RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE MINISTÈRE DE L ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITÉ MOULOUD MAMMERI DE TIZI-OUZOU

Faculté des Sciences Economiques, Commerciales et des Sciences de Gestion

Département des Sciences Economiques

Polycopié de cours
destiné aux étudiants de première année (LMD)

Statistique 1

Réalisé par :

OULD ABDESLAM-HAMAZ Sabrina

Objectif du cours

Le cours de statistique descriptive vise à initier l'étudiant en sciences économiques aux outils fondamentaux d'analyse des données quantitatives. Il a pour objet principal la collecte, l'organisation, la présentation et la synthèse de l'information statistique, en vue de décrire et de mieux comprendre les phénomènes économiques observés.

Face à la multiplication des données dans les domaines de l'économie, de la gestion ou des politiques publiques, la statistique descriptive fournit une méthode rigoureuse pour traiter l'information. Elle permet de transformer une masse de données brutes en indicateurs lisibles et utiles à la prise de décision. Ce cours constitue donc un socle méthodologique essentiel pour tout étudiant amené à interpréter des chiffres, construire des raisonnements appuyés sur des faits, ou encore valider des hypothèses à partir d'observations concrètes.

L'enseignement couvre plusieurs dimensions : la classification des variables, la construction de tableaux et de graphiques, le calcul et l'interprétation des paramètres statistiques (moyenne, médiane, mode, étendue, écart-type, etc.), ainsi que la mesure des formes et des concentrations (asymétrie, indice de Gini, courbe de Lorenz). L'objectif n'est pas seulement de maîtriser des techniques, mais de comprendre le sens économique des résultats obtenus.

Ce cours prépare également le terrain pour des disciplines plus avancées, telles que la statistique inférentielle, l'économétrie ou l'analyse de données. Il développe chez l'étudiant des capacités d'observation, de synthèse et d'analyse critique indispensables à tout travail quantitatif. Par son approche progressive et appliquée, il favorise une compréhension active des données économiques réelles issues des enquêtes, rapports ou bases statistiques des institutions (ONS, FMI, Banque mondiale, etc.).

En somme, la statistique descriptive constitue un langage commun entre les faits économiques et leur interprétation rationnelle. Elle est le point de départ de toute démarche scientifique en économie et un outil indispensable pour former des acteurs capables de lire, comprendre et utiliser les chiffres dans un monde de plus en plus piloté par les données.

Plan du cours :

Sommaire

Introduction générale

La statistique, en tant que discipline, ne date pas d'hier. Son origine remonte à l'Antiquité, où elle servait essentiellement à des fins de recensement des populations, de collecte d'informations fiscales ou militaires dans les sociétés égyptienne, chinoise ou romaine. Cependant, ce n'est qu'à partir du 17ème siècle que la statistique commence à prendre une forme plus structurée, notamment avec les travaux de John Graunt sur les registres de mortalité à Londres, posant les bases de la statistique démographique.

Au fil des siècles, la statistique s'est enrichie, professionnalisée et diversifiée, jusqu'à devenir aujourd'hui un outil incontournable dans presque tous les domaines : économie, gestion, médecine, ingénierie, sciences sociales, etc. À l'ère du numérique et des mégadonnées, elle joue un rôle stratégique dans la prise de décision, l'analyse des phénomènes et l'interprétation des tendances.

Ce cours de statistique descriptive constitue une première initiation à cet univers. Il vise à fournir aux étudiants les outils de base nécessaires pour organiser, représenter et résumer des données de manière claire et rigoureuse. Contrairement à la statistique inférentielle, qui cherche à tirer des conclusions à partir d'échantillons, la statistique descriptive se limite à l'observation et à l'analyse des données disponibles, sans extrapolation.

Le contenu du cours s'articule autour de plusieurs axes fondamentaux :

- **La classification et la typologie des variables statistiques** : qualitatives ou quantitatives, discrètes ou continues ;
- **Les outils de représentation des données** : tableaux, diagrammes, graphiques ;
- **Les mesures de tendance centrale et de dispersion** : moyenne, médiane, mode, variance, écart-type, etc. ;
- **Les indices statistiques**, qui permettent de mesurer l'évolution d'un phénomène dans le temps, notamment en économie (ex. : indices des prix, indices de volume ou de valeur) ;
- **L'analyse des relations entre deux variables** à travers l'étude de la **corrélation** et de la **régression linéaire**, qui visent à explorer les liens éventuels entre deux grandeurs statistiques.

Au-delà des calculs, l'objectif est aussi de développer chez l'étudiant une capacité d'interprétation et d'analyse critique des données. En effet, la statistique n'est pas seulement un ensemble de formules : c'est un langage de la réalité, un moyen structuré d'organiser l'information pour mieux la comprendre et l'exploiter.

Ce polycopié propose une approche progressive, illustrée et contextualisée, afin de permettre une assimilation claire des concepts, tout en montrant leurs applications pratiques dans la vie réelle ou dans des disciplines spécifiques.

Chapitre 1 : Vocabulaire statistique et définition des concepts de base

Introduction :

La statistique constitue un outil essentiel pour comprendre, analyser et interpréter les phénomènes observables dans de nombreux domaines : sciences sociales, économie, santé, environnement, etc. Pour rendre les données collectées compréhensibles, comparables et exploitables, il est nécessaire de disposer d'un langage commun et rigoureux : le **vocabulaire statistique**. Ce premier chapitre a pour but de présenter les concepts fondamentaux qui constituent la base de toute analyse statistique.

1. Population et unité statistique :

C'est l'ensemble des éléments sur lesquels porte une étude statistique. Par exemple ; l'ensemble des étudiants de première année, l'ensemble de personnes dans un ménage, etc. Les éléments qui constituent la population sont appelés unités statistiques ou individus. Ils peuvent être des êtres humains, des objets ou des événements, c'est-à-dire physiques ou immatériels. Ces individus sont de même nature et forment un ensemble homogène. La population statistique doit être définie avec précision car cela conditionne fortement l'étude statistique et ses résultats.

En pratique, lorsqu'il est difficile d'étudier toute la population (du fait de sa taille ou de contraintes logistiques), on utilise un échantillon**,** c'est-à-dire un sous-ensemble représentatif de la population.

2. Caractère

C'est la propriété possédée par les unités statistiques permettant de les décrire et de les distinguer les unes des autres. Par exemple ; l'âge des étudiants de première année, la couleur des voitures dans un parking, le nombre de pièces par logement dans une ville, etc. Chaque individu ou unité statistique peut être décrite selon un ou plusieurs caractères.

3. Modalités

Ce sont les diverses situations ou possibilités pouvant être prises par le caractère. Chaque caractère peut présenter deux (02) ou plusieurs modalités. Par contre, chaque individu n'appartient qu'à une et une seule modalité (un individu ne peut pas avoir en même temps 18 ans et 20 ans à la fois).

4. Les différents types de caractères

Un caractère peut être de type qualitatif ou de type quantitatif.

4.1. Le caractère qualitatif

Un caractère est qualifié de qualitatif lorsque ses modalités ne peuvent pas être mesurées ou quantifiées. Les modalités de ce caractère sont généralement exprimées à l'aide de mots ou de numéros sous forme d'étiquettes, de codes, etc.
Il est important de noter que les opérations arithmétiques ne peuvent pas être appliquées aux valeurs d'un caractère qualitatif, car elles conduiraient à des résultats dépourvus de sens et irrationnels.

4.2. Le caractère quantitatif

Un caractère est considéré comme quantitatif lorsqu'il est possible de quantifier ou de mesurer ses différentes valeurs (modalités), ce qui signifie qu'il peut être exprimé sous forme de mesures ou de quantités numériques. En conséquence, les modalités de ce caractère se traduisent toujours par des données numériques (chiffres), ce qui permet d'effectuer des opérations arithmétiques sur elles et d'obtenir des résultats rationnels.

C'est avec ce type de caractère que la notion de ***variable statistique*** prend tout son sens mathématique, et ses modalités sont les valeurs possibles de la variable. Ainsi, l'âge, la taille, le poids, la durée, le nombre d'enfants par ménage, le nombre d'étudiants par salle,… ; sont des caractères quantitatifs. Les caractères quantitatifs ou les variables statistiques sont de deux natures :

4.2.1. La variable statistique discrète (VSD)

Lorsque les modalités d'une variable statistique reflètent un *dénombrement* ou un *comptage*, c'est-à-dire désignent *le nombre de* quelque chose, on dit qu'il s'agit d'une variable statistique *discrète* ou *discontinue*. Ses modalités sont, généralement, exprimées sous forme de nombres entiers ou isolés appartenant à l'ensemble des *nombres naturels* (Ñ), reflétant des réalités indivisibles. Exemple ; le nombre d'enfants par ménage, le nombre d'étudiants par salle, le nombre de pièces par logement, etc.

4.2.2. La variable statistique continue (VSC)

Une variable statistique est considérée comme continue lorsque ses valeurs peuvent appartenir à l'ensemble des nombres réels (R), ce qui signifie qu'elle peut prendre une gamme infinie de valeurs dans un intervalle. Elle peut également être exprimée sous forme d'intervalles ou de classes. Par conséquent, à l'exception des situations de dénombrement, toutes les opérations de mesure, telles que la pesée, la mesure de longueur, le chronométrage, le calcul, etc., sont des exemples de caractères quantitatifs continus.

Conclusion :

Ce premier chapitre a posé les bases essentielles de l'analyse statistique en définissant les notions clés du vocabulaire statistique. Comprendre la différence entre population et échantillon, identifier les unités statistiques, distinguer les types de caractères (qualitatifs ou quantitatifs) ainsi que leurs modalités, sont autant d'éléments fondamentaux pour interpréter correctement les données. Cette terminologie rigoureuse constitue un langage commun indispensable pour éviter toute ambiguïté dans la collecte, l'organisation et l'analyse des données. La maîtrise de ces concepts permettra d'aborder avec clarté et précision les méthodes statistiques plus avancées présentées dans les chapitres suivants.

Chapitre 2 : Présentation des distributions statistiques

Introduction :

Les données collectées, dans leur forme brute, sont souvent difficiles à exploiter directement. Pour en faciliter l'analyse et l'interprétation, il est essentiel de les organiser et de les traiter de manière rigoureuse. En statistique, les tableaux et les graphiques constituent les principaux outils de présentation. Les tableaux offrent une vue structurée et synthétique des informations, facilitant la compréhension des données par le lecteur. Les graphiques, quant à eux, permettent une visualisation rapide et intuitive, rendant la lecture des tendances et des relations beaucoup plus accessibles.

Section 1 : Les tableaux statistiques

Une fois les données arrangées et triées, elles sont généralement présentées de manière synthétique dans un tableau statistique. Par définition, un tableau statistique vise à offrir une représentation claire et concise du phénomène étudié. Il met en évidence la variable observée, ses différentes modalités (souvent disposées dans la première colonne), ainsi que les effectifs et/ou les fréquences associés. Ainsi, le tableau permet de visualiser les couples (xi ; ni) ou (xi ; fi), facilitant l'analyse quantitative et comparative des données.

1.1 Structure d'un tableau statistique

Quel que soit le type de variable ou la nature des données observées, le tableau statistique obéit à une structure générale uniforme. Il respecte un principe de présentation standardisé qui facilite la lecture et l'interprétation. En règle générale, un tableau statistique de base se compose d'un agencement précis permettant de représenter les modalités d'une variable ainsi que les mesures associées, telles que les effectifs ou les fréquences correspondants.

Tableau n°… : « Intitulé du tableau et date des données……. »

(Unité de mesure)

Caractère (Xi)	x_1 x_2 . . x_k	**Total**
Effectifs (ni)	n_1 n_2 . . n_k	**N**
Fréquences (fi)	f_1 f_2 . . f_k	**1**

Source :

Dans le cadre d'un travail de recherche scientifique, la présentation d'un tableau statistique doit impérativement respecter une structure rigoureuse. Cette structure standardisée vise à garantir la lisibilité, la cohérence et la transparence des données. Un tableau statistique bien construit comprend généralement les éléments suivants :

- **Titre du tableau :** Il fournit une description synthétique du contenu, permettant au lecteur de saisir rapidement la nature des informations représentées.

- **Numérotation** : Chaque tableau se voit attribuer un numéro unique, ce qui facilite sa citation dans le texte et améliore la navigation dans le document.
- **En-têtes de colonnes** : Ces intitulés précisent le contenu de chaque colonne, apportant des détails complémentaires au titre pour une meilleure compréhension des données.
- **Unité de mesure** : Elle permet de savoir dans quelle unité les valeurs sont exprimées (par exemple : kilogrammes, pourcentages, dinars…), ce qui est essentiel pour interpréter correctement les résultats.
- **Période ou date de référence** : Mentionner la période de collecte ou l'année des données est fondamental, car de nombreuses variables évoluent avec le temps. Cette précision contribue à la validité et à la pertinence de l'analyse, tout en témoignant de la rigueur méthodologique du chercheur.
- **Source des données** : L'indication de la provenance des données renforce la crédibilité du travail en permettant au lecteur de vérifier leur fiabilité. Elle illustre également le respect des principes d'intégrité et de transparence scientifique.

1.2 Elaboration du tableau statistique : La construction d'un tableau statistique varie en fonction de la nature de la variable étudiée, qu'elle soit qualitative ou quantitative.

1.2.1 Cas d'un caractère qualitatif : Lorsqu'il s'agit d'un caractère qualitatif, le tableau statistique prend la forme d'une **nomenclature**. Dans ce type de présentation, chaque modalité de la variable est désignée sous le terme de **rubrique**. Ce format permet de classer les données selon des catégories non numériques, facilitant ainsi l'analyse descriptive des informations recueillies.

Tableau n°1 : « Répartition des étudiants de première année section B selon leur sexe »

Sexe (xi)	Féminin	Masculin	**Total**
Effectifs (ni)	7	3	**10**
Fréquences (fi)	0.7 (7/10)	0.3 (3/10)	**1**

1.2.2 Cas d'un caractère quantitatif : Une variable quantitative peut être soit **discrète**, soit **continue**, selon la nature de ses valeurs.

a. Présentation d'une variable statistique discrète (VSD)

Une variable est dite discrète lorsque l'ensemble des valeurs qu'elle peut prendre est fini ou dénombrable, généralement constitué de nombres entiers ou de valeurs distinctes. Contrairement à la présentation utilisée pour les variables qualitatives, le tableau statistique associé à une variable discrète intègre une colonne supplémentaire dédiée aux effectifs ou fréquences cumulés, car cette information permet une meilleure interprétation des distributions numériques. Par ailleurs, les modalités doivent impérativement être classées par ordre croissant, depuis la

première jusqu'à la dernière ligne (ou colonne), afin de garantir une lecture logique et structurée des données.

Tableau n°2 : « Répartition des ménages d'une ville X selon le nombre de personnes »

Nombre de personnes (xi)	1	2	3	4	5	6	**Total**
Nombre de ménages (ni)	15	35	25	10	5	20	**110**
Fréquences (fi)	0.14	0.32	0.23	0.09	0.04	0.18	**1**
Fréquences cumulées croissantes (fi↑)	0.1	0.46	0.69	0.78	0.82	1	**-**
Fréquences cumulées décroissantes (fi↓)	1	0.86	0.54	0.31	0.22	0.18	**-**
Effectifs cumulés croissants (ni ↑)	15	50	75	85	90	110	**-**
Effectifs cumulées décroissants (ni ↓)	110	95	60	35	25	20	**-**

Comment faire des lectures à partir du tableau d'une variable statistique discrète ?

- 85 ménages, soit 78 %, ont moins de 5 personnes, ou au plus 4 personnes.

- 90 ménages, soit 82 %, ont moins de 6 personnes, ou au plus 5 personnes.

- 60 ménages, 54%, ont plus de 2 personnes, ou au moins 3 personnes.

- 25 ménages, soit 22%, ont plus de 4 personnes, ou au moins 5 personnes.

b. Présentation d'une variable statistique continue (VSC) : Lorsqu'il s'agit d'une variable continue, les données sont regroupées en classes ou intervalles afin de faciliter leur traitement statistique. Contrairement au tableau d'une variable discrète, celui-ci comporte une colonne supplémentaire correspondant aux centres des classes (xi), qui représentent une approximation des valeurs contenues dans chaque intervalle. Comme pour les variables discrètes, l'ordre croissant doit être respecté, en commençant par la borne inférieure de la première classe jusqu'à la borne supérieure de la dernière, afin d'assurer une organisation cohérente des données.

Tableau n°3 : « Répartition des salaires mensuels en 10^3 DA des employés d'une entreprise »

Salaires (en 10^3 DA)	[40-50[	[50-60[	[60-70[	[70-80[	[80-90[	**Total**
Centres de classes (xi)	45	55	65	75	85	**-**
Nombre d'employés (ni)	15	5	20	7	3	**50**
Fréquences (fi)	0.3	0.1	0.4	0.14	0.06	**1**
Fréquences cumulées croissantes (fi↑)	0.3	0.4	0.8	0.94	1	**-**

Fréquences cumulées décroissantes (fi↓)	1	0.7	0.6	0.2	0.06	-
Effectifs cumulés croissants (ni↑)	15	20	40	47	50	-
Effectifs cumulées décroissants (ni↓)	50	35	30	10	3	-

Comment faire des lectures à partir du tableau d'une variable statistique continue ?

- 40employés, soit 80%, touchent un salaire inférieur à 70.000DA.

- 30 employés, soit 60%, touchent un salaire supérieur ou égal à 60.000 DA (ou 60.000DA et plus).

1.2.3 Construction des classes en statistique : principes et règles fondamentales

Dans l'analyse des données continues, les observations sont réparties en groupes appelés classes, chacun représentant une plage de valeurs. Ces classes permettent de synthétiser l'information en segments homogènes, facilitant ainsi l'organisation et l'interprétation des résultats. Chaque classe correspond à un intervalle délimité par deux bornes : une inférieure et une supérieure. En statistique descriptive, ces intervalles sont souvent définis comme semi-ouverts à droite (ex. : [a ; b[), afin d'éviter les recoupements entre classes adjacentes.

a. Amplitude et centre de classe : L'amplitude d'une classe représente la largeur de l'intervalle, c'est-à-dire la différence entre la borne supérieure et la borne inférieure :

$$a_i = b_{i+1} - b_i$$

Étant donné que chaque classe couvre une infinité de valeurs (puisqu'on travaille avec des données continues), il est courant d'en donner une représentation simplifiée à travers le centre de classe. Celui-ci est défini comme la moyenne des deux bornes :

$$X = \frac{bi+b_{i+1}}{2}$$

Ce centre permet de représenter l'ensemble des modalités contenues dans la classe de manière synthétique, notamment lors du calcul de moyennes ou d'autres indicateurs statistiques.

b. Détermination du nombre de classes : Le choix du nombre de classes est une étape essentielle dans la construction d'un tableau pour données continues. Il doit permettre une lecture fluide, tout en offrant une granularité suffisante pour refléter les variations significatives du phénomène observé.

Même si ce choix dépend souvent de l'objectif de l'analyse et de l'expérience du statisticien, on recommande en pratique de ne pas dépasser une fourchette comprise entre 6 et 12 classes, afin de concilier précision et lisibilité.

Voici deux méthodes couramment utilisées pour estimer ce nombre à partir de la taille de l'échantillon N :

- Règle de STURGES : Z = 1+3,33(logN).
- Racine carrée de l'effectif : Z = $\sqrt{N}$.

c. Harmonisation des classes et étendue de la série : Lorsque l'on choisit une amplitude constante pour toutes les classes, il devient possible d'établir un lien direct entre l'étendue de la série **(**e), le nombre de classes (Z) et l'amplitude (a) :

$$a = \frac{e}{Z} = \frac{Xmax - Xmin}{Z}$$

Il est également impératif que la valeur minimale de la série soit incluse dans la première classe, et que la valeur maximale figure dans la dernière. Cela garantit une couverture complète des données sans créer de classes artificielles non représentatives.

Section 2 : Les représentations graphiques

Les représentations graphiques constituent une traduction visuelle des données issues d'un tableau statistique, reposant sur les couples (xi ; ni). Elles permettent de saisir, d'un simple regard, la structure globale d'une distribution statistique ou les tendances d'un phénomène observé. Chaque type de variable (ou caractère) admet un ou plusieurs modes de représentation graphique adaptés à sa nature.

2.1 Représentation graphique d'un caractère qualitatif

Les caractères qualitatifs peuvent être représentés graphiquement de plusieurs façons. Parmi les méthodes les plus répandues, on retrouve notamment le diagramme circulaire et les diagrammes en barres (aussi appelés "tuyaux d'orgue"). Ces outils facilitent la visualisation des répartitions catégorielles.

2.1.1 Diagramme circulaire

Également désigné sous les termes de diagramme en secteurs ou camembert, ce type de graphique repose sur le principe de proportionnalité, appliqué dans un système de coordonnées polaires. Il consiste à répartir un cercle de 360° entre les différentes modalités du caractère étudié, en fonction de leurs fréquences relatives.

Chaque modalité est représentée par un secteur angulaire dont la mesure (α_i) est proportionnelle à sa fréquence (f_i), selon la formule suivante :

$$\boldsymbol{\alpha_i = (n_i / N) \cdot 360 = f_i \cdot 360}$$

2.1.2 Le diagramme en barres (tuyaux d'orgue)

Ce type de graphique associe à chaque modalité d'une variable qualitative un rectangle vertical dont la largeur est constante et la hauteur proportionnelle à l'effectif ou à la fréquence correspondante. L'axe des abscisses présente les différentes modalités, sans valeur numérique propre, tandis que l'axe des ordonnées indique les effectifs ou fréquences.

2.2- La représentation graphique d'un caractère quantitatif

Les variables à caractère quantitatif, qu'elles soient discrètes ou continues, s'interprètent souvent plus aisément à travers des représentations graphiques adaptées à leur nature. Ces outils visuels facilitent la lecture des distributions numériques et permettent de dégager des tendances générales d'un simple regard.

2.2.1 - Variables quantitatives discrètes (VSD)

Lorsqu'une variable ne prend que des valeurs isolées (entières ou discontinues), on parle de variable discrète. Sa représentation graphique se fait principalement selon deux modalités :

- A travers un diagramme en bâtons pour mettre en valeur la fréquence ou l'effectif de chaque modalité ;
- A l'aide d'une courbe cumulative par paliers (ou en escalier), utile pour observer la répartition progressive des données.

a. Le diagramme en bâtons

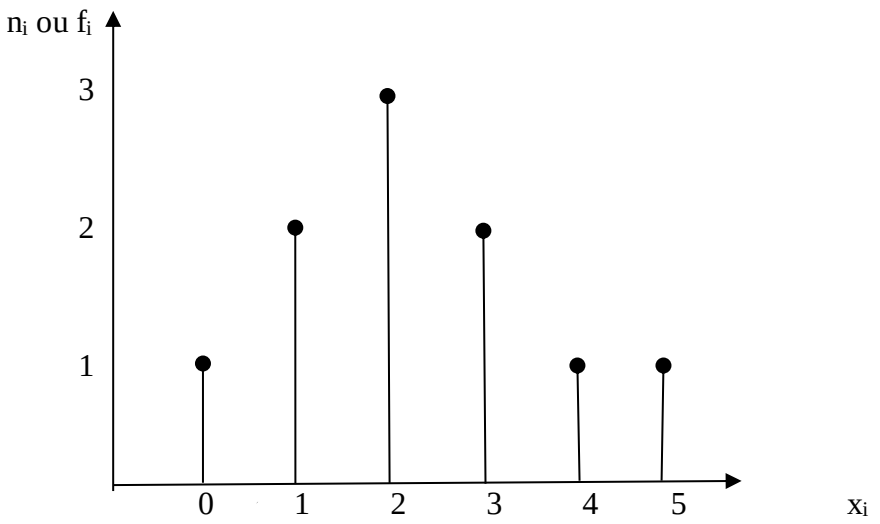
Dans ce type de graphique, chaque valeur de la variable est représentée par un segment vertical (ou "bâton") dont la hauteur traduit l'importance de cette valeur en termes d'effectif ou de fréquence. Les différentes modalités sont positionnées horizontalement sur l'axe des abscisses, tandis que les effectifs ou fréquences sont portés sur l'axe des ordonnées.

Exemple :

*Le tableau suivant donne l*a distribution du nombre de visites quotidiennes enregistrées à l'accueil d'un centre culturel sur une période de 10 jours :

- *Représenter graphiquement la distribution des fréquences.*

Nombre de visites (xi)	Nombre de jours (ni)
0	1
1	2
2	3
3	2
4	1
5	1
Total	10



« Diagramme en bâtons »

Ce diagramme en bâtons correspondant permettrait de visualiser, par exemple, que le pic de fréquentation se situe à 2 visites par jour (3 jours), et que les jours sans visite sont rares (1 seul jour).

b. La courbe cumulative en escaliers

Cette représentation graphique illustre la fonction de répartition cumulative d'une variable discrète, qu'elle soit exprimée en effectifs cumulés (Ni) ou en fréquences cumulées (Fi). Elle permet d'observer, pour chaque modalité xi, la proportion ou le nombre d'individus dont la valeur de la variable est inférieure strictement à xi — autrement dit, ceux qui satisfont la condition « moins de xi ».

Visuellement, la courbe prend la forme d'une succession de segments horizontaux, rappelant des marches d'escalier :

- Chaque palier s'étend horizontalement jusqu'à la valeur suivante de xi, indiquant la stabilité de l'effectif cumulé sur cet intervalle.
- Le segment est ouvert à gauche et fermé à droite, traduisant l'exclusion de la modalité actuelle du cumul (on ne compte que les valeurs inférieures à xi).

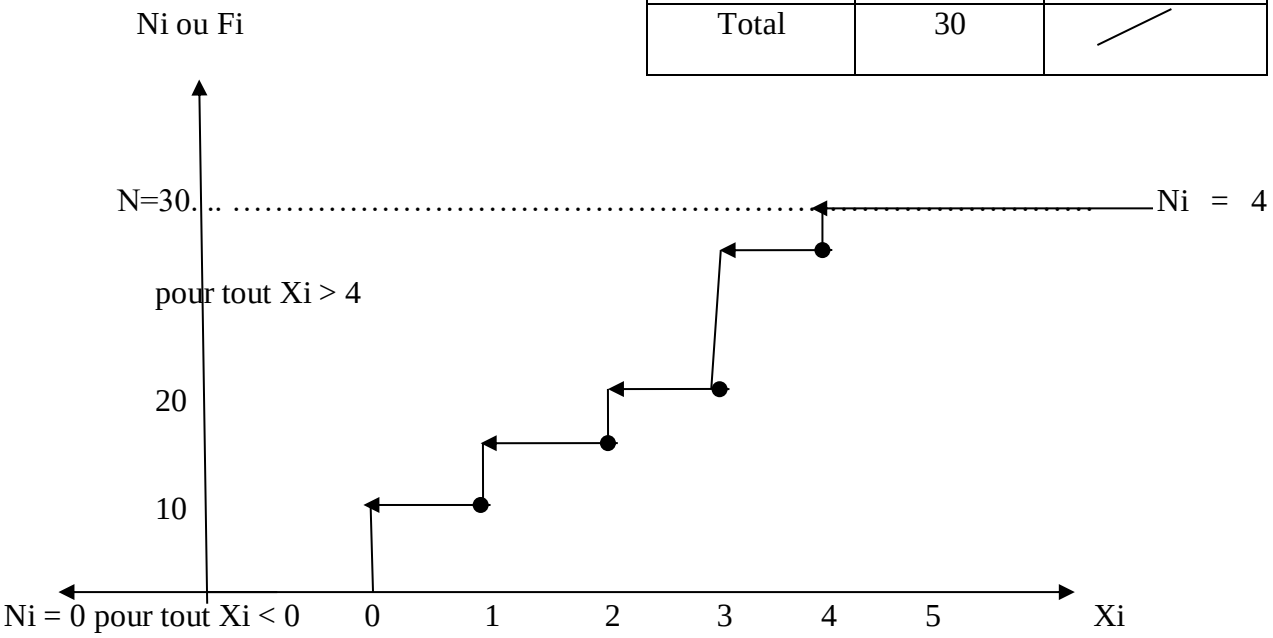
À noter :

- Avant la plus petite valeur observée, la courbe reste au niveau zéro**,** puisque aucun individu n'a de modalité inférieure.
- Ainsi, le premier palier coïncide toujours avec l'axe des abscisses, et reste ouvert vers l'infini négatif ($-\infty$), signalant l'absence de données pour ces valeurs.

Exemple : *Reprenant l'exemple précédent, et représentant graphiquement sa fonction de répartition.*

Avant de tracer la courbe cumulative, nous devons d'abord construire la colonne des effectifs (fréquences) cumulés (Ni ou Fi).

Nombre de terminaux (xi)	**Nombre de bureaux (ni)**	**Effectifs cumulés ($n_i^{\nearrow}$)**
0	10	10
1	8	18
2	2	20
3	4	24
4	6	30
Total	30	



« Courbe cumulative en escaliers »

2.2.2- La représentation graphique de la VSC

Dans ce cas, les données sont regroupées sous forme de classes. La fonction de distribution {xi ; ni } ou {xi ; fi} est représentée par un ***histogramme*** (à partir duquel on peut déduire le ***polygone***), et la fonction cumulative {xi ; Ni} ou {xi ; Fi} est représentée par la ***courbe cumulative en « S ».***

a. L'histogramme

Ce type de graphique repose sur un système de coordonnées cartésiennes. Dans ce cas, chaque classe est représentée par un rectangle vertical dont la largeur représente l'amplitude de la classe, et la longueur l'effectif ou la fréquence de la classe.

C'est la surface de l'histogramme (des rectangles) qui intéresse le chercheur. Elle doit être proportionnelle aux effectifs. Cette proportionnalité se vérifie suivant deux situations :

- **Cas des amplitudes égales :**

Dans ce cas tous les rectangles ont la même largeur. Donc les surfaces sont proportionnelles aux seules longueurs (n_i ou f_i). Alors, l'histogramme peut être tracé directement avec les n_i ou les f_i.

Il s'agit d'une variable statistique continue et les données sont présentées sous forme de classes. Donc la représentation graphique correspondante c'est automatiquement l'histogramme.

Cependant, avant de tracer cet histogramme, il faut vérifier le respect du principe de la proportionnalité des effectifs (ou fréquences) aux surfaces des rectangles. Pour cela, il faut vérifier si les rectangles ont tous la même largeur ou pas, c'est-à-dire la même amplitude de classe ou pas. Dans cet exemple, on remarque que toutes les classes ont la même amplitude ai = 5. Donc on tracer notre histogramme directement avec les ni ou fi, le principe de la proportionnalité des surfaces aux effectifs est respecté. Notre histogramme sera donc comme suit :

- **Cas des amplitudes inégales :**

Dans ce cas, la surface des rectangles n'est pas proportionnelle aux seuls effectifs ou fréquences (longueurs), mais aussi aux amplitudes de classes (largeurs). Il faudrait procéder à la correction des effectifs. En fait, il s'agit de rendre les surfaces proportionnelles aux densités. Ces densités se conçoivent de deux manières :

- soit on ramène chaque effectif à l'amplitude de la classe correspondant, c'est-à-dire aux densités (***$d_i = ni/a_i$ ou $d_i = f_i/a_i$***),

- soit on ramène tous les effectifs (ou fréquences) de classes à une *même* et *commune* amplitude, appelée *amplitude de base,* notée « ***a_0*** ». Celle-ci n'est autre que *la plus petite amplitude* de classe observée dans la distribution (parfois, c'est aussi le plus grand diviseur commun). On construit dans ce cas un histogramme avec les effectifs (ou fréquences) corrigés, notés « ***n_{ic}*** » ou « ***f_{ic}*** », où les surfaces sont proportionnelles à ces « n_{ic} ».

Il s'agit d'une distribution sous forme de classes (forme continue), donc la représentation graphique correspondante est l'histogramme. Cependant, avant de tracer celui-ci, il faut d'abord vérifier les amplitudes de classes si elles sont constantes ou non, c'est-à-dire vérifier le principe de la proportionnalité des surfaces aux effectifs ou fréquences.

On remarque dans cet exemple que l'amplitude de classe n'est pas constante : ex ; la première, la deuxième et la troisième classes ont une même amplitude (a_i = 5), alors que la quatrième classe a une amplitude ai = 10. Il faudrait donc, avant de tracer l'histogramme, calculer les densités (**ni/ai** ou fi/ai) ou bien corriger les effectifs **nic=** $\frac{ni \times ao}{ai}$ (on précisera que l'amplitude de base « ***a_0*** », qui est la plus petite amplitude de classe, est $a_0 = 5$).

b. Le polygone

Parfois l'histogramme, aussi parfait soit-il, ne permet pas des comparaisons entre différentes distributions. Aussi, dans ce cas, on a recours au ***« polygone »***, construit à partir de l'histogramme lui-même. Le polygone est un ensemble de segments qui relient les milieux des sommets des rectangles, à distance égale à « $\mathbf{a_i/2}$ » si l'amplitude de classe est constante, et à « $\mathbf{a_0/2}$ » si l'amplitude de classe n'est pas constante ; tout en ajoutant deux classes fictives l'une avant la première classe, l'autre après la dernière classe.

Le polygone nous donne une courbe continue dont la surface délimitée avec l'axe des abscisses représente la même surface que celle de l'histogramme, celle-ci étant égale à N ou 1, et ce, conformément au principe de *compensation des aires*.

c. Les courbes cumulatives en « S »

La fonction de répartition d'une variable statistique continue est représentée par deux courbes cumulatives en « s ». L'une croissante, pour les fréquences ou effectifs cumulés croissants, l'autre décroissante, pour les fréquences ou effectifs cumulés décroissants. Les deux courbes sont représentées sur un même graphique (même plan). La courbe croissante relie les limites supérieures des classes (Moins de), la courbe décroissante relie les limites inférieures des classes (Plus de).

Remarque

- Quelque soit l'amplitude de classe, constante ou non, cela n'a aucune incidence sur les courbes cumulatives. Les effectifs cumulés étant des rangs, des numéros de modalités, pas des densités.

- La courbe cumulative croissante stagne à N, pour toute valeur xi supérieure (>) à la limite supérieure de la dernière classe. Elle stagne aussi à 0 pour toute valeur xi inférieure(<) à la limite inférieure de première classe.

- La courbe cumulative décroissante stagne à N pour toute valeur xi inférieure à la limite inférieure de la première classe. Elle stagne aussi à 0 pour toute valeur xi supérieure à la limite supérieure de la dernière classe.

On peut reprendre l'exemple précédent dans lequel nous avons déjà calculé les effectifs cumulés, et tracer sa courbe cumulative.

Conclusion :

La présentation graphique et tabulaire des données constitue une étape fondamentale de toute analyse statistique. En structurant les informations brutes à travers des tableaux et en les

illustrant par des représentations visuelles adéquates, le statisticien facilite l'identification des tendances, des anomalies, et des relations essentielles entre variables.

Ces outils sont indispensables à la prise de décision éclairée, tant dans le domaine médical, économique, que dans la gestion des politiques de santé publique. Une bonne maîtrise de ces instruments permet de passer d'une lecture descriptive des données à une interprétation analytique solide, condition préalable à toute recherche empirique sérieuse.

Chapitre 3 : Caractéristiques des distributions à un caractère

Introduction

Après avoir appris à synthétiser les données statistiques dans le chapitre précédent, il devient essentiel de passer à une étape plus fine d'analyse : la caractérisation de la distribution. Cette étape vise à décrire une série statistique à l'aide d'indicateurs numériques représentatifs appelés paramètres. Ces paramètres, choisis et calculés par le statisticien en fonction du phénomène étudié, permettent de mieux cerner la structure et les particularités de la série observée.

En statistique descriptive, ces outils de mesure se répartissent en **quatre grandes catégories**, chacune jouant un rôle spécifique dans l'interprétation des données :

- Les **paramètres de tendance centrale**, qui indiquent vers quelles valeurs les données se regroupent.
- Les **paramètres de dispersion**, qui évaluent l'étalement ou la variabilité des données autour de cette tendance.
- Les **paramètres de forme**, qui donnent des indications sur l'asymétrie ou l'aplatissement de la distribution.
- Les **paramètres de concentration**, qui mesurent la manière dont les valeurs se répartissent entre les individus.

Section 1: Les paramètres de tendance centrale : Nous allons étudier successivement dans cette section: le mode, la médiane et la généralisation de la médiane, la moyenne arithmétique et les autres types de moyennes.

1.1. Le Mode (noté Mo): Le mode est par définition la valeur de la variable la plus dominante ou la plus fréquente, c'est-à-dire la modalité qui correspond au plus grand effectif (ou à la plus grande fréquence). La détermination de la valeur du mode diffère selon qu'il s'agisse de variable statistique discrète (VSD) ou de variable statistique continue (VSC).

1.1.1. Variable statistique discrète (VSD): lorsque la variable statistique étudiée est discrète, la valeur du mode est directement observable.

Exemple: *Soit la distribution statistique de 24 étudiants d'un groupe de TD selon leurs notes :*

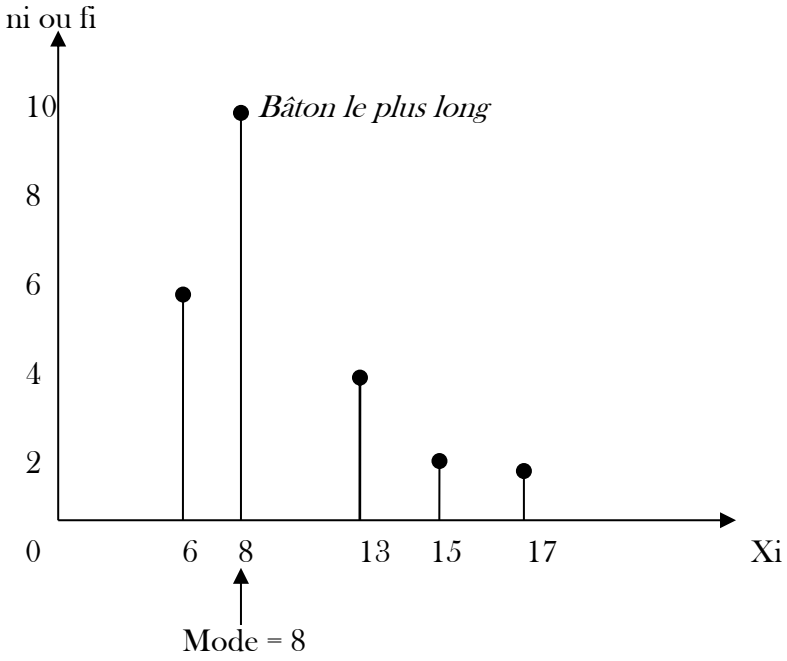
x_i	6	8	13	15	17	Total
n_i	6	10	4	2	2	24

- *Déterminer le mode de cette distribution.*

Il suffit de regarder la colonne (ni), le plus grand nombre c'est ni = 10, la modalité correspondante dans la colonne (xi) c'est 8. Donc le Mode = 8.

- De manière graphique :

Il s'agit d'une variable statistique discrète, donc le graphique correspondant est le *diagramme en bâtons*. Le principe est de repérer le bâton le plus long. On aura alors le graphique suivant :



1.1.2- Variable statistique continue (VSC)

Dans ce cas les données sont regroupées sous forme de classes. La valeur (xi) correspondant au mode appartient forcément à une classe. Cette classe s'appelle *classe modale*. Toutefois, pour trouver la valeur du mode, on doit distinguer deux situations:

a)- Cas des amplitudes égales

On détermine le mode en deux étapes: on trouve d'abord la classe modale (celle qui a le plus grand n_i ou f_i), puis on calcule la valeur du mode par la formule suivante:

$$\mathbf{Mo = X_o + \left(a \frac{(n_{mo} - n_{mo-1})}{(n_{mo} - n_{mo-1}) + (n_{mo} - n_{mo+1})} \right)}$$

Ou bien :

$$\mathbf{Mo = X_o + \left(a \frac{(f_{mo} - f_{mo-1})}{(f_{mo} - f_{mo-1}) + (f_{mo} - f_{mo+1})} \right)}$$

Avec :

- X_o = Limite inférieure de la classe modale.
- a = amplitude la classe modale.

- n_{mo} = effectif ou fréquence de la classe modale.
- $n_{mo\ -1}$ = effectif ou fréquence de la classe avant ou précédant la classe modale.
- n_{mo+1} = effectif ou fréquence de la classe après ou suivant la classe modale.

b)- Cas des amplitudes inégales: Quand les amplitudes des classes sont différentes, pour trouver la classe modale et déterminer le mode, on doit calculer les effectifs corrigés (nic) ou bien les densités (d_i).

$$\mathbf{Mo = Xo + \left[a \frac{(n_{cmo} - n_{cmo-1})}{(n_{cmo} - n_{cmo-1}) + (n_{cmo} - n_{cmo+1})} \right]}$$

Ou bien encore avec les densités :

$$\mathbf{Mo = Xo + \left[a \frac{(d_{mo} - d_{mo-1})}{(d_{mo} - d_{mo-1}) + (d_{mo} - d_{mo+1})} \right]}$$

Avec :

- Xo = Limite inférieure de la classe modale.
- a = amplitude la classe modale.
- n_{cmo} = effectif (ou fréquence) corrigé de la classe modale.
- $n_{cmo\ -1}$ = effectif (ou fréquence) corrigé de la classe avant ou précédant la classe modale.
- $n_{cmo+}1$ = effectif (ou fréquence) corrigé de la classe après ou suivant la classe modale.
- d_{mo} = densité de la classe modale.
- d_{mo-1} = densité de la classe avant ou précédant la classe modale.
- d_{mo+1} = densité de la classe après ou suivant la classe modale.

Exemple :

Calculer le mode de la distribution suivante :

Classes	Effectifs
10 - 20	10
20 - 30	20
30 - 50	30
50 - 80	25

Pour calculer le mode, il faut d'abord vérifier les amplitudes (ai). On construit comme précédemment une colonne (ai).

Classes	Effectifs	a_i	d_i	n_{ic}
10 - 20	10	10	1	10
20 - 30	20	10	**2**	**20**
30 - 50	30	20	1,5	15
50 - 80	25	30	0,83	8,33
Total	85	-	-	-

Si on regarde la colonne (a_i), on constate que l'amplitude de classe n'est pas constante. La classe modale est celle qui a la plus grande densité ou le plus grand effectif corrigé.

Dans le tableau ci-dessus, la densité la plus élevée $d_2 = 2$, correspond également sur la même ligne, dans la colonne (n_{ic}), à la plus grande valeur n_{ic}, $n_{2c} = 20$. La classe correspondante, à savoir ; [20 - 30[, est la classe modale.

$$Mo = Xo + \left[a \frac{(n_{cmo} - n_{cmo-1})}{(n_{cmo} - n_{cmo-1}) + (n_{cmo} - n_{cmo+1})} \right]$$

$$Mo = 20 + \left[10 \frac{(20 - 10)}{(20 - 10) + (20 - 15)} \right] = 26,67 \longrightarrow \underline{Mo = 26,67}$$

Ou bien ;

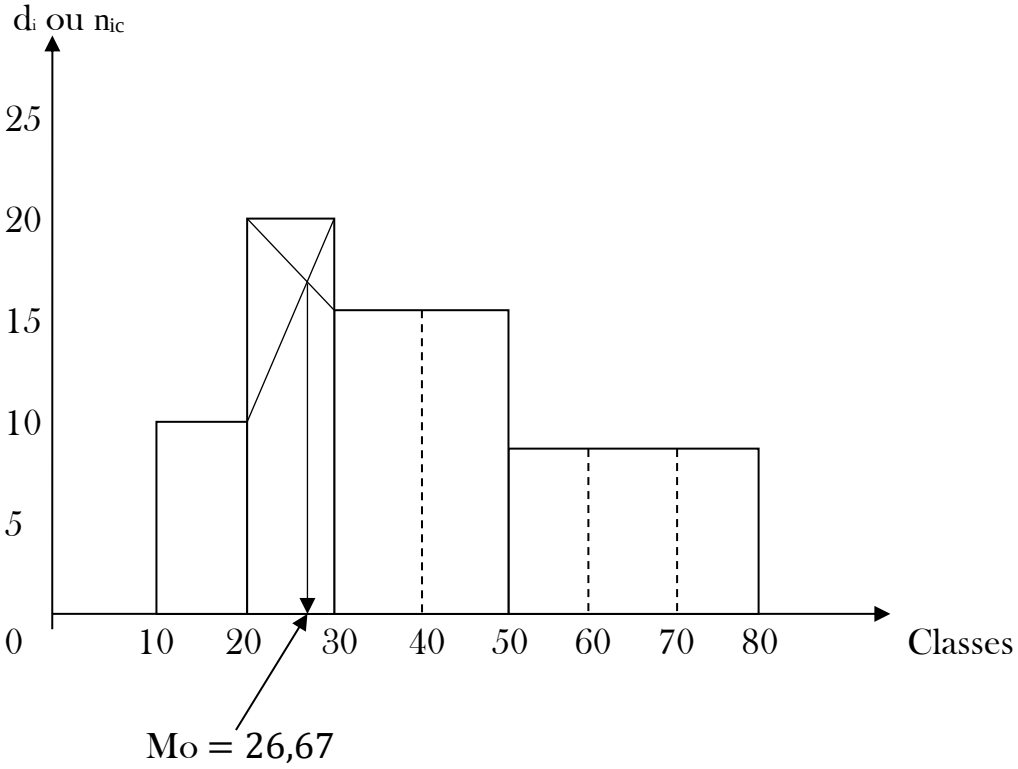
$$Mo = Xo + \left[a \frac{(d_{mo} - d_{mo-1})}{(d_{mo} - d_{mo-1}) + (d_{mo} - d_{mo+1})} \right]$$

$$Mo = 20 + \left[10 \frac{(2 - 1)}{(2 - 1) + (2 - 1,5)} \right] = 26,67 \longrightarrow M = 26,67$$

De manière graphique: Il s'agit d'une VSC, donc la fonction de distribution est représentée graphiquement par l'histogramme. Compte tenu de l'amplitude de classe (égale ou pas), le mode se détermine de la même manière à partir du rectangle de la classe modale dans l'histogramme. On obtient le mode, graphiquement, en joignant par une droite les limites supérieures de la classe modale et de la classe précédente. Avec une autre droite, on joint les limites inférieures de la classe modale et de la classe suivante. Le point d'intersection des deux droites sera projeté sur l'axe des abscisses. La valeur (xi) correspondante sur cet axe est le mode.

Reprenant l'exemple 2 précédent et traçant le graphique correspondant.

Il s'agit d'une VSC, donc le graphique correspondant est un histogramme, comme suit :



Remarque

- Une série statistique peut présenter un, deux ou plusieurs modes à la fois, comme elle peut ne pas en présenter du tout, on l'appelle alors une distribution ou série « *amodale* ».

1.2. La Médiane (notée Me): la médiane est par définition la valeur de la variable qui partage l'effectif total ou la fréquence totale de la distribution en deux parties égales. Le calcul de la médiane dépend de la nature de la variable (VSD ou VSC).

1.2.1- Cas d'une variable statistique discrète (VSD):

Dans ce cas la détermination de la médiane dépend du nombre d'individus (N).

- ***Si N est un nombre impair***

Dans ce cas il existe une modalité, parmi toutes les modalités de la série, qui divise la série en deux groupes de même effectif (N/2). Cette modalité est la médiane. On la détermine comme suit :

N est impair, ceci implique que mathématiquement N s'écrit : **N = 2k+1.**

NB/- k est un effectif cumulé croissant. Il donne la position (ou le numéro) de la médiane dans la série ordonnée.

Dans ce cas la médiane est la modalité correspondant à la position numéro ***(k+1)^{ème}*** . Il suffit alors de calculer k.

Exemple: *Soit la série ordonnée suivante : 3 - 6 - 12 - 18 - 20 - 23 - 28. Déterminer la médiane.*

N= 7, N est impair, il s'écrit donc N = 2k+1 ⟶ k = 3

Donc la médiane est la modalité numéro (k+1)^{ème}, soit (3 + 1 = 4)^{ème} ou la 4ème.

Dans la série ordonnée, on constate que la 4ème modalité c'est 18. Donc Me = 18.

- ***Si N est un nombre pair***

Dans ce cas, on parle d' *intervalle médian* délimité par ***[k^ème et (k+1)^ème]*** modalités.

Exemple: *Soit la série ordonnée suivante : 3 - 6 - 12 - 18 - 20 - 23 - 28 - 30. Déterminer la médiane de cette série.*

N = 8, N est pair, il existe donc un intervalle médian [k^ème ; (k+1)^ème]

N peut s'écrire mathématiquement : N = 2k.

Donc k = 4 ⟶ l'intervalle médian est délimité par la 4 ème et (4+1 = 5)^ème modalités.

La 4ème correspond à la modalité 18, et la 5ème à la modalité 20. Donc l'intervalle médian est :

[18 ; 20[, Me = $\frac{18+20}{2}$ = 19.

Remarque : *Dans le cas des valeurs répétitives (où N est trop élevé), on a recours alors au tableau statistique, où les effectifs cumulés nous permettent de repérer les positions des modalités qu'on cherche, en suivant la même logique.*

Exemple:
Déterminer la médiane de la distribution suivante :

Xi	Ni	Ni
0	10	10
1	32	42
2 ◄	36	78
3	15	93
4	5	98
5	2	100
Total	100	-

Dans cet exemple on déterminera la médiane en utilisant directement la colonne des Ni.

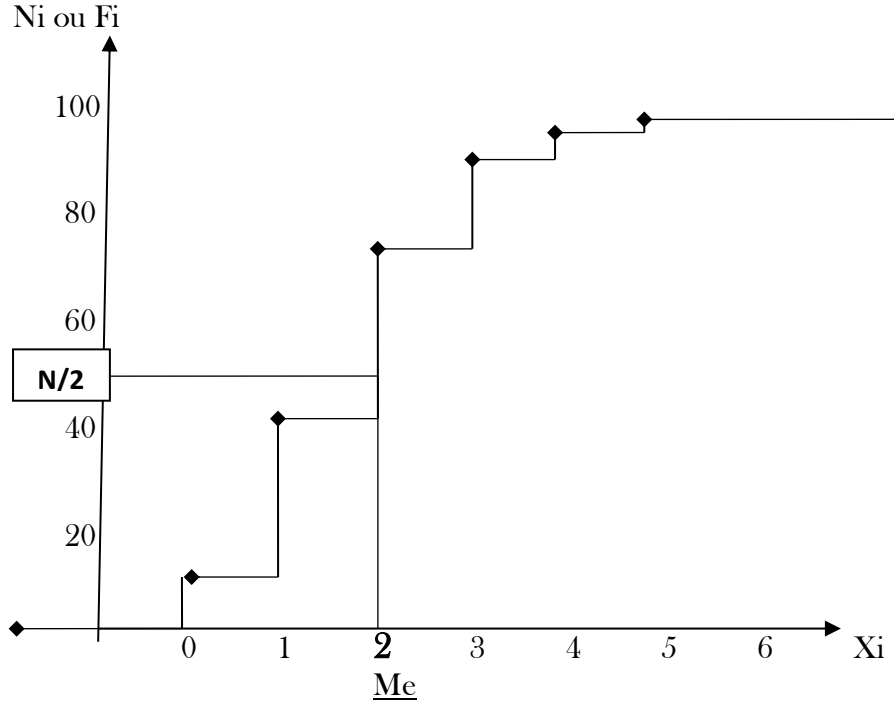
N = 100, donc N est un nombre pair. Il existe donc un intervalle médian.

N = 2k ⟶ k = 50 et (k+1) = 51. Me ∈ [50ème ; 51ème] modalités. C'est-à-dire Me se situe à la position entre la 50ème et la 51ème modalité du tableau. Me= $\frac{2+2}{2}$, Me= 2.

Détermination graphique

Dans le cas d'une VSD la fonction cumulative est représentée graphiquement par la *courbe cumulative en escalier*. La médiane se déterminant par les effectifs cumulés, elle est donc

logiquement déterminée graphiquement par la courbe représentant ces derniers. On peut illustrer la méthode de détermination à partir de la courbe cumulative de l'exemple précédent.



1.2.2- Variable statistique continue (VSC):

Dans ce cas, les données sont présentées sous forme de classes, la modalité médiane appartient forcément à une classe, appelée intervalle ou *classe médiane*, à partir de laquelle il faudrait la déterminer.

a)- Détermination de la médiane par le calcul: pour déterminer la valeur de la médiane, on doit d'abord trouver la classe médiane. Pour cela, on commence par cumuler les ni ou les fi et repérer dans la colonne Ni ou Fi, la valeur ou la position N/2 ou Fi = 0,5 (soit 50%), notée « Th_2 ». La valeur de la médiane sera trouvée à partir de la formule suivante:

$$\mathbf{Me = X_0 + \left[a \frac{(Th_2 - n_{me-1})}{n_{me}} \right]} \quad \mathbf{ou} \quad \mathbf{Me = X_0 + \left[a \frac{(Th_2 - f_{me-1})}{f_{me}} \right]}$$

Avec :

X_0= Limite inférieure de la classe médiane.

a = amplitude de la classe médiane.

Th_2 = N/2 ou 0,5 respectivement.

N_{me-1} ou F_{me-1} = effectif ou fréquence cumulé correspondant à la classe avant la classe médiane. n_{me} ou f_{me} = effectif absolu ou fréquence relative correspondant à la classe médiane.

Exemple: *Déterminer la médiane de la distribution suivante ;*

Classes	Effectifs
15 - 25	26
25 - 30	33
30 - 40	64
40 - 50	7
50 - 65	10

Pour déterminer la médiane, on a besoin des effectifs ou fréquences cumulées. On construit la colonne Ni. Le tableau complet nécessaire sera comme suit (on ajoutera d'emblée la colonne ni↓ décroissantes pour les besoins du graphique :

Classes	Xi	ni	ni↑	ni↓
15 - 25	20	26	26	140
25 - 30	27,5	33	9	114
30 - 40	35	64	123	81
40 - 50	45	7	130	17
50 - 65	57,5	10	140	10
Total		140	-	-

$Th_2 = N/2 = 70 \longrightarrow Me \in [30 - 40[.\quad Me = x_0 + \left[a\ \frac{N/2 - n^{\uparrow}_{me-1}}{nme}\right]$

Ainsi:

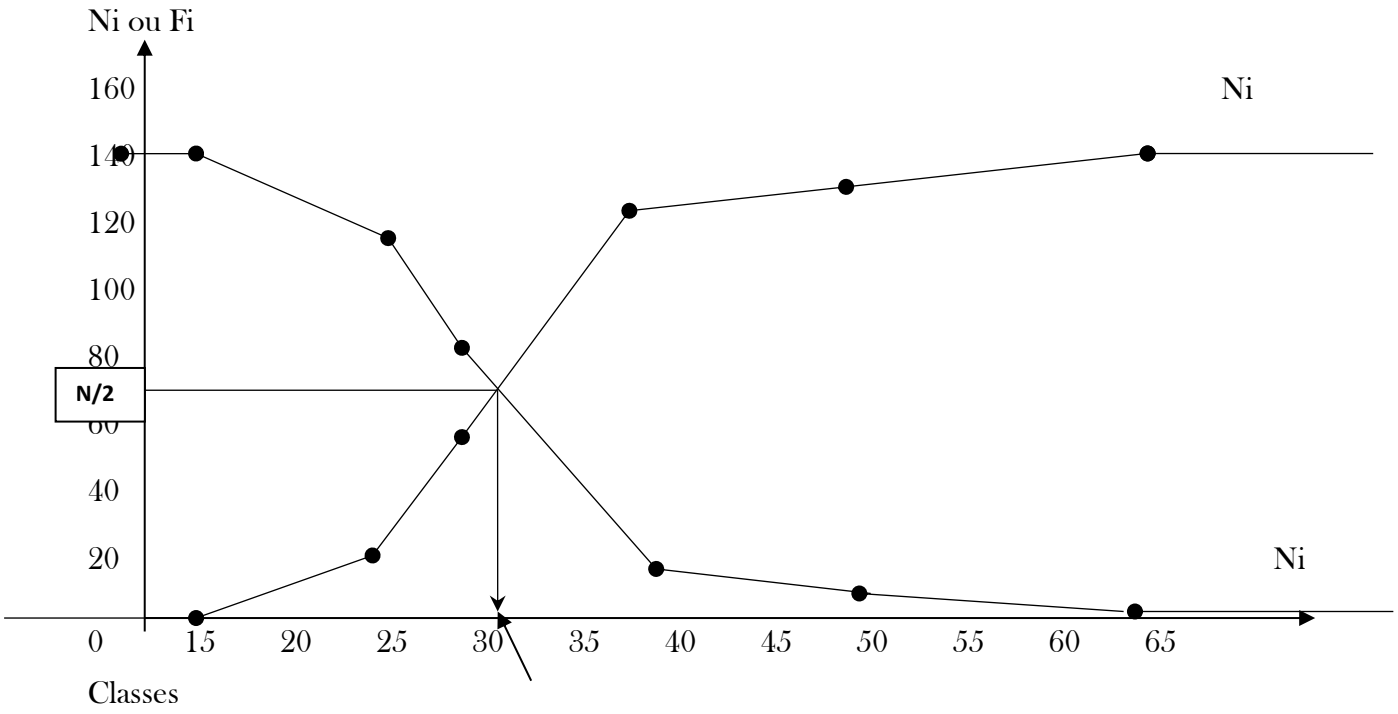
$$Me = 30 + \left[10\ \frac{70 - 59}{64}\right] = 31,72 \longrightarrow \mathbf{Me = 31,72}$$

b). La médiane graphiquement:

A partir des deux courbes cumulatives en « S », on détermine la médiane en projetant le point de rencontre des deux courbes sur l'axe horizontale. La valeur correspondante est la médiane.

Remarque

Logiquement le point où se rencontrent les deux courbes est le point correspondant sur l'axe vertical à Ni = N/2 ou Fi = 0,5 (50%).



Me = 31,72

NB: La médiane n'est pas affectée par les valeurs extrêmes, puisqu'elle ne dépend des valeurs que par leur position (leurs effectifs cumulés).

1.3 Généralisation de la notion de la médiane : les quantiles

En suivant la même logique de définition que celle de la médiane, on peut déterminer d'autres paramètres de position qui permettent de partager une série statistique, non seulement en deux groupes de mêmes effectifs, mais aussi en quatre, dix, cents,… ; groupes de mêmes effectifs. On appelle ces paramètres les « *Quantiles* ».

Nous définissons dans ce cours les trois quantiles les plus usités, à savoir ; les *quartiles*, les *déciles* et les *centiles*. Ils sont surtout calculés pour les VSC, en suivant les trois étapes que pour la médiane.

1.3.1 Les Quartiles

Notés « Qi », ce sont des paramètres de position et aussi des modalités xi qui, par leurs positions (notées « **Thi** ») , partagent la série statistique en quatre groupes de mêmes effectifs, soit 25% chacun. Ils sont donc au nombre de trois : Q_1 ; Q_2 ; Q_3.

- **Q_1**, appelé premier quartile, est la modalité de la série par rapport à laquelle 25% des modalités y sont inférieures (à Q_1) et 75% restantes y sont supérieures (à Q_1).

- Q_2 ; appelé aussi deuxième quartile, est la modalité xi par rapport à laquelle 50% des modalités de la série y sont inférieure et 50% restantes y sont supérieurs. Ce n'est donc rien d'autre que la Médiane que nous avons déjà définie. (**Q_2 = Me**).

- Q_3 ; appelé également troisième quartile, est la modalité de la série par rapport à laquelle 75% des modalités y sont inférieures et 25% restantes y sont supérieures. Il est donc le contraire de Q_1.

L'intervalle $[Q_1 ; Q_3[$ s'appelle ***l'intervalle interquartile***. Il contient **50%** des modalités centrales.

Formule

$$Qi = X_0 + a\left[\frac{Thi - N_{Qi-1}}{n_{Qi}}\right] \qquad \text{avec : } Thi = \frac{i.N}{4}$$

1.3.2 Les Déciles

Notés « Di », ce sont des paramètres de position et aussi des modalités xi qui partagent la série statistique en dix groupes de même effectif, soit 10% chacun. Ils sont donc au nombre de neuf (9) : D_1 ; D_2…….. ; D_9.

- D_1 ; appelé premier décile, est la modalité xi par rapport à laquelle 90% des valeurs y sont supérieures (à D_1), et les 10% restantes y sont inférieures.

- D_5 ; appelée aussi cinquième décile, est la modalité xi par rapport à laquelle 50% des modalités y sont supérieures à (D_5), et les 50% restantes y sont inférieures. Ce n'est donc rien d'autre que la médiane : **D_5 = Me**.

- D_9 ; appelé neuvième décile, est le contraire de D_1. C'est la modalité xi par rapport à laquelle 90% des modalités y sont inférieures, et les 10% restantes y sont supérieures.

L'intervalle [**D_1 ; D_9[** s'appelle ***intervalle inter-décile***. Il contient **80%** des modalités centrales.

Formule

$$Di = X_0 + a\frac{Thi - N_{Di-1}}{n_{Di}} \qquad \text{avec : } Thi = \frac{i.N}{10}$$

1.3.3 Les Centiles

Notés « **Ci** », ce sont des paramètres de position et aussi des modalités xi qui partagent la série statistique en cent groupes de même effectif, soit 1% chacun. Ils sont donc au nombre de 99 : C_1 ; C_2……C_{50} ;… ; C_{99}.

- C_1, appelé aussi premier centile, est la modalité de la série par rapport à laquelle 1% des modalités y sont inférieures (à C_1) et 99% restantes y sont supérieures (à C_1).

- C_{50}, appelé aussi cinquantième centile, est la modalité de la série par rapport à laquelle 50% des modalités y sont inférieures (à C_{50}) et 50% restantes y sont supérieures (à C_{50}). Ce n'est donc rien d'autre que la médiane : **C_{50} = Me**.

- C_{99}, appelé quatre-vingtième centile, est le contraire de C_1. C'est la modalité xi par rapport à laquelle 99% des modalités y sont inférieures, et les 1% restantes y sont supérieures.

L'intervalle **[C_1 ; C_{99}[** s'appelle ***intervalle inter-centile***. Il contient **98%** des modalités centrales**.**

Formule

$$Ci = X_0 + a \frac{Thi - N_{ci-1}}{n_{ci}} \qquad \text{avec : } Thi = \frac{i.N}{100}$$

Remarque

- **Me = Q_2 = D_5 = C_{50}.**
- **Q_3 = C_{75}.**

Exemple : *Calculer Q_1 ; D_2 et C_{80} de la distribution de l'exemple précédent.*

1/- Q_1 :

$Th_1 = 1.N/4 = 140/4 = 35 \longrightarrow Q_1 \in [25 ; 30[$

$$Q_1 = 25 + 5 \frac{35 - 26}{33} = 26,36 \qquad \boldsymbol{Q_1 \approx 26,36}$$

2/- D_2 :

$Th_2 = 2.N/10 = 28 \longrightarrow D_2 \in [25 ; 30[$

$$D_2 = 25 + 5 \frac{28 - 26}{33} = 25,30 \qquad \boldsymbol{D_2 \approx 25,30}$$

3/ C_{80} :

$Th_{80} = 80.N/100 = 112 \longrightarrow C_{80} \in [30 ; 40[$

$$C_{80} = 30 + 10 \frac{112 - 59}{64} = 38,28 \qquad \boldsymbol{D_2 = 38,28}$$

1.4. La moyenne arithmétique (notée $\bar{X}$):

Il s'agit dans ce point de découvrir le paramètre de tendance centrale le plus usuellement utilisé en statistique. En définissant ce paramètre, en donnant ses formules et, surtout, ses propriétés, on permet à l'étudiant de comprendre les soubassements qui en font un paramètre d'excellence.

1.4.1- Définition

Notée «$\bar{X}$ », la moyenne arithmétique correspond au rapport de la somme des modalités par leur effectif total.

- On dit qu'une moyenne arithmétique est « *simple* » ou non pondérée lorsque chaque modalité x_i ne se répète qu'une seule fois ($n_i = 1$). On écrit alors :

$$\overline{X} = \left(\sum_{i=1}^{N} Xi\right) / N$$

Cette formule se lit comme suit : « *X barre égale la somme des x_i ; (i allant de 1 jusqu'à n (N étant le nombre de modalités différentes dans ce cas égale au nombre d'individus k = N)* ».

- On dit qu'une moyenne arithmétique est « *pondérée* » lorsqu'à chaque modalité x_i correspond un effectif (n_i). On écrit alors :

$$\overline{X} = \sum_{i=1}^{k}(nixi)\ /N \ = \ \sum_{i=1}^{k} fixi$$

Cette formule se lit comme suit : « *X barre égale la somme des x_i ; (i allant de 1 jusqu'à k (k étant le nombre de modalités différentes dans ce cas différent du nombre d'individus k≠ N)* ».

La moyenne arithmétique pondérée s'emploie quand les données sont regroupées en classes ou quand les données discrètes se répètent, c'est-à-dire dans le cas de distribution statistique (xi ; ni).

1.4.2- Méthode de calcul

a. Variable statistique discrète

- **Cas d'une série simple**

Soit la série suivante : {10 - 20 - 24 - 28 - 30 - 32}.

N= 6 ; toutes les modalités se répètent une seule fois, c'est donc une série simple. Dans ce cas la moyenne arithmétique sera : $\overline{X} = \left(\sum_{i=1}^{N} Xi\right) / N$

$\overline{X}$ = (10 + 20 + 24 + 28 + 30 + 32) / 6 = 24 ⟶ $\overline{X} = 24$

- **Cas d'une série pondérée**

Soit la distribution suivante :

Xi	0	1	2	3	4
Ni	20	35	10	30	5

Dans ce cas à chaque modalité (x_i) est associé un effectif (n_i). Il s'agit donc d'une série pondérée ou distribution statistique. Dans ce cas la moyenne arithmétique sera :

$$\overline{X} = \sum_{i=1}^{k}(nixi)\ /N \ = \ \sum_{i=1}^{k} fixi$$

Pour pouvoir appliquer cette formule, nous devons compléter le tableau, en ajoutant une nouvelle colonne « n_iX_i » ou « f_iX_i ». Le tableau sera comme suit :

x_i	n_i	f_i	n_iX_i	f_iX_i
0	20	0,200	0	0
1	35	0,350	35	0,35
2	10	0,100	20	0,2
3	30	0,300	90	0,9

4	5	0,050	20	0,20
Total	100	1	165	1,65

$\overline{X}$ = le total de la colonne (n_ix_i) divisé par le total de la colonne ni ; ou directement $\overline{X}$ = total de la colonne (f_ix_i) :

$\overline{X}$ = 165/100 = 165. ⟶ $\underline{\overline{X} = 1,65}$

b. Variable statistique continue

Dans ce cas on retient pour les calculs les centres de classes (xi), et on ajoute les colonne (n_ix_i) ou (f_ix_i). Soit la distribution suivante :

Classes	[0 - 5[	[5 - 10[	[10 - 15[	[15 - 30[
Effectif	15	30	20	35

Il s'agit d'une distribution sous forme de classes. Pour calculer $\overline{X}$, on doit d'abord calculer les centres de classes, à partir desquels on calculera la colonne (n_ix_i). On aura le tableau suivant :

Classes	x_i	n_i	f_i	n_ix_i	f_ix_i
[0 ; 5[	2,5	15	0,150	37,5	0,375
[5 ; 10[	7,5	30	0,300	225	2,25
[10 ; 15[	12,5	20	0,200	250	2,5
[15 ; 30[	22,5	35	0,350	787,5	7,875
Total	**-**	**100**	**1**	**1300**	**13**

$\overline{X} = \sum_{i=1}^{k}(nixi)\ /N \ = \ \sum_{i=1}^{k} fixi$ = 1300/100 = 13 $\underline{\overline{X} = 13}$

1.4.3- Propriétés de la moyenne arithmétique : la moyenne arithmétique possède plusieurs propriétés qu'il convient de connaitre, car leur utilisation permet de faire des simplifications de calculs. On peut retenir les propriétés suivantes :

1- La somme des écarts à la moyenne arithmétique est nulle : $\Sigma(x_i - \overline{X}) = 0$.

2- La somme des carrés des écarts à la moyenne arithmétique est minimale (c'est le plus petit écart qu'on puisse calculer, on l'appelle le principe des moindres carrés).

$$\Sigma n_i(x_i - \overline{X})^2 \longrightarrow \textbf{Minimum}$$

3- La moyenne arithmétique d'une constante (a) est égale à la constante elle-même :

$$\overline{\mathbf{a}} = \mathbf{a}$$

4- La moyenne arithmétique d'une population scindée en deux ou plusieurs sous-populations est égale la moyenne arithmétique des moyennes des sous-populations, pondérées par leurs effectifs respectifs. Ainsi si une population P d'effectif total N est subdivisée en P_1, P_2, P_3,….Pn sous-populations de moyennes respectives m_1 ; m_2 ; m_3….., m_k et d'effectifs respectifs n_1 ; n_2 ; n_3 ;…… ; n_k ; alors la moyenne arithmétique de cette population est :

$$\overline{\mathbf{X}}_{\mathbf{P}} = \mathbf{1/N\ \Sigma[m_1n_1 + m_2n_2 + m_3n_3 + ;\ldots\ldots\ldots ;\ m_kn_k]}$$

Ou bien $$\overline{\mathbf{X}}_{\mathbf{P}} = \mathbf{\Sigma[m_1f_1 + m_2f_2 + m_3f_3 + ;\ldots\ldots\ldots + m_kf_k]}$$

Avec : $\Sigma n_i = N$ et la $\Sigma f_i = 1$.

5- La moyenne arithmétique est très sensible aux influences des valeurs (modalités) extrêmes qui la rendent peu significative. De même, dans le calcul des durées ou vitesses moyennes, des taux ou des pourcentages et des valeurs élevées au carré, elle est très mal appropriée. On lui préfère dans ce cas d'autres types de moyennes que nous développons dans le dernier point de la présente section.

6- Lorsqu'on ajoute ou l'on soustrait une quantité constante (a) aux modalités de la variable, la moyenne arithmétique de la série augmente ou diminue de la même quantité (a).

Démonstration :

$X_i' = x_i + a \longrightarrow \overline{X}' = \Sigma n_i x_i'/N = \Sigma n_i(x_i + a)/N = \Sigma n_i x_i/N + \Sigma n_i.a/N = \bar{X} + \mathbf{a}$

$\overline{X}$; a ,car $\Sigma n_i/N = 1$.

7- En divisant ou en multipliant les modalités d'une série statistique par une constante (a), la moyenne arithmétique sera aussi multipliée ou divisée par cette même constante (a).

Démonstration :

$Xi' = ax_i \longrightarrow \overline{X}' = \Sigma n_i x_i'/N = \Sigma n_i(a.x_i)/N \longrightarrow \overline{X}' = a(\Sigma n_i x_i/N) = \mathbf{a.\overline{X}}$.

$Xi' = x_i/a \longrightarrow \overline{X}' = \Sigma n_i(x_i/a)/N = \Sigma(n_i x_i)1/a/N = (1/a)(\Sigma n_i x_i)/N = (\Sigma n_i x_i/a).(1/N)$

$= (\Sigma n_i x_i)/a.N = 1/a\ .\ \overline{X} = \mathbf{\overline{X}/a}$.

NB/- *De ces deux dernières propriétés découle la méthode de calcul de* $\bar{X}$ *par le* <u>*changement de variable*</u> *que nous développons ci-dessous.*

1.4.4- Méthode de changement de variable

Cette méthode permet de calculer X en réduisant l'importance des modalités, notamment lorsque celles-ci sont trop volumineuses.

Cette méthode consiste à utiliser une nouvelle variable, notée « $\mathbf{X_i'}$ », qu'on obtient en faisant subir à la variable (x_i) :

- D'abord un *changement d'origine,* c'est-à-dire ramener tous les centres de classes à un même centre « x_0 », qui n'est autre que le centre de la classe modale, on obtient alors une première variable qu'on appelle *variable centrée*, notée ($\boldsymbol{x_i - x_0}$),

- Ensuite, on fait subir à cette *variable centrée* ($x_i - x_0$) un *changement d'échelle,* c'est-à-dire on ramène toutes les amplitudes de classes à une même amplitude « **a** » qui n'est autre que l'amplitude de la classe modale. On obtient alors : $\mathbf{x_i' = (x_i - x_0)\ /a,}$ appelée variable centrée et réduite, qu'on utilise pour calculer $\bar{X}$.

Autrement dit**,** on fait subir à la variable x_i, à la fois, un changement d'origine et d'échelle pour aboutir à la nouvelle variable « $\mathbf{x_i'}$ ». A partir de là, on détermine $\overline{X}$ comme suit :

$Xi' = (x_i - x_0)/a = \longrightarrow a.x_i' = (x_i - x_0) \longrightarrow x_i = a.\ x_i' + x_0$

$\longrightarrow \overline{x_i} = \overline{a.\ x_i' + x_0} \longrightarrow \overline{X} = \overline{a.\ x_i'} + \overline{x_0}$

Conformément aux propriétés de la moyenne arithmétique définies plus haut, on en déduit que :

$\overline{x_0} = x_0$ (x_0 étant une constante).

$\overline{a.x_i'} = a.\ \overline{X'}$ (a étant une constante).

On en déduit que :

$$\overline{\mathbf{X}} = \mathbf{a.}\overline{\mathbf{X}}\mathbf{' + x_0} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..(1)$$

Il suffit alors de calculer d'abord la moyenne arithmétique des x_i' ; ($\overline{X'}$), en ajoutant une colonne ($n_i x_i'$) au tableau statistique, et en déduire $\overline{X}$.

$\overline{X'} = (\sum_{i=1}^{k} xi')/N.$ ……..(2)

On remplace (2) dans (1), et on retrouve $\overline{X}$.

Exemple

Soit la distribution suivante :

Classes	[5 - 10[	[10 - 15[	[15 - 20[	[20 - 30[	[30 - 45[
Effectifs	4	6	20	30	40

- Calculer la moyenne arithmétique par la formule de définition et par la formule de changement de variable.

Pour calculer $\overline{X}$ avec la formule classique (de définition), il faut ajouter une colonne des centres de classes (x_i) et une colonne ($n_i x_i$). De plus pour calculer X avec le changement de variable, il faut ajouter une autre pour les densités (d_i) étant donné que les amplitudes de classes ne sont pas constantes, une autre colonne pour la nouvelle variable x_i' et une autre ($n_i x_i$'). Le tableau sera alors :

Classes	x_i	n_i	f_i	$n_i x_i$	$f_i x_i$	a_i	d_i	x_i'	$n_i x_i$'	$f_i x_i$'
[5 - 10[	7,5	4	0,040	30	0,3	5	0,8	-2	-8	-0,08
[10 - 15[	12,5	6	0,060	75	0,75	5	1,2	-1	-6	-0,06
[15 -- 20[	17,5	20	0,200	350	3,5	5	**4**	0	0	0
[20 - 30[	25	30	0,300	750	7,5	10	3	15	045	0,45
[30 - 45[	37,5	40	0,400	1500	15	15	2,66	4	16	1,6
Total	-	100	1	2705	27,05	-	-	-	191	1,91

NB/- On a ajouté la colonne (fi) à titre supplémentaire pour rappeler à l'étudiant que toutes les applications que l'on fait avec les ni, on peut aussi les faire avec les fi.

$\overline{X} = \Sigma n_i x_i / N = 2705/100 = \Sigma f_i x_i =$ **<u>27,05</u>**

<u>Par le changement de variable</u>

$\overline{X} = a.\overline{X'} + x_0$; (avec a et x_0 respectivement l'amplitude et le centre de la classe modale).

a_i n'est constante ⟶ la classe modale est celle qui correspond à la plus grande densité ($d_3 = 4$)

⟶ $M_0 \in [15\ ;\ 20[$ ⟶ ~~a = 5~~; $x_0 = 17,5$.

$\overline{X'} = \Sigma n_i x_i'/N = 191/100 = 1,91 = \Sigma f_i x_i'$ ⟶ $\overline{\mathbf{X}} = 5.\ 1,91 + 17,5 =$ **<u>27,05</u>**.

On retrouve donc la même moyenne calculée précédemment.

1.5. Généralisation de la moyenne

La moyenne arithmétique n'est qu'un cas particulier de la notion de moyenne. Il existe en mathématiques des phénomènes où la moyenne arithmétique ne donne pas des résultats fiables. Ainsi, on a recours à d'autres types de moyennes, construites suivant la même logique que celle de la moyenne arithmétique. Il s'agit de la moyenne géométrique, pour le calcul des taux ou pourcentages moyens et de la moyenne harmonique, pour le calcul des rapports ou vitesses moyennes.

1.5.1- La moyenne géométrique

a. Définition

Notée « G », elle est la racine N^ième du produit (multiplication) des N modalités positives du caractère. On l'emploi dans le calcul des *taux d'accroissement moyens*.

b. Méthode de calcul

- **Moyenne géométrique simple**

Dans ce cas où toutes les modalités se répètent une seule fois (ni = 1, Constante). On écrit alors :

$$G = \sqrt[N]{x_1.\ x_2.\ x_3.....x_k} = \sqrt[N]{\prod_{i=1}^{N} Xi} = \left[\prod_{i=1}^{N} Xi\right]^{1/N}$$

Le calcul peut également se faire par les logarithmes :

$$\log G = \log\left(\sqrt[N]{x_1.\ x_2.\ x_3.....x_k}\right) = \log\left(\sqrt[N]{\prod_{i=1}^{n} Xi}\right) = 1/N\ (\Sigma \log x_i)$$

Autrement dit, (log G) est une moyenne arithmétique des logarithmes de la variable x_i.

Exemple :

Soit la série suivante : 2 ; 3 ; 4 ; 5 ; 6. Calculer sa moyenne géométrique.

$G = \sqrt[N]{x_1.\ x_2.\ x_3.....x_k} = \sqrt[5]{2.3.4.5.6} = \sqrt[5]{720} = 3,73$ **$G = 3,73$**

- **La moyenne géométrique pondérée**

Dans ce cas, à chaque modalité est associé un effectif. On écrit alors :

$$G = \sqrt[N]{x_1^{n1}.\ x_2^{n2}.\ x_3^{n3}.....x_k^{nk}} = \sqrt[N]{\prod_{i=1}^{k}(Xi)^{ni}} = \left[\prod_{i=1}^{k}(xi)^{ni}\right]^{1/N}$$

$G = \prod_{i=1}^{k}(xi)^{fi}$. On peut écrire aussi :

$$\log G = 1/N\sum_{i=1}^{k} nilogxi = \sum_{i=1}^{k} filogxi$$

Exemple :

Soit la distribution suivante :

Xi	1	2	3	4
ni	4	6	8	2

- Calculer la moyenne géométrique.

$G = \sqrt[N]{x_1^{n1}.\ x_2^{n2}.\ x_3^{n3}.....x_k^{nk}} = \sqrt[20]{1^4.\ 2^6.\ 3^8.4^2} = 2,195$ $G = 2,195$

1.5.2- La moyenne harmonique

a. Définition

Notée « **H** », la moyenne harmonique est égale à l'inverse de la moyenne arithmétiques des inverses des x_i.

On l'utilise surtout pour le calcul des moyennes des rapports, notamment les vitesses et les densités moyennes.

b. Méthode de calcul

➢ **Moyenne harmonique simple**

$$H = N / \sum_{i=1}^{n}\left(\frac{1}{Xi}\right) = N / \frac{1}{X1} + \frac{1}{X2} + \cdots \frac{1}{xk} \longrightarrow \mathbf{H = \frac{N}{\sum(\frac{1}{xi})}}$$

Exemple : Trois camions effectuent le trajet Tizi Ouzou- Alger aux vitesses moyennes de 40 km/h pour le premier, 60 km/h pour le deuxième et 80 km/h pour le troisième camion.

- Quelle est la vitesse moyenne sur l'ensemble des 3 trajets ?

On sait que : La vitesse moyenne = $\frac{\text{Distance D}}{\text{temps T}}$ $\boxed{V_m = \frac{D}{T}}$

$V_m = \frac{D1+D2+D3}{T1+T2+T3}$; $T = \frac{D}{Vi}$, (D est la même dans cet exemple)

Alors : $\frac{D+D+D}{T1+T2+T3} = \frac{D+D+D}{\frac{D}{V1}+\frac{D}{V2}+\frac{D}{V3}}$, on met « D » en facteur : $Vm = \frac{D(3)}{D(\frac{1}{V1}+\frac{1}{V2}+\frac{1}{V3})} = \frac{3}{\frac{1}{V1}+\frac{1}{V2}+\frac{1}{V3}}$

Vm est la moyenne harmonique simple des vitesses moyennes, $Vm = \frac{3}{\frac{1}{40}+\frac{1}{60}+\frac{1}{80}} = 55{,}38$ km/h

➢ **Moyenne harmonique pondérée**

$$H = N/ \left(\sum_{i=1}^{k} ni/xi\right) = N /\Sigma\ (n_1/x_1 + n_2/x_2 + \ldots + n_K/x_K) \longrightarrow \mathbf{H = \frac{N}{\sum(\frac{ni}{xi})}}$$

Exemple :

Un courtier vend pour 20.000€ d'actions, au cours de 20€ l'action. Il vend une seconde fois pour 10.000€ d'actions au cours de 14€ l'action. Quel est le cours moyen subit par le courtier sur l'ensemble des deux opérations ?

Le cours moyen est = $\frac{\text{valeur totale des actions}}{\text{nombre total d'actions}}$

Soient :

Vi = valeur de l'action lors de chaque opération ⟶ la valeur totale des actions = ΣVi

Qi = nombre d'actions vendues par opération ⟶ le nombre total des actions vendues = ΣQi.

Le cours moyen (CM) sur les deux opérations sera :

$$CM = \frac{\Sigma Vi}{\Sigma Qi} \text{.....(1)}$$

Cependant, nous connaissons ΣVi mais nous ne connaissons pas ΣQi.

Dans chaque opération : $CMi = \frac{Vi}{Qi}$ ⟶ $Qi = \frac{Vi}{CMi}$; on remplace dans (1) :

$CM = \frac{\Sigma Vi}{\Sigma(\frac{Vi}{CMi})}$, on retrouve donc la formule de la moyenne harmonique : $H = \frac{N}{\sum(\frac{ni}{xi})}$

Donc, le cours moyen que nous recherchons est une moyenne harmonique.

$$CM = \frac{20.000 + 10.000}{\frac{20.000}{20} + \frac{10.000}{14}} = \underline{\mathbf{17,50€.}}$$

Remarque : On vérifie toujours la relation suivante :

$$\mathbf{H < G < \overline{X}}$$

Section 2 : Les paramètres de dispersion

La portée de l'analyse des séries statistiques à l'aide des paramètres de tendance centrale est limitée et insuffisante. Car, il peut exister des distributions qui présentent les mêmes paramètres de position ($\overline{X} = M_o = M_e$) et qui sont en réalité différentes en terme de dispersion des données.

Après les paramètres de tendance centrale étudiés dans la section précédente, il s'agit dans la présente section d'étudier d'autres paramètres qui consistent à évaluer ou à calculer l'éloignement des valeurs par rapport à leur valeur centrale, le plus souvent leur moyenne arithmétique. Ces paramètres sont appelés les paramètres de « *dispersion* ». Un paramètre de dispersion est un nombre qui permet d'estimer dans quelle mesure des observations s'écartent les unes des autres ou bien s'écartent de leurs valeurs centrales. Il existe plusieurs mesures de la dispersion, les plus courantes sont :

2.1 L'étendue :

Notée (e), elle est la différence entre la plus grande et la plus petite valeur de la série statistique ordonnée par ordre croissant. On comprend, par conséquent, qu'elle est sujette à des fluctuations considérables d'un échantillon à un autre et très sensible aux influences des valeurs extrêmes aberrantes. Aussi, on ne l'utilise que pour avoir une idée sommaire et rapide de la dispersion de la série. Pour éviter l'influence des valeurs extrêmes aberrantes, on choisit de les écarter de la série, on a alors recours aux intervalles inter-quantiles ; on perd en informations mais on gagne en homogénéité.

2.2 L'intervalle inter-quartile [Q_1 , Q_3] : Contrairement à l'étendue, l'intervalle interquartile élimine les valeurs extrêmes aberrantes, il faut savoir que ce paramètre est aussi imparfait. Il est simple et rapide à calculer et à interpréter mais il a l'inconvénient de ne tenir compte que de la position des modalités et pas de leurs valeurs. Or, il est indispensable d'avoir recours à des paramètres permettant de tenir compte de toutes les modalités de la série. C'est ce que nous permettent les écarts autour d'une valeur centrale ou *écarts moyens*.

Il existe plusieurs paramètres pour calculer ces écarts moyens. Au préalable, il faut savoir qu'en statistique la notion d'« écart » désigne la distance (Xi – X). L'écart moyen désigne la somme de ces (Xi – X) divisé par l'effectif total (N). Cependant, comme nous l'avons vu dans le section précédente, à savoir ; la première propriété de $\overline{X}$, la somme $\Sigma(Xi - X)$ est nulle. Aussi, pour contourner cette nullité, il existe mathématiquement deux moyens possibles :

- considérer la valeur absolue des écarts : $| Xi - \overline{X} |$, afin d'éviter les écarts négatifs,
- considérer les carrés des écarts $(Xi - \overline{X})^2$, permettant également d'éviter les écarts négatifs.

De ces deux possibilités résultent deux types d'écarts moyens : l'écart absolu moyen et la variance (et l'écart type).

2.3 L'écart absolu moyen

Noté **Ex**, il désigne la moyenne arithmétique des valeurs absolues des écarts des modalités par rapport à leur moyenne arithmétique $|Xi - \overline{X}|$. On écrit alors :

$$\mathbf{Ex} = \frac{\Sigma|Xi - \overline{X}|}{N}$$ pour l'écart absolu moyen simple.

$$\mathbf{Ex} = \frac{\Sigma|Xi - \overline{X}|ni}{N}$$ **ou** $\mathbf{E_x = \Sigma|Xi - \overline{X}|fi}$ pour l'écart absolu moyen pondéré.

On peut également déterminer la moyenne arithmétique des valeurs absolues des écarts des modalités par rapport à leur médiane qu'on appelle ***écart absolu médian***. On écrit alors :

$$\mathbf{E_{Me}} = \frac{\Sigma|Xi - Me|}{N}$$ pour l'écart absolu médian simple.

$$\mathbf{E_{Me}} = \frac{\Sigma|Xi - Me|ni}{N}$$ **ou** $\mathbf{E_{Me} = \Sigma|Xi - Me|fi}$ pour l'écart absolu médian pondéré.

L'écart absolu moyen est un bon paramètre si ce n'est la lourdeur des valeurs absolues à trainer dans les calculs. Aussi, lui préfère-t-on la *variance* et l'*écart type*.

2.4 La variance et l'écart type

L'autre façon d'éviter les écarts négatifs, comme souligné plus haut, est l'élévation au carré. On obtient alors la variance et l'écart type.

2.4.1. La Variance : notée Vx, la variance est la moyenne arithmétique des carrés des écarts à la moyenne arithmétique. On écrit alors dans un premier temps les formules suivantes : appelée ***formules de définition*** :

$$\mathbf{Vx} = \frac{\Sigma(\mathbf{Xi} - \overline{\mathbf{X}})^{\mathbf{2}}}{\boldsymbol{N}} \quad \text{……. ……...Série simple.}$$

$$\mathbf{Vx} = \frac{\Sigma\mathbf{ni}(\mathbf{Xi} - \overline{\mathbf{X}})^{\mathbf{2}}}{\boldsymbol{N}} \quad \textbf{ou} \quad \Sigma\mathbf{fi}(\mathbf{Xi} - \overline{\mathbf{X}})^{2} \text{……...Série pondérée.}$$

En plus des formules de définition que l'on vient d'expliquer, il existe d'autres formules plus simplifiées et moins lourdes permettant un calcul plus rapide avec des risques d'erreurs de calcul minimisés. Ces formules résultent du développement mathématique des formules de définition appelées de ce fait, ***formules développées ou simplifiées***

On aura donc : : $\mathbf{Vx} = \left[\frac{\sum \mathbf{xi}^{2}}{\mathbf{N}}\right] - \overline{\mathbf{X}}^{2}$ ……..Série simple

$$\mathbf{Vx} = \left[\frac{\sum \mathbf{ni}(\mathbf{xi}^{2})}{\mathbf{N}}\right] - \overline{\mathbf{X}}^{2} \quad \textbf{ou } \mathbf{V_x} = \Sigma\mathbf{fi}(\mathbf{xi}^{2}) - \overline{\mathbf{X}}^{2} \text{ …….Série pondérée}$$

2. 4.2. L'Ecart type

Noté δx, il constitue le paramètre de dispersion le plus pertinent. Il est la racine carré de la variance. On écrit alors :

$$\boldsymbol{\delta}\mathbf{x} = \sqrt{\boldsymbol{Vx}} = \sqrt{\frac{\Sigma(\mathbf{Xi} - \overline{\mathbf{X}})^{2}}{\boldsymbol{N}}} \quad \text{……..……. Série simple.}$$

$$\boldsymbol{\delta}\mathbf{x} = \sqrt{\boldsymbol{Vx}} = \sqrt{\frac{\Sigma\mathbf{ni}(\mathbf{Xi} - \overline{\mathbf{X}})^{2}}{\boldsymbol{N}}} \quad \textbf{ou} \quad \sqrt{\sum \boldsymbol{fi}(\boldsymbol{xi} - \overline{\boldsymbol{X}})^{2}} \text{……Série pondérée}$$

$$\boldsymbol{\delta}\mathbf{x} = \sqrt{\left[\frac{\sum \mathbf{ni}(\mathbf{xi}^{2})}{\mathbf{N}}\right] - \overline{\mathbf{X}}^{2}} \qquad \text{(formule développée ou simplifiée)}$$

δx et Vx sont deux indicateurs de dispersion de même nature, puisque l'un est tout simplement la racine carré de l'autre.

Les paramètres déterminés précédemment, notamment $\overline{X}$ et δx sont de même nature que la variable Xi et sont exprimés dans la même unité de mesure. Cependant, il arrive qu'on compare deux ou plusieurs séries statistiques exprimées dans des unités de mesure différentes.

Pour pouvoir faire ces comparaisons sans difficulté, on calcule un paramètre appelé « *Coefficient de variation* ».

2.5 Le Coefficient de variation : noté C.V, il est un nombre sans dimension (ou sans unité de mesure) exprimé de ce fait en pourcentage. Il mesure le rapport de la moyenne arithmétique à l'écart-type : $\mathbf{C.V = \frac{\delta x}{\bar{X}} . 100}$ Plus ce pourcentage (C.V) est élevé, plus la dispersion est forte, et inversement.

Exercice :

La distribution des salaires (en 10^3DA) des employés d'une entreprise est donnée comme suit :

Salaire (10^3)	Effectif
12 - 16	26
16 - 20	33
20 - 24	64
24 - 28	7
28 - 32	10
Total	**140**

Questions :

- Calculer l'écart absolu moyen
- Calculer la variance et l'écart-type de cette distribution en utilisant les formules de définition et les formules développées.
- Calculer le Coefficient de variation de cette série, puis, Comparer la dispersion des salaires de notre entreprise avec celle d'une autre entreprise dont le salaire moyen est de 30. 10^3 DA et l'écart-type est de 15,68 DA.

Réponses :

Salaire (10^3)	n_i	Xi	nixi	$\lvert Xi - \bar{X}\rvert$	$ni\lvert Xi - \bar{X}\rvert$	$(Xi - \bar{X})^2$	$ni(Xi - \bar{X})^2$	Xi^2	$ni(Xi^2)$
12 - 16	26	14	364	6,34	164,84	40,196	1045,086	196	5090
16 - 20	33	18	594	2,34	77,22	5,4756	180,695	324	10692
20 - 24	64	22	1408	1,66	106,24	2,7556	176,358	484	30976
24 - 28	7	26	182	5,66	39,62	32,0356	224,2492	676	4732
28 - 32	10	30	300	9,66	96,6	93,3156	933,156	900	9000
Total	**140**	**-**	**2848**	**-**	**484,52**	**-**	**2559,5442**	**-**	**60496**

1. **Ecart absolu moyen** : $\mathbf{Ex = \frac{\Sigma|Xi - \bar{X}|ni}{N}}$

$\bar{X} = \frac{\Sigma nixi}{N} = \frac{2848}{140} = \underline{20,343} \longrightarrow \underline{\bar{X} = 20,343}$

$$Ex = \frac{484,52}{140} = 3,46 \longrightarrow \underline{\boldsymbol{Ex = 3,46 .10^3 DA.}}$$

2. a) La Variance : $\mathbf{Vx = \frac{\Sigma ni(Xi - \overline{X})^2}{N}}$ **(formule de définition)**

$$\longrightarrow \quad Vx = \frac{2559,5442}{140} \longrightarrow \underline{\boldsymbol{Vx=18,282 .10^3 DA.}}$$

L'Ecart-type : $\boldsymbol{\delta x = \sqrt{Vx}} = \sqrt{18,282} = 4,275 \longrightarrow \underline{\boldsymbol{\delta x =4,275 .10^3 DA.}}$

b) La Variance : $\mathbf{Vx = \left[\frac{\sum ni(xi^2)}{N}\right] - \overline{X}^2}$ **(formule développée)**

$$Vx = \frac{60496}{140} - (20,343)^2 \longrightarrow \underline{\boldsymbol{Vx = 18,28 .10^3}}$$

L'Ecart-type : $\boldsymbol{\delta x = \sqrt{Vx}} = \sqrt{18,28} = 4,275 \longrightarrow \underline{\boldsymbol{\delta x = 4,275 .10^3.}}$

3. Calculons d'abord le Coefficient de variation de notre entreprise que l'on va noter CV_A :

$$\mathbf{CV_A = \frac{\delta x_A}{\bar{X}_A} . 100} = \frac{4,27}{20,34} . 100 = 21\%. \longrightarrow \underline{\boldsymbol{CV_A = 21\%}}$$

Pour pouvoir comparer, il faut calculer le Coefficient de variation de l'autre entreprise et qu'on va noter CV_B.

$$\mathbf{CV_B = \frac{\delta x_B}{\bar{X}_B} . 100} = \frac{5,68}{30} . 100 = 52,27\%. \longrightarrow \underline{\boldsymbol{CV_B = 52,27\%}}$$

CV_B étant supérieur à CV_A, on en déduit que la dispersion des salaires est plus élevée au niveau de l'entreprise (B).

Section 3 : Les paramètres de concentration

3.1 Méthode de calcul

L'analyse algébrique de la concentration revient à calculer et à comparer des paramètres, en suivant les quatre étapes suivantes :

1- Calculer le paramètre « *Médiane », noté* (Me).

2- Calculer le paramètre «*Médiale* », noté (ML) .

3- Calculer l'écart Médiale-Médiane, noté $\Delta M = |ML - Me|$.

4- Comparer ΔM à l'étendue de la distribution : [ΔM/e]. 100, (exprimé en %).

Les notions de Médiane et d'étendue étant déjà définie dans les chapitres précédents, il nous faut, cependant, définir la notion de « Médiale ».

3.1.1- Notion de Médiale

Notée ML, la Médiale est la modalité de série qui partage la somme des masses ou la masse totale, $(\Sigma n_i x_i)$, en deux parties identiques : $(\Sigma n_i x_i)/2$ chacune. Elle se détermine suivant le même principe que la médiane.

a. Calcul de la Médiale

Mis à part le fait que l'on se réfère à la colonne des $(n_i x_i)$ cumulés, notée $(n_i x_i)^{\nearrow}$, on calcule la Médiale en suivant les même étapes que la médiane :

- Calculer $TH_L = (\Sigma n_i x_i)/2$, il représente le $(n_i x_i)$ cumulé que l'on déterminera dans la colonne des $(n_i x_i)^{\nearrow}$. En terme relatif TH_L est toujours égale à 0,5, soit 50%, ce qui correspond à la fréquence cumulée des $(n_i x_i/\Sigma n_i x_i)^{\nearrow}$ ou $F_i' = 0,5$ ou 50%.
- Déterminer la classe correspondant à la position TH_L. On l'appelle la *classe médiale*.
- Appliquer la formule de la Médiale suivante :

$$\mathbf{ML = X_0 + a\left[\frac{THL - (n_i x_i)^{\nearrow}_{ML-1}}{(n_i x_i)_{ML}}\right]}$$

Ou bien :

$$\mathbf{ML = X_0 + a\left[\frac{0,5 - f'^{\nearrow}_{ML-1}}{f'_{ML}}\right]}$$

Avec:

X_0 = Borne inférieure de la classe médiale.

a = amplitude de la classe médiale.

$(n_i x_i)_{ML}$ = masse de la classe correspondant à la classe médiale.

$(n_i x_i)_{ML-1}$ = masse <u>cumulée</u> correspondant à la classe avant la classe médiale.

$f'^{\nearrow}_{ML-1} = (n_i x_i/\Sigma n_i x_i)^{\nearrow}$ = fréquence <u>cumulée</u> des masses de la classe avant la classe médiale. Elle est toujours égale à 0,5, soit 50%.

$f'_{ML} = (n_i x_i/\Sigma n_i x_i)$ = Fréquence <u>relative</u> de la masse de la classe correspondant à la classe médiale.

b. L'écart Médiale-Médiane

Noté ΔM, il mesure la différence $|ML - Me|$. Cet écart est généralement exprimé en valeur absolu pour rappeler qu'il est positif, puisque la Médiale est toujours supérieure à la Médiane.

3.1.2 Indicateur de concentration (ΔM/e)

Le rapport (ΔM/e) est un nombre sans dimension, exprimé en pourcentage (%), sert à donner une première idée, voire un premier constat sur la concentration des modalités. Ainsi, c'est autour de la valeur 50% que l'on évalue, à priori, la concentration :

- Si (ΔM/e) est inférieur à 50%, la concentration est dite ***faible***.
- (ΔM/e) est supérieur à 50%, la concentration est dite ***forte***.
- (ΔM/e) est égale à 1, cela signifie que ΔM = e, dans ce cas la concentration est dite ***nulle***. C'est une situation de parfaite répartition des modalités ou d'« équi-répartition ».

Remarque

La méthode algébrique nous donne un premier aperçu de la concentration des modalités. Cet aperçu est confirmé par la suite et mesuré avec plus de précision grâce à la méthode pratique.

Exemple 1

Le tableau suivant donne la répartition des salaires (en 10^3DA) dans une entreprise :

Salaire 10^3DA	0 - 4	4 - 8	8 - 12	12 - 16	16 - 22	22 - 30	30 - 42
Nombre d'employés	6	25	24	17	14	11	3

- Analyser la concentration des salaires en utilisant la méthode algébrique.

Réponse

Pour faire cette analyse, nous allons suivre les quatre étapes de la méthode algébrique. Mais, auparavant, il faut d'abord compléter le tableau statistique avec toutes les colonnes nécessaires dont nous avons besoins en fonction des formules que l'on va utiliser. Le tableau complet sera comme suit :

Classes	x_i	n_i	$n_i\uparrow$	f_i	$f_i\uparrow$ *ou bien* (F_i)	$n_i x_i$	$f_i' = nixi/\Sigma nixi$	$f_i'\uparrow = (nixi/\Sigma nixi)\uparrow$ *ou bien* (F'_i)
0 - 4	*2*	*6*	*6*	*0,06*	*0,06*	*12*	*0,0092*	*0,0092*
4 - 8	*6*	*25*	*31*	*0,25*	*0,31*	*150*	*0,1154*	*0,1246*
8 - 12	*10*	*24*	***55***	*0,24*	***0,55***	*240*	*0,1846*	*0,3092*
12 - 16	*14*	*17*	*72*	*0,17*	*0,72*	*238*	*0,1831*	*0,4923*
16 - 22	*19*	*14*	*86*	*0,14*	*0,86*	*266*	*0,2046*	***0,6969***

22 - 30	*26*	*11*	*97*	*0,11*	*0,97*	*286*	*0,2200*	*0,9169*
30 - 42	*36*	*3*	*100*	*0,03*	*1*	*108*	*0,0831*	*1*
Total	*-*	*100*	*-*	*1*	*-*	*1300*	*1*	*-*

1- Calcul de la médiane

$Th2 = N/2 = 50 \longrightarrow Me \in [8 - 12[$

$Me = 8 + 4 \left(\frac{50 - 31}{24}\right) =$ ***11,17. 10^3 DA***.

2- Calcul de la Médiale

$TH_L = \Sigma nixi/2 = 1300/2$ 650 (nixi) = 650) ; Ou directement $TH_L = 0,5$ (F' = 0,5).

$\longrightarrow ML \in [16 - 22[$

$ML = 16 + 6 \left(\frac{05 - 04923}{02046}\right) =$ ***16,23. 10^3 DA***

3- Comparer ΔM à l'étendue de la distribution

$\Delta M = |ML - Me| = |16,23 - 11,17| = 5,06.$

$(\Delta M/e). 100 = [506/(42\ 0)]. 100 =$ ***12,04%***

Donc (ΔM/e) est inférieur à 50%, la concentration est par conséquent ***faible****.*

Remarque

Ce résultat devra être confirmé par la méthode graphique. C'est l'objet du point suivant.

3.2 L'analyse graphique de la concentration

Cette méthode est complémentaire de la précédente. Elle est développée par les auteurs italiens Gini et Lorenz.

Elle consiste à représenter graphiquement la concentration en traçant une courbe, dite *courbe de Lorenz* ou *courbe de Gini*, ou encore, *courbe de concentration.* A partir de cette courbe est déduit, par des opérations géométriques, un paramètre permettant de mesurer l'intensité de la concentration, appelé « *indice de Gini* ».

3.2.1- La courbe de concentration

Cette courbe se trace sur un plan orthonormé, à partir des fréquences cumulées des masses (f'_i) notées aussi (F_i') et des fréquences cumulées des effectifs (f_i) notées aussi (F_i).

Les fréquences cumulées des individus, notées (Fi), sont portées sur l'axe horizontal (axe des abscisses). Les fréquences cumulées des masses, notées (F'_i), sont portées sur l'axe vertical (axe des ordonnées).

Les fréquences variant de 0 à 1, on obtient un carré, appelé carré de Gini, de coté égale à 1 et de surface également égale à 1.

En reliant les points de coordonnées (0 ;0) et (1 ;1), on obtient une diagonale qui divise le carré en deux triangles de même surface, soit ½ ou 0,5 chacun.

C'est à l'intérieur du triangle sous la diagonale que se trace la courbe de Gini.

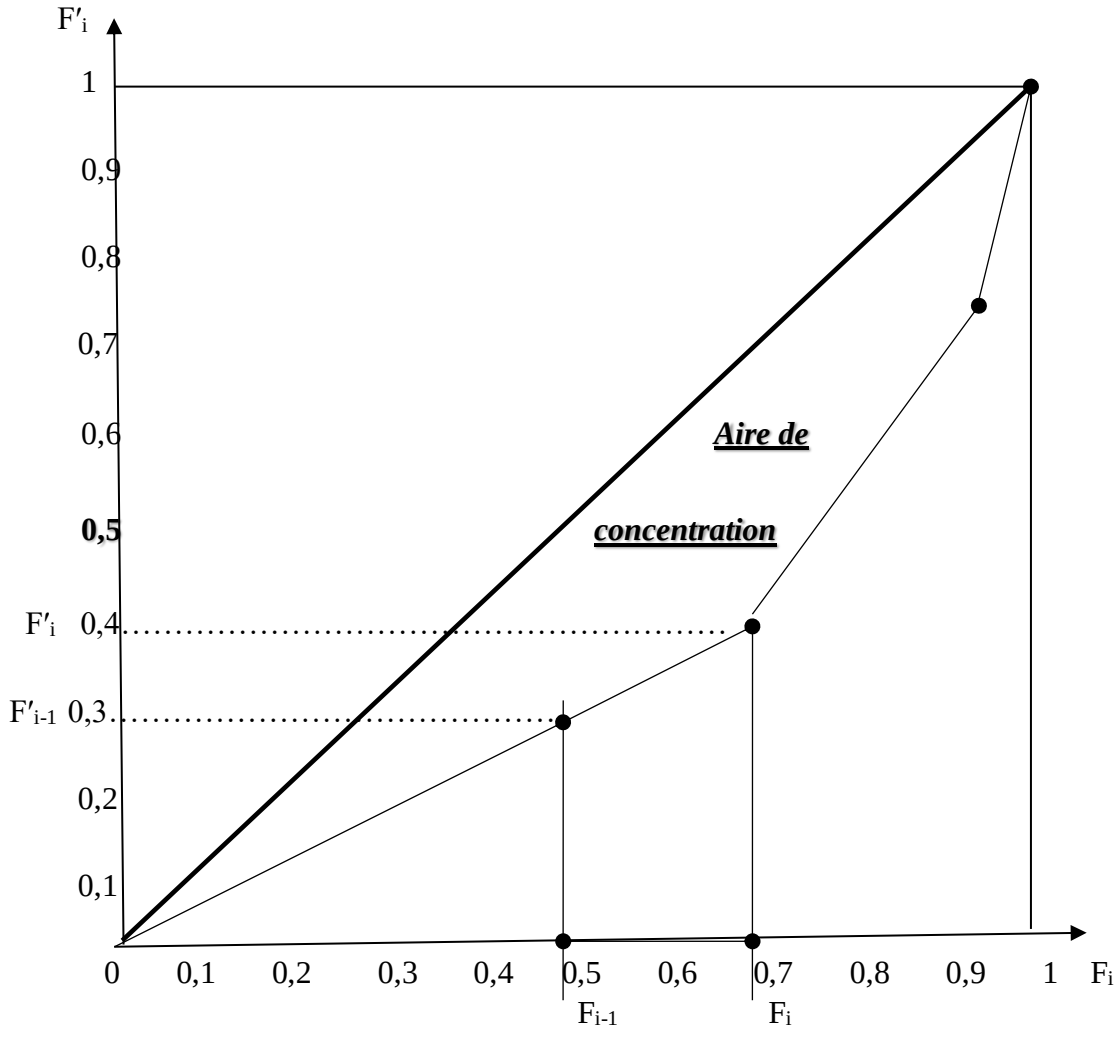
La diagonale représente la courbe où la concentration est nulle, c'est-à-dire, à chaque proportion d'individus correspond la même proportion de masse : $f_i = f'_i$. On l'appelle la droite d' « *équi-répartition* » ou droite de répartition équitable.

Par conséquent, plus la courbe de concentration se rapproche de la diagonale, plus la concentration est faible et plus l'*aire de concentration* se réduit.

Remarque

L'aire de concentration est la surface située entre la diagonale et la courbe de concentration.

La courbe de concentration se trace et se présente comme suit :



3.2.2- L'indice de Gini

Noté I_G, il mesure l'intensité de la concentration, en mesurant le rapport de l'aire de concentration à la surface du triangle du bas. On aura alors :

$$I_G = \frac{\text{Aire de concentration}}{1/2} \longrightarrow \boxed{I_G = 2.\ \text{Aire de concentration}}$$

$$\longrightarrow \boxed{I_G = 1 - \Sigma\,[(F'_{i-1}+F'_i)f_i]}$$

Interprétation de l'indice de Gini

I_G varie entre 0 et 1 : $I_G \in [0 - 1]$

- Si $I_G = 0$, ⟶ la concentration est nulle, l'aire de concentration est nulle. La courbe de concentration se confond avec la diagonale (la courbe de concentration est dans ce cas la diagonale elle même).
- Si = 1, la concentration est maximale, l'aire de concentration est égale à toute la surface du triangle sous la diagonale qui est elle même égale à 1/2.
  Dans ce cas, la courbe de concentration s'éloigne au maximum de la diagonale.
- Ainsi ;
  - Si I_G tend vers 1, ($I_G > 0,5$), la concentration est dite *forte*.
  - Si I_G tend vers 0, ($I_G < 0,5$), la concentration est dite *faible*.

2- Calcul de l'Indice de Gini

Pour cela il faut ajouter quelques colonnes au tableau précèdent, des colonnes pour les besoins de la formule de IG. Autrement dit, il faut ajouter trois autres colonnes, à savoir ;

F'_{i-1}	$F'_i+F'_{i-1}$	$(F'_i+F'_{i-1})f_i$
0	*0,0092*	*0,00052*
0,0092	*0,1338*	*0,03345*
0,1246	*0,4338*	*0,104112*
0,3092	*0,8015*	*0,136255*
0,4923	*1,1892*	*0,166488*
0,6969	*1,6139*	*0,177529*
0,9169	*1,9174*	*0,05752*
-	-	*0,6752*

Remarque

La colonne F'_{i-1} se détermine en remplaçant chaque valeur dans la colonne F'_i par sa valeur précédente (i par i_{-1}).

$IG = 1 - \Sigma (F'_{i-1} + F'_i) f_i) = 1 - 0{,}6752 = 0{,}325$

IG = 0,325

IG < 0,5 ⟶ *il en résulte que la concentration est faible. On confirme le résultat établi par la courbe de concentration et la méthode algébrique.*

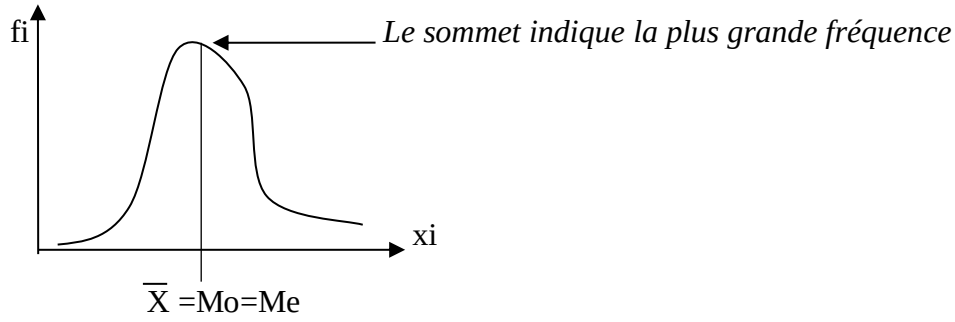
Section 4 : Les paramètres de forme

4.1 La mesure de la symétrie

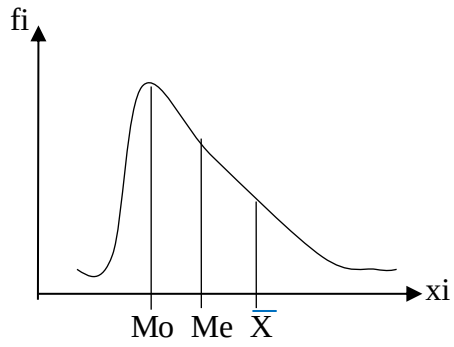
4.1.1. Définition

Une série statistique est dite symétrique si les modalités repérées par leurs fréquences sont également dispersées de part et d'autre de leur valeur centrale ($\bar{X}$, Me, Mo).

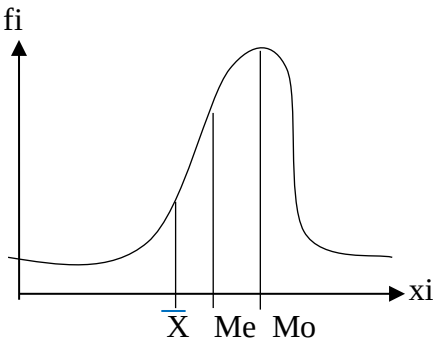
Lorsque la distribution est symétrique, les trois paramètres se confondent (ils sont égaux). La courbe des fréquences sera comme suit :



Lorsque la distribution statistique n'est pas symétrique, elle est dite *asymétrique* ou *oblique*. L'obliquité se repère du côté de la décroissance la plus forte (ou le coté tendant le plus vers la verticale) de la courbe des fréquences et l'étalement se repère du côté opposé.



Mo < Me < $\overline{X}$



Mo > Me > $\overline{X}$

4.1.2. Calcul des paramètres d'asymétrie

La statistique a mis au point plusieurs paramètres pour mesurer l'asymétrie d'une série. Ces paramètres sont appelés « coefficients d'asymétrie ». Nous en étudions le plus utilisé, à savoir ; le *coefficient de PEARSON.*

Pearson propose un coefficient qui analyse la position de deux valeurs centrales : $\overline{X}$ et Mo, relativisée par la dispersion de la série, mesurée par l'écart-type (δ_x). La formule est la suivante :

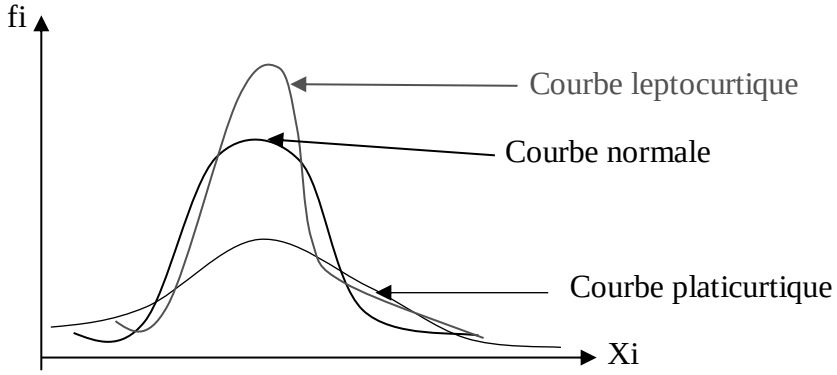
$$S = \frac{\overline{X} - Mo}{\delta_x}$$

- S = 0 ⟶ La distribution est symétrique, $\overline{X}$ = Mo (cette situation est très rare en pratique).
- S < 0 ⟶ $\overline{X}$ < Mo ⟶ La courbe est oblique à droite (le sommet de la courbe est située à droite après $\overline{X}$).
- S > 0 ⟶ $\overline{X}$ > Mo ⟶ La courbe est oblique à gauche (le sommet de la courbe est située à gauche avant $\overline{X}$).

4.2 La mesure de l'aplatissement

4.2.1 Définition

L'aplatissement indique si une faible variation des fréquences de la variable entraine ou non une forte variation des fréquences relatives et inversement. L'aplatissement étant mesuré par rapport à la courbe *normale* (appelée aussi courbe de la *loi normale*).



Ainsi, une distribution est dite aplatie (ou *platicurtique*) si une forte variation de la variable entraine une faible variation de la fréquence relative, et inversement.

4.2.2. Calcul des paramètres d'aplatissement

Comme pour la mesure de l'asymétrie, en statistique il existe plusieurs paramètres, appelés aussi coefficients, qui permettent de mesurer l'aplatissement d'une distribution statistique. La logique est de comparer si la distribution est plus ou moins aplatie par rapport à une courbe normale. Nous en proposons l'un des plus utilisés en statistique, à savoir *le coefficient d'aplatissement de Pearson*, il repose sur la notion de « *Moment* », plus précisément le *moment centré d'ordre 'r'*, noté $\boldsymbol{\mu}_r$, il est égal à : $\mu_r = \frac{\sum ni(Xi - \overline{X})^r}{N}$

Le coefficient d'aplatissement de Pearson, noté β_2, s'écrit comme suit :

$$\boldsymbol{\beta_2 = \mu_4/(\delta x)^4}$$

$\beta_2 = 3 \longrightarrow$ La distribution est normale.

$\beta_2 > 3 \longrightarrow$ La courbe est leptocurtique.

$\beta_2 < 3 \longrightarrow$ La courbe est platicurtique.

Conclusion :

Ce chapitre a permis d'approfondir l'analyse des séries statistiques en introduisant les principaux paramètres descriptifs. Chacun de ces indicateurs offre un éclairage complémentaire sur la distribution des données. Leur utilisation conjointe permet au statisticien de résumer efficacement une série, d'en comprendre la structure, et de mieux orienter les décisions ou les interprétations à tirer de l'observation statistique. Dans la continuité de cette analyse, le chapitre suivant portera sur les indices, outils fondamentaux pour mesurer les variations relatives d'un phénomène dans le temps ou dans l'espace.

Chapitre 4 : Les indices

Introduction :

Dans l'analyse des faits économiques et sociaux, il est fréquent de vouloir observer l'évolution de certaines grandeurs dans le temps ou d'en comparer les niveaux entre différentes zones géographiques. Qu'il s'agisse du prix du blé, de la quantité de maïs produite ou du volume des exportations automobiles, ces comparaisons s'appuient souvent sur le calcul de rapports, appelés *indices statistiques élémentaires*. Toutefois, il est parfois nécessaire d'apprécier l'évolution de grandeurs plus complexes — telles que le niveau général des prix ou la production industrielle. Ces dernières nécessitent l'élaboration d'*indices synthétiques*, lesquels condensent l'information de plusieurs indicateurs simples afin de révéler une tendance centrale globale.

Section 1 : Les indices élémentaires ou simples :

1.1 Définition :

Un indice simple noté $I_{t/0}$ de la grandeur G est le rapport de la valeur G_t, prise par la grandeur à l'époque t, à la valeur G_0 prise à la date 0, soit : $I_{t/0} = \frac{Gt}{G0} \times 100$.

Cet indice exprime donc la variation ou l'évolution de la grandeur G entre deux dates t et 0. La date t est la date finale (courante), la date 0 est la date de référence ou de base.

Dans le cas de comparaison géographique ou comparaison dans l'espace, on note l'indice de la région A par rapport à la région B : $I_{A/B}$. $I_{A/B} = \frac{GA}{GB} \times 100$.

B : région de référence ; A : région courante.

Applications :

a/ Indice dans le temps : le prix d'un litre d'huile est passé de 250 DA à 400 DA entre 2005 et 2008. L'indice du prix d'huile de l'année courante (2008) par rapport à l'année de base (2005) est : $I_{p2008/2005} = \frac{\text{prix en 2008}}{\text{prix en 2005}} \times 100 = \frac{400}{250} \times 100 = 160$ ($I_{p2008/2005}$ =1,60 ou 160). Cela signifie que le prix a été multiplié par 1,60 entre 2005 et 2008 ou bien le prix a augmenté de 60% [(1,60-1) ×100] entre 2005 et 2008, c-à-d en calculant le taux d'évolution ou le taux de croissance.

b/Indice dans l'espace : la densité de la population au KM^2 en 2003 est de 416H/ KM^2 sur l'ensemble de la wilaya de Tizi-Ouzou (WTO) et de 1319 H/ KM^2 pour la commune de Tizi-Ouzou (CTO). L'indice de la densité de la commune, la wilaya étant choisie comme la base est égal à : $I_{CTO/WTO} = \frac{1319}{416} = 3{,}17 = 317\%$. La densité de la commune représente 3,17 fois celle de la wilaya.

Remarque : on peut calculer l'indice de la densité de la wilaya en prenant comme base la commune.

1.2 Les propriétés des indices élémentaires :

1.2.1 La circularité ou la transférabilité : elle s'exprime de la manière suivante :

$I_{t/0} = I_{t/t-1} \times I_{t-1/t-2} \times I_{t-2/t-3} \times \ldots\ldots \times I_{2/1} \times I_{1/0}$

Ceci est aussi appelé le principe d'enchainement des indices.

Exemple : le chiffre d'affaires (CA) d'une entreprise a augmenté de 30% de 2010 à 2011 et diminué de 15% de 2011 à 2012. Le CA a-t-il diminué ou augmenté de 2010 à 2012 ?

Réponse :

On a $I_{11/10}$ = 100+30=130, $I_{12/11}$ = 100-15=85 et on cherche $I_{12/10}$ par la propriété de la circularité. $I_{12/10} = I_{12/11} \times I_{11/10}$= 0,85x 1,30= 1,105. Le CA a augmenté de [(1,105-1) x100] = 10,5 % de 2010 à 2012.

1.2.2 La réversibilité : elle s'exprime de la manière suivante : $I_{0/t} = \frac{1}{It/0}$

Exemple :

Si le prix d'un produit augmente de 20% de 2010 à 2012, calculer $I^{p}_{2010/2012}$.

Réponse :

On a $I^{p}_{12/10}$ = 1,20 ou 120 et on cherche $I^{p}_{10/12}$ par la propriété de réversibilité.

$I^{p}_{10/12} = 1/ I^{p}_{12/10} = 1/ 1,20= 0,83$. Cela veut dire que le prix de 2010 est inférieur à celui de 2012 de 17% soit [(0,83-1) x100]

1.2.3 L'identité: elle s'exprime de la manière suivante : $I_{0/0} = I_{t/t} = 1$ ou 100

Remarque : il est possible de calculer les indices élémentaires de prix, de quantité, de valeur et de pouvoir d'achat.

$I_{valeur} = I_{prix} \times I_{quantité}$

$I_{pouvoir\ d'achat} = \frac{1}{Iprix}$

Exemple 1 :

Soit P et Q les prix et quantités d'un produit vendu par une entreprise. Si le prix de ce produit a augmenté de 60% de 2000 à 2010 et si les quantités vendues ont diminué de 50% de 2000 à 2010, quelle est l'′évolution des recettes de 2000 à 2010 ?

Réponse :

$I^{v}_{10/00} = I^{P}_{10/00} \times I^{Q}_{10/00}$ =1,60x0, 50= 0,8 soit une baisse des recettes (valeur) de 20% [(0,80-1) ×100

Exemple 2 : si les prix augmentent de 10%, comment varie le pouvoir d'achat ?

Réponse :

$I_{pouvoir\ d'achat} = \frac{1}{Iprix}$; I_p= 1+0,10= 1,10. $I_{pouvoir\ d'achat} = \frac{1}{1,10}$ = 0,9090. Cela veut dire que le pouvoir d'achat baisse de 9,1% [(0,9090-1) ×100]

Section 2 : Les indices synthétiques:

Les *indices synthétiques* permettent de mesurer l'évolution d'une grandeur composite, c'est-à-dire composée de plusieurs éléments simples. Par exemple, la production agricole regroupe le blé, le maïs, etc. On distingue des indices de valeur**,** prix et quantité, chacun calculé selon trois méthodes classiques : *Laspeyres*, *Paasche* et *Fisher*.

2.1. L'indice de Laspeyres : c'est une moyenne arithmétique des indices élémentaires pondérés par les coefficients budgétaires α_0^j(coefficients de pondération) de la date ou période de base (0).

α_0^j est une fréquence relative. Elle représente l'importance du constituant j dans la grandeur complexe.

Exemple : la part dans la dépense totale de viande des ménages de chaque article entrant dans la catégorie de viande : le bœuf, le veau, le mouton, le cheval, le poulet, le chameau et le lapin.

$$\boldsymbol{\alpha_0^j = \frac{P_0^J . Q_0^J}{\sum P_0^j . Q_0^j}} \text{ ou } \boldsymbol{\alpha_0^j = \frac{P_0 . Q_0}{\sum P_0 . Q_0}}$$

$P_0^j . Q_0^j$: Valeur du produit j à l'année de base.

$\sum P_0^j . Q_0^j$: Total des valeurs des produits.

-L'indice de Laspeyres des prix : $L^P_{t/0} = \sum \alpha_0^j . I^P_{t/0}$. Cette formule est appelée « formule de définition ». Après simplification, on obtient la formule suivante : $L^P_{t/0} = \frac{\Sigma P_t . Q_0}{\Sigma P_0 . Q_0}$ (formule simplifiée).

-L'indice de Laspeyres des quantités : $L^Q_{t/0} = \sum \alpha_0^j . I^Q_{t/0}$ (formule de définition) ou

$L^Q_{t/0} = \frac{\Sigma P_0 . Q_t}{\Sigma P_0 . Q_0}$ (formule simplifiée).

Remarque : dans le cas où les coefficients budgétaires sont exprimés en pourcentage, la formule de définition devient : $L^P_{t/0} = \sum \alpha_0^j . I^P_{t/0}/100$. (Laspeyres des prix).

$L^Q_{t/0} = \sum \alpha_0^j . I^Q_{t/0}/100$ (Laspeyres des quantités).

2.2. L'indice de Paasche : c'est une moyenne harmonique des indices élémentaire pondérés par les coefficients budgétaires α^j_t de l'année courante.

$$\alpha_t^j = \frac{P_t^j . Q_t^j}{\sum P_t^j . Q_t^j} \text{ ou } \alpha_t^j = \frac{P_t . Q_t}{\sum P_t . Q_t}$$

-L'indice de Paasche des prix : $P^P_{t/0} = \frac{\sum \alpha_t^j}{\sum \alpha_t^j . \frac{1}{IPt/0}} = \frac{1}{\sum \alpha_t^j . \frac{1}{IPt/0}}$ (formule de définition). Après simplification, on obtient la formule suivante : $P^P_{t/0} = \frac{\Sigma P_t . Q_t}{\Sigma P_0 . Q_t}$ (formule simplifiée).

-L'indice de Paasche des quantités : $P^Q_{t/0} = \frac{1}{\sum \alpha_t^j . \frac{1}{IQt/0}}$ (formule de définition) ou bien

$P^Q_{t/0} = \frac{\Sigma P_t . Q_t}{\Sigma P_t . Q_0}$ (formule simplifiée).

2.3. L'indice de Fisher :

L'indice de Fisher est la moyenne géométrique simple des indices de Laspeyres et de Paasche.

-L'indice de Fisher des prix : $\mathbf{F^P = \sqrt{L^P . P^P}}$

-L'indice de Fisher des quantités : $\mathbf{F^Q = \sqrt{L^Q . P^Q}}$

2.4. L'indice des valeurs globales ($I^{VG}_{t/0}$) :

Nous avons calculé, précédemment, les indices de prix et de quantités de Laspeyres, de Paasche et de Fisher, avec des formules différentes. Par contre, les indices de valeur globales de Laspeyres, de Paasche et de Fisher se calculent tous de la même manière. Autrement dit

$$I^{VG}_{t/0} = \frac{\Sigma P_t.Q_t}{\Sigma P_0.Q_0} \times 100$$

Remarque :

On peut aussi calculer l'indice des valeurs globales comme suit :

$I^{VG}_{t/0} = L^{P}_{t/0} . P^{Q}_{t/0} = L^{Q}_{t/0} . P^{P}_{t/0}$

Exercice d'application :

Un responsable d'approvisionnement de rayon a relevé, au cours de deux années, les quantités et les prix de trois produits « A », « B » et « C » et a établi le tableau suivant :

Année	**Articles**	**Quantités (kg)**	**Prix (DA)**	**Valeurs. (Quantité x Prix)**
2000	A	15	10	150
	B	20	5	100
	C	25	8	200
2005	A	25	12	300
	B	25	6	150
	C	35	15	525

Questions :

1. Calculer les indices des prix de Laspeyres et de Paasche par les formules de définition et par les formules simplifiées.
2. Calculer les indices des quantités de Laspeyres et de Paasche par les formules de définition et par les formules simplifiées.
3. Calculer les indices de Fisher des prix et des quantités
4. Calculer l'indice des valeurs globales et vérifier que $I^{VG}_{t/0} = L^{P}_{t/0} . P^{Q}_{t/0} = L^{Q}_{t/0} . P^{P}_{t/0}$

1/Calcul des indices des prix :

	P_0	Q_0	P_t	Q_t	$P_0.Q_0$	α^j_0	$P_0.Q_t$	$P_t.Q_0$	$P_t.Q_t$	α^j_t	$I^P_{t/0}$	$I^P_{0/t}$	$I^Q_{t/0}$	$I^Q_{0/t}$
A	10	15	12	25	150	0,33	250	180	300	0,31	1,2	0,83	1,66	0,60
B	5	20	6	25	100	0,22	125	120	150	0,15	1,2	0,83	1,25	0,80
C	8	25	15	35	200	0,45	280	375	525	0,54	1,87	0,53	1,4	0,71
∑	/	/	/	/	450	1	655	675	975	1	/	/	/	/

A/L'indice de Laspeyres des prix :

-Formule de définition :

$L^{P}_{t/0}=\sum \alpha_0^j \,.\, I^{P}_{t/0}$ ➡ $L^{P}_{2005/2000}$ = (0,33.1, 2) + (0,22.1, 2) + (0,45.1,87) ≈ 1.5 ou 150

- Formule simplifiée :

$$L^{P}_{t/0}=\frac{\Sigma P_t.Q_0}{\Sigma P_0.Q_0} \Rightarrow L^{P}_{2005/2000}=\frac{675}{450}=1{,}5 \text{ ou } 150$$

Soit une augmentation des prix des trois (3) produits de 50%, entre 2000 et 2005.

B) *Indice de Paasche des prix*

- Formule de définition :

$$P^{P}_{t/0}=\frac{1}{\sum \alpha_t^j.\frac{1}{IPt/0}} \Rightarrow P^{P}_{2005/2000}=\frac{1}{(0{,}31.0{,}83)+(0.15.0{,}83)+(0{,}54.0{,}53)} \approx 1.49 \text{ ou } 149$$

-Formule simplifiée :

$$P^{P}_{t/0}=\frac{\Sigma P_t.Q_t}{\Sigma P_0.Q_t} \Rightarrow P^{P}_{2005/2000}=\frac{975}{655} \approx 1.49 \text{ ou } 149$$

Soit une augmentation des prix des trois (3) produits, de près de 49%, entre 2000 et 2005.

2/Indices des quantités :

***A/L'indice de* Laspeyres des quantités:**

-Formule de définition :

$L^{Q}_{t/0}=\sum \alpha_0^j \,.\, I^{Q}_{t/0}$ ➡ $L^{Q}_{2005/2000}$= (0,33.1, 66) + (0,22.1, 25) + (0,45.1, 4) = 1,45 ou 145

-Formule simplifiée)

$$L^{Q}_{t/0}=\frac{\Sigma P_0.Q_t}{\Sigma P_0.Q_0} \Rightarrow L^{Q}_{2005/2000}=\frac{655}{450}=1.45 \text{ ou } 145$$

Soit une augmentation des quantités demandées des trois (3) produits de 45%, entre 2000 et 2005.

B/ L'indice de Paasche des quantités :

-Formule de définition :

$$P^{Q}_{t/0}=\frac{1}{\sum \alpha_t^j.\frac{1}{IQt/0}} \Rightarrow P^{Q}_{2005/2000}=\frac{1}{(0{,}31.0{,}60)+(0{,}15.0{,}80)+(0{,}54.0{,}71)} = 1{,}44 \text{ ou } 144$$

-Formule simplifiée:

$$P^{Q}_{t/0}=\frac{\Sigma P_t.Q_t}{\Sigma P_t.Q_0} \Rightarrow P^{Q}_{2005/2000}=\frac{975}{675}= 1{,}44 \text{ ou } 144$$

Soit une augmentation des quantités demandées des trois (3) produits de 44%, entre 2000 et 2005.

3/Indices de Fisher :

A/Fisher des prix

$F^{P}=\sqrt{L^P.P^P}$ ➡ $F^{P}=\sqrt{1{,}5.1{,}49}$ = 1,495 ou 149,5

Soit une augmentation des prix des trois (3) produits, de près de 49%, entre 2000 et 2005.

B/Fisher des quantités

$F^Q = \sqrt{L^Q . P^Q}$ ➡ $F^Q = \sqrt{1{,}45.1.44} = 1{,}445$ ou 144,5

Soit une augmentation des quantités demandées des trois (3) produits de 44,5%, entre 2000 et 2005.

4/Indices des valeurs globales :

$$I^{VG}_{t/0} = \frac{\Sigma P_t . Q_t}{\Sigma P_0 . Q_0} = \frac{975}{450} = 2{,}16$$

On vérifie par ailleurs que :

$L^P_{t/0} . P^Q_{t/0} = L^Q_{t/0} . P^P_{t/0} = I^{VG}_{t/0}$ ➡ (1,5. 1,44)= (1,45. 1,49) = 2,16

Soit une augmentation de la valeur (dépenses) des trois (3) produits de 116% entre 2000 et 2005.

Conclusion :

Ce chapitre a mis en évidence l'importance des indices statistiques dans l'analyse des variations économiques et sociales. En distinguant les indices élémentaires des indices synthétiques, nous avons montré comment ces outils permettent de suivre l'évolution d'un phénomène dans le temps ou de comparer des situations dans l'espace. Leur bonne maîtrise est indispensable pour interpréter correctement les données économiques. Dans la suite, nous aborderons l'analyse de la corrélation, qui permet d'étudier les liens éventuels entre deux variables statistiques.

Chapitre 5 : Les distributions statistiques à deux caractères : étude de la régression, de l'ajustement et de la corrélation

Introduction

Dans les chapitres précédents, nous avons porté notre attention sur les distributions univariées, c'est-à-dire celles qui décrivent une population à partir d'un seul caractère. Or, dans de nombreuses situations, une population statistique peut être décrite simultanément à l'aide de deux, voire plusieurs caractères.

Dans ce présent chapitre, nous nous limiterons à l'étude de distributions à deux caractères, notés généralement **x** et **y**. Il peut s'agir, par exemple, de l'analyse d'une population de salariés selon :

- l'âge (**x**) et le salaire (**y**),
- le salaire (**x**) et le nombre d'enfants (**y**),
- ou encore, l'âge (**x**) et la qualification professionnelle (**y**).

De manière générale, ces deux caractères peuvent être quantitatifs**,** qualitatifs**,** ou bien l'un quantitatif et l'autre qualitatif. Leur mise en relation s'effectue à travers un tableau à double entrée**,** appelé aussi tableau de contingence, dans lequel les effectifs sont répartis selon les lignes et les colonnes correspondant aux modalités des deux caractères.

Par la suite, notre analyse portera plus particulièrement sur les relations susceptibles d'exister entre deux caractères quantitatifs. Nous chercherons à déterminer si ces caractères sont liés, autrement dit, si l'un influence l'autre, ou s'ils évoluent indépendamment. Ce type d'étude relève de ce qu'on appelle **l'**analyse de corrélation**.**

Section 1 : Les tableaux de contingence :

Dans cette section, nous allons explorer un outil fondamental de l'analyse statistique bivariée : le tableau de contingence. Il permet de représenter de manière synthétique la répartition conjointe de deux variables qualitatives.

1.1 Distributions conjointes, marginales et conditionnelles

Avant de tirer des conclusions à partir des données croisées dans un tableau de contingence, il est essentiel de comprendre les différents types de distributions que l'on peut en extraire. Nous commencerons par la distribution conjointe.

1.1.1 Distributions conjointes :

Soit une population composée de N individus sur lesquels on observe les variables x et y. Les k modalités de x sont désignées par x_1,x_2, …, $x_k$et les p modalités de y par y_1, y_2,…, y_p.

- La répartition de N observations appelée distribution conjointe se présente sous la forme d'un tableau (ci-dessous) à double entrée où figure en **ligne** les modalités de x et en **colonne** les modalités de y.

- L'effectif n_{ij} désigne le nombre de fois où les modalités de x et les modalités de y ont été observées simultanément.
- L'effectif $\boldsymbol{n_{i.}}$ désigne le nombre total d'observations des modalités de x quelle que soit la modalité de y et $n_{i.} = \sum n_{ij}$ $j = 1$ jusqu'à p ; « j » varie, donc devient « **.** »
- L'effectif $\boldsymbol{n_{.j}}$ représente le nombre total d'observations des modalités de y quelle que soit la modalité de x**;** et nous avons $\boldsymbol{n_{.j}} = \sum n_{ij}$ $i = 1$ jusqu'à k ; « i » varie, donc devient « **.** »

Le tableau de contingence peut être présenté comme suit :

$x \setminus y$	y_1	y_2	…	…	y_j	…	…	y_p	**Total**
x_1	n_{11}	n_{12}	…	…	n_{1j}	…	…	n_{1p}	$\boldsymbol{n_{1.}}$
x_2	n_{21}	n_{22}	…	…	n_{2j}	…	…	n_{2p}	$\boldsymbol{n_{2.}}$
⋮	⋮	⋮			⋮			⋮	⋮
x_i	n_{i1}	n_{i2}	…	…	n_{ij}	…	…	n_{ip}	$\boldsymbol{n_{i.}}$
⋮	⋮	⋮			⋮			⋮	⋮
⋮	⋮	⋮			⋮			⋮	⋮
x_k	n_{k1}	n_{k2}	…	…	n_{kj}	…	…	n_{kp}	$\boldsymbol{n_{k.}}$
Total	$\boldsymbol{n_{.1}}$	$\mathbf{n}_{.2}$	…	…	$\boldsymbol{n_{.j}}$	…	…	$\boldsymbol{n_{.p}}$	$\boldsymbol{n_{..}}$**=N**

1.1.2 Distributions marginales

Disposant d'une distribution conjointe, on peut déduire les distributions marginales qui permettent d'étudier séparément chaque variable, ce qui nous permettra de calculer leurs caractéristiques de tendance centrale ou de dispersion. Ainsi les k couples $(x_i, n_{i.})$ définissent la distribution marginale des observations selon la modalité $de\ x$quelle que soit la modalité de y. Cette distribution est représentée par la dernière **colonne** du tableau. De même les $p$couples$(y_j, n_{.j})$ définissent la distribution marginale des observations suivant la variable y quelle que soit la modalité de x. On obtient :

a. Distribution marginale de x :

x_i	x_1	x_2	…	x_i	…	x_k	Total
$n_{i.}$	$n_{1.}$	$n_{2.}$	…	$n_{i.}$	…	$n_{k.}$	$n_{..} = N$

b. Distribution marginale de$y$:

y_j	y_1	y_2	…	y_j	…	y_p	Total
$n_{.j}$	$n_{.1}$	$n_{.2}$	…	$n_{.j}$	…	$n_{.p}$	$n_{..} = N$

1.1.3. Distributions conditionnelles

La distribution conditionnelle correspondant à une modalité x_i de la variable $x$suivant les modalités de y est appelée distribution conditionnelle de y pour $x = x_i$. Cette distribution est donnée par le tableau suivant :

y / $x = x_i$	y_1	y_2	…	y_j	…	y_p	total
n_{ij}	n_{i1}	n_{i2}	…	n_{ij}	…	n_{ip}	$n_{i.}$

Symétriquement on peut définir la distribution conditionnelle de x pour $y = y_j$ comme suit :

x / $y = y_j$	x_1	x_2	…	x_i	…	x_k	Total
n_{ij}	n_{1j}	n_{2j}	…	n_{ij}	…	n_{kj}	$n_{.j}$

1.2. Notion de fréquences relatives :

Comme pour une distribution à un seul caractère, il est possible de calculer des fréquences relatives pour une distribution bi-variée.

1.2.1 Fréquences relatives partielles:

La fréquence relative partielle est définie comme étant le rapport du nombre d'individus possédant simultanément la modalité $x_i$de x et la modalité $y_j$de y sur l'effectif total.

$$f_{ij} = {}^{n_{ij}}/_{n_{..}} \text{ et } \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij} = 1$$

La somme des fréquences relatives partielles de tous les couples de valeurs$\{x_i, y_j\}$ est égale à un.

1.2.2 Fréquences relatives marginales

Pour la distribution marginale de x: $f_{i.} = \frac{n_{i.}}{n_{..}}$

Pour la distribution marginale de y : $f_{.j} = \frac{n_{.j}}{n_{..}}$

Et $\sum_{i=1}^{k} f_{i.} = 1 = \sum_{j=1}^{p} f_{.j}$. La somme des fréquences marginales est égale à un.

1.2.3 Fréquences relatives conditionnelles :

On a p fréquences relatives conditionnelles de $x$selon y puisque j varie de 1 jusqu'à p :

$f$de i si j $\quad f_{i/j} = \frac{n_{ij}}{n_{.j}}$

On a k fréquences relatives conditionnelles de y selon $x$puisque i varie de 1 jusqu'à k

f de j si i $f_{j/i} = \frac{n_{ij}}{n_{i.}}$

1.3. Les paramètres des lois marginales et conditionnelles :

1.3.1 Les paramètres des lois marginales (moyenne et variance)

a. Les paramètres des lois marginales selon x

Si x est un caractère quantitatif, on définit les paramètres marginaux (moyenne et variance) à partir de la **colonne** marginale où se trouvent les effectifs $n_{i.}$ correspondant respectivement aux k modalités de x.

- La moyenne marginale de x est $\bar{x}$. Elle est définie comme suit :

$$\bar{x} = \frac{1}{n..}\sum_{i=1}^{k} n_{i.}\, x_i = \sum_{i=1}^{k} f_{i.} x_i$$

Remarque: la moyenne marginale de x est notée $\bar{x}$ (x barre) ou $\bar{\bar{x}}$ (x deux barres).

- La variance marginale de x notée $V(x)$. On peut la calculer de deux manières :

Formule de définition : $V(x) = \frac{1}{n..}\sum_{i=1}^{k} n_{i.}\,(x_i - \bar{x})^2 = \sum_{i=1}^{k} f_{i.}\,(x_i - \bar{x})^2$

Formule développée : $V(x) = \frac{1}{n..}\sum_{i=1}^{k} x_i^2 n_{i.} - \bar{x}^2 = \sum_{i=1}^{k} f_{i.} x_i^2 - \bar{x}^2$

b.Les paramètres des lois marginales de y :

Si y est un caractère quantitatif, on définit les paramètres marginaux $\bar{y}$ et $V(y)$ à partir de la ligne marginale où se trouvent les effectifs $n_{.j}$ correspondant chacun respectivement aux p modalités y_j de y

- La moyenne marginale $\bar{y}$: $\bar{y} = \frac{1}{n..}\sum_{j=1}^{p} n_{.j} y_j = \sum_{j=1}^{p} f_{.j} y_j$

Remarque : la moyenne marginale de y est notée $\bar{y}$ (y barre) ou $\bar{\bar{y}}$ (y deux barres).

- La variance $V(y)$:

- Par la formule de définition: $V(y) = \frac{1}{n..}\sum_{j=1}^{p} n_{.j}\left(y_j - \bar{y}\right)^2 = \sum_{j=1}^{p} f_{.j}\left(y_j - \bar{y}\right)^2$ o u
- Par la Formule développée :$V(y) = \frac{1}{n..}\sum_{j=1}^{p}\left(n_{.j} y_j^2\right) - \bar{y}^2 = \sum_{j=1}^{p} f_{.j} y_j^2 - \bar{y}^2$

Pour les données groupées d'une distribution conjointe, on peut définir la covariance $V(xy)$comme suit :

- Formule de définition :

$$Cov(xy) = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{p}\left[\left(x_i - \bar{x}\right)\left(y_j - \bar{y}\right)\right] n_{ij}$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{p} \left[(x_i - \bar{x})(y_j - \bar{y}) \right] f_{ij}$$

- Formule développée :

$$Cov(xy) = \left(\frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{p} n_{ij} x_i y_j \right) - \bar{x}\bar{y}$$

$$= \left(\sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij} x_i y_j \right) - \bar{x}\bar{y}$$

1.3.2 les paramètres des lois conditionnelles

a. paramètres des distributions conditionnelles de x selon y

Il y a p distributions conditionnelles de x selon y auxquelles correspondent p paramètres conditionnels (p moyennes et p variances)

- Les moyennes conditionnelles de x selon y, $y = y_j$ (y_j fixe)

$$\bar{x}_j = \frac{1}{n_{.j}} \sum_{j=1}^{k} n_{ij} x_i = \sum_{i=1}^{k} f_{i/j}\, x_i$$

$\bar{x}_j$ est la moyenne conditionnelle de x sachant que y_j fixe ($y = y_j$)

- Les variances conditionnelles de x selon y ($y = y_j$)

Par définition : $V_j(x) = \frac{1}{n_{.j}} \sum_{i=1}^{k} \left[\left(x_i - \bar{x}_j \right)^2 n_{ij} \right]$

Ou$= \sum_{i=1}^{k} \left(x_i - \bar{x}_j \right)^2 f_{i/j}$

Formule développée:$V_j(x) = \frac{1}{n_{.j}} \sum_{i=1}^{k} n_{ij} x_i^2 - \bar{x}_j^2 = \sum_{i=1}^{k} x_i^2 f_{i/j} - \bar{x}_j^2$

b.Paramètres des distributions conditionnelles de y selon x

Ici on fixe $x = x_i$

De façon analogue, on a k distributions conditionnelles de $\boldsymbol{x}$ selon **y** auxquelles correspondent k paramètres conditionnels (moyennes, variances)

- Les moyennes conditionnelles de $\boldsymbol{y}$ selon $\boldsymbol{x}$

$$\bar{y}_i = \frac{1}{n_{i.}} \left[\sum_{j=1}^{P} n_{ij} y_j \right] = \sum_{j=1}^{P} f_{j/i} y_j$$

- Les variances de y selon x

Par définition

$$V_i(y) = \frac{1}{n_{i.}} \sum_{j=1}^{P} \left[\left(y_j - \bar{y}_j \right)^2 n_{ij} \right]$$

ou

$$= \sum_{j=1}^{P} \left(y_j - \bar{y}_j \right)^2 f_{j/i}$$

Formule développée :

$$V_i(y) = \frac{1}{n_{i.}} \sum_{j=1}^{P} \left(n_{ij} y_j^2 \right) - \bar{y}_j^2$$

Ou $= \sum_{j=1}^{P} \left(f_{j/i} y_j^2 \right) - \bar{y}_j^2$

Exercice d'application : Le tableau suivant donne la répartition de 1000 familles selon l'âge du père (X_i) et le nombre d'enfants (Yj)

Yj / Xi	Moins de 2 enfants	[2-5[	5 et plus	**Totaux ($n_{i.}$)**
Moins de 25ans	100	20	5	125
[25 -30[	50	25	15	90
[30-40[	30	100	100	230
40 et plus	20	200	335	555
Totaux ($n_{.j}$)	200	345	455	1000

Questions :

1- Quel est le nombre de familles ayant de 2 à 5 enfants et dont l'âge du père est compris entre 30 et 40 ans ?
2- Quel est le nombre de familles ayant moins de 2 enfants ?
3- Quel est le nombre de familles dont l'âge du père est égal à 40 ans et plus ?
4- Que signifie le nombre 5 de la première ligne et de la troisième colonne ?
5- Donner les valeurs de n_{11} ; n_{23} ; $n_{2.}$; $n_{.3}$
6- Déterminer la distribution marginale du caractère X.
7- Déterminer la distribution marginale du caractère Y.
8- Dégager la distribution conditionnelle de Y selon $X \in [30 - 40[$.
9- Dégager la distribution conditionnelle de X selon Y∈ [2-5[.
10- Calculer f_{33} ; $f_{2.}$; $f_{.3}$; $f_{1/2}$ avec i fixé et $f_{3/2}$ avec j fixé.

Réponses :

1- Le nombre de familles ayant de 2 à 5 enfants et dont l'âge du père est compris entre 30 et 40 ans est égal à 100 familles (l'effectif partiel $\mathbf{n_{32}}$**).**
2- Le nombre de famille ayant moins de 2 enfants est égal à 200 (l'effectif marginal de Y : $\mathbf{n_{.1}}$).

3- Le nombre de familles dont l'âge du père est égal à 40 ans et plus est 555 (L'effectif marginal de X : $\mathbf{n_{4.}}$)

4- Le nombre 5 de la première ligne et de la troisième colonne **($\mathbf{n_{13}}$)** représente le nombre de familles ayant 5 enfants ou plus et dont l'âge du père est inférieur à 25 ans.

5- $\mathbf{n_{11}}$= 100 ; $\mathbf{n_{23}}$= 15 ; $\mathbf{n_{2.}}$ = 90 ; $\mathbf{n_{.3}}$ = 455

6- Distribution marginale de X :

X âge du père	Moins de 25 ans	[25-30[	[30-40[	40 et plus	Total
$\mathbf{n_{i.}}$	125	90	230	555	1000

7- Distribution marginale de Y :

Y nombre d'enfants	Moins de 2 enfants	[2-5[	5 et plus	Total
$\mathbf{n_{.j}}$	200	345	455	1000

8- Distribution conditionnelle de Y selon X$\in [30-40[$:

Y/X$\in$ [30-40[	Moins de 2 enfants	[2-5[	5 et plus	Total
$\mathbf{n_{3j}}$	30	100	100	230 **($\mathbf{n_{3.}}$)**

9- Distribution conditionnelle de X selon Y$\in$ [2-5[:

X/Y$\in$ [2-5[	Moins de 25 ans	[25-30[	[30-40[	[40 et plus	Total
$\mathbf{n_{i2}}$	20	25	100	200	345 ($\mathbf{n_{.2}}$)

10- Les fréquences relatives :

$\mathbf{f_{33}} = \frac{n_{33}}{n_{..}} = \frac{100}{1000}$ = 0, 10 ou 10% (Fréquence partielle sur l'effectif total). Cela signifie qu'il y a 10% de familles ayant 5 enfants et plus et dont l'âge du père est compris entre 30 et 40 ans.

$\mathbf{f_{2.}} = \frac{n_{2.}}{n_{..}} = \frac{90}{1000}$ = 0,09 ou 9% (Fréquence marginale de X). Cela signifie qu'il y a 9% de familles dont l'âge du père est compris entre 25 et 30 ans quel que soit le nombre d'enfants.

$\mathbf{f_{.3}} = \frac{n_{.3}}{n_{..}} = \frac{455}{1000}$ = 0,455 ou 45,5% (Fréquence marginale de Y). Cela veut dire qu'il y a 45,5% de familles ayant 5 enfants et plus quel que soit l'âge du père.

$\mathbf{f_{1/2}}$ **avec i fixé** = $\frac{n_{21}}{n_{2.}} = \frac{50}{90}$ = 0,5555 ou 55,55% (la fréquence conditionnelle de Y avec j = 1 si i = 2. Cela signifie que parmi les familles dont l'âge du père est compris entre 25 et 30 ans, 55,55% ont moins de deux enfants.

$\mathbf{f_{3/2}}$ **avec j fixé** = $\frac{n_{32}}{n_{.2}} = \frac{100}{345}$ = 0,2898 ou 28,98% (la fréquence conditionnelle de X avec i = 3 si j = 2). Cela veut dire que parmi les familles ayant de 2 à 5 enfants, 28,98% des pères ont l'âge compris entre 30 et 40 ans.

Section 2 : Ajustement, régression et corrélation : Après avoir appris à organiser des données à deux caractères et à calculer certains indicateurs, nous abordons ici l'analyse des liens entre deux variables, notamment dans les situations de prévision ou d'estimation. Trois approches principales s'en dégagent :

- **Ajustement** : À partir d'un nuage de points expérimentaux, on cherche à définir une fonction mathématique qui décrit au mieux la relation entre deux variables, en estimant les paramètres de façon à minimiser l'écart entre la courbe et les données observées.
- **Régression** : On suppose que l'une des variables (x) influence l'autre (y). L'objectif est de modéliser cette relation causale potentielle, en considérant x comme variable explicative et y comme variable dépendante.
- **Corrélation** : Lorsque les deux variables sont aléatoires, on mesure l'intensité du lien entre leurs variations, sans postuler de relation causale.

On distingue par ailleurs trois formes de liaison statistique entre deux variables :

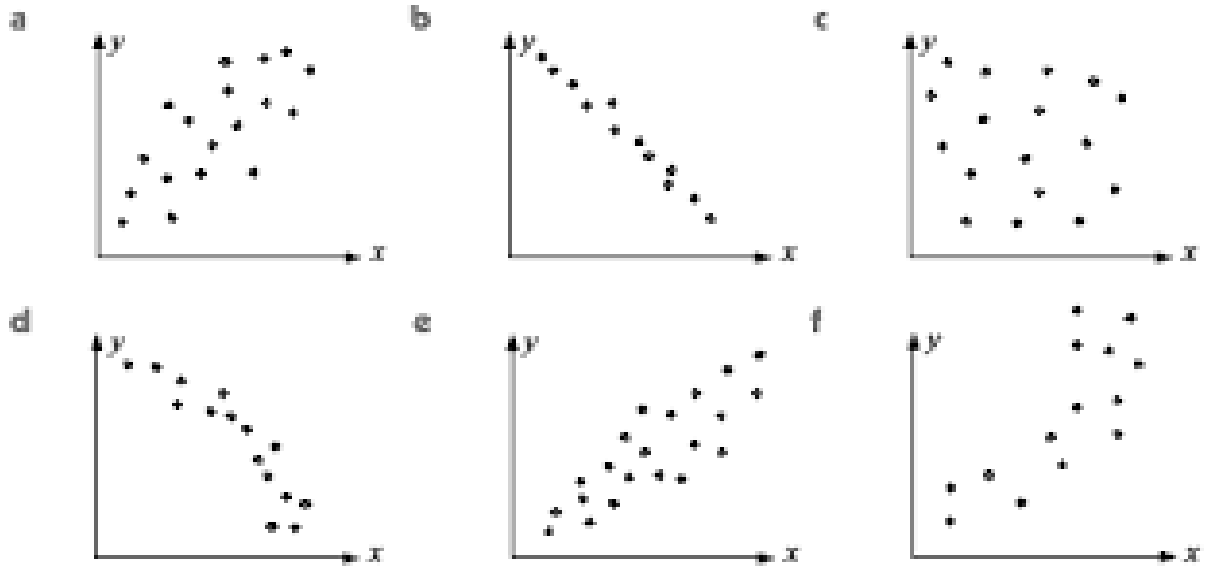
- **Absence de liaison** : les variables sont totalement indépendantes, comme le lien entre la taille d'un individu et son salaire.
- **Liaison fonctionnelle parfaite** : la variation de l'une détermine avec certitude celle de l'autre (ex. : rayon et périmètre d'un cercle).
- **Liaison partielle** : les variables sont liées de manière plus ou moins forte, sans relation systématique (ex. : revenu et consommation).

NB : Dans ce cours, destiné aux étudiants de première année, nous ne ferons pas de distinction rigoureuse entre ajustement et régression, bien que leurs finalités diffèrent. De plus, seules les séries à deux variables non pondérées seront étudiées, à la différence de la section précédente.

Lorsqu'on se retrouve face à une série de données à deux variables, notre premier réflexe n'est pas de plonger dans des calculs, mais bien de regarder ce que les données ont à nous dire. La représentation graphique, sous forme d'un nuage de points, devient alors un outil intuitif : elle offre une première lecture visuelle de la tendance, de la dispersion et, parfois, de la nature même du lien entre les variables.

C'est souvent à partir de cette simple image que l'on commence à formuler des hypothèses : les points suivent-ils une droite ? Une courbe ? Sont-ils désordonnés ?

Dans ce qui suit, plusieurs exemples concrets de nuages de points nous permettront d'explorer les différentes formes que peuvent prendre les relations statistiques.

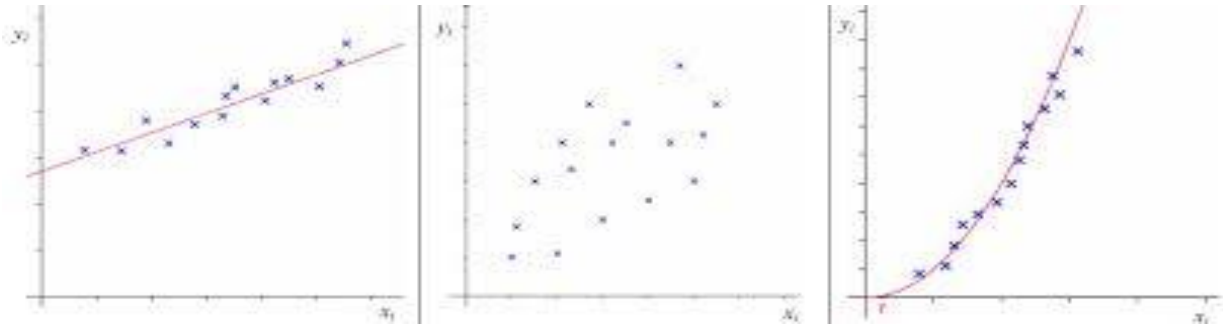


Les graphiques ci-dessus illustrent différentes formes de relations possibles entre deux variables x et y :

- **a)** Nuage de points indiquant une **corrélation positive faible à modérée** : à mesure que x augmente, y a tendance à augmenter également, mais la dispersion reste importante.
- **b) Corrélation négative forte** : les points sont bien alignés selon une pente descendante ; plus x augmente, plus y diminue.
- **c) Absence apparente de relation** : les points sont dispersés sans direction claire, suggérant une indépendance entre les variables.
- **d) Relation non linéaire décroissante** : les points suivent une courbe concave, indiquant une liaison mais pas de type linéaire.
- **e) Corrélation linéaire positive forte** : les points s'alignent quasiment sur une droite montante ; plus x augmente, plus y augmente.
- **f) Corrélation positive avec une dispersion importante** : tendance croissante mais avec une variabilité marquée.

Cette première exploration visuelle nous donne une idée générale de la nature du lien entre les deux variables. Mais pour aller au-delà de l'intuition, il est souvent nécessaire de chercher une courbe ou une droite qui résume au mieux cette tendance. C'est là qu'intervient l'ajustement. Il existe différentes manières d'ajuster une courbe aux données, selon le niveau de précision souhaité et les outils disponibles. La plus simple, et souvent la plus immédiate, est l'ajustement graphique**,** que nous allons examiner en premier.

2.1 Ajustement graphique :



L'ajustement graphique consiste à tracer, à main levée, une courbe qui suit au mieux la disposition des points du nuage.

Lorsque ces points semblent s'aligner globalement selon une droite, on parle alors de liaison linéaire entre les deux variables. Dans ce cas, une droite ajustée peut résumer efficacement leur relation. Cependant, la forme du nuage peut aussi suggérer une relation non linéaire (comme une courbe), ou aucune relation apparente**,** indiquant une indépendance statistique entre x et y. Le choix de la courbe dépend donc de l'aspect général du nuage de points.

2.2. Ajustement analytique :

L'ajustement analytique consiste à rechercher une droite mathématique qui traduit le mieux la relation entre les deux variables observées. Autrement dit, on cherche à déterminer une équation de la forme : $Y=aX+b$ où :

- a représente le **coefficient directeur** (la pente) de la droite,
- b est l'**ordonnée à l'origine** (le point où la droite coupe l'axe des ordonnées).

Dans la pratique, les données ne sont généralement pas parfaitement alignées : certains points sont au-dessus de la droite idéale, d'autres en dessous. C'est pourquoi on utilise une méthode objective pour définir la « meilleure » droite possible : **la méthode des moindres carrés**.

Cette méthode consiste à minimiser la somme des carrés des écarts verticaux entre les points observés et ceux prédits par la droite. La droite ainsi obtenue est appelée droite d'ajustement de Y en fonction de X.

On peut également inverser la relation et chercher à ajuster X en fonction de Y, en obtenant une équation de la forme : $X=a'Y+b'$. Le choix dépend du rôle attribué à chaque variable dans l'analyse.

2.2.1 Droite d'ajustement de Y en fonction de X

Comme évoqué précédemment, la méthode des moindres carrés repose sur un principe simple : réduire au minimum les écarts entre les données observées et les valeurs fournies par la droite

d'ajustement. Mais ces écarts peuvent être positifs ou négatifs, ce qui risquerait d'annuler leur somme. Pour éviter cela, on élève chaque écart au carré, de manière à les rendre tous positifs. C'est de là que vient l'expression **« moindres carrés »**.

Concrètement, cela revient à minimiser la fonction suivante :

$$S = \sum_{i=1}^{n}(y_{i-} ax_i - b)^2$$

Où :

- yi représente la valeur observée de la variable dépendante,
- axi+b est la valeur estimée par la droite d'ajustement,
- S est la somme des carrés des écarts entre les valeurs observées et celles prévues.

L'objectif est donc de trouver les valeurs de a et b qui rendent cette somme S aussi petite que possible.

Nous n'allons pas revenir, en détail, sur la méthode de détermination des coefficients d'ajustement *a* et *b.* Nous notons alors : a = $\frac{cov(xy)}{v(x)}$

Si les données ne sont pas pondérées (séries simples), on calcule la covariance et la variance de la manière suivante :

Cov(xy) = $\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{N}$ (formule de définition) ou

cov(xy) = $\frac{\sum_{i=1}^{n} x_i y_i}{N} - \bar{x}\bar{y}$ (formule développée).

La variance v(x) = $\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{N}$ (formule de définition) ou

V(x)= $\frac{\sum_{i=1}^{n} x_i^2}{N} - \bar{x}^2$ (formule développée). En simplifiant, on peut calculer le « a » comme suit :

$$a = \frac{\sum_{i=1}^{n}(x_{i-} \bar{x})(y_{i-} \bar{y})}{\sum_{i=1}^{n}(x_{i-} \bar{x})^2} \text{ Ou encore } a = \frac{\sum_{i=1}^{n} x_i\, y_i \; - N\bar{x}\,\bar{y}}{\sum_{i=1}^{n}(x_i)^2 - N\bar{x}^2}$$

Comme la droite passe par le point moyen ($\bar{x}$,$\bar{y}$) donc $\bar{y} = a\bar{x} + b \implies b=\bar{y} - a\bar{x}$

2.2.2 Droite d'ajustement de X en fonction de Y : On peut déterminer l'équation de la droite qui ajuste X en fonction de Y lorsque l'on peut estimer la valeur de X à partir de Y. Dans ce cas, X devient la variable dépendante (ou expliquée), tandis que Y est la variable indépendante (ou explicative).

L'équation de cette droite s'écrit généralement sous la forme :

$$X = \acute{a}\, Y + \acute{b}$$

$\grave{a} = \frac{cov(xy)}{v(y)}$, v(y) = $\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{N}$ (formule de définition) ou

v(y) = $\frac{\sum_{i=1}^{n} y_i^2}{N} - \bar{y}^2$ (formule développée). En simplifiant, on calcule le coefficient $\acute{a}$ comme suit :

$$\acute{a} = \frac{\sum_{i=1}^{n}(y_{i-} \bar{y})(x_{i-} \bar{x})}{\sum_{i=1}^{n}(y_{i-} \bar{y})^2} \; ou \; a' = \frac{\sum_{i=1}^{n} x_i\, y_i \; - N\bar{x}\,\bar{y}}{\sum_{i=1}^{n}(y_i)^2 - N\bar{y}^2} \; et \; \acute{b}=\bar{x} - \acute{a}\bar{y}$$

2.3.Corrélation :

Nous avons souligné plus haut que pour mesurer l'intensité de la relation entre deux variables x et y nous utilisons un indicateur appelé coefficient de corrélation. Le coefficient de corrélation linéaire *r* entre les deux variables X et Y se définit soit par la formule

$$r= \sqrt{a.\acute{a}}$$

Ou encore par la formule

$$r= \frac{Cov\ (x,y)}{\sqrt{V(x).V(y)}} ou\ \ r= \frac{Cov\ (x,y)}{\sigma(X).\sigma(Y)} ou\ r=a \times \frac{\sigma(X)}{\sigma(Y)}$$

Autrement dit, $r= \frac{\sum_{i=1}^{n}(x_{i-}\ \bar{x})(y_{i-}\ \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_{i-}\ \bar{x})^2.\sum_{i=1}^{n}(y_{i-}\ \bar{y})^2}}$

Ou $r= \frac{\sum_{i=1}^{n} x_{i.}y_{i-}\ N\bar{x}\bar{y}}{\sqrt{\sum x_i^2 - N\bar{x}^2}.\sqrt{\sum y_i^2 - N\bar{y}^2}}$

Le coefficient de corrélation varie entre -1 et +1.

- Si r =0, il y a absence de corrélation entre x et y.

- Si r = +1 ou -1, il y a une corrélation maximale entre x et y, c'est-à-dire que tous les points sont alignés. On parle alors d'une liaison fonctionnelle.
- Si r est proche de +1 ou de -1, cela indique une très forte corrélation linéaire entre les deux variables.
- Si r est proche de zéro, alors il s'agit d'une faible corrélation linéaire entre les deux variables.

NB : Le signe positif (+) signifie que les deux variables varient dans le même sens.
Le signe négatif (-) signifie que les deux variables varient en sens inverse.

Nous pouvons calculer le coefficient de détermination r^2, il exprime le pourcentage de variation de la variable y expliquée par la variable x.

$$r^2 = a \times \acute{a}$$

Exercice d'application :

Soit la série bi-variée suivante où X représente les résultats au test (noté sur 10) de six (6) employés et Y les rendements (en douzaine d'unités).

X_i	2	3	5	7	9	10
Y_i	1	3	7	11	15	17

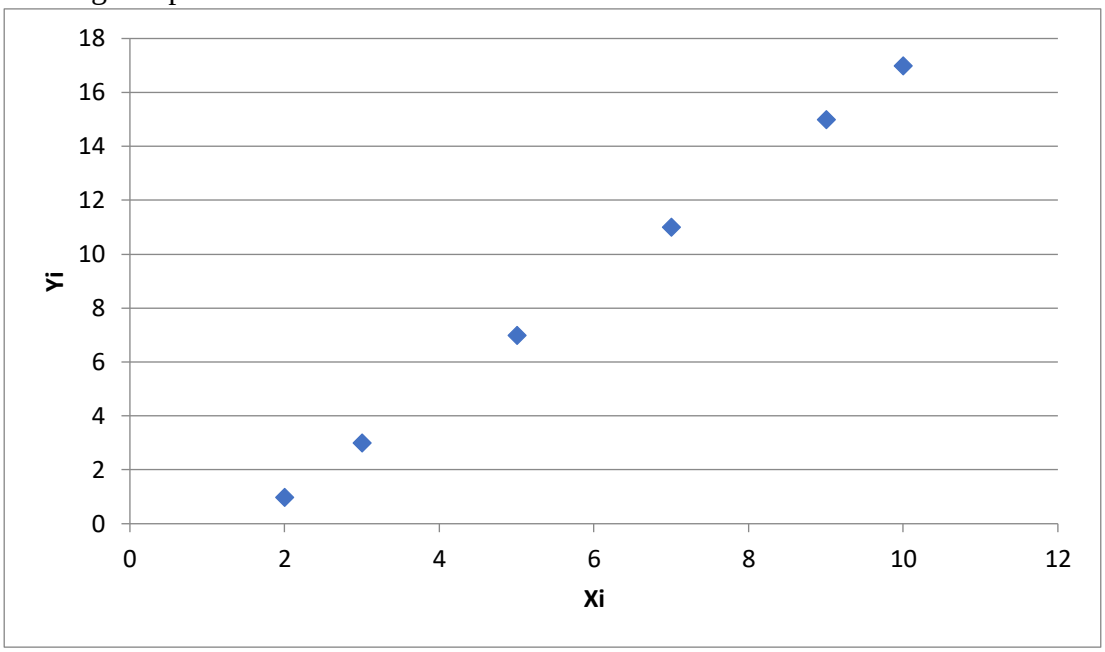
Questions :

1- Représenter le nuage de points.
2- Trouver l'équation de la droite de régression de Yen X par la méthode des moindres carrés.
3- Trouver l'équation de la droite de régression de X en Y.

4- Calculer les coefficients de corrélation et de détermination.
5- Estimer le rendement d'un employé ayant obtenu un résultat de 4 sur 10.

Solution :

1- Le nuage de points :



2- L'équation de la droite de régression de Yen X : Y = aX+b

X_i	Y_i	X_iY_i	X_i^2	Y_i^2
2	1	2	4	1
3	3	9	9	3
5	7	35	25	49
7	11	77	49	121
9	15	135	81	225
10	17	170	100	289
$\sum = 36$	$\sum = 54$	428	268	694

a = $\frac{cov\,(xy)}{v\,(x)}$ avec Cov (xy) = $\frac{\sum X_iY_i}{N} - \bar{X}\,\bar{Y}$ et V(x) = $\frac{\sum X_i^2}{N} - \bar{X}^2$

Calculons d'abord les moyennes marginales : $\bar{X}$ et $\bar{Y}$

$\bar{X} = \frac{\sum X_i}{N} = \frac{36}{6} = 6$ et $\bar{Y} = \frac{\sum Y_i}{N} = \frac{54}{6} = 9$

Cov (xy)= $\frac{428}{6} - (6)\times(9) = 17{,}33$ et V(x) = $\frac{268}{6} - (6)^2 = 8{,}66$

a= $\frac{17{,}33}{8{,}66} = 2$. On trouve le coefficient b comme suit : on a Y= aX+b comme la droite d'ajustement passe par le point moyen ($\bar{X}, \bar{Y}$) $\bar{Y}$= a$\bar{X}$+b $\Rightarrow$ b = $\bar{Y}$ − a$\bar{X}$ $\Rightarrow$

b= 9− (2) ×(6) = −3

L'équation est Y= 2X−3 (on peut la représenter sur le nuage de points)

3- La droite de régression de X en Y : X= $\acute{a}$Y +$\acute{b}$ avec $\acute{a}$ =$\frac{cov\,(xy)}{v(y)}$, calculons la variance de Y : v(y) =$\frac{\sum Y_i^2}{N}$−$\bar{Y}^2$ = $\frac{694}{6}$ –(9)^2 = 34,66

$\acute{a}$= $\frac{17,33}{34,66}$ = 0,5 et $\acute{b}$=$\bar{X}$−$\acute{a}\bar{Y}$ $\Longrightarrow$ $\acute{b}$ = 6− 0,5(9) = 1,5

L'équation est X= 0,5 Y+1,5

4- Coefficients de corrélation (r) et de détermination r^2 :

r= $\frac{cov\,(xy)}{\sigma(x)\times\sigma(y)}$ = $\frac{17,33}{\sqrt{v(X)}\sqrt{v(Y)}}$ =$\frac{17,33}{\sqrt{8,66}\sqrt{34,66}}$ = $\frac{17,33}{2,943\times5,887}$ $\cong$ 1 ou r = $\sqrt{a\acute{a}}$ =$\sqrt{2\times0,5}$ = 1

r est égal à 1, il y a une corrélation maximale entre les résultats du test et le rendement des employés.

r^2 = (1)^2 =1 ou 100%. Cela signifie que le rendement des employés est expliqué totalement (à 100%) par les résultats du test.

5- Si X=4 ; Y= ?, nous avons Y=2X−3 donc Y= 2(4) −3 = 5.

Conclusion :

Ce chapitre a permis d'élargir notre analyse statistique en étudiant les distributions à deux caractères et les relations qui peuvent exister entre eux. En explorant les notions de régression, d'ajustement et de corrélation, nous avons acquis des outils essentiels pour mesurer, quantifier et interpréter les liens entre deux variables quantitatives. Ces méthodes offrent une meilleure compréhension des interactions possibles au sein d'une population, facilitant ainsi la prise de décision et l'analyse prédictive.

Conclusion générale

Ce cours de statistique descriptive a permis de poser les fondements indispensables à toute démarche d'analyse quantitative rigoureuse. En progressant à travers les différentes étapes — de la définition des concepts de base à l'étude des relations entre deux variables — l'étudiant a acquis les outils nécessaires pour organiser, synthétiser et interpréter un ensemble de données de manière claire et structurée.

Les notions de population, de caractère, de modalité, ainsi que les différentes formes de représentations graphiques et tabulaires ont constitué une première approche concrète de l'observation statistique. Ensuite, l'étude des mesures de tendance centrale, de dispersion, de concentration et de forme a permis de caractériser plus finement les distributions. L'introduction aux indices statistiques a montré comment suivre l'évolution d'un phénomène dans le temps, notamment dans un contexte économique. Enfin, l'analyse bivariée, à travers la régression et la corrélation, a ouvert la voie à une compréhension plus approfondie des interactions entre variables.

Au-delà des méthodes et des calculs, ce parcours vise à développer chez l'étudiant une culture statistique critique : savoir interroger les données, en évaluer la signification, et en tirer des enseignements pertinents selon le contexte étudié.

La suite logique de cette formation sera abordée au second semestre, avec l'étude des probabilités, qui introduira les fondements mathématiques de l'incertitude, et ouvrira la voie vers la statistique inférentielle.

Références bibliographiques :

1- BAILLY P. : « L'économie et les chiffres, exercices corrigés de statistique descriptive », Ed : OPU, 1993.
2- BOUDIA M.C. : « Statistique descriptive », Ed : CASBAH, Alger, 2008.
3- BOUKELLA-BOUZOUANE M. : « Statistique descriptive, rappels de cours avec exercices corrigés », Ed. Casbah, Alger, 2001.
4- BOURSIN J.L. : « Comprendre la statistique descriptive », Ed : ELLIPSES, 1995.
5- CALOT Gérard : « Cours de statistique descriptive », Ed : DUNOD, 1993.
6- DUTHIL G. : « Initiation à la statistique descriptive », Ed : ELLIPSES, 1995.
7- GRAIS B. : « Statistique descriptive avec rappels de cours », Ed : DUNOD, Paris, 1998.
8- GRAIS B. : « Statistique descriptive », Ed : DUNOD, 3ème édition, Paris, 2000.
9- HAMDANI H. : « Statistique descriptive avec initiation aux méthodes d'analyse de l'information économique », Ed : OPU, 5ème édition, 2006.
10- LABROUSSE C. : « Statistique, exercices corrigés avec rappels de cours », Ed : BORDAS, Paris, 1977.
11- LE CORNU F. : « Statistique descriptive : Exercices corrigés », Ed : DUNOD, 2000.
12- LECOUTRE Jean-Pierre : « Statistique descriptive : Exercices corrigés avec rappels de cours », Ed : MASSON, 1990.
13- MONINO J.L. : « Statistique descriptive avec rappels de cours, questions de réflexion », Ed. DUNOD, 2000.

Table des matières