brun) – 3e édition – 2006
- L'essentiel de la comptabilité de gestion (B. et F. Grandguillot) – 3e édition – 2006
- L'essentiel de la comptabilité nationale (T. Tacheix) – 2e édition – 2004

Concours de la fonction publique

- Rédiger avec succès lettres et documents administratifs (R. Kadyss et A. Nishimata) – 3e édition – 2008
- L'essentiel des institutions scolaires et universitaires (F. Dupont-Marillia) – 2007
- L'essentiel pour accéder à la Fonction publique en France (F. Colin) – 2004
- L'essentiel de la note de synthèse aux concours de la fonction publique (M. Deyra) – 4e édition – 2006
- L'essentiel pour réussir l'épreuve de synthèse (A. Guilmoto) – 2004
- L'essentiel pour réussir l'épreuve orale de culture générale (A. Guilmoto) – 2004
- L'essentiel pour réussir les épreuves écrites fondamentales des concours (A. Guilmoto) – 2004
- L'essentiel de l'épreuve orale, conversation ou entretien avec un jury sur la carrière professionnelle aux concours de la fonction publique (R. Kadyss et A. Nishimata) – 2003
- L'oral aux concours de la fonction publique : commentaire de texte ou exposé sur un sujet de culture générale (A. Nishimata) – 2000
- L'essentiel du résumé aux concours de la fonction publique (A. Nishimata) – 2002
- L'essentiel de l'orthographe et de la grammaire française pour les candidats aux concours de la fonction publique (R. Kadyss et A. Nishimata) – 2004