MÉMENTOS LMD

Statistique descriptive

Séries statistiques à une et deux variables
Séries chronologiques
Indices

Une présentation synthétique et illustrée des séries statistiques à une ou deux variables, des séries chronologiques et des indices.

Fabrice MAZEROLLE
est Maître de conférences à la Faculté d'Aix-Marseille III. Il enseigne également la statistique descriptive dans divers établissements d'enseignement supérieur.

Site internet de l'auteur : www.mazerolle.fr

Du même auteur

- Exercices corrigés de statistique descriptive (*coll. Exercices corrigés*) – 1re édition 2006

Retrouvez tous nos titres
Defrénois - Gualino - Joly
LGDJ - Montchrestien
sur notre site
@ **www.eja.fr**

DANGER
LE
PHOTOCOPILLAGE
TUE LE LIVRE

© Gualino éditeur, EJA – Paris – 2006
ISBN 2 - 84200 - 891 - X

Dépôt légal : décembre 2005

MÉMENTOS LMD

Statistique descriptive

Séries statistiques à une et deux variables
Séries chronologiques
Indices

Une présentation synthétique et illustrée des séries statistiques à une ou deux variables, des séries chronologiques et des indices.

Fabrice MAZEROLLE

Gualino éditeur

Plusieurs séries de livres pour les **étudiants des facultés de droit, des sciences politiques, économiques et de gestion** ainsi que pour les **candidats aux concours de la Fonction publique** (catégorie A) :

- **Manuels**
- **Mémentos**
- **Les textes fondamentaux**
- **Panorama**
- **Abrégés illustrés**
- **Exercices corrigés**
  (collection en partenariat avec LGDJ)
- **AnnaDroit LMD**
  (édition annuelle des sujets d'examen)
- **Carrés Rouge**
- **Les glossaires**
- **QCM et QRC**

Catalogue général adressé gratuitement
sur simple demande :

Gualino éditeur
Tél. 01 56 54 16 00
Fax : 01 56 54 16 49
e-mail : gualino@eja.fr
Site Internet : www.eja.fr

Remerciements

Je tiens à remercier mon collègue Bernard PY pour m'avoir,
tout au long de la rédaction de ce *Mémento*,
fait bénéficier de sa grande expérience de la statistique.

Présentation

Ce mémento de **Statistique Descriptive** présente de façon synthétique, structurée et illustrée l'ensemble des connaissances et des techniques à maîtriser en sciences économiques et sociales.

Après un **chapitre introductif**, dans lequel le **vocabulaire des statistiques** est exposé, l'ensemble des connaissances nécessaires est développé en quatre parties. L'ouvrage contient de nombreux exemples permettant d'acquérir une pratique de cette matière :

- ***Les séries statistiques à une dimension*** : Qu'il s'agisse de la décomposition du Produit Intérieur Brut d'un pays par secteur d'activité, ou de l'évolution du chiffre d'affaires d'une entreprise à travers le temps, l'étudiant doit pouvoir en maîtriser la forme et la signification : présentation en tableaux, en graphiques et calcul des caractéristiques résumées d'une série de chiffres (moyenne, écart-type, mode, médiane, etc.).

- ***Les séries statistiques à deux dimensions*** : Le plus souvent, les tableaux et les graphiques présentent simultanément deux - voire plusieurs - dimensions d'un même phénomène, dans le but d'étudier leur interdépendance. Il existe pour cela des méthodes statistiques spécifiques, dont la plus connue est le coefficient de corrélation.

- ***Les séries chronologiques*** : L'évolution des phénomènes économiques et sociaux dans le temps joue un rôle si important en économie que l'étude des séries chronologiques mérite un traitement particulier, afin d'exposer en détail des outils tels que la décomposition d'une série sous forme d'un trend et d'une composante saisonnière.

- ***Les indices*** : Ils sont très utilisés en sciences sociales, de sorte qu'il est indispensable d'en connaître la construction, la manipulation et les propriétés.

- ***Un glossaire***, en fin d'ouvrage, reprend les principales formules étudiées dans le livre.

L'ouvrage s'adresse en priorité aux étudiants d'AEJ et de sciences économiques et gestion, mais aussi à tous les étudiants des formations dont le cursus comprend une initiation à la statistique descriptive.

Il peut être utilement complété par :

– Le livre *Exercices Corrigés de Statistique Descriptive*, publié dans la collection Fac-Université, du même auteur.

– Le site Internet de l'auteur, www.mazerolle.fr dont la rubrique « Statistique descriptive » est régulièrement mise à jour par des exercices corrigés, ainsi que des prolongements logiciels des exercices et des techniques statistiques exposés dans cet ouvrage.

Sommaire

Présentation 7

Chapitre 1 Vocabulaire de la statistique descriptive 15

1 Champ de la statistique descriptive 15
A – Définition 15
B – Statistique descriptive et statistique mathématique 15

2 Description d'une population statistique 16
A – Unités statistiques, population, échantillons 16
B – Caractères et variables 16
C – Modalités ordinales, modalités nominales 18
D – Valeurs discrètes, valeurs continues 19
E – Unités individuelles et unités groupées 19
F – Effectifs, fréquences, pourcentages, ratios, taux et indices 21
1) Effectifs ou fréquences absolues 21
2) Fréquences relatives et pourcentages 21
3) Ratio, taux et indices 22
G – Tableau récapitulatif 23

3 Taux de croissance 24
A – Définition 24
B – Évolutions successives 25
C – Taux de croissance moyen 25
D – Taux de croissance d'un produit 26
E – Taux de croissance d'un rapport 26

4 Opérateurs somme et produit 27
A – L'opérateur somme 27
B – L'opérateur produit 28

PARTIE 1 • Les séries statistiques à une dimension

Chapitre 2 Tableaux et graphiques 33

1 Tableaux 33

A – Tableaux de données qualitatives 33

B – Tableaux de données quantitatives 36

1) Variable quantitative discrète, valeurs connues individuellement 36
2) Variable quantitative discrète, valeurs regroupées 36
3) Variable quantitative continue, valeurs connues individuellement 37
4) Variable quantitative continue, données groupées 37

2 Graphiques 38

A – Importance des graphiques 38

B – Données individuelles 39

1) La ligne 39
2) Le graphique « tige et feuilles » 40

C – Données groupées par modalités ou valeurs 41

1) Diagramme en bâtons 41
2) Diagramme en barres 42
3) Nuage de points dans le cas d'une série unidimensionnelle 43

D – Camembert ou graphique « en tarte » ? 44

E – L'histogramme 45

F – L'utilisation des graphiques à des fins de comparaison 47

1) Le radar, excellent moyen d'effectuer des comparaisons visuelles 47
2) Comparaisons dans le temps 48
3) Les graphiques de séries chronologiques 48
4) Un beau graphique vaut mieux qu'un long discours 49
5) Les graphiques d'indices 50
6) Les échelles semi-logarithmiques 51

Chapitre 3 Les caractéristiques de tendance centrale 53

1 Les moyennes 53

A – La moyenne arithmétique 53

1) La moyenne arithmétique simple 53
2) La moyenne arithmétique pondérée 54
3) La moyenne élaguée 56

B – La moyenne quadratique 57

1) La moyenne quadratique simple 57
2) La moyenne quadratique pondérée 57

C – La moyenne géométrique 58

1) La moyenne géométrique simple 58
2) La moyenne géométrique pondérée 58

D – La moyenne harmonique 59
1) La moyenne harmonique simple 59
2) La moyenne harmonique pondérée 59

2 La médiane 60
A – Calcul de la médiane : effectif impair et aucune valeur n'est répétée 61
B – Calcul de la médiane : effectif pair et aucune valeur n'est répétée 61
C – Calcul de la médiane : effectifs groupés par valeurs 62
D – Calcul de la médiane : effectifs groupés par classes de valeurs 63

3 Le mode 65
A – Calcul du mode : série simple, aucune valeur n'est répétée 65
B – Calcul du mode : effectifs groupés par valeurs 65
C – Calcul du mode : effectifs groupés par classes d'amplitudes égales 65
D – Calcul du mode : effectifs groupés par classes d'amplitudes inégales 66

4 Comment caractériser la forme d'une distribution à l'aide de la moyenne arithmétique, de la médiane et du mode 68
A – Distribution parfaitement symétrique 68
B– Distribution étalée à droite 69
C – Distribution étalée à gauche 70

Chapitre 4 Dispersion et concentration 71

1 L'intervalle de variation 71

2 L'intervalle interquartile 72

3 La boîte à moustache 78
A – Définition 78
B – Utilité de la boîte à moustache pour comparer des séries 79
C – Utilité de la boîte à moustache pour déterminer la forme d'une distribution 80

4 Variance, écart-type et coefficient de variation 81
A – La variance 81
1) Définition 81
2) Mode de calcul de la formule (1-a) 82
3) Mode de calcul de la formule « développée » 83
B – L'écart-type et le coefficient de variation 84
1) L'écart-type 84
2) Le coefficient de variation 85

5 Les indicateurs de concentration 87
A – La médiale 87
B – La détermination de la concentration par la méthode graphique 88
C – L'indice de GINI 90
D – L'écart médiale-médiane rapporté à l'intervalle de variation 92

PARTIE 2 • Les séries statistiques à deux dimensions

Chapitre 5 Les séries statistiques à deux dimensions. I : tableaux, graphiques, vocabulaire 97

1 Tableaux et graphiques 97
A – Séries quantitatives connues individuellement 97
B – Séries quantitatives groupées 99
C – Séries qualitatives 100

2 Représentation abstraite d'un tableau de contingence 101

3 Effectifs marginaux et fréquences marginales 103

4 Moyennes et variances marginales 104
A – Moyennes marginales 104
B – Variances marginales 105

5 Fréquences partielles sur effectif total 106

6 Distributions conditionnelles 106

7 – Moyennes et variances conditionnelles 108
A – Moyennes conditionnelles 108
B – Variances conditionnelles 109

Chapitre 6 Les séries statistiques à deux dimensions. II : outils d'analyse 111

1 Séries quantitatives avec observations connues individuellement 111
A – Liaison linéaire, liaison non linéaire, absence de liaison 111
B – La droite de régression linéaire 114
1) Définition 114
2) Calcul des coefficients 115
3) Utilité de la droite de régression 117
C – Le coefficient de corrélation 117
1) Définition et calcul 117
2) Coefficient de corrélation et coefficient de détermination 118
3) Corrélation et causalité 118

2 Séries quantitatives avec observations groupées 120
A – Cas des données groupées par valeurs 120
B – Cas des données groupées par classes 121
1) Le coefficient de corrélation 121
2) Le test d'indépendance 124

3 Séries qualitatives 125
A – Le coefficient de corrélation de rang de SPEARMAN 125
B – Le test du Khi-carré de PEARSONS 127

PARTIE 3 • Les séries chronologiques

Chapitre 7 Les séries chronologiques 131

1 Introduction 131
A – Définition 131
B – Périodicité 132
C – Tendance, variations saisonnières et accidentelles 133
D – Modèle multiplicatif et modèle additif 134

2 Détermination du trend d'une série chronologique 135
A – La détermination du trend par la régression linéaire 135
B – La détermination du trend par la méthode des moyennes mobiles 137

3 Les variations saisonnières 140
A – Vocabulaire 140
B – Les étapes du calcul de la série CVS 141
1) Détermination de l'équation du trend 142
2) Calcul des coefficients saisonniers 143
3) Détermination de la série CVS 145

4 Les variations accidentelles 146

PARTIE 4 • Les indices

Chapitre 8 Les indices 151

1 Introduction 151
A – Définition et exemples 151
B – Indice temporel et indice de situation 152
C – Indice élémentaire et indice synthétique 154

2 Les indices synthétiques de LASPEYRES, PAASCHE et FISHER 156
A – Définition de la valeur d'un panier de biens 156
B – Les indices de LASPEYRES 156
1) L'indice de LASPEYRES des prix 156
2) L'indice de LASPEYRES des quantités 158
C – Les indices de PAASCHE 158
1) L'indice de PAASCHE des prix 159
2) L'indice de PAASCHE des quantités 159
D – Les indices de FISHER 160
1) L'indice de FISHER des prix 160
2) L'indice de FISHER des quantités 161

3 L'indice des prix à la consommation de l'INSEE 161

Glossaire des formules 163

Bibliographie 173

Avertissement

Les erreurs éventuelles qui subsisteraient dans cette première édition sont toutes de mon fait et seront corrigées dans les éditions ultérieures.

CHAPITRE 1

Vocabulaire de la statistique descriptive

Ce chapitre introductif est consacré à la définition de la statistique descriptive ainsi que des différents termes qui en constituent le vocabulaire de base.

1 • CHAMP DE LA STATISTIQUE DESCRIPTIVE

Il suffit d'allumer son ordinateur ou d'écouter les informations à la radio pour constater que les statistiques sont partout. Ceci révèle que le monde moderne est presque entièrement tourné vers le quantitatif et le mesurable. D'où l'intérêt de la statistique, discipline relativement récente, mais qui correspond parfaitement à cette orientation du monde moderne.

A – Définition

Il existe de nombreuses définitions (plusieurs centaines), celle que nous donnons ici est celle de Bernard PY,dans son livre *Statistique descriptive, nouvelle méthode pour bien comprendre et réussir* (éditions Economica) : *« La statistique [descriptive] est un ensemble de méthodes permettant de décrire et d'analyser, de façon quantifiée, des phénomènes repérés par des éléments nombreux, de même nature, susceptibles d'être dénombrés et classés.* »

Deux points importants ressortent de cette définition :

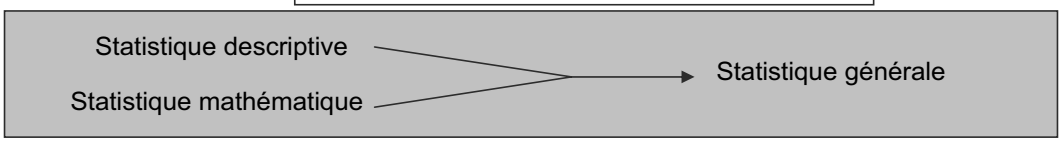
1) Ensemble de méthodes : la statistique descriptive ne contient aucune théorie, mais seulement des outils d'investigation et de mesure des données chiffrées.

2) Décrire et analyser, de façon quantifiée, des phénomènes repérés par des éléments nombreux : décrire, c'est-à-dire faire des tableaux, des graphiques, calculer des moyennes afin de faire ressortir la signification.

B – Statistique descriptive et statistique mathématique

La **statistique descriptive** appartient cependant à un ensemble plus vaste, la **statistique générale**, qui se divise en deux branches : statistique descriptive, objet de ce mémento, et la **statistique mathématique** (ou statistique "inférentielle"), dont l'objet est de formuler des lois de comportement à partir d'observation souvent incomplètes. Cette dernière intervient dans les enquêtes et les sondages. Elle s'appuie non seulement sur la statistique descriptive, mais aussi sur le **calcul des probabilités**.

Schéma 1 : Les deux branches de la statistique

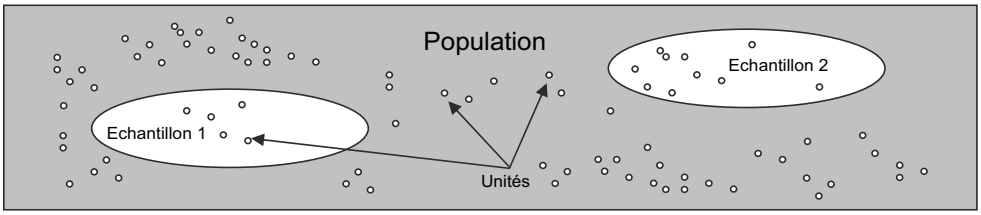


2 • DESCRIPTION D'UNE POPULATION STATISTIQUE

A – Unités statistiques, population, échantillons

Les éléments nombreux dont s'occupe la statistique descriptive sont appelés des **unités statistiques**. Ces unités sont regroupées dans une **population**. Lorsque la population est trop importante pour être connue entièrement, on prélève un **échantillon**. Les relations qui existent entre la population, les échantillons et les unités statistiques sont résumées dans le schéma ci-dessous.

Schéma 2 : Unités statistiques, population, échantillons



En théorie, on doit soigneusement distinguer la description d'un échantillon et la description d'une population. C'est d'ailleurs l'un des objets principaux de la statistique mathématique que de préciser les conditions dans lesquelles un échantillon est représentatif d'une population. De ce fait, certaines formules de calcul qui sont valables pour une population sont légèrement différentes quand on les applique à un échantillon. C'est le cas notamment de la variance (voir le chapitre 3). Cependant, **sauf mention contraire explicite**, nous considérons dans cet ouvrage que les séries étudiées constituent une population complète et non un échantillon.

B – Caractères et variables

Dans une population, par exemple celle des étudiants d'une faculté, les unités sont repérées par le nom et le prénom des étudiants (on a donc une liste). Si l'on souhaite étudier cette population, on va retenir certains critères d'étude comme le sexe, la filière principale à laquelle chaque étudiant se rattache, les matières optionnelles qu'il a choisi, l'âge, le poids, la taille, etc.

Parmi ces critères, certains sont **quantitatifs**, comme l'âge, le poids, la taille. On peut en effet effectuer des calculs numériques sur ces critères : poids moyen, taille maximale, taille minimale, etc. D'autres critères ne sont pas quantifiables, car on ne peut pas effectuer de calculs dessus. Ils sont **qualitatifs**. C'est le cas du sexe par exemple. On peut connaître l'effectif masculin et l'effectif féminin d'une population, mais la notion de « sexe moyen » n'a pas de sens et ne peut d'ailleurs pas être calculée.

Afin de différencier les deux type de critères, les critères qualitatifs sont appelés des **caractères** et les critères quantitatifs des **variables**. On désigne par **modalités** les différentes catégories d'un caractère qualitatif et on qualifie de **valeurs** les différents chiffres d'une variable.

Exemple 1 : soit une population de 600 étudiants, avec un effectif féminin de 230 et un effectif masculin de 370. Traduisons ces informations dans le vocabulaire de la statistique descriptive.

Tableau 1 : Exemple d'un critère qualitatif

P	Population	Effectif total : n = 600
i	unités statistiques	Chaque étudiant i = 1, 2, …, n
X	Caractère	Le sexe
X_F X_M	Modalités	Féminin ou Masculin
n_F n_M	Effectifs associés à chaque modalité	370 hommes, 230 femmes

L'effectif total, n, va se répartir entre l'effectif masculin et l'effectif féminin, ce qui nous permet décrire que $n = n_F + n_M$. Cette égalité, nous pouvons l'écrire parce que les différentes modalités d'un caractère sont à la fois **exhaustives** et **incompatibles**. Exhaustives, car elles décrivent toutes les valeurs ou états possibles d'un caractère. Incompatibles, car un individu ne peut pas avoir plus d'une modalité.

Exemple 2 : soit un échantillon de 10 étudiants ayant passé un examen. Ils ont obtenu les notes suivantes (sur 20) : {16, 8, 6, 14, 10, 18, 13, 9, 10, 15}.

Tableau 2 : Exemple d'un critère quantitatif

E	Échantillon	Effectif de l'échantillon : n=10
i	Unités statistiques	Chaque étudiant i = 1, 2, …, n
X	Variable	Notes
$\{x_1, x_2, \ldots, x_h\}$	Valeurs (*)	{6,8,9,10,13,14,15,16,18}
$\{n_1, n_2, \ldots, n_h\}$	Effectifs associés à chaque valeur	{1,1,1,2,1,1,1,1,1}

(*) Il n'y a que 9 valeurs, parce que le 10 est répété 2 fois. Ce qui montre l'importance de distinguer les valeurs de la variable et l'effectif de l'échantillon (ou de la population). L'effectif varie de 1 à n (avec n=10), tandis que les valeurs varient de 1 à 9 (avec h=9).

C – Modalités ordinales, modalités nominales

Les modalités d'un caractère qualitatif, si elles ne peuvent pas être mesurées quantitativement, sont parfois susceptibles d'être classées. Ce sont des **modalités ordinales**.

Exemple 1 : Un questionnaire de satisfaction demande aux consommateurs d'évaluer une prestation en cochant l'une des six catégories suivantes :

(a) nulle, (b) médiocre, (c) moyenne, (d) assez bonne, (e) très bonne, (f) excellente

Il s'agit de modalités ordinales puisqu'elles peuvent être hiérarchisées : une prestation excellente est meilleure qu'une prestation bonne, etc. La différence avec des valeurs quantitatives est qu'on ne peut dire, par exemple, si une prestation jugée excellente est deux fois ou quatre fois meilleure qu'une prestation décrite comme moyenne. On peut effectuer un classement, non une quantification.

Remarque : certaines modalités ordinales peuvent néanmoins être transformées valeurs quantitatives. Ce sont en fait des valeurs quantitatives qui prennent l'apparence de modalités qualitatives ordinales.

Exemple 2 : Des chemises sont classées par taille : XS, S, M, L, XL, XXL, XXXL. Il s'agit de modalités faussement ordinales. En réalité il existe un tableau de correspondance qui explicitera à quelle taille en cm chacune de ces catégories correspond.

Les modalités d'un caractère qualitatif qui ne peuvent pas être classées ou hiérarchisées sont dites **nominales**.

Exemple 3 : On demande à un échantillon de personnes ce qu'évoque pour elles un parfum. Plus précisément, elles doivent cocher une des cases suivantes :

(a) aventure, (b) sensualité, (c) confort, (d) nostalgie

Il est clair qu'aucune comparaison ni hiérarchisation ne peuvent être établies entre ces modalités. Elles sont nominales.

Remarque : Certaines modalités purement nominales sont parfois codées avec des chiffres. Par exemple, le sexe des individus d'une population sera codé par "1" pour les hommes et par "2" pour les femmes. Il s'agit bien là d'une tentative de quantification d'une variable purement nominale. On parle alors de variables **pseudo-numériques**. On peut en effet de cette façon calculer une moyenne, qui sera en fait la proportion des hommes dans la population ou dans l'échantillon.

D – Valeurs discrètes, valeurs continues

Une variable quantitative peut-être discrète ou continue. Lorsque le nombre de valeurs possibles est fini (exemple : le nombre d'enfants, le nombre de pièces d'un logement, etc.), la variable est **discrète**. Lorsque le nombre de valeurs possibles de la variable est infini (exemple : la taille, le poids ou le revenu des ménages), la variable est **continue**.

E – Unités individuelles et unités groupées

Les unités d'une population, que le critère soit qualitatif ou quantitatif (discret ou continu), peuvent être présentées individuellement (c'est généralement le cas lorsque les données sont saisies) ou regroupées. Le regroupement peut être effectué par modalités, par valeurs ou par classes de modalités ou de valeurs.

Exemple 1 : Un questionnaire de satisfaction demande à un échantillon de 10 consommateurs d'évaluer une prestation en cochant l'une des six catégories suivantes :

(a) nulle, (b) médiocre, (c) moyenne, (d) assez bonne, (e) très bonne, (f) excellente

On présenter les données individuellement (tableau 3), groupées par modalités (tableau 4) ou par classes de modalités (tableau 5).

Tableau 3 : Données présentées individuellement

Identificateur(*)	1	2	3	4	5	6	7	8	9	10
Évaluation	a	e	e	c	e	f	a	f	e	b

(*) Nom de la personne ou numéro si l'on veut préserver l'anonymat.

Tableau 4 : Données groupées par modalités

Modalités	a	b	c	d	e	f
Effectif	2	1	1	0	4	2

Tableau 5 : Données groupées par classes de modalités

Classes	De nulle à assez bonne (a – b – c – d)	De très bonne à excellente (e – f)
Effectif	4	6

Exemple 2 : On a mesuré 20 personnes et les résultats sont (en cm) :

{148, 165, 145, 173, 148, 145, 152, 180, 135, 170, 170, 170, 142, 148, 165, 175, 180, 180, 180, 180}

Il s'agit d'un variable continue (la taille), mais dont les valeurs sont ici connues individuellement. On peut aussi effectuer un regroupement par taille car certaines tailles, comme 170 ou 180, apparaissent plusieurs fois (tableau 6).

Tableau 6 : Données groupées par valeurs

Taille	**135**	**142**	**145**	**148**	**152**	**165**	**170**	**173**	**175**	**180**
Effectifs	1	1	2	3	1	2	3	1	1	5

Il est également possible d'effectuer un regroupement par classes de valeurs. On choisira, à titre d'exemple, un regroupement par classes **d'amplitudes égales** (tableau 7), puis un regroupement par **classes d'amplitudes inégales** (tableau 8). On désigne par a_i, l'amplitude d'une classe. Dans le tableau 7, l'amplitude de classe est la même pour toutes les classes (10 cm) alors qu'elle est de 20 cm, 20 cm et 10 cm dans le tableau 8.

Tableau 7 : Groupement par classes (amplitudes égales)

Classes	**Effectifs**
[130-140[	1
[140-150[	6
[150-160[	1
[160-170[	2
[170-180]	10

Tableau 8 : Données groupées par valeurs (amplitudes inégales)

Classes	**Effectifs**
[130-150[	7
[150-170[	3
[170-180]	10

Lorsque les unités statistiques sont groupées par classes, on calcule un **centre de classe**, désigné par c_i, qui est égal à la moyenne des extrémités de classes (voir le tableau 9 pour le calcul des centres de classe du tableau 8).

Tableau 9 : Calcul des centres de classe des données du tableau 8

Classes	**Centres de classe (c_i)**
[130-150[	(130+150)/2 = 140
[150-170[	(150+170)/2 = 160
[170-180]	(170+180)/2 = 175

Exemple 3 : On a questionné 100 ménages sur le nombre d'ampoules électriques utilisées dans leur domicile. Dans le premier tableau, les données sont regroupées par nombre d'ampoules. Dans le second tableau, elles sont regroupées par classes.

Tableau 10 : Regroupement par nombre d'ampoules

Nombre d'ampoules	2	3	4	5	6	7	8	9	11	12	13	15
Effectifs	5	8	8	10	18	16	10	9	6	5	3	2

Tableau 11 : Regroupement par classes

Classes	Effectifs
[2-5[	21
[5-10[	63
[10-15[	16

F – Effectifs, fréquences, pourcentages, ratios, taux et indices

Une fois les unités statistiques d'une population répertoriées, celles-ci sont présentées dans des tableaux (voir le chapitre 2), de diverses manières : effectifs ou fréquences absolues, fréquences relatives, pourcentages, ratios, indices et taux. Il convient de définir ces termes avec précision :

1) Effectifs ou fréquences absolues

Il s'agit de la répartition brute des données. Lorsque les données sont présentées individuellement, chaque donnée a la même fréquence unitaire d'apparition, leur **effectif** ou **fréquence absolue** est égal à 1. Lorsque les données sont regroupées par valeurs ou modalités, les effectifs ou fréquences absolues correspondent au nombre de données qui ont la valeur ou modalité, ou encore qui sont groupées dans une classe donnée.

Symboliquement, les effectifs ou fréquences absolues s'écrivent n_i. Et la somme des effectifs est égale à n. Ainsi, dans le cas du tableau 11, les effectifs ou fréquences absolues dont respectivement égaux à n_1=21, n_2=63 et n_3=16. De plus, on a :

$$n_1 + n_2 + n_3 = 21 + 63 + 16 = 100 = n \qquad (1)$$

2) Fréquences relatives et pourcentages

La **fréquence relative** est égale à la fréquence absolue divisée par l'effectif total :

$$f_i = \frac{n_i}{n} \qquad (2)$$

On a donc :

$$f_1 + f_2 + \ldots + f_h = \frac{n_1}{n} + \frac{n_2}{n} + \ldots + \frac{n_h}{n} = \frac{n_1 + n_2 + \ldots + n_n}{n} = \frac{n}{n} = 1 \qquad (4)$$

Le **pourcentage** des données qui correspondent à une modalité, à une valeur ou à une classe s'obtient en multipliant la fréquence relative correspondante par 100. C'est-à-dire:

Pourcentage de la valeur (modalité ou classe) i = f_i x 100 (5)

Le tableau 12 reprend l'exemple de la répartition des ménages en fonction du nombre d'ampoules utilisées à leur domicile, en ajoutant la colonne des fréquences relatives à côté de celle des fréquence absolues. La dernière ligne correspond aux totaux.

Tableau 12 : Répartition des ménages en fonction du nombre d'ampoules à leur domicile

Classes	Effectifs ou fréquences absolues	Fréquences relatives	Pourcentages
[2-5[	21	0,21	21
[5-10[	63	0,63	63
[10-15]	16	0,16	16
Total	100	1	100

Les colonnes 2 (fréquences absolues) et 4 (pourcentages) contiennent les mêmes valeurs car l'effectif total est égal à 100. Si celui-ci était différent de 100, les valeurs contenues dans les deux colonnes seraient différentes.

3) Ratio, taux et indices

Un **ratio** est une fraction qui divise deux quantités. Les fréquences relatives sont des ratios puisqu'elles divisent deux quantités. Plus généralement, les ratios sont très utilisés en statistiques.

Exemple 1 : Soit la série de pièces défectueuses produites par 10 machines au cours d'une semaine donnée.

{8, 16, 9, 33, 14, 5, 3, 7, 10, 7}

Le ratio du nombre de pièces défectueuses le plus élevé au nombre de pièces défectueuses le plus faible est 33/3 = 11. La machine numéro 4 a donc produit 11 fois plus de pièces défectueuses que la machine numéro 7.

Un **taux** est le ratio d'une quantité par unité (de temps, de surface, de poids, etc.)

Exemple 2 : Soit la série de pièces défectueuses produites par 10 machines au cours d'une semaine donnée.

{8, 16, 9, 33, 14, 5, 3, 7, 10, 7}

Ces chiffres sont des taux car ils sont exprimés dans l'unité « semaine ». Cette unité est « 1 ». On dit par conséquent 8 pièces **par semaine**, 16 pièces par semaine, etc.

Un **indice** est le ratio d'une quantité à une autre quantité qui sert de référence, multiplié par 100.

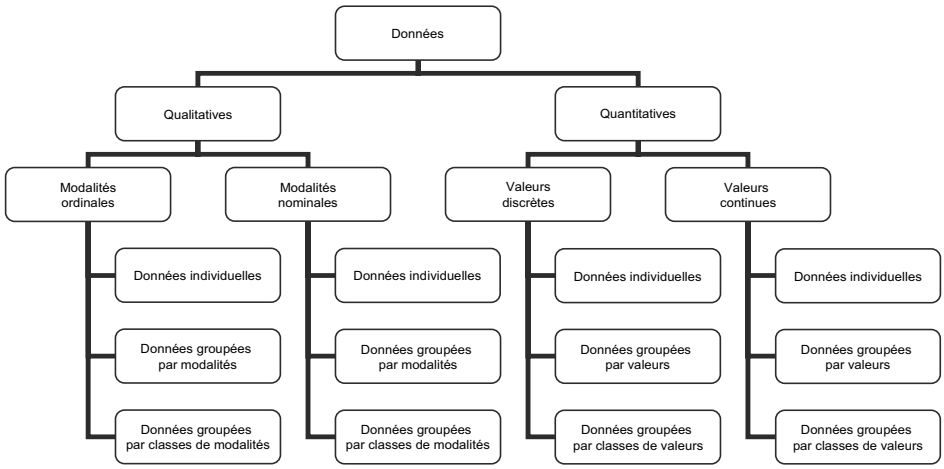
Exemple 3 : Soit la série de pièces défectueuses produites par 10 machines au cours d'une semaine donnée de l'exemple 1. Divisons chacune des valeurs de la série par la valeur la plus faible et multiplions ensuite chaque valeur par 100. Le résultat est une série d'indices, la « base 100 » étant la machine numéro 7.

{ 266,7 ; 533,3 ; 300 ; 1100 ; 466,7 ; 166,7 ; **100** ; 233,3 ; 333,3 ; 233,3 }

G – Tableau récapitulatif

Le Schéma 3 ci-dessous récapitule les différentes sortes de données que l'on rencontre en statistique, en partant de la distinction fondamentale entre données qualitatives et données quantitatives.

Schéma 3 : Différentes sortes de données statistiques



3 • TAUX DE CROISSANCE

A – Définition

Le **taux de croissance** est très utilisé en statistique et, plus généralement, en économie. Il se définit ainsi :

$$\text{Taux de croissance} = \frac{\text{Valeur d'arrivée}}{\text{Valeur de départ}} - 1 \qquad (3)$$

Soit g = taux de croissance, V_0 = valeur de départ et V_t =valeur d'arrivée. On a :

$$g = \frac{V_t}{V_0} - 1 = \frac{V_t - V_0}{V_0}$$

Le rapport V_t/V_0 est appelé **multiplicateur**. Dès lors, on peut écrire :

$$g = \text{multiplicateur} - 1 \qquad (5)$$

Ou encore :

$$\text{multiplicateur} = 1 + g \qquad (6)$$

Prenons un exemple :

$$\left.\begin{array}{l} V_t = 150 \\ V_0 = 100 \end{array}\right\} \longrightarrow \quad g = \frac{150}{100} - 1 = 0,5$$

Le taux de croissance, exprimé en pourcentage, est égal à 0,5 x 100 = 50%.

Ne pas confondre le taux de croissance, qui est une **variation relative**, et la **variation absolue** qui est $V_t - V_0$. Ici, la variation absolue est égale à 150-100 = 50.

Remarque :

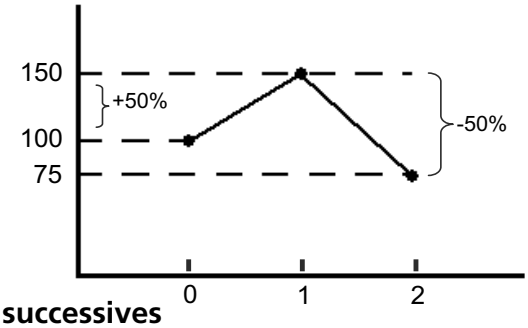
1) Ne pas confondre le taux de croissance, qui est une variation relative, avec la variation absolue, qui est égale à $V_t - V_0$. Dans l'exemple précédent, la variation absolue est égale à 150 - 100 = 50. En d'autres termes :

$$g = \frac{V_t}{V_0} - 1 = \frac{V_t - V_0}{V_0} = \frac{\text{variation absolue}}{\text{valeur de départ}} \qquad (7)$$

2) En matière de taux de croissance, il n'y a pas de symétrie entre les hausses et les baisses :

Lorsque je passe de 100 à 150, le taux de croissance, g est égal à (150/100) – 1 = 0,5, comme on l'a vu précédemment. Mais si maintenant on applique une baisse de 50% à 150, on obtient 150(1-0,5) = 75. On ne retrouve pas la valeur de départ. Le graphique ci-dessous illustre ce point.

Figure 1 : Une hausse de 50% suivie d'une baisse de 50%



B – Évolutions successives

Soient g_1, g_2,…, g_t des taux de croissance successifs. Le taux de croissance global sur la période 1,…,t est :

$$g = (1+g_1)(1+g_2)....(1+g_n)-1 \tag{8}$$

Exemple : soit une hausse de 5% suivie d'une hausse de 2%, puis d'une baisse de 3%. Quel est le taux de croissance global (sur les 3 périodes) ?

$$g = (1+0,05)(1+0,02)(1-0,03)-1 = 0,03887$$

C – Taux de croissance moyen

Soient g_1, g_2,…, g_t des taux de croissance successifs. Le taux de croissance moyen sur la période 1,…,t est :

$$\overline{g} = \sqrt[t]{(1+g)}-1 \tag{9}$$

C'est-à-dire :

$$\overline{g} = (1+g)^{\frac{1}{t}}-1 \tag{9-1}$$

Exemple : soit une grandeur qui a augmenté successivement de g_1 = 10%, g_2 = 20% et g_3 = 40% sur 3 ans. Son taux d'accroissement global est :

g= (1+0,1)(1+0,2)(1+0,4)-1=0,848

Et son taux de croissance moyen sur les trois périodes :

$$\overline{g} = (1+g)^{\frac{1}{3}} - 1 = 1{,}848^{\frac{1}{3}} - 1 \tag{10}$$

D – Taux de croissance d'un produit

Soient deux grandeurs à la date t :

$$V_t = (1+g_v)V_0 \quad \text{et} \quad U_t = (1+g_u)U_0 \tag{11}$$

La grandeur qui représente leur produit est :

$$W_t = V_t \times U_t = (1+g_v)(1+g_u)W_0 \tag{12}$$

Et son taux de croissance est :

$$g_w = \frac{W_t}{W_0} - 1 = (1+g_v)(1+g_u) - 1 \tag{13}$$

Exemple : Soit un commerçant qui augmente le prix d'un produit de 4%. À la suite de cette augmentation, la quantité vendue baisse de 3%. Le taux de croissance de la recette totale est alors donnée par :

(1 + 0,04)(1 - 0,03) – 1 = (1,04 x 0,97) – 1 = + 0,0088

Soit une hausse de 0,88% de la recette totale.

E – Taux de croissance d'un rapport

Soient deux grandeurs à la date t :

$$V_t = (1+g_v)V_0 \quad \text{et} \quad U_t = (1+g_u)U_0 \tag{14}$$

La grandeur qui représente leur rapport est :

$$Z_t = \frac{V_t}{U_t} = \frac{(1+g_v)}{(1+g_u)} Z_0 \tag{15}$$

Et son taux de croissance est :

$$g_z = \frac{(1+g_v)}{(1+g_u)} - 1 \qquad (16)$$

Exemple : soit un commerçant qui augmente le prix d'un produit de 4%. À la suite de cette augmentation, il constate que sa recette totale augmente de 0,88%. Étonné, il calcule le taux de croissance de la quantité vendue :

(1 + 0,0088)/(1 + 0,04) - 1= 0,97 – 1 = - 0,03

Il constate ainsi que la quantité vendue a baissé de 3%. Il comprend alors que si la recette totale a augmenté en dépit de la baisse de la quantité vendue, c'est parce que la baisse de la quantité vendue (3%) a été moins importante que l'augmentation du prix (4%) et s'endort content.

4 • OPÉRATEURS SOMME ET PRODUIT

A – L'opérateur somme

Pour exprimer une somme d'éléments de façon compacte, on utilise **l'opérateur somme**, symbolisé par la lettre grecque majuscule "Sigma".

Sigma ⟶ $\sum$ opérateur somme

Exemple 1 : soit quatre valeurs d'une variable x, indicées par i : x_1, x_2, x_3, x_4. Le produit de ces 4 valeurs est donné par l'expression :

$$\sum_{i=1}^{4} x_i = x_1 + x_2 + x_3 + x_4$$

L'expression de gauche se lit ainsi "somme des x_i pour i allant de 1 à 4". Plus généralement, pour une somme de n éléments, on écrit :

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_4$$

Exemple 2 : soit le tableau de valeurs suivant. Calculons les expressions :

x_i	y_i
1	2
-3	3
-4	4
2	5

$$\sum_{i=1}^{4} x_i \qquad \sum_{i=1}^{4} y_i \qquad \sum_{i=1}^{4} x_i^2$$

$$\sum_{i=1}^{4} (x_i + y_i) \qquad \sum_{i=1}^{4} x_i^2 y_i$$

D'où le tableau :

x_i	y_i	x_i^2	$x_i + y_i$	$x_i^2 y_i$
1	2	1	3	2
-3	3	9	0	27
-4	4	16	0	64
2	5	4	7	20
- 4	$\sum_{i=1}^{4} y_i = 14$	$\sum_{i=1}^{4} x_i^2 = 30$	$\sum_{i=1}^{4} (x_i + y_i) = 10$	$\sum_{i=1}^{4} x_i^2 y_i = 113$

$$\boxed{\sum_{i=1}^{4} x_i = \left[1 + (-3) + (-4) + 2\right] = -4}$$

B – L'opérateur produit

Pour exprimer un produit d'élément de façon compacte, on utilise **l'opérateur produit**, symbolisé par la lettre grecque majuscule Pi :

Pi ⟶ $\prod$ opérateur produit

Exemple 1 : soit quatre valeurs d'une variable x, indicées par i : x_1, x_2, x_3, x_4. Le produit de ces 4 valeurs est donnée par l'expression :

$$\prod_{i=1}^{4} x_i = x_1 \times x_2 \times x_3 \times x_4$$

L'expression de gauche se lit ainsi "produit des x_i pour i allant de 1 à 4". Plus généralement, pour un produit de n éléments, on écrit :

$$\prod_{i=1}^{n} x_i = x_1 \times x_2 \times \ldots \times x_4$$

Exemple 2 : soit le tableau de valeurs suivant. Calculons les expressions :

x_i	y_i
1	2
-3	3
-4	4
2	5

$$\prod_{i=1}^{4} x_i \qquad \prod_{i=1}^{4} y_i \qquad \prod_{i=1}^{4} x_i^2$$

$$\prod_{i=1}^{4} (x_i + y_i) \qquad \prod_{i=1}^{4} x_i^2 y_i$$

D'où le tableau :

x_i	y_i	x_i^2	$x_i + y_i$	$x_i^2 y_i$
1	2	1	3	2
-3	3	9	0	27
-4	4	16	0	64
2	5	4	7	20
24	$\prod_{i=1}^{4} y_i = 120$	$\prod_{i=1}^{4} y_i^2 = 576$	$\prod_{i=1}^{4} (x_i + y_i) = 0$	$\prod_{i=1}^{4} x_i^2 y_i = 69120$

$$\boxed{\prod_{i=1}^{4} x_i = [1 \times (-3) \times (-4) \times 2] = 24}$$

PARTIE 1

Les séries statistiques à une dimension

CHAPITRE 2

Tableaux et graphiques

Tableaux et graphiques constituent les deux moyens principaux de présentation des données statistiques. Étant donné l'abondance des présentations tabulaires et graphiques, nous n'étudierons ici que les principales.

1 • TABLEAUX

Un tableau statistique est juste une liste de chiffres relative au caractère de la population que l'on souhaite étudier, présentée de façon la plus compréhensible possible. Les données peuvent être présentées individuellement, sous forme d'effectifs, de fréquences ou de pourcentages et encore de bien d'autres façons.

Cette section propose d'étudier quelques exemples de tableaux-types, afin de familiariser le lecteur avec les modes de présentation les plus fréquents. L'analyse des tableaux à deux ou plusieurs caractères est renvoyée à la seconde partie de l'ouvrage.

A – Tableaux de données qualitatives

Le tableau (1) ci-dessous indique la répartition par continent des utilisateurs d'Internet en 2003. Le caractère étudié – la répartition continentale des utilisateurs d'Internet – est **qualitatif**. Il a sept modalités, listées dans la première colonne. La seconde colonne indique les effectifs, c'est-à-dire ici le nombre d'utilisateurs d'internet dans chacune des zones. La dernière ligne, en caractères gras, indique le total mondial.

Tableau 1 : Utilisateurs d'Internet par zones géographiques (Effectifs en mars 2005)

Zones géographiques (1)	Effectifs en millions
Asie	302,2
Europe	259,6
Amérique du Nord	221,4
Amérique du Sud/Caraïbes	56,2
Moyen-Orient	19,3
Océanie/Australie	16,2
Afrique	13,4
Total	**883,3**

Source : www.internetworldstats.com/stats

Note : Pour connaître la liste des pays inclus dans chaque zone, voir la source des données.

On prendra soin de toujours indiquer la source des données, afin que l'utilisateur du tableau puisse éventuellement s'y référer. Il est également important d'ajouter toute note utile pour la compréhension des données. Dans l'exemple des zones géographiques, il peut être nécessaire soit d'énumérer les pays qui figurent dans les zones, soit de référer à la source (à condition qu'elle le fasse, ce qui est le cas ici, mais il faut le vérifier).

Remarquons que les données ont été classées, non par ordre alphabétique des zones (ce qui est normalement le cas), mais par ordre croissant du nombre d'utilisateurs, ceci afin de faire apparaître les zones où l'utilisation d'Internet est la plus répandue.

Ce tableau peut être complété de plusieurs façons, afin d'en faciliter l'analyse.

Premièrement, on peut présenter les chiffres en pourcentages, dans une seconde colonne, afin de mieux apprécier la part de chaque zone dans le total des utilisateurs. C'est ce qui a été fait dans le tableau ci-dessous (colonne 3).

Deuxièmement, la colonne (4) présente la **somme cumulée des pourcentages**, de façon à mettre en évidence la contribution additionnelle de chaque zone ainsi que la concentration des utilisateurs. On voit ainsi que les 3 premières zones (Asie, Europe et Amérique du Nord) totalisent 88,7% des utilisateurs, les quatre autres zones (Amérique du sud/caraïbes, Moyen-Orient et Océanie/Australie) ne représentent quant à elles que 100 - 88,7 = 11,3% des utilisateurs.

Tableau 2 : Utilisateurs d'Internet par zones géographiques
(Effectifs, pourcentages et pourcentages cumulés en mars 2005)

Zones géographiques (1)	Effectifs en millions	Pourcentages	Pourcentages cumulés
Asie	302,2	34,02	34,02
Europe	259,6	29,22	62,24
Amérique du Nord	221,4	24,92	88,17
Amérique du Sud/Caraïbes	56,2	6,33	94,49
Moyen-Orient	19,3	2,17	96,67
Océanie/Australie	16,2	1,82	98,49
Afrique	13,4	1,51	100
Total	**883,3**	**100**	

Source : www.internetworldstats.com/stats
Note : Pour connaître la liste des pays inclus dans chaque zone, voir la source des données.

Troisièmement, il est souvent nécessaire de présenter des données complémentaires, quand elles sont disponibles, pour faciliter la compréhension des données principales. Ici, par exemple, on peut souhaiter connaître les populations des zones concernées, ainsi que la population mondiale, afin de rapporter le nombre d'utilisateurs d'internet à un indicateur des utilisateurs potentiels.

Le tableau ci-dessous donne le nombre d'utilisateurs d'Internet en pourcentage de la population de chaque zone, et la population mondiale de chaque zone en pourcentage de la population mondiale totale. Le tableau fournit également, sur la dernière ligne, le nombre total d'utilisateurs d'Internet, ce qui permet de retrouver les données brutes en multipliant les pourcentages par les totaux de la colonne correspondante.

Par exemple, si l'on veut retrouver le nombre d'utilisateurs d'internet en Asie, il suffit d'effectuer l'opération suivante :

Nombre d'utilisateurs d'internet en Asie = (34,02/100) * 888,3 = 302,2

De même, si l'on veut retrouver la population d'Asie, il suffit d'effectuer l'opération suivante:

Population d' Asie = (9,61/100) * 6411 = 3612

Tableau 3 : Utilisateurs d'Internet et population exprimés pour chaque zone géographique en pourcentage des totaux respectifs (Mars 2005)

Zones géographiques (1)	**Nombre d'utilisateurs d'Internet en % de la population de chaque zone**	**Population de chaque zone en % de la population mondiale**
Asie	34,02	9,61
Europe	29,22	11,48
Amérique du Nord	24,92	51,58
Amérique du Sud/Caraïbes	6,33	8,59
Moyen-Orient	2,17	4,07
Océanie/Australie	1,82	0,52
Afrique	1,51	14,14
Total (en millions)	888,3	**6411**

Source : www.internetworldstats.com/stats
Note : Pour connaître la liste des pays inclus dans chaque zone, voir la source des données.

Cette présentation des données d'utilisateurs d'internet et de la population mondiale, ainsi que des pourcentages qui en découlent, permet par exemple de faire apparaître que le classement par zones des pourcentages d'utilisateurs d'internet n'est pas identique à celui du classement par zones des pourcentages de la population mondiale. Par exemple, l'Afrique, qui constitue le 3ème groupe en termes de pourcentage de population, se trouve en dernière position pour ce qui est des utilisateurs d'internet. Inversement, l'Amérique du Nord, qui est au dernier rang en termes de pourcentage de population, est au troisième rang des utilisateurs d'Internet. Le **degré de corrélation** entre deux variables, ici le pourcentage d'utilisateurs d'internet et de la population totale, sera étudié dans la seconde partie de ce mémento.

B –Tableaux de données quantitatives

1) Variable quantitative discrète, valeurs connues individuellement

Exemple : on interroge 100 ménages sur le nombre de pièces de leur logement. La variable « nombre de pièces » est quantitative et discrète (les valeurs sont dénombrables). En outre, les valeurs, n'ayant pas été groupées, sont connues individuellement. On obtient le tableau ci-dessous, où x_i représente le nombre de pièces et n_i les effectifs correspondants :

Tableau 4 : **Nombre de pièces du logement (x_i)**

x_i	Effectifs (n_i)
1	5
2	30
3	40
4	20
5	5

2) Variable quantitative discrète, valeurs regroupées

Exemple : on interroge 100 ménages sur le nombre de pièces de leur logement. La variable « nombre de pièces » est quantitative et discrète (les valeurs sont dénombrables). Cette fois, les valeurs ont été groupées. On obtient le tableau ci-dessous :

Tableau 5 : Nombre de pièces du logement (x_i)
Groupement par classes

(x_i)	Effectifs (n_i)
[1-3[	35
[3-5]	65

Lorsque les données sont groupées, il faut porter attention aux crochets (les signes « [» et «] ») car ce sont eux qui indiquent si les valeurs limites sont incluses ou non dans la classe. Par exemple, dans le tableau ci-dessus, le groupe [1-3[inclut les ménages dont le logement n'a qu'une seule pièce (c'est le signe « [» qui marque l'inclusion, mais exclut les ménages qui ont 3 pièces (c'est le signe « [»).

La valeur « 3 » ayant été exclue du groupe [1-3[, elle sera nécessairement incluse dans le groupe [3-5]. Cela correspond à la propriété évoquée dans le chapitre 1, d'après laquelle les modalités d'un caractère (ici les valeurs d'une variable) sont exhaustives et incompatibles.

3) Variable quantitative continue, valeurs connues individuellement

Exemple : on dispose d'un échantillon de 122 réponses d'étudiants à la question « À quel âge avez-vous obtenu votre bac ? ». Bien qu'il s'agisse d'une variable quantitative continue, les données sont présentées par âge et non par groupe d'âge. On a donc le tableau ci-après :

Tableau 6 : Âge d'obtention du bac (x_i)
Groupement par valeurs

x_i	n_i
16	5
17	25
18	45
19	20
20	15
21	8
22	4

4) Variable quantitative continue, données groupées

Exemple 1 : on dispose d'un échantillon de 122 réponses d'étudiants à la question « À quel âge avez-vous obtenu votre bac ? ». Cette fois, les données sont présentées par groupe d'âge.

Tableau 7 : Âge d'obtention du bac (x_i)
Groupement par classes

x_i	n_i
[16-18[	30
[18-20[	80
[20-22]	12

2 • GRAPHIQUES

A – Importance des graphiques

Il est parfois indispensable de recourir à la présentation graphique des données. Le tableau 6 ci-dessous, connu sous l'appellation de quartet d'Anscombe, illustre parfaitement ce point.

Tableau 6 : Séries ayant des moyennes identiques (9 pour X et 7,5 pour Y)

Série 1		Série 2		Série 3		Série 4	
X_1	Y_1	X_2	Y_2	X_3	Y_3	X_4	Y_4
10	8,04	10	9,14	10	7,46	8	6,58
8	6,95	8	8,14	8	6,77	8	5,76
13	7,58	13	8,74	13	12,74	8	7,71
9	8,81	9	8,77	9	7,11	8	8,84
11	8,33	11	9,26	11	7,81	8	8,47
14	9,96	14	8,10	14	8,84	8	7,04
6	7,24	6	6,13	6	6,08	8	5,25
4	4,26	4	3,10	4	5,39	19	12,50
12	10,84	12	9,13	12	8,15	8	5,56
7	4,82	7	7,26	7	6,42	8	7,91
5	5,68	5	4,74	5	5,73	8	6,89

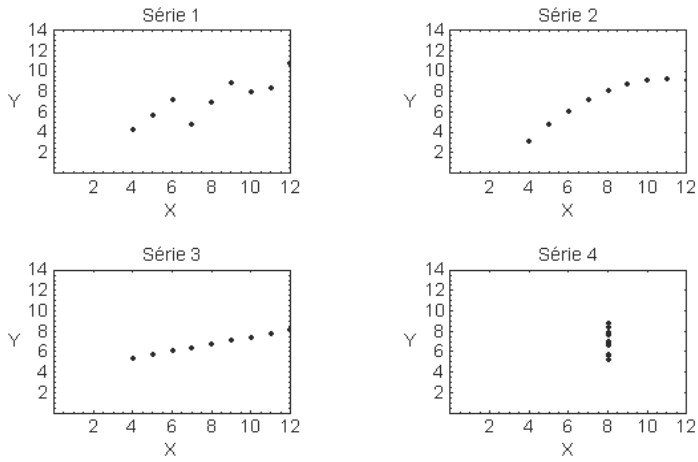
Source : Anscombe, Francis J. (1973) *Graphs in statistical analysis*. American Statistician, 27, 17–21.

Si l'on calcule les moyennes arithmétiques simples de ces deux séries (voir le chapitre 3 pour la définition de la moyenne arithmétique simple), on constate que la moyenne de X_1, X_2, X_3 et X_4 est égale à 9, tandis que la moyenne de Y_1, Y_2, Y_3, Y_4 est égale à 7,5.

Certes, il s'agit d'une curiosité, mais celle-ci illustre parfaitement que pour décrire une série de chiffres (ici deux séries de chiffres), il ne suffit parfois pas de calculer des indicateurs numériques. Dans cet exemple, l'usage d'un indicateur simple tel que la moyenne dissimule en fait une très grande diversité.

La figure 1 ci-après montre en fait les nuages de point associés à chacune des séries $\{X_1,Y_1\}$, $\{X_2,Y_2\}$, $\{X_3,Y_3\}$ et $\{X_4,Y_4\}$.

Figure 1 : Séries ayant des moyennes identiques
mais les nuages de points révèlent des formes extrêmement différentes



La présentation des données statistiques sous forme de graphiques joue un rôle essentiel pour permettre à un auditoire ou à des lecteurs de suivre une explication. Ne dit-on pas qu'un beau graphique vaut mieux qu'un long discours. On dit d'ailleurs que Michael DELL est arrivé un jour à une assemblée générale d'actionnaires avec pour tout document le graphique qui montrait l'évolution spectaculaire du cours de l'action des entreprises DELL au cours des 5 dernières années…

La diversité des présentations graphiques ne connaît d'autres limites que celles de l'imagination. Nous nous bornerons dans les pages qui suivent à passer en revue les graphiques les plus connus et les mieux adaptés aux données qu'il s'agit de représenter.

B – Données individuelles

Lorsque l'on veut représenter graphiquement toutes les unités statistiques d'une population à un caractère ou à une variable, on dispose de deux graphiques : la **ligne** et le graphique dit « **tige et feuilles** » (de l'anglais « stem and leaf »).

1) La ligne

Exemple 1 : Soit la série de chiffres :

$$\{8, 2, 3, 7, 4\}$$

où aucune unité n'a la même valeur.

On obtient alors la représentation graphique suivante :

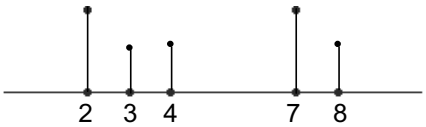
Figure 2 : Représentation graphique en ligne quand les unités statistiques sont peu nombreuses et connues individuellement et non répétées.

En revanche, si certaines données sont répétées, comme dans l'exemple ci-dessous, il faut passer à une représentation des données sous forme groupée, ce qui est l'objet de la partie C de cette sous-section 2.

Exemple 2 : Soit la série de chiffre où le 7 et le 2 sont répétés 2 fois :

{8, 2, 3, 7, 4, 7, 2}

Figure 3 : Représentation graphique quand les unités statistiques sont peu nombreuses et connues individuellement **mais répétées**.



Remarques :

1) À la représentation en **ligne horizontale**, on peut parfois préférer une représentation en **ligne verticale**.

2) Cette représentation en ligne peut être raffinée, pour donner naissance à un graphique analytique, dit « **boîte à moustaches** » (de l'anglais « **Box and Whiskers** »), que nous aborderons dans le chapitre 4, car sa compréhension nécessite l'acquisition de notions telles que la **médiane** et les **quartiles**.

2) Le graphique « tige et feuilles »

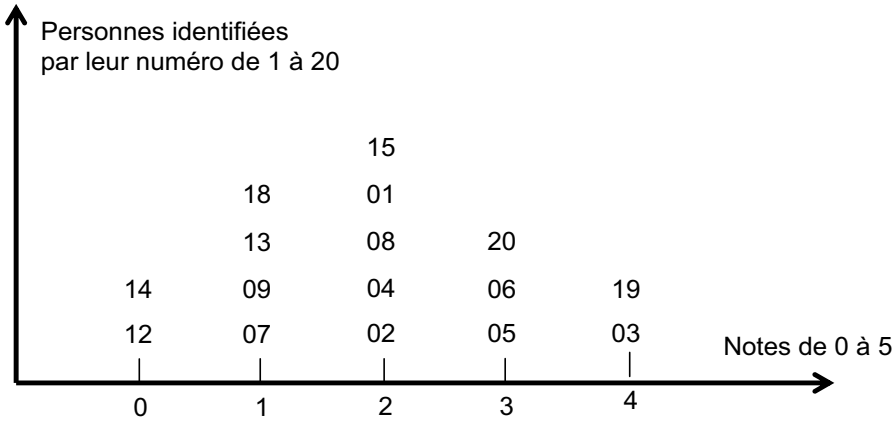
Ce graphique très original consiste à empiler des unités en conservant leur identification (un numéro, un nom, etc.). De cette façon, aucune donnée initiale n'est absente du graphique et chacune peut facilement être repérée.

Exemple 1 : Soit 20 personnes, repérées par un numéro de 1 à 20, à qui des notes allant de 0 à 5 ont été attribuées.

Notes = {{0, 12}, {0, 14}, {1, 7}, {1, 9}, {1, 13}, {1, 18}, {2, 4}, {2, 8}, { 2, 11}, {2, 15}, {2, 16}, {3, 17}, {3, 12}, {4, 5}, {4, 6}, {4, 20}, {5, 3}, {5, 19}}

Dans chaque couple de données, le premier chiffre correspond à la note (de 0 à 5), c'est la « tige » et le second sert à identifier la personne par un numéro allant de 1 à 20, c'est « les feuilles ». La représentation **tiges et feuilles** donne la figure 4.

Figure 4 : Diagramme « tige et feuilles »



C – Données groupées par modalités ou valeurs

Que les données soient regroupées par modalité, comme c'est le cas pour les groupements qualitatifs, ou par valeurs, comme c'est le cas pour les groupements quantitatifs, on dispose de nombreuses représentations graphiques. Nous limiterons notre présentation aux plus connues, à savoir : le diagramme en bâtons, le diagramme en barres et le nuage de points, de l'anglais « scatter plot ».

1) Diagramme en bâtons

C'est peut-être la représentation la plus simple qui soit. En réalité, le **diagramme en bâtons** s'inspire directement de la présentation tige et feuilles, mais le contenu en information est moins riche.

Exemple 1 : On interroge 11 personnes sur leurs préférences concernant les 4 produits A,B,C,D. Chaque personne doit choisir seulement un produit. On obtient les résultats groupés suivants :

{{A, 4}, {B, 4}, {C, 1}, {D,1}}

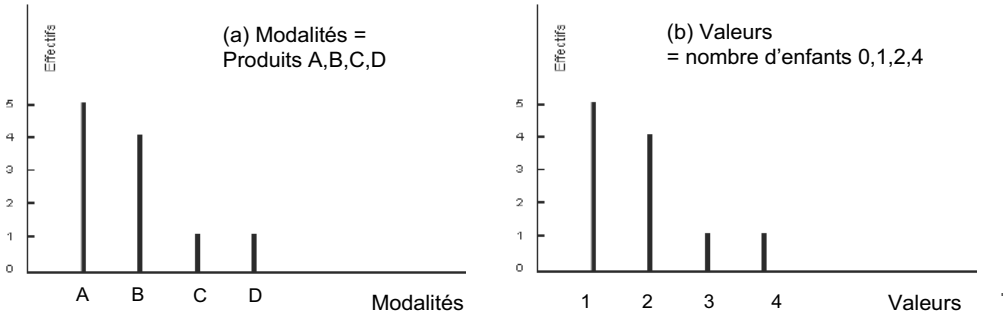
Dans chaque couple de données, le premier chiffre correspond au produit (A,B,C,D) et le second correspond au nombre de personnes qui ont choisi ce produit. La figure 5 (a) illustre le résultat.

Si le regroupement se fait par valeur, on a par exemple les couples :

{{1, 4}, {2, 4}, {3, 1}, {4, 1}}

Où le premier chiffre de chaque couple correspond par exemple au nombre d'enfants. On obtient alors le graphique de la figure 5(b).

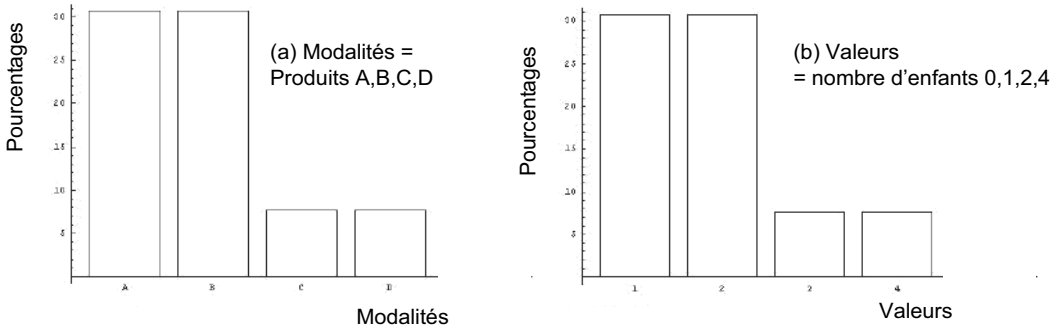
Figure 5 : Diagrammes en bâtons



2) Diagramme en barres

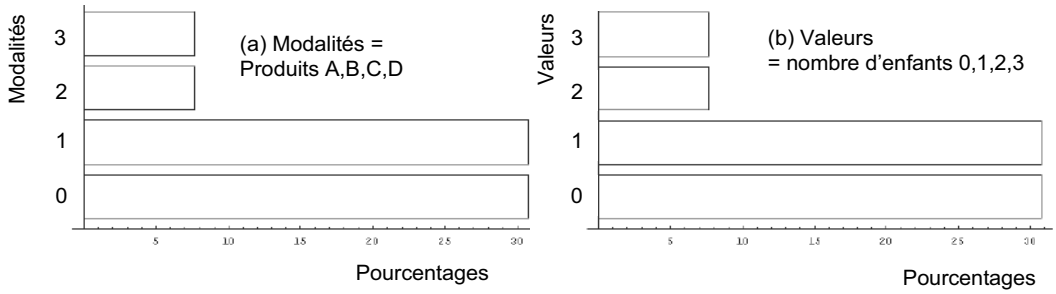
Le **diagramme en barres** repose sur le même principe que le diagramme en bâtons, sauf qu'au lieu de bâtons, on a des barres rectangulaires de base identique et identiquement espacées les unes des autres. La taille de la base, ainsi que celle de l'espacement n'ont pas de signification particulière. L'espacement n'est pas obligatoire. La figure 6 représente les mêmes données que la figure 5, mais ces données sont exprimées en pourcentage.

Figure 6 : Diagramme en barres verticales



Le diagramme en barre est souvent présenté de façon horizontale. Ainsi, le graphique de la figure 6 peut-il être présenté ainsi :

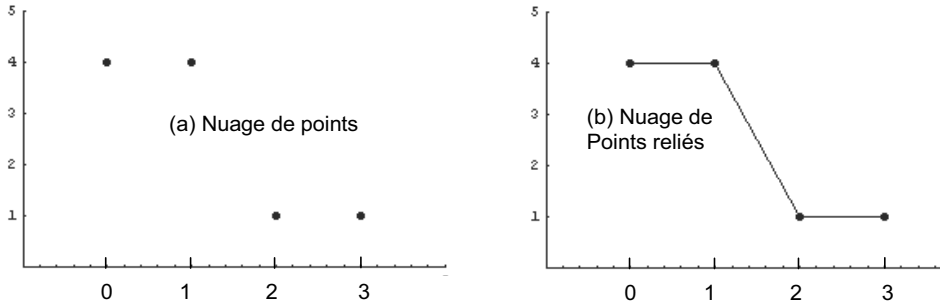
Figure 7 : Diagramme en barres horizontales



3) Nuage de points dans le cas d'une série unidimensionnelle

Pour des raisons pédagogiques, la figure 2 de ce chapitre a présenté des graphiques sous forme de nuages de points concernant des variables bidimensionnelles. Il y avait deux séries, et chaque point avait pour coordonnée un élément de chaque série. Mais le nuage de points peut aussi être employé pour représenter graphiquement une simple série de chiffres. Les données des figures 5 à 7 peuvent également être représentées par un nuage de points ou par une ligne joignant ces points (voir la figure 8, qui reprend les données précédentes dans l'hypothèse quantitative.)

Figure 8 : Nuage de points, reliés et non reliés – nombre d'enfants par foyer



D – Camembert ou graphique « en tarte » ?

Les anglo-saxons l'appellent « Pie Chart » c'est-à-dire, littéralement « graphique en tarte ». En France, on l'appelle le camembert. Ce graphique universel convient à toutes les données, dès l'instant où il s'agit d'exprimer des parts ou des pourcentages.

Exemple : Soit les chiffres d'affaires en millions d'euros des quatre principales entreprises du marché d'un produit (pour simplifier, on suppose que ces entreprises contrôlent la totalité du marché) :

Tableau 7 : Chiffre d'affaires en millions d'euros de quatre entreprises qui contrôlent un marché

Entreprise	Chiffre d'affaires	Part de marché
A	50	31,25
B	70	43,75
C	10	6,25
D	30	18,75
Total	**160**	**100**

La part de marché (colonne 3) n'est en fait qu'un pourcentage. Chaque ligne de la colonne 2 est divisée par la dernière ligne (total) et multipliée par 100.

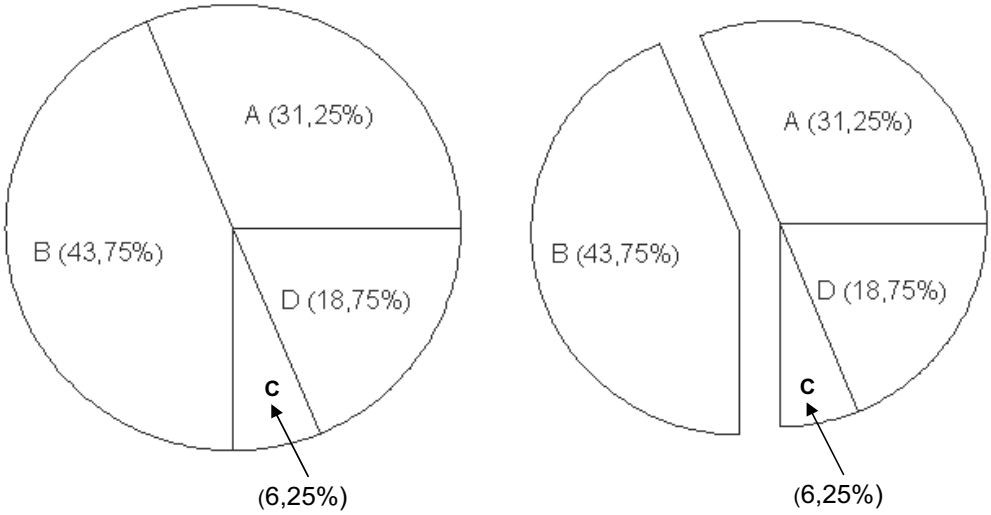
Notons qu'il s'agit d'un caractère qualitatif, les modalités étant les quatre entreprises. Pour faire le graphique en camembert, il reste à calculer la part que le chiffre d'affaires de chacune de ces entreprises représente dans 360° (voir le tableau 8 ci-dessous).

Tableau 8 : Chiffre d'affaires en millions d'euros de quatre entreprises qui contrôlent un marché

Entreprise	Part de marché	Degrés
A	31,25	(31,25 *360) / 100 = 112,5
B	43,75	(43,75 *360) / 100 = 157,5
C	6,25	(6,25 *360) / 100 = 22,5
D	18,75	(6,25 *360) / 100 = 22,5
Total	**100**	**360**

La dernière colonne du tableau 7 va nous permettre de dessiner le camembert, puis de « couper les parts ». Il suffit pour cela de tracer un cercle, puis au moyen d'un rapporteur, de déterminer les angles correspondant à chaque part. On obtient alors le résultat voulu. La figure ci-dessous illustre 2 variantes du même graphique. Dans la seconde variante, l'entreprise qui a la part de marché la plus élevée est détachée du lot.

Figure 9 : Le camembert ou « pie chart »



Le camembert peut aussi servir à représenter des variables quantitatives, y compris des variables quantitatives groupées par classes.

E – L'histogramme

L'histogramme convient particulièrement aux variables quantitatives quand celles-ci sont regroupées par classes. Parfois les classes ont des amplitudes égales. C'est le cas le plus évident. Parfois, cependant, les amplitudes des classes sont différentes. Il faut alors opérer une correction en suivant la méthode indiquée ci-après.

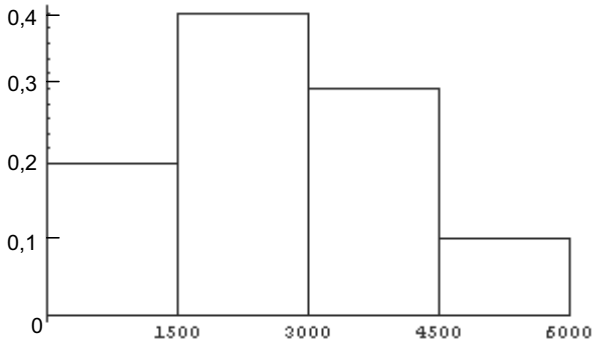
Exemple 1 : Soit 100 ménages distribués selon leur revenu mensuel en euros. On définit des classes d'amplitudes égales à 1 500 euros.

Tableau 9 : Répartition d'un échantillon de 100 ménages par classe de revenu mensuel (amplitude de classe = 1 500 euros)

Classe de revenu	n_i	f_i
[0;1500[	20	0,2
[1500;3000[	40	0,4
[3000;4500[	30	0,3
[4500;6000[	10	0,1

L'histogramme peut-être construit à partir des effectifs (les n_i) ou à partir des fréquences (et d'ailleurs aussi en prenant les pourcentages). Contrairement au diagramme en barre, avec lequel il ne faut pas le confondre, les rectangles qui composent l'histogramme ont une base qui est définie par l'amplitude de la classe qu'ils représentent et, de plus, ils sont collés les uns aux autres.

Figure 10 : Histogramme correspondant aux données du tableau 9



Exemple 2 : Supposons que l'on regroupe les données de l'exemple 1 en classes d'amplitudes inégales ([0-1500[;[1500-4500[,[4500-6000[).

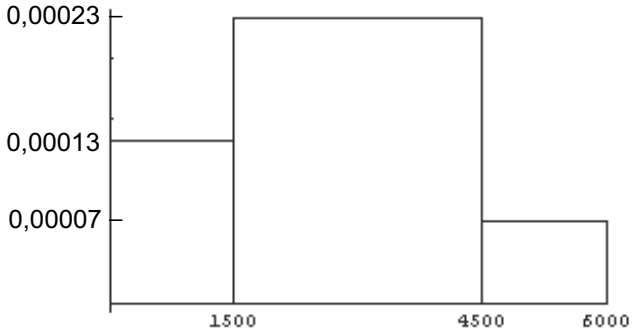
Il faut dans ce cas effectuer une correction pour tenir compte des différences d'amplitude. Il convient en fait de diviser la fréquence de chaque classe par l'amplitude correspondante. On obtient ainsi l'**amplitude corrigée** (h_i).

Tableau 10 : Calcul de l'amplitude corrigée

Classe de revenu	Amplitude de classe (a_i)	n_i	f_i	$h_i = f_i/a_i$
[0;1500[	1500	20	0,2	0,00013
[1500;4500[	3000	70	0,7	0,00023
[4500;6000[	1500	10	0,1	0,00007

Sur l'histogramme de la figure 11, on aura donc l'amplitude corrigée en abscisse et des classes d'inégales amplitudes en ordonnée.

Figure 10 : Histogramme avec amplitudes inégales
(voir le tableau 10 pour les calculs)



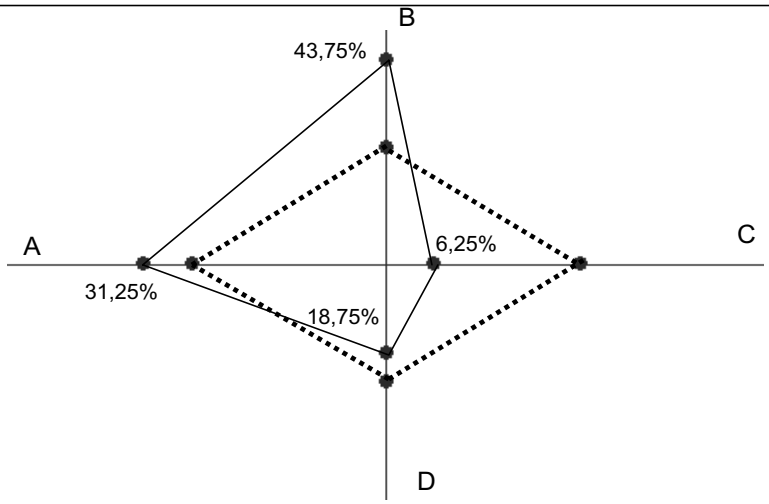
F – L'utilisation des graphiques à des fins de comparaisons

C'est dans les possibilités de comparaisons qu'ils offrent que les graphiques sont particulièrement utiles : comparaisons dans le temps, comparaisons spatiales, etc.

1) Le radar, excellent moyen d'effectuer des comparaisons visuelles

La figure 11 utilise le graphique dit « en radar » afin de comparer la répartition réelle des parts de marché des 4 entreprises A, B, C et D avec une répartition égalitaire où chacune aurait 25% du marché (cette répartition égalitaire est représentée par le losange en pointillé). Les parts de marché réelles sont indiquées sur chaque axe. On voit ainsi immédiatement que A et B ont une part de marché supérieure à la répartition égalitaire et B et C une part de marché inférieure. On peut à partir de là calculer combien il faut retrancher à A et à B (et combien par conséquent il faut redistribuer à C et D) pour revenir à une répartition égalitaire).

Figure 11 : Le graphique en radar pour représenter et comparer les parts de marché des entreprises du tableau 7

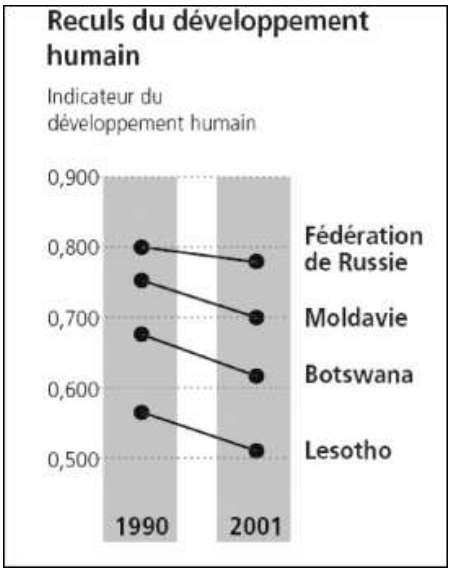


2) Comparaisons dans le temps

Il est facile de voir que le graphique en radar permet aussi de comparer les parts de marché des quatre entreprises A,B,C et D du tableau 7 en deux, voire trois ou quatre points du temps. On aboutirait ainsi à une « toile d'araignée » dont la complexité irait cependant grandissante avec le nombre d'années. Il est sage de se limiter à une comparaison de deux périodes.

Toutefois, le radar n'est pas le seul moyen d'effectuer des comparaisons temporelles, loin de là. La figure 12, ci-dessous illustre une façon très simple (et malheureusement très réaliste) de comparer deux situations éloignées dans le temps.

Figure 12 : Une façon très simple de représenter l'évolution du développement humain sur une décennie pour quatre pays peu développés. Ces quatre pays sont les seuls pour lesquels l'indice du développement humain a régressé au cours de la décennie 1990.



Source : PNUD, Rapport sur le développement humain 2003, p. 40. Sur la méthode de calcul de l'indicateur du développement humain, voir le chapitre 7 de ce mémento.

3) Les graphiques de séries chronologiques

Pour les comparaisons dans le temps, rien ne remplace cependant la **série chronologique**. Typiquement, les années sont en abscisse et la valeur qui évolue dans le temps est en ordonnée.

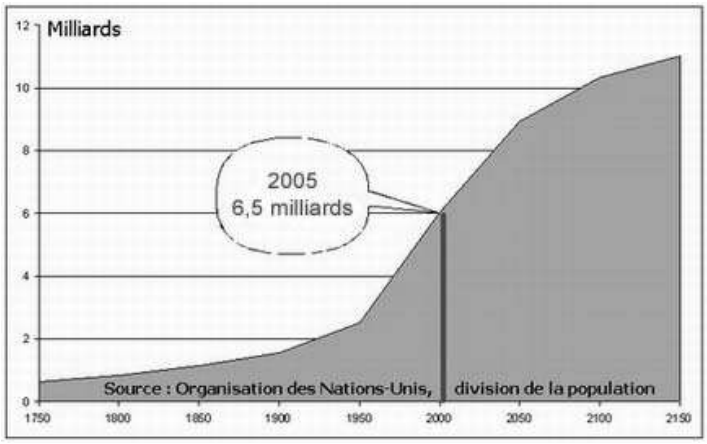
Les graphiques de séries chronologiques sont parmi les plus fréquents. Selon Edward R. TUFTE[1], qui a procédé à un tirage aléatoire de 4000 graphiques dans 15 magazines et journaux entre 1974 et 1980, il apparaît que plus de 75% d'entre eux sont des graphiques de séries chronologiques.

Le graphique de la figure 13 ci-après représente l'évolution de la population mondiale telle qu'elle a été reconstituée (pour les données les plus éloignées) et projetée (pour les données futures) par les démographes de la division de la population de l'ONU.

---

[1] Edward R. TUFTE, *The Visual Display of Quantitative Information*, Graphics Press, LLC, 2001, page 25

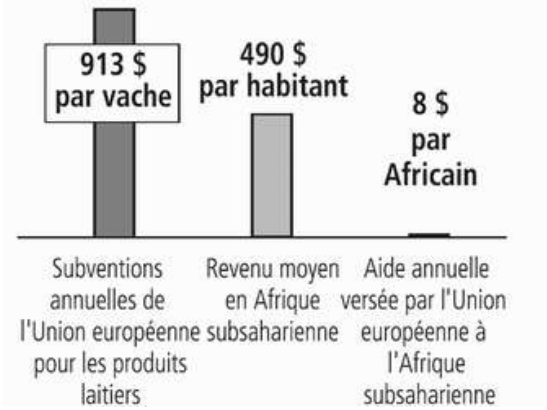
Figure 13 : Évolution de la population mondiale de 1750 à 2150 (projection)



4) Un beau graphique vaut parfois mieux qu'un long discours

Rien ne vaut un graphique lorsqu'on veut mettre en valeur une comparaison saisissante. La figure 14, par exemple, illustre de façon éclatante l'inefficacité (pour ne pas dire plus) de la répartition des aides dans le monde. On y voit que les subventions annuelles de l'Union Européenne par vache (et par an), sont presque deux fois supérieures au revenu moyen par habitant (et par an) en Afrique subsaharienne. Ce n'est pas les agriculteurs qui s'en plaindront.

Figure 14 : Un beau graphique vaut mieux qu'un long discours



Source : PNUD, Rapport sur le développement humain 2003, p. 155.

5) Les graphiques d'indices

Les indices se prêtent également particulièrement bien aux comparaisons sous forme graphique. Sans entrer dans le détail de leur étude (que nous réservons au chapitre 7), donnons-en une définition simple et illustrons-la par un exemple.

Un **indice** est un rapport de grandeurs exprimées dans la même unité, ce qui en fait un nombre sans dimension. Généralement, ce rapport est multiplié par 100. Lorsque l'on divise tous les éléments d'une série chronologique par l'un d'entre eux (et que l'on multiplie par 100) on transforme la série chronologique en indice. Ceci facilite les comparaisons avec une années de référence, laquelle aura alors pour valeur 100.

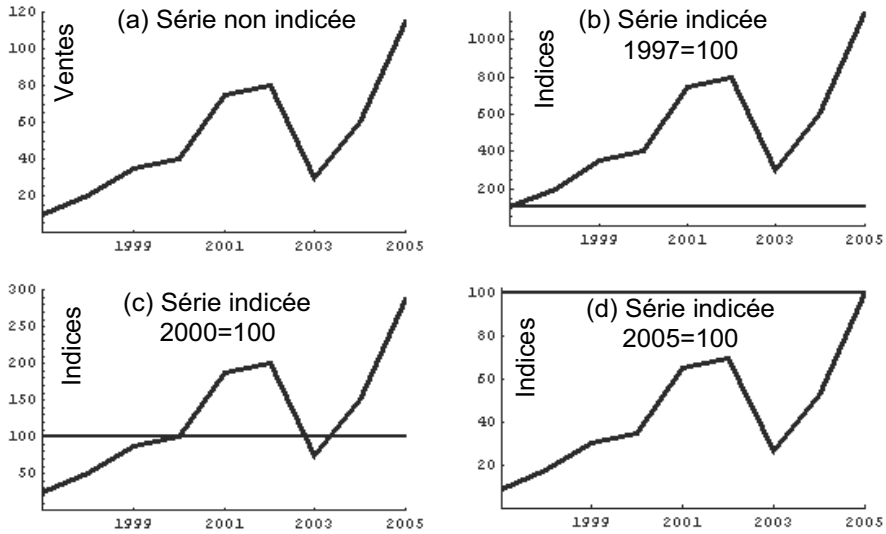
Exemple : Soit la série chronologique suivante qui indique le nombre d'avions d'un certain modèle, vendus par une grosse firme aéronautique.

Tableau 11 : Ventes annuelles d'un certain modèle d'avion

Années	1997	1998	1999	2000	2001	2002	2003	2004	2005
Ventes	10	20	35	40	75	80	30	60	115

La représentation graphique de base est celle d'une sérié chronologique. Toutefois, si on divise tous les chiffres par ceux de l'année 1997, « année de base » (et que l'on multiplie par 100) on obtient une série indice. La figure ci-dessous représente, outre la série initiale, trois choix d'indice : 1997, 2000 et 2005. À noter que le passage à un indice ne modifie que l'échelle de l'ordonnée, non la forme de la courbe.

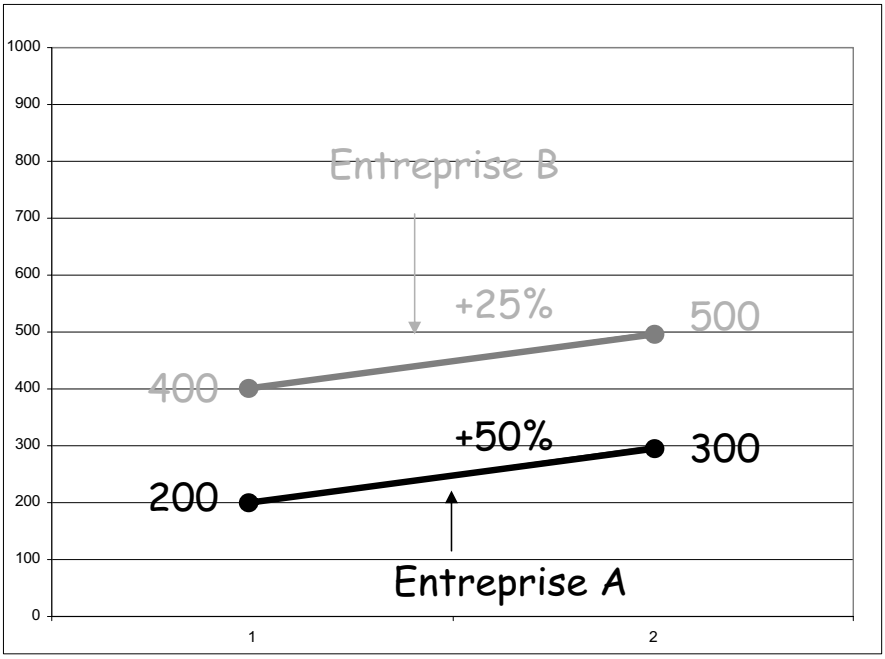
Figure 15 : Une série chronologique transformée en séries indicées



6) Les échelles semi-logarithmiques

Les échelles arithmétiques ne sont pas toujours les plus adaptées à la représentation graphique des caractères continus. Dans l'exemple suivant, les entreprises A et B ont augmenté leur production dans des proportions différentes et pourtant le graphique donne l'impression que la progression est identique en raison du parallélisme des progressions.

Figure 16 : Sur une échelle arithmétique les progressions parallèles semblent identiques

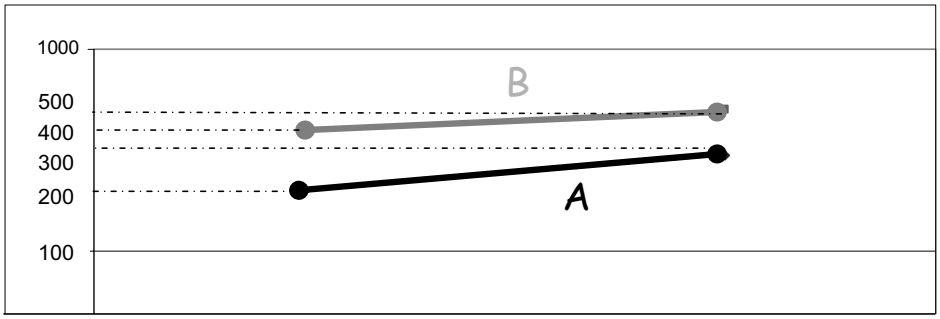


Pour remédier à cela, on peut prendre une échelle « semi-logarithmique » pour l'axe des ordonnées. Cela consiste à prendre le logarithme base 10 des valeurs en ordonnées. On obtient alors deux droites qui ne sont plus parallèles. La droite A est plus pentue, ce qui traduit une plus forte progression.

Tableau 12 : Quelques exemples de conversions de chiffres en logarithme décimal (de base 10)

10	100	200	300	400	500
log 10=1	log 100 =2	log 200 =2,3	log 300 = 2,477	log 400 =2,602	log 500 =2,698

Figure 17 : Sur une échelle logarithmique les différences de vitesse de progression se traduisent par des pentes différentes



CHAPITRE 3

Les caractéristiques de tendance centrale

Qu'elles soient non groupées ou au contraires groupées par valeurs ou par classes, les variables quantitatives peuvent être utilement résumées par des caractéristiques dites de « tendance centrale ». Ces nombres résumés sont ainsi appelés car ils privilégient les valeurs principales de la distribution, au détriment par exemple de ceux qui caractérisent la dispersion ou la concentration des valeurs d'une série.

Ces valeurs centrales sont les moyennes, la médiane et le mode. Nous exposerons leur mode de calcul et leur signification en distinguant pour chacune d'elles le cas des données non groupées et le cas des données regroupées (soit par valeurs, soit par classes).

1 • LES MOYENNES

A – La moyenne arithmétique

1) La moyenne arithmétique simple

Exemple : Soit la série de chiffres {8, 5, 9, 13, 25}. La **moyenne arithmétique** de cette série de chiffres se calcule ainsi :

$$\overline{x} = \frac{8+5+9+13+25}{5} = \frac{60}{5} = 12$$

Comme nous l'avons indiqué dans le chapitre 1, nous ne distinguerons pas la moyenne de la population et la moyenne de l'échantillon. Par conséquent, nous traitons ici la série de chiffres sans nous préoccuper de savoir s'il s'agit d'une population ou d'un échantillon.

Signification de la moyenne : Construisons un tableau avec pour première colonne la série de chiffres et pour seconde colonne l'écart de chacun des chiffres à la moyenne que nous venons de calculer ($\overline{x} = 12$) :

Tableau 1 : La somme des écarts à la moyenne est nulle

x_i	$x_i - \overline{x}$
8	-4
5	-7
9	-3
13	1
25	1
	$\sum_{i=1}^{5}(x_i - \overline{x}) = 0$

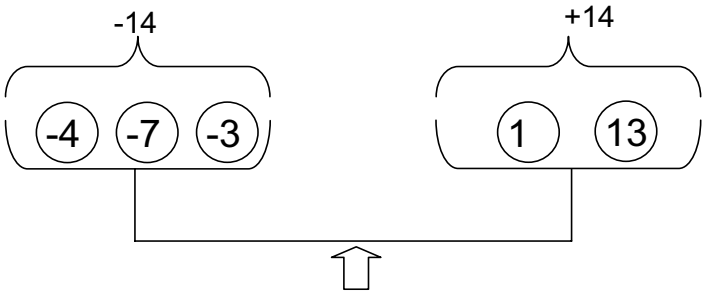
Quand on soustrait la moyenne arithmétique à chacun des chiffres de la série, on observe la propriété suivante :

1) La somme des écarts à la moyenne est nulle :

$$(-4)+(-7)+(-3)+(+1)+(+13)=0$$

2) Ou, ce qui revient au même, mais est plus imagé, la somme des écarts positifs est égale à la somme des écarts négatifs, au signe près.

Schéma 1 : En valeur absolue, la somme des écarts négatifs (panneau de gauche) est égale à la somme des écarts positifs (panneau de droite)



Moyenne arithmétique : 12

Formule générale de la moyenne arithmétique simple : Soit $\{x_1, x_2, \ldots, x_n\}$ une série de chiffres. La formule de la moyenne arithmétique de cette série est donnée par :

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (1)$$

2) La moyenne arithmétique pondérée

Exemple 1 : Soit la série de chiffres {8, 13, 5, 8, 5, 9, 13, 25, 13, 9}. Certains chiffres, comme le 8, le 9 ou le 13 sont répétés. On peut simplifier la présentation en regroupant les données par valeurs (voir le tableau 2). La troisième ligne est le produit des deux premières. En effet, on a par exemple :

$$x_1 = 5 \qquad n_1 = 2 \qquad n_1 \cdot x_1 = 2 \times 5 = 10$$

$$x_2 = 8 \qquad n_2 = 2 \qquad n_2 \cdot x_2 = 2 \times 8 = 16$$

Et ainsi de suite (voir le tableau 2).

Tableau 2 : Calcul de la moyenne arithmétique pondérée

x_i	5	8	9	13	25
n_i	2	2	2	3	1
$n_i \cdot x_i$	10	16	18	39	25

$$\sum_{i=1}^{5} n_i \cdot x_i = 108$$

La **moyenne pondérée** se calcule alors en faisant la somme pondérée c'est-à-dire la somme des $n_i \cdot x_i$ et en divisant par n. Elle est égale à :

$$\overline{x} = \frac{(5\times 2)+(8\times 2)+(9\times 2)+(13\times 2)+(25\times 1)}{10} = \frac{108}{10} = 10,8$$

Formule générale de la moyenne arithmétique pondérée : Soit $\{x_1, x_2, \ldots, x_h\}$ une série de chiffres et $\{n_1, n_2, \ldots, n_h\}$ les effectifs correspondants. La formule de la moyenne arithmétique pondérée de cette série est donnée par :

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{h}(n_i \cdot x_i) \qquad (2)$$

Exemple 2 : Soit la série de chiffres {8,13,5,8,5,9,13,25,13,9,35,44,54,28}. Supposons que l'on regroupe les valeurs en 3 catégories comme dans le tableau 3 ci-dessous. Dans ce cas, il faut calculer le centre de chaque classe, c_i, **c'est-à-dire la somme des extrémités de classe divisée par 2** et appliquer la formule de la moyenne pondérée.

Tableau 3 : Calcul de la moyenne arithmétique quand les valeurs sont groupées par classes

Classes	n_i	c_i	$n_i \cdot c_i$
[5-13[	6	9	54
[13-28[	3	7,5	22,5
[28-54[	5	41	205

$$\sum_{i=1}^{3} n_i \cdot c_i = 281,5$$

On applique donc la formule (2), mais en remplaçant x_i par c_i :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{h}(n_i \cdot c_i) \qquad (3)$$

Dans notre exemple, on a donc :

$$\bar{x} = \frac{(6\times 9)+(3\times 7{,}5)+(5\times 41)}{14} = \frac{54+22{,}5+205}{14} = \frac{281{,}5}{14} \cong 20{,}11$$

3) La moyenne élaguée

Exemple : Soit la série de notes d'un élève au cours de l'année {12, 13, 11, 14, 2}. Si l'on calcule la moyenne arithmétique simple on obtient :

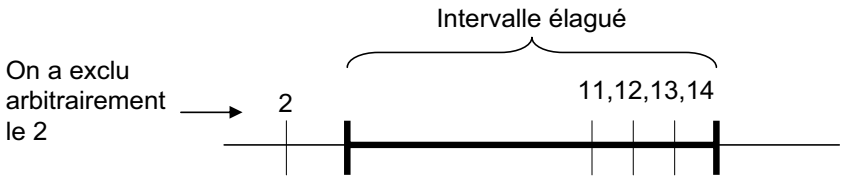
$$\bar{x} = \frac{12+13+11+14+2}{5} = \frac{52}{5} = 10{,}4$$

Par contre, si on retire le « 2 » et que l'on recalcule la **moyenne élaguée** sur 4 notes, on obtient :

$$\bar{x} = \frac{12+13+11+14}{4} = \frac{50}{4} = 12{,}5$$

Dans ce cas, on a retiré le « 2 », qui est considéré comme un accident, mais qui, si on le maintient dans la série, fait fortement baisser la moyenne. Dans certains cas, on retire les valeurs extrêmes et on calcule la moyenne uniquement sur un intervalle de valeurs élagué, conformément au schéma 2 ci-dessous. Le principe est identique quand les données sont groupées par valeurs ou par classes.

Schéma 2 : La moyenne élaguée



B – La moyenne quadratique

1) La moyenne quadratique simple

Exemple : Soit la série de chiffres {-4, -2, 0, 2, 4}. Si l'on calcule la moyenne arithmétique simple on obtient zéro.

Parfois, on souhaite obtenir une caractéristique de tendance centrale ayant une valeur positive là où le calcul de la moyenne arithmétique simple aurait donné zéro. On calcule alors la **moyenne quadratique simple** en additionnant le carré de toutes les valeurs de la série et en prenant la racine carrée du total. Autrement dit, dans notre exemple :

$$Q = \sqrt{\frac{(-4)^2 + (-2)^2 + (0)^2 + (2)^2 + (4)^2}{5}} = \sqrt{\frac{16+4+0+4+16}{5}} = \sqrt{\frac{40}{5}} = \sqrt{8} \cong 2{,}83$$

Formule générale de la moyenne quadratique simple : Soient $\{x_1, x_2, \ldots, x_n\}$ une série de chiffres. La formule de la moyenne quadratique simple de cette série est donnée par :

$$Q = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2} \qquad (4)$$

2) La moyenne quadratique pondérée

Soit $\{x_1, x_2, \ldots, x_h\}$ une série de chiffres et $\{n_1, n_2, \ldots, n_h\}$ les effectifs correspondants. La formule de la **moyenne quadratique pondérée** de cette série est donnée par :

$$Q = \sqrt{\frac{1}{n}\sum_{i=1}^{h}\left(n_i \cdot x_i^2\right)} \qquad (5)$$

Exemple : Soit le tableau 4 ci-dessous :

Tableau 4 : Calcul de la moyenne quadratique pondérée

x_i	n_i
25	10
8	16
4	25
12	20

Il suffit de rajouter deux colonnes, une pour x_i^2 et une pour $n_i \cdot x_i^2$ (voir le tableau 5)

Tableau 5 : Calcul de la moyenne quadratique pondérée

x_i	n_i	x_i^2	$n_i \cdot x_i^2$
25	10	625	6250
8	16	64	1024
4	25	16	400
12	20	144	2880

$$\sum_{i=1}^{4}\left(n_i \cdot x_i^2\right) = 10554$$

En appliquant la formule (5) on obtient :

$$Q = \sqrt{\frac{1}{n}\sum_{i=1}^{h}\left(n_i \cdot x_i^2\right)} = \sqrt{\frac{10554}{71}} \cong 12{,}1921$$

Lorsque les valeurs sont regroupées en classes, il faut calculer les centres de classes et appliquer ensuite la formule (5) en remplaçant x_i par c_i.

C – La moyenne géométrique

1) La moyenne géométrique simple

Soit $\{x_1, x_2, \ldots, x_n\}$ une série de chiffres. La formule de la **moyenne géométrique simple** de cette série est donnée par :

$$G = \left[\prod_{i=1}^{n} x_i\right]^{\frac{1}{n}} \qquad (6)$$

Exemple : Soit la série de chiffres {8, 5, 9, 13, 25}. La moyenne géométrique de cette série est égale à :

$$G = \left[8 \times 5 \times 9 \times 13 \times 25\right]^{\frac{1}{5}} = \sqrt[5]{117000} \cong 10{,}32$$

2) La moyenne géométrique pondérée

Soit $\{x_1, x_2, \ldots, x_h\}$ une série de chiffres et $\{n_1, n_2, \ldots, n_h\}$ les effectifs correspondants.
La formule de la **moyenne géométrique pondérée** de cette série est donnée par :

$$G = \left[\prod_{i=1}^{h} x_i^{n_i}\right]^{\frac{1}{n}} \qquad (7)$$

Exemple : Soit les chiffres du tableau 4

Pour calculer la moyenne géométrique pondérée, on peut passer par les logarithmes népériens (ln) :

$$G = \left[\prod_{i=1}^{h} x_i^{n_i}\right]^{\frac{1}{n}} = \left[25^{10} \times 8^{16} \times 4^{25} \times 12^{20}\right]^{\frac{1}{71}}$$

$$\ln G = \frac{1}{71}\left[10\ln 25 + 16\ln 8 + 25\ln 4 + 20\ln 12\right]$$

$$\ln G = \frac{1}{71}\left[32{,}1888 + 32{,}2711 + 34{,}6574 + 49{,}6981\right]$$

$$\ln G = \frac{149{,}815}{71} \cong 2{,}1100704$$

$$G = e^{2{,}1100704} = 8{,}2488$$

D – La moyenne harmonique

1) La moyenne harmonique simple

Soit $\{x_1, x_2, \ldots, x_n\}$ une série de chiffres. La formule de la **moyenne harmonique simple** de cette série est donnée par :

$$H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} \tag{8}$$

Exemple : Soit la série de chiffres {8, 5, 9, 13, 25}. La moyenne harmonique de cette série est égale à :

$$H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} = \frac{5}{\frac{1}{8} + \frac{1}{5} + \frac{1}{9} + \frac{1}{13} + \frac{1}{25}} = \frac{5}{0{,}5530342} \cong 9{,}04$$

2) La moyenne harmonique pondérée

Soit $\{x_1, x_2, \ldots, x_h\}$ une série de chiffres et $\{n_1, n_2, \ldots, n_h\}$ les effectifs correspondants. La formule de la **moyenne harmonique pondérée** de cette série est donnée par :

$$H = \frac{n}{\sum_{i=1}^{h} \frac{n_i}{x_i}} \tag{9}$$

Exemple 1 : Soit les chiffres du tableau 4. Pour calculer la moyenne harmonique pondérée, on applique la formule (9).

$$H = \frac{n}{\sum_{i=1}^{h} \frac{n_i}{x_i}} = \frac{71}{\frac{10}{25} + \frac{16}{8} + \frac{25}{4} + \frac{20}{12}} = \frac{71}{0,4 + 2 + 6,25 + 1,66667} = \frac{71}{10,3167} = 6,882$$

Exemple 2 : Une petite usine abrite 2 machines. La première machine a produit 500 pièces à la vitesse de 100 pièces par heure. Une seconde machine a produit 300 pièces à la vitesse de 60 pièces par heure. Calculer la vitesse moyenne (exprimée en nombre de pièces par heure) de production dans l'usine.

Vitesse moyenne = nombre total de pièces produites/nombre d'heures de production. La première machine a produit 500 pièces en (500/100) heures (5 heures) La seconde machine a produit 300 pièces en (300/60) heures (5 heures). La vitesse moyenne est donc donnée par :

$$\text{vitesse moyenne} = \frac{\overset{\boxed{n}}{800}}{\underset{\boxed{\frac{n_1}{x_1}}}{\frac{500}{100}} + \underset{\boxed{\frac{n_2}{x_2}}}{\frac{300}{60}}} = \frac{800}{10} = 80 \text{ pièces/heure}$$

2 • LA MÉDIANE

La **médiane** d'une série est la valeur qui partage cette série, préalablement classée, en deux séries aux effectifs égaux. Dans la première série, on trouve les valeurs inférieures à la médiane. Dans la seconde série on trouve les valeurs supérieures à la médiane.

La médiane ne se calcule que pour les données quantitatives et son mode de calcul dépend du type de données. On distinguera quatre cas :
– les séries non groupées dont l'effectif est impair et où aucune valeur n'est répétée,
– les séries non groupées dont l'effectif est pair et où aucune valeur n'est répétée,
– les séries groupées par valeurs,
– les séries groupées par classes de valeurs.

A – Calcul de la médiane : effectif impair et aucune valeur n'est répétée

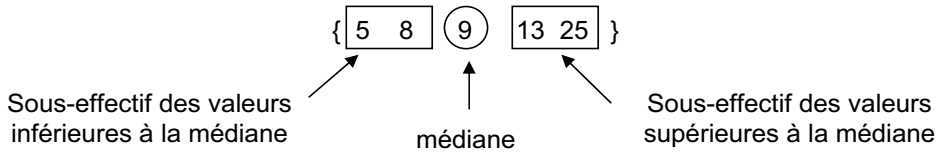
C'est le cas idéal, celui qui permet le mieux de comprendre c'est qu'est la médiane.

Exemple : Soit la série de 5 chiffres suivants : {8 ,5 9, 13 , 25}

Pour trouver la médiane, il faut :

a) Classer la série par ordre croissant des valeurs {5 ,8, 9, 13 , 25}

b) Localiser la valeur qui partage l'effectif total en deux sous effectifs égaux en appliquant la formule (n+1)/2, c'est-à-dire ici (5+1)/2=3. La troisième valeur de la série est le 9.



On vérifie qu'il y autant de valeurs inférieures à la médiane qu'il y a de valeurs supérieures à la médiane. L'effectif total est bien partagé en deux parties égales.

B – Calcul de la médiane : effectif pair et aucune valeur n'est répétée

Quand l'effectif est pair, la médiane n'est pas une valeur de la série. Il faut la calculer.

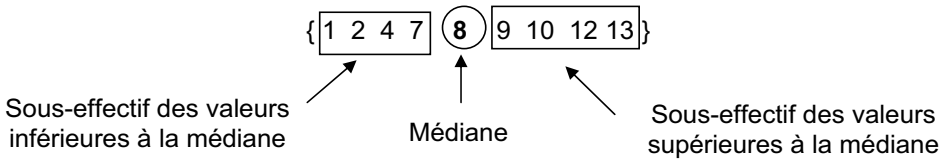
Exemple : Soit la série des 8 chiffres suivants : {13,1,9,10,2,4,12,7}

Pour trouver la médiane, il faut :

a) Classer la série par ordre croissant des valeurs {1,2,4,7,9,10,12,13}

b) Appliquer la formule (n+1)/2, c'est-à-dire ici (8+1)/2=4,5. Ceci nous indique que **l'intervalle médian** est constitué par les 4ème et la 5ème valeurs. La médiane est donc égale à la moyenne arithmétique simple de ces deux valeurs :

Me=(7+9)/2=8



On vérifie qu'il y autant de valeurs inférieures à la médiane qu'il y a de valeurs supérieures à la médiane. L'effectif total est bien partagé en deux parties égales.

C – Calcul de la médiane : effectifs groupés par valeurs

Dans ce cas, la procédure ne permet pas toujours de partager l'effectif total en deux parties égales.

Exemple : Dans le tableau 6 ci-dessous, les valeurs de la variable X ont déjà été classées La troisième colonne est celle des fréquences (f_i) et la quatrième est celle des fréquences cumulées F(x). La cinquième colonne, séparée du tableau, est celle des effectifs cumulés N(x).

Tableau 6 : Calcul de la médiane quand les données sont groupées par valeurs

x_i	n_i	f_i	$F(x)$	$N(x)$
2	2	0,066	0,066	2
8	3	0,1	0,167	5
9	4	0,133	0,3	9
10	4	0,133	0,433	13
11	5	0,167	0,6	18
12	3	0,1	0,7	21
13	6	0,2	0,9	27
15	1	0,033	0,933	28
18	2	0,067	1	30

Médiane =11

0,5

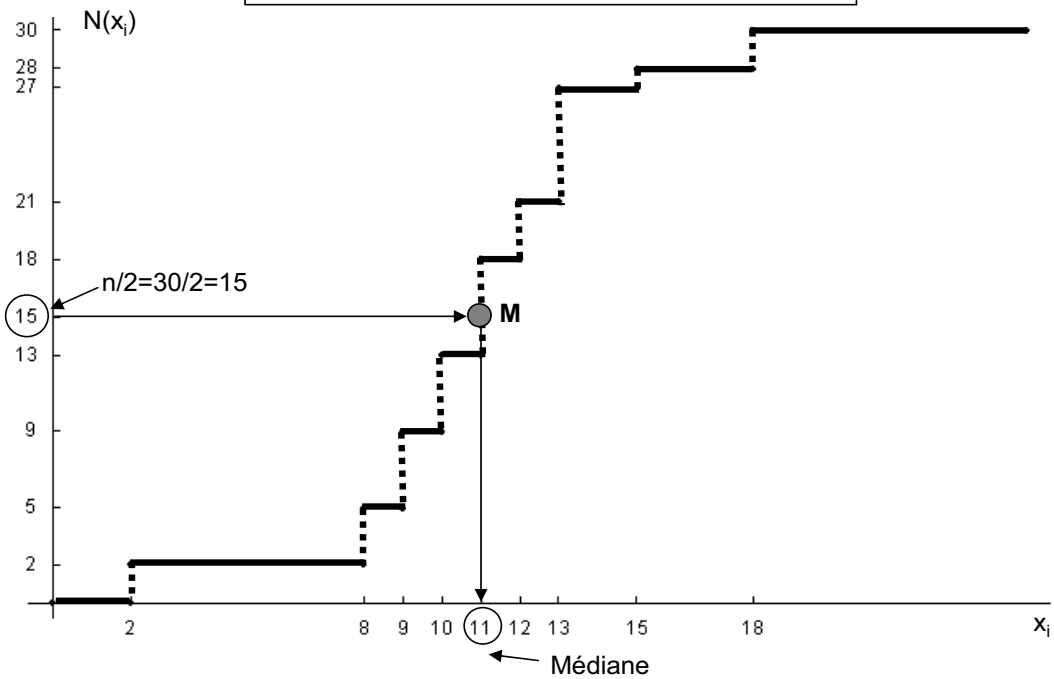
n/2=15

Pour déterminer la médiane, on repère 0,5 dans la colonne des fréquences cumulées F(x) ou bien n/2 dans la colonne des effectifs cumulés N(x). On choisit ensuite la valeur F(x) égale ou immédiatement supérieure à 0,5 (ou la valeur N(x) égale ou immédiatement supérieure à n/2) et l'on suit le sens des flèches comme indiqué sur le tableau 6. Dans notre exemple, il n'y a pas de valeur F(x) égale à 0,5, la valeur immédiatement supérieure à 0,5 est 0,6 (et la valeur immédiatement supérieure à n/2=30/2=15 est 18). Par conséquent, en suivant les flèches, on remonte à la valeur qui correspond à la médiane, soit 11. On remarque alors que la médiane ne sépare pas l'effectif en deux parties égales. En effet, il y a 13 valeurs qui sont inférieures à 11 (soit 43,3% de l'effectif) et 12 valeurs qui sont supérieures à 11 (soit 40% de l'effectif). En outre, que faire des 5 valeurs qui sont exactement égales à 11 (16,6% de l'effectif total). Faut-il les compter dans l'effectif des valeurs inférieures à la médiane ou dans l'effectif des valeurs supérieures à la médiane ? Il n'y a pas de réponse à cette question, chacun fait comme il l'entend [1].

(1) La méthode de calcul de la médiane proposée ici est celle décrite par Bernard PY, dans son ouvrage *Statistiques descriptives*, Éditions Economica, page 76.

Détermination graphique. La figure 1 ci-dessous illustre la détermination de la médiane à partir à partir de $N(x_i)$, la **courbe cumulative des effectifs**. Cette courbe « en escalier » a pour ordonnée les effectifs dont la valeur est strictement inférieure à x_i. Par exemple, l'effectif des valeurs strictement inférieures à 11 est égal à 13. De même, l'effectif des valeurs strictement inférieures à 12 est égal à 18.

Figure 1 : Détermination graphique de la médiane à partir de la courbe cumulative des effectifs



Pour trouver la médiane, il faut localiser n/2=30/2=15 sur l'axe des ordonnées, puis tracer une flèche horizontale jusqu'au point M. Une fois au point M, il faut tracer une flèche verticale en direction de l'abscisse. On lit alors la valeur de la médiane qui, dans notre exemple, est égale à 11.

D – Calcul de la médiane : effectifs groupés par classes de valeurs

Dans ce cas, le calcul de la médiane nécessite d'appliquer la formule suivante :

$$M_e = x_i^{\inf} + a_i \left[\frac{\frac{n}{2} - N\left(x_{i-1}\right)}{n_i} \right] \qquad (10)$$

Où : $x_i^{\inf}$ = Borne inférieure de la classe médiane.

$N(x_{i-1})$ = Effectif cumulé strictement inférieur à x_i

x_i = Classe médiane a_i = Amplitude de la classe médiane

Exemple : Dans le tableau 7 ci-dessous, les valeurs de la variable X du tableau 6 ont été groupées par classes de valeurs d'amplitudes égales (la procédure est la même si les classes sont d'amplitudes inégales).

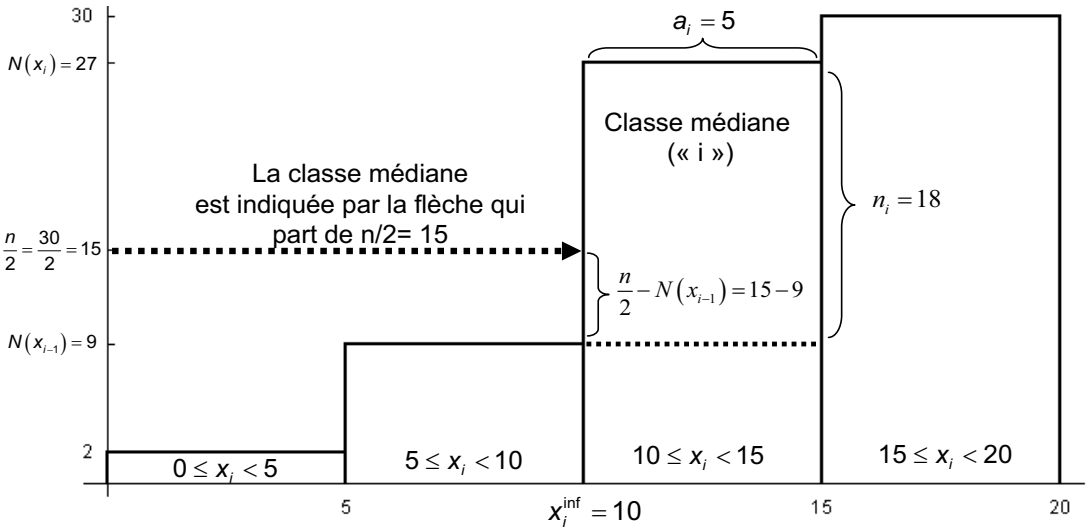
Tableau 7 : Valeurs groupées par classes de valeurs d'amplitude égales

x_i	n_i	$N(x_i)$
[0-5[	2	2
[5-10[	7	9
[10-15[	18	27
[15-20[	3	30

Appliquons la formule (10) en l'interprétant par rapport à la figure 2 qui représente le cumul des n_i en ordonnée [soit $N(x_i)$] et x_i en abscisse :

$$M_e = x_i^{\inf} + a_i \times \left[\frac{\frac{n}{2} - N(x_{i-1})}{n_i} \right] = 10 + 5 \times \left[\frac{15-9}{18} \right] = 11,666$$

Figure 2 : Histogramme des effectifs cumulés



3 • LE MODE

Le **mode** d'une série est la valeur la plus fréquente de cette série. Une série peut avoir plusieurs modes. Le calcul dépend du type de données. Prenons quelques exemples.

A – Calcul du mode : série simple, aucune valeur n'est répétée

Exemple : Soit la série de chiffres {8 ,5 9, 13 , 25}

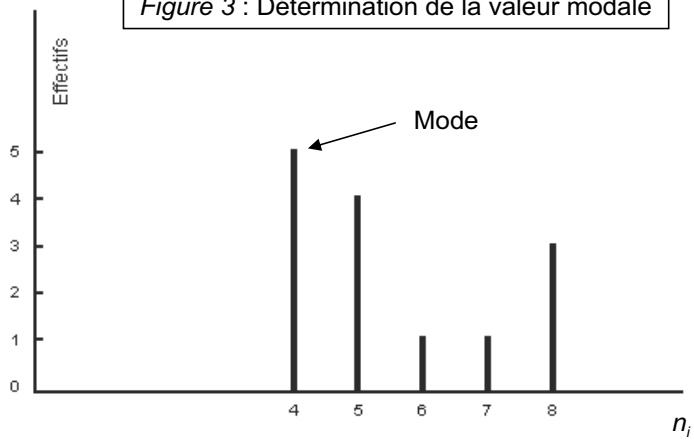
Il n'y a pas de mode car chaque valeur n'est répétée qu'une fois (la fréquence de chaque valeur est égale à 1).

B – Calcul du mode : effectifs groupés par valeurs

Exemple : Soit la série de chiffres {8, 8, 8, 7, 4, 4, 4, 4, 4, 5 ,5, 5, 5, 6}

La valeur la plus fréquente est le 4. Un diagramme en bâtons comme celui de la figure 3 permet de confirmer que le 4 apparaît 5 fois. C'est donc la valeur modale.

Figure 3 : Détermination de la valeur modale



C – Calcul du mode : effectifs groupés par classes d'amplitudes égales

Exemple : Soit le tableau 7 où des données sont présentées par classes d'amplitudes égales.

Dans ce cas, pour calculer le mode, il faut appliquer la formule suivante :

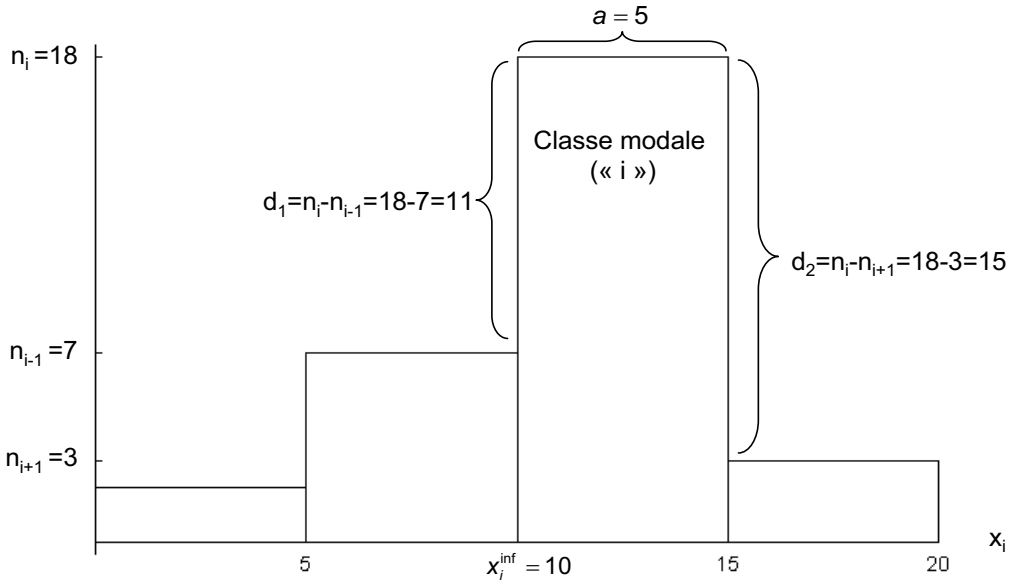
$$\text{Mode} = x_i^{inf} + a\frac{d_1}{d_1 + d_2} \qquad (11)$$

x_i^{inf} = Borne inférieure de la classe modale a = Amplitude de classe

$d_1 = n_i - n_{i-1}$ et $d_2 = n_i - n_{i+1}$

Appliquons la formule (11) en l'interprétant par rapport à la figure 4 qui représente l'histogramme correspondant au tableau 7 (en ordonnée on a les n_i et en abscisse on a les classes de valeurs d'amplitudes égales).

Figure 4 : Calcul du mode quand les classes sont d'égales amplitudes



$$\text{Mode} = x_i^{\inf} + a\frac{d_1}{d_1 + d_2} = 10 + 5 \times \left(\frac{11}{11+15}\right) = 12{,}115$$

D – Calcul du mode : effectifs groupés par classes d'amplitudes inégales

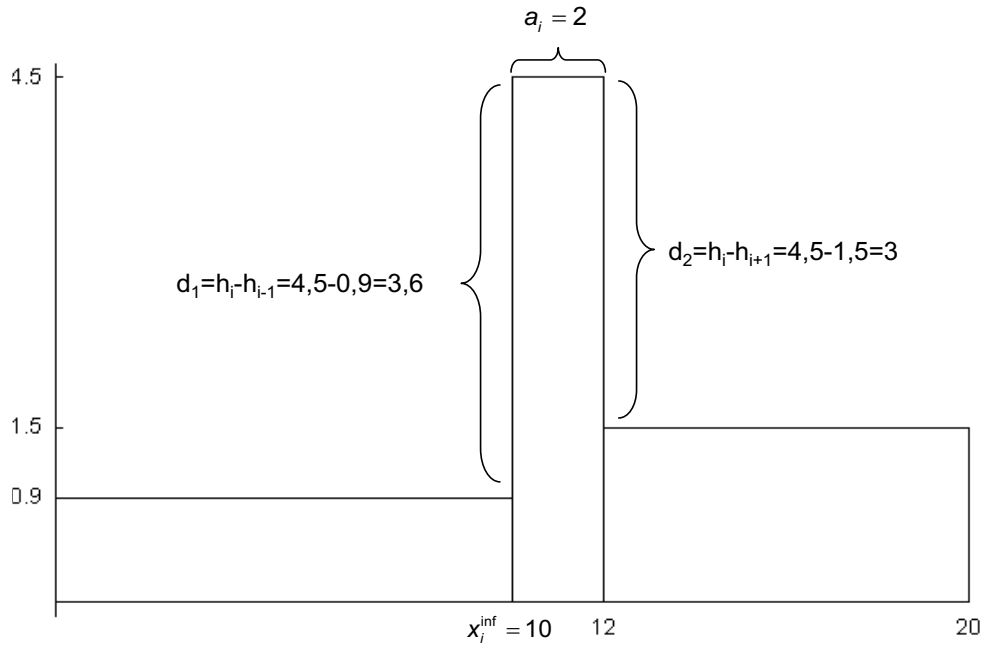
Exemple : Soit le tableau 8 où des données sont présentées par classes d'amplitudes inégales.

Tableau 8 : Valeurs groupées par classes de valeurs d'amplitudes inégales

x_i	n_i	a_i	$h_i = \frac{n_i}{a_i}$
[0-10[	9	10	0,9
[10-12[	9	2	4,5
[12-20[	12	8	1,5

Dans ce cas, pour calculer le mode, il faut appliquer la formule (11), mais la définition de d_1 et de d_2 change, car il faut remplacer les effectifs n_i par les amplitudes corrigées $h_i=n_i/a_i$. On a donc, en suivant par rapport à la figure 5 qui représente l'histogramme correspondant au tableau 8 (en ordonnée on a les n_i/a_i et en abscisse on a les classes de valeurs d'amplitudes inégales).

Figure 5: Calcul du mode quand les classes sont d'inégales amplitudes



$h_{i-1}=n_{i-1}/a_{i-1}=9/10= 0,9$ $h_i=n_i/a_i=9/2=4,5$ $h_{i+1}=n_{i+1}/a_{i+1}=12/8=1,5$

$d_1=h_i-h_{i-1} = 4,5-0,9=3,6$ $d_2=h_i-h_{i+1} =4,5-1,5=3$

et $$\text{Mode} = x_i^{inf} + a_i \frac{d_1}{d_1+d_2} = 10 + 2 \times \left(\frac{3,6}{3,6+3}\right) = 11,09$$

4 • COMMENT CARACTÉRISER LA FORME D'UNE DISTRIBUTION À L'AIDE DE LA MOYENNE ARITHMÉTIQUE, DE LA MÉDIANE ET DU MODE

La comparaison de la moyenne arithmétique, de la médiane et du mode permet de caractériser la forme d'une distribution. 3 cas sont possibles :

– Distribution parfaitement symétrique : Moyenne=Médiane=Mode
– Distribution étalée vers la droite : Moyenne > Médiane> Mode
– Distribution étalée vers la gauche : Moyenne<Médiane< Mode.

Considérons chacun de ces cas en l'illustrant par un exemple.

A - Distribution parfaitement symétrique

Exemple : soit le tableau 9 suivant et le diagramme en barre de la figure 6 qui l'illustre.

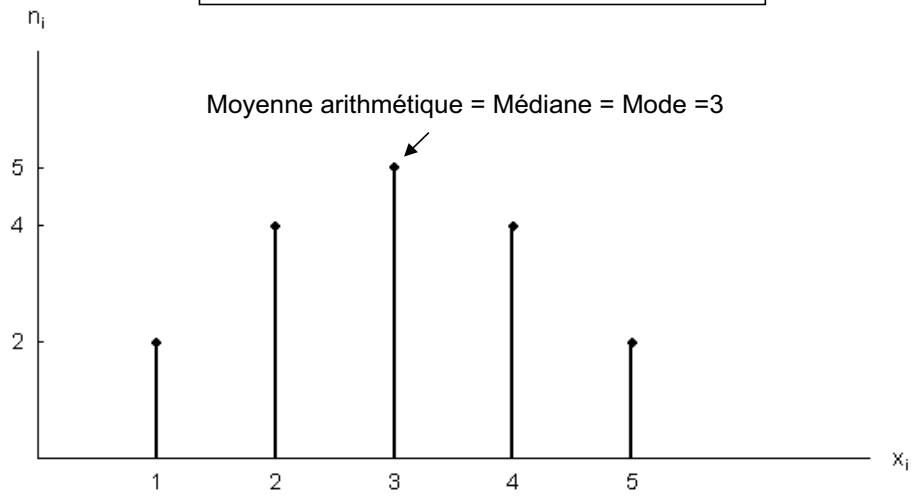
Tableau 9 : Distribution parfaitement symétrique

x_i	1	2	3	4	5
n_i	2	4	5	4	2

Le calcul des 3 indices révèle que $\overline{x} = Me = Mo = 3$

La distribution est parfaitement symétrique, comme l'illustre le diagramme en bâtons de la figure 6 ci-dessous.

Figure 6 : Distribution parfaitement symétrique



B - Distribution étalée à droite

Exemple : soit le tableau 10 suivant et le diagramme en barre de la figure 7 qui l'illustre.

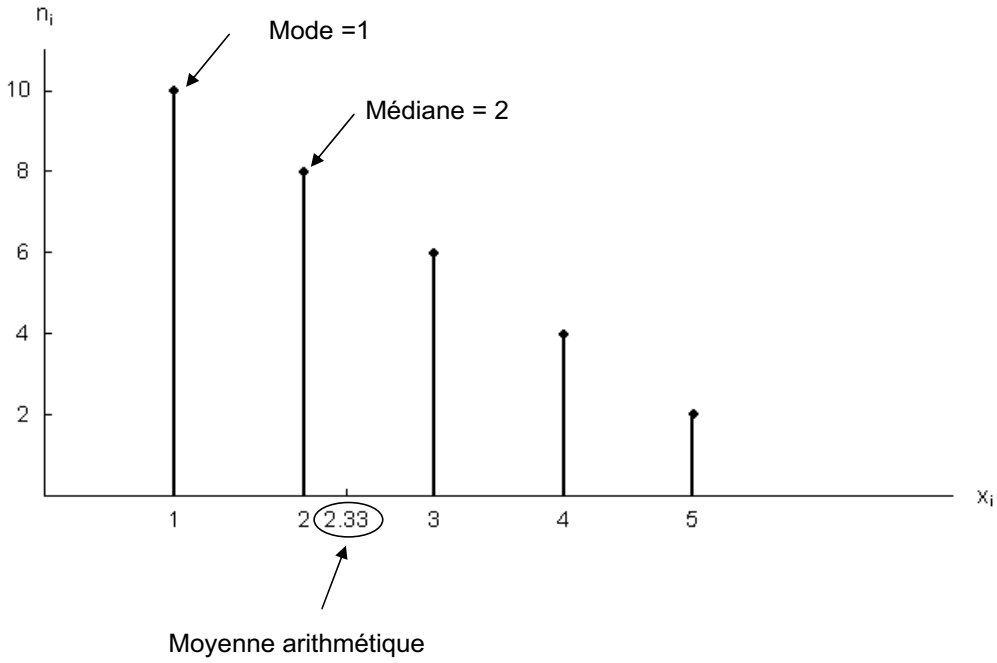
Tableau 10 : Distribution étalée à droite

x_i	1	2	3	4	5
n_i	10	8	6	4	2

Le calcul des 3 indices révèle que $\bar{x} = 2{,}33 > Me = 2 > Mo = 1$

La distribution est étale à droite, comme l'illustre le diagramme en bâtons de la figure 7 ci-dessous.

Figure 7 : Distribution étalée à droite



C - Distribution étalée à gauche

Exemple : soit le tableau 11 suivant et le diagramme en barre de la figure 8 qui l'illustre.

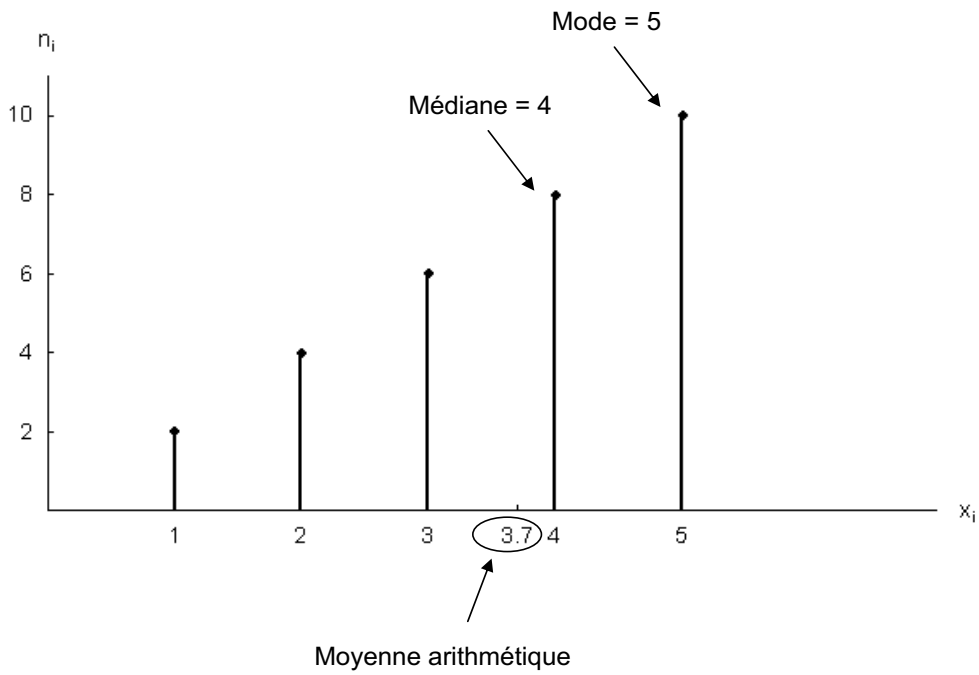
Tableau 11 : Distribution étalée à gauche

x_i	1	2	3	4	5
n_i	2	4	6	8	10

Le calcul des 3 indices révèle que $\bar{x} = 3{,}7 < Me = 4 < Mo = 5$

La distribution est étalée à gauche, comme l'illustre le diagramme en bâtons de la figure 8 ci-dessous.

Figure 8 : Distribution étalée à gauche



CHAPITRE 4

Dispersion et concentration

En complément du chapitre précédent qui étudiait les caractéristiques de tendance centrale d'une distribution, le présent chapitre s'intéresse à la **variabilité** des données au sein d'une série. Ainsi, une fois la moyenne connue, on peut compléter la connaissance d'une série pour apprécier dans quelle mesure les données sont **dispersées** ou au contraire **concentrées** autour de la moyenne.

Sauf dans le cas très rare d'une série statistique où toutes les valeurs sont identiques – par exemple un élève qui a 15 sur 20 dans toutes ses matières – il existe toujours une certaine variabilité des données dans une série. Ainsi, le prix au mètre carré varie plus ou moins d'une maison à l'autre, le prix d'un produit varie aussi d'un magasin à l'autre. Les salaires varient d'une entreprise à l'autre, de même que, en général, les notes d'un élève dans les différentes matières de son cursus.

Les caractéristiques de dispersion et/ou de concentration sont nombreuses. Nous étudierons ici les plus fréquemment utilisées : l'intervalle de variation, la variance, l'écart-type, le coefficient de variation, les intervalles interquartiles et interdéciles et l'écart médiale-médiane. Nous verrons également deux outils graphiques utiles pour l'analyse de la dispersion/concentration d'une distribution : le graphique « boîte à moustaches », ainsi que la courbe de concentration.

1 • L'INTERVALLE DE VARIATION

L'**intervalle**, ou « spread » c'est la différence entre la plus grande valeur et la plus petite valeur de la variable.

Exemple : soit deux élèves dont les notes dans quatre matières ont été les suivantes :

Élève A : { 8, 9, 10, 11, 12} Élève B : {2, 4, 16, 18}

L'étendue des notes de A est 12-8 = 4, tandis que l'étendue des notes de B est 18 – 2 = 16. On notera pourtant que la moyenne des deux élèves est de 10. Mais B a des notes beaucoup plus dispersées que A. En fait, si on fait le rapport 16/4, on voit que les notes de B sont 4 fois plus dispersées que celles de A.

Cet exemple montre l'utilité de l'intervalle de variation pour avoir une première idée de la dispersion. Mais l'indicateur est assez limité, car il est trop sensible aux valeurs extrêmes comme le montre l'exemple ci-après.

Exemple : soit la série suivante {1016, 774, 1008, 8, 1001, 999, 1100}

Il est commode de classer les chiffres par ordre croissant :

{8, 774, 999, 1001, 1008, 1016, 1100}

L'intervalle de variation est donc donné par IV = 1100 – 8 = 1092. On constate que la valeur de l'intervalle de variation est exagérément augmentée par la présence du chiffre 8.

2 • L'INTERVALLE INTERQUARTILE

L'intervalle interquartile est une mesure de la variation qui n'est pas influencée par les valeurs extrêmes, contrairement à l'intervalle de variation.

Sa définition est simple : **l'intervalle interquartile** mesure l'étendue des 50% de valeurs situées au milieu d'une série de données classées.

Il se calcule en procédant aux quatre étapes suivantes:

1) Classement des données de la série par ordre croissant.

2) Trouver la médiane de la série pour séparer celle-ci en deux séries : la pemière série contient les données inférieures à la médiane et la seconde les données supérieures à la médiane.

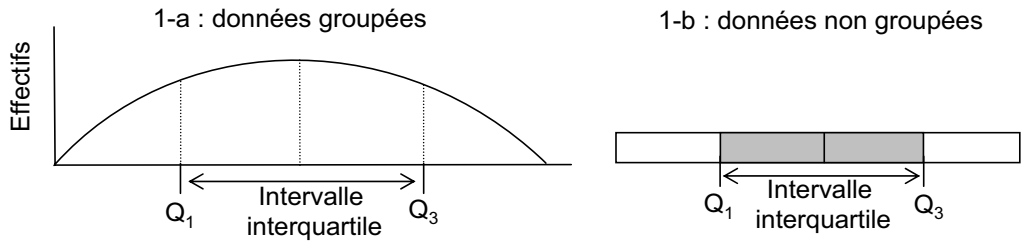
3) Déterminer la médiane des deux nouvelles séries, sans inclure dans aucune d'elle la médiane de la série initiale. La médiane de la première série est appelée « premier quartile » et désigné par Q_1. La médiane de la seconde série est appelée « second quartile » et désigné par Q_3.

4) Calculer IQ, l'intervalle interquartile par la formule :

$$IQ = Q_3 - Q_1$$

Les figures 1-a et 1-b, ainsi que les quatre exemples ci-après illustrent les notions de quartiles et d'intervalle interquartile dans le cas de données groupées (1-a) ou non groupées (1-b)

Figure 1 : La notion d'intervalle interquartile



Exemple 1 : soit la série de chiffres suivants, où aucune valeur n'est répétée. Le nombre de chiffres est impair.

{4, 13, 17, 7, 1, 3, 9, 14, 12, 20, 16, 15, 11, 6,5}

1) Afin de déterminer l'intervalle interquartile, classons d'abord les données de la plus petite à la plus grande.

{1, 3, 4, 5, 6, 7, 9, 11, 12, 14, 15, 16, 17, 19, 20}

2) Déterminons la médiane et séparons la série en deux « sous-séries » :

{1, 3, 4, 5, 6, 7, 9 , 11 ,12, 14, 15, 16, 17, 19, 20}

1ère série — Me — 2ème série

3) Déterminons ensuite la médiane de chacune de ces deux nouvelles séries

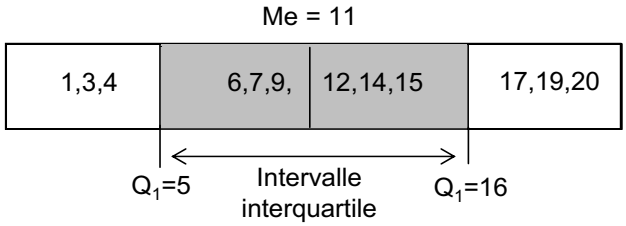
{1, 3, 4, 5 , 6, 7, 9} — Q_1

{12, 14, 15, 16 ,17, 19, 20} — Q_3

4) Il reste plus qu'à calculer l'intervalle interquartile :

$$IQ = Q_3 - Q_1 = 16 - 5 = 11$$

Figure 2 : L'intervalle interquartile
données non groupées, effectif impair



Remarque : Dans ce cas, particulier, la médiane est égale à 11 et l'intervalle interquartile a aussi pour valeur le chiffre 11. Mais c'est un hasard.

Exemple 2 : soit la série de chiffres suivants, où aucune valeur n'est répétée. Cette fois, le nombre de chiffres est pair.

{4, 13, 17, 7, 1, 3, 9, 14, 12, 20, 16, 15, 11, 6}

1) Afin de déterminer l'intervalle interquartile, classons d'abord les données de la plus petite à la plus grande.

{1,3,4, 6,7,9,11,12,14,15,16,17,19,20}

2) Déterminons l'intervalle médian, puis la médiane et séparons la série en deux séries. Ici, (n+1)/2=(14+1)/2=7,5. L'intervalle médian est donc constitué par la 7ème et la 8ème valeur, c'est-à-dire [11-12]. Et la médiane (11+12)/2=11,5.

{ 1,3,4,6,7,9,11 , 12,14,15,16,17,19,20 }

11,5 ↑ Me

1ère série — 2ème série

3) Déterminons ensuite la médiane de chacune de ces deux nouvelles séries

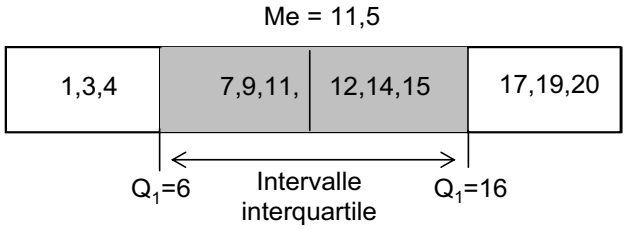
{1,3,4, 6 ,7,9,11} ↑ Q_1

{12,14,15, 16 ,17,19,20} ↑ Q_3

4) Il reste plus qu'à calculer l'intervalle interquartile :

$$IQ = Q_3 - Q_1 = 16 - 6 = 10$$

Figure 3 : L'intervalle interquartile
données non groupées, effectif pair



Exemple 3 : Soit la série de chiffres suivants :

{4,13, 6, 4,13, 17,7,15,7,16,9, 6,7,1,3,9,14,1,1,12, 11, 20,16,15,11,6, 11}

1) Afin de déterminer l'intervalle interquartile, classons d'abord les données de la plus petite à la plus grande et, comme certaines données sont répétées, construisons un tableau, en ajoutant une ligne pour les effectifs cumulés. ($\sum n_i \uparrow$ désigne le cumul croissant des valeurs).

Tableau 1 : Série groupée par valeurs

x_i	1	3	4	6	7	9	11	12	13	14	15	16	17	20
n_i	3	1	2	3	3	2	3	1	2	1	2	2	1	1
$\sum n_i \uparrow$	3	4	6	9	12	14	17	18	20	21	23	25	26	27

n/2=27/2=13,5

2) Déterminons la médiane de la série par la méthode étudiée dans le chapitre 3 dans le cas des données groupées par valeurs. On voit que puisque n=27, on a n/2=27/2=13,5, ce qui tombe entre 12 et 14. Par convention, nous choisissons la valeur de la variable qui correspond à 14, soit 9.

3) La médiane est donc égale à 9. Et nous avons deux séries, dont nous pouvons maintenant déterminer les médianes respectives, suivant la même méthode.

Tableau 2 : Calcul des quartiles

Q_1 Q_1

x_i	1	3	4	6	7	9	11	12	13	14	15	16	17	20
n_i	3	1	2	3	3	2	3	1	2	1	2	2	1	1
$\sum n_i \uparrow$	3	4	6	9	12	14	5	6	8	9	11	13	14	15

n/2 = 9/2 = 4,5 n/2 = 15/2 = 7,5

4) L'intervalle interquartile est donc :

$$IQ = Q_3 - Q_1 = 13 - 4 = 9$$

Remarques :

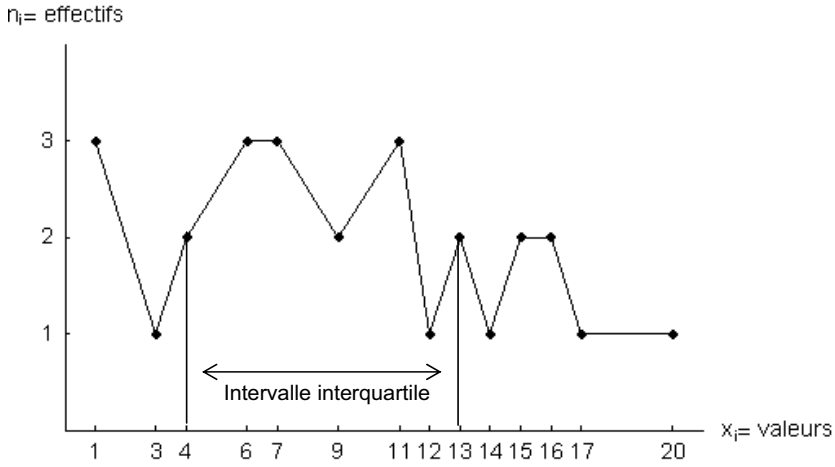
1) Normalement, 50% des effectifs devraient être concentrés dans l'intervalle interquartile. Ce n'est pas tout à fait le cas ici, en raison des approximations de la méthode. Il y a en effet 16 unités statistiques sur 27 qui sont dans cet intervalle, soit 16/27 = 0,59.

2) On peut rapporter l'intervalle interquartile à l'intervalle de variation :

$$\frac{\text{Intervalle interquartile}}{\text{Intervalle de variation}} \times 100 = \left(\frac{Q_3 - Q_1}{20-1}\right) \times 100 = \left(\frac{13-4}{19}\right) \times 100 = \left(\frac{9}{19}\right) \times 100 = 47{,}3\%$$

3) Enfin, on peut représenter les résultats sur un graphique :

Figure 4 : L'intervalle interquartile, données groupées



Exemple 4 : Soit le tableau suivant, où les valeurs de l'exemple précédent ont été regroupées par classes.

Tableau 3

x_i	[0-4[	[4-8[	[8-12[	[12-16[	[16-20]
n_i	4	8	5	6	4

1) Afin de déterminer l'intervalle interquartile, ajoutons une ligne pour les effectifs cumulés.

Tableau 4

x_i	[0-4[	[4-8[	[8-12[	[12-16[	[16-20]
n_i	4	8	5	6	4
$\sum n_i \uparrow$	4	12	17	23	27

n/2 = 27/2 = 13,5

2) Déterminons la médiane de la série par la méthode étudiée dans le chapitre 3 dans le cas des données groupées par classe. Il faut d'abord déterminer la classe médiane, qui est ici [8-12[. Il n'est pas nécessaire de connaître la valeur exacte de la médiane pour séparer les deux séries, mais calculons-là quand même en appliquant la formule étudiée au chapitre 3 pour le calcul de la médiane quand les données sont groupées par classe :

$$M_e = x_i^{\inf} + a_i \times \left[\frac{\frac{n}{2} - N\left(x_{i-1}\right)}{n_i} \right] = 8 + 4 \times \left[\frac{13,5 - 12}{5} \right] = 9,2$$

3) La classe médiane [8-12[permet de diviser le tableau en deux. Calculons les médianes respectives de chacun de ces tableaux :

Tableau 5

xi	[0-4[	[4-8[
ni	4	8
$\sum n_i \uparrow$	4	12

n/2 = 12/2 =6

xi	[12-16[	[16-20[
ni	6	4
$\sum n_i \uparrow$	6	10

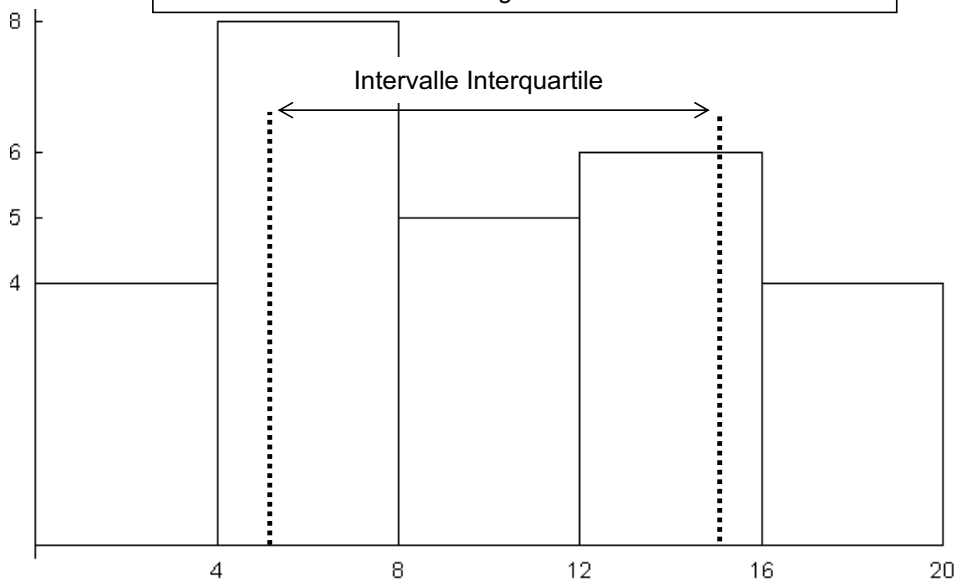
n/2 = 10/2 = 5

$$Q_1 = 4 + 4 \times \left[\frac{6-4}{8} \right] = 5$$

$$Q_3 = 12 + 4 \times \left[\frac{5-0}{6} \right] = 15,3$$

L'histogramme ci-dessous, permet d'illustrer l'intervalle interquartile dans le cas où les données sont groupées par classes.

Figure 5



3 • LA BOÎTE À MOUSTACHE

A – Définition

La **boîte à moustache**, de l'anglais « Box and Whiskers », parfois aussi désignée « box plot », est un graphique qui résume la dispersion d'une série à partir de 5 valeurs : la valeur minimale et la valeur maximale (ce sont les « moustaches »), l'intervalle interquartile (désigné par ses deux valeurs Q_1 et Q_3) et la médiane (ces trois dernières valeurs constituant la « boîte »).

Exemple : soit la série de chiffres suivante, où aucune valeur n'est répétée. Le nombre de chiffres est impair.

{4, 13, 17, 7, 1, 3, 9, 14, 12, 20, 16, 15, 11, 6, 5}

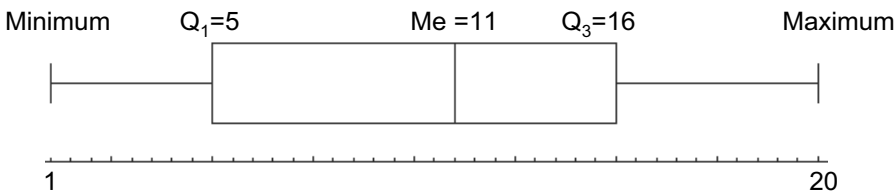
Nous savons que Me = 11, Q_1 = 5 et Q_3 = 16 pour les avoir calculés à l'exemple 1 de la section 2 de ce chapitre. Quant aux valeurs minimale et maximale, elles sont respectivement égales à 4 et 20. Classons la série par ordre croissant pour mieux faire apparaître les différentes valeurs impliquées dans la boîte à moustache.

{1, 3, 4, 5, 6, 7, 9, 11, 12, 14, 15, 16, 17, 19, 20}

Minimum — Q_1 — Me — Q_3 — Maximum

Le graphique dit de la « boîte à moustache » correspondant est donc :

Figure 6



B – Utilité de la boîte à moustache pour comparer des séries

La boîte à moustache permet de comparer des séries du point de vue de leur dispersion mais aussi de leur caractéristique de tendance centrale (puisque la médiane est repérée).

Exemple : soient les notes sur 20 de 4 groupes d'étudiants :

Groupe A {1, 2, 2, 12, 5, 5, 9, 5, 7, 11, 7, 8, 2}

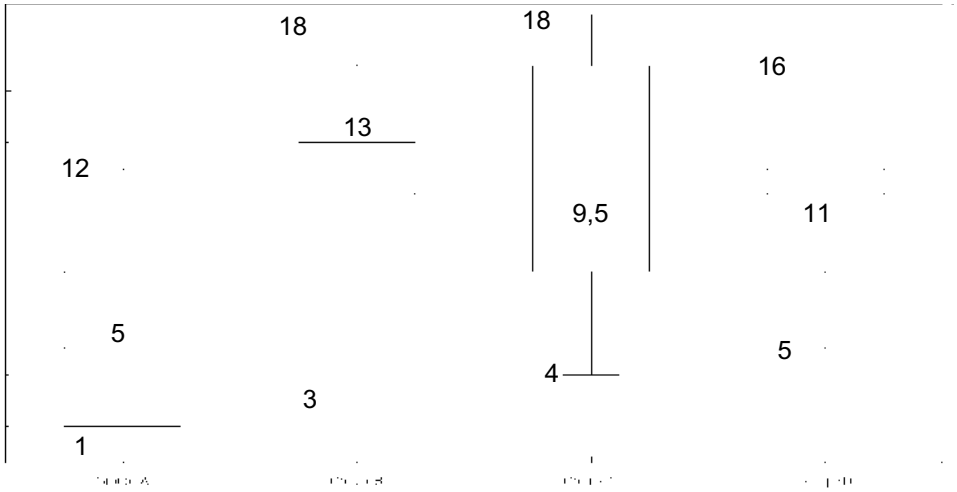
Groupe B {16, 13, 15, 13, 11, 13, 16, 3, 18, 11}

Groupe C {8, 8, 8, 7, 4, 16, 13, 16, 18, 11}

Groupe D {12, 10, 6, 8, 5, 16, 12, 15, 10, 15, 12, 10}

La comparaison des graphiques boîtes à moustaches de chaque groupe permet d'avoir une bonne idée de la dispersion des notes, tout en visualisant la note médiane (qui est souvent jugée préférable à la note moyenne).

Figure 7



C – Utilité de la boîte à moustache pour déterminer la forme d'une distribution

Suivant la position de la médiane au sein de la boîte, on peut en déduire des informations sur la forme de la distribution.

1) Si la médiane est proche du centre de la boîte, c'est que la distribution est symétrique.

2) Si la médiane est à gauche du centre de la boîte, c'est que la distribution est étalée à droite.

3) Si la médiane est à droite du centre de la boîte, c'est que la distribution est étalée à gauche.

De même, en comparant la longueur respective de chaque moustache, on peut en déduire des informations sur la forme de la distribution.

1) Si les moustaches sont à peu près de la même longueur, c'est que la distribution est symétrique.

2) Si la moustache de droite est plus longue que la moustache de gauche, c'est que la distribution est étalée à droite.

3) Si la moustache de gauche est plus longue que la moustache de droite, c'est que la distribution est étalée à gauche.

Exemple : Soit les trois séries utilisées dans la section 4 du chapitre 3, dont les distributions (voir les diagrammes en bâtons) sont respectivement symétrique (Me=3), étalée à droite (Me = 2) et étalée à gauche (Me = 4) :

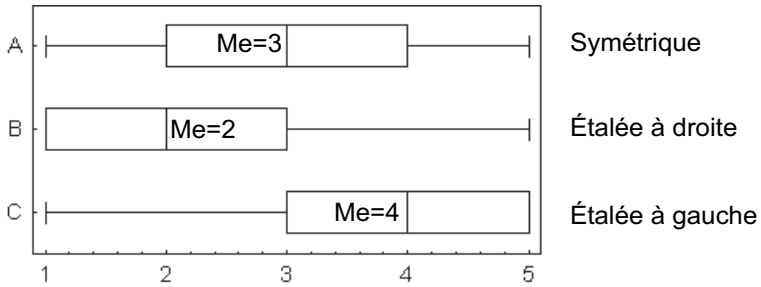
A = {1, 1, 2, 2, 2, 2, 3, 3, 3, 3, 3, 4, 4, 4, 4, 5, 5}

B = {1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 5, 5}

C = {1, 1, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5}

Les boîtes à moustaches correspondantes ont bien les caractéristiques précitées :

Figure 8



4 • VARIANCE, ÉCART-TYPE ET COEFFICIENT DE VARIATION

La variance, l'écart-type et le coefficient de variation sont les indicateurs les plus fréquemment utilisés pour mesurer la dispersion d'une série. Ces indicateurs renseignent sur la **dispersion des données autour de la moyenne**.

Plus les données sont concentrées autour de la moyenne, plus les valeurs de ces trois indicateurs sont faibles. Inversement, plus les données sont dispersées autour de la moyenne, plus ces trois indicateurs sont élevés.

A – La variance

1) Définition

Soit une série de valeurs d'une variable X : $\{x_1, x_2, \ldots, x_k\}$. Soit les effectifs associés : $\{n_1, n_2, \ldots, n_k\}$. La variance de cette série s'écrit :

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{k} n_i \left(x_i - \overline{x}\right)^2 \quad \text{, si l'effectif considéré est celui d'une population.} \qquad (1)$$

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{k} n_i \left(x_i - \overline{x}\right)^2 \quad \text{, si l'effectif considéré est celui d'un échantillon.} \qquad (2)$$

Ainsi que nous l'avons déjà indiqué dans le chapitre 1, sauf mention contraire explicite, nous ne considérons dans cet ouvrage que des populations. Par conséquent, la formule (1) sera utilisée dans la suite.

Remarque : Si $\{n_1, n_2, \ldots, n_k\} = \{1, 1, \ldots, 1\}$ et que k = n, la variance de la série s'écrira :

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

Autrement dit, lorsque les données sont connues individuellement ou qu'elles ne se répètent pas, c'est la formule (1-a) qui s'applique. En revanche, lorsque les données sont groupées par valeurs, c'est la formule (1) qui s'applique. Enfin, lorsque les données sont groupées par classe, c'est le centre de classe c_i, qui remplace x_i dans la formule (1).

2) Mode de calcul de la formule (1-a)

Pour calculer la variance à partir de la formule (1-a), on applique successivement les étapes suivantes :

a) Calcul de la moyenne
b) Calcul des écarts à la moyenne
c) Calcul des carrés des écarts à la moyenne
d) Somme des carrés des écarts à la moyenne
e) Division par n

L'exemple ci-après illustre cette méthode.

Exemple : soit la série {2, 5, 7, 1, 9, 13, 6, 15, 8, 16}

Les étapes a), b), c) et d) sont facilitées par la disposition en tableau :

Tableau 6

(a)

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{10} x_i = 8{,}2$$

x_i	**(b)** $(x_i - \bar{x})$	**(c)** $(x_i - \bar{x})^2$
2	-6,2	38,44
5	-3,2	10,24
7	-1,2	1,44
1	-7,2	51,84
9	0,8	0,64
13	4,8	23,04
6	-2,2	4,84
15	6,8	46,24
8	-0,2	0,04
16	7,8	60,84

(d)

$$\frac{1}{n}\sum_{i=1}^{10}(x_i - \bar{x})^2 = \frac{237{,}6}{10} = 23{,}76$$

3) Mode de calcul de la formule « développée »

La formule (1) peut aussi être calculé suivant la méthode précédente. Toutefois, pour faciliter les calculs, il est préférable d'utiliser la formule dite « développée ». On montre en effet que la formule (1) peut s'écrire :

(1-b)

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

Pour calculer la variance à l'aide de la formule « développée », on suit les étapes :

a) Calcul de la moyenne pondérée et élévation de celle-ci au carré
b) Calcul des x_i^2
c) Calcul des $n_i x_i^2$
d) Somme des $n_i x_i^2$
e) Division des $n_i x_i^2$ par n
f) Soustraction du carré de la moyenne au carré de la moyenne des $n_i x_i^2$

Exemple : soit le tableau suivant

Tableau 7

x_i	2	6	9	11	15
n_i	5	9	4	3	5

Les étapes a), b), c), d) et e) sont facilitées par la disposition en tableau :

Tableau 8

x_i	n_i	$n_i x_i$	x_i^2	$n_i x_i^2$
2	5	10	4	20
6	9	54	36	324
9	4	36	81	324
11	3	33	121	363
15	5	75	225	1125
	26	**208**		**2156**

Totaux

$$\overline{x} = \frac{1}{26}\sum_{i=1}^{5} n_i x_i = \frac{208}{26} = 8$$

$$\sigma^2 = \frac{1}{26}\sum_{i=1}^{5} n_i x_i^2 - \overline{x}^2$$

$$\sigma^2 = \frac{1}{26}2156 - (8)^2$$

$$\sigma^2 = 82{,}9231 - 64 = 18{,}9231$$

B – L'écart-type et le coefficient de variation

1) L'écart-type

L'écart-type est égal à la racine carrée de la variance :

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{k} n_i \left(x_i - \overline{x}\right)^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{k} n_i x_i^2 - \overline{x}^2} \qquad (3)$$

Naturellement, si aucune valeur n'est répétée ou si les données ne sont pas regroupées par valeur, on aura :

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2} \qquad (3\text{-a})$$

Exemple 1 : Soit la série {2, 5, 7, 1, 9, 13, 6, 15, 8, 16}

La variance de cette série a été calculée à la section 4-2. Elle est égale à :

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{10} \left(x_i - \overline{x}\right)^2 = \frac{237{,}6}{10} = 23{,}76$$

L'écart-type est :

$$\sigma = \sqrt{23{,}76} \cong 4{,}87$$

Exemple 2 : Soit les données du tableau 7

La variance a été calculée et est égale à :

$$\sigma^2 = 18{,}9231$$

On en déduit l'écart-type :

$$\sigma = \sqrt{18{,}9231} \cong 4{,}35$$

2) Le coefficient de variation

$$CV = \left(\frac{\sigma}{\bar{x}}\right) \times 100$$

Exemple : On connaît les salaires mensuels bruts en euros des 200 employés de la même entreprise, à 10 ans d'intervalle (voir le tableau 9). Les données sont groupées par classe. Le nombre d'employés est passé de 200 en 1994 à 280 en 2004. On veut savoir si la dispersion des salaires à augmenté. Pour cela on va calculer le coefficient de variation en 1994 et en 2004.

Tableau 9

Salaires	Effectifs 1994	Effectifs 2004
1000-2000	40	56
2000-3000	70	118
3000-4000	80	92
4000-5000	5	10
5000-10000	5	4

On notera tout d'abord que les données sont groupées par classes de valeurs. Dès lors, il convient de calculer c_i, le centre de chaque classe, qui tiendra lieu de x_i dans les différentes formules. Les tableaux 10 et 11 ci-après indiquent les calculs intermédiaires nécessaires pour obtenir le coefficient de variation des salaires, respectivement en 1994 et en 2004.

Tableau 10

Salaires	1994 (n_i)	c_i	n_ic_i	c_i^2	$n_ic_i^2$
1000-2000	40	1500	6000	2250000	90000000
2000-3000	70	2500	175000	6250000	43750000
3000-4000	80	3500	280000	12250000	980000000
4000-5000	5	4500	22500	20250000	101250000
5000-10000	5	7500	37500	56250000	281250000
	200		**575000**		**1890000000**

Totaux

Calculons la moyenne, la variance et l'écart-type à partir des calculs intermédiaires du tableau 10 :

$$\bar{x} = \frac{1}{200}\sum_{i=1}^{5} n_i c_i = \frac{575000}{200} = 2875$$

$$\sigma = \sqrt{\frac{1890000000}{200} - (2875)^2} = 1088,29$$

Et le coefficient de variation des salaires pour l'année 1994 est donc égal à :

$$CV_{1994} = \left(\frac{\sigma}{\bar{x}}\right) \times 100 = \frac{1088,29}{2875} \times 100 = 37,8536$$

Refaisons les calculs pour l'année 2004 :

Tableau 11

Salaires	2004 (n_i)	c_i	$n_i c_i$	c_i^2	$n_i c_i^2$
1000-2000	56	1500	84000	2250000	126000000
2000-3000	118	2500	295000	6250000	737500000
3000-4000	92	3500	322000	12250000	1127000000
4000-5000	10	4500	45000	20250000	202500000
5000-10000	4	7500	30000	56250000	225000000
	280		**776000**		**2418000000**

Totaux

$$\bar{x} = \frac{1}{280}\sum_{i=1}^{5} n_i c_i = \frac{776000}{280} = 2771,43$$

$$\sigma^2 = \frac{1}{280}\sum_{i=1}^{5} n_i c_i^2 - \bar{x}^2$$

$$\sigma^2 = \frac{2418000000}{280} - (2771,43)^2$$

$$\sigma^2 = 954898$$

$$\sigma = \sqrt{954898} = 977,189$$

$$CV_{2004} = \left(\frac{\sigma}{\bar{x}}\right) \times 100 = \frac{977,189}{2771,43} \times 100 = 35,2594$$

En comparant les deux coefficients de variation, on constate que la dispersion des salaires s'est réduite.

5 • LES INDICATEURS DE CONCENTRATION

C'est pour l'étude de la répartition des salaires, des revenus ou des patrimoines que les premiers indicateurs de concentration ont été élaborés. C'est en fait une autre façon de mesurer la dispersion puisque, par définition, plus une série est concentrée, moins elle est dispersée et réciproquement.

Cependant, contrairement à la dispersion, la concentration n'a de sens que pour des données positives et a des variables ou des caractères dont l'addition a un sens : ainsi pourra-t-on additionner des patrimoines, des surfaces, des chiffres d'affaires, etc. La notion de concentration appliquée à des variables telles que l'âge, la taille ou le poids d'une population, quoique envisageable en théorie, n'a pas nécessairement de signification.

Il existe deux méthodes pour mesurer la concentration : par le calcul et par les graphiques. Avant de les étudier, il faut d'abord introduire la notion de médiale.

A – La médiale

C'est un indicateur qui s'apparente à la médiane, mais appliquée à une série différente. En effet, alors que la médiane s'applique aux valeurs de la variable (les « x_i »), la médiale s'applique aux valeurs de la variables multipliées par leurs effectifs respectifs (les « $n_i.x_i$ »). C'est la valeur du caractère qui partage l'effectif cumulé des $n_i.x_i$ en deux parties égales. Elle sert à déterminer la concentration de la distribution par comparaison avec la médiane et avec l'intervalle de variation.

On a donc la formule suivante :

$$M_l = x_i^{\inf} + a_i \left[\frac{\frac{\sum_{i=1}^{k} n_i x_i}{2} - N(n_i x_i)}{n_i x_i} \right]$$

Où : $x_i^{\inf} =$ Borne inférieure de la classe médiale.

$N(x_{i-1}) =$ Effectif cumulé strictement inférieur à $n_i\, x_i$

$x_i =$ Classe médiale $a_i =$ Amplitude de la classe médiale

Exemple : Soit le tableau suivant

Tableau 12

Classes	[0-1[	[1-5[	[5-10[	[10-20[	[20-50]
Effectifs	6	39	30	27	24

Afin de calculer la médiale, il faut d'abord faire un tableau avec les fréquences cumulées et les **masses cumulées** (c'est-à-dire les $n_i c_i \uparrow$) . Comme les données sont regroupées par classe, c tient lieu de $x_{i.;}$

Tableau 13

Classes	Centres de classe	Effectifs (ni)	$n_i c_i$	$n_i c_i \uparrow$
[0-1[	0,5	6	3	3
[1-5[	3	39	117	120
[5-10[	7,5	30	225	345
[10-20[	15	27	405	750
[20-50]	35	24	840	1590

1590/2=795

$n = \sum n_i$ → **126** **1590** ← $\sum n_i c_i$

Moitié de la somme des $n_i x_i$ = 1590/2 = 795 ⟶ Classe médiane [20-50]

$$M_l = x_i^{\inf} + a_i \left[\frac{\frac{\sum_{i=1}^{k} n_i x_i}{2} - N(n_i x_i)}{n_i x_i} \right] = 20 + 30 \left[\frac{795 - 750}{840} \right] = 21{,}61$$

B – La détermination de la concentration par la méthode graphique

Il s'agit de construire une figure appelée « Courbe de concentration » ou encore « courbe de LORENZ », du nom de son inventeur, l'américain Max O. LORENZ (1880-1962) qui cherchait un moyen commode de comparer les inégalités de revenu entre diverses populations. Elle peut aussi servir à mesurer d'autres formes d'inégalité que celles des revenus.

La courbe de LORENZ se trace dans un carré de côté 1. En abscisse, figurent les fréquences relatives cumulées de la variable et en ordonnée figurent les $n_i.x_i$ cumulés rapportés à la la somme des $n_i.x_i$. Afin de fixer les idées, la courbe de LORENZ de la figure 9 est tracée avec les données du tableau 12 (et au moyen des calculs dérivés qui figurent dans le tableau 14 ci-après).

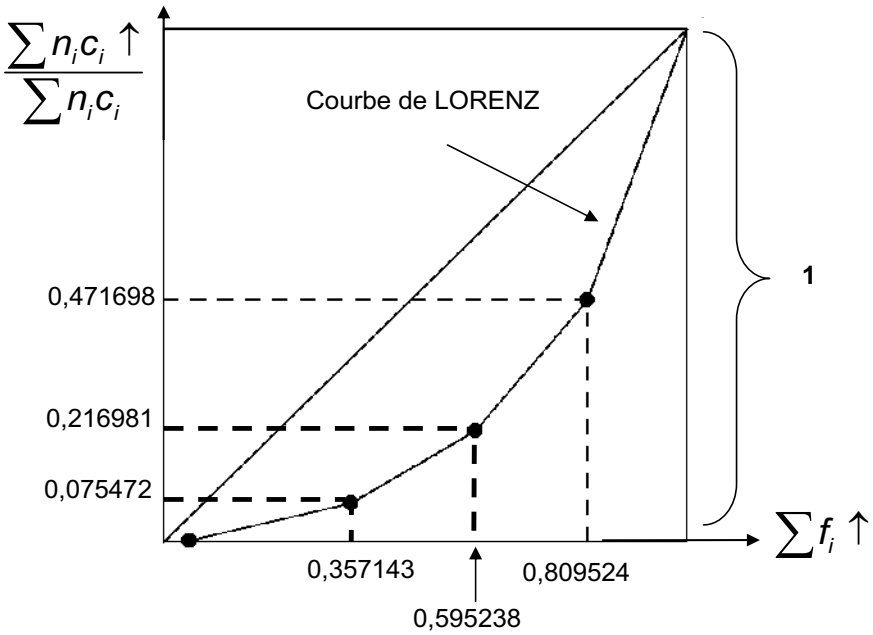
Tableau 14

Classes	Effectifs (n_i)	f_i	f_i cumulés	n_ic_i	n_ic_i cumulés	n_ic_i cumulés relatifs (division par 1590)
[0-1[	6	0,047619	0,047619	3	3	0,001887
[1-5[	39	0,309524	0,357143	117	120	0,075472
[5-10[	30	0,238095	0,595238	225	345	0,216981
[10-20[	27	0,214286	0,809524	405	750	0,471698
[20-50]	24	0,190476	1	840	1590	1

Abscisse de la courbe de LORENZ

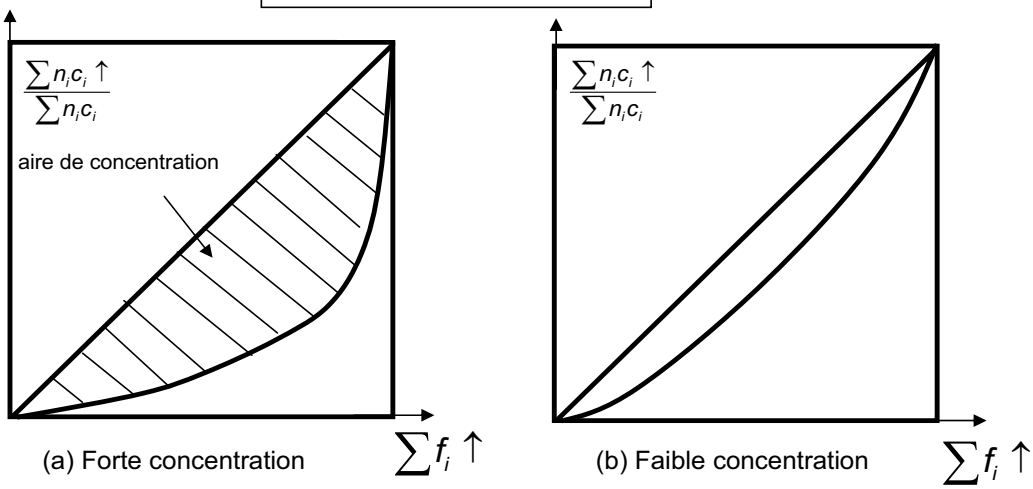
Ordonnée de la courbe de LORENZ

Figure 8 : Courbe de LORENZ



Plus la courbe de LORENZ est éloignée de la première diagonale, plus la distribution est concentrée. Plus au contraire elle se rapproche de la diagonale et moins la distribution est concentrée. Si la courbe de LORENZ se confond avec la diagonale, la répartition est dite « égalitaire ». La figure 9 ci-après illustre deux situations diamétralement opposées : forte concentration (a) et faible concentration (b). La courbe de LORENZ est donc un moyen géométrique et visuel d'observer la concentration d'une série. Elle permet d'effectuer des comparaisons de séries à un même moment (les salaires dans deux ou plusieurs entreprises) ou d'une série à plusieurs moments différents (l'évolution de la répartition des salaires dans une entreprise).

Figure 9 : Courbes de LORENZ



La surface hachurée est appelée « aire de concentration ». On peut la mesurer par une formule, appelée « Indice de concentration de GINI » ou plus communément « indice de GINI », que nous allons maintenant étudier.

C – L'indice de GINI

Géométriquement, l'indice de GINI, du nom du statisticien italien Corrado GINI (1884-1965), est égal à l'aire de concentration, divisée par la moitié de la surface du carré (c'est-à-dire ½) :

$$\text{Indice de GINI} = \frac{\text{aire de concentration}}{1/2} = 2 \text{ aires de concentration}$$

Si l'on dispose de papier millimétré, on peut compter les petits carrés et avoir une idée approximative de la surface de l'aire de concentration. Mais il est préférable d'utiliser la formule analytique.

La formule analytique de l'indice de GINI est donnée par :

$$I = \frac{\sum_i \sum_j |x_i - x_j| n_i n_j}{2n(n-1)\bar{x}}$$

Pour voir ce que représentent les x_i et les x_j, ainsi que les n_i et les n_j, le mieux est d'appliquer la formule à un exemple.

Exemple : Soit le tableau suivant d'un groupe de 15 individus répartis en fonction de la valeur de leur patrimoine (en millions d'euros). La troisième colonne indique les centres de classe.

Tableau 15

Gains	Effectifs (n_i)	Centres (c_i)
[0,5 -1[	1	0,75
[1-2[	2	1,5
[2-3[	6	2,5
[3-4[	4	3,5
[4-5[	2	4,5

Afin de calculer le **numérateur** de la formule, il faut disposer les chiffres dans un tableau, de la façon suivante :

Tableau 16 : Disposition des calculs pour la détermination de l'indice de GINI

$|x_i - x_j| n_i n_j$
$= |1,5 - 0,75| 1 \times 2$

	x_i	**0,75**	**1,5**	**2,5**	**3,5**	**4,5**	$\sum$
x_j	n_j / n_i	**1**	**2**	**6**	**4**	**2**	**15**
0,75	**1**	0	1,5	10,5	11	7,5	**30,5**
1,5	**2**	1,5	0	12	16	12	**41,5**
2,5	**6**	10,5	12	0	24	24	**70,5**
3,5	**4**	11	16	24	0	8	**59**
4,5	**2**	7,5	12	24	8	0	**51,5**
$\sum$	**15**	**30,5**	**41,5**	**70,5**	**59**	**51,5**	253

La somme de la dernière colonne est égale à la somme de la dernière ligne, ce qui confirme qu'il n'y a pas d'erreur. Par conséquent :

$$\sum_i \sum_j |x_i - x_j| n_i n_j = 253$$

Reste à calculer le dénominateur et en particulier la moyenne :

$$\bar{x} = \frac{1}{15}\left[(1\times 0,75)+(2\times 1,5)+(6\times 2,5)+(4\times 3,5)+(2\times 4,5)\right] = 2,78333$$

Par conséquent :

$$2n(n-1)\bar{x} = 2\times 15\times(15-1))\times 2,78333 = 1169$$

Et donc :

$$I = \frac{\sum_i \sum_j |x_i - x_j| n_i n_j}{2n(n-1)\bar{x}} = \frac{253}{1169} = 0,22$$

D – L'écart médiale/médiane rapporté à l'intervalle de variation

L'autre façon de mesurer la concentration consiste à calculer le ratio suivant :

$$IC = \frac{Ml - Me}{IV}$$

Où Ml est la médiale, Me la médiane et IV l'intervalle de variation.

Exemple : Reprenons les données du tableau 15. Disposons le tableau des calculs intermédiaires pour la médiane et la médiale :

Tableau 17

Gains	Effectifs (n_i)	Centres (c_i)	$n_i c_i$	$n_i \uparrow$	$n_i c_i \uparrow$
[0,5 -1[	1	0,75	0,75	1	0,75
[1-2[	2	1,5	3	3	3
[2-3[	6	2,5	15	9	18,75
[3-4[	4	3,5	14	13	32,75
[4-5]	2	4,5	9	15	41,75

$\frac{n}{2} = \frac{15}{2} = 7,5 \longrightarrow$ Classe médiane : [2-3[

$\frac{\sum_{i=1}^{5} n_i c_i}{2} = \frac{41,75}{2} = 20,88 \longrightarrow$ Classe médiale : [3-4 [

Calculons la médiale :

$$M_l = x_i^{\inf} + a_i \left[\frac{\frac{\sum_{i=1}^{k} n_i x_i}{2} - N(n_i x_i)}{n_i x_i} \right] = 3 + 1 \left[\frac{41{,}75 - 18{,}75}{14} \right] = 1{,}6429$$

Calculons la médiane :

$$M_e = x_i^{\inf} + a_i \times \left[\frac{\frac{n}{2} - N(x_{i-1})}{n_i} \right] = 2 + 1 \times \left[\frac{7{,}5 - 3}{9} \right] = 1$$

L'intervalle de variation est égal à :

$$IV = 5 - 0{,}5 = 4{,}5$$

Par conséquent on a :

$$IC = \frac{Ml - Me}{IV} = \frac{1{,}6429 - 1}{4{,}5} = 0{,}1429$$

PARTIE 2

Les séries statistiques à deux dimensions

CHAPITRE 5

Les séries statistiques à deux dimensions I : Tableaux, graphiques, vocabulaire

La diffusion dans le grand public de logiciels permettant de produire des **tableaux et des graphiques à deux dimensions**, ainsi que divers calculs sur les séries à deux dimensions a grandement facilité leur étude, autrefois considérée comme difficile. Parmi les logiciels absolument incontournables, citons le logiciel Excel, de la suite OFFICE de MICROSOFT, qui permet de réaliser un très large éventail de graphiques et de tableaux, avec simplement quelques minutes de formation.

La suite OPEN OFFICE, téléchargeable sur http://fr.openoffice.org/ a des fonctionnalités identiques à celles d'OFFICE, mais possède l'avantage d'être gratuite.

Le logiciel de calcul et de traitement graphique le plus complet reste cependant MATHEMATICA (www.wolfram.com) qui possède des fonctionnalités très étendues, tant au niveau des possibilités de production de graphiques et de tableaux, que des possibilités d'analyse statistique et mathématique. Il est malheureusement trop coûteux pour en envisager l'acquisition à titre individuel.

Avant d'utiliser ces logiciels, il est cependant indispensable d'acquérir les bases nécessaires à la compréhension des concepts et outils statistiques développés pour la présentation et l'analyse des séries statistiques à deux dimensions. C'est pourquoi, dans ce chapitre (et le suivant), nous étudierons en détails ces méthodes de présentation et ces outils, en simplifiant au maximum les exemples proposés, sachant qu'une fois ces bases maîtrisées, l'étudiant pourra demander à un logiciel de faire les graphiques et les calculs.

1 • TABLEAUX ET GRAPHIQUES

A – Séries quantitatives connues individuellement

Exemple : on dispose des mesures de taille et de poids de 19 adolescents. Les données sont présentées par paires. Le premier élément de la paire correspond à la taille et le second au poids.

{{140 ; 38,2} ; {161 ; 44,3} ; {155 ; 46,1} ; {148 ; 38,2} ;{155 ; 50,5} ;{123 ; 22,4} ; {160 ; 40,4} ; {140 ; 34,7} ; {165 ; 50,5} ; {172 ; 50,5} ; {155 ; 38,1} ; {160 ; 57,3} ; {142 ; 39,3} ; {157 ; 46,1} ; {142;37,1} ; {148 ; 45,9} ; {180 ; 66,3} ; {167 ; 60} ; {165 ; 50,5}}

La présentation des données dans un tableau à deux dimensions est données ci-dessous, avec la représentation graphique la plus courante qui est celle dite du « **nuage de points** ».

Tableau 1

Taille	Poids
140	38,2
161	44,3
155	46,1
148	38,2
155	50,5
123	22,4
160	40,4
140	34,7
165	50,5
172	50,5
155	38,1
160	57,3
142	39,3
157	46,1
142	37,1
148	45,9
180	66,3
167	60
165	50,5

Figure 1

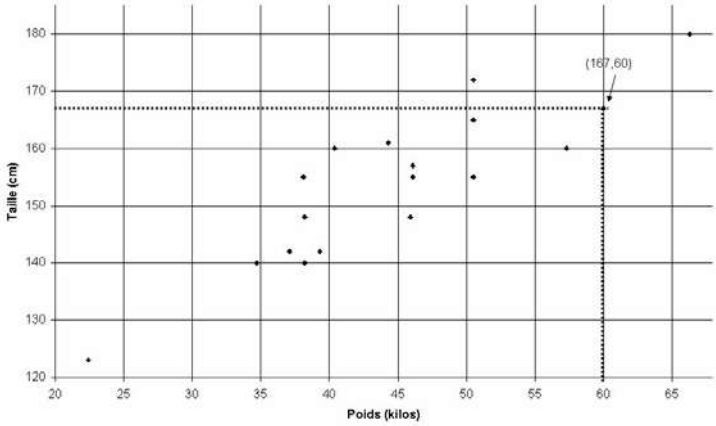
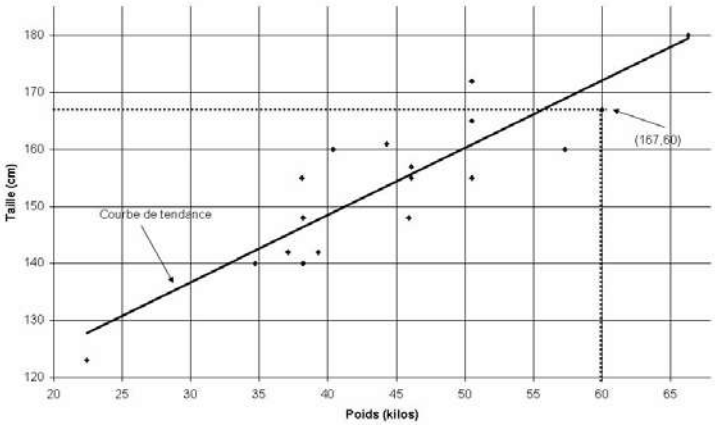


Figure 2



Ce graphique permet d'avoir un aperçu visuel de l'existence ou non d'une corrélation entre les deux variables, ici la taille et le poids. Ainsi, sur la figure 2, une **droite « de tendance »** a été ajoutée. Les coefficients de cette droite peuvent être calculés précisément (c'est l'objet du chapitre 6). On se contentera ici de noter que les points se regroupent assez bien autour de cette droite, ce qui semble confirmer que, toutes choses égales par ailleurs, il existe une relation positive entre la taille et le poids.

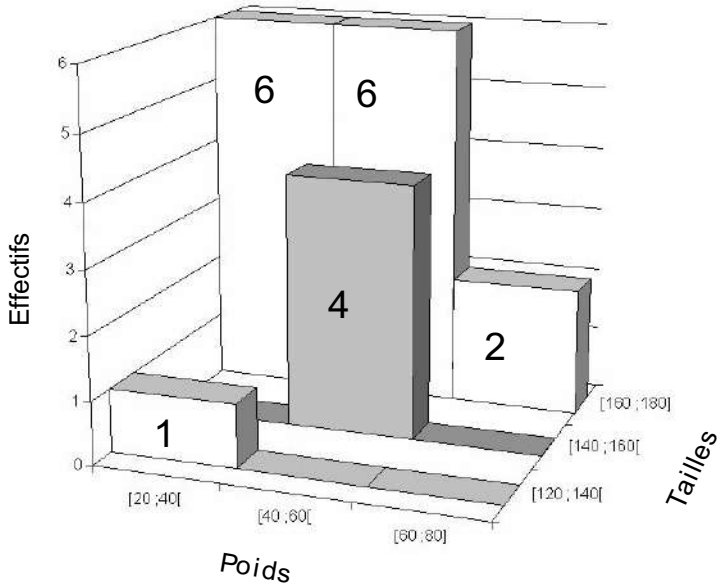
B – Séries quantitatives groupées

Exemple : Les données de l'exemple 1 concernant la taille et le poids de 19 adolescents ont été regroupées par classe dans le t**ableau de contingence** ci-après.

Tableau 2

Taille \ Poids	[20 ;40[	[40 ;60[	[60 ;80]
[120 ;140[	1	0	0
[140 ;160[	6	4	0
[160 ;180]	0	6	2

Figure 3



La figure 3 illustre la représentation classique sous forme d'un **histogramme à trois dimensions** : le poids, la taille et les effectifs. Les effectifs non nuls ont été reportés directement sur les barres.

C – Séries qualitatives

Exemple : supposons que l'on ait les données suivantes sur le sexe et le statut d'activité de 20 personnes. Les données sont présentées par paire. La première information concerne le sexe avec les deux modalités M et F. La seconde information concerne le statut d'activité, avec trois modalités (actif occupé [AO], chômeur [C], inactif [I]).

{{F ; AO} ; {M ; I} ; {F ; C} ; {F ; C} ; {M ; AO} ; {M ; AO} ; {M ; C} ; {F ; I} ; {F ;I } ; {F ;I } ; {M ; C} ;
{F ; AO} ; {F ; AO} ; {F ; AO} ; {M ; AO} ; {M ; C} ; {M ; AO} ; {F ; I} ; {F , C} ; {M , AO}}

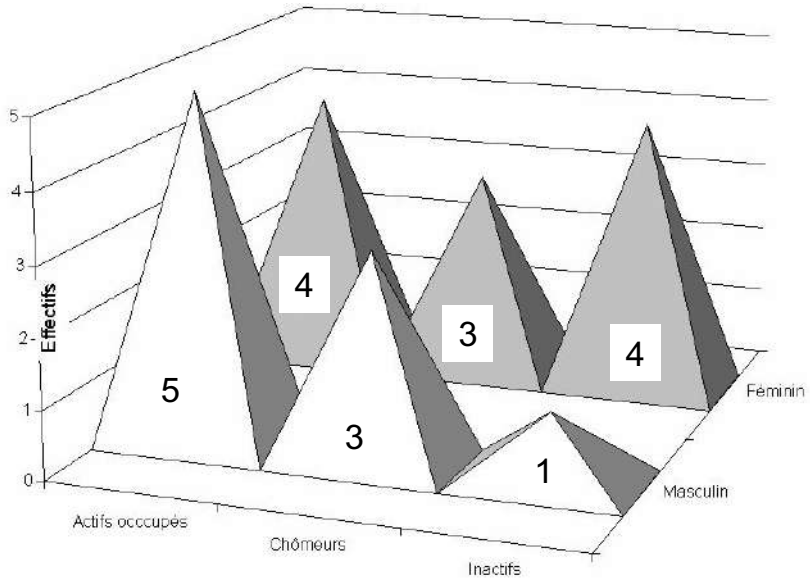
Regroupons ces données dans un tableau de contingence :

Tableau 3

Sexe \ Statut	Actifs occupés	Chômeurs	Inactifs
Masculin	5	3	1
Féminin	4	3	4

On obtient le graphique suivant, qui est une variante d'histogramme :

Figure 4



2 • REPRÉSENTATION ABSTRAITE D'UN TABLEAU DE CONTINGENCE

Le tableau 4 représente un tableau de contingence sous forme symbolique.
À l'intersection de la modalité x_i et de la modalité y_j se trouve l'effectif correspondant.

Tableau 4

Valeurs ou modalités de Y

x \ y	y_1	y_2	…	y_j	…	y_q	$n_{i\bullet}$
x_1							$n_{1\bullet}$
x_2		n_{22}				n_{2q}	$n_{3\bullet}$
…							…
x_i				n_{ij}			$n_{i\bullet}$
…							…
x_p						n_{pq}	$n_{p\bullet}$
$n_{\bullet j}$	$n_{\bullet 1}$	$n_{\bullet 2}$	…	$n_{\bullet j}$	…	$n_{\bullet q}$	$n_{\bullet\bullet}$

Valeurs ou modalités de X

Effectifs marginaux de x

Effectifs marginaux de y

L'effectif n_{ij} représente le nombre d'individus qui ont à la fois la modalité/valeur x_i et la modalité/valeur y_j. On a ensuite les symboles suivants :

n_{22} : effectif des individus qui ont la modalité/valeur 2 de x et la modalité 2 de Y.

Par convention, on notre toujours la modalité/valeur de X (i) avant celle de Y (j).

n_{2q} : effectif des individus qui ont la modalité/valeur 2 de x et la modalité q de Y.

n_{pq} : effectif des individus qui ont la modalité/valeur p de x et la modalité/valeur q de Y.

$n_{1\bullet}$: effectif des individus qui ont la modalité/valeur i (le « • » à la place du j signifie que l'on ne tient pas compte de Y). Exemple : $n_{1\bullet}$ désigne tout l'effectif des individus qui ont la modalité/valeur 1 de X.

$n_{\bullet j}$: effectif des individus qui ont la modalité j (le "•" à la place du i signifie que l'on ne tient pas compte de X). Exemple : $n_{\bullet 1}$ désigne **tout** l'effectif des individus qui ont la modalité/valeur 1 de Y.
$n_{\bullet\bullet}$: effectif total.

Dès lors : $$n_{i\bullet} = \sum_{j=1}^{q} n_{ij} = n_{i1} + n_{i2} + \cdots + n_{iq}$$

$$n_{\bullet j} = \sum_{i=1}^{p} n_{ij} = n_{1j} + n_{2j} + \cdots + n_{pj}$$

$$n_{\bullet\bullet} = \sum_{i=1}^{p} n_{i\bullet} = \sum_{i=1}^{p}\left(\sum_{j=1}^{q} n_{ij}\right) = \sum_{j=1}^{q} n_{\bullet j} = \sum_{j=1}^{q}\left(\sum_{i=1}^{p} n_{ij}\right)$$

Exemple : Soit le tableau de contingence suivant d'un groupe de 50 personnes réparties par groupe d'âge (« x ») et par sexe (« y »), tous âgés de 45 ans au plus.

Tableau 5

x \ y	H	F
[0-18 [	10	20
[18 -45]	5	15

En reprenant la notation du tableau 4 on a ici :

$$n_{11} = 10;\ n_{12} = 20; n_{21} = 5; n_{22} = 15$$

$$n_{1\bullet} = n_{11} + n_{12} = 10 + 20 = 30 \qquad n_{2\bullet} = n_{21} + n_{22} = 5 + 15 = 20$$

$$n_{\bullet 1} = n_{11} + n_{21} = 10 + 5 = 15 \qquad n_{\bullet 2} = n_{12} + n_{22} = 20 + 15 = 35$$

$$n_{\bullet\bullet} = n_{11} + n_{12} + n_{21} + n_{22} = 10 + 20 + 5 + 15 = 50$$
$$n_{\bullet\bullet} = n_{1\bullet} + n_{2\bullet} = 30 + 20 = 50$$
$$n_{\bullet\bullet} = n_{\bullet 1} + n_{\bullet 2} = 15 + 35 = 50$$

3 • EFFECTIFS MARGINAUX ET FRÉQUENCES MARGINALES

Ajoutons une ligne et une colonne au tableau 5, et remplissons-les par les résutats des sommes que nous venons juste de calculer.

Tableau 6

x \ y	H	F	$n_{i\bullet}$
[0-18 [	10	20	**30**
[18 -45]	5	15	**20**
	15	**35**	**50**

Cette ligne et cette colonne que nous venons d'ajouter, ce sont les distributions marginales du tableau de contingence. Ainsi, la colonne $n_{i\bullet}$ représente la **distribution marginale de x**, c'est-à-dire les valeurs possibles de x quel que soit y. De même la ligne $n_{\bullet j}$ représente la **distribution marginale de y**, c'est-à-dire les valeurs possibles de y quel que soit x.

Les **fréquences marginales de x** s'obtiennent en divisant la colonne par son total soit dans l'exemple 30+20 = 50. De même les **fréquences marginales de y** s'obtiennent en divisant la ligne par son total soit dans l'exemple 15+35 = 50. Le tableau 7 donne les fréquences marginales de x et de y dans le cas du tableau 6.

Tableau 7

x \ y	H	F	$n_{i\bullet}$
[0-18 [	10	20	**30/50=0,6**
[18 -45]	5	15	**20/50=0,4**
	15/50=0,3	**35/50=0,7**	**50**

Plus formellement, les définitions des fréquences marginales sont données par :

Fréquences marginales de x : $f_{i\bullet} = \dfrac{n_{i\bullet}}{n_{\bullet\bullet}}$ i=1,…,p

Fréquences marginales de y : $f_{\bullet j} = \dfrac{n_{\bullet j}}{n_{\bullet\bullet}}$ j=1,…,q

Ainsi, dans l'exemple du tableau 7, on a :

$$f_{1\bullet} = \frac{n_{1\bullet}}{n_{\bullet\bullet}} = \frac{30}{50} = 0{,}6 \quad f_{2\bullet} = \frac{n_{2\bullet}}{n_{\bullet\bullet}} = \frac{20}{50} = 0{,}4 \quad f_{\bullet 1} = \frac{n_{\bullet 1}}{n_{\bullet\bullet}} = \frac{15}{50} = 0{,}3 \quad f_{2\bullet} = \frac{n_{\bullet 2}}{n_{\bullet\bullet}} = \frac{35}{50} = 0{,}7$$

4 • MOYENNES ET VARIANCES MARGINALES

A – Moyennes marginales

Les **moyennes marginales** de x et de y se calculent à partir des distributions marginales suivant les formules suivantes :

$$\overline{\overline{x}} = \frac{1}{n_{\bullet\bullet}} \sum_{i=1}^{p} n_{i\bullet} x_i$$

$$\overline{\overline{y}} = \frac{1}{n_{\bullet\bullet}} \sum_{j=1}^{q} n_{\bullet j} y_j$$

Où le signe « = » situé sur x et y permet de rappeler qu'il s'agit de moyennes de distributions marginales.

Exemple : Soit le tableau de contingence suivant

Tableau 8

x \ y	1	4	$n_{i\bullet}$
2	3	5	**8**
8	4	12	**16**
$n_{\bullet j}$	**7**	**17**	**24**

Calculons la moyenne marginale de x :

$$\overline{\overline{x}} = \frac{1}{n_{\bullet\bullet}} \sum_{i=1}^{p} n_{i\bullet} x_i = \frac{1}{24}\left[(8\times 2)+(16\times 8)\right] = 6$$

Ainsi que la moyenne marginale de y :

$$\overline{\overline{y}} = \frac{1}{n_{\bullet\bullet}} \sum_{j=1}^{q} n_{\bullet j} x_j = \frac{1}{24}\left[(7\times 1)+(17\times 4)\right] = 3{,}125$$

B – Variances marginales

Les **variances marginales** de x et de y se calculent à partir des distributions marginales suivant les formules suivantes :

$$\sigma_x^2 = \frac{1}{n_{\bullet\bullet}} \sum_{i=1}^{p} n_{i\bullet}\left(x_i - \overline{\overline{x}}\right) = \frac{1}{n_{\bullet\bullet}} \sum_{i=1}^{p} n_{i\bullet} x_i^2 - \left(\overline{\overline{x}}\right)^2$$

$$\sigma_y^2 = \frac{1}{n_{\bullet\bullet}} \sum_{j=1}^{q} n_{\bullet j}\left(y_j - \overline{\overline{y}}\right) = \frac{1}{n_{\bullet\bullet}} \sum_{j=1}^{q} n_{\bullet j} y_j^2 - \left(\overline{\overline{y}}\right)^2$$

Exemple : Calculons les variances marginales de x et de y à partir des données du tableau 8. Disposons les calculs sous forme de tableaux.

Tableau 9

x_i	$n_{i\bullet}$	x_i^2	$n_{i\bullet}x_i^2$
2	8	4	32
8	16	64	1024

$\sum_{i=1}^{p} n_{i\bullet} x_i^2 \rightarrow 1056$

y_j	$n_{\bullet j}$	y_j^2	$n_{\bullet j}y_j^2$
1	7	1	7
4	17	16	272

$\sum_{j=1}^{q} n_{\bullet j} y_j^2 \rightarrow 279$

$$\sigma_x^2 = \frac{1}{n_{\bullet\bullet}} \sum_{i=1}^{p} n_{i\bullet} x_i^2 - \left(\overline{\overline{x}}\right)^2 = \frac{1}{24}(1056) - 6^2 = 8$$

$$\sigma_y^2 \frac{1}{n_{\bullet\bullet}} \sum_{j=1}^{q} n_{\bullet j} y_j^2 - \left(\overline{\overline{y}}\right)^2 = \frac{1}{24}(279) - (3{,}125)^2 = 1{,}859375$$

5 • FRÉQUENCES PARTIELLES SUR EFFECTIF TOTAL

Les fréquences partielles sur effectif total s'obtiennent en divisant chaque n_{ij} par l'effectif total.

Exemple : Calculons les fréquences partielles sur effectif total du tableau 8

Tableau 10

x \ y	1	4
2	(3/24) = 0,125	(5/24) = 0,208
8	(4/24) = 0,167	(12/24) = 0,5

$n_{\bullet\bullet} = \mathbf{24}$

On remarquera que la somme des effectifs partiels sur effectif total est égale à 1. En effet :

$$0{,}125 + 0{,}208 + 0{,}167 + 0{,}5 \cong 1$$

Plus précisément, l'effectif partiel sur effectif total se définit par la notation :

$$f_{ij} = \frac{n_{ij}}{n_{\bullet\bullet}}$$

On a donc :

$$f_{11} + f_{12} + f_{21} + f_{22} = 1$$

6 • DISTRIBUTIONS CONDITIONNELLES

Les distributions conditionnelles s'obtiennent en fixant la valeur d'une des deux variables (où la modalité d'un des deux caractères).

Exemple 1 : Dans le cas de chiffres du tableau 8, la distribution conditionnelle de x quand y = 1 est donnée par la première colonne du tableau. De même, la distribution conditionnelle de x quand y = 4 est donnée par la deuxième colonne du tableau. Le tableau 11 illustre les deux distributions conditionnelles de x pour y donné. Il y a deux distributions conditionnelles de x car y ne prend ici que deux valeurs. En général, sachant que j varie de 1 à q, il y a q distributions conditionnelles de x.

Tableau 11

x \ y	**1**	**4**	$n_{i\bullet}$
2	3	5	**8**
8	4	12	**16**
$n_{\bullet j}$	**7**	**17**	**24**

Distribution conditionnelle de x quand y = 1

Distribution conditionnelle de x quand y = 4

Exemple 2 : Toujours en prenant les chiffres du tableau 8, la distribution conditionnelle de y quand x = 2 est donnée par la première ligne du tableau. De même, la distribution conditionnelle de y quand x = 8 est donnée par la deuxième ligne du tableau. Le tableau 12 illustre les deux distributions conditionnelles de y pour x donné. Il y a deux distributions conditionnelles de y car x ne prend ici que deux valeurs. En général, sachant que i varie de 1 à p, il y a p distributions conditionnelles de y.

Tableau 12

x \ y	**1**	**4**	$n_{i\bullet}$
2	3	5	**8**
8	4	12	**16**
$n_{\bullet j}$	**7**	**17**	**24**

Distribution conditionnelle de y quand x = 2

Distribution conditionnelle de y quand y = 8

7 • MOYENNES ET VARIANCES CONDITIONNELLES

A – Moyennes conditionnelles

Pour chaque distribution conditionnelle, on peut calculer une moyenne. Ainsi, dans le cas du tableau 8, puisqu'il y a deux distributions conditionnelles de x, il y a deux moyennes conditionnelles de que nous noterons respectivement :

$\bar{x}_1$ pour désigner la moyenne conditionnelle de x quand y = 1

$\bar{x}_2$ pour désigner la moyenne conditionnelle de x quand y = 4

De la même façon, puisqu'il y a deux distributions conditionnelles de y, il y a deux moyennes conditionnelles de y que nous noterons respectivement :

$\bar{y}_1$ pour désigner la moyenne conditionnelle de y quand x = 2

$\bar{y}_2$ pour désigner la moyenne conditionnelle de x quand x = 8

Exemple 1 : Calculons les deux moyennes conditionnelles de x dans le cas des données du tableau 8 :

$$\bar{x}_1 = \frac{1}{7}\left[(3\times 2)+(4\times 8)\right] = 5{,}4286$$

$$\bar{x}_2 = \frac{1}{17}\left[(5\times 2)+(12\times 8)\right] = 6{,}23529$$

La formule des moyennes conditionnelles de x :est donc donnée par :

$$\bar{x}_j = \frac{1}{n_{\bullet j}}\sum_{i=1}^{p} n_{ij}x_i \qquad 1\le j\le p$$

Exemple 2 : Calculons les deux moyennes conditionnelles de y dans le cas des données du tableau 8 :

$$\bar{y}_1 = \frac{1}{8}\left[(3\times 1)+(5\times 4)\right] = 2{,}875$$

$$\bar{y}_2 = \frac{1}{16}\left[(4\times 1)+(12\times 4)\right] = 3{,}25$$

La formule des moyennes conditionnelles de y est donc donnée par :

$$\overline{y}_j = \frac{1}{n_{i\bullet}} \sum_{j=1}^{q} n_{ij} y_j \qquad 1 \leq i \leq q$$

B – Variances conditionnelles

Pour chaque distribution conditionnelle, on peut calculer une variance. Ainsi, dans le cas du tableau 8, puisqu'il y a deux distributions conditionnelles de x, il y a deux variance conditionnelles de x, que nous noterons respectivement :

$V(x_1)$ pour désigner la variance conditionnelle de x quand y = 1

$V(x_2)$ pour désigner la variance conditionnelle de x quand y = 4

De la même façon, puisqu'il y a deux distributions conditionnelles de y, il y a deux variance conditionnelles de y que nous noterons respectivement :

$V(y_1)$ pour désigner la variance conditionnelle de y quand x = 2

$V(y_2)$ pour désigner la variance conditionnelle de x quand x = 8

Exemple 1 : Calculons les deux variances conditionnelles de x dans le cas des données du tableau 8 :

$$V(x_1) = \frac{1}{7}\left[\left(3 \times 2^2\right) + \left(4 \times 8^2\right)\right] - \left(5{,}428\right)^2 = 8{,}816$$

$$V(x_2) = \frac{1}{17}\left[\left(5 \times 2^2\right) + \left(12 \times 8^2\right)\right] - \left(6{,}2353\right)^2 = 7{,}474$$

La formule des variances conditionnelles de x est donc donnée par :

$$V(x_j) = \frac{1}{n_{\bullet j}} \sum_{i=1}^{p} n_{ij} \left(x_i - \overline{x}_j\right)^2 = \frac{1}{n_{\bullet j}} \sum_{i=1}^{p} n_{ij} x_i^2 - \overline{x}_j^2$$

Exemple 2 : Calculons les deux variances conditionnelles de y dans le cas des données du tableau 8 :

$$V(y_1) = \frac{1}{8}\left[\left(3 \times 1^2\right) + \left(5 \times 4^2\right)\right] - \left(2{,}875\right)^2 = 2{,}1094$$

$$V(y_2) = \frac{1}{16}\left[\left(4 \times 1^2\right) + \left(12 \times 4^2\right)\right] - \left(3{,}25\right)^2 = 1{,}6875$$

La formule des variances conditionnelles de x est donc donnée par :

$$V(y_i) = \frac{1}{n_{i\bullet}} \sum_{j=1}^{q} n_{ij}\left(y_j - \overline{y}_i\right)^2 = \frac{1}{n_{i\bullet}} \sum_{j=1}^{q} n_{ij} y_j^2 - \overline{y}_i^2$$

CHAPITRE 6

Les séries statistiques à deux dimensions

II : Outils d'analyse

Il est fréquemment nécessaire d'étudier les liens qui peuvent exister entre les deux (ou plus de deux) dimensions qui caractérisent une population statistique. Pour qualifier ces liens on parle de liaison statistique, de corrélation mais, c'est important de le préciser, il n'est jamais question de causalité, la statistique descriptive n'ayant pas pour objet de prouver des causalités.

Ce chapitre se limite à l'étude des séries à deux dimensions, X et Y. Cela offre déjà un large éventail de possibilités si l'on se souvient que chacune de ces dimensions peut être quantitative, qualitative et que les données peuvent être groupées dans chaque cas par valeur ou groupes de valeurs. À ces différents cas, correspondent des outils d'analyse appropriés que nous allons évoquer successivement.

1 • SÉRIES QUANTITATIVES AVEC OBSERVATIONS CONNUES INDIVIDUELLEMENT

A – Liaison linéaire, liaison non linéaire, absence de liaison

On s'intéresse à une statistique ayant deux dimensions que nous désignons par les variables X et Y. La notion de **courbe de régression** est un concept général qui va nous permettre de mettre en évidence au moyen d'un graphique s'il existe une relation entre ces deux variables et quelle est la nature de cette relation.

La courbe de régression est en fait un tracé que l'on fait passer entre les observations d'un nuage de points. Le plus souvent, on essaie de tracer une droite (voir la figure 2 du chapitre 5) que l'on désigne alors par **droite de régression** ou, plus simplement par l'expression **droite de tendance**.

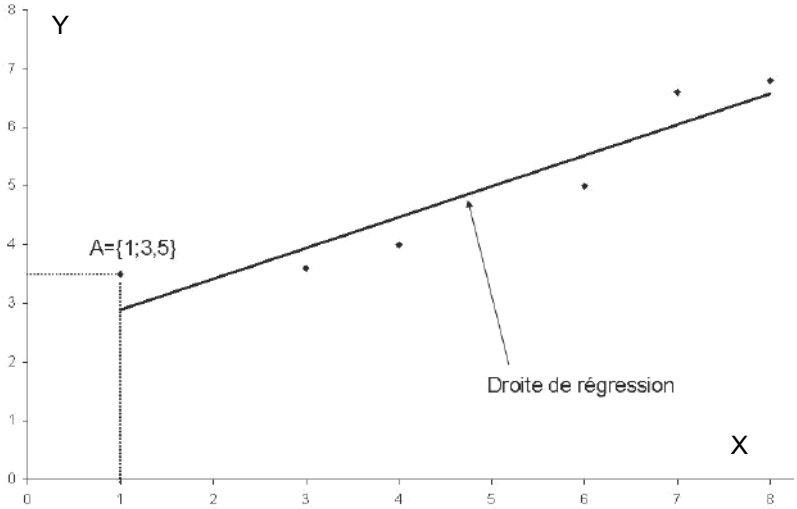
Exemple 1 : Soit S la série de données ci-dessous relatives aux deux variables X et Y, présentées par paires. Le premier élément de la paire correspond à la valeur de X et le second à la valeur de Y. Les éléments de chaque paire sont séparés par des points virgules afin de ne pas confondre la séparation des valeurs au sein de la paire, avec les décimales d'une valeur.

$$S = \{\{1 ; 3,5\} ; \{3 ; 3,6\} ; \{4 ; 4\} ; \{6 ; 5\} ; \{7 ; 6,6\} ; \{8 ; 6,8\}\}$$

Représentons ces données à l'aide d'un **nuage de points** (figure 1) où, par convention, la valeur X se lit en abscisse et la valeur Y en ordonnée. Ainsi, la paire qui correspond au point A sur le nuage de points est la première paire de S.

La valeur X = 1 se lit en abscisse et la valeur Y = 3,5 se lit en ordonnée. Il en va de même des cinq autres paires. Une main « experte » (celle du logiciel) a également tracé une droite, entre les points ; c'est la droite de régression ou droite de tendance.

Figure 1



Nous verrons un peu plus loin comment le tracé de cette droite peut s'effectuer mathématiquement et quelles sont les propriétés de la droite de régression. Toutefois, il convient de noter dès maintenant que la relation ainsi établie entre X et Y n'est pas nécessairement linéaire. Pour le montrer, prenons un nouvel exemple.

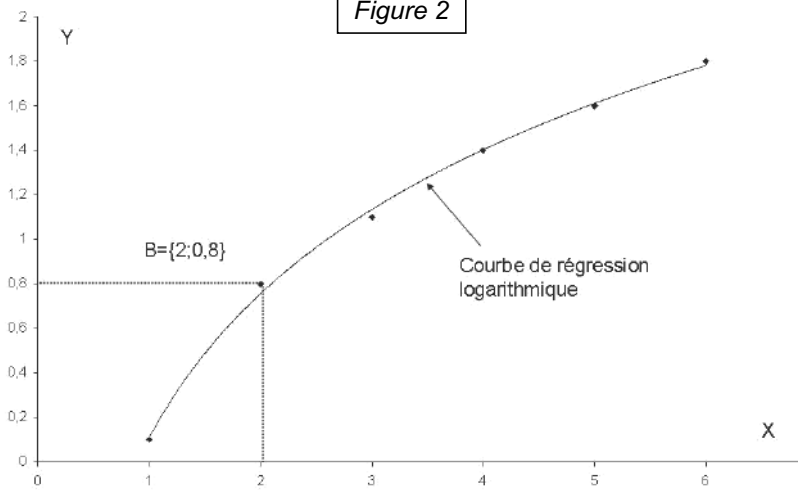
Exemple 2 : Soit les données ci-dessous relatives aux deux variables X et Y. Cette fois le nuage de points évoque davantage une courbe logarithmique qu'une droite linéaire. C'est pourquoi l'on a demandé à EXCEL de tracer une **courbe de régression** et que le logiciel a choisi un ajustement par une **courbe de régression logarithmique**, donc **non linéaire**.

T = {{1 ; 0,1} ; {2 ; 0,8} ; {3 ; 1,1} ; {4 ; 1,4} ; {5 ; 1,6} ; {6 ; 1,8}}

Quoique la très grande majorité des relations réelles entre variables ne soient pas linéaires, c'est néanmoins l'ajustement linéaire qui est retenu dans de nombreux cas, pour trois raisons :

1) L'ajustement linéaire est beaucoup plus simple à traiter mathématiquement.
2) Beaucoup de relations sont approximativement linéaires si l'on prend un intervalle de variation suffisamment petit.
3) Certaines relations peuvent être rendues linéaires par un changement de variable approprié (généralement une transformation logarithmique).

Figure 2

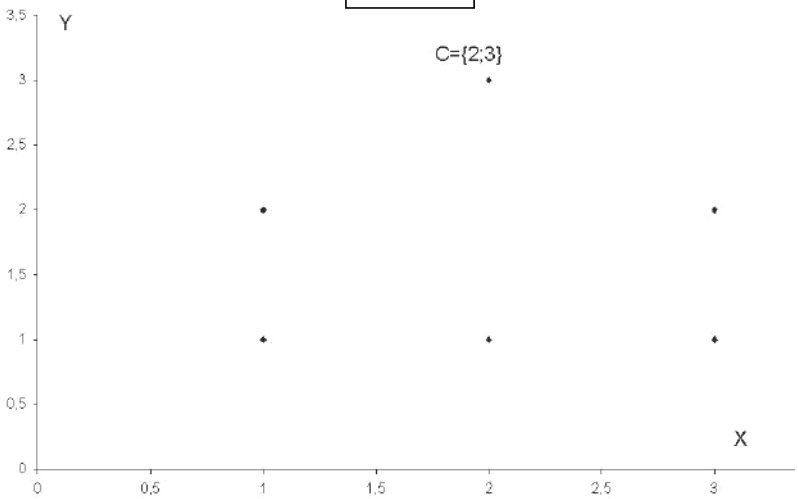


Pour finir, notons qu'il n'existe pas nécessairement de liaison entre deux variables, comme l'illustre l'exemple suivant d'**absence de relation**.

Exemple 3 : Soit les données ci-dessous relatives aux deux variables X et Y. Cette fois le nuage de points évoque davantage un amas de points. On peut certes y voir une forme non linéaire (si on relie les points on obtient un dessin de maison), mais il resterait alors à interpréter cette relation.

U = {{1 ; 1} ; {1 ; 2} ; {2 ; 3} ; {3 ; 2} ; {3 ; 1} ; {2 ; 1}}

Figure 3



B – La droite de régression linéaire

1) Définition

Le **point moyen** est le point qui a pour coordonnées la moyenne de X et la moyenne de Y. On l'appelle aussi le **centre de gravité**.

La **droite de régression** est une droite qui passe par le **point moyen**. C'est aussi la droite qui **minimise la somme des carrés des écarts des observations**. Une fois connue, l'équation de cette droite permet de résumer la série et de faire des prévisions.

Exemple : Soit la série S déjà étudiée au paragraphe A

S = {{1 ; 3,5} ; {3 ; 3,6} ; {4 ; 4} ; {6 ; 5} ; {7 ; 6,6} ; {8 ; 6,8}}

La moyenne de X est donnée par (le « double barre » sur le X indique qu'il s'agit d'une moyenne marginale) :

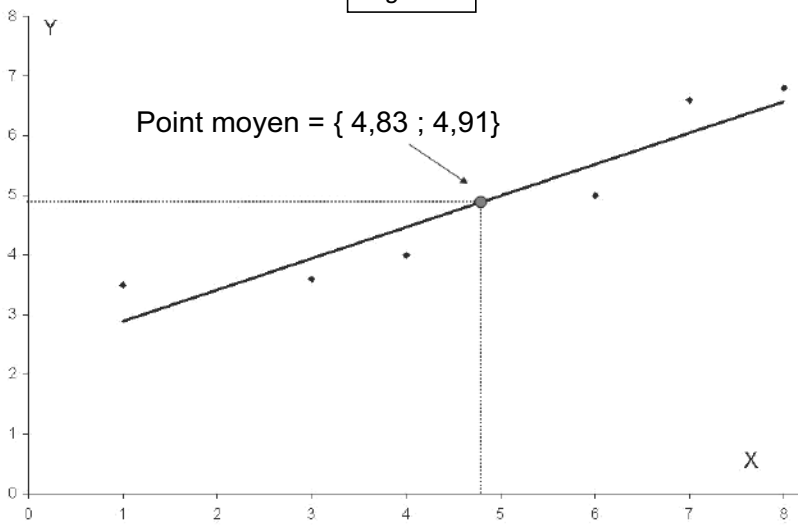
$$\bar{\bar{x}} = \frac{1+3+4+6+7+8}{6} = \frac{29}{6} = 4{,}833$$

La moyenne marginale de Y est donnée par :

$$\bar{\bar{y}} = \frac{3{,}5+3{,}6+4+5+6{,}6+6{,}8}{6} = \frac{29{,}5}{6} = 4{,}91$$

Le graphique de la figure 4, illustre le point moyen :

Figure 4



2) Calcul des coefficients

L'équation de la droite de régression se calcule ainsi. Soit la droite d'équation :

$$y = ax + b$$

Si nous voulons que cette droite soit ajustée à un nuage de points dans le plan {X,Y}, il faut calculer les coefficients a et b en appliquant les formules suivantes :

$$a = \frac{\operatorname{cov}(x,y)}{\sigma_x^2} \qquad b = \overline{\overline{y}} - a\overline{\overline{x}}$$

où cov(x,y) représente la covariance de (x,y) et se calcule ainsi :

$$\operatorname{cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{\overline{x}}\,\overline{\overline{y}}$$

Par conséquent, la formule détaillée de *a* est :

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

Exemple : calculons *a* et *b* dans le cas de la série S :

S = {{1 ; 3,5} , {3 ; 3,6} , {4 ; 4} , {6 ; 5} , {7 ; 6,6} , {8 ; 6,8}}

Pour faciliter les calculs, adoptons la disposition en tableau suivante :

Tableau 1

	X	Y	XY	X^2	Y^2
	1	3,5	3,5	1	12,25
	3	3,6	10,8	9	12,96
	4	4	16	16	16
	6	5	30	36	25
	7	6,6	46,2	49	43,56
	8	6,8	54,4	64	46,24
Sommes →	**29**	**29,5**	**160,9**	**175**	**156**

Ensuite, calculons les sommes dont nous avons besoin dans la formule de *a* :

$$\sum_{i=1}^{n} x_i = 29 \qquad \sum_{i=1}^{n} y_i = 29,5 \qquad \sum_{i=1}^{n} x_i y_i = 160,9 \qquad \sum_{i=1}^{n} x_i^2 = 175 \qquad \sum_{i=1}^{n} y_i^2 = 156$$

calculons *a* :

$$a = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{\overline{xy}}}{\frac{1}{n}\sum_{i=1}^{n} x^2 - \left(\overline{\overline{x}}\right)^2} = \frac{\frac{160,9}{6} - \frac{29}{6} \times \frac{29,5}{6}}{\frac{175}{6} - \left(\frac{29}{6}\right)^2} = 0,5258$$

Une fois *a* connu, on en déduit *b* :

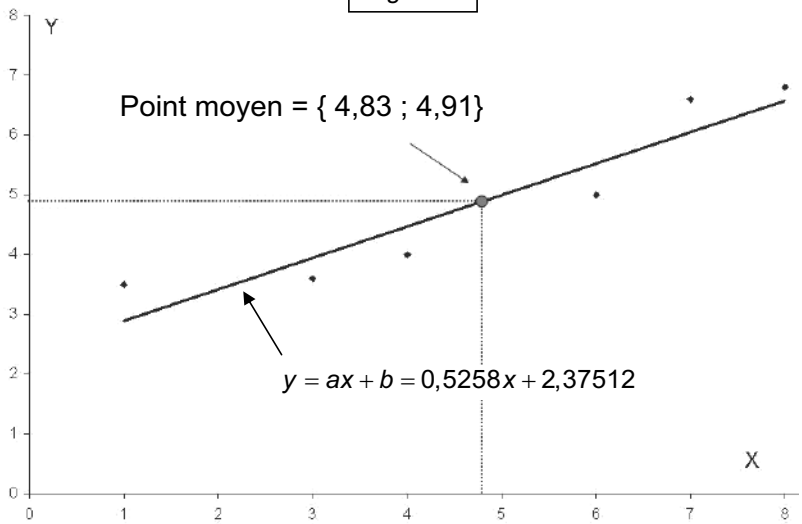
$$b = \overline{\overline{y}} - a\overline{\overline{x}} = \left(\frac{29,5}{6}\right) - 0,5258 \times \left(\frac{29}{6}\right) = 2,37512$$

L'équation de la droite de régression est donc :

$$y = ax + b = 0,5258x + 2,37512$$

La figure 5 ci-dessous illustre l'équation de cette droite. Nous vérifions à nouveau que cette droite passe par le point moyen.

Figure 5



3) Utilité de la droite de régression

La droite de régression sert d'abord à **vérifier l'existence d'une relation linéaire** et la nature de celle-ci. Ainsi, dans notre exemple, le coefficient directeur de la droite a=0,5258 est positif ce qui dénote une relation positive : x et y varient dans le même sens.

La droite de régression sert ensuite à **faire des prévisions**. Ainsi, nous pouvons utiliser l'équation de la droite de régression pour calculer des valeurs de Y associées à une valeur de X que l'on se donne.

Exemple 1 : Soit la série S, déjà étudiée précédemment et supposons que l'on veuille connaître la valeur Y qui correspond à X = 12 que l'on se donne et qui ne figure pas dans S. Dans ce cas, il suffit de remplacer X par dans l'équation de la droite pour obtenir Y :

$$y = 0{,}5258 \times (12) + 2{,}37512 = 8{,}6847$$

Exemple 2 : Soit la série S, déjà étudiée précédemment et supposons que l'on veuille connaître la valeur X qui correspond à Y = 5 que l'on se donne. Dans ce cas, il suffit de remplacer Y par dans l'équation de la droite pour obtenir X :

$$5 = 0{,}5258x + 2{,}37512 \Leftrightarrow x = 4{,}99212 \cong 5$$

C – Le coefficient de corrélation

1) Définition et calcul

Le coefficient de corrélation mesure la plus ou moins grande dépendance entre les deux caractères X et Y. On le désigne par la lettre "r" et il varie entre -1 et +1 :

$$r = \frac{\operatorname{cov}(x, y)}{\sigma_x \sigma_y}$$

Plus r est proche de +1 ou de -1, plus les deux caractères sont dépendants. Plus il est proche de 0, plus les deux caractères sont indépendants.

Exemple : Calculons le coefficient de corrélation de la série S :

$$r = \frac{\operatorname{cov}(x, y)}{\sigma_x \sigma_y} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{\overline{xy}}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} x^2 - \left(\overline{\overline{x}}\right)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n} y^2 - \left(\overline{\overline{y}}\right)^2}} = \frac{\frac{160{,}9}{6} - \frac{29}{6} \times \frac{29{,}5}{6}}{\sqrt{\frac{175}{6} - \left(\frac{29}{6}\right)^2}\sqrt{\frac{156}{6} - \left(\frac{29{,}5}{6}\right)^2}} = 0{,}9371$$

2) Coefficient de corrélation et coefficient de détermination

Il existe un lien entre le coefficient de corrélation et la droite de régression. Ce lien est donné par la formule :

$$R^2 = a \times a'$$

où *a* est le coefficient de la droite de régression de *y* en *x* (c'est-à-dire la droite de régression de la forme *y* = *ax*+*b*) et où *a'* est le coefficient de la droite de régression de *x* en *y* (c'est-à-dire le coefficient de la droite de régression de *x* en *y*).

Le terme R^2 est appelé **coefficient de détermination**. En pratique, il n'est pas nécessaire de passer par la formule $R^2 = a \times a'$. Il suffit en effet de calculer *r* et de l'élever au carré.

Exemple : Calculons le coefficient de détermination de la série S :

$$R^2 = r \times r = 0,9371^2 = 0,8781$$

Contrairement ou coefficient de corrélation, qui varie entre -1 et +1, le coefficient de corrélation varie entre 0 et 1. Il sert aussi à mesurer la corrélation des deux variables, mais ne donne aucune indication sur le sens (positif ou négatif) de la corrélation. Plus il est proche de 0, plus la corrélation est faible. Plus il est proche de 1, plus la corrélation est élevée.

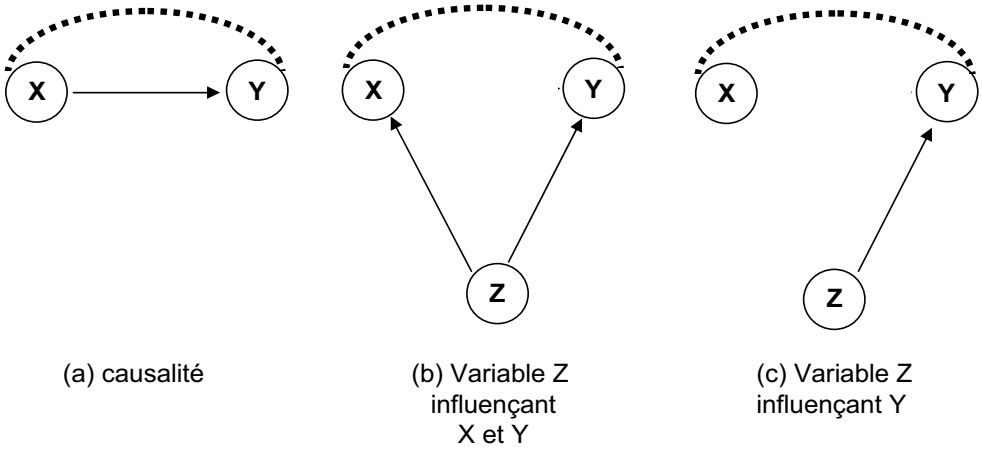
3) Corrélation et causalité

Le plus souvent, l'étude des relations entre deux variables a pour but plus ou moins avoué d'apprécier dans quelle mesure l'une des deux variables – dite variable explicative – exerce une influence causale sur l'autre – dite variable expliquée.

Malheureusement, ainsi que nous l'avons indiqué en introduction, la corrélation n'implique pas la causalité, pour diverses raisons que nous allons maintenant approfondir.

La figure 6 illustre trois liens possibles entre les deux variables X et Y, liens qui sont tous compatibles avec un coefficient de corrélation identique, lequel ne permettra donc pas de discriminer entre les trois.

Figure 6



Source : D'après David S. MOORE et George P. McCABE, 2001, *Introduction to the Practice of Statistics*, W.H. Freeman & Company, New York, 3ème édition, page 208.

Sur la figure 6, les lignes en pointillés indiquent l'existence d'une corrélation ente les variables X et Y. Les lignes en trait plein indiquent l'existence d'une causalité et la flèche indique le sens de la causalité. Dans le cas (a), nous voyons que la causalité sous-jacente va de X vers Y, c'est-à-dire que les variations de X expliquent celles de Y. La corrélation observée est donc bien le résultat d'une causalité directe.

Cependant, comme la causalité n'est pas observable, on ne peut pas conclure à l'existence d'une causalité de X vers Y à la simple mise en évidence d'une corrélation. En effet, comme l'illustrent les cas (b) et (c) de la figure 6, la corrélation peut aussi s'expliquer différemment.

Dans le cas (b), c'est une variable Z, qui peut être inconnue ou connue mais non prise en compte, qui influence simultanément X et Y. Dans ce cas, on observera effectivement une corrélation entre X et Y, mais cette corrélation n'impliquera pas de causalité de X vers Y.

Dans le cas (c), c'est une variable Z, qui peut être inconnue ou connue mais non prise en compte, qui influence uniquement Y. Dans ce cas, on observera effectivement une corrélation entre X et Y, mais cette corrélation n'impliquera pas de causalité de X vers Y, puisque la variation de X est autonome et celle de Y causée par la variable Z.

En conclusion, il faut retenir que corrélation n'est pas causalité.

2 • Séries quantitatives avec observations groupées

Lorsque les observations sont fournies groupées par valeurs, il est possible de se ramener au cas d'une série quantitative où les observations sont connues individuellement.

En revanche, lorsque les observations sont fournies groupées par classes, une partie de l'information a été détruite au profit d'une information plus synthétique, mais on ne peut pas revenir aux paires de valeurs individuelles. On est alors obligé d'analyser les tableaux avec des formules modifiées, que ce soit pour la droite ou la courbe de régression ou pour le coefficient de corrélation. En fait, étant donné la lourdeur des calculs, il est sage de s'en tenir au calcul du **coefficient de corrélation**. En outre, comme nous le verrons, il existe un autre moyen de vérifier l'existence d'une corrélation, basé sur le calcul du produit des fréquences conditionnelles et appelé « **test d'indépendance** ».

A – Cas des données groupées par valeurs

Lorsque les observations sont fournies groupées par valeurs, on a le choix entre deux procédures :

1) Se ramener au cas des données connues individuellement.

2) Utiliser la même procédure que lorsque les données sont groupées par classes.

Exemple : Soit le tableau de données suivant où les observations sont groupées par valeurs :

Tableau 2

x \ y	**4**	**6**
2	3	1
4	2	5

On peut sans difficulté se ramener au tableau 1 et calculer le coefficient de corrélation comme expliqué dans la section 1.

Tableau 3

X	2	2	2	2	4	4	4	4	4	4	4	**36**
Y	4	4	4	6	4	4	6	6	6	6	6	**56**
XY	8	8	8	12	16	16	24	24	24	24	24	**188**
X^2	4	4	4	4	16	16	16	16	16	16	16	**128**
Y^2	16	16	16	36	16	16	36	36	36	36	36	**296**

On calcule ensuite le coefficient de corrélation :

$$r = \frac{\operatorname{cov}(x,y)}{\sigma_x \sigma_y} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{\overline{xy}}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} x^2 - \left(\overline{\overline{x}}\right)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n} y^2 - \left(\overline{\overline{y}}\right)^2}} = \frac{\frac{188}{11} - \frac{36}{11} \times \frac{56}{11}}{\sqrt{\frac{128}{11} - \left(\frac{36}{11}\right)^2}\sqrt{\frac{296}{11} - \left(\frac{56}{11}\right)^2}} \cong 0,4485$$

B – Cas des données groupées par classes

Lorsque les observations sont fournies groupées par classes, on peut soit calculer un coefficient de corrélation avec une formule modifiée pour tenir compte des effectifs groupés, soit faire un test d'indépendance.

1) Le coefficient de corrélation

La formule du coefficient de corrélation devient :

$$r = \frac{\frac{1}{n_{\bullet\bullet}}\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij} x_i y_j - \overline{\overline{xy}}}{\sqrt{\frac{1}{n_{\bullet\bullet}}\sum_{i=1}^{p} n_{i\bullet} c x_i^2 - \left(\overline{\overline{x}}\right)^2}\sqrt{\frac{1}{n_{\bullet\bullet}}\sum_{j=1}^{q} n_{\bullet j} c y_j^2 - \left(\overline{\overline{y}}\right)^2}}$$

Exemple : Soit le tableau statistique ci-dessous :

Tableau 4

x \ y	[0-3[	[3-9]
[0-4[	2	4
[4-12]	8	3

Pour effectuer les calculs, il est nécessaire de faire un tableau disposé comme ci-après :

Tableau 5

x \ y	(1,5) [0-3[	(6) [3-9]	$n_{i\bullet}$	$n_{i\bullet}cx_i$	$n_{i\bullet}cx_i^2$	$\sum_{j=1}^{q} n_{ij}cy_j$	$\sum_{j=1}^{q} n_{ij}cy_j^2$	$\sum_{j=1}^{q} n_{ij}cx_icy_j$
(2) [0-4[	2 / **3**	4 / **12**	6	12	24	(2x1,5) + (4x6)=27	(2x1,5²) + (4x6²)=148,5	(2x3) + (4x12)=54
(8) [4-12]	8 / **12**	3 / **48**	11	88	704	(8x1,5) + (3x6)=30	(8x1,5²) + (3x6²)=126	(8x12) + (3x48)=240
$n_{\bullet j}$	10	7	17	100	728			294
$n_{\bullet j}cy_j$	15	42	57					
$n_{\bullet j}cy_j^2$	22,5	252	274,5					
$\sum_{i=1}^{p} n_{ij}cx_i$	(2x2) + (8x8)=68	(4x2) + (3x8)=32						
$\sum_{i=1}^{p} n_{ij}cx_i^2$	(2x2²) + (8x8²)=520	(4x2²) + (3x8²)=208						
$\sum_{i=1}^{p} n_{ij}cx_icy_j$	(2x3) + (8x12)=102	(4x12) + (3x48)=192	294					

a) Moyennes marginales

$$\bar{\bar{x}} = \frac{100}{17} = 5{,}88 \qquad \left(\bar{\bar{x}}\right)^2 = 34{,}6 \qquad \sigma_x^2 = \frac{728}{17} - 34{,}6 = 8{,}22$$

$$\bar{\bar{y}} = \frac{57}{17} = 3{,}35 \qquad \left(\bar{\bar{y}}\right)^2 = 11{,}24 \qquad \sigma_y^2 = \frac{274{,}5}{17} - 11{,}24 = 4{,}91$$

b) Variances marginales

c) Covariance

$$\operatorname{cov}(x,y) = \frac{1}{n_{\bullet\bullet}} \sum_{i=1}^{p} \sum_{j=1}^{q} n_{ij} x_i y_j - \bar{\bar{x}}\bar{\bar{y}} = \frac{294}{17} - \left(\frac{100}{17} \times \frac{57}{17}\right) = -2{,}429$$

d) Coefficient de corrélation

$$r = \frac{\operatorname{cov}(x,y)}{\sigma_x \sigma_y} = \frac{-2{,}429}{\sqrt{8{,}22} \times \sqrt{4{,}91}} = \frac{-2{,}429}{2{,}87 \times 2{,}21} = -0{,}38$$

La partie en pointillés du tableau 5, reprise ci-après dans le tableau 6, contient les informations initiales du tableau 4, ainsi que :

Tableau 6

x \ y	1,5 [0-3[	6 [3-9]	
2 [0-4[	**3**	**12**	2 / 4
8 [4-12]	**12**	**48**	8 / 3

1) Les centres de classes qui ont été cerclés.

2) Le produit des centres de classes en gras à l'intersection des lignes et des colonnes.

À noter que le tableau 5 facilite également les calculs des moyennes et des variances conditionnelles (voir les calculs ci-après) :

c) Moyennes conditionnelles

$$\bar{x}_1 = \frac{68}{10} = 6{,}8 \qquad \bar{y}_1 = \frac{27}{6} = 4{,}5$$

$$\bar{x}_2 = \frac{32}{7} = 4{,}57 \qquad \bar{y}_2 = \frac{30}{11} = 2{,}73$$

d) Variances conditionnelles

$$(\bar{x}_1)^2 = 46{,}24 \qquad \sigma^2_{x_1} = \frac{520}{10} - 46{,}24 = 5{,}76$$

$$(\bar{x}_2)^2 = 20{,}89 \qquad \sigma^2_{x_2} = \frac{208}{7} - 20{,}89 = 8{,}82$$

$$(\bar{y}_1)^2 = 20{,}25 \qquad \sigma^2_{y_1} = \frac{148{,}5}{6} - 20{,}25 = 4{,}5$$

$$(\bar{y}_2)^2 = 7{,}44 \qquad \sigma^2_{y_1} = \frac{126}{11} - 7{,}44 = 4{,}01$$

Comme nous l'avons déjà indiqué, lorsque les données sont groupées par valeurs, on peut aussi appliquer la procédure juste décrite pour le cas des données groupées par classe. On obtient alors le même résultat qu'en appliquant la procédure d'identification des données individuelles, mais les calculs sont plus fastidieux.

2) Le test d'indépendance

Deux variables sont indépendantes si et seulement si :

$$n_{ij} = \frac{n_{i\bullet} \times n_{\bullet j}}{n_{\bullet\bullet}}$$

Il suffit donc a contrario qu'un n_{ij} quelconque soit tel que :

$$n_{ij} = \frac{n_{i\bullet} \times n_{\bullet j}}{n_{\bullet\bullet}}$$

Pour que l'on puisse conclure à l'absence d'indépendance. Il est donc généralement plus rapide de vérifier l'absence d'indépendance que d'établir l'indépendance.

Exemple : Soit le tableau statistique ci-dessous :

Tableau 7

x \ y	y_1	y_2	$n_{i\bullet}$
x_1	6	10	**16**
x_2	12	120	**32**
$n_{\bullet j}$	**18**	**30**	**48**

Vérifions que les deux variables X et Y sont totalement indépendantes :

$$6 = \frac{18 \times 16}{48} \qquad 20 = \frac{30 \times 32}{48} \qquad 12 = \frac{18 \times 32}{48} \qquad 10 = \frac{30 \times 16}{48}$$

Remarques :

1) Le test d'indépendance convient bien pour des petits tableaux. Il devient fastidieux pour tableaux supérieurs à 2 x 2.
2) Le test d'indépendance peut être utilisé aussi bien pour des séries quantitatives que pour des séries qualitatives.

3 • SÉRIES QUALITATIVES

A – Le coefficient de corrélation de rang de SPEARMAN

Lorsque les séries sont qualitatives, il arrive que les modalités d'un des deux caractères soient ordinales (voir le chapitre 1), autrement dit que l'on puisse opérer un classement sur ces modalités. Dans ce cas, au lieu de calculer la corrélation entre les valeurs comme on le fait pour une variable, on calcule la corrélation entre les rangs des modalités. On calcule alors un coefficient appelé **coefficient de corrélation de rang de SPEARMAN**.

Voici la formule :

$$r_{sp} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n\left(n^2 - 1\right)}$$

où d_i est la différence entre les rangs des valeurs correspondantes de X et de Y et n le nombre d'observations.

Exemple : ci-dessous, les notes attribuées par deux enseignants à 5 copies.

Tableau 8

	Enseignant 1	Enseignant 2
A	10	11
B	12	15
C	8	6
D	5	7
E	16	14

On veut savoir si le classement qui résulte de la notation de l'enseignant 1 est cohérent avec le classement qui résulte de la notation de l'enseignant 2

On crée alors un tableau où les rangs des notes remplacent les notes. On calcule ensuite la formule de SPEARMAN.

Tableau 9

Rang		**Classement de 1**
1	D	5
2	C	8
3	A	10
4	B	12
5	E	16

Rang		**Classement de 2**
1	C	6
2	D	7
3	A	11
4	E	14
5	B	15

	Enseignant 1	**Enseignant 2**	**di**	**di²**
A	3	3	0	0
B	4	5	-1	1
C	2	1	1	1
D	1	2	-1	1
E	5	4	1	1

$$r_{sp} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n\left(n^2-1\right)} = 1 - \frac{6\times 4}{5\left(25-1\right)} = \frac{4}{5} = 0,8$$

Interprétation : si la corrélation est parfaite, r_{sp}=1. Plus les rangs sont différents, plus r_{sp} tend vers 0.

B – Le test du Khi-carré de PEARSONS

Lorsque les caractères sont qualitatifs l'étude de la corrélation se fait par un test statistique développé par Karl PEARSONS et appelé test d'indépendance du "Khi deux". Pour introduire ce test, considérons l'exemple suivant.

Exemple : 100 consommateurs sont questionnés sur leurs préférences à l'égard de 4 variétés d'un produit (A, B, C et D). On leur demande : "*Parmi ces 4 produits, quel est celui que vous préférez ?*". Ces consommateurs sont groupés en deux catégories, les moins de 20 ans et les plus de 20 ans, afin de déterminer si l'âge a une influence sur la préférence.

Tableau 10

Produits	Moins de 20 ans	Plus de 20 ans	Total
A	10	15	25
B	10	25	35
C	15	5	20
D	20	0	20
Total	55	45	100

Le tableau se lit ainsi : 10 personnes de moins de 20 ans préfèrent le produit A, 15 personnes de plus de 20 ans préfèrent le produit A, 25 en tout préfèrent le produit A.

Si l'âge n'a aucune influence sur le choix, les 2 premières colonnes devraient être proportionnelles à la troisième. On va donc calculer deux colonnes fictives, mais proportionnelles à la troisième, afin d'avoir les effectifs qui correspondent à une indépendance de l'âge sur le choix.

Dans la formule ci-après, la fréquence des plus de 20 ans est 45/100. Celle des moins de 20 ans : est 55/100. N_i est l'effectif théorique correspondant à une répartition homogène. Enfin, n_i est l'effectif observé.

Tableau 11

produits		N_i	n_i	$N_i - n_i$	$(N_i - n_i)^2$	$\frac{(N_i - n_i)^2}{N_i}$
- de 20	A	(55/100) x 25 =13,75	10	3,75	14,0625	1,02272
	B	(55/100) x 35 =19,25	10	9,25	85,5625	4,448
	C	(55/100) x 20 = 11	15	-4	16	1,4545
	D	(55/100) x 20 = 11	20	-9	81	7,3636
+ de 20	A	(45/100) x 25 =11,25	15	-3,75	14,0625	1,25
	B	(45/100) x 35 =15,75	25	-9,25	85,5625	5,4325
	C	(45/100) x 20 = 9	5	4	16	1,7777
	D	(45/100) x 20 = 9	0	9	81	9

Par définition : $\chi^2(\text{calculé}) = \sum_{i=1}^{n} \frac{(N_i - n_i)^2}{N_i}$

En appliquant cette définition aux données du tableau 11, on obtient : $\chi^2(\text{calculé}) = 31{,}74$

Une fois que l'on connaît le khi-carré calculé, on doit le comparer avec la valeur du khi-deux issue de la distribution du khi-carré (voir le tableau 12 ci-dessous). Ici, le nombre de « degrés de liberté » est égal à [8 (nombre d'observations) moins 2 (nombres de variables)], ce qui donne 6. Ensuite, nous devons choisir la probabilité de fiabilité du test : 5% de chances de se tromper (deuxième colonne), 1% (troisième colonne) et 1 pour 1000 (quatrième colonne). Si nous choisissons P = 0,05, nous avons donc :

Tableau 12

Probabilités

degrés de liberté	P = 0.05	P = 0.01	P = 0.001
1	3.84	6.64	10.83
2	5.99	9.21	13.82
3	7.82	11.35	16.27
4	9.49	13.28	18.47
5	11.07	15.09	20.52
6	**12.59**	**16.81**	**22.46**
7	14.07	18.48	24.32
8	15.51	20.09	26.13
9	16.92	21.67	27.88
10	18.31	23.21	29.59
11	19.68	24.73	31.26
12	21.03	26.22	32.91

$$\chi^2_{0,05} = 12,59 < \chi^2_{\text{calculé}} = 31,74$$

Ce qui nous permet de conclure que la répartition des préférences est suffisamment différente d'une répartition homogène pour qu'on puisse raisonnablement se fier à l'idée que l'âge a une influence sur le choix du produit (avec 5% de chances de nous tromper).

PARTIE 3

Les séries chronologiques

CHAPITRE 7

Les séries chronologiques

1 • INTRODUCTION

A – Définition

Une **série chronologique** est une variable statistique dont les observations sont repérées dans le temps.

Les séries chronologiques sont extrêmement utilisées dans les sciences sociales et, en particulier, en économie.

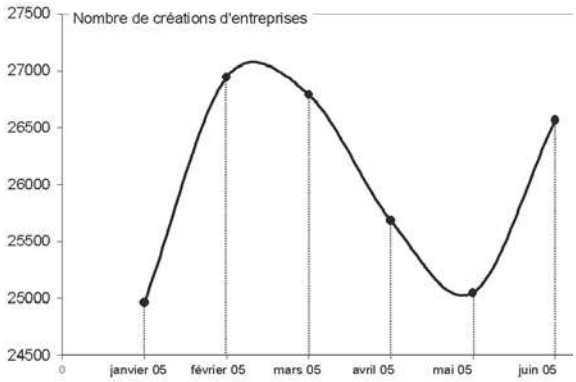
Exemple : Le tableau 1 et le graphique 1 ci-dessous retrace le nombre mensuel de créations d'entreprises en France de janvier à juin 2005.

Tableau 1 : Évolution mensuelle des créations d'entreprises en France

	Nombre de créations d'entreprises
janvier	24966
février	26942
mars	26790
avril	25684
mai	25050
juin	26566

Source : Insee Conjoncture, *Bulletin d'informations rapides*, numéro 2005, 11 juillet 2005.

Figure 1 : Évolution mensuelle des créations d'entreprises en France

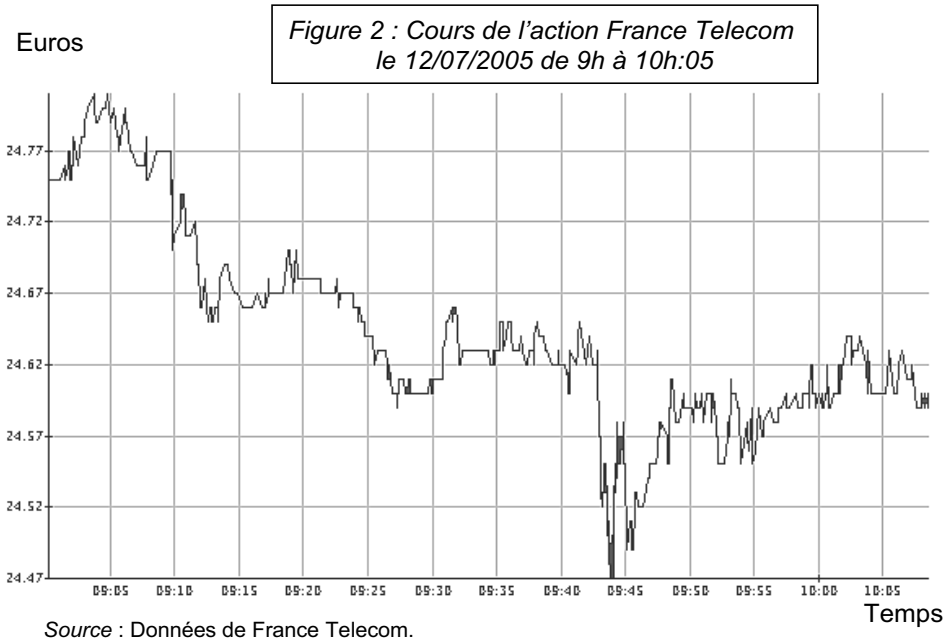


B – Périodicité

Les séries chronologiques peuvent être **annuelles**, **trimestrielles**, **mensuelles**, **hebdomadaires**, **journalières** et même infra-journalières.

Exemple 1 : Le cours d'une action peut être connu heure après heure et même minute après minute, voire de façon instantanée. Le graphique de la figure 2 ci-après retrace ainsi l'évolution du cours de l'action France Telecom, de minute en minute, le 12 juillet 2005, entre 9 h et 10 h.

Figure 2 : Cours de l'action France Telecom le 12/07/2005 de 9h à 10h:05

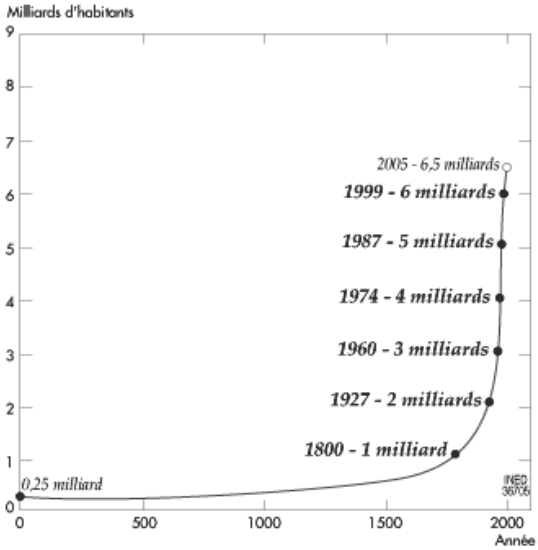


Source : Données de France Telecom.

À l'inverse, certaines données sont disponibles beaucoup plus rarement. On aura alors des observations sporadiques qui permettront de retracer l'évolution sur une longue période, mais avec une périodicité irrégulière.

Exemple 2 : Le graphique de la figure 3 ci-après, extrait d'une étude de l'Institut Nationale d'Études Démographiques (INED), montre l'évolution du nombre des hommes depuis l'an zéro. Un graphique fascinant… Et qui en même temps fait sourire. Il illustre en tous cas notre propos : certaines séries chronologiques n'ont pas une périodicité régulière. Dans ce cas particulier, le graphique présenté a nécessité le concours et l'ingéniosité de centaines de chercheurs en sciences sociales (paléontologues, historiens, statisticiens, etc.). Il reste approximatif mais il est significatif de la volonté insatiable de l'homme de connaître ses origines… Et du rôle indispensable de la statistique descriptive dans cette entreprise.

Figure 3 : Évolution du nombre des hommes depuis l'an 0



Source : François HERAN et Laurent TOULEMON, « La population mondiale… et moi ? » Une exposition à la Cité des sciences et de l'industrie à Paris, INED, *Population et Sociétés*, n° 412, mai 2005

Pour représenter graphiquement les séries chronologiques, on mettra toujours le temps en abscisse et les valeurs de la variable en ordonnée. La représentation la plus habituelle est le nuage de points. Mais il est fréquent que l'on relie les points entre eux. Les exemples des figures 1 à 3 illustrent ce dernier point.

C – Tendance, variations saisonnières et accidentelles

L'observation des séries chronologiques permet de distinguer trois composantes principales. La première de ces composantes, la **tendance** ou **trend**, donne le sens de l'évolution sur la durée. La seconde composante, ce sont les **variations saisonnières ou périodiques**. La troisième composante, ce sont les **variations accidentelles**.

Ces trois composantes ne sont pas toujours simultanément présentes dans une série chronologique. Certaines séries n'ont pas de tendance, d'autres n'ont aucune composante périodique. D'autres enfin, ne connaissent aucune variation accidentelle.

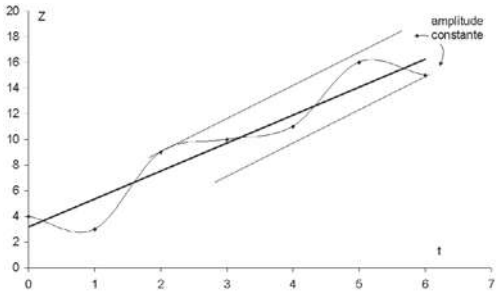
Dans la suite de ce chapitre, nous étudions les méthodes qui permettent d'identifier et de quantifier ces trois composantes.

D – Modèle multiplicatif et modèle additif

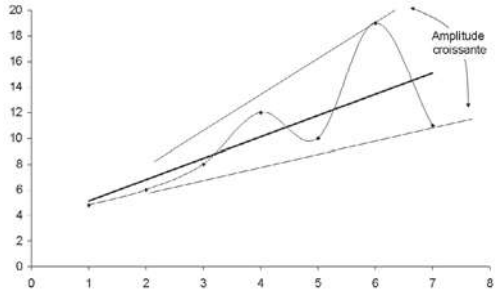
L'observation des séries chronologiques permet de distinguer deux grand types de série : celles qui se conforment au **modèle multiplicatif** et celles qui se conforment au **modèle additif.** Dans le modèle additif, les variations autour du trend demeurent dans une bande de variation à peu près constante (voir la partie (a) de la figure 6). Dans le modèle multiplicatif, au contraire, les variations autour du trend s'amplifient (voir la partie (b) de la figure 6).

Figure 4

(a) Modèle additif



(b) Modèle multiplicatif



Le plus simple pour déterminer le modèle le mieux adapté à une série chronologique particulière est de faire un graphique, d'y ajouter le trend linéaire et d'observer les fluctuations autour du trend. Si ces fluctuations sont régulières, il s'agit d'un modèle additif. Si, au contraire, elles s'amplifient, il s'agit d'un modèle multiplicatif.

Remarque : Dans le cas de données saisonnières (par exemple des données trimestrielles), on peut aussi calculer la moyenne annuelle de la variable et, ensuite, pour chaque trimestre, on retranche de la valeur du trimestre la valeur de la moyenne annuelle et on obtient un écart. Il suffit alors de comparer les écarts. Si les écarts ne cessent d'augmenter avec le temps, on en conclut que le modèle est multiplicatif. Sinon, c'est que le modèle est additif.

2 • DÉTERMINATION DU TREND D'UNE SÉRIE CHRONOLOGIQUE

Le « **trend** », autrement dit la **tendance**, est ce qui, au-delà des **variations saisonnières** ou **accidentelles** d'une série, indique le sens de son évolution. Autrement dit, le trend nous renseigne sur le fait de savoir si la variable augmente, diminue ou reste stable de façon tendancielle.

Pour déterminer le trend ou la tendance d'une série, il y a deux méthodes principales : 1) la **régression linéaire**, où l'on calcule les coefficients a et b d'une droite, qui représentera la tendance, et 2) la méthode des **moyennes mobiles**.

A – La détermination du trend par la régression linéaire

On calcule les coefficients a et b de la droite de régression comme expliqué au chapitre 6.

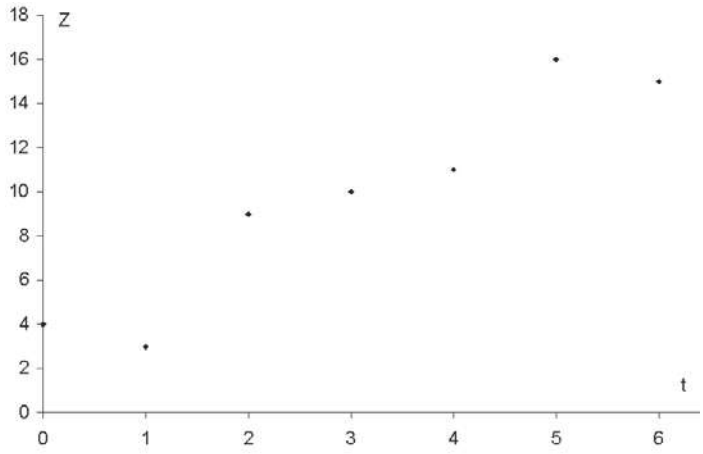
Exemple : Soit le tableau suivant, qui donne l'évolution d'une série chronologique en fonction du temps, repéré par l'indice t.

Tableau 2

t	0	1	2	3	4	5	6
z	4	3	9	10	11	16	15

Le graphique en « nuages de points » de cette série chronologique est illustré par la figure 5.

Figure 5



Nous allons déterminer le trend de cette série par une droite y = ax+b, en calculant les coefficients d'après les formules du chapitre 6, rappelées ci-après (ou t tient le rôle de x et z celui de y).

$$a = \frac{\sum_i t_i z_i - n\bar{\bar{t}}\bar{\bar{z}}}{\sum_i t_i^2 - n\left(\bar{\bar{t}}\right)^2} \qquad b = \bar{\bar{z}} - a\bar{\bar{t}}$$

Tableau 3

t_i	z_i	$z_i t_i$	t_i^2	$z_i t_i^2$
0	4	0	0	0
1	3	3	1	3
2	9	18	4	36
3	10	30	9	90
4	11	44	16	176
5	16	80	25	400
6	15	90	36	540
21	**68**	**265**	**91**	**1245**

$$\sum_{i=1}^{7} t_i z_i = 265 \qquad \bar{\bar{t}} = \frac{\sum_{i=1}^{7} t_i}{7} = \frac{21}{7} = 3 \qquad \bar{\bar{z}} = \frac{\sum_{i=1}^{7} z_i}{7} = \frac{68}{7} = 9{,}714 \qquad \sum_{i=1}^{6} t_i^2 = 91$$

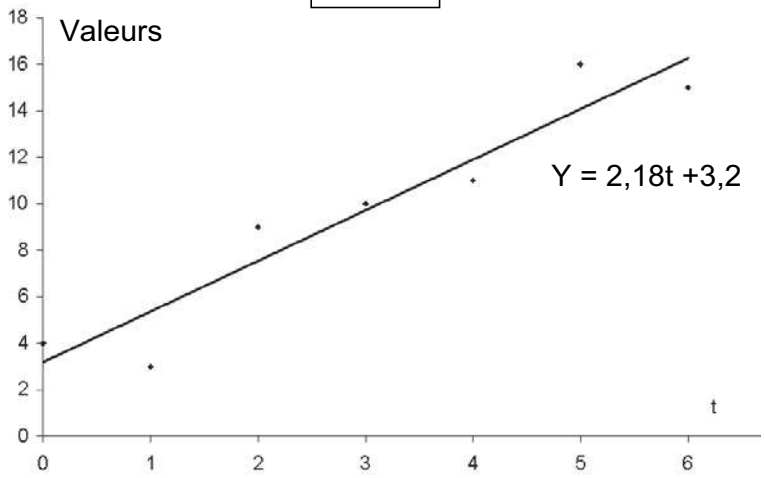
$$a = \frac{\sum_i z_i t_i - n\bar{\bar{t}}\bar{\bar{z}}}{\sum_i t_i^2 - n\left(\bar{\bar{t}}\right)^2} = \frac{265 - 7\times 3\times 9{,}714}{91 - 7\times 3^2} = \frac{61}{28} = 2{,}18 \qquad b = \bar{\bar{z}} - a\bar{\bar{t}} = 9{,}714 - 2{,}17\times 3 = 3{,}2$$

On obtient donc l'équation du trend suivante :

$$f_t = at + b = 2{,}17 + 3{,}2$$

La figure 6 ci-après montre à la fois le nuage de point et la droite de régression qui représente le « trend ».

Figure 6



B – La détermination du trend par la méthode des moyennes mobiles

La méthode des **moyennes mobiles** consiste à calculer la moyenne des valeur qui entourent chaque valeur et à remplacer la valeur par cette moyenne.

Exemple : Soit les données du tab leau 4 qui donne l'évolution du cours de clôture de l'action France Telecom du 13/06/05 au 13/07/05 (en euros).

La troisième colonne donne les moyennes mobiles d'ordre 2 qui sont calculées en prenant les moyennes des cours deux à deux.

À titre d'exemple, les deux premières moyennes mobiles d'ordre 2 s'obtiennent ainsi :

$$\frac{24,66+24,61}{2} = 24,635 \qquad \frac{24,61+24,73}{2} = 24,67$$

Et ainsi de suite pour les autres moyennes mobiles.

Tableau 4 : Cours de clôture de l'action France Telecom

Date	Cours de clôture(€)	Ordre 2	Ordre 3
13/07/2005	24,66		
12/07/2005	24,61	24,635	
11/07/2005	24,73	24,67	24,666
08/07/2005	24,53	24,63	24,623
07/07/2005	24,01	24,27	24,42
06/07/2005	24,16	24,09	24,23
05/07/2005	24	24,08	24,06
04/07/2005	24,18	24,09	24,11
01/07/2005	24,27	24,23	24,15
30/06/2005	24,16	24,22	24,20
29/06/2005	23,8	23,98	24,08
28/06/2005	22,6	23,20	23,52
27/06/2005	22,58	22,59	22,99
24/06/2005	22,66	22,62	22,61
23/06/2005	22,93	22,80	22,72
22/06/2005	22,97	22,95	22,85
21/06/2005	23,02	23,00	22,97
20/06/2005	22,85	22,94	22,95
17/06/2005	22,94	22,90	22,94
16/06/2005	22,68	22,81	22,82
15/06/2005	22,48	22,58	22,70
14/06/2005	22,6	22,54	22,59
13/06/2005	22,74	22,67	22,61

Source : Données de France Telecom.

La troisième colonne donne les moyennes mobiles d'ordre 3 qui sont calculées en prenant les moyennes des cours trois à trois.

À titre d'exemple, les deux premières moyennes mobiles d'ordre 3 s'obtiennent ainsi :

$$\frac{24,66+24,61+24,73}{3}=24,666 \qquad \frac{24,61+24,73+24,53}{3}=24,623$$

Et ainsi de suite pour les autres moyennes mobiles.

Pour avoir le trend mobile, il suffit de reporter sur un graphique les moyennes obtenues. La figure 7(a) représente la série initiale et le trend obtenu à l'aide de la méthode des moyennes mobiles d'ordre 2. La figure 7(b) représente la série initiale et le trend obtenu à l'aide de la méthode des moyennes mobiles d'ordre 2.

Figure 7 (a) : Moyenne mobile d'ordre 2

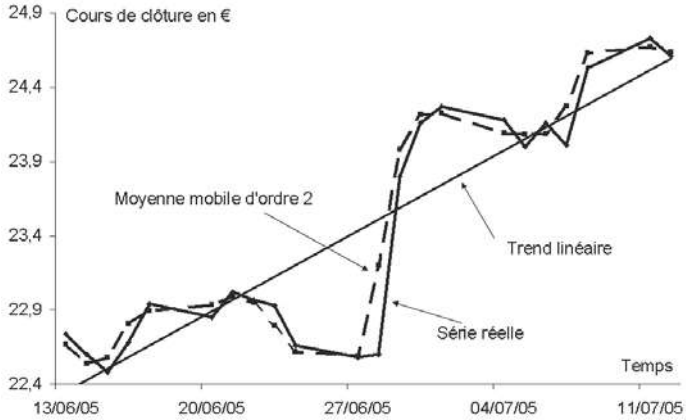
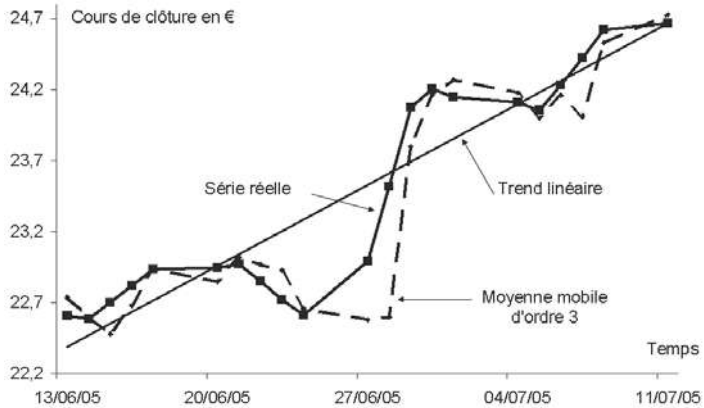


Figure 7 (b) : Moyenne mobile d'ordre 3



Lorsque la méthode de détermination par le trend linéaire apparaît trop grossière, ou lorsque par exemple il n'y a pas de raison de penser qu'il existe une composante saisonnière et qu'on veut juste gommer les variations accidentelles, alors la méthode des moyennes mobiles peut être un bon moyen d'obtenir une série ajustée ou une série lissée comme on dit parfois. D'autant plus que la méthode est facile d'emploi et disponible dans les fonctions des logiciels comme EXCEL.

Le plus simple, lorsque l'on fait les calculs avec un tableur, est de déterminer le trend par les deux méthodes.

À noter que plus la série est longue, plus on peut augmenter l'ordre de calcul des moyennes.

3 • LES VARIATIONS SAISONNIÈRES

A – Vocabulaire

Beaucoup de phénomènes, en particulier les phénomènes économiques, ont une **composante saisonnière**. Certains produits se vendent mieux l'été que l'hiver, d'autres se vendent mieux aux périodes de vacances scolaires. L'appellation de variation saisonnière ne signifie pas pour autant que la composante saisonnière se répartisse sur l'année, même si c'est souvent le cas. Il y a aussi des récurrences de type saisonnier à l'intérieur d'un mois, d'une semaine, voire d'un jour. Certains produits se vendent mieux certains jours et à certaines heures…

On est ainsi amené à calculer une composante saisonnière, puis un **coefficient saisonnier**, afin de déterminer la **série corrigée des variations saisonnière** ou **série CVS**. L'intérêt de ce calcul est d'obtenir une série chronologique dont l'évolution est débarrassée de la composante saisonnière qui parfois masque la tendance. Dans le cas souvent cité du chômage, par exemple, on peut avoir l'impression d'une augmentation ou d'une diminution tendancielle du chômage alors qu'il y a seulement des embauches ou des mises à pied qui ont lieu chaque année à la même période et avec la même ampleur.

On parle ainsi de « **désaisonnalisation** du taux de chômage », laquelle atténue les variations dues aux embauches pendant l'été et aux mises à pied pendant l'hiver dans des secteurs d'activité comme l'agriculture et la construction.

Pour obtenir une série corrigée des variations saisonnières, ou série CVS, on procède en trois étapes : (1) on calcule la composante saisonnière, (2) on en déduit le coefficient saisonnier et (3) on retranche le coefficient saisonnier de la série originale.

Dans l'exemple qui suit, nous supposerons que la série suit un **modèle additif,** l'application au cas multiplicatif étant légèrement différente (voir le livre de Bernard PY, *Statistique descriptive*, mentionné en bibliographie, pour l'étude du cas multiplicatif).

B – Les étapes du calcul de la série CVS

Ci-après, les étapes du calcul de la série CVS sont détaillées, puis appliquées à un exemple concret :

1) Détermination de l'équation du trend par régression linéaire.

2) Calcul des coefficients saisonniers.

3) Détermination de la série CVS.

Exemple : Soit le tableau suivant, qui donne l'évolution d'une série chronologique trimestrielle.

Tableau 5

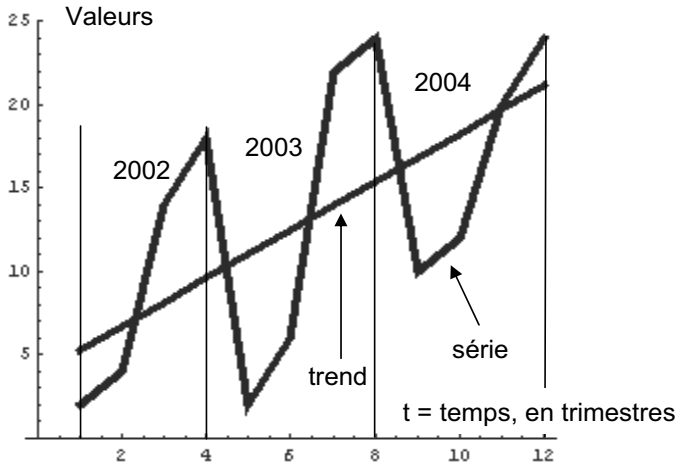
2002				2003				2004			
T1	**T2**	**T3**	**T4**	**T1**	**T2**	**T3**	**T4**	**T1**	**T2**	**T3**	**T4**
2	4	14	18	2	6	22	24	10	12	20	24

Le graphique de la figure 6, qui montre la série et son trend (pour le calcul de l'équation du trend, voir ci-après), révèle deux caractéristiques, qu'il nous est nécessaire de vérifier pour employer la méthode proposée :

1) D'une part, la série étudiée suit un **modèle additif**. En effet, les variations autour du trend ne semblent pas s'amplifier avec le temps.

2) D'autre part, il existe bien une **composante saisonnière**, ici trimestrielle, qui se superpose à une tendance à la hausse. On note en effet qu'à l'intérieur de chacun des trois cycles annuels, la variable débute à un niveau faible au premier trimestre, puis augmente à chaque trimestre pour atteindre un maximum au dernier trimestre, avant de repartir à la baisse au début de l'année suivante.

Figure 7



1) Détermination de l'équation du trend

Les calculs intermédiaires sont aisément effectués à l'aide du tableau 6 ci-après.

$$\sum_{i=1}^{12} t_i y_i = 1234 \qquad \bar{\bar{y}} = \frac{\sum_{i=1}^{12} y_i}{12} = \frac{158}{12} = 13{,}1667$$

$$\bar{\bar{t}} = \frac{\sum_{i=1}^{12} t_i}{12} = \frac{79}{12} = 6{,}583 \qquad \sum_{i=1}^{12} t_i^2 = 650$$

$$a = \frac{\sum_i y_i t_i - n\bar{\bar{t}}\bar{\bar{y}}}{\sum_i t_i^2 - n\left(\bar{\bar{t}}\right)^2} \cong 1{,}44755 \qquad b = \bar{\bar{y}} - a\bar{\bar{t}} = 3{,}757$$

On obtient donc l'équation du trend suivante :

$$f_i = at_i + b = 1{,}44755t_i + 3{,}75757576$$

Tableau 6

ti	y_i	y_it_i	t_i^2	$y_it_i^2$
1	2	2	1	2
2	4	8	4	16
3	14	42	9	126
4	18	72	16	288
5	2	10	25	50
6	6	36	36	216
7	22	154	49	1078
8	24	192	64	1536
9	10	90	81	810
10	12	120	100	1200
11	20	220	121	2420
12	24	288	144	3456
79	**158**	**1234**	**650**	**11198**

2) Calcul des coefficients saisonniers

Pour calculer les **coefficients saisonniers**, il faut d'abord **isoler la composante saisonnière** de la série. Pour ce faire, il convient de calculer les valeurs tendancielles, soit f_i pour i =1 à 12, grâce à l'équation du trend, puis de retrancher f_i de y_i.

Par exemple, quand i = 1, on a :

$$f_1 = 1,44755245 \times 1 + 3,75757576 = 5,20512821$$

La composante saisonnière quand t=1 est donc :

$$S_1 = y_1 - f_1 = 2\text{-}5,20512821 = -3,205128205$$

En réitérant le calcul pour les 12 valeurs, on obtient le tableau 7 :

Tableau 7

t_i	y_i	f_i	$S_i = y_i - f_i$
1	2	5,205128205	-3,205128205
2	4	6,652680653	-2,652680653
3	14	8,1002331	5,8997669
4	18	9,547785548	8,452214452
5	2	10,995338	-8,995337995
6	6	12,44289044	-6,442890443
7	22	13,89044289	8,10955711
8	24	15,33799534	8,662004662
9	10	16,78554779	-6,785547786
10	12	18,23310023	-6,233100233
11	20	19,68065268	0,319347319
12	24	21,12820513	2,871794872

Les 4 coefficients saisonniers s'obtiennent en faisant la moyenne arithmétique des composantes saisonnières (dernière colonne du tableau, Si) pour 2002, 2003 et 2004. On obtient :

C1 = (1/3)(S1+S5+S9) = -3,205128205+ -8,995337995+-6,785547786 = - 6,328671329

C2 = (1/3)(-2,652680653+-6,442890443+ -6,233100233) = - 5,10955711

C3 = (1/3)(5,8997669+8,10955711+2,871794872) = 4,776223776

C4 = (1/3)(8,452214452 +8,662004662+2,871794872) = 6,662004662

On remarquera que la somme C1+C2+C3+C4 est pratiquement égale à zéro. Dans le cas contraire, il faudrait appliquer un **coefficient correcteur** à chaque coefficient saisonnier. La formule de ce coefficient correcteur est :

$$\rho = \frac{1}{4}\sum_{j=1}^{4} C_j$$

On obtient donc un coefficient saisonnier corrigé, C_i' :

$$C_i' = C_i - \rho$$

3) Détermination de la série CVS

La série corrigée des variations saisonnières, dite « série CVS » s'obtient en retranchant les coefficients saisonniers du trend. Désignons par y_i^* la série CVS :

$$y_i^* = y_i - C_i'$$

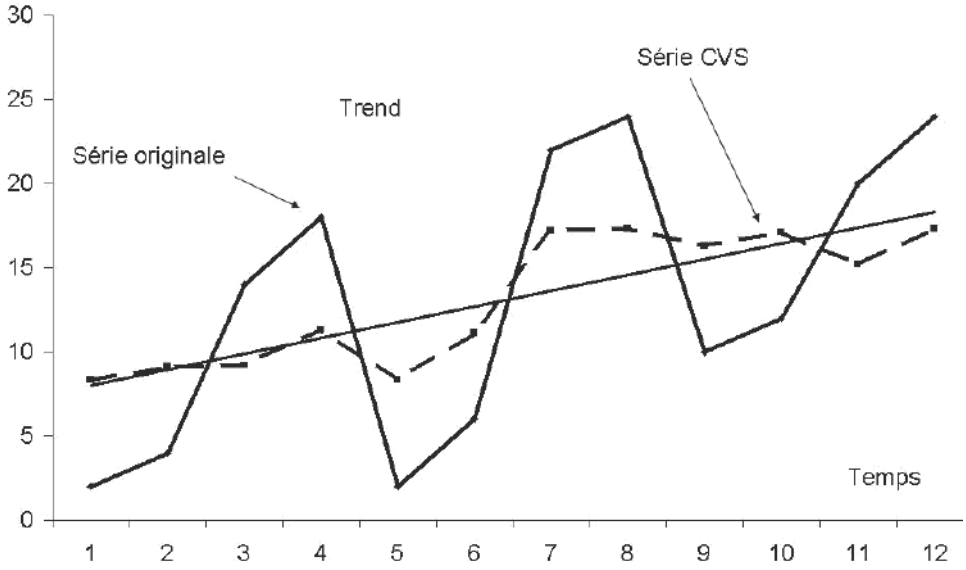
où C_i' représente le coefficient saisonnier, éventuellement corrigé (ici cela n'a pas été nécessaire). La dernière colonne du tableau 8 ci-après donne la série CVS.

Le graphique illustré par la figure 9 fait apparaître que la série CVS épouse davantage le trend que la série originale. C'est normal puisque l'on a effacé les variations saisonnières.

Tableau 8

t_i	y_i	$y_i^* = y_i - C_i'$
1	2	8,328671329
2	4	9,10955711
3	14	9,223776224
4	18	11,33799534
5	2	8,328671329
6	6	11,10955711
7	22	17,22377622
8	24	17,33799534
9	10	16,32867133
10	12	17,10955711
11	20	15,22377622
12	24	17,33799534

Figure 9



On notera néanmoins que la méthode est loin d'être parfaite. En effet, les variations saisonnière sont atténuées mais non supprimées. Cela vient du fait que la méthode ne permet pas de décomposer très finement les variations saisonnières et les variations accidentelles que nous allons étudier maintenant.

4 • LES VARIATIONS ACCIDENTELLES

Les **variations accidentelles** sont ce qui reste lorsqu'on a enlevé le trend de la série ajustée des variations saisonnières. Comme on vient de le voir, la décomposition entre les variations accidentelles et les variations saisonnières est loin d'être parfaite.

$$\varepsilon_i = y_i^* - f_i$$

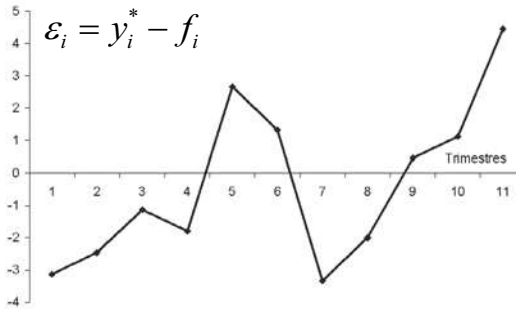
Exemple : Reprenons les données de l'exemple précédent et calculons la série des variations accidentelles en appliquant la formule.

On obtient alors le tableau 9 et la figure 10 ci-après :

tableau 9

Temps (t_i)	$\varepsilon_i = y_i^* - f_i$
1	-3,123543124
2	-2,456876457
3	-1,123543124
4	-1,79020979
5	2,666666667
6	1,333333333
7	-3,333333333
8	-2
9	0,456876457
10	1,123543124
11	4,456876457
12	3,79020979

Figure 10



La somme des 12 éléments de cette série donne un nombre pratiquement égal à zéro. Cela signifie qu'il y a **conservation des aires** (c'est-à-dire que les hausses sont compensées par les baisses). On peut d'ailleurs le vérifier sur le graphique.

PARTIE 4

Les indices

CHAPITRE 8

Les indices

1 • INTRODUCTION

A – Définition et exemples

Un **indice** est une mesure de la variation d'une grandeur comparée à une valeur de référence égale à 100 et appelée « **base** ».

Exemple 1 : Selon l'INSEE, **l'indice des prix à la consommation** de la France est égal à 112,5 en 2005 (base 100 en 1998).

L'avantage de cette formulation est de permettre une lecture immédiate de la variation des prix entre 1998 et 2005 : entre ces deux dates, les prix ont augmenté de 12,5%.

Remarque : certains indices ne sont pas exprimés par rapport à une base 100, mais par rapport à une base 1.

Exemple 2 : L'indice de trafic routier en Ile-de-France, dit « indice SIER », (Service Interdépartemental d'Exploitation Routière) est égal à 1 quand le trafic est fluide, c'est-à-dire quand il faut en moyenne 1 minute pour faire 1 km. Si l'indice est égal à 2, cela signifie que les temps de parcours sur le réseau sont 2 fois plus longs que si le trafic est fluide. S'il est égal à 3, ils sont 3 fois plus longs et ainsi de suite (*Source* : www.sytadin.equipement.gouv.fr).

Une **série indice** est une série divisée par une de ses valeurs et éventuellement multipliée par 100.

Exemple 3 : Soit la série :

{1, 3, 7 ,4 ,8 ,6 , 11, 9}

Supposons que l'on divise tous les éléments de la série par son troisième élément et que l'on multiplie par 100. La nouvelle série est une série indice, la base est le troisième élément de la série :

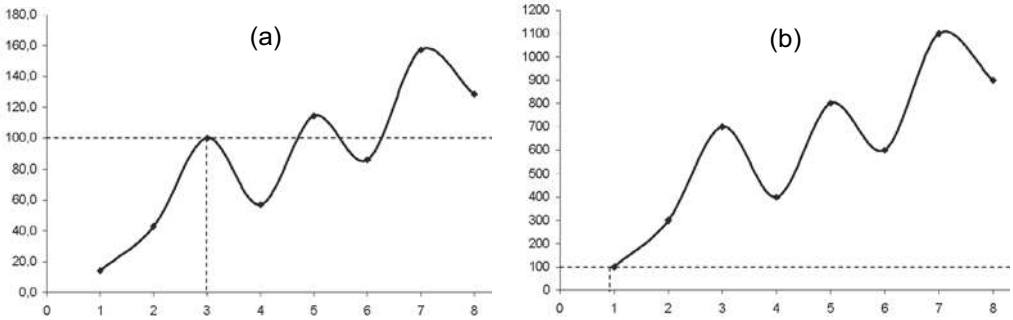
{14,3 ; 42,9 ; 100 ; 57,1 ; 114,3 ; 85,7 ; 157,1 ; 128,6}

On peut effectuer un changement de base en divisant la série par la premier chiffre de la série plutôt que par le troisième :

{100, 300, 700, 400, 800, 600, 1100, 900}

Le graphique (a) de la figure 1 illustre la série indice quand la base est le troisième chiffre et le graphique (b) illustre la série indice quand la base est le premier chiffre.

Figure 1 : Représentation graphique d'une série indice



On remarquera que le changement de base n'a pas d'incidence sur la forme de la courbe, mais seulement sur l'échelle de l'ordonnée.

B – Indice temporel et indice de situation

Un **indice temporel** est un indice qui concerne une comparaison de valeurs dans le temps. La base est dans ce cas la date de référence.

Exemple 1 : Le 15/07/2005, l'action CNP Assurances (ISIN FR0000120222) a coté 54,10 euros en ouverture et 54 euros en fermeture. L'indice de variation du cours de l'action sur la séance, donné par (54,1/54)*100 = 100,185, est un indice temporel, la base étant l'heure de l'ouverture de la séance du 15/07/2005.

Un **indice de situation**, également appelé indice spatial, est un indice qui concerne n'importe quelle comparaison de valeurs, hormis les comparaison temporelles.

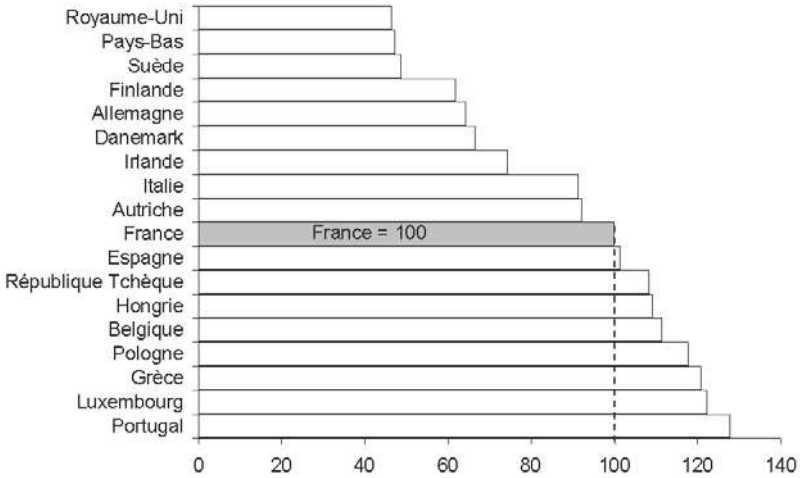
Exemple 2 : En 2002, le nombre de victimes d'accidents de la route en France a été de 129 par million d'habitants, alors qu'au Portugal il a été de 165 par million d'habitants. L'indice de situation du nombre de victimes d'accidents est égal à (165/129)*100 = 127,9, si l'on prend le nombre d'accidents en France comme base.

Bien entendu, les notions d'indices temporel et de situation peuvent s'étendre à toute une série. Le tableau 1 et la figure 2 ci-après illustrent la série indice de situation du nombre de victimes d'accidents de la route en Europe en 2002, en prenant le nombre de victimes en France comme base.

Tableau 1 : Série indice du nombre de victimes d'accidents de la route en 2002 (France=100)

Pays européens	**Indice (France =100)**
Portugal	128
Luxembourg	122
Grèce	121
Pologne	118
Belgique	112
Hongrie	109
République Tchèque	109
Espagne	102
France	**100**
Autriche	92
Italie	91
Irlande	74
Danemark	67
Allemagne	64
Finlande	62
Suède	49
Pays-Bas	47
Royaume-Uni	47

Figure 2 : Série indice du nombre de victimes d'accidents de la route en 2002 (France=100)



Source : Insee, Tableaux de l'Économie Française 2004-05, page 65.

C – Indice élémentaire et indice synthétique

Un **indice élémentaire** est un indice qui renseigne sur l'évolution temporelle ou situationnelle (spatiale) d'une seule valeur. Il a pour définition :

$$I_{t/0} = \frac{V_t}{V_0} \times 100$$

Où V_0 représente la valeur de référence et V_t la valeur qui est comparée à la valeur initiale.

Dans le cas d'un indice temporel, « 0 » représente la période référence (la base) et « t » la période que l'on compare à la période de référence.

Dans le cas d'un indice de situation ou indice spatial, « 0 » représente la situation de référence (la base) et «t » la situation que l'on compare à la situation de référence.

Exemple 1 : le « Ph », ou potentiel hydrogène de l'eau d'une piscine a été mesuré à 8 h du matin. La mesure révèle qu'il est égal à sa valeur de neutralité (soit 7 sur une échelle qui varie de 1 à 14). Le soir à 18 h, on mesure à nouveau le Ph et cette valeur est alors de 5. L'indice élémentaire de la variation du Ph entre 8 h et 18 h est donné par :

$$I_{18h/8h} = \frac{5}{7} \times 100 = 71{,}43$$

Un **indice synthétique** est un indice qui résume l'évolution de plusieurs valeurs ou qui mesure l'évolution de valeurs liées par un produit ou un rapport.

Exemple 2 : Le prix d'un bien x est égal à 1,5 euro à la date 0. À la date t, il est égal à 2,3 euros. Le prix d'un bien y est égal à 2 euros à la date 0 et à 1,8 euro à la date t. Nous pouvons calculer les indices élémentaires d'évolution des prix du bien x et du bien y. Mais nous pouvons aussi calculer l'indice synthétique d'évolution du prix des deux biens. Pour calculer cet indice synthétique, nous allons faire une moyenne. Cette moyenne peut être une moyenne arithmétique ou non. De plus, nous pouvons choisir de pondérer chaque bien par ½ (moyenne arithmétique simple) ou par des coefficients α_x et α_y différents de ½ mais tels que $\alpha_x + \alpha_y = 1$.

Les indices élémentaires de l'évolution des prix des biens x et y sont donnés par :

$$Ix_{t/o} = \frac{2,3}{1,5} \times 100 = 153,3 \qquad Iy_{t/o} = \frac{1,8}{2} \times 100 = 90$$

L'indice synthétique le plus simple de l'évolution du prix de ces deux biens est une moyenne pondérée, soit :

$$I_{t/o} = \alpha_x Ix_{t/0} + \alpha_y Iy_{t/0}$$

Si l'on prend α_x =1/2 et α_y =1/2 on obtient :

$$I_{t/o} = \frac{Ix_{t/0} + Iy_{t/0}}{2} = \frac{153,3 + 90}{2} = 121,65$$

Soit une évolution de l'indice synthétique égale à +21,65%.

Si l'on prend α_x =1/4 et α_y =3/4 on obtient :

$$I_{t/o} = \frac{1}{4} Ix_{t/0} + \frac{3}{4} Iy_{t/0} = \frac{1}{4} \times 153,3 + \frac{3}{4} \times 90 = 38,325 + 67,5 = 105,8$$

Soit une évolution de l'indice synthétique égale à 105,8 -100= + 5,8 %

Si l'on prend α_x =3/4 et α_y =1/4 on obtient :

$$I_{t/o} = \frac{3}{4} Ix_{t/0} + \frac{1}{4} Iy_{t/0} = \frac{3}{4} \times 153,3 + \frac{1}{4} \times 90 = 114,975 + 22,5 = 137,475$$

Soit une évolution de l'indice synthétique égale à +37,475.

2 • LES INDICES SYNTHÉTIQUES DE LASPEYRES, PAASCHE ET FISHER

Les indices synthétiques les plus utilisés en économie sont les indices qui résument l'évolution de la valeur d'un panier de produits. Trois économistes, LASPEYRES, PAASCHE et FISHER, ont proposé des indices synthétiques différents pour mesurer l'évolution de cette valeur.

A – Définition de la valeur d'un panier de biens

Comment mesurer l'évolution d'une variable synthétique, la **valeur d'un panier de produits**, sachant que la valeur de chaque produit est elle-même le produit d'un prix par une quantité ? Pour clarifier cette question, posons quelques définitions.

Soit $V_t^i = p_t^i\, q_t^i$ la valeur du bien *i*, à la date i où p_t^i représente le prix du bien *i* à la date t et q_t^i sa quantité. Par exemple, si p_t^i = 2 euros et que q_t^i =4 unités, on aura :

$$V_t^i = p_t^i q_t^i = 2 \times 4 = 8 \text{ euros}$$

Maintenant, s'il y a *n* produits dans le panier (i= 1, n), la valeur totale du panier à la date t s'écrira :

$$V_i = \sum_{i=1}^{n} p_t^i q_t^i \qquad (1)$$

L'évolution de la valeur du panier entre deux dates dépend de l'évolution du prix de chaque bien et de l'évolution de la quantité de chaque bien. Il faut donc construire un indice synthétique qui permette d'imputer l'évolution de la valeur du panier à la composante prix ou à la composante quantité. Plusieurs indices peuvent être envisagés.

Nous étudierons successivement les indices proposés par LASPERES, PAASCHE et FISHER. Dans chaque cas, nous définirons l'indice et nous illustrerons son mode de calcul par un exemple.

B – Les indices de LASPEYRES

L'économiste allemand Ernst Louis Etienne LASPEYRES (1834-1913) a proposé de calculer deux indices synthétiques qui portent son nom : l'indice de LASPEYRES des prix et l'indice de LASPEYRES des quantités.

1) L'indice de LASPEYRES des prix

L'indice de LASPEYRES des prix mesure l'évolution entre deux dates 0 et t, des prix des biens qui composent un panier, en prenant comme référence la valeur du panier à la date initiale (t = 0) et en supposant que les quantités de biens dans le panier n'ont pas varié entre 0 et t.

L'indice de LASPEYRES des prix se définit comme suit :

$$L_{t/0}^{P} = \frac{\sum_{i=1}^{n} p_t^i q_0^i}{\sum_{i=1}^{n} p_0^i q_0^i} \times 100$$

On voit ainsi que si les prix ne changent pas entre 0 et t (c'est-à-dire si $p_t^i = p_0^i$), l'indice synthétique de LASPEYRES des prix demeure égal à 100.

Exemple : Soit le tableau 2, qui donne les prix et les quantités de deux produits 1 et 2, aux périodes 0 et t.

Tableau 2

	Période 0		Période t	
Produit 1	$p_0^1=10$	$q_0^1=4$	$p_t^1=14$	$q_t^1=8$
Produit 2	$p_0^2=6$	$q_0^2=12$	$p_t^2=5$	$q_t^2=9$

Calculons l'indice de LASPEYRES des prix :

$$L_{t/0}^{P} = \frac{\sum_{i=1}^{n} p_t^i q_0^i}{\sum_{i=1}^{n} p_0^i q_0^i} \times 100 = \frac{p_t^1 q_0^1 + p_t^2 q_0^2}{p_0^1 q_0^1 + p_0^2 q_0^2} = \frac{(14 \times 4) + (5 \times 12)}{(10 \times 4) + (6 \times 12)} = 103{,}57$$

Dans notre exemple, le prix du bien 1 a augmenté (de 10 à 14) et le prix du bien 2 a baissé. L'indice, qui synthétise ces deux variations contraires, nous permet de conclure à une « inflation », c'est-à-dire une augmentation du niveau général des prix égale à 3,57%.

2) L'indice de LASPEYRES des quantités

L'indice de LASPEYRES des quantités mesure l'évolution entre deux dates 0 et t, des quantités des biens qui composent un panier, en prenant comme référence la valeur du panier à la date initiale (t=0) et en supposant que les prix des biens dans le panier n'ont pas varié entre 0 et t.

On a donc la formule suivante de l'indice de LASPEYRES des quantités :

$$L^{Q}_{t/0} = \frac{\sum_{i=1}^{n} p_0^i q_t^i}{\sum_{i=1}^{n} p_0^i q_0^i} \times 100$$

On voit ainsi que si les quantités ne changent pas entre 0 et t (c'est-à-dire si $q_t^i = q_0^i$, l'indice synthétique de LASPEYRES des quantités demeure égal à 100.

Exemple : Soit le tableau 2, qui donne les prix et les quantités de deux produits 1 et 2, aux périodes 0 et t.

Calculons l'indice de LASPEYRES des quantités :

$$L^{P}_{t/0} = \frac{\sum_{i=1}^{n} p_0^i q_t^i}{\sum_{i=1}^{n} p_0^i q_0^i} \times 100 = \frac{p_0^1 q_t^1 + p_0^2 q_t^2}{p_0^1 q_0^1 + p_0^2 q_0^2} = \frac{(10 \times 8) + (6 \times 9)}{(10 \times 4) + (6 \times 12)} = 119{,}64$$

Dans notre exemple, la quantité du bien 1 a augmenté (de 4 à 8) et la quantité du bien 2 a baissé. L'indice, qui synthétise ces deux variations contraires, nous permet de conclure à une augmentation des volumes égale à 19,64%.

C – Les indices de PAASCHE

L'économiste allemand Hermann PAASCHE (1851-1925) a proposé de calculer deux indices synthétiques qui portent son nom : l'indice de PAASCHE des prix et l'indice de PAASCHE des quantités.

1) L'indice de PAASCHE des prix

L'indice de PAASCHE des prix mesure l'évolution entre deux dates 0 et t, des prix des biens qui composent un panier, en prenant comme référence la valeur du panier à la date terminale (t) et en supposant que les quantités de biens dans le panier n'ont pas varié entre 0 et t.

On a donc la formule suivante de l'indice de PAASCHE des prix :

$$P_{t/0}^{P} = \frac{\sum_{i=1}^{n} p_t^i q_t^i}{\sum_{i=1}^{n} p_0^i q_t^i} \times 100$$

Exemple : Soit le tableau 2, qui donne les prix et les quantités de deux produits 1 et 2, aux périodes 0 et t.

Calculons l'indice de PAASCHE des prix :

$$P_{t/0}^{P} = \frac{\sum_{i=1}^{n} p_t^i q_t^i}{\sum_{i=1}^{n} p_0^i q_t^i} \times 100 = \frac{p_t^1 q_t^1 + p_t^2 q_t^2}{p_0^1 q_t^1 + p_0^2 q_t^2} = \frac{(14 \times 8) + (5 \times 9)}{(10 \times 8) + (6 \times 9)} = 117{,}16$$

Dans notre exemple, le prix du bien 1 a augmenté (de 10 à 14) et le prix du bien 2 a baissé. L'indice, qui synthétise ces deux variations contraires, nous permet de conclure à une « inflation », c'est-à-dire une augmentation du niveau général des prix égale à 17,6% (contre 3,57% quand on utilise la formule de LASPEYRES).

2) L'indice de PAASCHE des quantités

L'indice de PAASCHE des quantités mesure l'évolution entre deux dates 0 et t, des quantités des biens qui composent un panier, en prenant comme référence la valeur du panier à la date terminale (t) et en supposant que les prix des biens dans le panier n'ont pas varié entre 0 et t.

On a donc la formule suivante de l'indice de PAASCHE des quantités :

$$P_{t/0}^{Q} = \frac{\sum_{i=1}^{n} p_t^i q_t^i}{\sum_{i=1}^{n} p_t^i q_0^i} \times 100$$

Exemple : Soit le tableau 2, qui donne les prix et les quantités de deux produits 1 et 2, aux périodes 0 et t.

Calculons l'indice de PAASCHE des quantités :

$$P_{t/0}^{Q} = \frac{\sum_{i=1}^{n} p_t^i q_t^i}{\sum_{i=1}^{n} p_t^i q_0^i} \times 100 = \frac{p_t^1 q_t^1 + p_t^2 q_t^2}{p_t^1 q_0^1 + p_t^2 q_0^2} = \frac{(14 \times 8) + (5 \times 9)}{(14 \times 4) + (5 \times 12)} = 135{,}34$$

Dans notre exemple, la quantité du bien 1 a augmenté (de 4 à 8) et la quantité du bien 2 a baissé. L'indice, qui synthétise ces deux variations contraires, nous permet de conclure à une augmentation des volumes égale à 35,34% (contre 19,64% quand on utilise la formule de LASPEYRES).

D – Les indices de FISHER

L'économiste américain Irving FISHER (1867-1947) a proposé de calculer deux indices synthétiques qui portent son nom : l'indice de FISHER des prix et l'indice de FISHER des quantités. En fait, chacun de ces deux indices est une moyenne géométrique des indices de LASPEYRES et de PAASCHE correspondant.

1) L'indice de FISHER des prix

L'indice de FISHER des prix est la moyenne géométrique des indices de prix de LASPEYRES et de PAASCHE

On a donc la formule suivante de l'indice de FISHER des prix :

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

Exemple : Soit le tableau 2, qui donne les prix et les quantités de deux produits 1 et 2, aux périodes 0 et t.

Calculons l'indice de FISHER des prix :

$$F_{t/0}^{P} = \sqrt{L_{t/0}^{P} \times P_{t/0}^{P}} = \sqrt{103{,}57 \times 117{,}16} = 110{,}16$$

2) L'indice de FISHER des quantités

L'indice de FISHER des quantités est la moyenne géométrique des quantités de prix de LASPEYRES et de PAASCHE.

On a donc la formule suivante de l'indice de FISHER des prix :

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

Exemple : Soit le tableau 2, qui donne les prix et les quantités de deux produits 1 et 2, aux périodes 0 et t.

Calculons l'indice de FISHER des quantités :

$$F_{t/0}^{Q} = \sqrt{L_{t/0}^{Q} \times P_{t/0}^{Q}} = \sqrt{119,64 \times 135,34} = 127,39$$

3 • L'INDICE DES PRIX À LA CONSOMMATION DE L'INSEE

L'un des indices synthétiques les plus connus et les plus utilisés est l'indice des prix à la consommation (IPC) publié chaque mois par l'INSEE. L'IPC permet de mesurer l'inflation, c'est-à-dire la variation du niveau général des prix des biens et des services consommés par les ménages sur le territoire français entre deux périodes données. **C'est une mesure synthétique des évolutions de prix à qualité constante**.

Pour le calculer, l'INSEE applique la formule de l'indice de LASPEYRES des prix à un échantillon de quelques 21 000 indices élémentaires. Ces 21 000 indices élémentaires sont calculés à partir de prix recueillis dans 106 agglomérations de plus de 2 000 habitants réparties sur tout le territoire. L'indice couvre plus de 1 000 variétés de produits, regroupées en 161 groupes. Pour éviter toute tentative de manipulation des prix, la liste précise de ces 1 000 variétés de produits reste confidentielle. Actuellement, la période de référence, ou « base » de l'IPC, est 1998.

L'IPC est publié aux environs du 13 de chaque mois et porte sur l'évolution des prix du mois précédent. Ce chiffre, régulièrement relayé par les médias, est très attendu car il sert de multiples fonctions économiques parmi lesquelles la connaissance de l'inflation, la définition des objectifs de la politique monétaire, mais aussi le versement de pensions et de divers revenus, tels le SMIC, dont le montant est « indexé » sur l'évolution de l'IPC.

Glossaire des formules

Les formules sont classées par leur ordre d'apparition dans le Mémento

Fréquence relative : Elle est égale à la fréquence absolue divisée par l'effectif total :

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

Taux de croissance : Soit g = taux de croissance, V_0 = valeur de départ et V_t =valeur d'arrivée. On a :

$$g = \frac{V_t}{V_0} - 1 = \frac{V_t - V_0}{V_0}$$

Évolutions successives : Soit g_1, g_2,…, g_t des taux de croissance successifs. Le taux de croissance global sur la période 1,…,t est :

$$g = (1+g_1)(1+g_2)....(1+g_n) - 1$$

Taux de croissance moyen : Soit g_1, g_2, …, g_t des taux de croissance successifs. Le taux de croissance moyen sur la période 1, …, t est :

$$\overline{g} = \sqrt[t]{(1+g)} - 1$$

Taux de croissance d'un produit : Soi t deux grandeurs à la date t :

$$V_t = (1+g_v)V_0 \qquad U_t = (1+g_u)U_0$$

La grandeur qui représente leur produit est :

$$W_t = V_t \times U_t = (1+g_v)(1+g_u)W_0$$

Et son taux de croissance est :

$$g_W = \frac{W_t}{W_0} - 1 = (1+g_v)(1+g_u) - 1$$

Taux de croissance d'un rapport : Soit deux grandeurs à la date t :

$$V_t = (1 + g_v)V_0$$

La grandeur qui représente leur rapport est :

$$Z_t = \frac{V_t}{U_t} = \frac{(1 + g_v)}{(1 + g_u)} Z_0$$

Et son taux de croissance est :

$$g_z = \frac{(1 + g_v)}{(1 + g_u)} - 1$$

Moyenne arithmétique simple : Soit $\{x_1, x_2, \ldots, x_n\}$ une série de chiffres. La formule de la moyenne arithmétique de cette série est donnée par :

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

Moyenne arithmétique pondérée : Soit $\{x_1, x_2, \ldots, x_h\}$ une série de chiffres et $\{n_1, n_2, \ldots, n_h\}$ les effectifs correspondants. La formule de la moyenne arithmétique pondérée de cette série est donnée par :

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{h} (n_i \cdot x_i)$$

Moyenne quadratique simple : Soit $\{x_1, x_2, \ldots, x_n\}$ une série de chiffres. La formule de la moyenne quadratique simple de cette série est donnée par :

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

Moyenne quadratique pondérée : Soit $\{x_1, x_2, \ldots, x_h\}$ une série de chiffres et $\{n_1, n_2, \ldots, n_h\}$ les effectifs correspondants. La formule de la moyenne quadratique pondérée de cette série est donnée par :

$$Q = \sqrt{\frac{1}{n} \sum_{i=1}^{h} (n_i \cdot x_i^2)}$$

Moyenne géométrique simple : Soit $\{x_1, x_2, \ldots, x_n\}$ une série de chiffres. La formule de la moyenne géométrique simple de cette série est donnée par :

$$G = \left[\prod_{i=1}^{n} x_i \right]^{\frac{1}{n}}$$

Moyenne géométrique pondérée : Soit $\{x_1, x_2, \ldots, x_h\}$ une série de chiffres et $\{n_1, n_2, \ldots, n_h\}$ les effectifs correspondants. La formule de la moyenne géométrique pondérée de cette série est donnée par :

$$G = \left[\prod_{i=1}^{h} x_i^{n_i} \right]^{\frac{1}{n}}$$

Moyenne harmonique simple : Soit $\{x_1, x_2, \ldots, x_n\}$ une série de chiffres. La formule de la moyenne harmonique simple de cette série est donnée par :

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

Moyenne harmonique pondérée : Soit $\{x_1, x_2, \ldots, x_h\}$ une série de chiffres et $\{n_1, n_2, \ldots, n_h\}$ les effectifs correspondants. La formule de la moyenne harmonique pondérée de cette série est donnée par :

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

Médiane quand les effectifs groupées par classes de valeurs

$$M_e = x_i^{\inf} + a_i \left[\frac{\frac{n}{2} - N\left(x_{i-1}\right)}{n_i} \right]$$

où : $x_i^{\inf} =$ Borne inférieure de la classe médiane.

$N\left(x_{i-1}\right) =$ Effectif cumulé strictement inférieur à x_i

Classe médiane

$x_i =$ Classe médiane $a_i =$ Amplitude de la classe médiane

Mode quand les effectifs sont groupés par classes d'amplitudes égales

$$\text{Mode} = \mathrm{x}_{\mathrm{i}}^{\inf} + a\frac{d_1}{d_1 + d_2}$$

$\mathrm{x}_{\mathrm{i}}^{\inf}$ = Borne inférieure de la classe modale a=amplitude de classe

$d_1=n_i-n_{i-1}$ et $d_2=n_i-n_{i+1}$

Variance : Soit une série de valeurs d'une variable X : $\{x_1,x_2,\ldots,x_k\}$. Soit les effectifs associés à cette modalité : $\{n_1,n_2,\ldots,n_k\}$. La variance de cette série s'écrit :

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{k} n_i\left(x_i - \overline{x}\right)^2$$, si l'effectif considéré est celui d'une population

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{k} n_i\left(x_i - \overline{x}\right)^2$$, si l'effectif considéré est celui d'un échantillon

Remarques : 1) Si $\{n_1,n_2,\ldots,n_k\} = \{1,1,\ldots,1\}$ et que k=n, la variance

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2$$, si l'effectif considéré est celui d'une population

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2$$, si l'effectif considéré est celui d'un échantillon

2) La formule développée de la variance est :

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

Écart-type : L'écart-type est égal à la racine carrée de la variance :

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

Si aucune valeur n'est répétée ou si les données ne sont pas regroupées par valeur, on a :

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

Coefficient de variation : Il est donné par le rapport de l'écart-type à la moyenne, multiplié par 100.

$$CV = \left(\frac{\sigma}{\overline{x}}\right) \times 100$$

Médiale : C'est un indicateur qui s'apparente à la médiane, mais appliquée à une série différente. En effet, alors que la médiane s'applique aux valeurs de la variable (les «x_i »), la médiale s'applique aux valeurs de la variables multipliées par leurs effectifs respectifs (les « $n_i.x_i$ »). C'est la valeur du caractère qui partage l'effectif cumulé des $n_i.x_i$ en deux parties égales. Elle sert à déterminer la concentration de la distribution par comparaison avec la médiane et l'intervalle de variation. D'où la formule :

$$M_I = x_i^{\inf} + a_i \left[\frac{\frac{\sum_{i=1}^{k} n_i x_i}{2} - N\left(n_i x_i\right)}{n_i x_i} \right]$$

où : $x_i^{\inf} =$ Borne inférieure de la classe médiale.

$N\left(x_{i-1}\right) =$ Effectif cumulé strictement inférieur à $n_i\ x_i$

$x_i =$ Classe médiale $a_i =$ Amplitude de la classe médiale

Indice de GINI : La formule analytique de l'indice de GINI est donnée par:

$$I = \frac{\sum_i \sum_j \left| x_i - x_j \right| n_i n_j}{2n\left(n-1\right)\overline{x}}$$

Moyennes marginales : Soit deux variables X et Y, dont on étudie la liaison. Les moyennes marginales de X et de Y sont données par :

$$\overline{\overline{x}} = \frac{1}{n_{\bullet\bullet}} \sum_{i=1}^{p} n_{i\bullet} x_i$$

$$\overline{\overline{y}} = \frac{1}{n_{\bullet\bullet}} \sum_{j=1}^{q} n_{\bullet j} y_j$$

où :

$$n_{i\bullet} = \sum_{j=1}^{q} n_{ij} = n_{i1} + n_{i2} + \cdots + n_{iq}$$

$$n_{\bullet j} = \sum_{i=1}^{p} n_{ij} = n_{1j} + n_{2j} + \cdots + n_{pj}$$

$$n_{\bullet\bullet} = \sum_{i=1}^{p} n_{i\bullet} = \sum_{i=1}^{p}\left(\sum_{j=1}^{q} n_{ij}\right) = \sum_{j=1}^{q} n_{\bullet j} = \sum_{j=1}^{q}\left(\sum_{i=1}^{p} n_{ij}\right)$$

Variances marginales : Les variances marginales de x et de y se calculent à partir des distributions marginales suivant les formules suivantes :

$$\sigma_x^2 = \frac{1}{n_{\bullet\bullet}} \sum_{i=1}^{p} n_{i\bullet}\left(x_i - \overline{\overline{x}}\right) = \frac{1}{n_{\bullet\bullet}} \sum_{i=1}^{p} n_{i\bullet} x_i^2 - \left(\overline{\overline{x}}\right)^2$$

$$\sigma_y^2 = \frac{1}{n_{\bullet\bullet}} \sum_{j=1}^{q} n_{\bullet j}\left(y_j - \overline{\overline{y}}\right) = \frac{1}{n_{\bullet\bullet}} \sum_{j=1}^{q} n_{\bullet j} y_j^2 - \left(\overline{\overline{y}}\right)^2$$

Moyennes conditionnelles : La formule des moyennes conditionnelles de x et de y est donnée par :

$$\bar{x}_j = \frac{1}{n_{\bullet j}} \sum_{i=1}^{p} n_{ij} x_i \quad 1 \le j \le p$$

$$\bar{y}_j = \frac{1}{n_{i\bullet}} \sum_{j=1}^{q} n_{ij} y_j \quad 1 \le i \le q$$

Variances conditionnelles : La formule des variances conditionnelles de x et de y est donnée par :

$$V(x_j) = \frac{1}{n_{\bullet j}} \sum_{i=1}^{p} n_{ij} \left(x_i - \bar{x}_j\right)^2 = \frac{1}{n_{\bullet j}} \sum_{i=1}^{p} n_{ij} x_i^2 - \bar{x}_j^2 \quad 1 \le j \le p$$

$$V(y_i) = \frac{1}{n_{i\bullet}} \sum_{j=1}^{q} n_{ij} \left(y_j - \bar{y}_i\right)^2 = \frac{1}{n_{i\bullet}} \sum_{j=1}^{q} n_{ij} y_j^2 - \bar{y}_i^2 \quad 1 \le i \le q$$

Droite de régression linéaire : Soit la droite d'équation :

$$y = ax + b$$

Pour ajuster par une droite un nuage de points dans le plan {X,Y}, il faut calculer les coefficients a et b en appliquant les formules suivantes :

$$a = \frac{\operatorname{cov}(x,y)}{\sigma_x^2} \qquad b = \bar{\bar{y}} - a\bar{\bar{x}}$$

Où cov(x,y) représente la covariance de (x,y) et se calcule ainsi :

$$\operatorname{cov}(x,y) = \frac{1}{n} \sum_{i=1}^{n} x_i y_i - \bar{\bar{x}}\bar{\bar{y}}$$

Par conséquent, la formule détaillée de *a* est :

$$a = \frac{\frac{1}{n} \sum_{i=1}^{n} x_i y_i - \bar{\bar{x}}\bar{\bar{y}}}{\frac{1}{n} \sum_{i=1}^{n} x^2 - \left(\bar{\bar{x}}\right)^2}$$

Coefficient de corrélation (données non groupées) : il mesure la plus ou moins grande dépendance entre les deux caractères X et Y. On le désigne par la lettre "r" et il varie entre -1 et +1 :

$$r = \frac{\operatorname{cov}(x,y)}{\sigma_x \sigma_y} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{\overline{xy}}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} x^2 - \left(\overline{\overline{x}}\right)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n} y^2 - \left(\overline{\overline{y}}\right)^2}}$$

Plus r est proche de +1 ou de -1, plus les deux caractères sont dépendants. Plus il est proche de 0, plus les deux caractères sont indépendants

Test d'indépendance : Deux variables X et Y sont indépendantes si et seulement si :

$$n_{ij} = \frac{n_{i\bullet} \times n_{\bullet j}}{n_{\bullet\bullet}}$$

Il suffit donc a contrario qu'un n_{ij} quelconque soit tel que :

$$n_{ij} = \frac{n_{i\bullet} \times n_{\bullet j}}{n_{\bullet\bullet}}$$

pour que l'on puisse conclure à l'absence d'indépendance. Il est donc généralement plus rapide de vérifier l'absence d'indépendance que d'établir l'indépendance.

Coefficient de corrélation (données groupées) : Quand les données sont groupées, le coefficient de corrélation s'écrit :

$$r = \frac{\frac{1}{n_{\bullet\bullet}}\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij} x_i y_j - \overline{\overline{xy}}}{\sqrt{\frac{1}{n_{\bullet\bullet}}\sum_{i=1}^{p} n_{i\bullet} cx_i^2 - \left(\overline{\overline{x}}\right)^2}\sqrt{\frac{1}{n_{\bullet\bullet}}\sum_{j=1}^{q} n_{\bullet j} cy_j^2 - \left(\overline{\overline{y}}\right)^2}}$$

Indice élémentaire : Un indice élémentaire renseigne sur l'évolution temporelle ou situationnelle (spatiale) d'une seule valeur. Il a pour formule :

$$I_{t/0} = \frac{V_t}{V_0} \times 100$$

Valeur d'un panier de produits : Soit $V_t^i = p_t^i \, q_t^i$ la valeur du bien *i*, à la date i où p_t^i représente le prix du bien *i* à la date t et q_t^i sa quantité. S'il y a *n* produits dans le panier (i=1,n), la valeur totale du panier à la date t s'écrit :

$$V_i = \sum_{i=1}^{n} p_t^i q_t^i$$

Indice de LASPEYRES des prix : Il mesure l'évolution entre deux dates 0 et t, des prix des biens qui composent un panier, en prenant comme référence la valeur du panier à la date initiale (t=0) et en supposant que les quantités de biens dans le panier n'ont pas varié entre 0 et t. Sa formule est :

$$L_{t/0}^{P} = \frac{\sum_{i=1}^{n} p_t^i q_0^i}{\sum_{i=1}^{n} p_0^i q_0^i} \times 100$$

où p_t^i représente le prix du bien *i* à la date t et q_t^i sa quantité.

Indice de LASPEYRES des quantités : Il mesure l'évolution entre deux dates 0 et t, des quantités des biens qui composent un panier, en prenant comme référence la valeur du panier à la date initiale (t=0) et en supposant que les prix des biens dans le panier n'ont pas varié entre 0 et t. Sa formule est :

$$L_{t/0}^{Q} = \frac{\sum_{i=1}^{n} p_0^i q_t^i}{\sum_{i=1}^{n} p_0^i q_0^i} \times 100$$

où p_t^i représente le prix du bien *i* à la date t et q_t^i sa quantité.

Indice de PAASCHE des prix : Il mesure l'évolution entre deux dates 0 et t, des prix des biens qui composent un panier, en prenant comme référence la valeur du panier à la date terminale (t) et en supposant que les quantités de biens dans le panier n'ont pas varié entre 0 et t. Sa formule est :

$$P_{t/0}^{P} = \frac{\sum_{i=1}^{n} p_t^i q_t^i}{\sum_{i=1}^{n} p_0^i q_t^i} \times 100$$

où p_t^i représente le prix du bien *i* à la date t et q_t^i sa quantité.

Indice de PAASCHE des quantités : Il mesure l'évolution entre deux dates 0 et t, des quantités des biens qui composent un panier, en prenant comme référence la valeur du panier à la date terminale (t) et en supposant que les prix des biens dans le panier n'ont pas varié entre 0 et t. Sa formule est :

$$P_{t/0}^{Q} = \frac{\sum_{i=1}^{n} p_t^i q_t^i}{\sum_{i=1}^{n} p_t^i q_0^i} \times 100$$

où p_t^i représente le prix du bien *i* à la date t et q_t^i sa quantité.

Indice de FISHER des prix : C'est la moyenne géométrique des indices de prix de LASPEYRES et de PAASCHE :

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

Indice de FISHER des quantités : C'est la moyenne géométrique des quantités de prix de LASPEYRES et de PAASCHE :

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

Bibliographie

Bernard PY, 1996, *Statistique descriptive, nouvelle méthode pour bien comprendre et réussir*, 4e édition, Economica.

Bernard PY, 1994, *Exercices corrigés de statistique descriptive*, 2e édition, Economica.

Alain PILLER, 2004, *Statistique Descriptive*, éditions Premium.

Maurice LETHIELLEUX, 2003, *Statistique Descriptive*, éditions Dunod, collection « Express ».

INSEE, 2005, *Tableaux de l'économie française*, Insee éditeur, collection « Références ».

INSEE, (site Internet) : www.insee.fr

Deborah RUMSEY, 2003, *Statistics for Dummies*, Wiley Publishing inc.

Lloyd R. JAISINGH, 2000, *Statistics for the Utterly Confused*, McGraw-Hill.

David S. MOORE & George P. MCCABE, 2002, *Introduction to the Practice of Statistics*, 4e édition, W.H. Freeman & Company .

Edward R. TUFTE, 2001, *The Visual Display of Quantitative Information*, Graphics Press.

Trevor BOUNDFORD, 2000, *Digital Diagrams*, Watson-Guptill Publications.