SOMMAIRE

AVANT-PROPOS

Ce présent document constitue un support de cours préliminaire de la statistique descriptive. Il s'adresse plus particulièrement aux étudiants en sciences économiques, sciences de gestions et sciences commerciales première année de licence. Il peut également être utile à toute personne conduite à utiliser des données statistiques, qu'il s'agisse de faire un rapport ou de rédiger un mémoire. Il a été conçu pour être accessible au plus grand nombre. On veut développer le sens critique nécessaire lors de la mise en œuvre et de l'interprétation d'un traitement statistique. Pour cela, on introduira et utilisera un cadre mathématique rigoureux. Nous fournirons autant d'exemples et de figures nécessaires afin d'obtenir une meilleure compréhension du cours.

Le cours a pour but d'initier les étudiants aux principes de base de la statistique 1 (statistique descriptive) et vise principalement à introduire et faire méditer les concepts fondamentaux et méthodes élémentaires de la statistique pour permettre un apprentissage autonome ultérieur de méthodes complémentaires.

Objectifs du cours :

La statistique descriptive a pour but d'étudier un phénomène à partir de données. Cette description se fait à travers la présentation des données (la plus synthétique possible), leur représentation graphique et le calcul de résumés numériques.

- Comprendre la différence entre la statistique descriptive et la statistique inférentielle
- Savoir quelle est l'utilité de la statistique descriptive lors de toute analyse préliminaire des données
- Explorer les données graphiquement pour les caractériser et identifier des problèmes
- Utiliser les statistiques descriptives pour décrire les données
- Effectuer le test statistique approprié à l'objectif de leur étude
- Traiter statistiquement les données plus rapidement et plus facilement
- Interpréter les résultats avec plus de confiance et de fiabilité.

Les cours, exemples et exercices d'applications sont organisés en cinq grands points :

- notions fondamentales de la statistique descriptive ;
- présentation des données statistiques (Tableau statistique, Représentation graphique) ;

- indicateurs statistiques concernant l'étude d'une variable (position, dispersion, indices, concentration) ;
- étude de distributions statistiques à deux variables (régression, corrélation) ;
- étude de séries chronologiques.

D'autres exercices de révisions sans laissés sans solutions, qui vont d'être traiter sur d'autres manuels.

Objectifs généraux de l'enseignement de la Statistique :

En licence, master et en maîtrise de Sciences de l'Education:

Cet enseignement vise à apporter à chaque étudiant quelques outils techniques et conceptuels efficaces et bien identifiés qui devraient l'aider à :

- expliciter les questions d'une problématique dont l'émergence et/ou la validation des réponses relèvent d'une approche statistique en liaison avec le modèle dans lequel cette problématique est posée,

- décrire, traiter, analyser des données de manière pertinente dans le cadre d'une étude en particulier dans le domaine éducatif,

- faire le lien entre la réflexion analytique sur des questions relevant du champ de l'éducation, leur formalisation et leur traitement quantitatif,

- lire avec un regard critique et distancié, les conclusions de diverses études statistiques,

- poursuivre de façon autonome et personnalisée un apprentissage en statistique afin d'enrichir les acquis personnels actuels,

- poser un regard plus positif à l'égard d'un domaine largement exploité dans les media, dans le sens de ne pas considérer les résultats dans l'ordre du tout ou rien mais en les replaçant judicieusement dans leur domaine de validité et en relativisant la portée,

- exploiter des notions et des démarches mathématiques à des fins d'outils, et de ce fait à modifier dans un sens positif le rapport souvent négatif que nombre entretient avec cette science,

- s’exercer à un raisonnement intégrant l’idée de "risque d’erreur" dans l’énoncé des conclusions.

- s’exercer à l’interprétation de phénomènes éducatifs sur la base de données statistique recueillie sur des "faits",

- s’exercer à la communication des résultats des analyses des données en faisant une distinction bien nette entre le modèle utilisé et la réalité qu’il est supposé représenter, entre les traitements conduits au sein du modèle et les interprétations reformulées dans le contexte au sein duquel est posé le problème.

Donc l'objectif de la statistique descriptive est de décrire, c'est-à-dire de résumer ou représenter, par des statistiques, les données disponibles quand elles sont nombreuses, pour les rendre plus lisibles. (C’est-à-dire parmi ces objectives c’est d'extraire des informations pertinentes d'une liste de nombres complexe à interpréter par une simple lecture)

En 1922, un des plus grands statisticiens de la première moitié du XXème siècle, **Ronald A. Fisher**, écrivit "L’objet de la méthode statistique est la réduction des données. Une masse de données doit être remplacée par un petit nombre de quantités représentant correctement cette masse, et contenant autant que possible la totalité de l’information pertinente contenue dans les données d’origine. Ces méthodes elles consistent à essayée de résumer un échantillon de données par des graphiques ou des caractéristiques numériques. Donc la statistique descriptive dispose deux moyens **l’approche graphique** et **l’approche numérique**.

Considérons par **exemple** les notes globales à un examen. Il peut être intéressant d'en tirer une valeur centrale qui donne une idée synthétique sur le niveau des étudiants. Celle-ci peut être complétée par une valeur de dispersion qui mesure, d'une certaine manière, l'homogénéité du groupe. Si on veut une information plus précise sur ce dernier point, on pourra construire un histogramme ou, d'un point de vue un peu différent, considérer les déciles. Ces notions peuvent être intéressantes pour faire des comparaisons avec les examens analogues passés les années précédentes ou en d'autres lieux. Ce sont les problèmes les plus élémentaires de l'analyse des données qui concernent une population finie. Les problèmes portant sur des statistiques multidimensionnelles nécessitent l'utilisation de l'algèbre linéaire. Indépendamment du caractère, élémentaire ou non, du

problème il s'agit de réductions statistiques de données connues dans lesquelles l'introduction des probabilités perfectionnerait difficilement l'information obtenue. Il est raisonnable de regrouper ces différentes notions.

Introduction générale:

« *Tout est nombre* » affirmait **Pythagore**, il ya quelques 25 siècles. Depuis quelques décennies, beaucoup d'informations, jusqu'alors latentes, ont pris corps, et l'entreprise, plus que toute autre institution, s'est trouvée à la tête d'une masse d'informations disponibles qu'il a semblé, le plus souvent à juste titre, intéressant d'exploiter .

L'amélioration de la gestion des entreprises ou bien n'importe qu'elle établissement passe par la substitution d'une régulation a priori, a une régulation a posteriori. Cette régulation a priori repose sur la connaissance précise des phénomènes que l'on cherche à maitriser. C'est à ce niveau qu'intervient le statisticien dont l'art consiste à utiliser ''au mieux'' l'information disponible pour tirer cette connaissance des faits. Pour le dirigeant la statistique est une technique mise à sa disposition pour l'aider à améliorer ses décisions (pour prendre des décisions).

L'utilisation de la statistique comme outil d'aide à la décision se propose de montrer concrètement comment construire et utiliser des processus statistiques pour une meilleure efficacité et pour améliorer les produits, les services et les systèmes de toute nature.

Ainsi avec Savoir utiliser la statistique. Outil d'aide à la décision et à l'amélioration de la qualité, étudiants, techniciens, cadres et responsables de tout secteur d'activité trouveront des outils pratiques pour savoir comment résoudre les nombreux problèmes auxquels ils sont confrontés.

Chapitre I : Généralités sur la statistique

I.1 Historique et définitions de la statistique :

Pour montrer ce rôle de la statistique dans différents domaine (comme par exemple la prise de décision etc.) commençant d'abord par les déférentes définitions et son contenu.

Historiquement, le mot statistique (du **latin status** « état ») n'aurait guère été utilisé avant le XVIIIème siècle. Il semble qu'il désignait alors l'ensemble des renseignements, en particulier quantitatifs, nécessaires à l'autorité politique pour diriger l'état. On peut cependant considérer que la statistique, au moins sous forme de dénombrements, a été pratiquée par toutes les civilisations. Les grands empires de l'antiquité nous en apportent les premiers témoignages ; les monarchies européennes centralisées du XVIIème, dans le cadre de la mise en place de l'état moderne, ont toutes eu ce même souci du dénombrement. Quant aux sociétés contemporaines, elles ont institutionnalisé des pratiques qui semblent aujourd'hui bien ordinaires, et nombre de secteurs de l'activité humaine font appel, sous une forme ou sous une autre, à la statistique.

Les informations qui sont apparues très tôt indispensable à l'état étaient celles qui permettaient de recueillir des impôts et de recruter des conscrits pour les besoins de guerres. Ce n'est donc pas étonnant si les premiers chiffres recueillis concernaient les populations et l'économie.

Apres l'économie et la démographie, la statistique s'est étendue à l'ensemble des sciences et elle devenue une discipline scientifique faisant appel aux mathématiques (calcul algébriques, analyse, calcul des probabilités…) et à l'informatique pour les applications pratiques.

I.1.1Définitions de la statistique (C'est quoi la statistique ?) Il existe tellement de définitions différentes de la statistique qu'on pourrait presque en faire une étude

La statistique est à la fois une science formelle, une méthode et une technique. Elle comprend la collecte, l'analyse, l'interprétation de données mais aussi la présentation de ces données pour les rendre lisibles.

Ce domaine des mathématiques ne doit pas être confondu avec une statistique qui est un nombre calculé à partir d'observations.

Nous avons l'habitude de rencontrer, dans des domaines très divers, ce que l'on appelle des statistiques ; dans le langage courant le mot « **statistiques** » **au pluriel** désigne des

collections des chiffres présentées sous forme de tableaux ou de graphiques ; ces chiffres représentent des observations portant sur des faits nombreux.

Le mot **« statistique » au singulier** désigne l'ensemble des méthodes ayant pour l'étude numérique des faits nombreux (on peut parler de la statistique en tant que science)

Les statistiques sont le produit des analyses reposant sur l'usage de la statistique. Cette activité regroupe trois principales branches :

- la collecte des données ;
- le traitement des données collectées, aussi nommé la statistique descriptive ;
- l'interprétation des données, aussi nommée l'inférence statistique, qui s'appuie sur la théorie des sondages et la statistique mathématique.

Dans le but :

- d'informer
- d'aider à la prise de décision

Cette distinction ne consiste pas à définir plusieurs domaines étanches. En effet, le traitement et l'interprétation des données ne peuvent se faire que quand celles-ci ont été récoltées. Réciproquement, la statistique mathématique précise les règles et les méthodes sur la collecte des données, pour que celles-ci puissent être correctement interprétées.

Donc la statistique en tant que méthode d'analyse comporte deux niveaux :
I.1.1.1. La statistique descriptive qui regroupe les nombreuses techniques et méthodes utilisées pour décrire avec des outils appropriés un ensemble relativement important de données et dégagé l'essentiel de l'information ; elle utilise des modes de représentations graphiques et également des caractéristiques (indicateur de valeur centrale : moyenne arithmétique, médiane et le mode) et indicateur de dispersion autour de la valeur centrale ;
A cet égard on peut présenter la statistique descriptive **comme suit :**

- La notion de variable
- Types de variables : valeurs prises par les variables: discrètes ou continues
- **Résumer une distribution :** la statistique dispose deux moyens l'approche graphique et la méthode numérique.
- **Présentation d'une série statistique**

- ✓ Tableau / Graphique
- ✓ Fréquences et effectifs / fréquences et effectifs cumulés croissants
- ✓ Regroupement de données

Il est souvent utile de résumer une distribution par un ou deux nombres, par exemple, pour comparer des distributions différentes.

Ces deux nombres sont

- **la valeur centrale**
- **La dispersion**

Il existe plusieurs indicateurs possibles pour la valeur centrale ou la dispersion

1 : **Mesures de tendance centrale** : mode, médiane, quartiles, moyenne, etc.

2 : **Mesure de dispersion** : variance, écart-type, coefficient de variation, étendue, étendue inter-quartile, etc. et mesures de forme (analyse de la symétrie d'une distribution et analyse de l'aplatissement d'une distribution).

Mise en relation de deux variables :

Continues: **Le coefficient de corrélation** et les graphiques en nuage de points (**la fonction de régression** pour faire **des estimations**).

Discrètes**: Tableaux de contingence**.

Pour les prévisions on se base sur l'étude des **séries chronologiques**

I.1.1.2. Statistique théorique ou mathématique qui prend la suite la statistique descriptive lorsque l'on peut élaborer des lois, (dont l'objet est de formuler des lois à partir de l'observation d'échantillons, c'est-à-dire de tirages limités effectués au sein d'une population. La statistique mathématique intervient dans les enquêtes et les sondages). Elle s'appuie sur la statistique descriptive, mais aussi sur le calcul des probabilités.

Pour notre part nous avons cherché à expliciter notre conception de la statistique on prenant appui sur **la définition suivante** : La statistique est la science qui procède à l'étude méthodique à partir de modélisations mathématiques, des modes d'utilisation et de traitement de données, c'est à dire de l'information, dans le but de conduire et d'étayer une réflexion ou de prendre une décision en situation concrète soumise aux aléas de l'incertain.

La statistique descriptive étudie ces modes d'utilisation et de traitement de données, à un premier niveau, dans la perspective de produire essentiellement des descriptions des informations.

La statistique inférentielle les étudie à un second niveau dans la perspective d'étendre ces informations décrites à un domaine de validité non exploré directement, avec, si possible, un contrôle des risques encourus dans ce raisonnement inductif.

Alors cette technique trouve son application dans des domaines très déférents

I.1.2. Domaines d'application :

En 1963, **W. Weaver** écrit " la statistique et La théorie des probabilités sont deux domaines importants, intégrés à nos activités quotidiennes. Le monde de l'industrie, les compagnies d'assurance sont largement tributaires des lois probabilistes. La physique elle-même est de nature essentiellement probabiliste.

Les statistiques sont utilisées dans des domaines particulièrement variés comme :

- en géophysique, pour les prévisions météorologiques, la climatologie, la pollution, les études des rivières et des océans ;
- en démographie : le recensement sert à faire une photographie à un instant donné d'une population et permettra ensuite des sondages dans des échantillons représentatifs ;
- en sciences économiques et sociales, et en économétrie : l'étude du comportement d'un groupe de population ou d'un secteur économique s'appuie sur des statistiques. Les questions environnementales s'appuient aussi sur des données statistiques ;
- en sociologie : les sources statistiques forment des matériaux d'enquête, et les méthodes statistiques sont utilisées comme techniques de traitement des données ;
- en marketing : le sondage d'opinion devient un outil pour la décision ou l'investissement ;
- en physique : l'étude de la mécanique statistique et de la thermodynamique statistique (Physique statistique) sert à déduire du comportement de particules individuelles un comportement global (passage du microscopique au macroscopique) ;
- en métrologie, pour tout ce qui concerne les dispositifs de mesure et les mesures elles-mêmes ;
- en médecine et en psychologie, tant pour le comportement des maladies que leur fréquence ou la validité d'un traitement ou d'un dépistage ;
- en assurance et en finance (calcul des risques, ...)
- etc.…

I.1.3.La démarche de la statistique

I.1.3.1.Recueil des données

L'enquête statistique est toujours précédée d'une phase où sont déterminés les différents caractères à étudier.

L'étape suivante consiste à choisir la population à étudier. Il se pose alors le problème de l'échantillonnage : choix de la population à sonder, la taille de la population et sa représentativité.

Le pré traitement des données est extrêmement important, en effet, une transformation des données initiales (un passage au logarithme, par exemple), peuvent énormément favoriser les traitements statistiques suivants.

I.1.3.2.Traitement des données

Le résultat de l'enquête statistique est une série de chiffres (tailles, salaires) ou de données qualitatives (langues parlées, marques préférées). Pour pouvoir les exploiter, il va être indispensable d'en faire un classement et un résumé visuel ou numérique. Il sera quelquefois indispensable d'opérer une compression de données. C'est le travail de la statistique descriptive. Il sera différent selon que l'étude porte sur une seule variable ou sur plusieurs variables.

- Étude d'une seule variable

Le regroupement des données, le calcul des effectifs, la construction de graphiques permet un premier résumé visuel du caractère statistique étudié. Dans le cas d'un caractère quantitatif continu, l'histogramme en est la représentation graphique la plus courante.

Les valeurs numériques d'un caractère statistique se répartissent dans $\mathbb{R}$, il est indispensable de définir leurs positions. En statistiques, on est généralement en présence de la plupart de valeurs. Or, si l'intégralité de ces valeurs forme l'information, il n'est pas aisé de manipuler plusieurs centaines ou alors milliers de chiffres, ni d'en tirer des conclusions. Il faut par conséquent calculer quelques valeurs qui vont permettre d'analyser les données : c'est le rôle des réductions statistiques. Celles-ci peuvent être extrêmement concises, réduites à un nombre : c'est le cas des valeurs centrales et des valeurs de dispersion. Certaines d'entre

elles (comme la variance) sont élaborées pour permettre une exploitation plus théorique des données

- Étude de plusieurs variables

Les moyens informatiques permettent actuellement d'étudier plusieurs variables simultanément. Le cas de deux variables va donner lieu à la création d'un nuage de points, d'une étude de corrélation (mathématiques) éventuelle entre les deux phénomènes ou étude d'une régression linéaire.

Mais on peut rencontrer des études sur plus de deux variables : c'est l'analyse multidimensionnelle dans laquelle on va trouver l'analyse en composantes principales, l'analyse en composantes indépendantes, la régression linéaire multiple s'appuie sur la statistique pour découvrir des relations entre les variables de très vastes bases de données. Et quant on veut faire des estimations et des prévisions, par le bais des séries chronologiques.

En résumé on peut dire que le but essentiel de la statistique est de présenter des faits, de les analyser pour en tirer des conclusions permettant :

- Des estimations
- Des comparaisons
- Des prévisions.

Dans un sens général, la statistique est l'ensemble des méthodes scientifiques à partir des quelles sont recueillies, présentées, résumées et analysées des données, mais ce n'est pas l'application systématiques de formules mathématiques, il n'est pas nécessaire d'avoir un niveau très élevé en mathématique pour comprendre les statistiques, il suffit d'avoir le bon sens et de la logique, et surtout de savoir s'en servir et interpréter.

A cet égard on peut dire que la statistique repose sur les trois principes suivants :

1- **l'apprentissage de l'outil statistique** : c'est la phase d'acquisition et d'assimilation des concepts statistiques (les concepts de base).
2- **L'application de l'outil statistique** : c'est la phase de vérification des capacités d'application de manière intelligente (collecte d'information, construction de tableaux, traitement graphique, calculs des paramètres etc…)
3- **L'interprétation des résultats de calculs** : c'est la phase de réflexion à partir de laquelle on peut mesurer le degré de maitrise de l'outil statistique, car il s'agit d'expliquer, de

commenter et d'analyser les phénomènes mesurés pour en tirer des conclusions et en fin une décision pertinente

I.2. Notions de base de la statistique :

Nous allons commencer par définir les termes utilisés en statistiques pour désigner les observations chiffrées.

I.2.1. Populations et unités statistiques

En statistique, on travaille sur des populations. Ce terme vient du fait que la démographie, étude des populations humaines, a occupé une place centrale aux débuts de la statistique, notamment au travers des recensements de population. Mais, en statistique, le terme de population s'applique à tout objet statistique étudié, qu'il s'agisse d'étudiants (d'une université ou d'un pays), de ménages (pour employer un terme utilisé en comptabilité nationale), du parc de micro-ordinateurs dans une entreprise ou de n'importe quel autre ensemble sur lequel on fait des observations statistiques.

Une population est composée d'individus. Les individus qui composent une population statistique sont appelés unités statistiques. Donc Les éléments de la population sont appelés les individus ou unités statistiques.

Par exemple si l'on veut faire des observations chiffrées sur l'ensemble des étudiants composant un amphithéâtre, la population statistique étudiée est cet ensemble d'étudiants, chaque étudiant étant une unité statistique.

La statistique étudie les caractéristiques des individus. C'est donc sur eux que portent les observations. Mais elle ne s'intéresse pas aux individus en tant que tels ; elle s'y intéresse seulement dans la mesure où ils contribuent à une meilleure connaissance de la population, puisque la statistique « descriptive », comme son nom l'indique cherche à décrire une population donnée.

I.2.2. Caractères et variables

I.2.2. 1. Définition

Pour étudier une population, le statisticien ne retient que les caractères qui l'intéressent, un caractère étant une variable qui caractérise les individus de cette population.

C'est une propriété possédée par les unités statistiques permettant de les décrire et de les distinguer les unes des autres. Toute unité statistique peut être étudiée selon un ou plusieurs caractères

Ainsi, si l'on s'intéresse à la population des étudiants d'un amphithéâtre, on peut le faire d'un point de vue démographique (c'est le cas par exemple, si l'on s'intéresse à l'âge des étudiants), d'un point de vue économique, quand par exemple on s'intéresse aux revenus des étudiants, d'un point de vue sociologique (en s'intéressant aux loisirs des étudiants), d'un point de vue anthropométrique (en s'intéressant à la taille) ou de tout autre point de vue.

Dans chaque exemple cité, c'est un caractère différent qui est étudié :

Âge, revenus, loisirs, taille.

Dans une population donnée, un caractère peut varier d'un individu à l'autre. On dit que ce caractère présent différent modalités.

Exemple : Si l'on étudie la population des étudiants d'un amphithéâtre et que le caractère étudié est l'âge, les modalités du caractère seront 18 ans, 19 ans, 20 ans, etc.

Si l'on étudie une population de voitures et que le caractère étudié est la couleur, les modalités du caractère seront des couleurs : bleu, vert, blanc, etc.

On emploie également le terme de variable statistique pour désigner un caractère, les modalités du caractère étant les valeurs prises par cette variable.

I.2.2. 2. Modalité :

On désigne par modalités les différentes catégories d'un caractère (Ce sont les différentes situations possibles du caractère) et on qualifie de valeurs les différents chiffres d'une variable.

Ce sont les diverses situations (cas, état, valeur) susceptibles d'être prises par le caractère. Un caractère peut posséder une ou plusieurs modalités.

I.2.2. 3. Caractères qualitatifs et quantitatifs

Il existe deux grandes catégories de caractères : les caractères qualitatifs et les caractères quantitatifs.

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

Afin de différencier les deux type de critères, les critères qualitatifs sont appelés des caractères et les critères quantitatifs des variables.

Les caractères qualitatifs sont ceux dont les modalités ne peuvent pas être ordonnées c'est-à-dire que si l'on considère deux caractères pris au hasard, on ne peut pas dire de l'un des caractères qu'il est inférieur ou égal à l'autre. Ainsi, la catégorie socioprofessionnelle des individus d'une population donnée (artisan, ouvrier, etc.) est un caractère qualitatif, la situation matrimoniale (célibataire, veuf, etc.) aussi. On appelle également caractères nominaux, les caractères qualitatifs. A cet effet on peut classer les modalités d'un caractère qualitatif comme suit : **Modalités ordinales, modalités nominales**

Les modalités d'un caractère qualitatif, si elles ne peuvent pas être mesurées quantitativement, sont parfois susceptibles d'être classées dans un certain ordre les unes par rapport aux autres, les modalités sont alors hiérarchisées. Ce sont des **modalités ordinales**.

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

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

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

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

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

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

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

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

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

Remarque : Certaines modalités purement nominales sont parfois codées avec des chiffres. Par exemple, le sexe des individus d'une population sera codé par "1" pour les hommes et par "2" pour les femmes. Il s'agit bien là d'une tentative de quantification d'une variable

purement nominale. On parle alors de variables pseudo-numériques. On peut en effet de cette façon calculer une moyenne, qui sera en fait la proportion des hommes dans la population ou dans l'échantillon.

I.2.2. 4. Caractères discrets et caractères continus

Selon la forme des valeurs de la variable, on distingue deux types de caractères quantitatifs, ceux qui sont discrets et ceux qui sont continus.

I.2.2.4.1. Les caractères discrets (ou discontinues) ne prennent que des valeurs isolées, sont ceux dont le nombre de modalités est fini ou dénombrable. Leurs valeurs peuvent être ou non des nombres entiers. Par exemple Le nombre de pages d'un livre, le nombre de personnes dans une famille, le nombre d'enfants par ménage ne peut être que 0, ou 1, ou 2, ou 3, ... ; il ne peut jamais prendre une valeur strictement comprise entre 0 et 1, ou 1 et 2, ou 2 et 3,sont des caractères discrets.

I.2.2.4.2. Les caractères continus sont ceux qui ont une infinité de modalités peuvent prendre toute valeur dans un intervalle (exemple : la taille, le poids, le revenu des ménages ou le chiffre d'affaire par PME peut être 2900000,1DA, 2900000,12DA, ...,).

I.3. Présentation des données statistiques :

I.3.1. Effectifs et fréquences

L'effectif total est le nombre d'individus appartenant à la population statistique étudiée. L'effectif total sera noté N.

Exemple : Considérons un groupe comprenant trente étudiants et observons l'âge des étudiants dans cette population.

L'effectif total de la population statistique étudiée est trente (N= 30).

I.3.1.1. Définition et notation :

L'effectif d'une modalité xi d'un caractère x est le nombre d'individus présentant cette modalité. L'effectif correspondant à la ième modalité du caractère x est noté ni.

Exemple : Considérons de nouveau le groupe de trente étudiants et construisons un tableau pour regrouper les différentes informations que l'on a sur leur âge.

La première information que l'on va noter dans ce tableau est l'effectif de chaque âge observé.

Âge de 30 étudiants d'un groupe de TD

Âge	Effectif n_i
18	2
19	4
20	10
21	11
22	3
Total	30

De façon générale, pour une variable qui a k modalités, l'effectif total N est égal à la somme des effectifs de chaque modalité du caractère, ce que l'on peut écrire : n1 ; n2 ; ... ;nk N pour une variable qui a k modalités.

I.3.1.2.OPÉRATEURS SOMME

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

Pour simplifier l'écriture, on note cette somme n1 ; n2 ; ... ; nk par $\sum_{i=1}^{k}$ ni Cette notation se lit somme des ni pour i variant de 1 à k

De façon générale la notation $\sum_{i=1}^{n}$ ai ai se lit somme des ai pour i variant de 1 à n et signifie que l'on ajoute les ai en faisant varier i de 1 en 1 en partant de la borne inférieure i=1 et en allant jusqu'à la borne supérieure i=n, les bornes inférieure et supérieure étant respectivement mentionnées en dessous et au-dessus du signe ∑ qui se lit « somme » et correspond à la lettre grecque sigma.

$\sum_{i=1}^{n} ai$= a_1, a_2,…, a_n

Remarque : Pour exprimer une somme d'éléments de façon compacte, Nous avons utilisé la lettre grecque majuscule "Sigma" « ∑ »; Il s'agit tout simplement d'une notation permettant de raccourcir certaines écritures.

Ainsi, lorsqu'on fait la somme des valeurs indicées ni (les effectifs de la série), au lieu d'écrire N = n1 +n2 +…+ ni +… + nk, il est plus commode d'écrire N =$\sum_{i=1}^{k}$ ni .

En ce qui concerne les effectifs, on a donc :

$\sum_{i=1}^{n} ni$ = N pour une variable présentant k modalités.

I.3.1.3.L'opérateur produit

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

Exemple :

Soit quatre valeurs d'une variable x, indicées par i : x1, x2, x3, x4. Le produit de ces 4 valeurs est donné par l'expression :

$$\prod_1^4 xi = x_1 \text{ x } x_2 \text{ x } x_3 \text{ x } x_4$$

L'expression de gauche se lit ainsi "produit des xi pour i allant de 1 à 4".

I.3.2. Présentation des données statistiques

Les informations statistiques recueillies à l'état brut sont incompréhensibles et pratiquement inexploitables. Pour leurs donner un sens et une utilité, il faut les ordonner, les classer et les analyser en utilisant comme support des tableaux et des graphes.

I.3.2.1.Tableaux de distribution des effectifs et les différents types de fréquences : Les tableaux statistiques sont un moyen de classification et des présentations des unités d'une population statistique.

Si l'on considère une population statistique N décrite selon un caractère constitué de k modalités, la présentation de ces informations dans un tableau consiste à dénombrer (compter) pour chacune des modalités l'effectif (le nombre d'unité statistique) qui lui correspond.

On peut construire un tableau d'effectifs quelque soit la nature du caractère (qualificatif ou quantitatif).

Les observations ordonnées et classées dans un tableau forment une distribution

Un tableau de distribution des effectifs est formé de deux colonnes.

Modalités / valeurs du caractère		Effectifs Ni
m1	$\varkappa$ 1	N1
m2	$\varkappa$2	N2
m3	$\varkappa$3	N3
.	.	.
.	.	.
.	.	.
.	.	.
.	.	.
Mk	Kk	Nk
Total		N = $\sum_{i=1}^{\square} ni$

- L'effectif = nombre d'observation pour chaque modalités est appelée aussi fréquence absolue, il est noté ni
- Les modalités du caractère qualificatif (mots) sont notées : mi
- Les valeurs du caractère quantitatif (valeurs) sont notées : $\varkappa$i

N = $n_1 + n_2 + n_3 + \ldots . n_k = \sum_{i=1}^{\square} ni$ = Population

Exemple 1 : La série statistique suivante représente la mention obtenue par 20 étudiants dans un examen.

D-B-E-C-D-B-D-C-E-A-B-E-C-D-B-D-D-A-E-C

(A: Excellent; B: Très bien; C: Bien; D: Assez bien; E: Moyen)

Travail à faire : construise au tableau de distribution des effectifs

- Population statistique : constituée de 20 étudiants → N = 20
- Nature du caractère : qualitatif ordinal car le caractère mention peut être ordonné
- Nombre de modalités : 5 (A, B, C, D, E) → C_1 …. C_5
- Effectif ni : nombre d'étudiants ayant obtenues l'une des modalités (n_1 …. N_5)

Modalités Ci	Effectifs ni
C_1 : A	$N_1 = 2$
C_2 : B	$N_2 = 4$
C_3 : C	$N_3 = 4$
C_4 : D	$N_4 = 6$
C_5 : E	$N_5 = 4$
Total	$N = \sum_{i=1}^{\square} ni = 20$

Exemple2 :

La série statistique suivante représente le nombre d'élèves par classe dans une école primaire

30-30-25-32-26-30-26-26-28-30-30-27-26-27-25

- la population statistique est constituée des 15 classes de l'école ⟹ N = 15

- Le caractère étudiée est le nombre d'élèves c'est un caractère quantitatif discret ou discontinu

- Les valeurs de ce caractère sont au nombre de 6 → 6 valeurs → $\varkappa_1$ …..$\varkappa_6$

- Les effectifs ni : nombre de classes correspondants à l'une des valeurs (n_1 …. N_6)

Valeurs de $\varkappa_i$	Effectifs n_i
$\varkappa_1$=25	n_1=2
$\varkappa_2$=26	n_2=4
$\varkappa_3$=27	n_3=2
$\varkappa_4$=28	n_4=1
$\varkappa_5$=30	n_5=5
$\varkappa_6$=32	n_6=1
Total	$N = \sum_{i=1}^{\square} ni = 15$

Remarque :

Les **variables continue** sont représentées par un très grand nombre de valeurs c'est pourquoi on est contraint de regrouper ces valeurs dans des classes, pour que le tableau ne soit pas trop long et ennuyeux.

Pour déterminer ces classes et construire le tableau de distribution des fréquences absolues (effectifs) on doit suivre les étapes suivantes :

① définir l'étendue générale (E) des différentes valeurs

E = Max (xi) – Min (xi)

② définir le nombre de classes (N C)). En général on peut définir le nombre de classes des deux manières différentes selon la taille de la population

- So N ≤ 100 on calcule (NC) en utilisant le théorème De YULE : NC = 2,5 $\sqrt[4]{N}$
- Si N> 100, on calcule NC en utilisant le théorème de STURGE:NC = 1+ 3,32 log N

③ définir l'amplitude (a_i) c'est-à-dire le nombre de valeur comprises dans chaque classe.

Les classes qu'on déterminera auront des amplitudes égales ai = $\frac{E+1}{N\ C}$

④ Ecrire les limites de la première classe [$L_0 – L_1$ [

L_0 = limite inférieure ; L_0 = Min (x_i)

L_1 = limite supérieure ; $L_1 = L_0 + a_i$

⑤ Enfin on procède à l'écriture de toutes les autres classes.

Remarque : quand la variable est continue, on forme des classes ayant des limites communes [10-14[; [14-18[; [18-22[

- Quand la variable est discontinue mais avec de nombreuses valeurs, on peut aussi fermer des classes mais, celles-ci seront discontinues (éloignées les unes des autres, n'ayant aucune valeurs commune) [10-14]; [15-19] ; [20-24]
- Quand le tableau de distribution comporte des classes, on doit calculer le centre de chaque classe x_i (limite inférieure + limite supérieure divisé par 2)

Exemple 3: les donnés suivantes représentent le revenu horaire (DA) de 50 employés d'une entreprise (E)

62-63-82-63-64-65-66-68-65-81

62-67-70-68-79-62-72-75-69-63

63-71-70-68-77-68-62-73-76-70

68-80-65-62-65-64-85-65-64-67

65-70-67-66-63-88-70-73-62-68

Travail à faire : mettre ces données dans un tableau de distribution des effectifs.

Dans cet exemple ; la population étudiée est représentée par les 50 employés de l'entreprise (E). Le caractère étudiée est le revenu horaire (DA) c'est un caractère quantitatif continu.

Puisque le nombre de valeurs de la variable est assez important (50) ; il est préférable de regrouper ces valeurs dans des classes avant d'établir le tableau de distribution des effectifs.

Pour former ces classes ont suit les étapes suivantes :

① Calcul de l'étendue générale (E)

E = Max (xi) – Min (xi) $\Longrightarrow$ E = 88-62 $\Longrightarrow$ E = 26

② Calcul du nombre de classe (NC)

N = 50 $\rightarrow$ N $\leq$ 100 = théorème de Yule

NC = 2,5 $\sqrt[4]{50}$ $\Longrightarrow$ NC = 2,5x2, 659 $\Longrightarrow$ NC = 6,64 $\Longrightarrow$; NC $\approx$ 7

Remarque : quand on calcule NC, on doit essayer d'avoir un nombre entier, c'est pourquoi il faut arrondir le résultat à l'unité la plus proche.

③ Amplitude des classes (ai) : ai = $\frac{E+1}{NC}$ $\Longrightarrow$ ai = $\frac{26+1}{7}$ $\Longrightarrow$ ai = 3,857 $\Longrightarrow$ ai $\approx$ 4

④ Ecriture de la première classe [L_0-L_1 [

L_0 = Min (xi) = 62 ; L1 = L_0+a $\Longrightarrow$ L_1 = 62+4 $\Longrightarrow$ L_1 = 66 $\Longrightarrow$ 1ère classe : [62-66[

⑤ Tableau de distribution des effectifs

Classes	Effectifs ni	Centre de classe xi
[62-66[	20	$\varkappa 1 = \frac{62+66}{2}$ = 64
[66-70[	12	$\varkappa 2 = \frac{66+70}{2}$ = 68
[70-74[	9	$\varkappa 3 = \frac{70+74}{2}$ = 72
[74-78[	3	$\varkappa 4 = \frac{74-78}{2}$ = 76
[78-82[	3	$\varkappa 5 = \frac{78-82}{2}$ = 80
[82-86[	2	$\varkappa 5 = \frac{82-86}{2}$ = 84
[86-90[	1	$\varkappa 6 = \frac{86-90}{2}$ = 88
Total	$\sum$ni = 50 = N	

I.3.2.2. Tableau de distribution des fréquences relatives et des pourcentages

Il est fréquent que l'on mette une distribution par effectifs sous forme d'une distribution de fréquences relatives ou sous forme de distribution de pourcentages.

La fréquence d'une modalité est la proportion d'individus de la population totale qui présentent cette modalité : elle est obtenue en divisant l'effectif de cette modalité du caractère ni par l'effectif total N et notée fi, soit : fi = $\frac{ni}{N}$ = $\frac{ni}{\sum \text{ni}}$ (**fréquence relative**)

La notion d'effectif d'une modalité est une notion absolue, elle ne permet pas directement les comparaisons. Par contre la notion de fréquence est une notion relative, elle permet directement les comparaisons.

Exemple :

Considérons l'exemple du groupe de trente étudiants. On a regroupé les fréquences correspondant à l'âge des étudiants dans le tableau suivant :

Âge	Effectif ni	Fréquence fi
18	2	2/30=0,067
19	4	4/30=0,133
20	10	10/30=0,333
21	11	11/30=0,367
22	3	3/30=0,100
Total	30	30/30=1

Pour faciliter la lecture des fréquences relatives, il est souvent préférable d'utiliser des pourcentages qui sont beaucoup plus pratiques et pertinents dans les interprétations.

$$fi\% = fix100 = \frac{ni}{N} \times 100$$

Exemple 4

Soit le tableau suivant représentant la taille en cm de 50 enfants

Taille en cm classes	Effectifs ni	Fréquence relative fi	Pourcentage fi%
[110-113 [	10	$f1 = \frac{10}{50} = 0,2$	0,2 x 100=20
[113-116 [	15	$f\ 2 = \frac{15}{50} = 0,3$	0,3x100=30
[116-119 [	12	$f\ 3 = \frac{12}{50} = 0,24$	0,24x100=24
[119-122 [	7	$f\ 4 = \frac{7}{50} = 0,14$	0,14x100=14
[122-125 [	3	$f\ 5 = \frac{3}{50} = 0,06$	0,06x100=6
[125-128 [	2	$f\ 6 = \frac{2}{50} = 0,04$	0,04x100=4
[128-131 [	1	$f\ 7 = \frac{1}{50} = 0,02$	0,02x100=2
Total	50	∑fi = 1	∑fi% = 100

Remarque :

- ∑fi = 1

  ∑fi% = 100

  $\sum fi = \frac{\sum ni}{N}$
- 30% des enfants ont une taille comprise entre 113 et116 cm

I.3.2.3.Tableau de distribution de fréquences cumulées :

On utilise la fréquence cumulée pour déterminer le nombre d'observations qui se situent au-dessus d'une valeur ou classe particulière dans un ensemble de données.

Il existe deux types de fréquences cumulées :

Les fréquences cumulées croissantes et les fréquences cumulées décroissantes, on calcule les fréquences cumulées (absolues, relatives ou pourcentages) à partir du tableau de distribution des fréquences.

Les modalités d'un caractère variant de 1 à k, l'effectif cumulé d'une modalité i est le nombre d'individus de la population présentant une modalité d'indice inférieur ou égal à i.

Exemple 1 : Le tableau suivant représente le nombre de pièces par logement

Travail à faire : 1- **calculer les fréquences absolues cumulées croissantes** NCi↑

Inférieur ou égal ≤

Nbre de pièces x_i	Nbre de logement n_i	Effets cumulés croissants NCi↑
1	⑤	⑤
2	10	5+10=15
3	20	15+20=35
4	30	35+30=65
5	25	65+25=90
6	10	90+10=100
Total	100	

Quel est le nombre de logements qui ont un nombre de pièces inférieur ou égal à 1 : c'est 5

2- avec les mêmes données de l'exemple.

Calculons **les fréquences absolues cumulées décroissants** NCi↓ et les fréquences relatives cumulées décroissants FCi↓ et les fréquences en pourcentages cumulées décroissantes FCi%↓

≥ Supérieur ou égal

Nombre de pièces xi	Nombre de logements ni	Effectifs cumulés décroissants NCi↓	Fréquence relative fi	fréquences relatives cumulées décroissantes FCi↓	Fréquence en pourcentage fi%	fréquences en pourcentages cumulées décroissantes FCi%↓
1	5	100	0,05	1	5	100
2	10	100-5=95	0,1	0,95	10	95
3	20	95-10=85	0,2	0,85	20	85
4	30	85-20=65	0,3	0,65	30	65
5	25	65-30=35	0,25	0,35	25	35
6	10	35-25=10	0,1	0,1	10	10
Total	100		1		100%	

Exemple 2 : le tableau suivant représente le retard (mn) des employés à rejoindre leur lieu de travail durant le début de la matinée. Calculer NCi↑, NCi↓, FCi↑, FCi%↑

Le retard en mn (classes)	Nbre employés ni	Limité inférieure -Plus de -	NCi↓	Limité supérieure -moins de-	NCi↑	Fi	FCi↑	FCi%↑
[0- 5 [	4	Plus de 0	72	Moins de 5	4	0,05	0,05	5
[5-10 [	7	Plus de 5	68	Moins de 10	11	0,1	0,15	15
[10-15[	10	Plus de 10	61	Moins de 15	21	0,14	0,29	29
[15-20[	15	Plus de 15	51	Moins de 20	36	0,21	0,50	50
[20-25[	18	Plus de 20	36	Moins de 25	54	0,25	0,75	75
[25-30[	12	Plus de 25	18	Moins de 30	66	0,17	0,92	92
[30-35[	6	Plus de 30	6	Moins de 35	72	0,08	1	100
Total	72					$\bar{\bar{1}}$		

Du Calcul de différentes fréquences cumulées on peut savoir :

- Nbre d'employés dont le retard est supérieur à 10Min : 61 employés
- Nbre d'employés dont le retard est inférieur à 20Min : 36 employés
- Proportion des employés dont le retard est inférieur à 20Mn : 0,5
- Pourcentage des employés dont le retard inférieur à 25Mn : 75%
- Nombre d'employés qui ont un retard compris entre 10 et 25 mn

De NCi↑ : 54-11 = 43 employés

De NCi↓ : 61-18 = 43 employés.

Chapitre II : Les représentations graphiques :

Le classement des données dans un tableau nous permet souvent de comprendre le phénomène étudié, mais dans certaines situations, ces tableaux ne suffisent pas pour avoir une idée claire, c'est pourquoi on utilise une autre manière de présentation sous forme de graphes, qui permettent de donner une idée rapide et mette du phénomène étudié et visualiser d'un seul coup d'œil les principales caractéristiques (mais on perd une quantité d'informations),

On distingue les méthodes de représentation d'une variable statistique en fonction de la nature de la variable (qualitative ou quantitative). Les représentations recommandées et les plus fréquentes sont les tableaux et les diagrammes (graphe).

Les graphes différents selon la nature du caractère étudié.

II.1.La représentation graphique d'un caractère qualitatif :

Dans cette situation les principaux graphes qu'on peut utiliser sont :

II.1.1. Le diagramme en barres (tuyaux d'orgues)

Le graphe est constitué de rectangles ayant des largeurs constants mais des longueurs proportionnelles aux fréquences (absolues, relatives ou pourcentages). Sur l'axe des abscisses on représente les modalités et sur l'axe des ordonnées on porte les fréquences ou aux effectifs de chaque modalité.

Exemple1 : Le tableau suivant représente la distribution d'un échantillon de personnes d'une ville d'Algérie, selon leur situation matrimoniale

Travail à faire :

Représentez ces données en utilisant un diagramme en barres

Modalités -mi-	Effectifs ni	fi	fi%
Célibataire	150	0,42	42
Marié	120	0,33	33
Veuf	10	0,03	3
Divorcé	80	0,22	22
Total	360	≈ 1	100

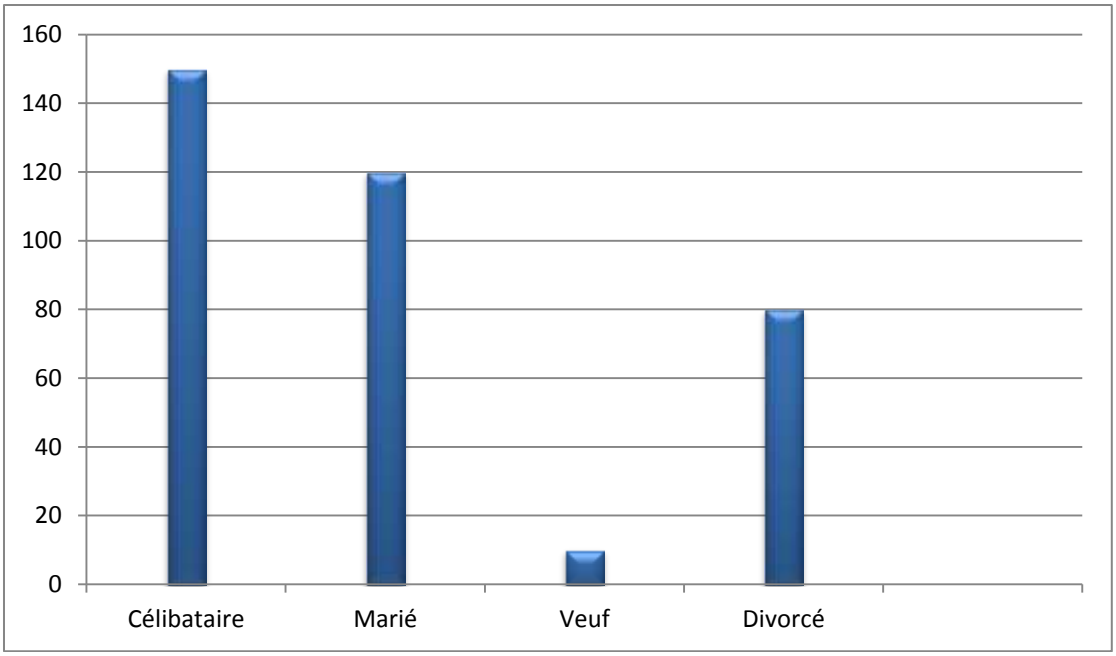
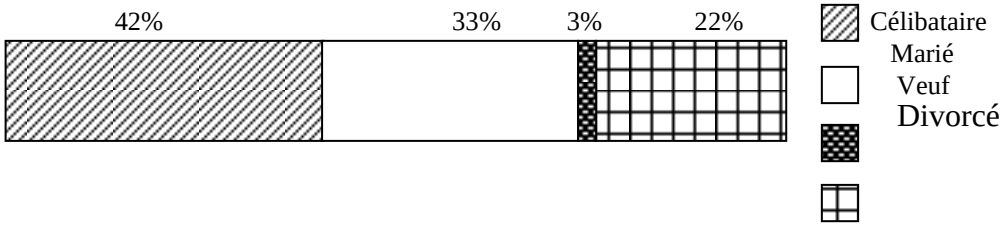


Diagramme en barre

II.1.2. Le diagramme rectiligne :

C'est un rectangle divisé en parties, chaque partie représente une modalité du caractère exprimée en pourcentage.

Exemple2 : reprenons l'exemple précédent et traçons le diagramme rectiligne



II.1.3. Le diagramme circulaire (en secteurs)

C'est un cercle divisé en secteurs, chaque secteur représentant une modalité du caractère, Ces diagrammes conviennent très bien pour des données politiques ou socio-économiques. on calcule l'angle que représente chaque secteur, en utilisant :

$$\alpha = fi \times 360$$

Exemple 3: le tableau suivant représente la distribution des étrangers en France selon leur nationalité. Représentez ces données à l'aide d'un diagramme circulaire.

Mi	ni	fi	α = fi x360°
Marocains	150	0,25	90
Algériens	200	0,333	120
Tunisiens	125	0,208	75
Moyen orient	50	0,083	30
Portugais	75	0,125	45
Total	600	≈1,00	360°

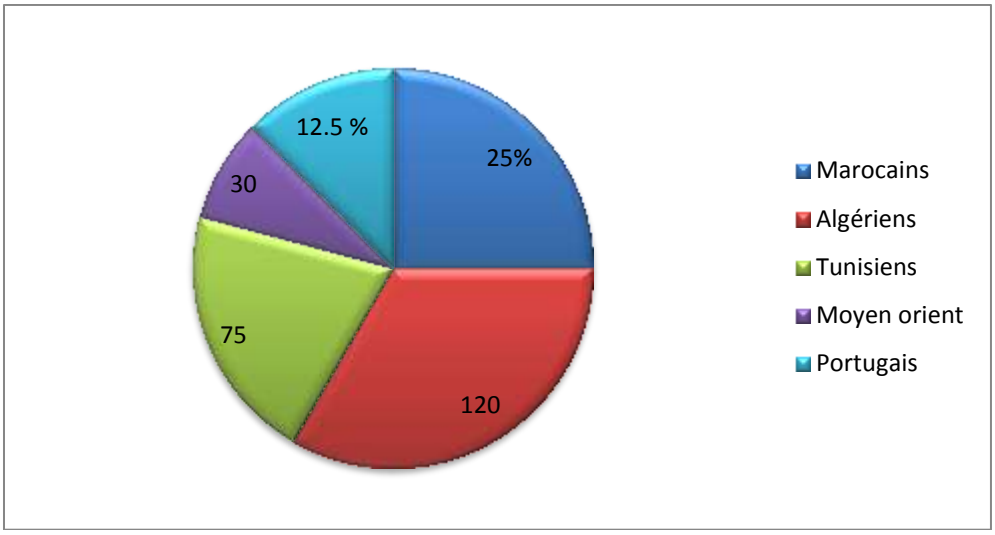


Diagramme circulaire(en secteur)

II.2. La représentation graphique d'un caractère quantitatif

Dans cette situation on doit distinguer entre une variable discrète et une variable continue

II.2.1.Cas d'une variable discrète (discontinue) : Les différentes représentations sont :

II.2.1.1.Le diagramme en bâtons (différentiel)

C'est de simple bâtons dont la hauteur est proportionnelle à la fréquence (absolue, relative ou en pourcentage) de chaque valeur de la variable.

Exemple 4: Le tableau suivant donne la distribution d'un échantillon de logements d'un quartier selon le nombre de pièces. Représentez graphiquement ces données.

xi	**ni**	**Nci↑**	**Nci↓**
1	5	5	100
2	10	15	95
3	20	35	85
4	30	65	65
5	25	90	35
6	10	100	10
Total	**100**		

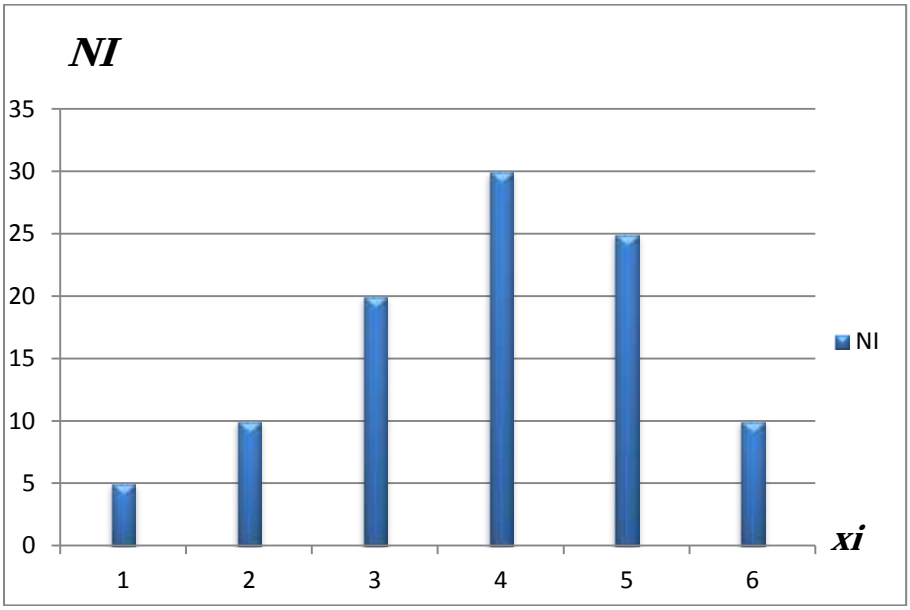


Diagramme en bâtons

I.2.1.2.Diagramme cumulatif (intégral)

C'est des segments de droites qui représentent les fréquences cumulées (ascendantes « croissantes » ou descendantes « décroissantes » des valeurs de la variable.

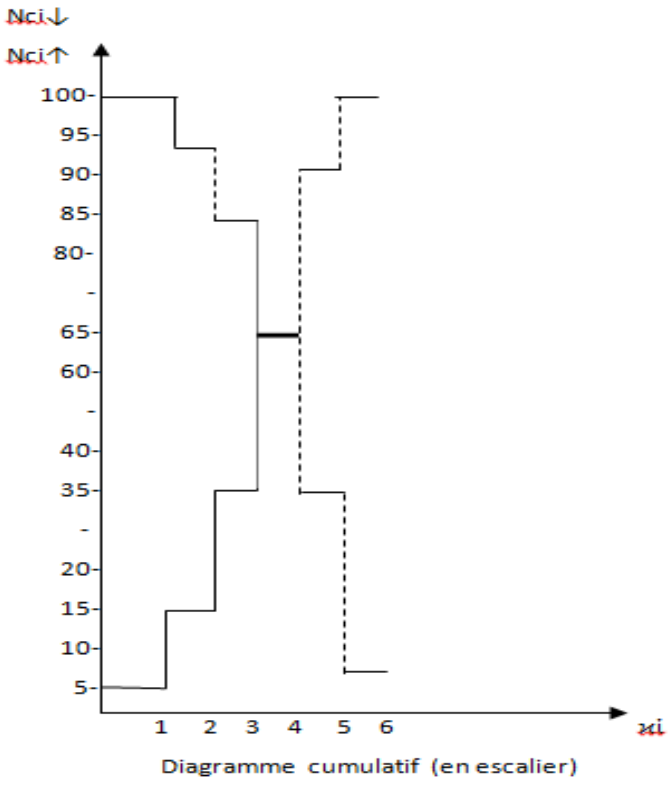


Diagramme cumulatif (en escalier)

II.2.2.Cas d'une variable continue

Parmi les différents graphes on distingue :

II.2.2.1. L'histogramme :

C'est la représentation graphique des donnés lorsqu'elles sont groupées en classe autrement dit c'est la représentation graphique d'une série statistique dont le caractère est quantitatif continu.

L'histogramme est composé de rectangles accolés, de hauteurs variables correspondantes à l'effectif ou fréquence relative et de largeurs correspondantes aux classes (amplitude).

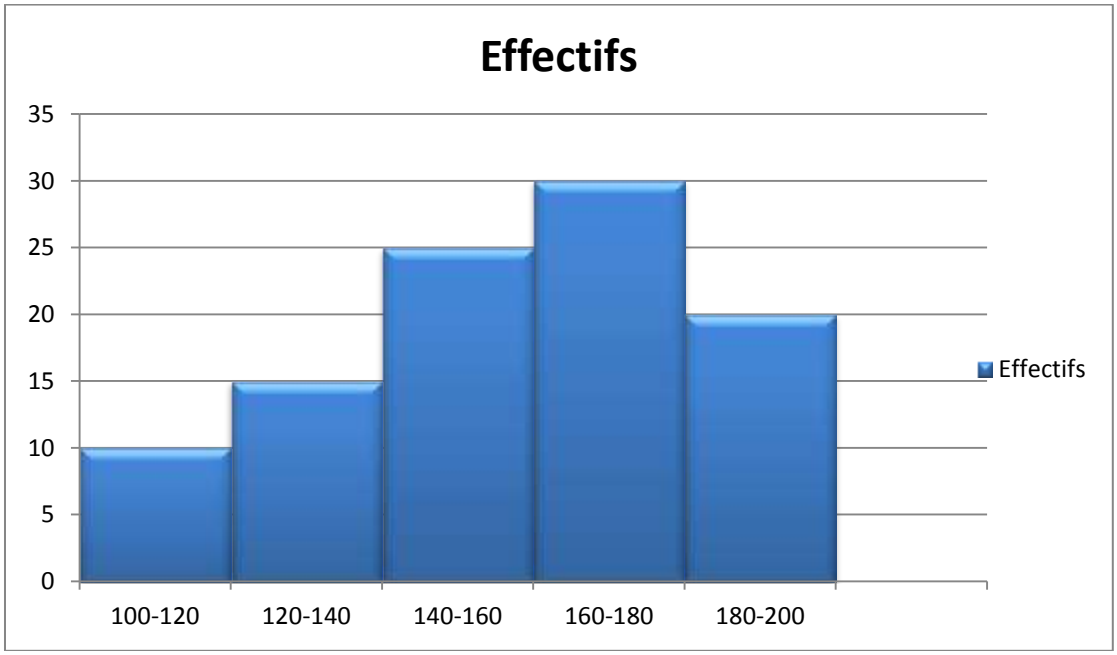
Remarque : Lors de la construction d'un histogramme on peut se trouver face à deux situations :

1ère situation : Les amplitudes sont égales ; On construit l'histogramme directement

Exemple 5 : Le tableau suivant représente le revenu horaire en DA d'un échantillon d'employés

Représentez ces données à l'aide d'un histogramme.

Classes	***Effectifs ni***
[100-120[	10
[120-140[	15
[140-160[	25
[160-180[	30
[180-200[	20
Total	100



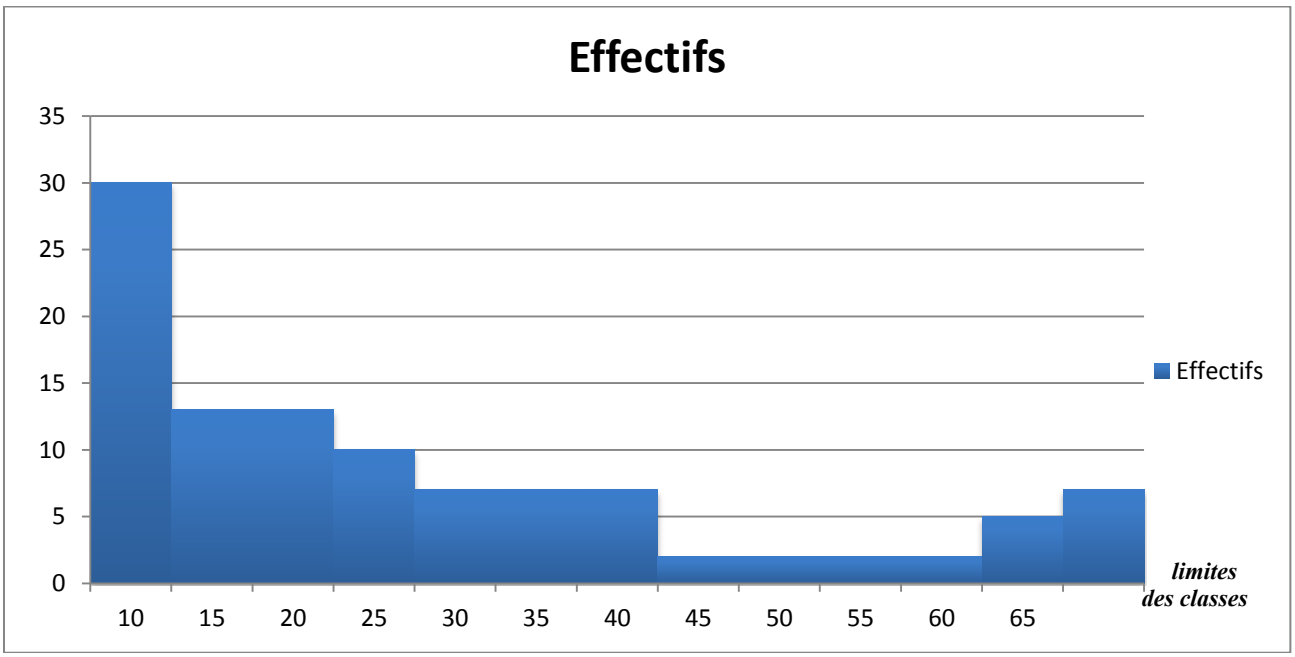
2ème situation : Les aptitudes sont inégales : il faut corriger les effectifs (ni) avant de tracer l'histogramme de la manière suivante :

- Calculer les amplitudes (ai) de chaque classe.
- Déterminer l'amplitude unité (a_u) → PGDC des a_i / a_i le plus fréquent / a_i minimum
- Calculer les coefficients (Ci) : $Ci = \frac{ai}{au}$
- Calculer les effectifs corrigés (h_i) : $h_i = \frac{ni}{ci}$

Exemple 6 : Le tableau suivant représente la distribution du revenu mensuel en milliers de DA, d'un échantillon de travailleurs dans une société privée. Tracez l'histogramme

Classes	ni	ai	$c_i=\frac{ai}{au}$	$h_i=\frac{ni}{ci}$
[10-15[	30	5	1	30
[15-25[	26	10	2	13
[25-30[	10	5	1	10
[30-45[	21	15	3	7
[45-65[	8	20	4	2
[65-70[	5	5	1	5
Total	100			

Remarque : on est dans la situation où les classes ont des amplitudes inégales on doit donc corriger les effectifs avant de tracer l'histogramme
Le plus grand diviseur commun (PGDG) des amplitudes est 5 → $a_u = 5$



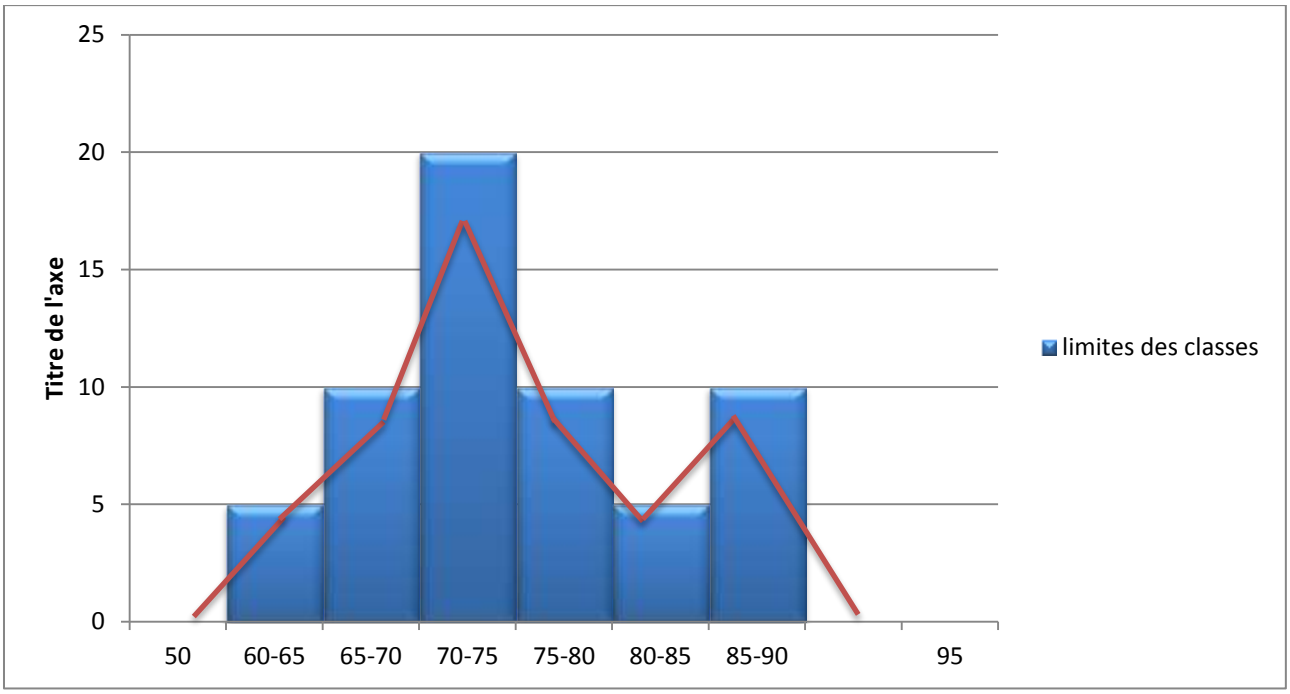
L'histogramme corrigé des effectifs

II.2.2.2. Le polygone de fréquence

On construit le polygone de fréquences en joignant les points situées au milieu de chaque sommet des rectangles par des segments de droite. Le polygone est fermé aux deux bouts en le prolongeant sur l'axe horizontal

Exemple 7 : Le tableau suivant représente la distribution d'un échantillon d'étudiants selon leur poids en Kg. Tracez le polygone de fréquence de cette distribution

Classes	ni	xi
[60-65[	5	62,5
[65-70[	10	67,5
[70-75[	20	72,5
[75-80[	10	77,5
[80-85[	5	82,5
[85-90[	10	87,5
Total	60	



Le Polygone de Fréquences

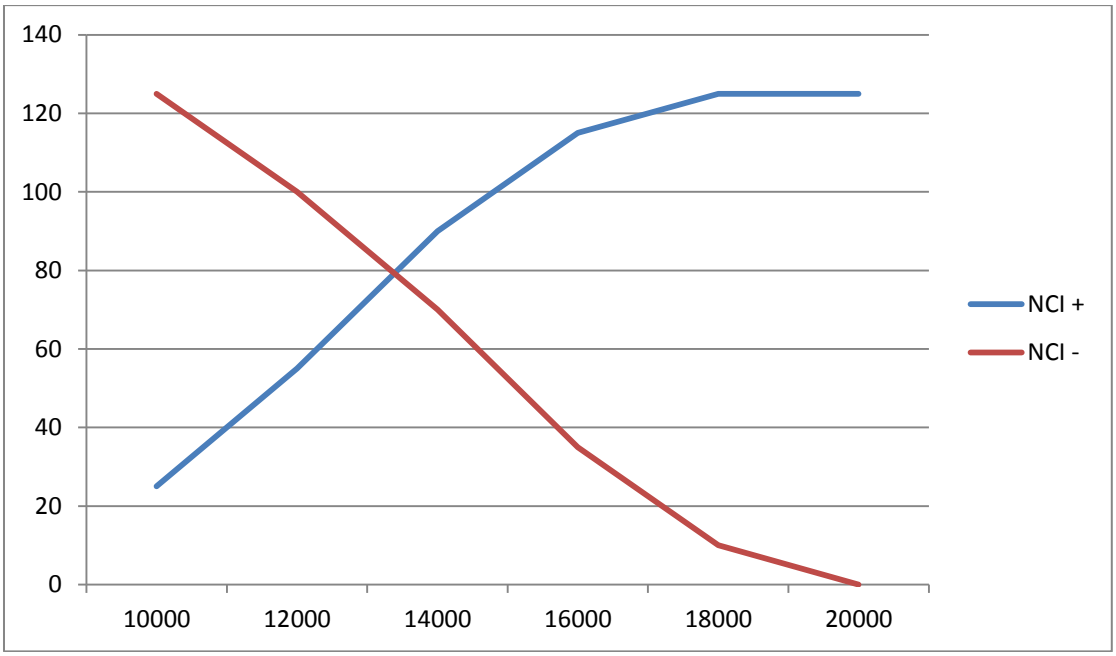
Le polygone de fréquence donne une vision plus réaliste de la distribution en éliminant les ruptures entre les classes.

II.2.2.3. Les courbes cumulatives (croissante et décroissante)

On obtient les courbes cumulatives en joignant les points obtenus grâce aux limites de classes en abscisses et les fréquences cumulées en ordonnées.

Exemple 8 : Le tableau suivant nous donne la répartition des employés d'une entreprise selon leur salaire mensuel en DA.

Classes	ni	Limite sup (moins de)	Nci↑	Limite inf (plus de)	Nci↓
[10.000-12.000[	25	Moins de 12000	25	Plus de 10000	125
[12.000-14.000[	30	Moins de 14000	55	Plus de 12000	100
[14.000-16.000[	35	Moins de 16000	90	Plus de 14000	70
[16.000-20.000[	25	Moins de 20000	115	Plus de 16000	35
[20.000-25.000[	10	Moins de 25000	125	Plus de 20000	10
Total	125				



Courbes cumulatives ascendants et descendants

Chapitre III : Caractéristique de tendance centrale :

Introduction :

Devant la distribution statistique d'un caractère quantitatif, le traitement statistique des informations ne se limite pas aux seules représentations graphiques, le premier objectif du statisticien est de caractériser cette distribution par un chiffre et de résumer les informations par des indicateurs numériques bien choisis. Dites de « tendance centrale ». Ces nombres résumés sont ainsi appelés car ils privilégient les valeurs principales de la distribution.

Tout le problème est de définir le chiffre le plus pertinent pour cela. On peut en effet construire plusieurs indicateurs pour caractériser une série statistique. C'est ce que nous allons voir dans ce chapitre avant d'étudier les façons de comparer deux situations au moyen des indices.

Le rôle des paramètres de position ou de dispersion est de transmettre, par un calcul, une information liée à une réalité statistique.

L'objet de ce chapitre est de présente quelques-uns de ces indicateurs. Les plus utilisés sont : Le mode, la médiane et la moyenne arithmétique, connues sous le nom de caractéristiques de tendances central ou caractéristiques de position. Nous exposerons leur mode de calcul et leur signification en distinguant pour chacune d'elles le cas des données non groupées et le cas des données regroupées (soit par valeurs, soit par classes).

III.1. Le Mode : on appelle mode d'une série statistique, et on note « Mo »la modalité qui admet **la plus grande fréquence** (ou ayant le plus grand effectif).

Il est parfaitement défini pour une variable qualitative ou une variable quantitative discrète ;

Le mode est un indicateur de position insensible aux valeurs extrêmes.

III. 1.1 cas d'une variable discrète :

La détermination du mode est immédiate à partir du tableau de dénombrement (valeur de la variable correspondant à l'effectif maximal).

Exemple 1 : dans une PME on a comptabilisé pendant un an le nombre de jours d'absence arrêt-maladie de chacun des 12 employés : 11-5-3-5-7-3-11-13-8-3-0-3

La valeur de la variable « nombre de jours d'absence » qui est la plus fréquente est la valeur3.

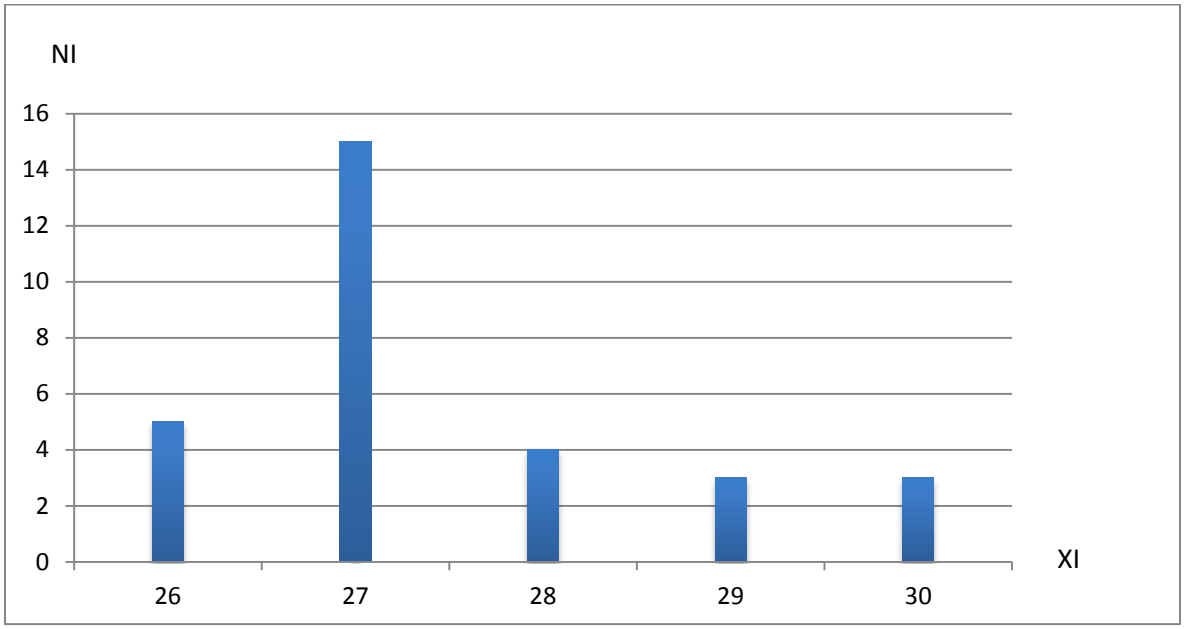
Le mode de la série est donc 3. Mo = 3 jours

***Exemple 2* :** on considère une distribution de l'âge au mariage de 30 personnes.

Age au mariage xi	26	27	28	29	30
Effectif ni	5	15	4	3	3

L'effectif maximal est « 15 » associé à la valeur « 27 » de la variable. Le mode est donc Mo=27

Graphiquement le mode correspond à l'abscisse du bâton le plus élevé.



III .1. 2 cas d'une variable continue :

On appelle mode d'une variable statistique continue et on note Mo la valeur de la variable associée à l'effectif (on a la fréquence) le plus élevé par unité d'amplitude.

Pour une variable quantitative continue nous parlons de **classe modale** : c'est la classe dont la densité de fréquence est maximum.

- La détermination est faite à partir du tableau de dénombrement. Elle s'effectue en deux temps :
- Détermination de la classe modale
- Détermination du mode

Cependant, on distingue deux cas selon que les amplitudes des classes sont égales on inégales.

III .1. 2.1. Cas d'amplitudes identiques (égales) :

Dans ce cas, la classe modale est la classe d'effectif n_i le plus élevé, soit $[L_0, L_1[$

- l'effectif de la classe qui précède la classe modale est n_{i-1}
- L'effectif de la classe qui suit la classe modale est n_{i+1}

Donc : $$M_0 = L_0 + a_i \left[\frac{m1}{m1+m2}\right]$$

- L_0 : Limite inférieur de la classe modale
- L_1 : Limite supérieure de la classe modale
- a_i : amplitude de la classe modale
- $m_1 = n_i - n_{i-1}$
- $m_1 = n_i - n_{i+1}$

On peut aussi déterminer le mode avec la fréquence relative :

$$M_0 = L_0 + a_i \left[\frac{f_i - f_{i-1}}{f_i - f_{i-1} + f_i - f_{i+1}}\right]$$

Exemple 3: soit la distribution de la population de 40 nourrissons selon la taille à la naissance :

	Classe (taille en cm)	Amplitudes	Effectifs ni
	[48-50[	2	3
Classe Modèle →	[50-52[	2	10
	[52-54[	2	9
	[54-56[	2	7
	[56-58[	2	7
	[58-60[	2	3
	[60-62[	2	1
	Total		**40**

$$Mo = 50+2\left[\frac{(10-3)}{(10-3)+(10-9)}\right] = 50+2\left[\frac{7}{7+1}\right] = 51{,}75$$

III .1. 2.2. Cas d'amplitudes inégales :

Dans le cas où les amplitudes sont différentes la classe modale est la classe de l'effectif corrigé le plus élevé.

Il faut corriger les effectifs (mi) avant de calculer le mode.

L'effectif corrigés (h_i) : $h_i = \frac{ni}{Ci}$

Les coefficients (C_i) : $C_i = \frac{ai}{au}$

L'amplitude unité (au) → ai la plus fréquente

→ ai la plus faible

Dans ce cas : $M_1 = h_i - h_{i-1}$

$M_2 = h_i + h_{i-1}$

Le mode est donné par : $M_0 = L_0 + a_i \left[\frac{m_1}{m_1+m_2}\right]$

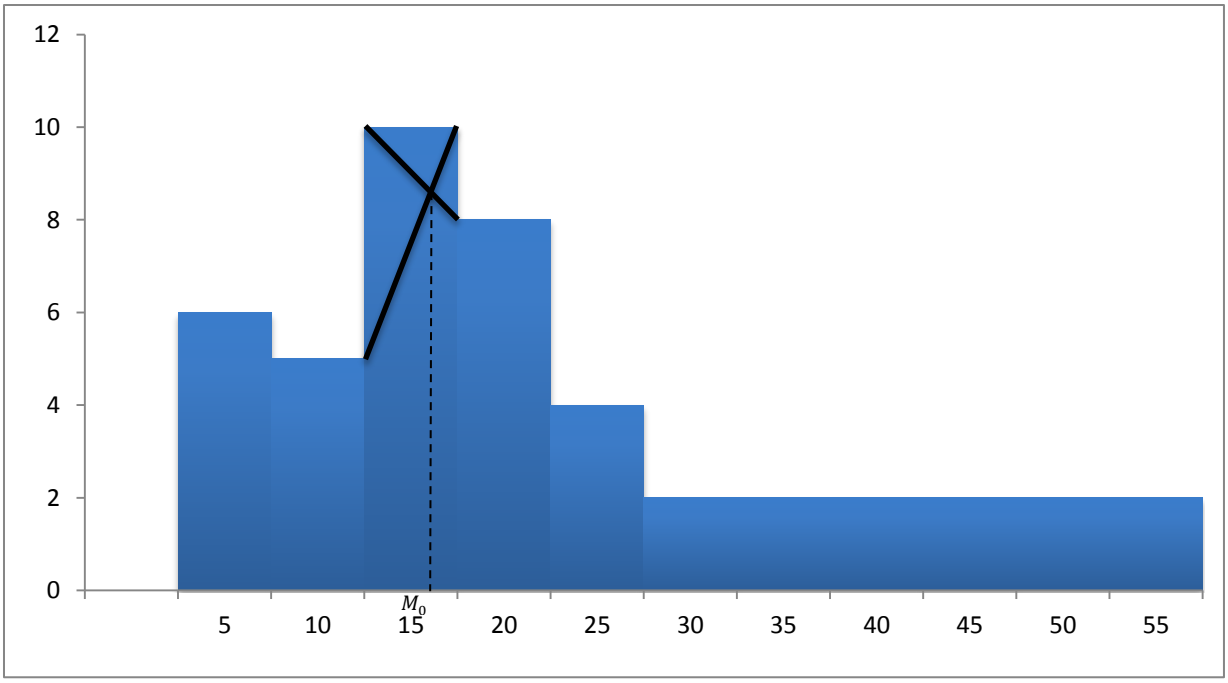
Exemple 4*:* soit la répartition de 100 personnes selon leur âge :

	Classes d'âges	**Effectif n_i**	**Amplitudes a_i**	C_i	h_i
	[5 ; 10[	11	5	1	11
	[10 ; 15[	10	5	1	10
Classe modèle →	[15-20[	20	5	1	20
	[20-30[	30	10	2	15
	[30-40[	18	10	2	9
	[40-60[	11	20	4	2,75
	Total	100			

$a_u = 5$

$Mo = 15+5\left[\frac{20-10}{[20-10]+[20-15]}\right] = 15+\frac{50}{10+5} = 15+\frac{50}{15}$

Mo = 18,33



L'estimation du Mo peut être faite graphiquement en considérant l'intersection des diagonales du trapèze ABCD, Ce trapèze est construit à partir des sommets du rectangle associé à la classe modale et des deux rectangles contigus.

Remarque :

Si toutes les modalités on le même effectif la distribution statistique est donc uniforme (elle ne possède pas de mode).

Si la série possède un seul mode, la distribution statistique est unimodale.

Si la série possède deux ou plusieurs mode la distribution est respectivement qualifiée de bimodale et multimodale.

III.2. la Médiane :

On appelle médiane d'une distribution et on note Me la valeur de la variable qui partage la population statistique étudiée en deux effectifs égaux, les individus étant ordonnés selon les valeurs de la variable. Ce sera donc la valeur de la variable telle que 50% de la population se situe au-dessus et 50% se situe en dessous.

III. 2 .1. Cas d'une variable discrète :

Pour déterminer la médiane d'une série de données non groupées, on ordonne dans un premier temps les valeurs observées en ordre croissant. La médiane est la valeur de la variable située « au milieu » de la série ordonnée.

- Lorsque le nombre d'observation n est impair, la médiane est la valeur de la série ordonnée située à la position $\frac{n+1}{2}$

- Lorsque le nombre d'observation n est pair, la médiane se trouve à l'intérieur de l'intervalle médian compris entre les deux valeurs centrales situées aux positions $\frac{n}{2}$ et $\frac{n}{2}$+1 donc la médiane est le centre de l'intervalle médian.

Exemple 5 *:* soit la répartition d'un groupe de neuf ménages selon le nombre de personnes par ménage :

3 – 1 – 4 – 6 – 2 – 4 – 3 – 5-7

En ordonnant la série suivant un ordre croissant, on obtient

1 – 2 – 3 – 3 – **4** – 4 – 5 – 6-7

La médiane ici est la cinquième valeur ($\frac{n+1}{2}$= $\frac{9+1}{2}$ = 5) donc Me = 4 personnes.

Autre méthode pour déterminer la médiane :

- Si le nombre d'observation n est impair, n prend la forme n= 2p+1 et donc Me se trouve au rang (p+1).
- le nombre d'observation n est pair, n prend la forme n= 2p, dans ce cas Me se trouve à l'intérieur de l'intervalle médian compris entre les deux valeurs centrales situées aux positions entre (p) et (p+1).

Exemple 6:

Soit la répartition d'un groupe de ménage 12 ménages selon le nombre de personnes par ménage.

5-3-2-3-6-3-5-4-7-2-1-4

En ordonnant la série suivait un ordre croissant on obtient :

$$\underbrace{1-2-2-3}_{\text{5 valeurs}}-\underbrace{3-3-4-4}_{\text{|Intervalle médiane|}}-\underbrace{5-5-6-7}_{\text{5 valeurs}}$$

La médiane se situe ici entre la sixième et septième valeur (c.-à-d. entre les positions$\frac{n}{2}$ = $\frac{12}{2}$=6 et$\frac{n}{2}$ +1= $\frac{12}{2}$ +1=7) donc Me = $\frac{3+4}{2}$ = 2,5 personnes.

Dans le cas d'une variable discrète, le calcul de la médiane se fait à partir des effectifs ou des fréquences cumulées : la médiane est alors lavaleur de la variable à laquelle est associé un effectif cumulé deN/2 (N étant l'effectif total) ou une fréquence cumulée de 50% de la population. Prenons l'exemple de la répartition par âge d'un groupe de trenteétudiants dans un cours.

Âge	Effectif n_i	Fréquence f_i	Effectif cumulé	Fréquence cumulée
18	3	2/30 = 06,7%	2	2/30 = 06,7%
19	4	4/30 = 13,3%	6	6/30 = 20,0%
20	9	9/30 = 30,0%	15	15/30 = 50,0%
21	11	11/30 = 36,0%	26	26/30 = 86,7%
22	4	3/30 = 10,0%	30	30/30= 100%
Total	30	1=100%		

Dans l'exemple ci-dessus, l'effectif total de la population étudiée est de 30.La médiane est la valeur de la variable qui partage la population en deux, c'est-à-dire la valeur en dessous de laquelle on trouve 30/2 = 15 étudiants.La moitié de l'effectif est atteint pour l'âge de 20 ans. En effet, on a un effectif cumulé de 15 étudiants ayant 20 ans et moins.

Quand la médiane ne tombe pas sur une valeur exacte de la variable, par convention on retient la valeur de la variable immédiatement supérieure. Quand la variable est continue et classée, le calcul ne peut se faire que par approximation : on traite les variables par interpolation linéaire comme si les effectifs étaient uniformément répartis à l'intérieur d'une classe.

Propriété de la médiane :

La médiane donne des indications utiles sur la tendance centrale d'une distribution statistique. Elle n'est pas influencée par les valeurs extrêmes de la variable.

II .2.2. Cas d'une variable continue :

Il n'y a aucune différence de calcul pour la médiane, selon que les classes sont d'amplitudes égales ou inégales.

Le calcul de la médiane passe d'abord par la détermination de la classe médiane.

Soit [L0, L1[la classe médiane, a_i l'amplitude de la classe médiane, Ni l'effectif cumulé croissant de la classe médiane, Ni-1 l'effectif cumulé croissant de la classe avant la classe médiane et N l'effectif total.

$$Me = L0 + a_i \left| \frac{\frac{N}{2} - Ni - 1}{Ni - Ni - 1} \right|$$

Ou avec la fréquence relative :

$$Me = L0 + a_i \left| \frac{0{,}5 - f\,i - 1}{f\,i - f\,i - 1} \right|$$

Exemple 7: en reprenant l'exemple sur la répartition des 100 personnes selon leur âge :

Classes d'âge	Effectif n_i	Effectifs cumulé croissants $N_i\uparrow$	Fréquence relative f_i	Fréquence cumulée croissants $f_i\uparrow$
[5-10[	11	11	0,11	0,11
[10-15[	10	21	0,10	0,21
[15-20[	20	41	0,20	0,41
[20-30[	30	71	0,30	0,71
[30-40[	18	89	0,18	0,89
[40-60[	11	100	0,11	1
Total	100		1	

$$Me = 20+10\left|\frac{50-41}{71-41}\right| = 20+10\left[\frac{9}{30}\right]$$

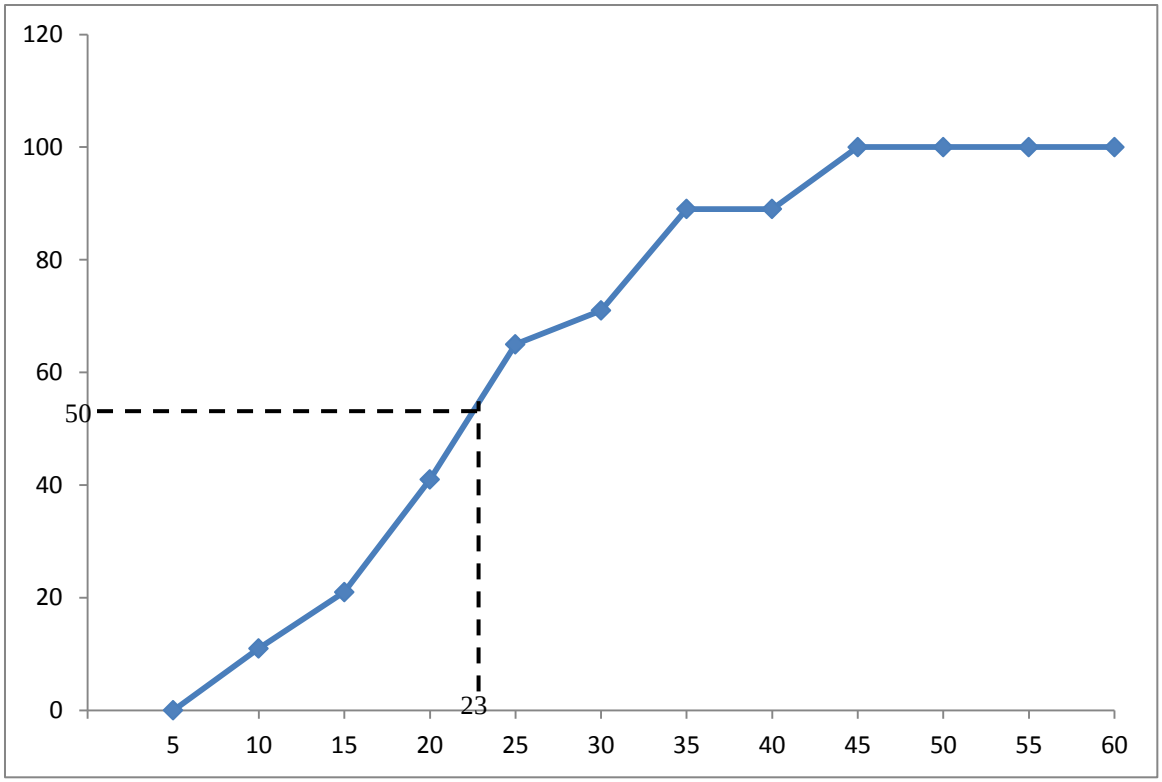
$$Me = 23 \text{ ans}$$

Ou par la fréquence relative

$$Me = 20+10\left|\frac{0,5-0,41}{0,71-0,41}\right| = 20+3 = 23 \text{ ans}$$

Détermination graphique de la médiane :

$N_i\uparrow$



III.3. Les moyennes :

On suppose dans cette section que X est une variable quantitative définie sur une population composée de N individu.

III. 3.1. La moyenne arithmétique :

La moyenne arithmétique est le paramètre le plus utilisé. Il a l'inconvénient d'être sensible à des valeurs aberrantes, ce qui le rend moins significatif dans certains cas.

III.3.1.1. Donnée non groupée (La moyenne arithmétique simple) :

On appelle moyenne arithmétique la somme de toutes les données statistiques divisée par le nombre de ces données. La moyenne arithmétique conserve la somme totale des valeurs observées : si on modifie les valeurs de deux observations d'une série statistique tout en conservant leur somme, la moyenne de la série sera inchangée.

Soit x1, x2, …. Xn les N observations de la variable X on appelle moyenne arithmétique simple des N valeurs (x1, x2, ….. xn), ou encore moyenne arithmétique de la variable Xi et on note $\bar{X}$ le rapport défini par :

$$\bar{X} = \frac{x1+x2+\cdots.+xn}{N}$$

Soit, en utilisant le signe $\sum$: $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} Xi$

Exemple 8: les données du tableau suivant précisent l'évolution de la production d'un produit donnée d'une entreprise en millier de 2010 à 2017.

Année	2010	2011	2012	2013	2014	2015	2016	2017
Production	5,45	5,65	5,29	5,67	5,55	4,94	5,12	5,84

$$\bar{X} = \frac{5{,}45+5{,}65+5{,}29+5{,}67+5{,}55+4{,}99+5{,}12+5{,}84}{8} = 5{,}445 \approx 5{,}45$$

III. 3.1.2. Données groupées :

III. 3.1.2.1.cas d'une variable discrète :

Lorsque les données d'une variable quantitative X sont organisées dans un tableau de dénombrement, chacune des valeurs observées xi de la variable est affectée d'une pondération (ou coefficient) égale au nombre d'observations distinctes égale à K.

On appelle moyenne arithmétique pondérée de K valeurs (x_1, x_2, …. x_k) affectée de K coefficient (n_1, n_2, …. n_K) ou encore moyenne arithmétique de la variable X, et on note $\bar{X}$ le rapport défini par :

$$\bar{X} = \frac{n1\varkappa 1+n2\varkappa 2+\cdots\ldots+nK\varkappa K}{n1+n2+\cdots\ldots nK} = \bar{X} = \frac{1}{N}\sum_{i=1}^{K} \text{ni}\varkappa\acute{\imath}$$

La moyenne des utilisant les fréquences f_i relatives :

$$\bar{X} = \sum_{i=n}^{k} f_i x_i$$

Exemple 9: on considère la répartition de 10 ménages selon le nombre d'enfant par ménage :

x_i	n_i	f_i	Produit $n_i x_i$	$f_i x_i$
0	2	0,2	0	0
1	2	0,2	2	0,2
2	1	0,1	2	0,2
3	3	0,3	9	0,9
4	2	0,2	8	0,8
Total	10	1	21	2,1

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{K} \text{ni}\varkappa\acute{\imath} = \frac{21}{10} = 2{,}1 \text{ enfants}$$

$$\bar{X} = \sum_{i=1}^{K} \text{ﻑ ί}\varkappa\acute{\imath} = 2{,}1 \text{ enfants}$$

III. 3.1.2.2.cas d'une variable continue :

Pour calculer la moyenne d'une variable continue dont les valeurs observées sont regroupées dans des classes statistiques on considère que les valeurs sont réparties uniformément à l'intérieur de chaque classe. Cette hypothèse revient à supposer que la moyenne des valeurs d'une classe est égale au centre de classe. Pour une classe donnée de centre Xi, contenant ni valeurs observées, on peut alors remplacer la somme des xi valeurs de la classe par le produit $n_i x_i$

On opère comme si les valeurs étaient concentrées au centre de classe.

Exemple10 *:* soit la répartition de 100 personnes selon leur âge :

Classes d'âges	Effectif n_i	Centre de classe x_i	$n_i x_i$
[10-20[	10	15	150
[20-30[	30	25	750
[30-40[	20	35	700
[40-50[	40	45	1800
Total	100		3400

Soit : $\bar{X} = \frac{3400}{100} = 34$ ans.

Propriétés de la moyenne arithmétique

Somme des écarts à la moyenne est nulle soit

$\sum_{i=1}^{n}(xi - \bar{x}) = 0$.

En effet : $\sum_{i=1}^{n}(xi - \bar{x}) = \sum_{i=1}^{n} xi - \sum_{i=1}^{n} \bar{x} = \sum_{i=1}^{n} xi - n\bar{x} = n\bar{x} - n\bar{x} = 0$

La moyenne arithmétique est un indicateur qui est facilement calculable et interprétable. La moyenne arithmétique est la plus fréquente mais elle ne s'applique pas à toutes les variables

III.3.2.Généralisation de la moyenne : Nous allons voir maintenant qu'il existe d'autres moyennes que la moyenne arithmétique et dans quelles circonstances utiliser celle-ci plutôt que celles-là ?

III.3.2.1. La moyenne géométrique :

La moyenne géométrique est très utilisée dans l'analyse de l'évolution d'une variable dans le temps.

Moyenne géométrique simple d'une variable x, notée G On doit distinguer les données groupées des données non groupées, lorsque l'on a des données non groupées on parle de

moyenne géométrique simple et pour des données groupées on parle de moyenne géométrique pondérée.

On appelle moyenne géométrique d'une distribution statistique, et on note G la racine nième du produit des N valeurs observées, soit :

III.3.2.1. 1.Moyenne géométrique simple :

G = $\sqrt[N]{\varkappa 1 \varkappa 2, \ldots . \varkappa n} = (\varkappa 1 \varkappa 2 \ldots \ldots \varkappa n)^{1/n}$

Ou : log G = $\sum_{i=1}^{K} \frac{log \varkappa i}{N}$

Exemple 11*:* calculer la moyenne géométrique des données suivantes : 2-4-8-10

G = $\sqrt[4]{2.4.8.10} = \sqrt[4]{640}$ = 5,029

Méthode 2 :

Log G = $\frac{1}{4}$[log 2+log 4+ log 8 +log 10]

Log G = $\frac{1}{4}$[0,30+0,60+0,90+1] = $\frac{2,8}{4}$ = 0,7

G = $10^{0,7}$ = 5,0118

III.3.2.1.2. Moyenne géométrique pondéré :

G = $\sqrt[N]{x_1^{n1}.x_2^{n2} \ldots \ldots X_K^{nK}} = (x_1^{n1}.x_2^{n2} \ldots \ldots . x_K^{nK})^{1/n}$

=$\varkappa 1^{f 1}.\varkappa 2^{f 2} \ldots . \varkappa_{iK}^{f K}$

Ou: log G =$\sum_{i=1}^{K} \frac{(ni\ log \varkappa i)}{N}$

Champ d'application de la moyenne géométrique :

Le calcul de G s'applique lorsque les valeurs de la variables se multiplient, c'est-à-dire suivent une progression géométrique. Elle permet l'estimation du rapport moyen d'une variable évolutive. Par exemple le taux d'accroissement moyen des bénéfices réalisés par une entreprise au cours d'une période donnés, le taux de croissance démographique (moyen d'une population etc.….

Exemple12 :

Au cours des 4 dernières années, les taux de croissance annuels du revenu national ont été les suivants.

1ère année → + 7,2%

2ème année → + 6,3%

3ème année → + 7,0%

4ème année → + 4,8%

Quel est le taux moyen de croissances du RN au cours de ces 4années ?

Solution :

G = $\sqrt[4]{107,2x106,3x107x104,8}$

Log G = $\frac{1}{4}[\log 107,2 + \log 106,3 + log107 + log104,8]$

Log G = 2,027

$\Longrightarrow$ G = $10^{2,027}$ = 106,3%

Au cours de cette période, le taux de croissance du RN est de 6,3%

Exemple 13:

Trois équipes se sont succédé à la direction d'une entreprise. Pendant la première période qui a duré quatre ans, les bénéfices réalisés ont augmenté de 50% par an, pendant la seconde période de trois ans, le 17% par an et pendant la 3ème période, qui a duré deux ans, les bénéfices ont enregistré une baisse de 30% par an. Quel est le taux de croissance annuel moyen des bénéfices réalisés au cours de ces 9 années.

Solution :

G = $\sqrt[9]{150^4.117^3.70^2}$

Log G = $\frac{1}{9}[4\log 150 + 3\ log117 + 2log70] = 2,06$

D'où : G = $10^{2,06}$=114,8%

L'augmentation annuelle moyenne des bénéfices est de 14,8%

Remarque :

1- Une moyenne géométrique nulle si une seule valeur est nulle. Elle n'est généralement définie que pour des séries à valeurs positives.
2- En pratique, la moyenne géométrique est essentiellement utilisée pour calculer une moyenne de ratios ou d'indices.

III.3.2.2. Moyenne harmonique : La moyenne harmonique se calcule dans le cadre de rapport, surtout vitesse ou distance moyenne.

On appelle moyenne harmonique de la distribution statistique et on note H la moyenne arithmétique des inverses de n valeurs observées, soit :

III.3.2.2.1.Moyenne harmonique simple :

H = $\frac{1}{\frac{1}{N}\left(\frac{1}{X_1}+\frac{1}{X_2}+ \ldots\ldots\frac{1}{X_K}\right)} = \frac{N}{\sum_{i=1}^{K}\frac{1}{Xi}}$

Exemple 14: un étudiant a consacré la même somme de 3600 DA pendant trois ans à l'achat de livre aux prix respectifs de 400 DA, 600 DA et 900 DA le livre.

Calculer la moyenne harmonique des différents prix des livres.

- L'étudiants a dépensé diront les trois ans 3x3600 = 10800 DA

Il a acheté :

$\frac{3600}{400}$ = 9 livres pendant la première année

$\frac{3600}{600}$ = 6 livres pendant la deuxième année

Et $\frac{3600}{900}$ = 4 livres pendant la troisième année

Il a donc acheté : $360\left[\frac{1}{400}+\frac{1}{600}+\frac{1}{900}\right]$ = 9+6+4 = 19 livres durant les trois années et le prix moyen d'un livre est donc :

H = $\frac{3x3600}{3600\left(\frac{1}{400}+\frac{1}{600}+\frac{1}{900}\right)} = \frac{3}{\left(\frac{1}{400}+\frac{1}{600}+\frac{1}{900}\right)}$= 568,42 ≅ 568 DA

$\frac{1}{H} = \frac{1}{10800}\left(\frac{3600}{400}+\frac{3600}{60}+\frac{3600}{900}\right) \rightarrow \frac{1}{H}$ = 0,00176 → H= 568,42

$$\frac{1}{H} = \frac{1}{n}\sum\frac{1}{\varkappa i}$$

III.3.2.2.2.Moyenne harmonique pondérée :

$$H = \frac{1}{\frac{1}{N}\left(\frac{n1}{\varkappa 1}+\frac{n2}{\varkappa 2}+\cdots\frac{nK}{\varkappa K}\right)} = \frac{N}{\sum_{i=1}^{K}\frac{ni}{\varkappa i}} = \frac{1}{\sum_{i=1}^{K}\frac{fi}{\varkappa i}}$$

$$\frac{1}{H} = \frac{1}{n}\sum\frac{ni}{\varkappa i}$$

Remarque :

1- Une moyenne harmonique ne peut être calculée que si toutes les valeurs observées sont non nulles.

2- En pratique, une moyenne harmonique est essentiellement utilisée lorsqu'une même somme est investie dans des biens de prix différents et que l'on souhaite déterminer le prix moyen des biens par exemple.

III.3.2.3.La moyenne quadratique : On utilise les moyennes quadratiques pour calculer des moyennes d'écart ou des moyennes d'amplitudes.

On appelle moyenne quadratique d'une distribution statistique et on note Q la racine carrée de la moyenne arithmétique des carres de N valeurs observées, soit :

III.3.2.3.1.Moyenne quadratique simple :

$$Q = \sqrt{\frac{\varkappa_1\ 2+\varkappa_2 2+\ \ +\varkappa K^2}{N}} = \sqrt{\frac{1}{N}\sum_{i=1}^{K}\varkappa i^2}$$

Exemple15 *:* quelle est la mesure du « côté moyenne » de trois plaques métalliques carrées dont les côtés mesurent 3 cm, 6 cm et 9 cm.

- Les superficies des plaques sont : 9cm^2, 36cm^2 et 81cm^2

$$\bar{X} = \frac{9 + 36 + 81}{3} = 42cm^2$$

Coté moyen : $\sqrt{42cm^2}$ = 6,48 cm

$$Q = \sqrt{\frac{3^2 + 6^2 + 9^2}{3}} = \sqrt{42} = 6{,}48\ cm$$

III.3.2.3.2. Moyenne quadratique pondérée

$$Q = \sqrt{\frac{n1\varkappa 1^2 + n2\varkappa 2^2 + \cdots..+ nK\varkappa K^2}{3}} = \sqrt{\frac{1}{N}\sum_{i=n}^{K} ni\varkappa i^2}$$

$$Q = \sqrt{\sum_{i=n}^{K} f_i \varkappa i^2}$$

Remarque : comparaison des moyennes

Soient $x_1, x_2, \ldots \ldots x_n$ les n observations relatives à une variable statistique x.

Lorsque les quatre moyennes (arithmétique, géométrique harmonique et quadratique) de la distribution peuvent être calculées, on a :

$$X_{min} \leq H \leq G \leq \bar{X} \leq Q \leq H_{max}$$

Exemple16: soit la répartition de 100 observations relatives à une variable statistique X

Calculer les quatre moyennes (arithmétique, géométrique, harmonique et quadratique)

Classes	**Effectifs** n_i	**Centre de classe** x_i	$n_i x_i$	$\frac{n_i}{x_i}$	$\log n_i x_i$	$x_i^2 n_i$
[10-20[	10	15	150	0,66	11,76	2250
[20-30[	30	25	750	1,2	41,93	18750
[30-40[	20	35	700	0,57	30,88	24500
[40-50[	40	45	1800	0,88	66,12	81000
Total	100		3400	3,31	150,69	126500

- $\bar{X} = \frac{\sum n_i x_i}{N} = \frac{3400}{100} = 34$

- $\log G = \frac{\sum_{i=1}^{4}(ni \log \varkappa i)}{\sum_{i=1}^{4} ni} = \frac{150,69}{100} = 1,5 \rightarrow G = 10^{1,5} = 31,62$

- $H = \frac{\sum_{i=1}^{4} ni}{\sum_{i=1}^{4}\left(\frac{ni}{Xi}\right)} = \frac{100}{3,31} = 30,21$

- $Q = \sqrt{\frac{\sum_{i=1}^{4}(ni\varkappa i^2)}{\sum_{i=1}^{4} ni}} = \sqrt{\frac{126500}{100}} = \sqrt{1265} = 35,56$

$$10 \leq 30,21 \leq 31,62 \leq 34 \leq 35,56 \leq 50$$

$$X_{min} \leq H \leq G \leq \bar{X} \leq Q \leq X_{max}$$

Chapitre IV : Les caractéristiques de dispersion

Introduction :

Très souvent les indicateurs de tendance centrale (mode, médiane et moyenne) s'avèrent insuffisants pour permettre de résumer à eux seuls et comparer deux ou plusieurs séries statistiques.

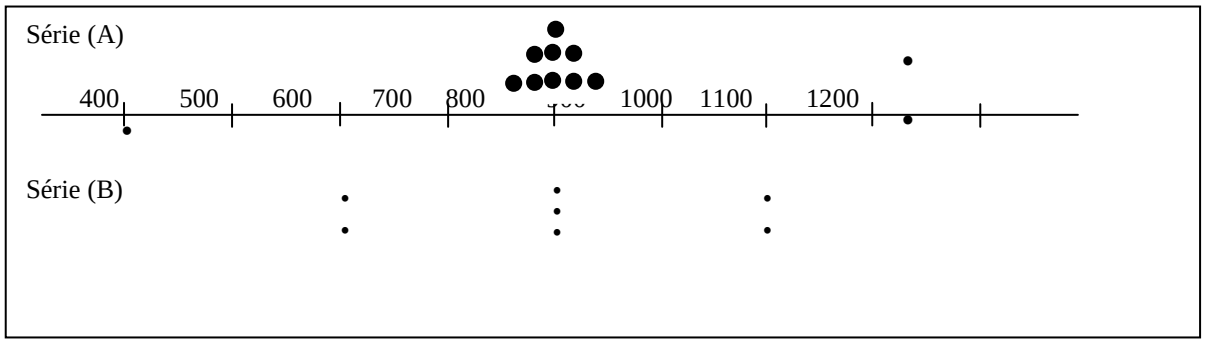
A titre indicatif la médiane est ce qu'on appelle une caractéristique de valeur centrale résumant la répartition d'une population selon un caractère, mais un tel résumé ne donne qu'une vision restreinte de cette répartition. Ainsi, on peut avoir deux répartitions très différentes, même si les caractéristiques de valeur centrale sont proches, l'une étant plus regroupée que l'autre. Pour préciser l'étude statistique, on va donc construire des indicateurs qui peuvent mesurer cet aspect des choses ; c'est ce qu'on appelle les indicateurs de dispersion. Les principaux sont l'étendue, les quantiles, le rapport inter-décile et l'écart-type.

Prenons à titre **d'exemple**, deux séries (A) et (B) de 9 lampes électriques, dont on a étudié la durée de vie en heures :

Série (A) : 780-790-790-800-800-800-810-810-820

Série (B) : 400-600-600-800-800-800-1000-1000-1200

Nous pouvons constater que les deux séries ont un même mode (Mo = 800), une même médiane (Me=800) et une même moyenne arithmétique ($\bar{x}$ = 800). Et pourtant si l'on représente les durées sur un axe, on constate que la deuxième série est plus dispersée que la première.



Pour la série (A), les durées de vie des lampes ne s'écartent pas trop des valeurs centrales (Mo= Me = $\bar{x}$ = 800). Ce qui n'est pas le cas pour la série (B).

D'où la nécessité de calculer d'autres indicateurs capables de rendre compte des écarts entre les différentes valeurs observées et la valeur centrale. Ces valeurs sont appelées indicateurs de dispersion, ils nous informent sur la variabilité des valeurs observées.

Définition : une caractéristique de dispersion est un indicateur qui permet d'estimer dans quelle mesure des observations s'écartent les unes des autres (ou s'écartent de leur valeur centrale généralement la moyenne).

Il existe plusieurs mesures de dispersion, le plus courantes sont l'étendue, l'écart type.

IV.1.Paramètre de dispersion absolue

IV.1.1.Paragraphe 1 : caractéristiques de dispersion n'utilisant pas de valeur centrale.

Les paramètres de dispersion (étendue, intervalle interquartile) sont calculés pour les variables statistiques quantitatives.

IV.1.1.1.L'étendue : c'est la différence entre la plus grande et la plus petite valeur de la série statistique.

$$E = Max(x_i) - Min(x_i)$$

Dans les séries (A) et (B) :

$E_A = 820 - 780 = 40$

$E_B = 1200 - 400 = 800$

L'étendue est simple et facile à calculer toutefois, il est très sensibles aux valeurs extrêmes qui peuvent être « aberrantes ».

IV.1.1.2.Les quantiles :

Comme pour la médiane ou l'on s'est intéressé à la valeur de la variable qui partage la population en deux parties d'effectifs d'égal effectif, on s'intéresse ici aux valeurs qui partagent la population en quatre, en dix ou en cent parties de même effectif.

Ces valeurs sont appelées respectivement quartiles الربيعيات déciles الأعشار et percentiles ou centiles المئينات

Remarque :

La détermination des différents quartiles se fait de la même manière que la médiane.

IV.1.1.2.1.Les quartiles : Les quartiles sont des valeurs du caractère qui partagent la série statistique ordonnée en 4 parties égales, ils sont au nombre de 3 que l'on désigne par Q1, Q2 et Q3

Le 1er quartile : Q1 est la valeur du caractère telle que 25% lui soient inférieure et 75% lui soient supérieures.

Rang du 1er quartile : $RQ1 = \sum ni \ \frac{1}{4} = \frac{\sum ni}{4} = \frac{N}{4}$

Le second quartile : Q2 correspond à la médiane

Le troisième quartile : Q3 est la valeur du caractère telle que 75% des observations lui soient inférieures et 25% des observations lui soient supérieures.

Rang du 3ème quartile : RQ3 = ∑ni $\frac{3}{4}$

L'intervalle interquartile ou l'écart interquartile noté IQ est égale à **IQ = Q_3 - Q_1**

Cet intervalle contient 50% des observations, en laissant 25% des observations de part et d'autre de l'intervalle et présente l'avantage d'éliminer l'influence des valeurs extrêmes.

<u>Exemple 1</u> :

Le tableau suivant donne la répartition de 100 salariés selon le salaire horaire en DA :

Classes DA	**Effectifs ni**	**Limites supérieures**	**Nci↑ ت م ص**
[140-150[	10	Moins de 150	10
[150-160[	26	Moins de160	36
[160-170[	40	Moins de170	76
[170-180[	16	Moins de180	92
[180-190[	8	Moins de190	100
Total	100		

Détermination de l'intervalle interquartile.

IQ = Q_3 - Q_1

Calcul de Q_1 : on suit 3ètapes

Rang de $Q_1 = \frac{\sum ni}{4} = \frac{100}{4} = 25$

Classe contenant Q_1 = [150-160[

<u>*Calcul de Q_1 :*</u>

Q_1 = 150 + $\left[\frac{25-10}{36-10}\right] x(160-150)$

→ Q_1 = 155,769 DA

Interprétation : 25% des salariés ont un salaire horaire inférieur à 155,769 DA et 75% des salariés ont un salaire horaire supérieur à 155,769 DA.

Calcul de Q_3 : RQ_3 = ∑ni $\frac{3}{4} = \frac{100x3}{4}$ = 75

Classe contenant Q_3 = [160-170[

Q_3 = 160 + $\left[\frac{75-36}{76-36}\right] x(170-160) = 169{,}75\ DA$

Interprétation : 75% des salariés ont un salaire horaire inférieur à 169,75 DA et 25% des salaires ont un salaire horaire supérieure 169,75 DA.

<u>*Interval interquartile*</u>*:*

IQ = Q_3 - Q_1

IQ = 169,75 – 155,769 = 13,981 DA

La comparaison de cette valeur à l'étendue nous permet de l'interpréter comme suit :

$\frac{Q_3-Q_1}{E} = \frac{13,981}{189-140} = \frac{13,981}{49} = 0,285\ ou\ 25,5\%$

Cela veut dire que 28,5% de l'étendue correspondant à 50% de la population c'est-à-dire les salariés dont les salaires sont compris entre 155,769 DA et 169,75 DA.

Remarque :

Parfois on calcul l'intervalle semi-interquartile défini par la formelle suivante :

I$Q_{/2}$ = $\frac{Q_3-Q_1}{2}$

Il est aussi connu par l'expression déviation-quartile

IV.1.1.2.2.Les déciles : sont des valeurs qui partagent la série statistique en 10 parties égales. Ils sont au nombre de 9 : D1, D2, …, D9

Les plus caractéristiques sont le 1er décile (D1) et le 9ème décile (D9).

D1 : est la valeur de caractère telle que 10% des observations lui soient inférieurs et 90% des observations lui soient supérieures.

RD1 = ∑ni $\frac{1}{10}$ (Rang du 1er décile)

D9 : est la valeur du caractère telle que 90% des observations lui soient inférieures et 90% des observations lui soient supérieures

RD9 = ∑ni $\frac{9}{10}$ (Rang du 9ème décile)

Intervalle inter-décile : $I_D = D_9 – D_1$

Cet intervalle contient 80% des observations

***IV.1.1.2.3.Les percentiles ou les centiles* :** Ils partagent la série ordonnée en 100 parties égales.

Ils sont au nombre de 99 :

Intervalle inter-percentiles : $Ip = P_{99} - P_1$

Il contient 98% des observations

Remarque :

La détermination des déciles, et des percentiles se fait de la même manière que la médiane et les quartiles à partir des effectifs (ou des fréquences cumulés).

IV.1.2.Paragraphe 2 : Indicateurs de dispersion par rapport à une valeur centrale (généralement la moyenne)

Exemple 2 :

Considérons les notes suivantes en statistique d'un groupe de 6 étudiants :

2-17-7-18-3-13

La moyenne des notes est $\bar{\varkappa} = \frac{\sum \varkappa i}{6} = \frac{60}{6} = 10$

Comment apprécier la dispersion de cette série ? كيف يتم تقييم التشتت

En calculant les différences entre les observations et la moyenne, on obtient :

$\varkappa i$	2	17	7	18	3	13	$\sum$
$(\varkappa i-\bar{\varkappa})$	-8	+7	-3	+8	-7	+3	0

Or la somme $\sum_{i=1}^{6}(\varkappa i - \bar{\varkappa})$ de ces différences est nulle

Une des propriétés de la moyenne arithmétique ; est la somme des écarts des effectifs à leur moyenne est nulle $\sum_{i=1}^{n}(\varkappa i - \bar{\varkappa}) = 0$.

Ce calcul se fait en valeur absolue pour que le résultat soit exploitable

IV.1.2.1. Ecart absolu moyen

Une autre idée consisterait à utiliser **les valeurs absolues** de ces différences soit $|\varkappa i - \bar{\varkappa}|$

On a ainsi établi la moyenne des valeurs absolues des écarts à la moyenne .cette caractéristique rend convenablement compte de la dispersion entre les deux séries

On pose donc : e = $\frac{1}{n}\sum_{i=1}^{n}|\varkappa i - \bar{\varkappa}|$ si les données ne sont pas groupées

e = $\frac{1}{n}\sum_{i=1}^{n} ni|\varkappa i - \bar{\varkappa}|$ si les données sont regroupées

Mais les valeurs absolues se prêtant mal aux calculs algébriques et la plupart des livres affirment que son maniement algébrique difficile en est la cause, c'est pour ca l'écart absolu est peu utilisé ; on préfère introduire les carrés des différences $(\varkappa i - \bar{\varkappa})$

Leur moyenne arithmétique s'appelle la variance التباين

IV.1.2.2.Variance notée V (x)

Données non-groupées :

$V(\varkappa) = \frac{\sum(\varkappa i-\bar{\varkappa})^2}{N} = \frac{1}{N}.\sum(\varkappa i - \bar{\varkappa})^2$

Données groupés :

$V(\varkappa) = \frac{\sum ni(\varkappa i-\bar{\varkappa})^2}{\sum ni} = \frac{1}{\sum ni}.\sum ni(\varkappa i - \bar{\varkappa})^2$

$V(\varkappa) = \sum fi.\varkappa i - \bar{\varkappa}$ avec les fréquences relatives

<u>Remarque :</u>

La variance est exprimée dans le carré de l'unité de la variable par exemple, la variance de la variable salaire est exprimée en dinars au carré (DA^2). C'est la raison pour la quelle on ne doit pas interpréter la variance, mais plutôt sa **racine carrée**.

IV.1.2.3.Ecart-type : σx

On appelle écart-type notéσx, la racine carrée de la variance. Il est utilisé comme un indicateur de la dispersion de la série statistique

$$\sigma x = \sqrt{V(\varkappa)}$$

Remarque :
L'écart-type est exprimé dans la même unité de mesure que la variable. Plus l'écart-type est grand, plus la dispersion des observations autour de la moyenne de la variable est forte.

Exemple 2 : **(suite)**

Calculer l'écart-type des notes de l'exemple précédent. Interpréter

$$\sigma x = \sqrt{\frac{1}{6}\sum(\varkappa i - \bar{\varkappa})^2} = \sqrt{\frac{1}{6}[(-8^2) + 7^2 + (-3)^2 + 8^2 + (-7)^2 + 3^2}$$

$$\sigma x = \sqrt{40{,}66} = 6{,}37$$

Interprétation :

Certains étudiants (les bons) auront approximativement la note moyenne (10) plus 6,37, les autres (les mauvais) auront la note moyenne (10) moins 6,37.

Exemple 3 :

Considérons les notes suivantes en statistique d'un deuxième groupe de 4 étudiants :
8-12-9-11

Calculer l'écart-type des notes et comparer le résultat obtenu avec le résultat de l'exemple.

$$\bar{\varkappa} = \frac{1}{4}\sum \varkappa i = \frac{40}{4} = 10$$

$$V(\varkappa)_2 = \frac{1}{4}\sum(\varkappa i - \bar{\varkappa})^2$$

$$V(\varkappa)_2 = \frac{1}{4}[(8-10)^2 + (12-10)^2 + (9-10)^2 + (11-10)^2]$$

$$V(\varkappa)_2 = 10/4 = 2{,}5 \qquad \sqrt{2{,}5} = 1{,}581$$

$$\rightarrow \sigma x_2 = \sqrt{V(\varkappa)_2} = \sqrt{10} = 3{,}16$$

Comparaison :

$$\sigma x_1 = 6{,}37 \quad \text{Et } \sigma x_2 = 3{,}16$$

La dispersion des notes dans l'exemple 2 est deux fois plus importante que celle de l'exemple 3. Le second groupe d'étudiant est un groupe plus homogène que le groupe 1.

Formules développées ou relation de Koenig de la variance et de l'écart-type (méthodes pratiques).

Dans les exemples précédents 2 et 3 les calculs de la variance et de l'écart-type ont été relativement aisés car la moyenne ($\bar{\varkappa}$ = 10) étant, comme les valeurs de la variable un nombre

entier, on a pu calculer les carrés $(\varkappa i - \bar{\varkappa})^2$. On conçoit que lorsqu'il n'en est pas ainsi, ce qui est le cas général, les calculs deviennent pénibles et on utilise de préférence une formule développée qui est plus pratique.

Formules développées (Koenig) :

Données non-groupées : $V(\varkappa) = \frac{\sum \varkappa i^2}{n} - \bar{\varkappa}^2$

Données groupées : $V(\varkappa) = \frac{\sum ni \varkappa i^2}{\sum ni} - \bar{\varkappa}^2$

Ou avec les fréquences relatives : $V(\varkappa) = \sum fi\ \varkappa i^2 - \bar{\varkappa}^2$

Remarque : Si $\sigma x = \sqrt{V(\varkappa)} \Rightarrow V(\varkappa) = \sigma x^2$

Exemple4 :

Soit les données suivantes : 6-9-10-11 et 15

Calculer l'écart-type par la formule de définition puis par la formule développée.

$\varkappa i$	$\varkappa i - \bar{\varkappa}$	$(\varkappa i - \bar{\varkappa})^2$	$\varkappa i^2$
6	-4,2	17,64	36
9	- 1,2	1,44	81
10	- 0,2	0,04	100
11	0,8	0,64	121
15	4,8	23,03	225
Total 51	0,00	42,8	563

$\bar{\varkappa} = \frac{\sum \varkappa i}{n} = \frac{51}{5} = 10{,}2$

Formule de définition :

$\sigma \varkappa^2 = \frac{1}{n} . \sum (\varkappa i - \bar{\varkappa})^2 = \frac{42{,}8}{5} = 8{,}56$

$\Rightarrow \sigma \varkappa = \sqrt{8{,}56} = 2{,}92$

Formule développée (KONIG)

$\sigma \varkappa^2 = \frac{\sum \varkappa i^2}{n} - \bar{\varkappa}^2 = \frac{563}{5} - 10{,}2^2 = 8{,}56$

$\Rightarrow \sigma \varkappa = \sqrt{8{,}56} = 2{,}92$ Même résultat

Exemple5 *:* soit le tableau suivant :

Classes	ni	$\varkappa i$	ni.$\varkappa i$	$\varkappa i - \bar{\varkappa}$	$(\varkappa i - \bar{\varkappa})^2$	Ni. $(\varkappa i - \bar{\varkappa})^2$	ni.$\varkappa i^2$
[10-20[	3	15	45	-10	100	300	675
[20-30[	4	25	100	0	0	0	2500
[30-40[	3	35	105	+10	100	300	3675
$\sum$	10	/	250	/	/	600	6850

$$\bar{\varkappa} = \frac{\sum ni.\varkappa i}{N} = \frac{250}{10} = 25$$

Formule de définition :

$$\sigma\varkappa^2 = \frac{\sum ni.(\varkappa i - \bar{\varkappa})}{\sum ni} = \frac{600}{10} = 60$$

$$\Rightarrow \sigma\varkappa = \sqrt{60} = 7{,}745$$

Formule développée (KONIG) :

$$\sigma\varkappa^2 = \frac{\sum ni.\varkappa i^2}{\sum ni} - \bar{\varkappa}^2 = \frac{6850}{10} - (25)^2 = 60$$

$\Rightarrow \sigma\varkappa = \sqrt{60} = 7{,}745$ Même résultat

IV.2.Les coefficients de dispersion relative :

L'écart-type s'exprimant dans la même unité de mesure que la variable, ce qui rend difficile la comparaison de deux ou plusieurs distributions exprimées dans des unités différentes :

C'est ainsi si l'on veut comparer les salaires des fonctionnaires algériens et ceux des fonctionnaires français, les salaires des algériens sont en dinars et ceux des français en euros.

Pour éliminer l'influence de l'unité de mesure, on calcul un coefficient de dispersion relatif sans dimension. Il en existe plusieurs sortes, citons :

IV.2.1. Le coefficient interquartile relatif :

$$CQ = \frac{Q_3 - Q_1}{Q_2}\ x\ 100$$

IV.2.2.Le coefficient de variation CV (utilisé le plus fréquemment dans la comparaison de 2 ou plusieurs séries) :

Remarque :

Pour comparer la distribution entre séries dont les éléments sont mesurés à partir d'unités différentes, ou dont l'ordre de grandeur n'est pas le même, on utilise des caractéristiques de coefficient de variation.

Parfois on explicité la comparaison de la moyenne et de l'écart-type en utilisant le C.V défini comme le rapport $\sigma x/\bar{\varkappa}$

Ce coefficient permet de la comparaison des distributions de nature similaire mais correspondant à des observations faites en des lieux et /on des dates différentes

C'est un coefficient sans dimension indépendants des unités de mesures, il sert à rendre les comparaisons entre des séries statistiques différentes plus aisées (plus logique).

Par exemple la comparaison de distribution de salaires exprimés dans deux manières différentes.

Exprime en pour cent, il dépend de choix des unités de mesure ex : on peut par exemple comparer l'écart des salaires entre le Maroc (dirham) et l'Algérie (DA), (le résultat est exprimé en pourcentage).

$\alpha = CV = \frac{\sigma x}{\bar{x}} x100$

Plus de C.V est élevé, plus la dispersion est forte.

Exemple si Le C.V dans l'entreprise A est de 40% et dans l'entreprise B est de35% on peut conclure que la dispersion relative est plus important de l'entreprise A que l'entreprise B

Si dans une même entreprise la dispersion relatives des salaires des ouvriers est de 0,41 alors que celles des employés est de 0,78 que peut on conclure.

On peut conclure que la répartition des salaires est dans plus homogène chez les ouvriers que chez les employés dans cette entreprise.

Exercice6:

On veut comparer les salaires dans deux pays Algérie et France. Nous avons calculé $\bar{x}$ et σx des salaires dans chaque pays et on a obtenu les résultats suivants :

Algérie : $\bar{x} = 25000\ DA$ avec σx = 35 DA

France : $\bar{x} = 7000\ euros$ avec σx = 25 euros

Quelle est la plus forte dispersion, celle des travailleurs algériens ou celle des travailleurs français ?

Solution :

Il est clair qu'on ne peut pas comparer des dinars et des euros. On utilise alors le coefficient de variation :

$CV_{Algérie} = \frac{\sigma x}{x}$x100 = $\frac{350}{25000} x100 = 1{,}4\%$

$CV_{France} = \frac{25}{7000} x100 = 0{,}357\%$

$VC_F < CV_{Algérie}$

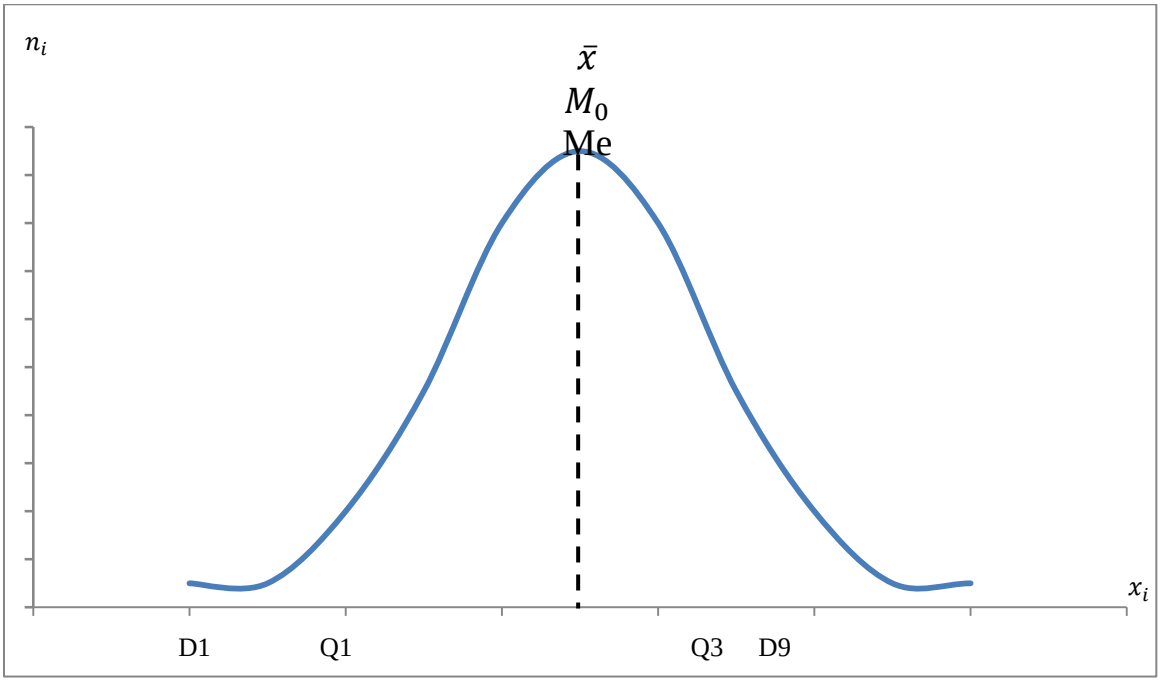
Nous pouvons en conclure qu'il y a d'avantage d'uniformité et d'homogénéité en ce qui concerne les salaires des français qu'en ce qui concerne ceux des algériens.

IV.4. Les caractéristiques de forme :

Ces caractéristiques permettent de préciser l'allure de la courbe des effectifs (ou des fréquences) sans la tracer.

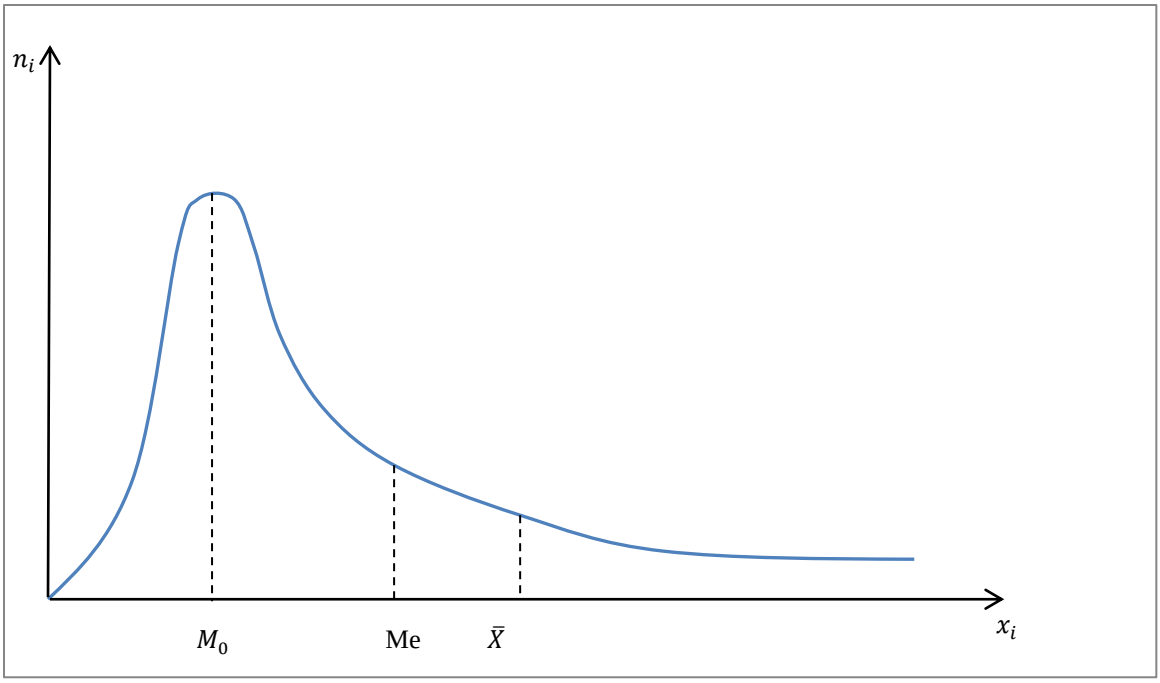
Une distribution est dite symétrique si les observations sont également dispersées de part et d'autre de la valeur centrale. Dans le cas contraire, la distribution est dite asymétrique ou dissymétrique.

Courbe symétrique $\bar{x}$ = M0 = Me

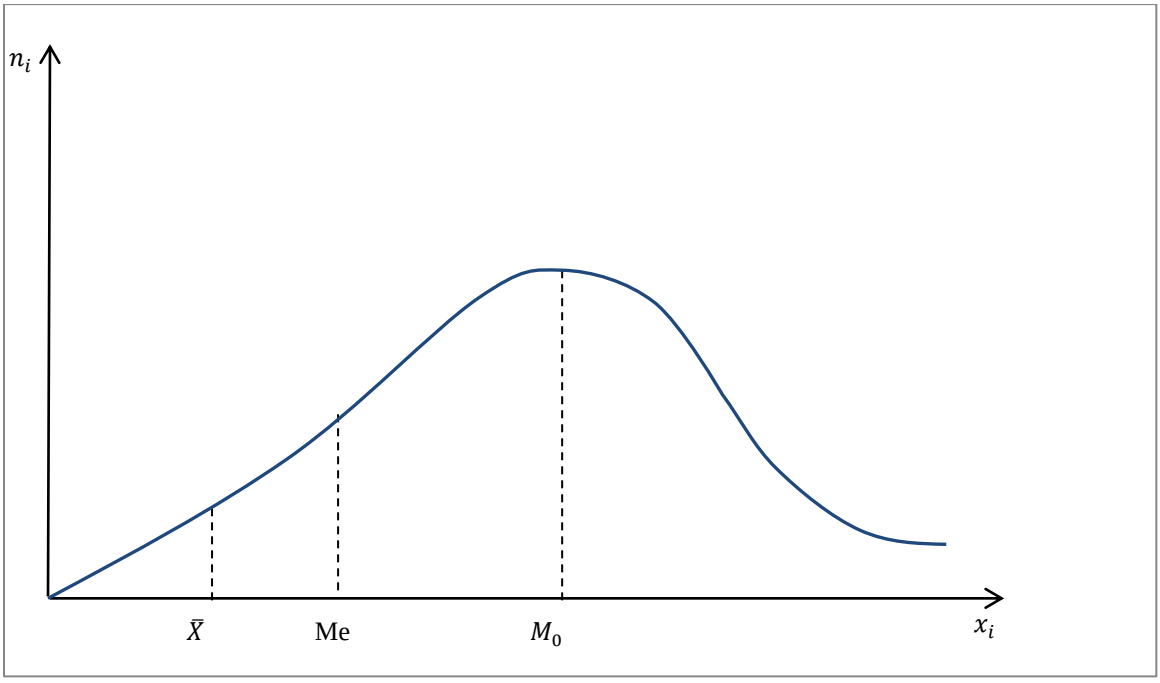


Dans une distribution symétrique, les trois valeurs typiques sont confondues. En outre les premiers et troisième quartiles, les premiers et neuvième déciles …, sont équidistants de la valeur centrale.

Courbe asymétrique ou oblique à droite (Etalement à droite) Mo < Me < $\bar{x}$



Courbe asymétrique ou oblique à gauche (Etalement à gauche) $\bar{x} <$ Me $<$ Mo



Pour caractériser l'asymétrique de la série on retient dans ce cours comme coefficient celui de Pearson, basé sur la moyenne, le mode et l'écart-type :

Si $cp_1 = \frac{\bar{x} - M0}{\sigma x}$ avec $-1 < cp_1 < +1$

Si $cp_1 = 0 \Longrightarrow$ courbe symétrique

si $cp_1 > 0 \Longrightarrow$ courbe oblique à droite ($\bar{x} >$ M0)

si $cp_1 < 0 \Longrightarrow$ courbe oblique à gauche ($\bar{x} <$ M0)

Remarque :

Pour les distributions unimodales légèrement asymétriques (peu asymétriques), les 3 valeurs de tendance centrale se trouvent liées par la relation empirique de Pearson :

$\bar{x} - M0 = 3(\bar{x} - Me)$

Cette relation permet de calculer une des valeurs à partir de la connaissance des deux autres.

Cette relation : $\bar{x}$ - M0 = 3 ($\bar{x}$ - Me) permet de calculer un second coefficient d'asymétrie de Pearson :

$Cp_2 = \frac{3.(\bar{x} - M0)}{\sigma x}$ $\quad -1 < cp_2 < +1$

si $cp_2 = 0 \Longrightarrow$ courbe symétrique

si $cp_2 > 0 \Longrightarrow$ Etalement (oblicité) à droite

si $cp_2 < 0 \Longrightarrow$ Etalement (oblicité) à gauche

Exemple 7 :

xi	2	3	4	5	6	7	8	∑
ni	1	6	15	20	15	6	1	64

Calculer le coefficient d'asymétrie de Pearson algébriquement et graphiquement.

Avec interprétation du résultat.

Solution :

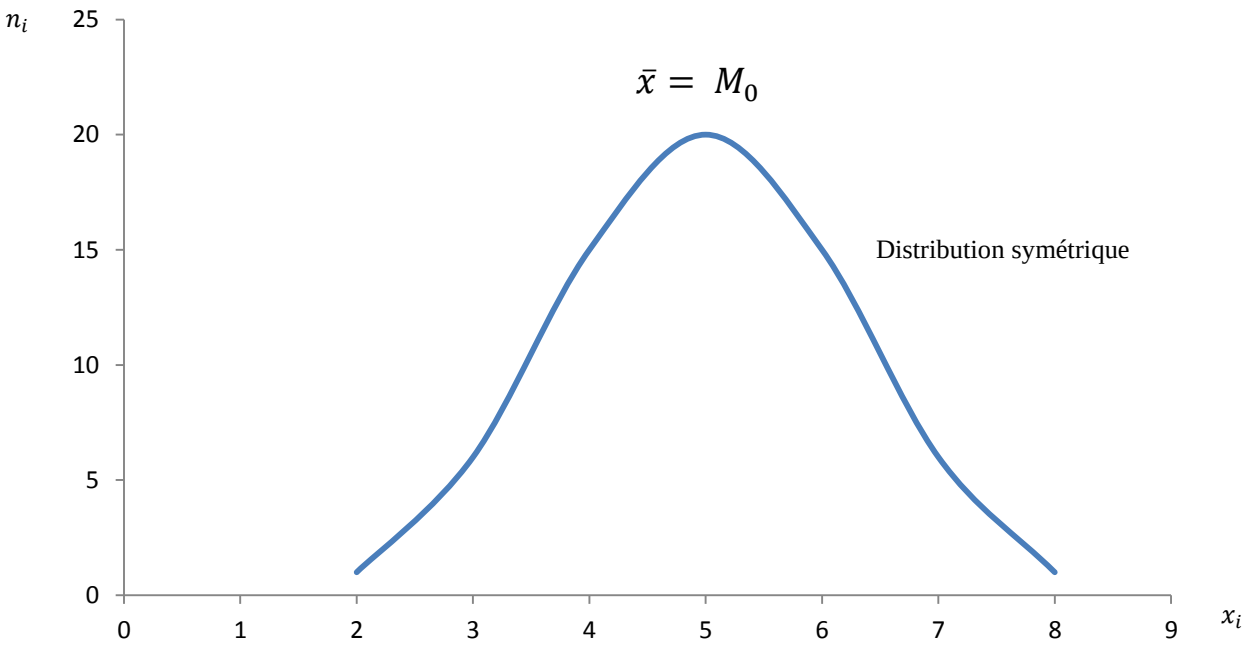
$cp_1 = \frac{\bar{x} - M0}{\sigma x}$

$\bar{x} = \frac{\sum ni.xi}{\sum ni} = \frac{320}{64} = 5$

M0 = 5

$\Rightarrow cp_1 = \frac{5-5}{\sigma x} = 0$

Graphiquement :



On joint les points de coordonnées ni et ϰi Puis on représente dans le graphe le mode et la moyenne.

Exemple 8 :

Classes	10-20	20-30	30-40	40-50	50-60	60-70	70-80
n_i	11	19	21	30	10	5	4

Travail à faire : Même question que l'exemple7.

$\sum ni = 100$ $\sum ni.\varkappa i = 3900$

$\sum ni.\varkappa i^2 = 174700$

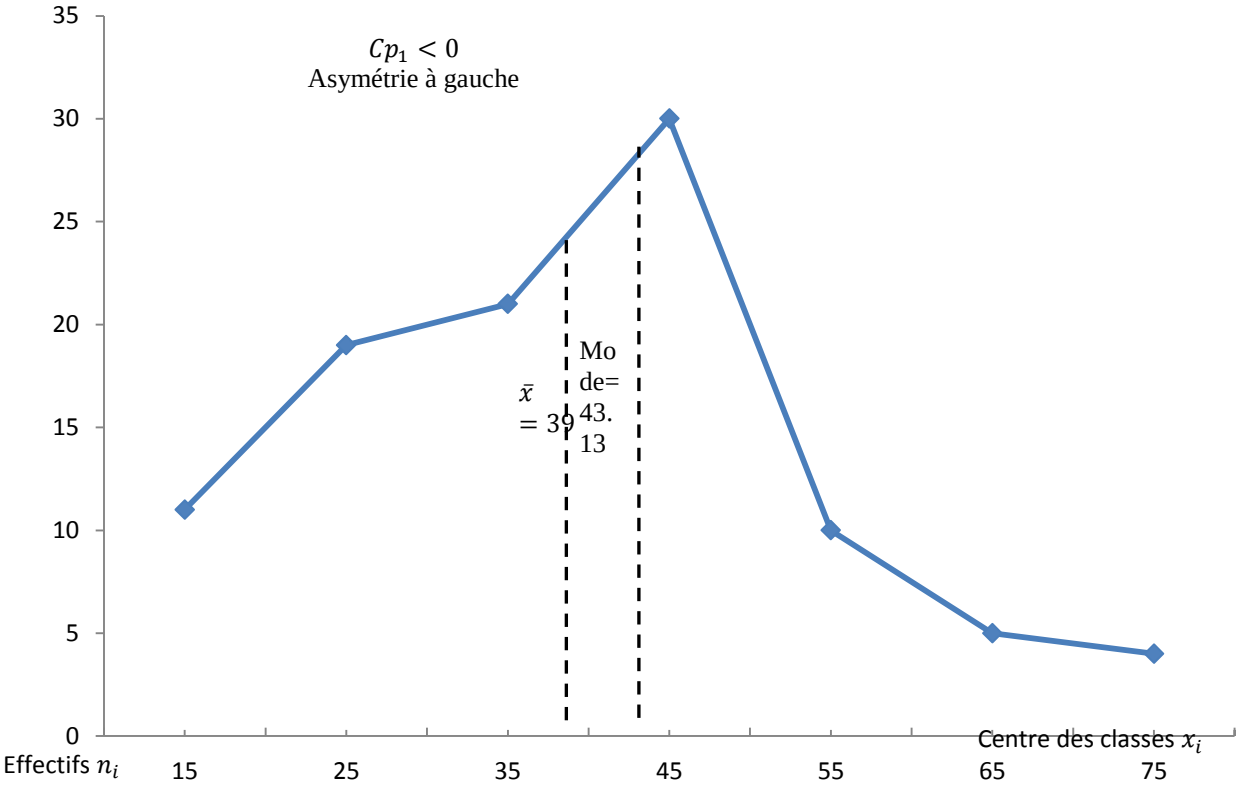
Après calcul on trouve :

$\bar{\varkappa} = 39$ Mo = 43,103 $\sigma x = 15{,}033$

$cp_1 = \frac{39-43{,}103}{15{,}033} = -0{,}272 < 0$

$Cp_1 < 0$ Oblique à gauche => Asymétrie a gauche

Graphiquement



Chapitre V : Les distributions statistiques à deux dimensions :

Dans les chapitres précédents on a étudié les distributions statistiques à un seul caractère (nous avons présenté les méthodes qui permettent de résumer et représenter les informations relatives à une variable). Cependant une population statistique peut être décrite selon un, deux ou plusieurs caractères(ou variables). Dans ce chapitre nous nous limiterons aux distributions à deux caractères ou deux variables ($\varkappa$, y).

On étudiera par exemple une population d'ouvriers selon deux caractères : l'âge ($\varkappa$) et les salaires (y) ou le salaire ($\varkappa$) et le nombre d'enfants (y) ou bien l'âge ($\varkappa$) et la qualification (y), les salaries en regardant leur ancienneté et leur niveau d'étude

Ce chapitre est consacré à l'étude des outils de base permettant d'étudier la liaison existante entre les deux caractères ainsi que le degré de l'intensité de cette relation, on a un caractère indépendante et une autre variable qui est dépendante.

V.1. La régression simple:

La régression est un des méthodes les plus connues et les plus appliquées en statistique pour l'analyse de données quantitatives. Elle est utilisée pour établir une liaison entre une variable quantitative et une ou plusieurs autres variables quantitatives, sous la forme d'un modèle. Si on s'intéresse à la relation entre deux variables, on parlera de régression simple en exprimant une variable en fonction de l'autre. Si la relation porte entre une variable et plusieurs autres variables, on parlera de régression multiple. La mise en œuvre d'une régression impose l'existence d'une relation de cause à effet entre les variables prises en compte dans le modèle.

Il est fréquent de s'interroger sur la relation qui peut exister entre deux grandeurs en particulier dans les problèmes de prévision et d'estimation.

Trois types de problèmes peuvent apparaître:

1. On dispose d'un certain nombre de points expérimentaux (xi, yi) où xi et yi sont les valeurs prises par les grandeurs x et y et on essaye de déterminer une relation fonctionnelle entre ces deux grandeurs x et y. Cette relation, pour des raisons théoriques ou pratiques s'écrira y = f(x, a,b,c...) et le problème sera d'ajuster au mieux les paramètres a,b,c.... pour que la courbe représentative de f passe au plus près des points (xi ,yi). Il s'agit d'un problème d'ajustement analytique.
2. On essaye de déterminer la relation statistique qui existe entre les deux grandeurs X et Y. Ce type d'analyse s'appelle analyse de régression. On considère que la variation de l'une des deux variables (par exemple X) explique celle de l'autre (par exemple Y).

Chaque domaine d'application a baptisé de noms différents ces deux variables : On trouve ainsi :

X	Y
Variable explicative	Variable expliquée
Variable contrôlée	Réponse
Variable indépendante	Variable dépendante
Régresser	....

Dans ce type d'analyse, on fixe *a priori* les valeurs de X. X n'est donc pas une variable aléatoire. Mais la deuxième grandeur Y, elle, est une variable aléatoire et sa distribution est influencée par la valeur de X. On a alors du point de vue statistique une relation de cause à effet. Le problème sera d'identifier cette relation.

3- Les deux grandeurs X et Y sont aléatoires et on cherche à savoir si leurs variations sont liées. Il n'y a pas ici de variable explicative ni de variable expliquée. Les variables peuvent avoir des causes communes de variation, parmi d'autres, qui expliquent leur relation d'un point de vue statistique : on est en présence d'un problème de corrélation. On cherche alors à mesurer le degré d'association entre les variables.

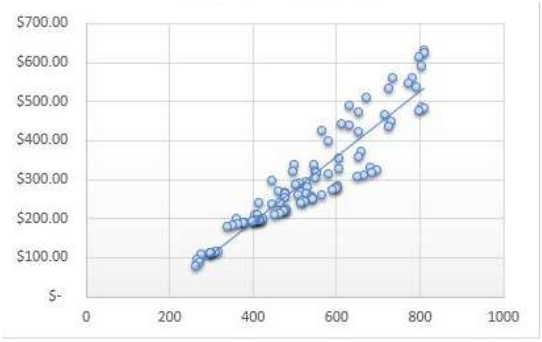
Exemple : poids et taille d'un individu, résultats obtenus à deux examens par des étudiants...

Régression linéaire simple :

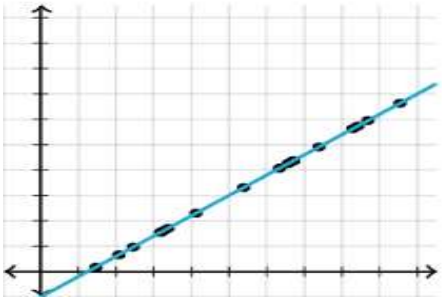
C'est l'étude de la relation qui existe entre deux variables quantitatives, en étudiant les variations de l'une en fonction de l'autre.

Pour étudier la relation qui existe entre les deux variables ; on doit définir la variable dépendante et cette qui est indépendante c.-à-d. celle qui a le plus d'influence. Pour étudier la liaison entre deux variables quantitatives (discrètes), on commence par faire un graphique du type nuage de points, Le plus souvent on utilisera la représentation cartésienne des couples (xi, yi) par un ensemble de n points, appelé nuage de points (**scatter diagram**), chaque point correspondant à une ligne du tableau de données, du types de celui qui figure ci-dessous. La forme générale de ce graphique indique s'il existe ou non une liaison entre les deux variables.

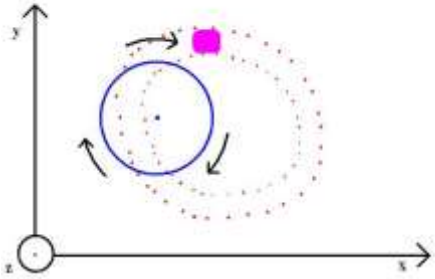
Pour déterminer le type de relation existante entre les deux caractères, on rencontre plusieurs situations :



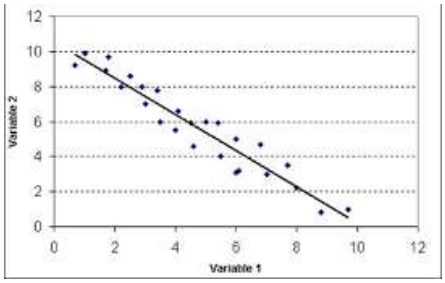
Relation Linéaire Positive



Relation Linéaire Parfaite



Relation Nulle



Relation Linéaire Négative

Dans ces diagrammes, chacun des individus à l'étude est représenté par un point dont les coordonnées ($\varkappa$, y) représentant les valeurs des deux variables. La droite tracée est celle rejoignant le maximum de ces points est appelée droite de régression.

Remarque : Dans le cadre de ce chapitre, seules les relations du premier degré c'est-à-dire linéaires sont considérées.

On doit écrire l'équation mathématique qui va nous permettre de trouver la droite de régression linéaire cette droite est de la forme : $\gamma i = a\varkappa i + b$ Elle nous permet de connaitre la valeur de la variable dépendante (y) grâce à la valeur de la variable indépendante ($\varkappa$).

L'idée est de transformer un nuage de point en une droite. Celle-ci doit être la plus proche possible de chacun des points. On cherchera donc à minimiser les écarts entre les points et la droite.

Pour cela, on utilise la méthode des moindres carrées. Cette méthode vise à expliquer un nuage de points par une droite qui lie *Y* à *X*, c'est à dire, $Y = aX + b$,

Telle que la distance entre le nuage de points et droite soit minimale. Cette distance matérialise l'erreur, c'est à dire la différence entre le point réellement observé et le point prédit par la droite. Si la droite passe au milieu des points, cette erreur sera alternativement positive et négative, la somme des erreurs étant par définition nulle. Ainsi, la méthode des

moindres carrés consiste à chercher la valeur des paramètres *a* et *b* qui minimise la somme des erreurs élevées au carré

D : y/ϰ c'est la droite de régression $\acute{Y}i$ = **a**ϰi + **b** où **a** représente la pente de la droite ou le coefficient de la droite de régression et b représente une constante.

V.1.1.Le modèle de régression linéaire simple :

Soit un échantillon de ***n*** individus. Pour un individu ***i*** (*i* = 1, ...*n*), on a observé :

- $\mathbf{x_i}$ la valeur de la variable quantitative **x**. (variable explicative ou indépendante).
- $\mathbf{y_i}$ la valeur de la variable quantitative y (variable réponse ou dépendante).

On veut étudier la relation entre ces deux variables, et en particulier, l'effet de **X** (variable indépendante) sur **Y** (variable dépendante). Dans un premier temps, on peut représenter graphiquement cette relation en traçant le nuage des n points de coordonnées (x_i, y_i). Dans le cas où le nuage de points est de forme "linéaire", on cherchera à ajuster ce nuage de points par une droite.

La relation entre $\mathbf{y_i}$ et $\mathbf{x_i}$ s'écrit alors sous la forme d'un modèle de régression linéaire simple (la droite de régression des moindre carrés) *:* $\boldsymbol{Y' = ax + b}$

- Le coefficient ***a*** est appelé la **pente**. C'est le changement sur **y** lorsque **x** change d'une unité.
- Le coefficient ***b*** est appelée l'ordonnée à l'origine (constante). C'est la valeur prédite de **y** quand **x = 0**.

Mais quelle méthode utiliser pour déterminer au mieux les paramètres(Les coefficients a et b) du modèle ?

V.1.2.MÉTHODE DES MOINDRES CARRES

Pour trouver les valeurs a et b on utilise la méthode de l'ajustement linéaire appelée **« méthode des moindres carrés ordinaires (MCO) ».**

- $\acute{Y}i$ = la valeur de y située sur la droite

-b = l'ordonnée à l'origine (la valeur de $\acute{Y}i \ \ quand \ \ \varkappa = 0$)

La méthode MCO consiste à retenir la droite qui va minimiser la somme des carrés des écarts entre les points situés sur le nuage et les points situés sur la droite.

Notons ℓi l'écart → $\ell i = (\gamma i - \gamma i)$

$\Longrightarrow \sum \ell i^2 = \sum(\gamma i - \gamma i)^2 =$ Min

Remplaçons γi par sa valeur

$\sum \ell i^2 = \sum(\gamma i - (axi + b))^2 = \sum(\gamma i^2 + (axi + b))^2 - 2\gamma(axi + b)$

$\sum \ell i^2 = \sum(\gamma i^2 + a^2 \varkappa i^2 + b^2 + 2ab\varkappa i - 2a\varkappa i\gamma i - 2b\gamma i)$

La condition nécessaire pour atteindre le minimum d'une fonction est d'annuler ses dérivées partielles

La méthode des moindres carrées consiste donc à minimiser la fonction U (la somme des erreurs commises). Nous avons la condition de minimisation suivante,

$\frac{\Delta \sum \ell i^2}{\Delta a} = 0$ Et $\frac{\sum \Delta \ell i^2}{\Delta b} = 0$

$\frac{\sum \Delta \ell i^2}{\Delta a} = \sum(2ax i^2 + 2b\varkappa i - 2\varkappa i\gamma i) = 0$

$\Longrightarrow 2a\sum \varkappa i^2 + 2b\sum \varkappa i - 2\sum \varkappa i\gamma i = 0$

$\Leftrightarrow 2(a\sum \varkappa i^2 + b\sum \varkappa i - \sum \varkappa i\gamma i = 0$

$\Leftrightarrow a\sum \varkappa i^2 + b\sum \varkappa i - \sum \varkappa i\gamma i = 0$

$\Leftrightarrow \sum \varkappa i\gamma i = a\sum \varkappa i^2 + b\sum \varkappa i \quad (1)$

$\frac{\Delta \ell i^2}{\Delta b} = 0 \Rightarrow \sum(2b + 2a\varkappa i - 2\gamma i) = 0$

$\Leftrightarrow n2b + 2a\sum \varkappa i - 2\sum \gamma i = 0$

$\Leftrightarrow 2(nb + a\sum \varkappa i - \sum \gamma i = 0$

$\Leftrightarrow nb + a\sum \varkappa i - \sum \gamma i = 0$

$\Longrightarrow \sum \gamma i = nb + a\sum \varkappa i \qquad (2)$

De (2) $\Longrightarrow b = \frac{\sum \gamma i - a\sum \varkappa i}{n} \Longrightarrow b = \frac{\sum \gamma i}{n} - \frac{a\sum \varkappa i}{n}$

$\Longrightarrow b = \bar{\gamma} - a\,\bar{\varkappa}$

Pour trouver la valeur de a remplaçons b par sa valeur dans①

$\sum \varkappa i\gamma i = a\sum \varkappa i^2 + (\bar{\gamma} - a\bar{\varkappa})\sum \varkappa i$

$\Leftrightarrow \sum \varkappa i\gamma i = a\sum \varkappa i^2 + \bar{\gamma}\sum \varkappa i - a\bar{\varkappa}\sum \varkappa i$

$\Leftrightarrow \sum \varkappa i\gamma i - \bar{\gamma}\sum \varkappa i = a(\sum \varkappa i^2 - \bar{\varkappa}\sum \varkappa i)$

$\Longrightarrow \mathbf{a} = \frac{\sum \varkappa i\gamma i - \bar{\gamma}\sum \varkappa i}{\sum \varkappa i^2 - \bar{\varkappa}\sum \varkappa i}$ si on divise le numérateur et le dénominateur par n :

On obtient **: a** $= \frac{\frac{\sum \varkappa i\gamma i}{n} - \overline{\varkappa\,\gamma}}{\frac{\sum \varkappa i^2}{n} - \bar{\varkappa}^2} = \frac{cov(x,y)}{S^2 x}$

Remarque : La droite d'ajustement passe par le point moyen $(\bar{\varkappa}, \bar{y})$

Dans le cas ou la variable y est indépendante et la variable $\varkappa$ dépendante ou obtient la droite de régression $\varkappa$ en Y $\Longrightarrow$ **D : $\varkappa$/y : $\varkappa i = \acute{a}\gamma i + \acute{b}$**

$\acute{a} = \frac{COV(\varkappa,\gamma)}{S^2\gamma} \Longrightarrow \frac{\frac{\sum \varkappa i\gamma i}{n} - \overline{\varkappa\,\gamma}}{\frac{\sum \gamma i^2}{n} - \bar{\gamma}^2}$, $\qquad \acute{b} = \bar{\varkappa} - \acute{a}\acute{\gamma}$

Exemple 1: soit le tableau suivant. Trouver les droits de régression Y/$\varkappa$ et $\varkappa$/Y, estimez la valeur de la consommation si le revenu est égale à 20UM et estimez le revenu quand la consommation est de 35UM.

$\varkappa$i(revenu)	γi (consommation)	$\varkappa i^2$	γi^2	$\varkappa i\, \gamma i$
5	4	25	16	20
6	5	36	25	30
7	6	49	36	42
9	7	81	49	63
12	8	144	64	96
15	12	225	144	180
54	42	560	334	431

1) Droite de régression y/x

D: Y/$\varkappa$: Yi = axi +b

$a = \frac{COV(\varkappa,\gamma)}{S^2\gamma}$

$COV(\varkappa,\gamma) = \frac{\sum \varkappa i \gamma i}{n} - \overline{\varkappa\,\gamma}$

$S^2\varkappa = \frac{\sum \varkappa i^2}{n} - \bar{\varkappa}^2$

$\bar{\varkappa} = \frac{\sum \varkappa i}{n} = \frac{54}{6} \Longrightarrow \bar{\varkappa} = 9$

$\bar{\gamma} = \frac{\sum \gamma i}{n} = \frac{42}{6} \Longrightarrow \bar{\gamma} = 7$

$COV(\varkappa,\gamma) = \frac{431}{6} - 9^2 \text{x} 7 = 8{,}833 \implies COV(\varkappa,\gamma) = 8{,}833$

$S^2\varkappa = \frac{560}{6} - 9^2 = 12{,}333 \implies S^2\varkappa = 12{,}333$

$a = \frac{8{,}833}{12{,}333} \implies a = 0{,}716$

$b = \bar{\gamma} - a\bar{\varkappa} \implies b = 7 - 0{,}716 \text{x} 9 \implies b = 0{,}556$

$\gamma i = 0{,}716\, \varkappa i + 0{,}556$ la droite de régression

Si $\varkappa = 20 \implies \gamma = 0{,}716 \text{ x } 0{,}556 \implies \gamma = 14{,}876$

2) Droite de régression $\varkappa/\gamma$

D: $\varkappa/\gamma$: $\varkappa i = \grave{a}\gamma i + \grave{b}$

$\grave{a} = \frac{COV(\varkappa,\gamma)}{S2y}$ / $COV(\varkappa,\gamma) = \frac{\sum \varkappa i \gamma i}{n} - \frac{\overline{\varkappa\,\gamma}}{S^2}$ / $S^2\gamma = \frac{\sum \gamma i^2}{n} - \bar{\gamma}$

$COV(\varkappa,\gamma) = \frac{431}{6} - (9 * 7) \implies COV(\varkappa,\gamma) = 8{,}833$

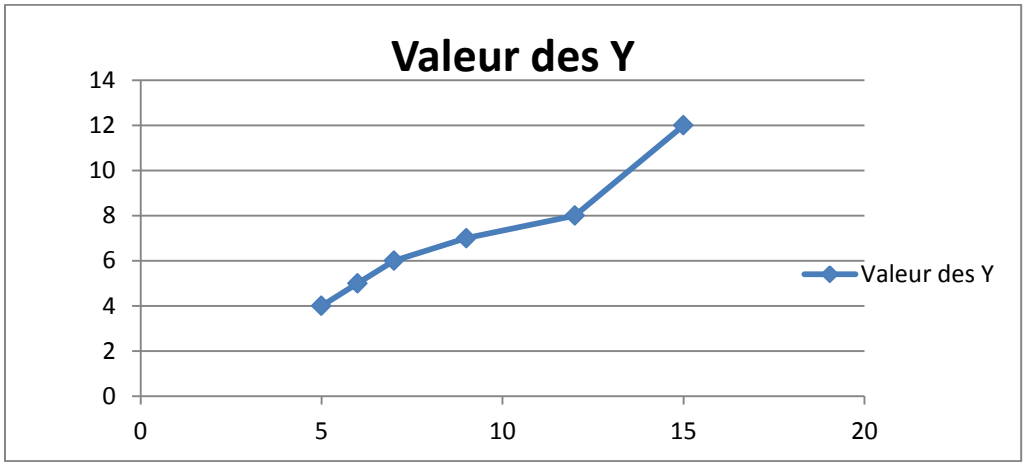
$S^2\gamma = \frac{334}{6} - 7^2 \Rightarrow S^2\gamma = 6{,}666$

$\grave{a} = \frac{8,833}{6,666} \implies \grave{a} = 1,325$

$\grave{b} = \bar{\varkappa} - \grave{a}\bar{\gamma} \implies \grave{b} = 9 - (1,325 * 7) \implies \grave{b} = -0,275$

La droite de régression $\varkappa\grave{\imath}$ = 1,325$\gamma\acute{\imath}$ - 0,275

Si γ=35 $\implies$ $\varkappa$=1,325x35- 0,275 $\implies$ $\varkappa$ = 46,1



V.2. La corrélation linéaire :

Dans le premier paragraphe on a pu par la méthode des moindres carrés déterminer les constantes de l'équation de régression et voir quelle est la nature de la relation entre les deux caractères. Dans ce deuxième paragraphe on va mesurer l'intensité de la relation entre les deux variables en utilisant le coefficient de corrélation noté (r) qui nous informera sur le degré de dépendance.

$\Gamma = \sqrt{a.\grave{a}} \iff \begin{cases} \gamma i = axi + b \\ \varkappa\acute{\imath} = \grave{a}\gamma i + \grave{b} \end{cases}$

$\Gamma = \sqrt{\frac{COV(\varkappa,\gamma)}{S2\varkappa} . \frac{COV(\varkappa,\gamma)}{S2y}} \rightarrow \boldsymbol{\Gamma = \frac{COV(\varkappa,\gamma)}{S\varkappa . S\gamma}}$

V.2.1.Qu'est-ce que la covariance :

La corrélation est une quantification de la relation linéaire entre des variables continues. Le calcul du coefficient de corrélation de Pearson repose sur le calcul de la covariance entre deux variables continues. Le coefficient de corrélation est en fait la standardisation de la covariance. Cette standardisation permet d'obtenir une valeur qui variera toujours entre -1 et +1, peu importe l'échelle de mesure des variables mises en relation.

La covariance est une mesure de l'association ou du lien qui existe entre deux variables. Pour comprendre la covariance, revenons à la notion de variance. La variance d'une

variable est une mesure qui quantifie la dispersion moyenne des valeurs prises par cette variable autour de sa moyenne.

Deux variables covariants ensemble lorsqu'un écart à la moyenne d'une variable est accompagné par un écart dans le même sens ou dans le sens opposé de l'autre pour le même sujet. Plus ce pattern est présent dans l'ensemble des observations, plus les deux variables semblent partager une association entre elles. Autrement dit, deux variables covariants lorsque la variation d'une des variables autour de sa moyenne semble influencer la manière dont l'autre variable varie autour de sa moyenne. La covariance exprime donc une quantité de variance partagée entre deux variables. En effet, tout comme la variance, la covariance peut se quantifier. Plus la valeur de la covariance est élevée, plus les deux variables partagent une portion importante de variance.

Voici la formule permettant de calculer la covariance entre deux variables continues :

$$Cov(X,Y) = \frac{\sum[(x_i - \bar{X})(y_i - \bar{Y})]}{N} = \quad Cov(X,Y) = \frac{\sum(x_i y_i)}{N} - (\overline{XY})$$

Remarque :

Le coefficient de corrélation est toujours compris dans l'intervalle [-1-1]

Quand r > 0 la relation est positive

Toute variation de $\varkappa$ induit une variation de Y dans le même sens.

-si r = 1 : la relation est positive très forte ou parfaite

-si r = 0,5 : la relation est position moyenne

-si r =∈] 0 ; +0,5[la relation est positive mais faible

-si r =∈] 0,5 ; +1[la relation est position relativement forte

Quand r < 0 la relation est négative : toute variation de $\varkappa$ implique une variation de Y dans le sens opposé

-si r = -1 : la relation est négative très forte ou parfaite ;

-si r = -0,5 : la relation est négative moyenne ;

-si r =∈] -0,5 ; 0[la relation est négative faible ;

-si r =∈] -1 ; -0,5[la relation est négative relativement forte.

Quand r est égale à zéro : Il n'y a aucune relation entre les deux variables, elles sont totalement indépendantes l'une de l'autre.

V.2.2. Le coefficient R^2 (coefficient de détermination):

Le coefficient R^2 est défini comme le carré du coefficient de corrélation de **x** et **y** est une mesure de qualité de l'ajustement.

Le $\mathbf{R^2}$ (r^2 %) est le pourcentage de variation totale de **Y** s'expliquant par la liaison de Y par rapport X. il s'agit du coefficient de détermination.

Exemple 2: Le tableau suivant représente le chiffre d'affaire et le bénéfice annuels de 5 entreprises en millions de dinars.

- calculer le coefficient de corrélation et interpréter le résultat.
- Trouvez les droits de régression $\varkappa$/Y et Y/$\varkappa$

$\varkappa$i(CA)	γi(benifices)	$\varkappa$i γi	γi^2	$\varkappa$i^2
10	1	10	1	100
20	3	60	9	400
30	2	60	4	900
40	5	200	25	1600
50	4	200	16	2500
150	15	530	55	

$\bar{\varkappa} = \frac{\sum \varkappa i}{n} = \frac{150}{5} \implies \bar{\varkappa} = 30 \quad \bar{\gamma} = \frac{\sum \gamma i}{n} = \frac{15}{5} \implies \bar{\gamma} = 3$

$$\boldsymbol{\Gamma(\varkappa,\gamma) = \frac{COV(\varkappa,\gamma)}{S\varkappa . S\gamma}}$$

$COV(\varkappa,\gamma) = \frac{\sum \varkappa i \gamma i}{n} - \bar{\varkappa}\bar{\gamma} \implies COV(\varkappa,\gamma) = \frac{530}{6} - 35x3 = 16$

$S_\varkappa = \sqrt{\frac{\sum \varkappa i^2}{n} - \bar{\varkappa}^2} = \sqrt{\frac{5500}{5} - 30^2} = \sqrt{200} \implies S_\varkappa = 14{,}142$

$S_\gamma = \sqrt{\frac{\sum \gamma i^2}{n} - \bar{\gamma}^2} = \sqrt{\frac{55}{5} - 3^2} = \sqrt{2} \implies S_\gamma = 1{,}414$

$\Gamma(\varkappa,\gamma) = \frac{16}{14{,}142 * 1{,}414} \implies \Gamma(\varkappa,\gamma) = 0{,}8$

La relation entre le chiffre d'affaire de le bénéfice est positive forte

Droite de régression Y/$\varkappa$ D : Y/$\varkappa \implies$ Yi = a$\varkappa$i + b

$a = \frac{COV(\varkappa,\gamma)}{S^2\varkappa} = \frac{16}{200} \implies$ a = 0,08

$b = \bar{\gamma} - a\bar{\varkappa}$ = 3-(0,08x30) $\implies$ b=0,6

$\underline{\boldsymbol{\gamma i = 0{,}08\ \varkappa i + 0{,}6}}$

Droite de régression $\varkappa$ /Y

D : $\varkappa$ /Y $\implies \varkappa i = \grave{a}\gamma i + \grave{b}$

$\grave{a} = \frac{COV(\grave{\varkappa},\gamma)}{S^2\varkappa} = \frac{16}{2} \implies \grave{a} = 8 \quad ; \quad \grave{b} = \bar{\varkappa} - \grave{a}\bar{\gamma}$ = 30-(8x3) $\implies \grave{b} = 6$

$$\underline{\boldsymbol{\varkappa i = 8\,\gamma i + 6}}$$

Exemple 3: le tableau ci-dessous donne l'évolution du prix des actions et des obligations.

- On demande s'il existe une corrélation entre l'évolution du prix des actions et l'évolution du prix des obligations
- Donner la droite de régression de **X** et **Y**.

Années	Actions X	Obligations Y
2001	352	1024
2002	360	998
2003	358	980
2004	361	970
2005	366	982
2006	382	972
2007	398	935
2008	406	902
2009	450	895
2010	445	900
∑	3878	9558

Solution:

Années	Actions X	Obligations Y	(xi -$\bar{X}$)	(xi -$\bar{X}$) ^{2}	(yi -$\bar{Y}$)	(yi -$\bar{Y}$) ^{2}	(xi yi)
2001	352	1024	-35,8	1281,64	68,2	4651,24	-2441,56
2002	360	998	-27,8	772,84	42,2	1780,84	-1173,16
2003	358	980	-29,8	888,04	24,2	585,64	-721,16
2004	361	970	-26,8	718,24	14,2	201,64	-380,56
2005	366	982	-21,8	475,24	26,2	686,44	-571,16
2006	382	972	-5,8	33,64	16,2	262,44	-93,96
2007	398	935	10,2	104,04	-20,8	432,64	-212,16
2008	406	902	18,2	331,24	-53,8	2894,44	-979,16
2009	450	895	62,2	3868,84	-60,8	3696,64	-3781,76
2010	445	900	57,2	3271,84	-55,8	3113,64	-3191,76
∑	3878	9558		11745,6		18305,6	-13546,4

La moyenne :

$\bar{X}$=3878 / 10= **387.8**

$\bar{Y}$=9558 / 10= **955.8**

La variance :

***V(x)*=** 11745.6 / 10= 1174.56

***V(y)*=** 18305,6 / 10= 1830.56

L'écart type :

$\sigma(x) = \sqrt{V(x)}$= 34.27

$\sigma(y) = \sqrt{V(y)}$= 42.79

La covariance :

***Cov(X, Y)*=** -13546,4/10**= -1354.64**

Le coefficient de corrélation :

r=1505.16/ (34.27*42.79) =**- 0.92** →Corrélation négative forte.

Le coefficient R^2 :

r^2**=** $(-.92)^2$**= 0.8535**

R^2= $(-.92)^2$*100=**85.35%.**

85.35% de la variation des obligations(Y) se trouvent expliquée par le lien entre actions(X) et obligations

La droite de régression de **Y** en **X** :

Elle est de la forme : **Y'=*ax+b***

a= Cov(X, Y)/ V(x)= -1354.64/1174.56= **-1.15**

b = $\bar{Y}$ **-** ***a***$\bar{X}$ → ***b***= **955.8** - (1.15)*(387.8) = **1403.06**

C'est-à-dire : **Y'**=-1.15+1403.06

Chapitre VI : Etude des séries chronologiques :

Introduction :

Dans la fiche précédente, nous avons étudié les distributions statistiques à deux variables c'est-à-dire très souvent dans une population il est nécessaire d'étudier des distributions pour lesquels les individus sont décrits selon deux ou plusieurs variables. Nous nous limiterons à l'étude des distributions à deux variables. Lorsque ces deux variables sont quantitatives il est possible de représenter les individus par des points dans un système d'axe, sur le premier axe on repère la première variable et sur le deuxième axe on repère la deuxième variable. Ces points étant représentés sur un graphique il est possible de réaliser des ajustements par des courbes en utilisant des méthodes du type moindres carrés d'étudier des liaisons et des corrélations entre deux variables. Lorsque l'une des variables est le temps on obtient la représentation des séries dites «chronologiques».

Une série chronologique ou temporelle, est constitué par une suite d'observation au cours du temps, ces observations sont régulièrement espacés dans le temps, le temps sera défini comme une variable discrète et les donner observées pourront être journalières, hebdomadaires, mensuelles ou trimestrielles. On pourrait traiter le temps de façon continue mais, dans les sciences humaines et en économie, on raisonne le plus souvent en temps discret.

VI.1.L'objectif essentiel de séries chronologiques :

Est de faire des prévisions en s'appuyant sur la connaissance du passé.

En économie par exemple : l'état fait des prévisions sur le niveau de croissance de la production à court et à moyen terme car la croissance a une influence sur les recettes de l'État, sur le chômage,... etc.

En médecine, par exemple on pourrait chercher à prévoir l'évolution du nombre de personnes atteintes d'une maladie et en conséquence essayer de mettre en place des mesures de prévention pour avoir une évolution moins défavorable.

Ces prévisions reposent sur la décomposition de la variable observée au cours de temps en plusieurs composantes.

Remarque : une série chronologique noter yt est une suite d'observation indexées par un ensemble de ordonné. T= $\{t_1, t_2, \dots, t_n\}$

Cette série se définit aussi comme une série statistique bidimensionnelle (t,y_t) avec $t \in T$, où la première composante du couple « t » est le temps et la deuxième composante est une variable numérique y_t prenant ses valeurs aux instants t.

Cette variable y_t est appelé aussi variable dépendante ou expliquée (c'est-à-dire c'est une variable qui doit être estimée dans le temps), par contre la variable « t » est considéré comme toujours indépendante (explicative).

Notation : N et le nombre total d'observation.

La variable est notée « y » que l'on porte en ordonnée sur les graphes rectangulaires et le temps est noté « t » que l'on porte en abscisses.

La variable y est en liaison donc fonctionnelle avec la variable « temps » c'est-à-dire qu'à chaque période « t » correspond a une valeur unique de y mais une valeur de y peut correspondre à plusieurs dates (la réciproque n'est pas varie).

On peut écrire donc $y=f(t)$.

Où la variable y, prend les valeurs $y_1, y_2, \dots, y_t$ (avec i variant de 1 à t).

Et la variable temps t, prend les valeurs $t_1, t_2, \dots, t_n$ (avec i variant de 1 à n).

Exemple1 : soit le tableau suivant donnant le nombre de vente effectuées par l'entreprise « X » par trimestre :

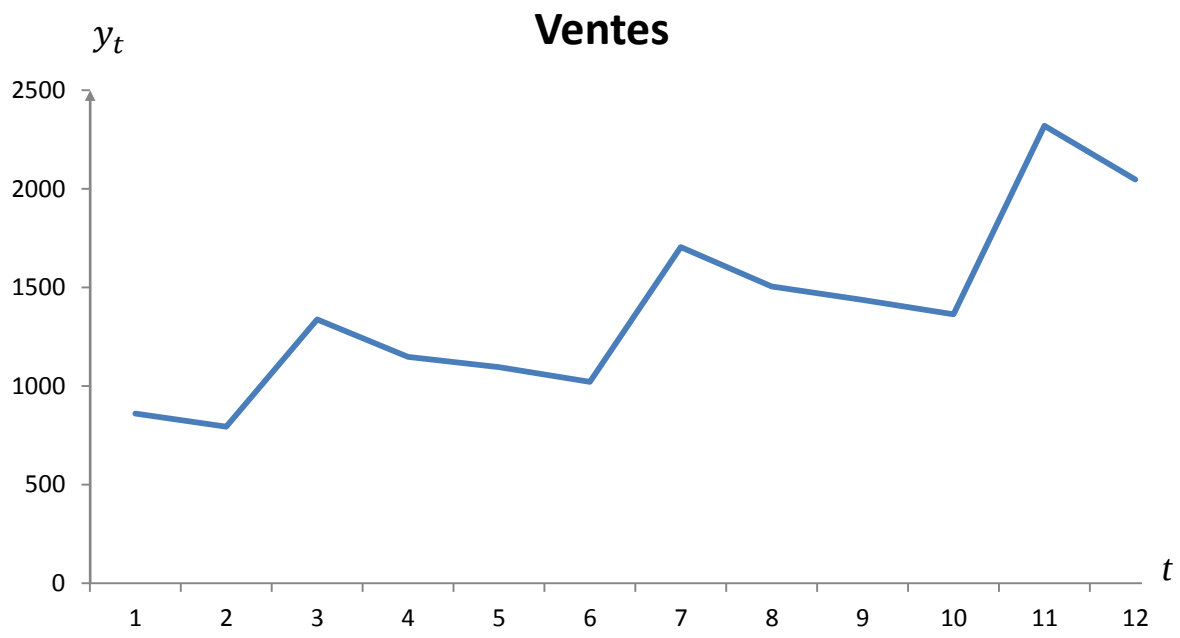
Trimestre / Année	1	2	3	4
2016	860	794	1338	1148
2017	1096	1021	1705	1505
2018	1436	1363	2319	2047

$t=1$ correspond au 1er trimestre 2016.
$t=2$ correspond au deuxième trimestre 2017……..
$t=12$ correspond au 4eme trimestre 2018.

$y_1 = 860\ (ventes)$; $y_2 = 794\ (ventes)$……

$y_{12} = 2047\ (ventes)$

Représentation graphique : on trace la ligne brisée reliant les valeurs de y_t en fonction du temps« t ».



Les points ne sont jamais reliés par des courbes mais par des segments

Compléments :

Les séries chronologiques sont très utilisées, elles représentent l'évolution dans le temps de variables de stock ou de flux qu'il convient de savoir distinguées.

– ***Une variable de stock :*** est représentative d'une quantité à une date instantané « t » ce sera par exemple la population algérienne à la date de recensement de 1995 ; le patrimoine immobilière des ménages algérienne au 1er janvier 2001

– ***Une variable de flux :*** est représentative d'une quantité consommée ou fabriquée durant une période t, ce sera par exemple l'ensemble des naissances enregistrées en Algérie au cours du mois de janvier 2001, l'accroissement des importations alimentaire de l'Algérie entre 2000 et 2010.

Exemple : Le montant de votre compte en banque est une variable de stock, la variation du montant de votre compte en banque au premier jour d'un mois au premier jour du mois suivant est une variable de flux.

L'analyse d'une série chronologique permet de décrire le comportement de la variable étudiée et de faire des prévisions. Dans ce chapitre, nous mettrons l'accent sur la prévision. Les modèles utilisés pour faire des prévisions sont souvent complexes, mais notre propos étant de présenter ce domaine de la statistique, nous n'utiliserons que des modèles simples

On peut considérer une série chronologique comme la résultante de l'action de différentes composantes fondamentales.

VI.2.Les composantes d'une série chronologique:

Les éléments constitutifs d'une série chronologique:

VI.2.1.La tendance ou mouvement de langue durée ou "Trend" : Noté T_t

Traduit l'évolution globale du phénomène étudier cette composante à langue terme concerne l'évolution moyenne sur une période de quelques années, en général 5 ans et plus, cette tendance générale est appeler le "TREND" c'est elle qui sert à ajuster l'ensemble des points de graphe par exemple le taux de croissance en économie sur une longue période.

VI.2.2.Le mouvement cyclique ou la composante cyclique: Noté C_t

Qui regroupent des variations à période moins précise autour de la tendance, ces mouvements peuvent être périodiques (par exemple récession et expansion économique…etc.), sont appelés des cycles ; ces phases durent généralement plusieurs années, mais n'ont pas de durée fixe, indépendamment de l'effet saisonnier, on ignore le plus souvent cette composante cyclique, soit parce que les données statistiques ne remontent pas suffisamment dans le temps soit presque la composante n'existe pas.

VI.2.3.Les mouvements saisonniers ou la composante saisonnière : Noté S_t

Sont des variations se reproduisant périodiquement a des moments bien déterminées se déploient généralement sur des périodes à l'intérieur d'une année :(semaine, mois, trimestre).
En général c'est un phénomène saisonnier d'où le terme de variations saisonnières c'est-à-dire que ses mouvements sont liés au rythme imposé par les saisons métrologique (par exemple production agricole, l'habillement, le tourisme etc.), ou encore pas des activités économique et sociales (par exemple fêtes, vacances,.. etc.).

VI.2.4.Les mouvements accidentels ou la composante accidentelle (résiduelles) : noté ε_t

Ce sont des fluctuations irrégulières en général de faible intensité mais de nature aléatoire; sont dues le plus souvent aux irrégularités de la conjecture et sont donc totalement imprévisible (exemple : grèves, panne, catastrophe naturelle, soulèvement populaire, etc.).

VI.3. Décomposition d'une série chronologique:

Après avoir déterminé les éléments constitutifs de mouvement brut il existe deux façons de décomposer cette série brute par le modèle additif ou par le modèle multiplicatif selon le modèle choisi.

VI.3.1.Le modèle additif :

Dans un modèle de type additif, on considère que le phénomène étudié en fonction du temps se décompose en éléments indépendant les uns des autres et s'ajoutent les uns aux autres dans ce cas nous pouvons écrire y_t etant une valeur observée de la série chronologique (a la date *t* ou durant la période *t*).

$$y_t = T_t + C_t + S_t + \varepsilon_t$$

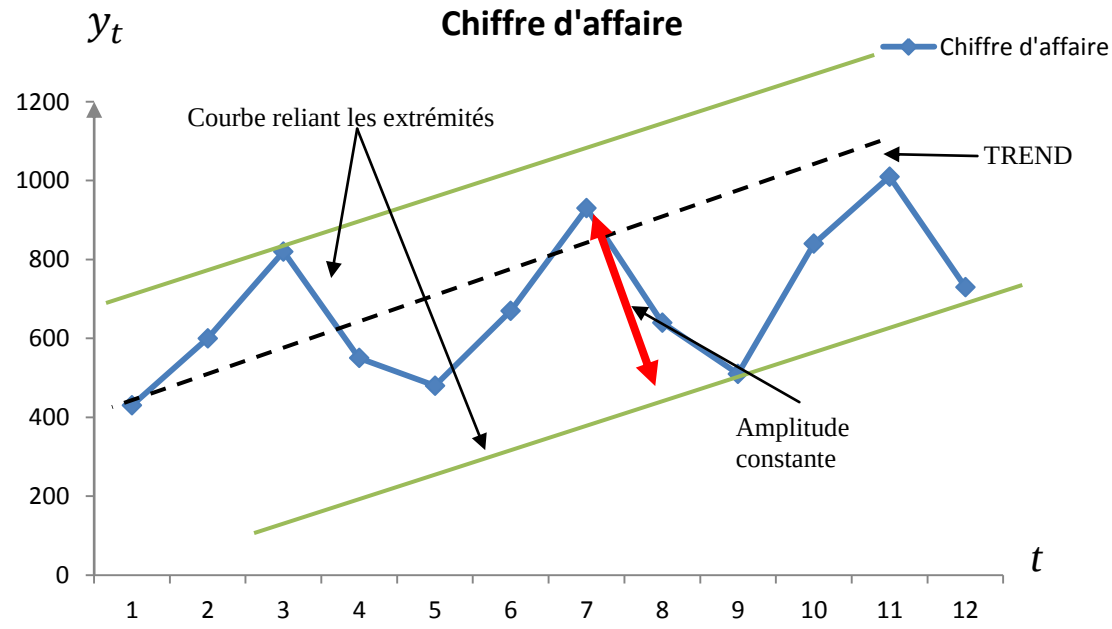
Graphiquement les amplitudes des composantes saisonnières sont constantes par rapport à la tendance « TREND ».

Exemple de représentation des années brutes :

Le tableau ci- contre donne le chiffre d'affaire en millions D.A d'une PME sur 3 années.

	Année 1	Année 2	Année 3
1er trimestre	430	480	510
2eme trimestre	600	670	840
3eme trimestre	820	930	1010
4eme trimestre	550	640	730

On représente cette série chronologique dans un repère orthogonale :



Dans ce modèles il est difficile de distinguer le cycle du TREND c'est-à-dire on considère que le cycle est confondu avec la tendance générale d'une part. D'autre part les variations

accidentelles ou résiduelles sont définies comme étant des mouvements aléatoires de courts périodes souvent imprévisibles et irréguliers.

Par une série d'observations sur *n* années on considère que les variations accidentelles ε_t se composent :

$\sum_{t=1}^{n} \varepsilon_{t=0}$ Dont l'hypothèse $\sum_{t=1}^{n} \varepsilon_{t=0}$ se vérifie.

Donc résumera la formule précédente en :

$$y_t = \mathrm{T_t} + \mathrm{S_t}$$

C.-à-d. que les données brutes se décomposent par le TREND plus la variation saisonnière.

$\Longrightarrow \mathrm{St} = \mathrm{Yt} - \mathrm{Tt}$

VI.3.2 .Modèle multiplicatif:

Il est le plus utilisé car il correspond à des variations saisonnières qui s'ajoutent dans le temps. Le phénomène étudié en fonction du temps se décompose en éléments dépendant les uns des autres, l'amplitude de la composante saisonnière n'est plus constante au cours du temps, elles varient au cours du temps proportionnellement à la tendance.

Dans ce cas nous pouvons écrire y_t de la façon suivante :

La série brute se décompose par la multiplication des ces composantes :

1er Cas : $y_t = T_t \times C_t \times S_t + \varepsilon_t$

2eme Cas : $y_t = T_t \times C_t \times S_t \times \varepsilon_t$

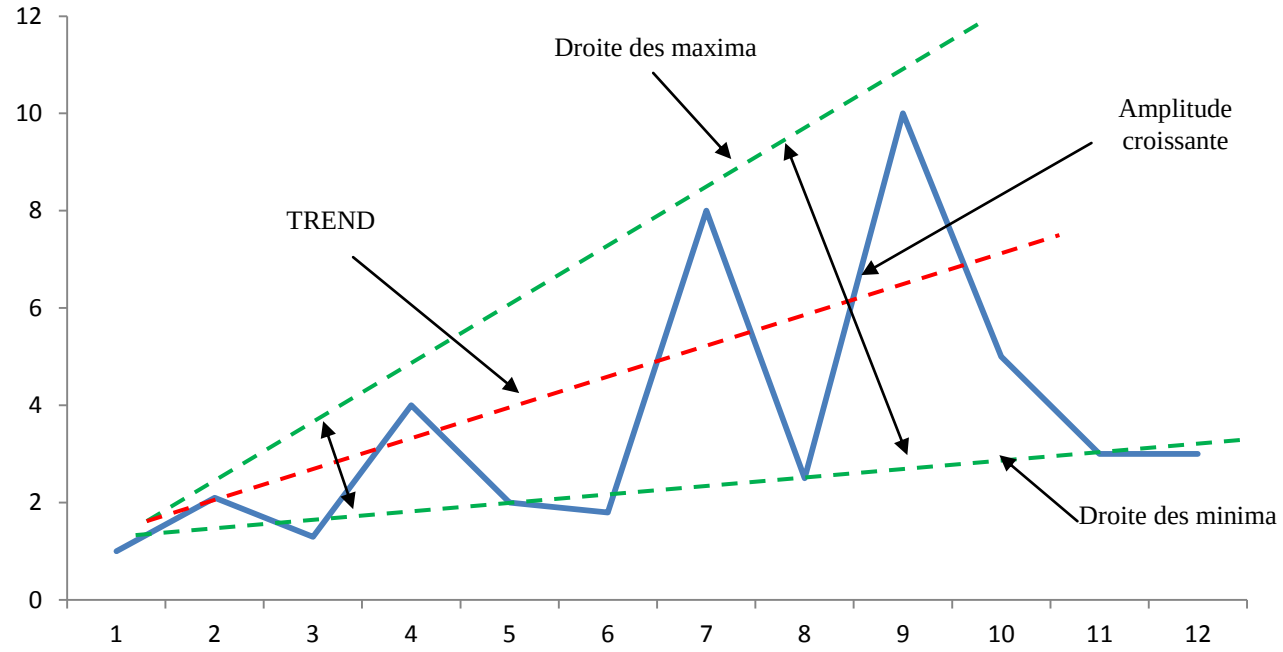
La démarche ressemble à celle utilisée dans le modèle additif c'est-à-dire dans ce cas on impose que le cycle est confondu avec la tendance générale et la somme des variations accidentelles sont nulles :

$$\sum \varepsilon_t = 0.$$

Donc la série brute ne se décompose pas la multiplication du TREND et de la composante saisonnière :

$$y_t = T_t \times S_t$$

Type de représentation graphique de la série brute

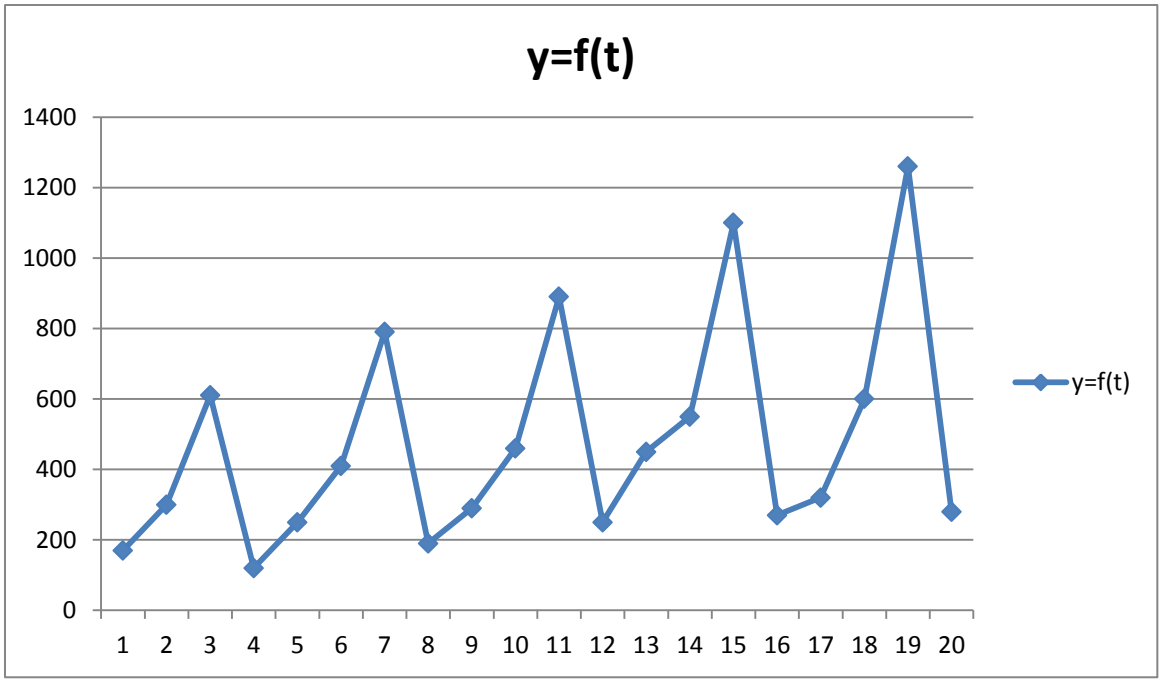


Dans ce cas si nous traçons les deux traits passons par la minima et les maxima de la série, alors nous obtenons deux droites non parallèles, le modèle est donc de type multiplicatif.

Exemple 1 : représentation graphique des données

Les ventes trimestrielles du jus de fruits dans un grand magasin en milliers de litres est comme suit :

Trimestre / Année	**I**	**II**	**III**	**IV**
2014	170	300	610	120
2015	250	410	790	190
2016	290	460	890	250
2017	450	550	1100	270
2018	320	600	1260	280



Le choix du schéma additif ou multiplicatif de l'allure générale de la courbe du phénomène à étudier sur le diagramme et des connaissances que l'on a par ailleurs sur le plan économique et autres.

Pour le schéma multiplicatif du second cas peut être ramené à un schéma additif, c'est-à-dire le schéma devenant linéaire grâce au logarithme.

$$y_t = T_t \times S_t$$

$$\Rightarrow \log y_t = \log T_t + \log S_t$$

En général, les schémas additive est représenté en coordonnées arithmétique et le schéma multiplicatif en coordonnées semi-logarithmiques.

VI.4. Analyse d'une série chronologique :

Les séries chronologiques présentent généralement des variations dues à des causes qui se renouvellent plus ou moins régulièrement, on est donc amené à corriger les données brutes pour obtenir une série corrigée des variations saisonnières.

On rappelle que la série chronologique se décompose en 4 composantes qui se combinent:

Estimation de la tendance:

Il est clair qu'afin de pourvoi estimer la tendance c'est-à-dire le mouvement d'un phénomène observé sur un grand intervalle de temps. il faut disposer d'une série statistique sur une longue période.

Disposant de ces données le premier travail consiste à effectuer une représentation graphique adéquate permettant d'avoir une vue globale de phénomène étudié.

Afin d'éliminer ou d'amortir les mouvements cyclique saisonnières et accidentelles, on utilise la technique des moyennes mobiles ou lissage de la courbe.

VI.4.1. La méthode des moyennes mobiles (ajustement par la MM) :

Le principe de cette méthode est de construire une nouvelle série obtenue en calculant des moyens arithmétiques successifs de la longueur « P » fixe à partir des données originales, chacun de ces moyens obtenus correspondant au " milieu" de la période par laquelle la moyen arithmétique vient d'être calculée.

La méthode de calcul selon que l'on dispose d'un nombre pair ou impaire de données, la méthode de calcul change légèrement, nous allons étudier ces deux cas :

VI.4.1.1. cas de moyen mobile d'ordre impair:

Exemple le tableau ci-dessous contient des mesures d'un phénomène relevées à 9 instants différents :

T	1	2	3	4	5	6	7	8	9
y_t	4	6	5	3	7	5	4	3	6

Nous calculons les moyens mobiles d'ordre 3, nous obtenons les valeurs suivantes:

T	1	2	3	4	5	6	7	8	9
y_t	4	6	5	3	7	5	4	3	6
$\mathbf{MM_3}$		5.00	4.67	5.00	5.00	5.33	4.00	4.33	

On constate que, vu la façon de calculer ses moyens, les deux valeurs extrêmes pour t1 et t9 ont disparu c'est-à-dire une de chaque côté : si nous voulons l'écrire sous forme de formule :

$$MM_3 = \frac{y_{t-1} + y_t + y_{t+1}}{3}$$

Ex : $MM_3(t_6) = \frac{7+5+4}{3} \approx 5.33$

En calculant les moyens mobiles d'ordre 5, nous aurons :

T	1	2	3	4	5	6	7	8	9
y_t	4	6	5	3	7	5	4	3	6
$\mathbf{MM_5}$			5.00	5.20	4.80	4.40	5.00		

Ex : $MM_5(t_4) = \frac{6+5+3+7+5}{5} \approx 5.2$

Cette fois-ci les deux valeurs extrêmes de la série sont perdus on constate que si p est impair donc la forme : $p = 2r + 1$

À chaque extrémité r valeurs sont perdues, c'est-à-dire les valeurs perdues sur l'extrémité de la série est de : $\frac{p-1}{2}$

Et la moyenne mobile relative à la date *t* prend la forme :

$$MM_p(t) = MM_{2r+1}(y_t) = \frac{1}{2r+1}\sum_{k=-r}^{k=+r} y_{t+k}$$

$$MM_p(y_t) = \frac{1}{p}\sum_{k=-r}^{k=+r} y_{t+k}$$

Exemple :

Si p=3=2r+1=> r=1 : donc $MM_3\ (y_t) = \frac{1}{3}(y_{t-1} + y_t + y_{t+1})$

Si p=5=2r+1=> r=2 : donc $MM_5\ (y_t) = \frac{1}{5}(y_{t-2} + y_{t-1} + y_t + y_{t+1} + y_{t+2})$

VI.4.1.2. Cas de moyennes mobile d'ordre pair :

Si p pair donc de la forme P = 2r et les valeurs perdues sur l'extrémité de la série est de : $\frac{p}{2}$

Exemple 2:

Dans la série chronologique précédente si nous calculons les moyens mobiles d'ordre 4 on raisonne sur 5 données en prenant seulement la moitié de la première et la moitié de la 5eme, en y ajoutant la deuxième, la troisième et la quatrième et divisant le tout par 4.

Nous aurons donc les moyens mobiles d'ordre 4 suivantes :

T	1	2	3	4	5	6	7	8	9
y_t	4	6	5	3	7	5	4	3	6
$\mathbf{MM_4}$			4.88	5.13	4.88	4.75	4.63		

Exemple : $MM_4(y_3) = \frac{\frac{1}{2}(4)+6+5+3+\frac{1}{2}(7)}{4} = 4.875$

Si nous voulons l'écrire sous forme de formule :

$$MM_4(y_t) = \frac{1}{4}(\frac{1}{2}(y_{t-2}) + y_{t-1} + y_t + y_{t+1} + \frac{1}{2}(y_{t+2}))$$

Donc si p pair : p=2r et la moyenne mobile relative a la date *t* prend la forme :

$$MM_P(t) = MM_{2r}(t) = \frac{1}{2r}[\frac{1}{2}(y_{t-r}) + \sum_{k=-r+1}^{k=r-1} y_{t+k} + \frac{1}{2}y_{t+r}]$$

Exemple si p = 6 = 2r => r=3

$$MM_6(y_t) = \frac{1}{6}[\frac{1}{2}y_{t-3} + y_{t-2} + y_{t-1} + y_t + y_{t+1} + y_{t+2} + \frac{1}{2}y_{t+3}]$$

Donc le choix pratique dans l'ordre d'une moyenne mobile nous rappelons que le but d'un lissage par moyen mobile et de faire apparaître l'allure de la tendance.

On fait disparaître la composante saisonnière de période p avec une moyenne mobile d'ordre p on gomme d'autant le bruit que l'ordre de la moyenne mobile est élevé.

VI.4.2. Ajustement analytique de la série chronologique:

On rappelle que l'objet essentiel des séries chronologiques et de faire des prévisions, ces prévision sont obtiennent dissociant les principales composantes de la série.

Le schéma d'évolution de mouvement global peut-être décomposé en deux composantes *:*

Le TREND et la variation saisonnière, on admet que les variations accidentelles sont intégrées au trend ou bien nulles.

Détermination des coefficients a et b de TREND linéaire:

On étudiera successivement le cas du schéma additif et celui du schéma multiplicatif.

VI.4.2.1. Ajustement dans le cas de schéma additif:

En supposant les variations accidentelles Ɛt nulles, la série s'écrit :

$$y_t = T_t \times S_t$$

Lorsque la tendance générale suggère un ajustement linéaire le trend à pour équation une droite de la forme:

$$T_t = at + b$$

La série sera ajustée par une équation linéaire de la forme :

$$y_t = at + b + S_t$$

Les données de la série statistique se présentent sous la forme d'un tableau de contingence constituée seulement de deux colonnes t_i et y_i.

Les coefficients de trend Tt= at+b sont déterminés par la méthode des moindres carrés comme au chapitre précédent on a remplacé la variable X par la variable T temps

La pente de la droite $\boldsymbol{T_t = at + b}$ est égal a :

$$a = \frac{cov(t,y)}{\delta^2 t}$$

Ou

$$a = \frac{\frac{1}{n}\sum t_i y_i - \bar{t}\bar{y}}{\frac{1}{n}\sum t_i^2 - \bar{t}^2} = \frac{\sum t_i y_i - n\bar{t}\bar{y}}{\sum t_i^2 - n\bar{t}^2}$$

Son ordonnée à l'origine est égal à : $b = \bar{y} - a\bar{t}$

Pour pourvoir calculer les coefficients a et b, on doit faire le tableau de calculs permettant de calculer la somme $t_i y_i$ et de t_i^2 .Mais celui de t_i ety_i pour calculer leurs moyennes arithmétique respective.

VI.4.2.2. ajustement dans le cas du schéma multiplicatif :

Lorsque le schéma est multiplicatif de la forme $y_t = T_t \times S_t$ (avec ε_t négligeable), il peut être exprimé de la manière suivante :

$$\log y_t = \log T_t + \log S_t$$

Le schéma devenant linéaire grâce au logarithme, la détermination des coefficients a et b se détermine de la même façon que précédemment.

Exemple 3:

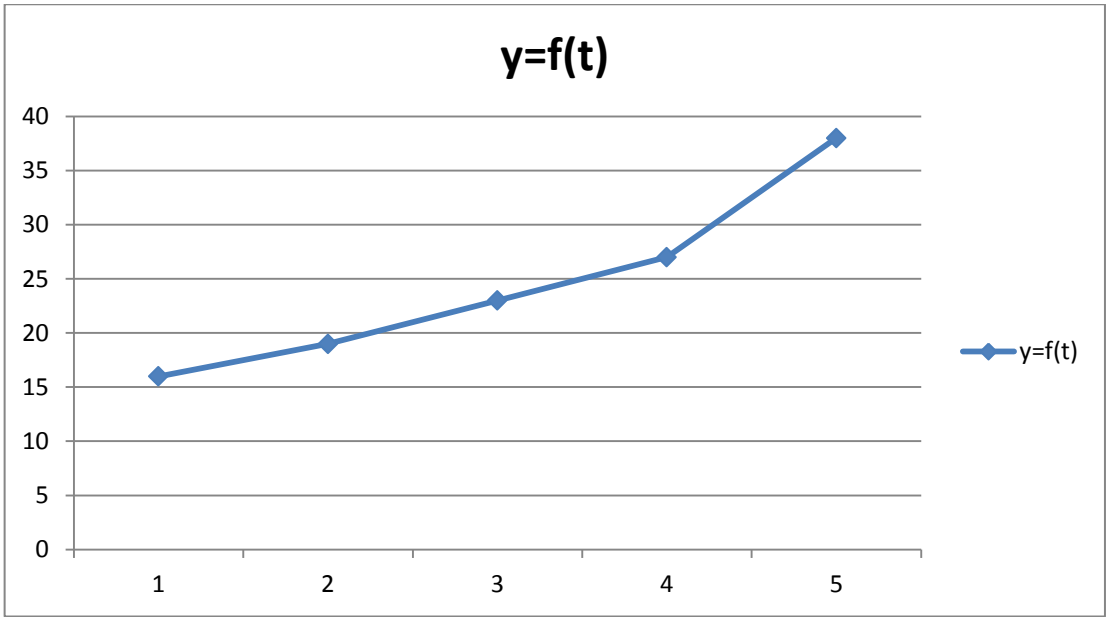
Le tableau suivant donnant les données d'une période de 2012 à 2016.

Estimer l'équation de la droite de la tendance générale (l'équation du trend) de ces données :

Année	y_i	t_i	$t^2{}_i$	$t_i y_i$	$\boldsymbol{\acute{Y}_i = 5.2t_i + 9}$
2012	16	1	1	16	14.2
2013	19	2	4	38	19.4
2014	23	3	9	69	24.6
2015	27	4	16	108	29.8
2016	38	5	25	190	35
Total	123	15	55	421	123

Solution :

Représentation graphique de la série chronologique :



D'après la Représentation graphique On constate qu'il y'a accroissement du phénomène. La tendance pourrait être représentée par une droite de pente positive de la manière suivante

$y_i = at_i + b$

Pour déterminer les constantes a et b de la droite **$y_i = at_i + b$** on utilise la méthode de moindre carrée :

$\bar{t} = \frac{\sum ti}{n} = \frac{15}{5} = 3$; $\bar{y} = \frac{\sum yi}{n} = \frac{123}{5} = 24.6$

a= $\frac{cov(t,y)}{\sigma 2\, t} = \frac{10.4}{2} = 5.2$

$cov(t,y) = \frac{\sum tiyi}{n} - \overline{ty} = \frac{421}{5}$ – (3*24.6) = 84.2 – 73.8 = 10.4

$\sigma^2{}_t = \frac{\sum t^2}{n} - \bar{t}^2 = \frac{55}{5}$ - 9 = 2

b = $\bar{y}$ - a$\bar{t}$ → b = 24.6 – (5.2*3) = 9.

Donc l'équation de la tendance est : **$\acute{Y}_i = 5.2t_i + 9$**

Remarque $\sum \acute{Y}i = \sum y_i$

VI.5. Série désaisonnalisée ou corrigé des variations saisonnières (CVS):

Si on reprend l'exemple précédent qui donne le chiffre d'affaire d'une PME sur 3 années

On distingue sur le graphique:

- une tendance à la augmentation du chiffre d'affaires c'est la tendance générale à long terme (appelée aussi le trend).
- des variations saisonnières : le chiffre d'affaires augmente chaque année au 2eme et 3eme trimestre, il baisse au 1er et 4eme trimestre.

La tendance générale peut se représenter par une droite D l'équation de la droite D peut-être trouvé par une méthode d'ajustement linéaire.

Pour tenir compte des variations saisonnières (augmentation chaque année au 2ème et 3eme trimestres et baisse au 1er et 4ème trimestres), on va calculer des donner qui vont tenir compte de ces variations pour ajuster au plus prés les prévisions, on parle alors de données CVS (corrigé des variations saisonnières).

On appelle série désaisonnalisée ou série corrigé des variations saisonnières notée série CVS, la série chronologique y_t à laquelle on a enlevé les variations saisonnières.

Intérêts:

- La particularité de la CVS et que les données de cette série sont directement comparable on a enlevé l'effet des saisons et donc le caractère propre de chaque mois on peut donc par exemple comparer les données d'un mois de janvier et celle du mois de juillet.
- À partir de la série CVS, on peut réévaluer le la tendance par ajustement (moindre carré, ou par moyen mobile, etc.);afin d'avoir une meilleure estimation de la tendance.

VI.5.1. Méthode de calcul de la série CVS:

Soit une série chronologique y_t

1- on trace le graphe de y_t de la série brute.

2- on estime la tendance par les moyens mobile sur p période, c'est-à-dire calcule les moyens mobiles$MM_P(y_t)$.

3- on choisit le modèle de composition additif ou multiplicatif :

VI.5.1.1. Schéma additif :

$$y_t = T_t + S_t$$

On calcule les différences de : $y_t - MM_p(y_t)$.

On a (n-1) différences : $(y_t - MM_p(y_t))$.

4- On calcul les coefficients saisonnières selon le modèle choisi.

Exemple : modèle additif on prend comme s_j : médiane ou moyenne pour chaque saison des (n-1) différences (c.-à-d. pour $y_t - MM_p(y_t)$) ce sont des coefficients bruts.

Puis on calcul $\bar{s} = \frac{\sum s_j}{n}$

5- Puis on fait, ou en ressort les coefficients définitifs : $S_j = s_j - \bar{s}$

Remarque : pour le schéma additif : $\sum S_j = 0$

Ces coefficients saisonniers vont nous servir à déterminer la série corrigés des variations saisonnières de la manière suivante :

$$y_{ij}^{CVS} = y_{ij} - S_j$$

C'est-à-dire nous obtenons la série corrigée (CVS), on faisant la différence entre la série brute y_t est les coefficients saisonnières S_j.

VI.5.1.2.Pour le schéma multiplicatif :

$$y_t = T_t \times S_t$$

1- La dessaisonalisation s'effectuera en calculant les moyens mobiles sur p période.

2 Les coefficients bruts sont obtenus en faisant le rapport entre la moyenne mobile et la valeur brute : $\frac{y_t}{MM_p(y_t)}$

3- on Calcule les coefficients saisonniers, on prend comme s_j médiane ou moyenne des rapports $\frac{y_t}{MM_p(y_t)}$ pour chaque saison.

Puis on calcul $\bar{s} = \frac{\sum s_j}{n}$

Puis on définit alors $S_j = s_j - \bar{s}$

Remarque : pour le schéma multiplicatif : $\sum S_j = 1$

5- Ces coefficients saisonniers vont nous servir à déterminer la série corrigés des variations saisonnières de la manière suivante :

$$y_{ij}^{CVS} = \frac{y_{ij}}{Sj} = \frac{y_t}{S_j}$$

Exercice 4

Soit la série chronologique suivante donnant l'évolution de chiffre d'affaire (en milliers DA) d'une entreprise donnée sur 3 années.

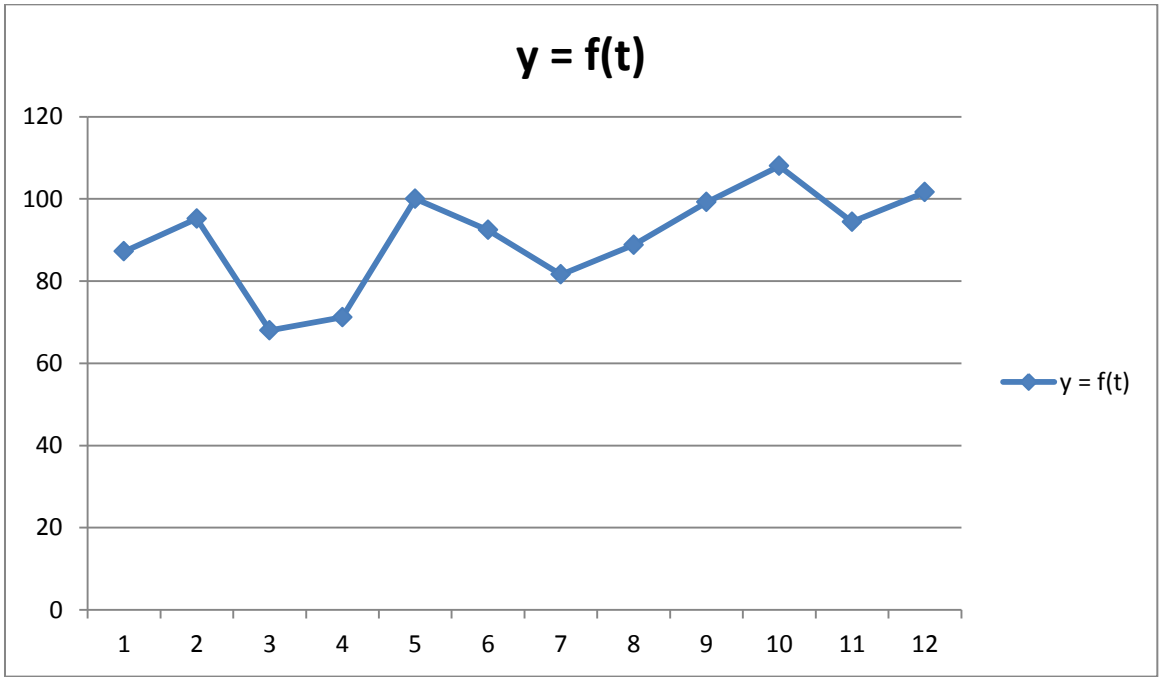
- déterminer les coefficients saisonniers et la série désaisonnalisée CVS.

Yt :

	I	II	III	IV
2016	87.2	95.2	68	71.2
2017	100	92.4	81.6	88.8
2018	99.2	108	94.4	101.6

Solution :

Représentation graphique de la série chronologique :



D'après la Représentation graphique On constate qu'il y'a accroissement du phénomène. La tendance pourrait être représentée par une droite de pente positive. Il existe des variations saisonnières et les amplitudes des composantes saisonnières sont constantes par rapport à la tendance, dans ce cas le modèle est **de type additif** ; on considère que le phénomène étudié en fonction du temps se décompose en élément indépendants les uns des autres.

Donc nous pouvons écrire la composante de la manière suivante :$\mathbf{yt = T_t + C_t + S_t + \mathcal{E}_t}$

La dessaisonalisation s'effectuer en calculant les moyens mobiles sur 4 trimestres supposant la composition additive des mouvements.

Moyen mobile sur 4 trimestres:

$$MM_4(y_t) = \frac{1}{4}\left(\frac{y_t - 2}{2} + y_{t-1} + y_t + y_{t+1} + \frac{y_{t+2}}{2}\right).$$

	I	II	III	IV
2016	-	-	82	83.25
2017	84.6	88.5	90.6	92.45
2018	96	99.2	-	-

$MM_4(y_1)$= - ; $MM_4(y_2)$ = -

$MM_4(y_3) = \frac{1}{4}\left(\frac{87.2}{2} + 95.2 + 68 + 71.2 + \frac{100}{2}\right) = 82$ …….

$$MM_{10}(y_t) = \frac{1}{4}\left(\frac{88.2}{2} + 99.2 + 108 + 94.4 + \frac{101.6}{2}\right) = 99.2$$

Les coefficients bruts sont obtenus en faisant la différence entre y_t (la série brute) et les moyenne mobiles.

$$y_t - MM_4(y_t)$$

	I	II	III	IV
2016	-	-	-14	-12.05
2017	15.4	3.9	-9	-3.65
2018	3.2	8.8	-	-

Les moyens s_j (les moyennes des coefficients bruts) :

s_j:	9.3	6.35	-11.5	-7.85

$s_1 = \frac{15.4+3.2}{2} = 9.5$ $\quad$ $s_2 = \frac{3.9+8.8}{2} = 6.35$

$s_3 = \frac{-14+(-9)}{2} = -11.5$ $\quad$ $s_4 = \frac{-12.05+(-3.65)}{2} = -7.85$

D'où les coefficients $S_j = s_j - \bar{s}$; $\bar{s} = \frac{\sum s_j}{n}$ = - 0.925

S_j	10.2	7.3	-10.6	-6.9

$$\sum S_j = 0$$

D'où la série CVS $y_{ij}^{CVS} = y_{ij} - S_j$

	I	**II**	**III**	**IV**
2016	77	87.9	78.6	78.1
2017	89.8	85.1	92.2	95.7
2018	89	100.7	105	108.5

Exercice 5

Soit le tableau statistique suivant donnant l'évolution du nombre de travailleurs dans la branche textile :

Année	2012	2013	2014	2015	2016	2017	2018	2019
Nbre d'employés	495	482	468	447	428	411	391	370

1- Déterminer l'équation qui ajuste les données selon le principe des moindres carrés.

2- Est-ce que la droite d'ajustement constitue une bonne estimation de la tendance de l'emploi dans la branche textile?

3- Quel serait le nombre d'emploi pour l'année 2022.

Solution :

On effectue un changement d'origine pour le temps pour simplifier les calculs, l'année 2012 devient t=1 et ainsi de suite.

t	y_t	t_i^2	$y_i t_i$	y_i^2
1	495	1	495	245025
2	482	4	964	232324
3	468	9	1404	219024
4	447	16	1788	199809
5	428	25	2140	183184
6	411	36	2466	168921
7	391	49	2733	152881
8	370	64	2960	136900
36	3492	14954	14954	1538068

La droite d'ajustement *y* en *t : y=at +b*

- *le coefficient « a » de la droite :*

$$a = \frac{\sum t_i y_i - n\bar{t}\bar{y}}{\sum t_i^2 - n\bar{t}^2} = \frac{\frac{1}{n}\sum t_i y_i - \bar{t}\bar{y}}{\frac{1}{n}\sum t_i^2 - \bar{t}^2}$$

$$\bar{t} = \frac{\sum t_i}{n} = \frac{36}{8} = 4.5 \qquad \bar{y} = \frac{\sum y_i}{n} = \frac{3492}{8} = 436.5$$

$$a = \frac{14954 - 8(4.5)(436.5)}{204 - 162} = -18.095$$

- *le coefficient b :*

$$b = \bar{y} - a\bar{t} = 517.92$$

Donc l'équation de l'évolution du taux de pénétration en fonction du temps est :

$$y = -18.1+517.9$$

2- La droite d'ajustement constitue une bonne estimation de la tendance de l'emploi dans la branche textile si le coefficient de détermination est supérieure à 0,9 c'est-à-dire le coefficient r il faut qu'il soit proche de 1.

Pour calculer le coefficient il faut calculer la variance de *y* et de *t.*

$\delta_y^2 = \frac{\sum y_i^2}{n} - \bar{y}^2$ $\qquad$ $\delta_t^2 = \frac{1}{n}\sum t_i^2 - \bar{t}^2$

$\delta_y = 41.55$ $\qquad$ $\delta_t = 2.29$

$$r = \frac{cov(y_t)}{\delta_y \delta_t} = a\frac{\delta_t}{\delta_y}$$

r= -0.998

Le coefficient de corrélation linéaire est de 0.998, le coefficient de détermination est de 0.996, la régression est satisfaisante.

3- La valeur de coefficient de détermination permet une estimation correcte de l'emploi en 2022, de plus cette date n'est pas fondamentalement changé, le nombre d'années qui s'écoule de 2012 à 2022 est égale à t = 11, c'est-à-dire l'année 2022 est l'année t=11

Donc l'emploi estimé pour l'année 2022 sera alors de :

$$y=- 18,1(11) + 517,9 = 318,8 \approx 319$$

Exercices de révision :

Comme on a signalé dans l'introduction, d'autres exercices de révisions sans laissés sans solutions, qui vont d'être traiter sur d'autres manuels dans les prochaines publications

Exercice n°1 :

La série statistique ordonnée suivante représente le poids en KG d'un ensemble d'individus.

50 - 55 - 61 - 62 - 64 - 65 - 66 - 67 - 68 - 69 - 70 - 72 - 73 - 74- 75 - 76 - 76 - 77 -78 - 79 - 80 - 81 - 81- 82 - 83- 83 - 84- 85 - 85 - 85 - 86 - 87 - 87 - 88 - 90 - 92 - 92 - 93 - 93 - 95 - 95 - 96 - 97 -98 - 100 - 102 - 102 - 104 -104 - 105 -107 - 109 - 111 - 118

1- Déterminer la population et le caractère étudié ?
2- Mettre ces données dans un tableau de distribution des effectifs ?
3- Calculer les paramètres de la tendance centrale. Que pouvez-vous en conclure ?

Exercice n°2 :

Le tableau suivant nous montre la distribution d'un groupe de 15 étudiants selon leurs dépenses quotidiennes :

Dépense quotidienne(DA)	0 – 10	10 - 50	50 - 100	100 et plus
Nombre d'étudiants	2	6	6	1

1- Déterminer la population, l'unité statistique (individus) ; et la nature du caractère ?
2- Donner la proportion % des étudiants selon leurs dépenses quotidiens ?
3- Donner la proportion % des étudiants qui dépense plus de 100 unités monétaires par jour ?
4- Donner la proportion % des étudiants qui dépense moins de 10 unités monétaires par jour ?
5- Donner la proportion % des étudiants qui dépense moins de 100 unités monétaires par jour ?
6- Quelle l'importance d'utilisation des méthodes empirique : la méthode de YULE et de STURGES et comment les déterminer ?
7- Déterminer le mode de cette distribution ; interpréter les résultats, on montrant les avantages et les inconvénients du mode ?

Exercice n°3 :

Le tableau suivant donne la répartition des 200 salariés selon leurs salaires

420-300	300-220	220-L	L-120	120-80	80-40	40-0	Salaire
6	22	28	34	N_3	N_2	12	Employer

1- déterminer la population, l'unité statistique (individus) ; et la nature du caractère ? Retrouver N2 et N3 sachant que le quatrième décile est égale 95 ?
2- Retrouver la limite L sachant que la moyenne arithmétique est égale à 130 ?
3- calculer : $\sum (2i+1)$; puis écrire la relation suivante on utilisant le signe $\sum$: 5+7+9+…. +17+19
4- Quelle est l'importance d'utilisation des méthodes empirique : la méthode de YULE et de STURGES et comment les déterminer ?
5- Salaires mensuels dans une petite entreprise de 5 salariés est comme suit:2500,3200,3800, 4500,8700.

   * calculer la médiane ? Puis la moyenne arithmétique ?

   * Nous modifions le dernier salaire à 22500, quelle sera alors la valeur de la médiane et la moyenne arithmétique de cette série, que peut-on conclure si on compare la médiane avec la moyenne ? Quelles est les avantages et les désavantages des valeurs de la tendance centrale (paramètre de position) ? quant-est on peut considérer que la moyenne géométrique, la moyenne arithmétique et la moyenne quadratique comme caractéristique de la tendance centrale ?

Exercice n°4 :

Dans une faculté d'économie de 3000 étudiants. Le choix des spécialités des étudiants de 2e année se présente comme suit : Finances 40%, gestion 25%, Comptabilité 20%, Marketing 10%, commerce international 5%.

1- Dresser le tableau de distribution des fréquences.
2- Quelles sont les représentations graphiques possibles.
3- Tracer une représentation graphique de votre choix.

Exercice n°5

Le tableau suivant représente la distribution des travailleurs en Algérie dans les différents secteurs, pendants deux périodes différentes (1960 - 1985)

Année / secteurs	1985	1960
Agriculture	14	6
Industrie	28	17
Bâtiments	16	11
Administration	25	62
Services	17	4
TOTAL	100	100

SOURCE : Abdekader sid Ahmed, revue **Tiers-Monde, Mars 1991.** Page 12 .

1- Quelles sont les différentes représentations graphiques possibles pour illustrer ce tableau.
2- Représentez graphiquement les données de 1960 de deux manières différentes.
3- Représentez les données du tableau dans un seul graphe, que pouvez vous conclure.

Exercice 6 :

Dans une société, les membres du personnel ont été classés d'après leur ancienneté dans l'établissement :

X : ancienneté en année	< 1	1 - 5	5 – 10	10 - 15	15 - 20	20 - 25	25 – 30
N : nombre de salariés	22	24	52	30	40	20	12

1- Déterminer la population, l'unité statistique (individus) ; et la nature du caractère ?
2- Déterminer par le calcule le pourcentage des salariés dont l'ancienneté est :
- Inférieure à 5 ans.
- Supérieure à 20 ans.
3- Déterminer la médiane par la méthode graphique ?
4- Calculer le premier quartile Q1 et le troisième quartile Q3 et en déduire l'intervalle interquartile ?

Exercice n°7 :

Désirant connaitre mieux sa clientèle, le directeur d'une entreprise fait procéder à un sondage, il obtient entre autre informations, la répartition par âge des clients

Age :ans	15 -20	20 – 25	25 - 30	30 – 35	35 - 40	40 - 45	+ 45	Total
fi%	13	26	28	15	10	5	3	%100

1- Tracer l'histogramme, en déduire le mode ?
2- Tracer les courbes de fréquences cumulées ascendantes et descendantes, en déduire la médiane ?
3- Retrouver les valeurs du mode, la médiane et la moyenne arithmétique par le calcul ?
4- Donner la proportion % des clients dont l'âge est ≥ 40 ans ?

Exercice n°8:

Une population statistique se présente comme suit :

Classes	[0 - 4[	[4 - 10[	[10 - 20[	[20 - 40[
Effectifs	4	20	14	2

1- Calculer la moyenne arithmétique et la variance.
2- Déterminer le coefficient d'asymétrie et interpréter le résultat.
3- Chacune des classes de la distribution précédente est divisée en deux classes de même amplitude, auxquelles on fait correspondre un effectif moitié de l'effectif initial de la classe qui a été divisée. Faire un nouveau tableau puis calculer la moyenne arithmétique et la variance. Comparer les résultats obtenus et interpréter.

Exercice n°9

Le tableau suivant représente la distribution de 80 employée selon le salaire mensuel en 10^3 ;

Salaires	[5-10[	[10-15[	[15-20[	[20-25[	[25-30[
Employés	8	10	12	30	20

1- Tracez l'histogramme puis trouvez graphiquement la valeur du mode.
2- Calculez le premier quartile et le troisième quartile puis commentez les résultats.
3- Quel est le pourcentage des travailleurs ayant un salaire compris entre 15000 et 25000 DA.

Exercice N°10 :

Le tableau suivant représente la répartition des salaires (en 103 DA) de 50 ouvriers d'une entreprise.

Salaires (en 10^3 DA)	[20-40[	[40-60[	60-80[	80-100[	[100-120[
Fréquence relative	0.18	0.2	0.3	0.2	0.12

1- calculez la fréquence absolue de chaque classe, les fréquences cumulées croissantes et les fréquences cumulées décroissantes
2- Calculez la moyenne arithmétique en utilisant la formule de définition et ensuite en utilisant la méthode de changement de variable.
3- Trouvez la valeur de la médiane et la valeur du mode ; comparez les trois valeurs centrales et en déduire la forme de la distribution

Exercice n°11

Désirant connaitre mieux sa clientèle, le directeur d'une entreprise fait procéder à un sondage, il obtient entre autre informations, la répartition par âge des clients

Age :ans	15 -20	20 – 25	25 - 30	30 – 35	35 - 40	40 – 45	+ 45	Total
fi%	13	26	28	15	10	5	3	%100

1- Tracer l'histogramme, en déduire le mode ?
2- Tracer les courbes de fréquences cumulées ascendantes et descendantes, en déduire la médiane ?
3- Retrouver les valeurs du mode par le calcul ?
4- Donner la proportion % des clients dont l'âge est ≥ 40 ans ?

Exercice n°12 :

Le tableau ci-après réunit les informations des salaires horaires perçus par 50 salariés d'une entreprise au cours du mois de mars 2018 :

340	**360**	**450**	**620**	**370**	**430**	**420**	**1020**	**310**	**420**
510	**300**	**610**	**630**	**470**	**1050**	**520**	**430**	**810**	**950**
920	**770**	**600**	**360**	**480**	**490**	**650**	**710**	**780**	**810**
430	**520**	**630**	**710**	**430**	**420**	**510**	**550**	**610**	**410**
930	**820**	**830**	**470**	**540**	**610**	**1020**	**330**	**480**	**550**

1- Déterminer la population étudiée, le caractère, sa nature et ses modalités.

2- Dépouiller les observations et présenter les résultats dans un tableau statistiques à l'aide de classes.

3- Présenter l'histogramme et la courbe des effectifs cumulés croissants et décroissants.

4- Déterminer le mode, la médiane et la moyenne arithmétique. Comparer.

5- Calculer la variance et l'écart-type.

Exercice n°13 :

On donne la répartition de 40 employés dans une entreprise selon le nombre d'enfants à charge :

Nombre d'enfants à charge (x_i)	0	1	2	3	4	5	6
Effectifs (n_i)	4	7	8	8	6	4	3

1- Représenter graphiquement cette série.

2- Déterminer le mode. A quoi correspond-il graphiquement ?

3- Déterminer la fonction de répartition et la représenter graphiquement.

4- Calculer la variance et l'écart-type de cette distribution

Exercice n°14 :

Le tableau suivant donne la répartition de N employés selon leur âge dans une entreprise A :

Age	[20-25[	[25-30[	[30-35[	[35-40[
Employés	20	?	50	10

1- Sachant que la médiane est égale à 30.5. Compléter le tableau.
2- Calculer les quartiles Q1, Q2 et Q3, le cinquième décile D5 et le vingt cinquième percentile P25.
3- Calculer les moyennes harmonique (H), géométrique (G), arithmétique ($\bar{x}$) et quadratique (Q). Comparer.
4- Dans une entreprise **B**, la moyenne des âges des employés est égale à 30 ans. Si dans cette entreprise le nombre d'employés est de 90. Calculer l'âge moyen des employés deux entreprises ensemble.

Exercice n°15 :

Calculer le taux de croissance annuelle moyen de la production d'un produit durant la période 2013-2018 qui a connu les taux de croissance successifs suivants :

Année	2013	2014	2015	2016	2017	2018
Taux de croissance%	9.6	7.4	5.8	2.7	3.1	2.7

Exercice n°16 :

La série suivante représente le nombre d'absences de 06 étudiants aux TD d'une matière : 4 -5 -6 - 2 - 1 -0.

Calculer la variance (puis l'écart-type) en utilisant : La formule de définition puis la formule développée (Théorème de Koenig).

Exercice n°17 :

Le tableau suivant donne la répartition de 50 salariés d'une entreprise selon leur salaire mensuels en 102 DA:

Salaires	[120-130[	[130-140[	[140-150[	[150-160[	[160-170[	[170-180[	[180-190[
Effectifs	4	7	10	15	7	4	3

1- Calculer les quartiles Q1 et Q3, l'étendue, l'écart interquartile, le cinquième décile (D5), le neuvième décile (D9), le quatre-vingt dixième percentile (P90). Interpréter les résultats.

2- Calculer la variance (puis l'écart-type) en utilisant : La formule de définition puis la formule développée (Théorème de Koenig).
3- Calculer le coefficient d'asymétrie algébriquement et graphiquement. Interpréter le résultat.

Exercice n18:

On a relevé le salaire horaire moyen de 100 employés hommes dans une entreprise qui est égale à 24 DA avec un écart-type de 1,56 DA.

On considère également un échantillon d'employées femmes dans la même entreprise avec leurs salaires horaires en DA :

Classes	20-22	22-24	24-26	26-28	28-30	30-32
Effectifs	10	20	30	25	8	7

Comparer les deux distributions à l'aide du coefficient de variation. Que peut-on conclure ?

Exercice n° 19 :

On donne les données suivantes :

$$\sum_{i=1}^{10} x_i^2 = 1060 \qquad \sum_{i=1}^{10} (x_i - \bar{x})^2 = 250$$

Calculer la moyenne arithmétique puis le coefficient de variation.

Exercice n°20:

On donne les données suivantes concernant une distribution symétrique des salaires de 50 employés :

$$\text{Variance} = 18{,}6 \qquad \sum n_i . x_i^2 = 1820$$

- Calculer la médiane et le mode de la distribution.

Exercice n°21:

Une étude sur le chiffre d'affaires d'une population de P.M.E a permis d'obtenir les résultats suivants en milliers de dinars algériens (DA) :

Minimum	3500
Moyenne arithmétique	4900
Ecart-type	650
Mode	4550
Ecart interquartile	1100
Médiane	4600
Premier quartile	4100
Premier décile	3700
Ecart inter décile	2800
Etendue	5000

Travail à faire :

1- Classer ces paramètres en deux catégories :

- tendance centrale et position
- dispersion

2- Quel est le chiffre d'affaires le plus grand dans cette population de P.M.E. ?

3- Calculer le troisième quartile et le neuvième décile.

4- Placer sur un axe les paramètres caractérisant cette série.

Exercice n°22 :

Un automobiliste effectue un trajet total de 900 Km en trois étapes réalisées à des vitesses différentes :

* 162 Km à 60 km/h (18% du trajet total)
* 528 Km à 110 km/h (58.67 du trajet total)
* 210 Km à 140 km/h (23.33 du trajet total)

- Calculer sa vitesse moyenne ?

Exercice 23

Le tableau suivant donne la répartition des 600 salariés d'une entreprise selon la durée, exprimé en minute, du trajet domicile-travail.

Durée du trajet en x minutes	[0à10 [	[10 à 20[	[20 à 30[	[30 à 40[	[40 à 50[	[50 à 60[	[60 à 70[
Effectifs	50	95	127	151	83	54	40

1- Parmi ces trois durées : 15 mn, 35 mn, 60 mn laquelle semble susceptible d'être proche de la durée moyenne de trajet ?
2- Déterminer la durée moyenne par le calcule ?
3- Quel est l'écart absolu moyen du temps du trajet ? le comparer à l'écart type.
4- Calculer le premier décile D1 et le neuvième décile D9 et en déduire un indicateur de dispersion autour de la médiane Me.

Exercice 24 :

On procède à l'achat d'action :

- Pour 26000 euro au cours de 520 euro ;
- Pour 40000 euro au cours de 500 euro ;
- Pour 37100 euro au cours de 530 euro ;
  - Quel est le cours moyens d'une action ?

Exercice n°25 :

On a relevé le salaire horaire moyen de 100 employés hommes dans une entreprise qui est égale à 24 DA avec un écart-type de 1,56 DA.

On considère également un échantillon d'employées femmes dans la même entreprise avec leurs salaires horaires en DA :

Classes	20-22	22-24	24-26	26-28	28-30	30-32
Effectifs	10	20	30	25	8	7

Comparer les deux distributions à l'aide du coefficient de variation. Que peut-on conclure ?

Exercice n°26 :

Une population statistique se présente comme suit :

Classes	[2 - 6[	[6 - 12[	[12 - 22[	[22 - 42[
Effectifs	4	20	14	2

1- Calculer la moyenne arithmétique et la variance.
2- Déterminer le coefficient d'asymétrie et interpréter le résultat.
3- Chacune des classes de la distribution précédente est divisée en deux classes de même amplitude, auxquelles on fait correspondre un effectif moitié de l'effectif initial de la classe qui a été divisée. Faire un nouveau tableau puis calculer la moyenne arithmétique et la variance. Comparer les résultats obtenus et interpréter.

Exercice n° 27 :

Soit le tableau suivant donnant la production de lait pendant 10 ans dans une entreprise donnée :

Année t	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Production (10^2L)	10	12	12	13	14	16	17	18	20	22

1- Faites une représentation graphique des données ; que constatez vous ?
2- Déterminer l'équation qui ajuste les données selon le principe des moindres carrés.
3- Si rien ne change quelle serait le niveau de la production de lait en 2023 ?

Exercice N°28:

On a relevé le salaire horaire moyen de 100 employés hommes dans une entreprise qui est égale à 24 DA avec un écart-type de 1,56 DA.

On considère également un échantillon d'employées femmes dans la même entreprise avec leurs salaires horaires en DA :

Classes	20-22	22-24	24-26	26-28	28-30	30-32
Effectifs	10	20	30	25	8	7

Comparer les deux distributions à l'aide du coefficient de variation. Que peut-on conclure ?

Exercice n° 29 :

On observe sur 11 mois, la consommation de deux biens X et Y. Le tableau est le suivant :

Consommation de X	70	50	51	40	41	55	56	52	58	60	65
Consommation de Y	50	69	70	80	81	65	66	67	64	65	67

On cherche à étudier le sens et l'intensité de la liaison (si elle existe) entre ces deux grandeurs par le calcul de :

1- La covariance entre X et Y.
2- Du Coefficient de corrélation linéaire de Pearson rx,y.
3- Du Coefficient de détermination r^2x,y .

Exercice n°30 :

La droite d'ajustement linéaire de y en fonction de x a pour équation :

Y= X+30

Et celle de x en y a pour équation : X= $\frac{1}{4}$ Y+60

1- Calculer le coefficient de corrélation linéaire.
2- Calculer les moyennes arithmétiques de la variable X et de Y.
3- Calculer la covariance entre X et Y, et la variance de X sachant que la variance de Y est égale à 40.
4- Si le coefficient de corrélation est égale à zéro entre les deux variables, comment peut-on estimer X à partir de Y ?

Exercice n°31:

On a relevé le salaire horaire moyen de 100 employés hommes dans une entreprise qui est égale à 24 DA avec un écart-type de 1,56 DA.

On considère également un échantillon d'employées femmes dans la même entreprise avec leurs salaires horaires en DA :

Classes	20-22	22-24	24-26	26-28	28-30	30-32
Effectifs	10	20	30	25	8	7

- Comparer les deux distributions à l'aide du coefficient de variation. Que peut-on conclure ?

Exercice N°32 :

Le tableau suivant représente l'évolution trimestrielle du chiffre d'affaires en millier de dinars d'une entreprise donnée.

Trimestres / Années	I	II	III	IV
2015	170	300	610	120
2016	250	410	790	190
2017	290	460	890	250
2018	450	550	1 100	270
2019	320	600	1 260	280

Travail à faire :

1- Représenter cette série chronologique dans un repère orthogonal ?
2- À partir de ce graphe, quel est le modèle qu'on peut choisir ?
3- Calculer les moyennes mobiles d'ordre 4 ?
4- Déterminer la série corrigée des variations saisonnières (CVS) ?

Bibliographie :

- Alain PILLER, 2004, Statistique Descriptive, éditions Premium.
- Aragon Y., 2016: Rappels de statistique mathématique ». Polycopié de cours, Université Toulouse,
- Antoine Ayache, Julien Hamonier , 2014: Cours de statistique descriptive. http://math.univlille1. fr/ ayache/cours sd.pdf,
- Baccini A., 2010: Statistique descriptive élémentaire. Publications de l'Institut de Mathématiques de Toulouse,
- Basirat A., 2009: Initiation aux statistiques descriptives. Cours,
- Bressoud E., Kahané J.-C., 2010: Statistique descriptive. 2ème Ed. Pearson France,
- B. Oukacha and M. Benmessaoud, Statistique descriptive et calcul des probabilités, 2013.
- Carricano M., Poujol F., Bertrandias L., 2010: Analyse de Données avec SPSS». 2ème Ed. Pearson France,
- Castillo I., 2012: Statistique descriptive. Cours 1, École des Ponts.
- Chassagnon A., 2010: Introduction à l'analyse statistique: Paramètres de dispersion d'une distribution. Cours L3 LISS - Université Paris-Dauphine, LEDa-SDFi,
- Chaumont L., 2010-2011: Statistique descriptive et prévision. Cours.
- Dakhmouche M., 2011 « Introduction à la Statistique Descriptive ». Ecole Préparatoire en Sciences Economiques Commerciales et des Sciences de Gestion de Constantine.
- De Sede-Marceau M.-H., 2010-2011: De la donnée à la connaissance: traitement, analyse et transmission: Introduction à la statistique descriptive. Élément 424b, Master AGPS, CTU.
- Desgraupes B., 2017: Statistiques descriptives: Indicateurs de forme et de concentration. Université Paris Ouest Nanterre La Défense, U.F.R. Segmi, L1 Économie.
- El Kacimi A., 1983-1984 : Elément de statistique descriptive. Edition Maghrébines, collection arabisation et connaissance, Université de Lille III.
- F. Mazerolle, Statistique descriptive, 2009.
- J. Blard-Laborderie, L'essentiel des outils de statistique descriptive pour aborder des études en sciences humaines et sociales, 2015.
- J. Vaillant, Eléments de Statistique descriptive, 2015.

- G. Calot, Cours de statistique descriptive, Dunod, 1969.
- G. Chauvat and J.-P. Reau, Statistiques descriptives, Armand Colin, 2002.
- Grenier E., Goupy J., Aubert H. P., 2007: Quelle est la bonne formule de l'écart-type? Reims Management School, Revue MODULAD N°37, pp: 102-105.
- Guérin H. 2001: Introduction à la statistique descriptive. Université Rennes 1, https://perso.univ-rennes1.fr/helene.guerin/enseignement/capes/statdes01.pdf
- Hocine hamdani, statistique descriptive avec initiation aux méthodes d'analyse de l'information économique, OPU 2010.
- Lenoir J. P.: Chapitre 1: statistique descriptive. Fiifo 3, probabilités et statistiques, https://www.math.u-psud.fr/pansu/web-ifips/statistique descriptive ch1.pdf, p14.
- L. Leboucher and M.-J. Voisin, Introduction à la statistique descriptive, 2013.
- Long D.: La variance. Université de Moncton, Canada E1A 3E9, pp: 858-488.
- Lucien LEBOUCHER, Marie-José VOISIN Introduction à la statistique descriptive, Cours et exercices avec tableur CÉPADUÈS-ÉDITIONS Toulouse -France août 2011
- Mzali H., 2013: Statistique et calcul de probabilité. Cours de l'Ecole Nationale de L'Administration, Tunis, p98.
- M. Tenenhaus, statistique : Méthodes pour décrire, expliquer et prévoir, Dunod, 2006
- Noel P., 2006 : Statistiques descriptive. http://amphimaths.chezalice. fr/N1/statistique descriptive, poly.pdf, p57.
- P. Roger, Probabilités, statistique et processus stochastiques, Pearson Education,, 2004.
- Pierre Bailly, Christine Carrère, Statistiques descriptives et Exercices ; Collection « Libres Cours Économie » ; Presses universitaires de Grenoble
- Poirrier J. E., 2006: Notes de cours de Statistiques uni variées http://www.poirrier.be/jeanetienne/ notes/statistiques ; univariees.pdf, p19.
- Putois B., 2009: Statistique descriptive: décrire, synthétiser, mettre en forme vos données. Statistica, Psychologie Niv L3.03, p12.
- Rakotomalala R., 2011: Tests de normalité: Techniques empiriques et tests statistiques. Université Lumière Lyon 2, p59.
- Sébastien Gerchinovitz, Polycopié du cours-TD Méthodes et outils de calcul « Introduction `a la méthodologie statistique, L2 Biochimie Année 2014 – 2015.
- Some S. A., 2005: Statistique: Les distributions à un caractère / Quelques applications à l'économie burkinabè. Série documents de travail, DT-CAPES N° 2005-.

- Tillé Y., 2010: Résumé du Cours de Statistique Descriptive. https://www.unine.ch/files/live/sites/statistics/files/shared/documents/cours statistique descriptive.pdf,
- Tugaut J.: Probabilités, Variance, Ecart-type. http://tugaut.perso.math.cnrs.fr/pdf/enseignement/2014/PFI/CM06.pdf.
- جلاطو جيلالي الإحصاء مع تمارين و مسائل محلولة OPU 2007

TABLE DES MATIERES