INSTITUT SUPERIEUR PEDAGOGIQUE

B.P : 127
MBANZA-NGUNGU

Section des Sciences et Technologies
Département de l'Informatique

MAT121 : Statistique descriptive

L1 Informatique et Technologie

Rabin Elembo Z.

Avant-propos

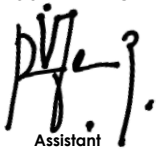
DESCRIPTION TECHNIQUE			
Enseignant	: ELEMBO Rabin	**Unité d'Enseignement**	: MAT121 : Statistique descriptive
Grade	: Assistant 1	**Nbre de Crédit**	: 3 Crédits – 45 heures
Titre	: Licencié	**Partie Magistrale**	: 30 heures en 4 Séances
Coordonnée	Tél. : +243992621890 Email : rabbielembo@gmail.com	**TP / TPE**	: 15 heures
Mention	: Informatique et Technologie	**Mode d'enseignement**	: Présentiel
Site	: MBANZA-NGUNGU	**Semestre**	: Deuxième

La statistique descriptive est une discipline fascinante qui nous permet de donner du sens aux données ; En résumant et en visualisant l'information de manière claire, elle offre une première vision synthétique des caractéristiques d'une distribution. Outil essentiel en recherche et en analyse quantitative, elle facilite l'interprétation des résultats. Son apprentissage constitue la base de tout travail statistique rigoureux. Maîtriser ces méthodes est indispensable pour exploiter correctement des données dans divers domaines scientifiques.

Pour l'élaboration de ce cours nous nous sommes basés sur des ouvrages, résultats et bien des ressources sur le sujet. Des lignes y ont été tiré, parfois entièrement, paraphrasé ou mixé avec nos idées sur le sujet pour plus d'originalité. Aussi, les travaux antérieurs en statistique du C.T Mbaka Ruffin de l'ISP Mbanza-Ngungu, en qui nous présentons nos vifs remerciements pour sa collaboration fructueuse.

Dans le cadre de ce UE, pour rester pragmatique et habituer les étudiants dans l'utilisation des outils logiciels pour statistique, nous avons jugé utile des faire nos rendus statistiques sur Python avec google Colab (en ligne) ou sur anaconda en local. Cette condition bien que nécessaire n'est cependant pas une obligation pour suivre cet enseignement.

Rabin ELEMBO



Assistant

Sommaire

Introduction

Objectifs

L'objectif de ce cours est de présenter aux étudiants un ensemble de méthodes de statistiques descriptives qui leur permet d'exploiter et d'analyser des fichiers de données quantitatives et qualitatives.

A la fin de ce cours, l'étudiant sera capable de collecter les données, d'organiser les données collectées et de synthétiser les données organisées à l'aide d'un certain nombre de paramètres.

A cet effet, les exercices, proposés dans le cadre des travaux dirigés (TD), devraient faciliter la réalisation de cet objectif.

Eléments bibliographiques

- ✓ Baggio, S., Deline, S. et Rothen, S. (2017), *Statistique Descriptive*, De Boeck, Paris.
- ✓ Bressoud E. et Kahane J.C. (2010), *Statistique descriptive*, 2ème édition, Pearson, Paris
- ✓ Hubler, J. (2007), *Statistique descriptive appliquée à la gestion et à l'économie*, Bréal, Paris.
- ✓ Leboucher, L. et Voisin, M. (2015), *Introduction à la Statistique Descriptive*, Cépaduès-Éditions, Paris.
- ✓ Lethielleux, M. (2016), *Statistique Descriptive*, Dunod
- ✓ Makhlouf, F. (2022), *Fiches de Statistique Descriptive*, Ellipses, Paris.
- ✓ Spiegel M. et Stephens L., *Statistique : Cours et problèmes*, 3ème édition, Série Schaum/McGraw Hill
- ✓ Mazerolle, F. (2006), *Statistique Descriptive : Séries statistiques à une et deux variables, séries chronologiques et des indices, Gualino éditeur, Paris.*
- ✓ Krickerberg, K. (1996), *Petit cours de statistique, Springer, Berlin.*

Partie I : STATISTIQUE UNIVARIEE

Chapitre 1 : NOTIONS FONDAMENTALES

1.1. Généralités

1.1.1. Définition

La statistique est la science qui a pour objet de recueillir, organiser, classer, présenter et interpréter les données.

La statistique (science) est à distinguer des statistiques (généralement employée au pluriel) qui désigne un chiffre ou une collection de chiffres se rapportant à un sujet quelconque et élaborés grâce à des outils et des méthodes statistiques.

1.1.2. Objet et utilité de la statistique

L'objet de la statistique est l'étude des faits pour prendre des décisions. Elle utilise des outils mathématiques pour étudier les propriétés numériques des ensembles de faits nombreux.

Elle permet donc de :

- ✓ Décrire les caractéristiques d'une population ainsi que les relations entre les critères qui caractérisent la population. *Exemple* : décrire le lien entre l'ancienneté des employés et leur salaire ;
- ✓ Estimer des paramètres et prendre des décisions ;
- ✓ Prévoir et éventuellement expliquer.

La statistique aide un pays à mesurer des grandeurs importantes (comme le chômage, la population ou la croissance économique) pour comprendre la situation actuelle d'un phénomène, son évolution dans le temps, de prévoir son état futur (prévision des recettes de l'Etat), de comparer des entités, de décider de l'action à mener.

Elle se pose des questions comme : *Combien de personnes sont au chômage ? quel est le taux de pauvreté cette année ? la population augmente-t-elle ou diminue-t-elle ?*

En somme, la statistique donne des chiffres clés pour décrire une réalité et agir en conséquence.

L'enseignement de la statistique présente essentiellement deux grandes branches :

- ✓ **Les *méthodes descriptives* :** elles comprennent les statistiques descriptives et l'analyse des données (analyses factorielles et classification). Elles servent à simplifier un ensemble de données (généralement vaste) sans trop perdre d'information ; par de graphes, de tableaux et de nombres qui résument les données ;

- ✓ **La *statistique mathématique*** dont l'objet et de formuler *découvrir des lois ou des tendances générales* à partir d'échantillons et de sous-ensembles d'une population statistique.

1.2. Définition des concepts usuels de la statistique

1.2.1. Population, individu et unités statistiques

L'ensemble sur lequel porte une étude statistique est appelé ***population***. Chaque élément de cet ensemble est appelé ***individu*** ou ***unité statistique***.

Note :

- ✓ On emploie les termes population et individu aussi bien lorsqu'il s'agit d'un ensemble d'êtres humains (les salariés d'une entreprise) ou d'objets inanimés ou bien d'un ensemble plus ou moins abstrait (ensemble des accidents de la route au cours d'une période donnée, parc de micro-ordinateurs dans une entreprise).

- ✓ La statistique ne s'intéresse pas à chaque individu personnellement, mais à ce que leurs caractéristiques révèlent sur le groupe entier (*la population*). Pour cela, la population doit être définie avec précision enfin que tous les intervenants qui concourent à l'observation, au traitement, à l'analyse ou à l'utilisation de l'information en aient la même compréhension et des conclusions cohérentes.

Exemples :

- ✓ Tous les salariés à temps plein de la S.A SNEL Mbanza-Ngungu en 2024
- ✓ Les étudiants de L1 Architecture de l'UK inscrits l'année académique 2025-2026.

1.2.2. Echantillon/Population mère

Il est souvent difficile voire impossible de mener une étude statistique sur une population toute entière. On choisit alors de travailler sur une partie de cette population. La sous-population choisie est appelée *échantillon*. La population initiale d'où est tire l'échantillon est la population mère.

La *taille* d'un échantillon *n* (ou d'une population *N*) est le nombre d'unités statistiques qui le composent.

1.2.3. Variable statistique ou caractère

Une variable statistique (ou caractère) est une caractéristique mesurable ou observable qui varie d'un individu à l'autre dans une population étudiée. C'est le critère ou la propriété suivant lequel on étudie la population statistique.

Exemple :

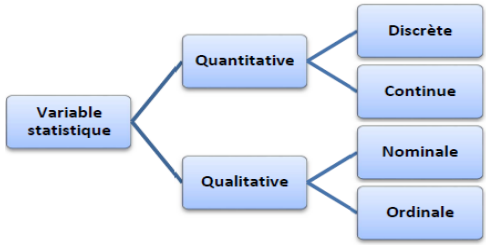
- ✓ Nombre d'enfants dans un ménage ;
- ✓ Revenus mensuels des travailleurs d'une entreprise ;
- ✓ Le niveau d'éducation des policiers Ngungois ;

La variable statistique prend des valeurs différentes pour les individus de la population. Les valeurs possibles d'une variable statistique sont ses *modalités*.

Exemple : Groupe sanguin : A, B, AB, O^+

La variable statistique peut être *qualitative* ou *quantitative*.

1.2.4. Types de variables statistiques



- ➢ ***Variable quantitative*** : numérique, mesurable ou repérable
  Exemples : âge, poids, ancienneté, température, taille, nombre d'étudiant.

  - ✓ *Variable quantitative discrète* : variable dont les modalités sont des valeurs isolées (par exemple des valeurs entières).

    Exemple : nombre d'enfants à charge, taille des entreprises (en nombre d'employés), nombre de pièces des logements des ménages.

  - ✓ *Variable quantitative continue* : variable pouvant prendre toute valeur dans un intervalle donné. En général, ses modalités sont des nombres à virgule.

    Exemple : âge, poids (en kilogrammes), taille (en mètres), PIB par tête des pays, salaire des employés.

    En pratique, on considère qu'une variable quantitative est continue lorsqu'elle prend un très grand nombre de valeurs possibles.

    Exemple : le revenu, le salaire des employés d'une entreprise.

- ➢ ***Variable qualitative*** : les modalités sont non mesurables. Elles sont généralement représentées par des noms qui traduisent des états. Elles sont catégorielles.

Exemple : Couleurs des yeux, situation matrimoniale, appréciation d'un cours par les étudiants.

Note : Les modalités peuvent être représentées par des chiffres qui représentent le codage et non une mesure.

Exemple : Situation matrimoniale :

1 = Célibataire 2 = Marié 3 = Divorcé 4 = Veuf

- ✓ *Variable qualitative nominale* : les modalités ne présentent aucun ordre, aucune hiérarchie entre elles.
  Exemple : situation matrimoniale, couleur des yeux

- ✓ *Variable qualitative ordinale* : les modalités respectent un certain ordre.

  Exemple : Appréciation d'un cours : Mauvais < Bon < TB

1.3. Elaboration de statistiques

L'étude statistique des phénomènes commence par la collecte des données de base. Cette collecte se fait à partir d'*enquêtes* (collecte auprès de personnes physiques ou morales), de résultats d'expériences ou d'exploitation de fichiers administratifs.

L'observation des faits peut se faire de façon instantanée ou ponctuelle (comme un sondage politique avant des élections, recensement étatique décennal de la population) ou de façon continue (enregistrement des naissances à l'état-civil, comptabilité d'une entreprise, suivi quotidien des ventes d'un supermarché).

1.3.1. Recensement

C'est une méthode exhaustive, c'est-à-dire que toute la population, pas un échantillon, fait l'objet d'observation suivant le ou les caractères étudiés. Important pour obtenir une photo complète de la population (pas d'estimations).

Exemple : recensement de la population de la RDC en 2023 suivant des caractères démographiques (âge, sexe, etc.), économiques (activités économiques), sociaux (niveau d'éducation, alphabétisation, etc.), géographiques (lieu de résidence).

1.3.2. Enquête par sondage

Elle porte que sur un échantillon. Elle est bien plus rapide mais toujours précise.

Exemples :
- Enquête sur les conditions de vie des ménages ;
- Enquête démographique et de santé

1.3.3. Les grandes étapes d'une enquête statistique

Le déroulement d'une enquête statistique peut être résumé en quatre (4) grandes étapes :

- ✓ ***La conception*** : Elle consiste à définir les objectifs de l'étude, définir l'ensemble de l'étude ainsi que les critères à étudier, à concevoir les outils nécessaires à la collecte des informations (questionnaires, guide d'entretien, manuels des agents, etc.). Elle doit également définir les résultats attendus, notamment les indicateurs essentiels à calculer.

- ✓ ***La phase de collecte*** : C'est le passage à l'action où l'on récolte les données sur le terrain. Elle comprend la formation des acteurs, la sensibilisation des personnes cibles, l'observation et l'enregistrement de l'information à l'aide de questionnaires. La collecte peut se faire par interview directe, par courrier (poste, e-mail), par téléphone, etc.

- ✓ ***La phase de traitement*** : C'est le nettoyage et l'organisation des informations collectées. Elle consiste à la validation des questionnaires, la codification des réponses, le dépouillement (manuel ou automatique) et le traitement éventuel des données manquantes, des erreurs de saisie, etc.

- ✓ ***La phase d'analyse et de diffusion*** : C'est le moment de donner du sens aux données. Calcul des indicateurs, critique et interprétation des résultats, présentation des résultats obtenus.

1.4. Exercices

1) On a demandé aux employés d'une entreprise pour quel parti politique ils avaient voté lors des dernières élections. Voici les données brutes obtenues :
UNC PPRD UNC AFDC UNC UDPS UNC UDPS PPRD UNC MPR AFDC UDPS PPRD MPR UDPS UDPS UDPS PPRD UNC PPRD AFDC PPRD AFDC UDPS AFDC UNC UDPS UDPS UDPS

   a) Identifier la population.
   b) Identifier la variable statistique.
   c) Donner l'ensemble des modalités.
   d) De quel type est cette variable statistique ?

2) Un professeur a noté le nombre de points (strictement positif) obtenus par 80 étudiants lors d'un test de statistiques : 2 3 5 5 4 6 6 5 4 3 7 7 7 6 2 7 7 9 8 10 5 6 6 8 6 6 3 7 3 5 9 7 6 4 7 5 9 9 6 9 6 3 9 8 8 7 5 6 10 6 9 7 7 7 4 7 10 8 7 10 3 5 8 5 8 7 4 8 10 7 4 6 6 8 7 7 7 8 8 9.

   a) Identifier la population.
   b) Identifier la variable statistique.
   c) Donner l'ensemble des modalités.
   d) De quel type est cette variable statistique ?

SERIE D'EXERCICES
CHAP.01

Chapitre 2 : PRESENTATION DES DONNEES

A l'issue de la collecte des données (lors d'une enquête par exemple), les informations recueillies ne sont pas immédiatement exploitables. Il est alors nécessaire de les organiser, les ordonner et les présenter de façon lisible et facilement compréhensible.

Pour cela la statistique descriptive offre des techniques pour la représentation des données sous forme de *tableaux ou de graphiques*.

Une *série statistique* est la liste des valeurs de la variable statistique observées sur les individus d'un échantillon d'une population donnée.

1.5. Tableaux statistiques

Un tableau statistique constitue un résumé ou une synthèse numérique des résultats d'une série statistique.

On distingue trois formes de tableaux statistiques qui sont fonction de l'objectif envisagé et de la nature du caractère étudié.

1.5.1. Tableau de dénombrement

Le tableau de dénombrement donne un résumé numérique d'une distribution statistique. Il reprend, sur la première colonne, les modalités du caractère étudié, et sur la deuxième l'effectif de la modalité correspondante.

Modalités	Effectif
M_1	n_1
M_2	n_2
…	…
M_n	nn
Somme	n

L'*effectif* (*fréquence absolue*) est le nombre des éléments statistiques relatifs à une modalité donnée, noté n_i. Autrement dit, l'effectif d'une valeur donnée (n_i) d'une variable est l'ensemble d'individus présentant cette valeur. L'effectif total (N ou n) est la somme de tous les effectifs d'une variable.

$$N = n_1 + n_2 + ... + n_n$$

On écrira alors :

$$N = \sum_{i=1}^{n} n_i$$

Note : Le tableau doit avoir un titre et une source d'où ont été puisées les données.

1.5.1.1. Cas d'une variable qualitative nominale

Soient les données suivantes provenant d'une étude sur la couleur préférée des étudiants menée auprès des étudiants de première année de Licence en Informatique à l'I.S.P. Mbanza-Ngungu :

Rose	Blanche	Jaune	Blanche	Rose	Noire	Verte
Orange	Verte	Rose	Jaune	Blanche	Jaune	Orange
Jaune	Jaune	Orange	Noire	Verte	Verte	Orange
Orange	Rose	Orange	Noire	Noire	Verte	Noire
Orange	Noire	Rose	Rose	Jaune	Blanche	Orange

Pour élaborer un tableau, il convient de procéder par un *dépouillement* des données brutes en vue de bien les organiser :

- ✓ Blanche : |||| = 4
- ✓ Jaune : 𝍸 | = 6
- ✓ Noire : 𝍸 || = 7
- ✓ Orange : 𝍸 ||| = 8
- ✓ Rose : 𝍸 | = 6
- ✓ Verte : |||| = 4

Voici le tableau :

Couleur	n_i
Blanche	4
Jaune	6
Noire	7
Orange	8
Rose	6
Verte	4
Somme	**35**

1.5.1.2. Cas d'une variable qualitative ordinale

Avec le caractère qualitatif ordinal, on tient compte de l'ordre.

On appelle *effectif cumulé croissant* le nombre d'individus qui correspondent au même caractère (modalité) et aux caractères précédents.

Exemple : On mène une étude sur le dernier diplôme obtenu par la population des salariés d'une entreprise. Les données sont les suivantes :

D	M	L	S	S	L	P	S	S	L
L	S	L	S	P	P	M	L	M	P
S	M	M	M	D	S	M	L	S	P
P	L	D	L	S	M	S	S	L	D

Le dépouillement indique :

- ✓ P 𝍸 | = 6
- ✓ S 𝍸 𝍸 || = 12
- ✓ L 𝍸 𝍸 = 10
- ✓ M 𝍸 ||| = 8
- ✓ D |||| = 4

Le tableau sera :

Diplôme	n_i	N_i^+	N_i^-
		0	40
Primaire	6	6	36
Secondaire	12	18	22
Licence	10	28	12
Maîtrise	8	36	4
Doctorat	4	40	0
Somme	40		

1.5.1.3. Cas de variable quantitative discrète

Le même tableau est valable pour la variable quantitative discrète.

Exemple : Voici les données provenant d'une enquête portant sur le nombre de pièces dans l'habitation des étudiants en Informatique :

4	6	3	4	2	2	4	4	2	3
5	4	1	1	3	4	3	5	3	4
3	5	2	3	4	3	2	5	3	2
2	2	4	2	1	1	4	2	1	3
4	3	5	3	3	5	6	3	4	5

Le dépouillement indique :

✓ 1. 𝍸 = 5 ✓ 3. 𝍸 𝍸 | | | | = 14 ✓ 5. 𝍸 | | = 7
✓ 2. 𝍸 𝍸 = 10 ✓ 4. 𝍸 𝍸 | | = 12 ✓ 6. | | = 2

Le tableau sera :

x_i	n_i	N_i^+	N_i^-
		0	50
1	5	5	45
2	10	15	35
3	14	29	21
4	12	41	9
5	7	48	2
6	2	50	0
∑	**50**		

1.5.1.4. Cas d'une variable continue

Dans le cas d'une variable quantitative continue, l'établissement du tableau statistique implique d'effectuer au préalable une répartition en *classes* des données. Cela nécessite de définir le *nombre de classes* attendu *J* et donc l'*amplitude* associée à chaque classe ou *intervalle* de classe.

En règle générale, on choisit des classes de même amplitude. Pour que la distribution en fréquence ait un sens, il faut que chaque classe comprenne un nombre suffisant de valeurs (n_i).

Certaines formules empiriques permettent d'établir le nombre de classes pour un échantillon de taille n.

- ✓ La règle de Sturge : $J = 1 + 3{,}3 \log n$
- ✓ La règle de Yule : $J = 2{,}5\sqrt[4]{n}$
- ✓ Si le nombre de classes est connu, on calcule l'amplitude de chaque classe (dans le cas où elle est constante) de la manière suivante :

$$a = \frac{X_{max} - X_{min}}{J}$$

Où :
- X_{max} est la plus grande valeur de la série
- X_{min} est la plus petite valeur de la série
- J est le nombre de classes.

Exemple : On a mesuré la taille (en centimètres) des mannequins de l'ISAM Kisantu et les données collectées sont les suivantes :

154	156	169	164	152	162	157	169	162	153
159	164	167	155	165	166	160	159	173	164
173	158	162	165	174	163	166	165	163	155
162	169	154	162	159	171	164	172	170	163
158	173	156	168	166	150	167	163	164	151

Questions :

1) Combien de classes doit-on former selon la formule de Sturge ?
2) Combien de classes doit-on former selon la formule de Yule ?
3) Tracez un tableau des données avec 5 classes.

Solution :

1) Nombre de classes selon la formule de Sturge :

$$J = 1 + 3{,}3 \log n = 1 + 3{,}3 \log 50$$
$$J = 1 + 3{,}3 \log 50 = 1 + 3{,}3 \times 1{,}69897$$
$$J = 5{,}6 \approx 6\ classes$$

2) Nombre de classes selon la formule de Yule :

$$J = 2{,}5\sqrt[4]{n} = 2{,}5\sqrt[4]{50}$$
$$J = 2{,}5 \times 2{,}659148$$
$$J = 6{,}64787 \approx 7\ classes$$

3) Tracez un tableau des données avec 5 classes.

Avec les données du tableau : $x_{max} = 174$ et $x_{min} = 174$

$$IV = x_{max} - x_{min} = 174 - 150 = 24$$
$$a = \frac{IV}{J} = \frac{24}{5} = 4{,}8 \approx 5$$

Le dépouillement donne :

- ✓ [150 ; 155[𝍸 | = 6
- ✓ [155 ; 160[𝍸 𝍸 = 10
- ✓ [160 ; 165[𝍸 𝍸 𝍸 = 15
- ✓ [165 ; 170[𝍸 𝍸 = 10
- ✓ [170 ; 175[𝍸 || = 7

Note : Avec le regroupement en classes, nous devons reconnaître les éléments ci-après :

- La borne inférieure : c_i^-
- La borne supérieure : c_i^+
- Le centre de classe $x_i = \frac{c_i^- + c_i^+}{2}$

Le tableau statistique devient :

Classes	n_i	c_i	N_i^+ (0)	N_i^- (50)
150 – 155	6	152,5	6	44
155 – 160	10	157,5	16	36
160 – 165	15	162,5	31	19
165 – 170	12	167,5	43	7
170 – 175	7	172,5	50	0
∑	**50**			

1.5.2. Tableaux des fréquences

La fréquence (fréquence relative) d'une valeur est le rapport de l'effectif de cette valeur par l'effectif total.

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

- ✓ On peut remplacer f_i par $f_i \times 100$ qui représente alors un *pourcentage*.
- ✓ La somme des fi est toujours égale à 1 :

$$f_1 + f_2 + \cdots + f_n = \sum_{i=1}^{n} f_i = 1$$

Exemple : Le tableau précédent devient ce qui suit :

Classes	n_i	f_i	%	c_i	N_i^+ (0)	N_i^- (50)	F_i^+ (0)	F_i^- (1)
150 – 155	6	0,12	12	152,5	6	44	0,12	0,88
155 – 160	10	0,20	20	157,5	16	36	0,32	0,68
160 – 165	15	0,30	30	162,5	31	19	0,62	0,38

165 – 170	12	0,24	24	167,5	43	7	0,86	0,14
170 – 175	7	0,14	14	172,5	50	0	1	0
∑	**50**	**1**	**100**					

Avec les données regroupées en classes, il faut retenir que :

- ✓ Chaque classe possède une certaine *amplitude*, qui est la longueur de l'intervalle définissant la classe.
- ✓ Le rapport entre l'effectif d'une classe et son amplitude s'appelle *la densité d'effectif*.
- ✓ Le rapport entre la fréquence d'une classe et son amplitude s'appelle *la densité de fréquence*.

1.6. Les graphiques

Les représentations graphiques ont l'avantage de renseigner immédiatement sur l'allure générale de la distribution. Elles facilitent l'interprétation des données recueillies et dépendent de la nature des données en présence.

1.6.1. Cas qualitatif

1.6.1.1. Le camembert

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.

Dans ce graphique, l'effectif total est représenté par un disque. Chaque modalité est représentée par un secteur circulaire dont la surface (pratiquement : l'angle au centre) est proportionnelle à l'effectif correspondant.

L'angle α_i d'une modalité d'effectif n_i est donné en degrés par :

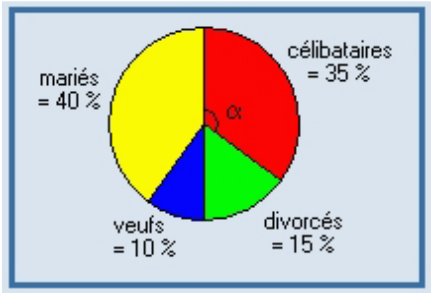
$$\alpha_i = \frac{n_i}{n} \cdot 360 = f_i \cdot 360$$

Exemple : Si l'on considère maintenant que parmi les travailleurs de l'I.S.P. il y a 35% de célibataires, 40% de mariés, 15% de divorcés et 10% de veufs. Tracer le camembert.

Solution : Il faudrait commencer par calculer l'angle correspondant à chaque modalité. Partant de la formule, nous présentons les angles dans un tableau :

Situation	f_i	α_i
Célibataires	35	126
Mariés	40	144
Divorcés	15	54
Veufs (ves)	10	36
Total	**100**	**360**

Ce sont ces angles qui nous aideront à tracer le camembert.



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

Application : La répartition des candidats convoqués pour participer au Test d'Admission à l'I.S.P. Mbanza-Ngungu pour l'année académique 2016-2017, selon le choix d'option, se présente comme suit :

Option	Effectif n_i
Informatique	250
Sciences exactes	200
Section Technique	300
Sciences humaines	150
Psychopédagogie	100
Total	**1000**

Travail : représentez cette distribution en Tuyaux d'orgues et Diagramme circulaire.

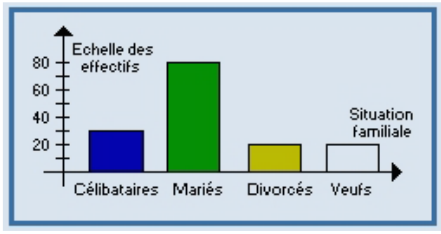
1.6.1.2. Le diagramme à tuyaux d'orgue

Cette représentation fait figurer les différentes modalités du caractère sous forme de rectangle ou de cylindres dont la base est constante et dont la hauteur est proportionnelle à l'effectif (ou à la fréquence).

Exemple : Considérons la situation familiale des 150 salariés de la *First Bank* représenté dans le tableau ci-après :

Situation	n_i
Célibataires	30
Marié(e)s	80
Divorcé(e)s	20
Veufs (ves)	20
Total	**150**

Représentons les données de ce tableau dans un graphique.



1.6.2. Cas quantitatif discret

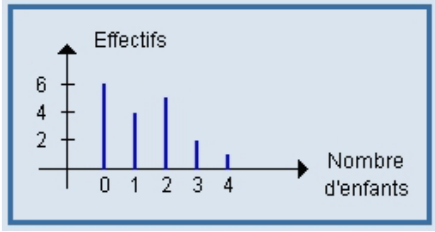
1.6.2.1. Le diagramme en bâtons

Les valeurs *discrètes* x_i prises par les variables sont placées sur l'axe des abscisses, et les effectifs (ou les fréquences) sur l'axe des ordonnées. La hauteur du bâton est proportionnelle à l'effectif.

Exemple : Une enquête sur le nombre d'enfants dans une famille menée auprès d'une population a révélé ce qui suit :

Nombre d'enfants	n_i
0	6
1	4
2	5
3	2
4	1
Total	**18**

Le diagramme en bâtons se présentera comme suit :



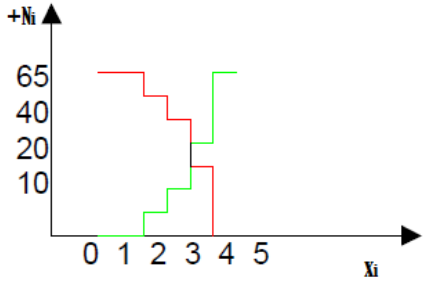
1.6.2.2. La courbe cumulative

La courbe cumulative est la représentation graphique des effectifs cumulés ou des fréquences cumulées.

C'est un graphique en escalier dont les paliers horizontaux ont pour ordonnées respectivement F_i ou N_i. Les marches de l'escalier correspondent aux valeurs possibles x_i de la variable statistique x et sont à des hauteurs proportionnelles aux effectifs cumulés ou aux fréquences cumulées.

Exemple : Soit la distribution du nombre d'enfants dans les ménages de Mbanza-Ngungu donnée par le tableau ci-après (x_i étant le nombre d'enfants et n_i le nombre de ménages) :

x_i	n_i	$+N_i$	$-N_i$
		0	65
1	5	5	60
2	10	15	50
3	30	45	20
4	20	65	0
Total	**65**		



Remarque : les deux courbes sont symétriques par rapport à un axe horizontal d'ordonnée *n*/2 pour les effectifs, 1/2 pour les fréquences.

- ✓ On utilise l'effectif (fréquence) cumulé croissant pour répondre aux questions du style : Quel est le nombre (%) d'individus dont la valeur du caractère est inférieure ou égale à x ?

- ✓ On utilise l'effectif (fréquence) cumulé décroissant pour répondre aux questions du style : Quel est le nombre (%) d'individus dont la valeur du caractère est strictement supérieure à x ?

1.6.3. Cas quantitatif continu

Comme pour les variables discrètes, il existe pour les variables statistiques continues deux types de représentations graphiques utilisés fréquemment.

1.6.3.1. L'histogramme

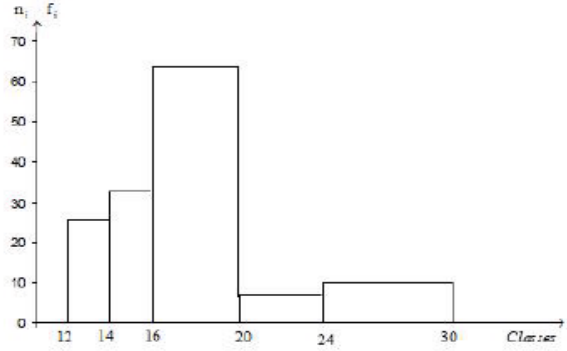
C'est un ensemble de rectangles contigus, chaque rectangle associé à chaque classe ayant une surface proportionnelle à l'effectif (fréquence) de cette classe.

Attention : Avant toute construction d'histogramme, il y a lieu de regarder si les classes sont d'amplitudes égales ou inégales.

Exemple : Considérons la répartition des ouvriers d'une entreprise suivant leur salaire mensuel net :

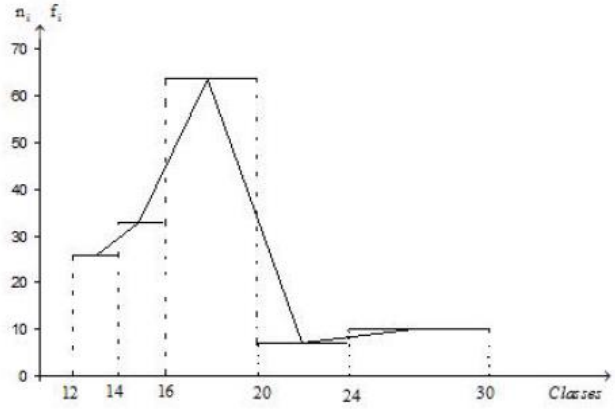
Classes de Salaire	n_i	f_i	N_i	F_i
12000 – 14000	26	0,186	26	0,186
14000 – 16000	33	0,235	59	0,421
16000 – 20000	64	0,458	123	0,879
20000 – 24000	7	0,050	130	0,929
24000 – 30000	10	0,071	140	1,000
Total	**140**	**1,000**		

Traçons l'histogramme des fréquences de cette distribution.



Le *polygone des fréquences* est la ligne brisée qui relie les milieux des cotés supérieurs des rectangles de l'histogramme des fréquences.

Exemple : Reprenons l'exemple de la répartition des ouvriers d'une entreprise suivant leur salaire mensuel net et traçons la courbe des fréquences et le polygone des fréquences de cette distribution.



1.6.3.2. La courbe cumulative

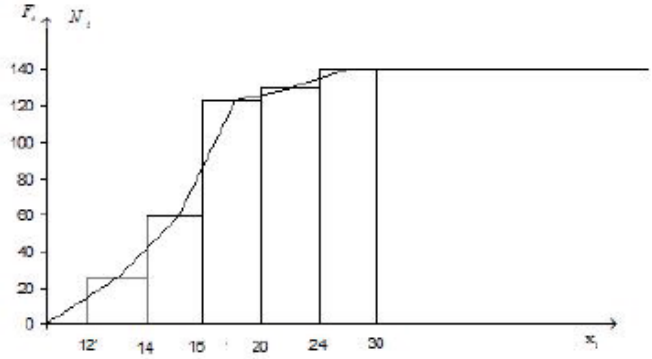
Comme pour les variables discrètes, la courbe cumulative ou histogramme des fréquences cumulées, est la représentation graphique de la fonction cumulative ou fonction de répartition *F(x)*.

Les observations étant regroupées en classes, on ne connait de cette fonction que les valeurs correspondant aux extrémités supérieures des classes :

$$F(e_i) = F_i \quad i = 1, 2, \ldots, k$$

Elle est estimée par le polygone des fréquences cumulées qui est la ligne brisée joignant les milieux des cotés supérieurs des rectangles de l'histogramme des fréquences cumulées.

Exemple : Reprenons l'exemple précédent et traçons l'histogramme des fréquences cumulées et le polygone des fréquences cumulées.



La notion de courbe des fréquences (ou la courbe des fréquences cumulées) découle de l'idée suivante : si les amplitudes des classes diminuent et si le nombre des observations est suffisamment grand pour éviter les irrégularités dues à la faiblesse des effectifs, alors l'histogramme des fréquences (l'histogramme des fréquences cumulées) tend, en tant que fonction en escalier, vers une courbe continue appelée courbe des fréquences (ou courbe des fréquences cumulées) et qui, à la limite, converge vers la densité de la distribution théorique (ou la fonction de répartition théorique) de la population.

Application : Représentez graphiquement la distribution de 50 étudiants en fonction de leur taille suivante :

Taille en cm x_i	155-160	160-165	165-170	170-175	175-180
n_i	6	12	16	14	2

SERIE D'EXERCICES
CHAP.02

Chapitre 3 : PARAMETRES STATISTIQUES

Après avoir collecter les différentes valeurs d'une variable dans un échantillon choisi, il est nécessaire de résumer les valeurs obtenues en quelques nombres appelés ***paramètres*** afin de les exprimer et de les utiliser dans la comparaison, l'estimation, etc.

Nous étudierons les paramètres de position, de dispersion et de concentration.

3.1. Paramètres de position

Ce type de paramètre sert à exprimer la position d'une distribution en fonction des valeurs associées à la variable étudiée.

3.1.1. Le mode

Le mode (ou la dominante) d'une distribution est la valeur la plus fréquente ou celui qui a le plus grand effectif.

Le mode peut ne pas exister, et s'il existe, il peut ne pas être unique :

- ✓ Quand on a un seul mode, c'est une distribution *uni-modale*.
- ✓ Quand on a deux modes, la distribution est dite *bimodale*.
- ✓ Quand on a trois modes, la distribution est *tri-modale*.
- ✓ …

Selon la nature de la variable quantitative (discrète ou continue), la manière de déterminer le mode n'est pas la même :

- ✓ *Dans le cas discret*, on peut déterminer la valeur du mode en regardant la fréquence (absolue ou relative) la plus élevée. Le mode étant la valeur de la variable qui a cette fréquence.

Nombre d'enfants	n_i
0	5
1	15
2	45
3	35
Total	**100**

On peut retrouver ce même résultat à partir du diagramme en bâtons.

- ✓ En revanche, dans le *cas continu*, puisque les données ont été regroupées en classes, on parlera plutôt de classe modale. A partir de la classe modale, on peut estimer une valeur modale. Pour y arriver, on utilise la formule suivante :

$$Mo = c_i^- + a_i \cdot \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right)$$

- c_i^- est la borne inférieure de la classe modale ;
- a_i est l'amplitude de la classe modale
- $\Delta_1 = n_i - n_{i-1}$ est la différence entre l'effectif de la classe modale et celui de la classe précédente ;
- $\Delta_2 = n_i - n_{i+1}$ est la différence entre l'effectif de la classe modale et celui de la classe suivante.

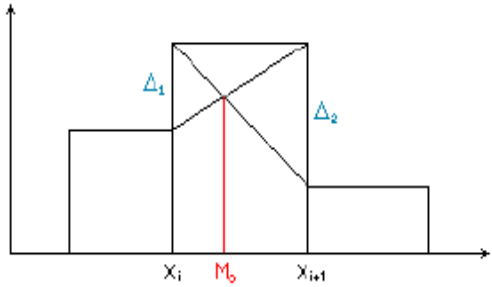
Exemple : Quel est le mode de la série suivante :

Classes	n_i	N_i
[142 - 150[	4	4
[150 - 158[	3	7
[158 - 166[	6	13
[166 - 174[	16	29
[174 - 182[	13	42
[182 - 190[	5	47
[190 - 198[	3	50
Total	**50**	

La classe modale est [166 - 174[$c_i^- = 166$ $a_i = 8$ $\Delta_1 = 16 - 6 = 10$ $\Delta_2 = 16 - 13 = 3$

$$Mo = 166 + 8 \cdot \left(\frac{10}{10 + 3} \right) = 172{,}15$$

Graphiquement, la méthode des diagonales donne la valeur du mode :



3.1.2. La médiane (M_e)

La médiane est la valeur qui partage la série, préalablement classée, en deux séries d'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.

Exemple : Soit la série statistique suivante : 15, 7, 22, 4, 12, 30, 9, 18, 6. Pour déterminer la médiane, il faut ordonner la série :

4 6 7 9 12 15 18 22 30

La médiane est 12 car dans cette série, il y a 4 nombres inférieurs et 4 supérieures de 12.

La médiane ne se calcule que pour les données quantitatives et sa logique de calcul dépend du type de données. On distinguera quatre cas :

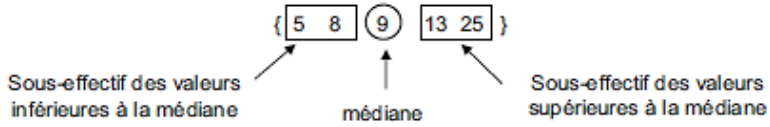
3.1.2.1. Effectif impair et aucune valeur n'est répétée

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 a autant de valeurs inférieures à la médiane 9 qu'il y a de valeurs supérieures à la médiane. L'effectif total est bien partagé en deux parties égales.

3.1.2.2. Effectif pair et aucune valeur n'est répétée

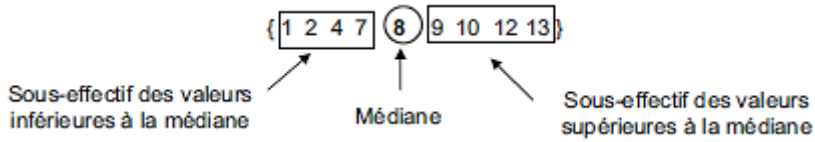
Quand l'effectif est pair, la médiane n'est pas une valeur de la série. Il faut la calculer.

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 la 4ème et la 5ème valeurs. La médiane est donc égale à la moyenne arithmétique simple de ces deux valeurs :

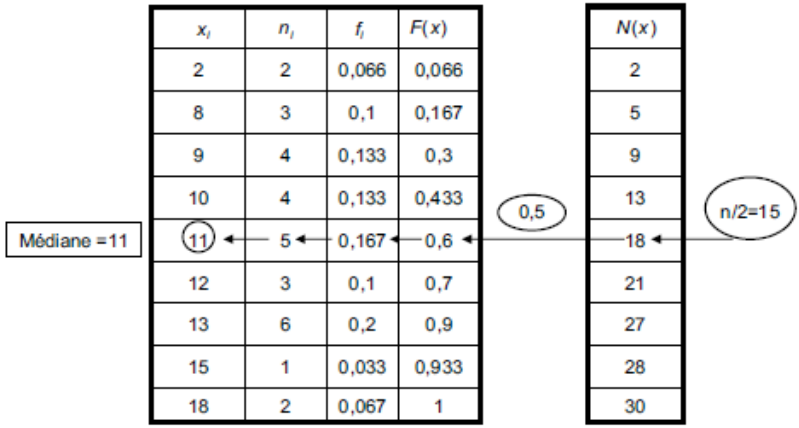
$$Me = (7+9)/2 = 8$$



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.

3.1.2.3. 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 ci-dessous, les valeurs de la variable *X* ont déjà été groupées. 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)*.



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

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 précédent.

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**.

N.B. : 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). On a perdu en termes de précision.

3.1.2.4. Effectifs groupés par classes de valeurs

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

$$Me = c_i^- + a_i \cdot \cdot \left(\frac{\frac{n}{2} - N_{i-1}}{n_i} \right)$$

Avec :

- n_i l'effectif de la classe médiane ;
- N_{i-1} l'effectif cumulé de la classe qui précède la classe médiane ;
- C_i^- la borne inférieure de la classe médiane ;
- a_i L'amplitude de classe (Borne Supérieur – Borne Inférieur)

Exemple : Soit la distribution de la bourse aux étudiants de l'I.S.P. Mbanza-Ngungu, de la première année de licence à la deuxième année de master (en milliers de francs congolais) :

Classe	n_i	N_i
[17 ; 22[	45	45
[22 ; 27[	22,5	67,5
[27 ; 32[	16	83,5
[32 ; 37[	9	92,5
[37 ; 42[	7,5	100
Total	**100**	

La fréquence cumulée 50 (N/2=100/2) se retrouve après l'effectif 45, c'est donc cet intervalle [22 ; 27[que l'on nomme classe médiane.

$$Me = c_i^- + a_i \cdot \cdot \left(\frac{\frac{n}{2} - N_{i-1}}{n_i} \right) \quad Me = 22 + 5 \cdot \left(\frac{\frac{100}{2} - 45}{22,5} \right) \quad Me = 22 + 5\left(\frac{5}{22,5} \right) = 23,11$$

Note : On peut aussi calculer la médiane en procédant à une interpolation linéaire.

De manière générale, si a et b sont les bornes de la classe contenant la médiane, F(a) et F(b) les valeurs de la fréquence cumulée croissante en a et b, càd la valeur de la fréquence cumulée de la classe médiane moins celle de la classe la précèdent, alors :

$$Me = a + (b - a) \times \frac{0,5 - F(a)}{F(b) - F(a)}$$

Application : Déterminez la valeur médiane de la distribution des tailles suivantes :

Taille en cm x_i	**155-160**	**160-165**	**165-170**	**170-175**	**175-180**
n_i	6	12	16	14	2

Ordonner les éléments du tableau :

Taille en cm x_i	n_i	f_i	F_i^+
155-160	6	0,12	0,12
160-165	12	0,24	0,36
165-170	16	0,32	0,68
170-175	14	0,28	0,96
175-180	2	0,04	1
Total	**50**	**1**	

0,5 se retrouve sur la classe [165 – 170[

$$Me = a + (b-a) \times \frac{0,5 - F(a)}{F(b) - F(a)} = 165 + (170 - 165) \times \frac{0,5 - 0,36}{0,68 - 0,36}$$

$$Me = 165 + 5 \times \frac{0,14}{0,32} = 167,1875$$

3.1.3. La moyenne ($\bar{X}$)

La moyenne représente la valeur centrale d'un ensemble de données, obtenue en distribuant équitablement la somme des valeurs observées (N_i) entre le nombre total d'observations (X_i).

Nous allons étudier quatre types de moyenne qui sont :

1. Moyenne Arithmétique
2. Moyenne Géométrique
3. Moyenne Quadratique
4. Moyenne Harmonique

3.1.3.1. La moyenne arithmétique

La moyenne arithmétique d'une série statistique est la somme des valeurs divisée par le nombre total des valeurs.

Exemple : Les 8 ouvriers d'une PME ont perçu en Mai 2022 les salaires suivants : 7500 ; 8300 ; 9100 ; 9600 ; 10700 ; 11300 ; 12000 ; 12500.

Le salaire moyen des ouvriers de cette PME est alors :

$$\bar{x} = \frac{7500 + 8300 + 9100 + 9600 + 10700 + 11300 + 12000 + 12500}{8} = 10125$$

Cette moyenne arithmétique simple se calcule par une formule qui est donnée par l'expression :

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

Lorsqu'aux valeurs sont affectées de coefficients (ici d'effectifs), on parle de la **moyenne pondérée**. Elle se calcule par :

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

Exemple : Calculer la moyenne de la distribution suivante :

x_i	n_i	$x_i n_i$
0	1	0
1	3	3
2	5	10
3	5	15
4	4	16
5	2	10
$\sum$	**20**	**54**

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i = \frac{54}{20} = 2{,}7$$

Lorsque l'on a une distribution par classes de valeurs, la moyenne se calcule en prenant la formule de la moyenne pondérée, mais "x_i" est remplacé par "c_i", **où c_i représente le *centre de la classe i***, c'est-à-dire la moyenne arithmétique des extrémités de classe.

Exemple : Calculer la moyenne de cette série classifiée :

Classes	n_i
[0 - 2[	4
[2 - 4[	10
[4 - 6[	6
Total	**20**

Pour calculer la moyenne, nous devons **déterminer les centres de classe**, puis faire la somme des $n_i c_i$ et diviser par *n*. Autrement dit, nous devons appliquer la formule :

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

On a donc le tableau de calcul suivant :

Classes	n_i	c_i	$n_i.c_i$
[0 - 2[	4	1	4
[2 - 4[	10	3	30
[4 - 6[	6	5	30
Total	**20**		**64**

Et finalement :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i c_i = \frac{64}{20} = 3{,}2$$

Nous avons donc une marge d'erreur non négligeable par rapport à la vraie moyenne, à savoir 2,7. La marge d'erreur en pourcentage est donnée par :

$$\frac{3{,}2 - 2{,}7}{2{,}7} \cdot 100 = 18{,}5\%$$

La marge d'erreur dépend de la définition des classes.

3.1.3.2. La moyenne géométrique (***G***)

La moyenne géométrique, notée *G*, est très utilisée dans l'analyse de l'évolution d'une variable dans le temps. Elle est donc utilisée pour tout ce qui concerne le calcul des taux de croissance.

Exemple : Taux de croissance du chiffre d'affaires d'une entreprise commerciale.

Tout comme avec la moyenne arithmétique, on distingue la moyenne géométrique simple de la moyenne géométrique pondérée.

Elles s'écrivent respectivement :

$$G = \prod_{i=1}^{n} x_i^{\frac{1}{n}} = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

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

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

Ou bien

$$G = \prod_{i=1}^{k} x_i^{f_i} = \sqrt[n]{\prod_{i=1}^{k} x_i^{n_i}}$$

Il peut arriver que l'on introduise la notion de logarithme pour calculer la moyenne géométrique.

$$\ln G = \sum_{i=1}^{k} f_i \ln x_i = \frac{\sum_{i=1}^{k} n_i \ln x_i}{N} \qquad G = \prod_i x_i^{f_i}$$

Pour la série statistique des pointures, nous obtenons G = 40,342 à partir du tableau ci-dessous :

x_i	n_i	$n_i \ln(x_i)$
38,0	100	363,75862
39,0	150	549,53425
40,0	250	922,21986
40,5	150	555,19530
41,0	120	445,62865
42,0	200	747,53392
44,0	30	113,52569
Total	**1000**	**3697,39628**

$$\ln G = \sum_{i=1}^{k} f_i \ln x_i = \frac{\sum_{i=1}^{k} n_i \ln x_i}{N} = \frac{3\,697,39628}{1\,000} = 3,69739628$$

$$\ln x = a \Leftrightarrow \log_e^x = a \quad \Rightarrow x = e^a \qquad \ln G = 3,69739628 \quad \Rightarrow G = e^{3,69739628}$$

Note : La valeur de « e » est **2,718281**

La moyenne géométrique s'utilise, par exemple, quand on veut calculer la moyenne de taux d'intérêt.

Exemple : Supposons que les taux d'intérêt pour 4 années consécutives soient respectivement de 5, 10, 15, et 10%. Que va-t-on obtenir après 4 ans si je place 100 francs ?

Solution :

- ✓ Après 1 an on a, 100 × 1.05 = 105
- ✓ Après 2 ans on a, 100 × 1.05 × 1.1 = 115.5
- ✓ Après 3 ans on a, 100 × 1.05 × 1.1 × 1.15 = 132.825
- ✓ Après 4 ans on a, 100 × 1.05 × 1.1 × 1.15 × 1.1 = 146.1075

Si on calcule la moyenne arithmétique des taux on obtient :

$$\bar{x} = \frac{1{,}05 + 1{,}10 + 1{,}15 + 1{,}10}{4} = 1{,}10$$

Si on calcule la moyenne géométrique des taux, on obtient :

$$G = \prod_{i=1}^{n} x_i^{\frac{1}{n}} = \sqrt[n]{\prod_{i=1}^{n} x_i} = \sqrt[4]{1{,}05 \cdot 1{,}10 \cdot 1{,}15 \cdot 1{,}10} = 1{,}099431377$$

Le bon taux moyen est bien *G* et non $\bar{x}$ car si on applique 4 fois le taux moyen *G* aux 100 francs, on obtient :

$$100\ Fr \times G^4 = 100 \times 1.099431377^4 = 146.1075$$

*3.1.3.3. La moyenne quadratique (**Q**)*

La moyenne quadratique (simple ou pondérée), notée Q, est très couramment utilisée en physique (tout comme la moyenne harmonique que nous allons définir ci-dessous).

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 :

$$Q = \sqrt{\frac{(-4)^2 + (-2)^2 + (0)^2 + (2)^2 + (4)^2}{5}} \approx 2{,}83$$

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 :

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

2) La moyenne quadratique pondérée

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

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

Exemple : Soit la distribution de 1000 étudiants habitant à Mbanza-Ngungu, suivant leurs pointures de chaussures.

x_i	n_i	f_i	F_i	X_i^2	$n_i x_i^2$
38,0	100	0,10	0,10	1 444,25	144400,00
39,0	150	0,15	0,25	1 521,25	228150,00
40,0	250	0,25	0,50	1 600,25	400000,00
40,5	150	0,15	0,65	1 640,25	246037,50
41,0	120	0,12	0,77	1 681,00	201720,00
42,0	200	0,20	0,97	1 764,00	352800,00
44,0	30	0,03	1,00	1 936,00	58080,00
Total	**1000**	**1,00**			**1631187,50**

$$Q = \sqrt{\frac{1}{1000} \cdot 1\,631\,187{,}50} = 40{,}388$$

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

*3.1.3.4. La moyenne harmonique (**H**)*

La moyenne harmonique _simple ou pondérée_ est utilisée dans le cadre de l'étude des rapports telle la vitesse, qui est exprimée en nombre de kilomètres parcourus par heure. Elle est égale à l'effectif total divisé par la moyenne de l'inverse des x_i.

En gros, on fait la moyenne des inverses, puis on prend l'inverse du résultat.

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 :

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

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

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

2) La moyenne harmonique pondérée

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

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

Il est judicieux d'appliquer la moyenne harmonique sur des vitesses.

Exemple : Un cycliste parcourt 4 étapes de 100km. Les vitesses respectives pour ces étapes sont de 10 km/h, 30 km/h, 40 km/h, 20 km/h. Quelle a été sa vitesse moyenne ?

Solution :

- ✓ Un raisonnement simple nous dit qu'il a parcouru la première étape en 10h, la deuxième en 3h20 la troisième en 2h30 et la quatrième en 5h. Il a donc parcouru le total des 400km en $10 + 3h20 + 2h30 + 5h = 20h50 = 20.8333h$, sa vitesse moyenne est donc :

$$Moy = \frac{400}{20{,}8333} = 19{,}2km/h$$

- ✓ Si on calcule la moyenne arithmétique des vitesses, on obtient :

$$\bar{x} = \frac{10 + 30 + 40 + 20}{4} = 25km/h$$

- ✓ Si on calcule la moyenne harmonique des vitesses, on obtient :

$$H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} = \frac{4}{\frac{1}{10} + \frac{1}{30} + \frac{1}{40} + \frac{1}{20}} = 19{,}2km/h$$

La moyenne harmonique est donc la manière appropriée de calculer la vitesse moyenne.

Pour la série statistique des pointures, nous obtenons **H** = 40,319 (1000/24,80199) à partir du tableau ci-dessous :

x_i	n_i	n_i / x_i
38,0	100	2,63158
39,0	150	3,84615
40,0	250	6,25000
40,5	150	3,70370
41,0	120	2,92683
42,0	200	4,76190
44,0	30	0,68182
Total	**1000**	**24,80199**

Notons que, quelle que soit la forme de la distribution de la variable statistique étudiée, il est admis que les inégalités suivantes sont toujours vérifiées :

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

3.2. Paramètres de dispersion

Les paramètres de dispersion absolue indiquent de combien les valeurs d'une distribution s'éloignent de la valeur centrale de référence (comme la moyenne ou la médiane). Un paramètre de dispersion absolue s'exprime toujours dans l'unité de mesure de la variable considérée.

Les quatre paramètres de dispersion absolue les plus courants sont **l'étendue**, **l'intervalle interquartiles**, **l'écart absolu moyen** et **l'écart type**.

3.2.1. L'étendue ou l'intervalle de variation (IV)

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

$$IV = x_{\max} - x_{\min}$$

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

Élève A : {8, 9, 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.

3.2.2. L'intervalle interquartile (IIQ)

L'intervalle interquartile est une mesure de la variation qui n'est pas influencée par les valeurs extrêmes, contrairement à l'intervalle de variation. Il indique l'étalement des données autour de la médiane.

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 première 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 IIQ, l'intervalle interquartile par la formule :

$$IQ = Q_3 - Q_1$$

Pour la détermination pratique des quartiles, on peut se référer à celle de la médiane – ce sont les mêmes procédés. On notera d'ailleurs que $Q_2 = Me$.

$$Q_1 = c_i^- + a_i\left(\frac{\frac{n}{4} - N_{i-1}}{n_i}\right)$$

$Q_2 = Me$

$$Q_3 = c_i^- + a_i\left(\frac{\frac{3n}{4} - N_{i-1}}{n_i}\right)$$

Exemple : La série ordonnée par ordre croissant S a 12 termes :

$$S = \{11,\ 12,\ 13,\ 15,\ 16,\ 16,\ 17,\ 17,\ 18,\ 19,\ 20,\ 22\}$$

Nous pouvons faciliter la lecture de ce tableau comme suit :

Rang	1	2	3	4	5	6	7	8	9	10	11	12
Série	11	12	13	15	16	16	17	17	18	19	20	22

Un quart (25%) des données correspond à : **12** × 0,25 = 3. Le premier quartile est alors la plus petite valeur Q_1 pour laquelle les valeurs de 3 termes de la série sont inférieures ou égales à Q_1. Le premier quartile est donc la valeur du 3ème terme de la série, c'est-à-dire 13.

Trois quarts (75%) des données correspondent à : 12 × 0,75 = 9. Le troisième quartile est alors la plus petite valeur Q_3 pour laquelle les valeurs de 9ème termes de la série sont inférieures ou égales à Q_3. Le troisième quartile est donc la valeur du 9ème terme de la série, c'est-à-dire 18.

L'intervalle interquartile est [13 ; 18]. L'écart interquartile est : 18-13 = 5.

3.2.3. Le diagramme en boîte (ou boîte à moustaches)

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.

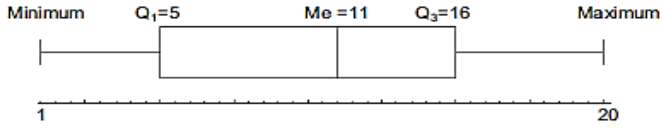
{4, 13, 17, 7, 1, 3, 9, 14, 12, 20, 16, 15, 11, 6, 5}

Nous savons que Me = 11, Q_1= 5 et Q_3= 16 pour les avoir calculés précédemment. 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 :

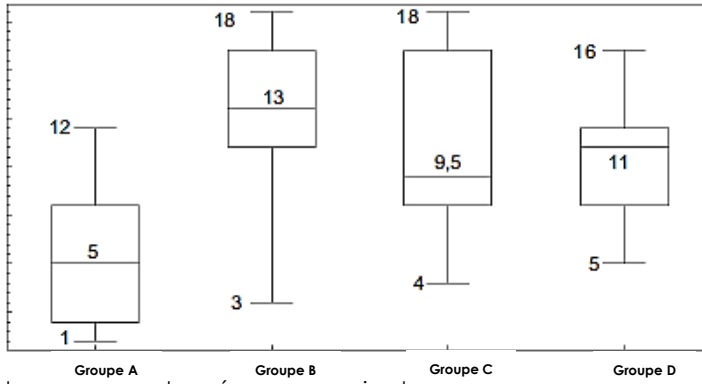


La boîte à moustache permet de comparer des séries du point de vue de leur dispersion mais aussi de leur caractéristique de tendance centrale (puisque la médiane est repérée).

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).



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

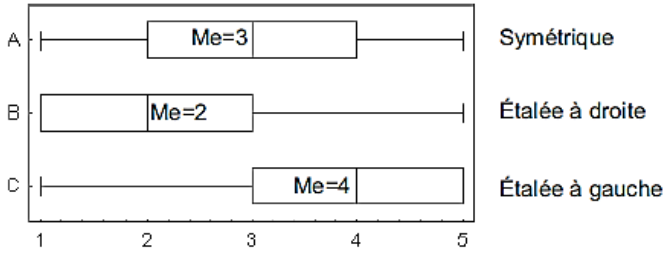
1) Si la médiane est proche du centre de la boîte, c'est que la distribution est symétrique.
2) Si la médiane est à gauche du centre de la boîte, c'est que la distribution est étalée à droite.
3) Si la médiane est à droite du centre de la boîte, c'est que la distribution est étalée à gauche.

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 symétriques (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, 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}



3.2.4. L'écart absolu moyen ($\bar{e}$)

Soit x une variable statistique pouvant prendre les *k* valeurs x_1, x_2, …, x_k auxquelles correspondent les effectifs respectifs n_1, n_2, …, n_k. L'écart absolu moyen, noté $\bar{e}$, est alors la moyenne arithmétique des valeurs absolues des écarts à la moyenne arithmétique.

$$\bar{e} = \frac{1}{n}\sum_{i=1}^{n} n_i \left| x_i - \bar{x} \right|$$

L'écart absolu moyen est minimum lorsqu'on prend les écarts par rapport à la médiane.
Exemple : Soit les âges d'un groupe de 5 enfants {3, 7, 8, 5, 12} :

Moyenne = (3 + 7 + 8 + 5 + 12) / 5 = 7

Écarts à la moyenne :

1. | 3 – 7 | = 4
2. | 7 – 7 | = 0
3. | 8 – 7 | = 1
4. | 5 – 7 | = 2
5. | 12 – 7 | = 5

$\bar{e}$ = (4 + 0 + 1 + 2 + 5) / 5 = 12 / 5 = 2,4 ans

En moyenne, les âges s'écartent de 2,4 ans par rapport à 7 ans.

3.2.5. La variance et l'écart-type

La variance, l'écart-type et le coefficient de variation 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.

3.2.5.1. La variance

1) Définition

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

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{k} n_i (x_i - \bar{x})^2$$ si l'effectif considéré est celui d'une population.

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{k} n_i (x_i - \bar{x})^2$$ si l'effectif considéré est celui d'un échantillon

Ainsi que nous l'avons déjà indiqué au début de ce cours, sauf mention contraire explicite, nous ne considérons que des populations. Par conséquent, la première formule sera utilisée dans la suite.

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

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

Autrement dit, lorsque les données sont connues individuellement ou qu'elles ne se répètent pas, c'est cette dernière formule qui s'applique. En revanche, lorsque les données sont groupées par valeurs, c'est la première formule 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.

2) Mode de calcul de la dernière formule

Pour calculer la variance lorsque les données ne se répètent pas, 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 :

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
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

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{10} x_i = 8{,}2 \qquad \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{237{,}6}{10} = 23{,}76$$

3) Mode de calcul de la formule « développée »

La première formule peut aussi être calculée suivant la méthode précédente. Toutefois, pour faciliter les calculs, il est préférable d'utiliser la formule dite « développée » :

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

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 :

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 :

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
Σ	**26**	**208**		**2156**

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

$$\sigma^2 = \frac{1}{26}\sum_{i=1}^{5} n_i x_i^2 - \bar{x}^2 = \frac{1}{26}\cdot 2156 - (8)^2 = 18{,}9231$$

Exemple 2 : Calculer la variance de la série suivante :

Classes	n_i
2 – 3	2
3 – 4	4
4 – 5	6
5 – 6	5
6 – 7	3
Total	**20**

Classes	n_i	c_i	$n_i c_i$	$n_i c_i^2$	$c_i - \bar{x}$	$(c_i - \bar{x})^2$	$n_i (c_i - \bar{x})^2$
2 – 3	2	2,5	5,0	12,50	-2,15	4,6225	9,2450
3 – 4	4	3,5	14,0	49,00	-1,15	1,3225	5,2900
4 – 5	6	4,5	27,0	121,50	-0,15	0,0225	0,1350
5 – 6	5	5,5	27,5	151,25	0,85	0,7225	3,6125
6 – 7	3	6,5	19,5	126,75	1,85	3,4225	10,2675
Total	**20**		**93,0**	**461**			**28,55**

$$V(x) = \frac{1}{n}\sum_{i=1}^{n} n_i x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} n_i x_i\right)^2 = \frac{461}{20} - \left(\frac{93}{20}\right)^2 = 1{,}42$$

$$\bar{x} = \frac{93}{20} = 4{,}65 \qquad V(x) = \frac{1}{n}\sum_{i=1}^{n} n_i (x_i - \bar{x})^2 = \frac{28{,}55}{20} = 1{,}42$$

4.1.5.2. L'écart-type

L'écart-type est égal à la racine carrée de la variance. Sa définition est donnée par la formule :

$$\sigma = \sqrt{\sigma^2}$$

De façon générale :

- ✓ *si l'écart-type est faible*, cela signifie que les valeurs sont assez concentrées autour de la moyenne ;
- ✓ *si l'écart-type est élevé*, cela veut dire au contraire que les valeurs sont plus dispersées autour de la moyenne.

3.3. Paramètres de concentration

La mesure de la concentration revient à celle de la conséquence de la dispersion. Très importante en économie (concentration des salaires, des revenus, de la taille des entreprises, …), elle concerne des variables continues ne pouvant prendre que des valeurs positives.

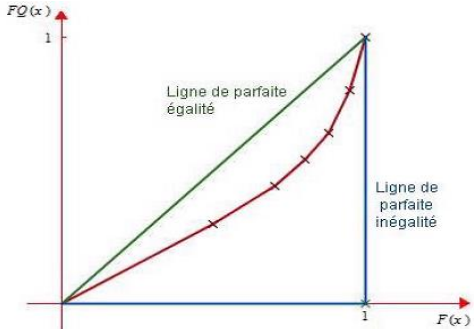
On la détermine par deux méthodes : le calcul et les graphes.

3.3.1. La courbe de concentration

L'objectif est de mesurer les inégalités dans la répartition d'une variable à l'intérieur d'une population.

La courbe de concentration exige comme pour la médiale, la connaissance pour chaque classe du nombre d'observations et de la somme des valeurs correspondantes.

Soit F_i la fréquence cumulée des observations et FQ_i le pourcentage cumulé de la somme des valeurs. Alors, la courbe de concentration est obtenue en traçant le graphe de FQ_i (ordonnée) en fonction de F_i (abscisse). On obtient alors la courbe suivante :



La courbe de concentration ou courbe de Lorenz est notamment utilisée en économie pour mesurer les inégalités de possession de richesse (on supposera donc que x représente un certain bien possédé par les individus de la population). Elle est fabriquée de la façon suivante :

Soit x_i une valeur prise par x. On note $F(x)$ la proportion de la population pour laquelle $x < x_i$ (F est donc la courbe cumulative (fonction de répartition) de x). On note $FQ(x_i)$ la proportion du bien possédé par ces individus par rapport au bien total. Alors la courbe de Lorenz est la courbe joignant tous les points $(F(x_i), FQ(x_i))$. La courbe de Lorenz joint donc toujours le point (0, 0) au point (1, 1). Elle est située sous le segment joignant ces deux points. La diagonale du carré circonscrit à la courbe de Lorenz s'appelle droite d'équi-répartition.

Remarque : La diagonale principale du graphique (droite d'équi-répartition) représente une distribution parfaitement égalitaire. Plus la courbe de concentration s'écarte de la droite d'équi-répartition, plus la distribution est inégalitaire. D'autre part, plus la dispersion est faible plus la courbe de concentration s'aplatit sur la diagonale.

3.3.2. L'indice de Gini ou indice de concentration

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 1/2) :

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

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 :

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

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 milliers de dollars). La troisième colonne indique les centres de classe.

Gains	n_i	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 :

$|x_i - x_j| n_i n_j$

$|1{,}5 - 0{,}75| 1 \times 2$

	x_i	0,75	1,5	2,5	3,5	4,5	∑
x_j	n	**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
∑	**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 :

Gains	n_i	c_i	$n_i c_i$
[0,5 -1[	1	0,75	0,75
[1-2[	2	1,5	3
[2-3[	6	2,5	15

[3-4[	4	3,5	14
[4-5[	2	4,5	9
Total	**15**		**41,75**

$$\bar{x} = \frac{1}{15}.41{,}75 = 2{,}78333$$

Par conséquent :

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

Et donc :

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

Exercice : Calculer l'indice de Gini pour la répartition des employés d'une entreprise selon leur salaire mensuel net, donnée par le tableau suivant :

Salaires en $	n_i
[800 ; 900[	25
[900 ; 1000[	30
[1000 ; 1100[	28
[1100 ; 1500[	25
[1500 ; 2000[	10
Total	**118**

SERIE D'EXERCICES
CHAP.03

Partie II : STATISTIQUE BIVARIEE

Chapitre 4 : SERIES A DEUX VARIABLES

Pour l'étude de certains phénomènes complexes, il s'avère insuffisant de prendre en compte un seul caractère. Alors il en faut considérer deux caractères ou plus. L'analyse et la représentation des tableaux statistiques obtenus deviennent évidemment plus complexes.

La représentation graphique, par exemple, n'est possible que dans un espace à trois dimensions au plus. En définissant les distributions marginales et conditionnelles, on peut ramener la représentation d'une distribution à plusieurs dimensions à quelques représentations unidimensionnelles. Dans la suite, on ne considérera que les séries statistiques à deux dimensions.

4.1. Présentation générale d'un tableau à deux dimensions

Le tableau qui suit 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.

	y_1	y_2	…	y_j	…	y_l	Σ
x_1	n_{11}	n_{12}	…	n_{1j}	…	n_{1l}	$n_{1\bullet}$
x_1	n_{21}	n_{22}	…	n_{2j}	…	n_{2l}	$n_{2\bullet}$
. . .	. . .	. . .		… … …		. . .	. . .
x_i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{il}	$n_{i\bullet}$
. . .	. . .	. . .		. . .		. . .	. . .
x_k	n_{k1}	n_{k2}	…	n_{kj}	…	n_{kl}	$n_{k\bullet}$
Σ	$n_{\bullet 1}$	$n_{\bullet 2}$	…	$n_{\bullet j}$	…	$n_{\bullet l}$	$n_{\bullet\bullet}$

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

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

Intervalle d'âge	**H**	**F**
[0 -18[	10	20
[18 - 45]	5	15

En reprenant la notation du tableau 4 on a ici :

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

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

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

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

$$n_{\bullet\bullet} = n_{1\bullet} + n_{2\bullet} = 30 + 20 = 50$$

$$n_{\bullet\bullet} = n_{\bullet 1} + n_{\bullet 2} = 15 + 35 = 50$$

4.2. Distributions marginales

4.2.1. Notion

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

Intervalle d'âge	**H**	**F**	$n_{i\bullet}$
[0 -18[	10	20	**30**
[18 - 45]	5	15	**20**
$n_{\bullet j}$	**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 suivant donne les fréquences marginales de x et de y dans le cas du tableau précédent :

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

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

Fréquences marginales de x : $f_{i\bullet} = \frac{n_{i\bullet}}{n_{\bullet\bullet}}$

Fréquences marginales de y : $f_{\bullet j} = \frac{n_{\bullet j}}{n_{\bullet\bullet}}$

4.2.2. Moyennes et variances marginales

4.2.2.1. Moyennes marginales

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

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

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

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 :

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 :

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

Ainsi que la moyenne marginale de y :

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

5.2.2.2. Variances marginales

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

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

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

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

x_i	$n_{i\bullet}$	x_i^2	$n_{i\bullet} x_i^2$
2	8	4	32
8	16	64	1024
			1056

y_j	$n_{\bullet j}$	y_j^2	$n_{\bullet j} y_j^2$
1	7	1	7
4	17	16	272
			279

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

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

4.3. 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 précédent, 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 suivant 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.

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

- ✓ 3 et 4 représentent la distribution conditionnelle de x quand y = 1
- ✓ 5 et 12 représentent la distribution conditionnelle de x quand y=4

Exemple 2 : Toujours en prenant les chiffres du même tableau, 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 suivant 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*.

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

- ✓ 3 et 5 représentent la distribution conditionnelle de y quand x=2
- ✓ 4 et 12 représentent la distribution conditionnelle de y quand *y=8*

4.4. Moyennes et variances conditionnelles

4.4.1. 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 :

- ✓ $\overline{x}_1$ pour désigner la moyenne conditionnelle de x quand y = 1
- ✓ $\overline{x}_2$ 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 y quand x = 8

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

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

$$\bar{x}_2 = \frac{1}{17}\left[(5\times 2)+(12\times 8)\right] = 6{,}2353$$

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 de notre tableau

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

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

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

$$\bar{y}_j = \frac{1}{n_{i\bullet}}\sum_{j=1}^{q} n_{ij}y_i \qquad 1 \le i \le q$$

4.4.2. Variances conditionnelles

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

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

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

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

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

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

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

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

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

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

SERIE D'EXERCICES
CHAP.04

Chapitre 5 : REGRESSION ET CORRELATION

En présence d'une distribution statistique de deux variables (X, Y), il est possible d'étudier les distributions marginales, les distributions conditionnelles, mais cette étude ne fournit pas d'interprétation des résultats.

Dans certains cas, nous pouvons nous poser la question suivante : La connaissance d'une modalité de la variable X apporte-t-elle une information supplémentaire sur les modalités de la variable Y ?

La réponse à cette question est du domaine de la **régression** : dans un tel cas, on dit que *X* est la *variable explicative* et *Y* la *variable expliquée*.

Dans d'autres cas, aucune des deux variables ne peut être privilégiée : la liaison stochastique entre X et Y s'apprécie alors de façon symétrique par la mesure de la **corrélation**.

Ainsi, lorsqu'on observe deux variables quantitatives sur les mêmes individus, on peut s'intéresser à une liaison éventuelle entre ces deux variables.

- ✓ La régression fournit une expression de cette liaison sous la forme d'une fonction mathématique.
- ✓ La corrélation renseigne sur l'intensité de cette liaison.

Exemple : *X* est le prix du litre de carburant à la pompe, *Y* est le prix d'une course en taxi dans la cité de Mbanza-Ngungu.

Dans cet exemple, *X* est la variable explicative et *Y* la variable expliquée. Il est à noter qu'une variable explicative X peut être une variable qualitative.

5.1. La régression linéaire

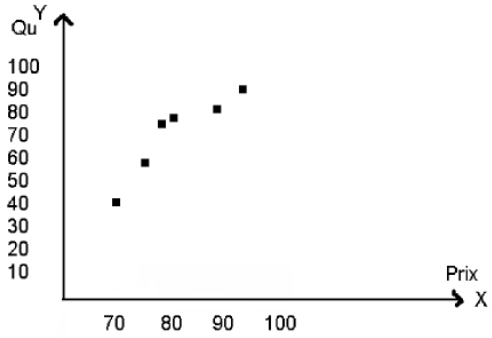
5.1.1. Présentation du problème

Soit le tableau suivant :

P \ Q	42	51	60	62	72	83	Total
70	1						1
75		1					1
77			1				1
80				1			1
86					1		1
93						1	1
Total	1	1	1	1	1	1	6

Ce tableau est un tableau de contingence ou les observations sont connues individuellement, on peut présenter plus simplement ce tableau de la manière suivante :

Prix	Qté
70	42
75	57
77	60
80	62
86	74
93	83
Total	



Nous avons un ensemble de points « un nuage statistique » qui nous indique que les prix et les quantités évoluent selon la même tendance.

Il est possible de schématiser ce nuage par une fonction simple : la fonction linéaire (Droite) dont les paramètres (points) sont inconnus et qu'il faudra trouver :

- ✓ *a* est la pente de la droite
- ✓ *b* est l'ordonnée à l'origine

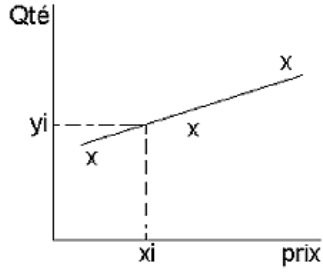
Une telle droite est appelée droite de régression D(x) et a le coefficient de régression. La régression est le fait de relier *y* à x par une fonction. Les paramètres de la droite de régression sont à calculer.

5.1.2. La méthode des moindres carrés ordinaires MCO

Ce nom vient du fait que la méthode consiste à déterminer la droite d'ajustement *en minimisant la somme du carré des écarts* entre cette droite et les observations.

5.1.2.1. Notion

Partons d'un nuage statistique théorique :



- ✓ Il s'agit de résumer ce nuage par une droite.
- ✓ Soit $y' = ax + b$ l'équation de la droite recherchée.
- ✓ Pour toute valeur de x (x_i) nous avons une valeur réellement observée y'.
- ✓ Pour toute valeur x_i, nous avons une valeur calculée sur la droite y'.
- ✓ Pour toute une valeur x_i, on a une erreur d'estimation égale à $| y_i - y'_i |$.
- ✓ La droite de régression idéale doit être de telle manière que la somme des erreurs d'estimation doit être la plus faible possible, | yi – y'i | doit être minimum.
- ✓ Pour éviter les valeurs absolues, on convient de calculer les carrés des erreurs. La droite de régression doit être telle que :
  $\Sigma (yi - y'i)^2$ minimum, et on appelle cela la condition des moindres carrés.

5.1.2.2. Calcul des paramètres de la droite de régression

Ces deux coefficients ***a*** et ***b*** sont calculés en appliquant les formules suivantes :

$$a = \frac{\text{cov}(x, y)}{\sigma_x^2} \qquad b = \hat{y} - a\bar{x}$$

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

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

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

$$a = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{xy}}{\frac{1}{n}\sum_{i=1}^{n} x^2 - (\bar{x})^2}$$

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 :

X	***Y***	***XY***	***X²***	***Y²***
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
29	**29,5**	**160,9**	**175**	**156**

Calculons *a* :

$$a = \frac{\frac{1}{6} \times 160{,}9 - \left(\frac{29}{6} \times \frac{29{,}5}{6}\right)}{\frac{1}{6} \times 175 - \left(\frac{29}{6}\right)^2} = 0{,}5258$$

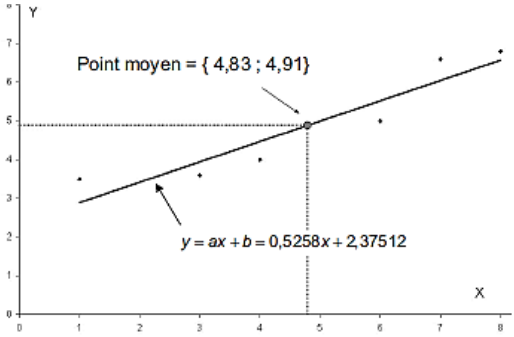
Une fois *a* connu, on en déduit *b* :

$$b = \bar{\bar{y}} - a\bar{\bar{x}} = \left(\frac{29{,}5}{6}\right) - 0{,}5258\left(\frac{29}{6}\right) = 2{,}37512$$

L'équation de la droite de régression est donc :

$$y = ax + b = 0{,}5258\,x + 2{,}37512$$

La figure ci-dessous illustre l'équation de cette droite. Nous vérifions à nouveau que cette droite passe par le point moyen.



Exemple 2 : Soit les ventes mensuelles et le prix unitaire correspondant. Estimer l'équation de la régression de ces données :

Mois	Qté (q_i)	Prix (p_i)
Janvier	5	14
Février	15	6
Mars	9	10
Avril	14	9
Mai	3	11
Juin	9	13

Juillet	10	9
Août	17	6
Septembre	11	5
Octobre	16	3
Novembre	7	11
Décembre	3	15
Total	**119**	**112**

Solution : Ce sont les prix qui influencent la vente (la quantité est la variable expliquée X et le prix est la variable expliquée Y) :

$$a = -0,6426 \qquad b = 15,7058$$

5.1.2.3. 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 premier 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 sa valeur dans l'équation de la droite pour obtenir *Y* :

$$y = 0,5258\ (12) + 2,37512 = 8,6847$$

Exemple 2 : Supposons maintenant 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 sa valeur dans l'équation de la droite pour obtenir *X* :

$$5 = 0,5258x + 2,37512 \Leftrightarrow x = 4,99212 = 5$$

5.2. La corrélation linéaire

5.2.1. 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 :

$$r = \frac{\mathrm{cov}(x, y)}{\sigma_x \sigma_y}$$

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 :

$$r = \frac{\text{cov}(x,y)}{\sigma_x \sigma_y} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{\bar{x}}\bar{\bar{y}}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} x^2 - \left(\bar{\bar{x}}\right)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n} y^2 - \left(\bar{\bar{y}}\right)^2}}$$

$$r = \frac{\text{cov}(x,y)}{\sigma_x \sigma_y} = \frac{\frac{160,9}{6} - \frac{29}{6} \times \frac{29,5}{6}}{\sqrt{\frac{175}{6} - \left(\frac{29}{6}\right)^2}\sqrt{\frac{156}{6} - \left(\frac{29,5}{6}\right)^2}} = 0,9371$$

5.2.2. 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 au coefficient de corrélation, qui varie entre -1 et +1, le coefficient de détermination 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

SERIE D'EXERCICES
CHAP.05

Chapitre 6 : LES INDICES STATISTIQUES

Pour l'étude de certains phénomènes économiques et sociaux, on est souvent amené à décrire ou à comparer les variations de grandeurs simples telles que le prix du blé, la production d'acier ou le taux de fécondité d'une certaine population, etc.

Pour les comparaisons dans le temps et dans l'espace de ces grandeurs, on introduit la notion d'indice statistique élémentaire. Ce sont généralement des rapports de ces grandeurs. Mais il est plus instructif de pouvoir suivre les évolutions de grandeurs plus complexes telles que le niveau général des prix, la production industrielle, le volume des importations, etc. Ces évolutions sont résumées par l'une ou l'autre des caractéristiques de tendance centrale de la série des indices élémentaires correspondants. On parle dans ce cas d'indices synthétiques.

6.1. Les indices élémentaires

Exemples :

a) Le prix du kilogramme d'un certain produit a été de 15$ en moyenne en 1990 et il est de 32$ en Octobre 2008. L'indice élémentaire du prix de ce produit en Octobre 2008, base 100 en 1990, est le rapport des deux prix exprimé en pourcentage :

$$I_{Oct98/Moy80} = \frac{32}{15} \cdot 100 = 213{,}33$$

b) La consommation d'électricité a été de 16500 Millions de Kwh en 1998 et de 6200 Millions de Kwh en 1983. L'indice élémentaire de la consommation d'électricité en 1998, base 100 en 1983, est le rapport des consommations des deux années exprimé en % :

$$I_{1998/1983} = \frac{16500}{6200} \cdot 100 = 266{,}13$$

Plus généralement, considérons la variation dans le temps d'une grandeur simple X, prenant les valeurs $X_0, X_1, \ldots, X_t,$ aux dates (ou périodes) successives 0, 1, 2, …, t.

On appelle indice élémentaire de la grandeur X à la date (ou période) t par rapport à la date (ou période) 0, le rapport :

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

Remarque : La date ou période 0 est appelée date de référence ou base de l'indice. La date ou période t est appelée date courante. En général, ce rapport est exprimé en % tel que :

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

On dit alors que l'indice à la date t est exprimé en base 100 à la date de référence 0.

Les indices statistiques élémentaires sont utilisés surtout pour retracer l'évolution des grandeurs simples dans le temps. Mais ils peuvent aussi servir à des comparaisons dans l'espace.

Exemple : La densité de la population congolaise a été de 14,6 h/Km^2 en 1996, alors que pour la capitale Kinshasa elle a été de 1540 h/Km^2. L'indice de densité de la capitale, l'ensemble de la R.D.C. étant choisi comme base, est :

$$I_{Kin/RDC} = \frac{1540}{14,6} \cdot 100 = 10580$$

L'indice de densité du Kongo Central dont la densité de la population est de 0,5 h/Km^2, par rapport à celle du pays, est alors :

$$I_{KC/RDC} = \frac{0,5}{14,6} \cdot 100 = 3,4$$

Propriétés : Les indices élémentaires possèdent les trois propriétés de réversibilité et de circularité (transitivité) :

- ✓ La **circularité** :

$$i_{t/0} = i_{t/t'} \cdot i_{t'/0}$$

En effet,

$$\frac{x_t}{x_0} = \frac{x_t}{x_{t'}} \cdot \frac{x_{t'}}{x_0}$$

La circularité est une propriété fondamentale qui permet de comparer non seulement les dates 0 et t d'une part, 0 et t' d'autre part, mais aussi t et t'.

$$i_{t/t'} = \frac{i_{t/0}}{i_{t'/0}}$$

Exemple 2 : Prix de gros du cobalt.

1962 : 404,90 $ les 100 kg
1965 : 494,58 $ les 100 kg
1967 : 687,19 $ les 100 kg

Les indices élémentaires du prix du cuivre, base 100 en 1962 sont :

$$I_{65/62} = 100 \times \frac{494,58}{404,90} = 122,15$$

$$I_{67/62} = 100 \times \frac{687,19}{404,90} = 169,72$$

$$I_{67/65} = 100 \times \frac{687,19}{494,58} = 138,94$$

Mais on peut aussi calculer l'indice $I_{67/65}$ en ignorant les prix aux différentes périodes :

$$I_{67/65} = 100 \times \frac{i_{67/62}}{i_{65/62}} = 100 \times \frac{1,6972}{1,2215} = 138,94$$

$$I_{62/67} = 100 \times \frac{1}{i_{67/62}} = 100 \times \frac{1}{1,6972} = 58,92$$

On a utilisé la propriété de réversibilité.

- ✓ **La réversibilité**

$$i_{0/t} = \frac{1}{i_{t/0}}$$

En effet,

$$\frac{x_0}{x_t} = \frac{1}{\frac{x_t}{x_0}}$$

Cette propriété est intéressante dans le cas de comparaison géographique, car le choix du lieu de référence est arbitraire.

Remarque : L'évolution d'un phénomène est souvent présentée sous forme d'une augmentation ou d'une diminution en pourcentage à l'aide de la formule suivante :

$$\frac{\text{V aleur nouvelle - V aleur primitive}}{\text{V aleur primitive}} \times 100$$

Le pourcentage de variation ne possède pas les propriétés de circularité et de réversibilité des indices, et est donc moins maniable. Les pourcentages de variation ne se rajoutent pas.

En effet, on appelle pourcentage de variation le nombre :

$$PV_{t/o} = \frac{x_t - x_o}{x_o} \times 100$$
$$= \left(\frac{x_t}{x_o} - 1 \right) \times 100$$
$$= 100 \left(i_{t/o} - 1 \right) = I_{t/o} - 100$$

L'évolution d'une variable est souvent présentée sous forme d'une augmentation ou diminution en pourcentage.

- $PV > 0$: augmentation de la variable.
- $PV < 0$: diminution de la variable.

Exemple 1 : Le prix de gros du cuivre est passé de 494,58 F les 100 kg en 1962 à 687,19 F les 100 kg en 1967.

$$PV_{67/62} = \left(\frac{687{,}19 - 404{,}90}{404{,}90} \right) \times 100$$
$$= 100 \cdot \left(i_{67/62} - 1 \right)$$
$$= 100 \cdot (1{,}6972 - 1)$$
$$= 69{,}72 \ \%$$

Ce qui correspond bien à l'indice $I_{67/62} = 169{,}72$

Les pourcentages de variation ne possèdent pas les propriétés de circularité et de réversibilité des indices élémentaires. De plus, les pourcentages de variation ne s'ajoutent pas.

Exemple 2 : Prix d'un bien quelconque

Novembre	1993	200 FC
Novembre	1994	250 FC
Novembre	1995	400 FC

- Augmentation de novembre 1993 à novembre 1994

$$=100\left(\frac{250\ -\ 200}{200}\right)=25\,\%$$

- Augmentation de novembre 1994 à novembre 1995

$$=100\left(\frac{400\ -\ 250}{250}\right)=60\,\%$$

- Augmentation de novembre 1993 à novembre 1995

$$=100\left(\frac{400\ -\ 200}{200}\right)=100\,\%$$

Et non pas : 60 +25 = 85 %

6.2. Les indices synthétiques

Les grandeurs complexes sont fonction de quelques grandeurs simples. Ainsi le niveau général des prix est constitué des prix des divers aliments et boissons, du logement, de l'équipement ménager, de l'habillement, des services médicaux, des transports, des loisirs, etc. La construction d'un indice synthétique relatif à la variation d'une grandeur complexe consiste à résumer une série d'indices élémentaires.

Un **indice synthétique** est un indice qui résume l'évolution de plusieurs grandeurs : plusieurs prix, plusieurs quantités, plusieurs valeurs (prix - quantités), etc.

6.2.1. Position du problème

Soit X une grandeur complexe composée des éléments $X_1, X_2, \ldots, X_j, \ldots, X_h$. La variable complexe X est, par exemple, le niveau général des prix, et $X_1, X_2, \ldots, X_j, \ldots, X_h$ représentent les prix des différents produits ou services offerts au public. Les indices élémentaires des constituants X_j, j = 1, 2, …, h, de X sont calculés par la formule $I_{t/0}^{j} = \frac{x_t^j}{x_0^j}$ j = 1, 2, …, h. Mais cette suite d'indices n'apporte aucune information sur l'évolution du niveau général des prix. Il serait judicieux de les résumer ou de les synthétiser par un seul indice qu'on appellera indice synthétique de la grandeur complexe X.

6.2.2. Les différentes formules d'indices synthétiques

Trois formules d'indices synthétiques sont utilisées en pratique. Ces sont les formules de LASPEYRES, de PAASCHE et de FISHER.

6.2.2.1. *Indice 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 :

$$L_{t/0} = \frac{\sum_{i=1}^{n} p_t q_0}{\sum_{i=1}^{n} p_0 q_0} \times 100$$

Exemple : Soit le tableau ci-après, qui donne les prix et les quantités de deux produits 1 et 2, aux dates 0 et t. On peut supposer que le produit 1 est un pantalon et le produit 2 un T-shirt. Calculer l'indice de LASPEYRES de prix.

	Date 0		Date t	
Produit 1	p_0=15	q_0=3	p_t=22	q_t=10
Produit 2	p_0=7	q_0=9	p_t=5	q_t=8

$$L_{t/0} = \frac{\sum_{i=1}^{n} p_t q_0}{\sum_{i=1}^{n} p_0 q_0} \times 100 = \frac{(22 \times 3) + (5 \times 9)}{(15 \times 3) + (7 \times 9)} \times 100 = 102{,}8$$

Dans notre exemple, le prix du bien 1 a augmenté (de 15 à 22) 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 2,8%.

2) L'indice de LASPEYRES des quantités

Cet indice 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 :

$$L_{t/0} = \frac{\sum_{i=1}^{n} p_0 q_t}{\sum_{i=1}^{n} p_0 q_0} \times 100$$

On voit ainsi que si les quantités ne changent pas entre 0 et t (c'est-à-dire si $q_t = q_0$), l'indice synthétique de LASPEYRES des quantités demeure égal à 100.

Exemple : reprenons le tableau précédent, qui donne les prix et les quantités de deux produits 1 et 2, aux dates 0 et t :

	Date 0		Date t	
Produit 1	p0=15	q0=3	pt=22	qt=10
Produit 2	p0=7	q0=9	pt=5	qt=8

Dans cet exemple, la quantité du bien 1 augmente (de 3 à 10 unités) tandis que celle du bien 2 baisse (de 9 à 8 unités). Pour savoir si l'indice synthétique des volumes augmente ou baisse, appliquons la formule de LASPEYRES d'évolution des quantités :

$$L_{t/0} = \frac{\sum_{i=1}^{n} p_0 q_t}{\sum_{i=1}^{n} p_0 q_0} \times 100 = \frac{(15\times 10)+(7\times 8)}{(15\times 3)+(7\times 9)} \times 100 = 190{,}74$$

On enregistre donc une évolution des volumes du panier de bien de 90,74 % selon la formule de LASPEYRES.

6.2.2.2. *Indice 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 :

$$P_{t/0} = \frac{\sum_{i=1}^{n} p_t q_t}{\sum_{i=1}^{n} p_0 q_t} \times 100$$

Exemple : Soit le tableau 2, qui donne les prix et les quantités de deux produits 1 et 2, aux périodes 0 et t.

	Période 0		Période t	
Produit 1	$p_0=10$	$q_0=4$	$p_t=14$	$q_t=8$
Produit 2	$p_0=6$	$q_0=12$	$p_t=5$	$q_t=9$

Calculons l'indice de PAASCHE des prix :

$$L_{t/0} = \frac{\sum_{i=1}^{n} p_t q_t}{\sum_{i=1}^{n} p_0 q_t} \times 100 = \frac{(14\times 8)+(5\times 9)}{(10\times 8)+(6\times 9)} \times 100 = 117{,}16$$

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 :

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

Calculons l'indice de PAASCHE des quantités à partir du tableau précédent :

$$P^q{}_{t/0} = \frac{\sum_{i=1}^{n} p_t q_t}{\sum_{i=1}^{n} p_t q_0} \times 100 = \frac{(14\times 8)+(5\times 9)}{(14\times 4)+(5\times 12)} \times 100 = 135{,}34$$

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).

6.2.2.3. *Indice 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, lesquels représentent 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 :

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

Reprenons notre exemple et calculons l'indice de FISHER des prix :

$$F_{t/0} = \sqrt{L_{t/0} \times P_{t/0}} = \sqrt{103{,}57 \times 117{,}16} = 110{,}16$$

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.

$$F(q)_{t/0} = \sqrt{L(q)_{t/0} \cdot P(q)_{t/0}}$$

Exemple : Calculons les indices de valeur, de LASPEYRES, de PAASCHE et de FISHER pour 1991 par rapport à 1990 sur l'ensemble des quatre produits décrits ci-dessous :

Produits	1990		1991	
	Prix p_0	Quantités q_0	Prix p_1	Quantités q_1
A	9,00	27	9,25	37
B	4,90	31	5,20	40
C	3,65	40	5,00	28
D	8,10	15	7,70	30

1990 : période 0 j = 1 à n
1991 : période t n = 4

Les calculs sont effectués à l'aide du tableau suivant :

	j	$p_0 q_0$	$p_0 q_1$	$p_1 q_0$	$p_1 q_1$

A	1	243,00	333,00	249,75	342,25
B	2	151,90	196,00	161,20	208,00
C	3	146,00	102,20	200,00	140,00
D	4	121,50	243,00	115,50	231,00
	Total	662,40	874,20	726,45	921,25

- Indice de LASPEYRES des prix

$$L(p)_{t/0} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100 = \frac{726,45}{662,40} \times 100 = 109,67$$

- Indice de LASPEYRES des quantités

$$L(q)_{t/0} = \frac{\sum p_0 q_1}{\sum p_0 q_0} \times 100 = \frac{874,20}{662,40} \times 100 = 131,97$$

- Indice de PAASCHE des prix

$$P(p)_{t/0} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100 = \frac{921,25}{874,20} \times 100 = 105,38$$

- Indice de PAASCHE des quantités

$$P(q)_{t/0} = \frac{\sum p_1 q_1}{\sum p_1 q_0} \times 100 = \frac{921,25}{726,45} \times 100 = 126,81$$

- Indice de FISHER des prix

$$F(p)_{t/0} = \sqrt{L(p)_{t/0} . P(p)_{t/0}} = \sqrt{109,67 \times 105,38} = 107,50$$

- Indice de FISHER des quantités

$$F(q)_{t/0} = \sqrt{L(q)_{t/0} . P(q)_{t/0}} = \sqrt{131,97 \times 126,81} = 129,36$$

SERIE D'EXERCICES
CHAP.06