Université Mouloud Mammeri de Tizi-Ouzou
Faculté des Sciences Economique, Commerciale et des Sciences de Gestion
Département des Sciences Commerciales

Polycopié

Cours de Statistique I

Licence 1 (Socle commun) - Semestre 1

Année universitaire 2023/2024

Réalisé par Dr. HADJEM Madjid

Maitre de conférences B

Syllabus

Nom du module : Statistique I

Domaine : Sciences Economique, Commerciale et des Sciences de Gestion

Niveau : 1ère année Licence

Filière : Toutes les filières

Spécialité : Socle commun

Niveau : Licence 1

Semestre : 1

Année universitaire 2023/2024

Identification du module

Nom du module : Statistique I

Unité d'enseignement : Méthodologique

Crédits : 5

Coefficient : 3

Module annuel dispensé en deux semestres et sous forme de cours et travaux dirigés

Volume horaire hebdomadaire : 4,5 heures

- Cours magistral : 3h
- Travaux dirigés : 1h30

Responsable du module

Nom et prénom : HADJEM Madjid (chargé de cours et TD)

Grade : Maître de conférence B

Affiliation : Département des sciences commerciales

Courriels : madjid.hadjem@yahoo.fr ou madjid.hadjem@ummto.dz

Téléphone : 0791 31 19 26

Description du module

Connaissances préalables recommandées

Connaissances et notions élémentaires en mathématiques, notamment les opérations et les règles dispensées en cycle de seconde, voire de moyen.

Objet de l'enseignement

L'objet de la statistique descriptive est de savoir collecter les données, concernant un phénomène social ou économique donné à étudier, les ordonner, les classifier, les résumer et les présenter, de façon claire et lisible, sous forme de tableaux et/ou graphes afin de pouvoir en calculer des paramètres divers permettant de décrire et d'analyser les variations du phénomène en question.

Objectifs de l'enseignement

Compte tenu de son objet, le présent cours, conçu conformément au programme ministériel (Arrêté n° 808 du 22 juillet 2022), vise à ce que l'étudiant acquiert des compétences en la matière, à savoir :

- maitriser les concepts clés de la statistique descriptive,

- résumer et présenter des données sous forme de tableaux et graphes,

- calculer et analyser les différents paramètres statistiques (de tendance central, de dispersion, de forme et de concentration),

- analyser et quantifier la relation entre deux variables et mesurer leur corrélation,

- calculer les indices de la vie économique et la compréhension de leur signification et leur utilité et leur usage,

- s'initier à l'usage des logiciels statistiques utilitaires, notamment le logiciel de base Excel, pour les graphiques et le calcul des paramètres.

Les chapitre dispensés en cours magistral font l'objet chacun d'un traitement appliqué, sous forme de séries d'exercices, en séances de travaux dirigés. Lors de ces dernières des éclaircissements et des informations supplémentaires, surtout d'ordre pratique, sont fournis aux étudiants.

Une bibliographie révisée et mise à jour est fournie en annexe, dans le but de permettre aux étudiants d'approfondir leurs connaissances.

Les chapitres du présent cours sont également publiés en ligne, sur la plateforme Moodle (E-learning) où un Forum aux questions et un e-mail (coustatistiques2020@yahoo.com) sont mis à la disposition des étudiants pour toute question ou renseignement éventuel.

Mode d'évaluation

Noté sur 20: contrôle continu (40%) et examen (60%).

Moyenne du module = (Note Examen x 0,6) + (Note TD X 0,4)

Plan du cours

Introduction générale

Dans la vie moderne, l'information statistique est très vaste et variée. Aucun domaine n'échappe à l'usage des statistiques, plus particulièrement dans le domaine économique où celles-ci sont plus qu'indispensables. Ce n'est donc pas un hasard que l'enseignement de la statistique soit généralisé.

En Algérie, les réformes structurelles engagées impliquent des changements qui imposent un recours plus fréquent et plus rigoureux à la statistique.

Aussi, à la lumière de cette réalité, les objectifs du présent cours de statistique descriptive, dispensé au semestre 1 (Licence 1) LMD, sont multiples, à savoir :

- initier l'étudiant aux concepts et applications de base de la statistique descriptive ;

- dans la mesure où les applications pratiques fournies sont associées à des développements méthodologiques consistant et utilisant les notations usuelles, le cours fournit une base préparatoire solide pour aborder les problématiques plus complexes de la statistique mathématique et de la modélisation enseignées au semestre 2 et en Licence 2 et Licence 3 ;

- apprendre à l'étudiant à mettre en application les techniques de la statistique descriptive de manière appropriée dans les domaines économique et commercial ;

- pouvoir évaluer le plus correctement possible l'information véhiculée régulièrement par les médias, les revues, les ouvrages et autres publications ;

- enfin, accroitre l'aptitude de l'étudiant, notamment dans sa future vie professionnelle, à prendre les meilleures décisions dans un large éventail de décisions.

Ce cours ne nécessite aucune connaissance préalable en statistique, ni des connaissances en mathématiques supérieures à celles requises en classe de seconde. Les notions et méthodes y sont exposées de manière succincte et seront développées en cours magistral et en T.D.

Le cour est structuré en neuf chapitres, conformément au programme ministériel de l'année en cours (Arrêté n° 808 du 22 juillet 2022). Les chapitres dispensés en cours magistral font l'objet, chacun, d'un traitement appliqué, sous forme de séries d'exercices, en séances de travaux dirigés. Lors de ces dernières des éclaircissements et des informations supplémentaires, surtout d'ordre pratique, sont fournis aux étudiants.

L'étudiant a la possibilité, mais aussi le devoir, de compléter ses connaissances par des lectures supplémentaires, à partir, notamment, de la bibliographie fournie et des informations complémentaires apportées lors des séances de cours et TD.

Chapitre 1 : Généralités sur la statistique et notions fondamentales

Introduction au chapitre 1

L'objet du présent chapitre est de faire découvrir à l'étudiant deux fondamentaux en statistique descriptive : le vocabulaire de base et les soubassements théorique et pratique de l'analyse statistique. Seront ainsi présentés, en trois sections successives, trois volets de définition, en allant du général vers le détail.

Section 1 : Notions de données, statistiques et de Statistique

Toutes les statistiques véhiculées et diffusées par les médias et les différentes publications sont le produit du traitement des données (celles-ci pouvant être d'ordre numérique, alphabétique et/ou alphanumérique) par les méthodes et les techniques de la Statistique (Anderson *& al.,* 2003).

1. Les Données

Ce sont les faits, les lettres ou les chiffres, porteurs d'informations cachées, collectés en vue d'un traitement statistique pour en extraire ou révéler ces informations (Anderson *& al.,* 2003). On les appelle aussi les données ou les informations brutes. Ce sont, en fait, les premiers éléments de l'enquête ou du recensement.

2. Les statistiques

Ce sont les chiffres, lettres et autre, contenus dans les journaux, rapports, revues, mémoires, livres et autre publication, destinés à un large éventail de lecteurs, spécialistes ou non en la matière. Ce sont, par conséquent, des données traitées, résumées et présentées sous une forme facilement compréhensible par le lecteur.

3. La Statistique

C'est à la fois l'art, la discipline ou l'ensemble des techniques et méthodes permettant de collecter, ordonner, résumer, présenter et interpréter les données, afin d'en révéler le message informationnel caché. La Statistique est l'élément privilégié de l'analyse économique.

Une étude de statistique descriptive se déroule globalement en cinq étapes obligatoires, à savoir ;

a- La collecte des données

Soit de manière exhaustive par un recensement, effectué généralement par des institutions spécialisées en déployant de gros moyens, soit par des enquêtes (recensement partiel), généralement effectuées par le chercheur lui-même.

b- Le dépouillement et l'ordonnancement des données

Cette étape consiste à trier et à ordonner, de manière logique et compréhensible, les données collectées. Le plus souvent, on adopte l'ordre alphabétique pour les données alphabétiques, et l'ordre de croissance pour les données numériques.

c- Le classement des données

C'est la première synthèse des données. Soit sous forme individualisée où à chaque donnée on affecte le nombre de fois qu'elle est observée (xi $\longrightarrow$ ni)., soit sous forme d'intervalles ou classes ([a ; b[ ) afin d'en réduire le volume.

d- La présentation des données

C'est la deuxième synthèse des données. Celles-ci sont résumées sous forme de tableaux et/ou de graphiques qui les rendent facilement et rapidement lisibles par le lecteur.

e- Le calcul de paramètres et l'interprétation des résultats

Dans cette étape le calcul de paramètres synthétiques et pertinents devient possible.

L'interprétation de ces derniers offre au chercheur une panoplie de renseignements.

Section 2 : Notions de population, caractère et de modalités

Le soubassement pratique ou le principe d'une étude de statistique descriptive consiste en l'observation d'un ensemble d'éléments, d'individus ou d'unités statistique, appelé ***population statistique***, sur lequel on repère ou on étudie une ou plusieurs propriété(s) ou caractéristique(s), appelées ***caractère(s)***. Celui-ci ou ceux-ci peuvent prendre plusieurs situations ou valeurs possibles, appelées ***modalités***. Ces termes sont empruntés à la démographie (Leboucher & Voisin, 2011).

1. Population statistique

C'est l'ensemble des éléments soumis à l'étude. On l'appelle aussi univers statistique, ensemble fondamental ou ensemble statistique. Par exemple; l'ensemble des étudiants de première année, l'ensemble des logements dans une ville, l'ensemble des voitures dans un parking, etc.

Les éléments qui constituent la population sont appelés unités statistique, éléments ou individus (Baccini, 2010). Ils peuvent être des êtres humains, des objets ou des événements, c'est-à-dire physiques ou immatériels. Ces individus sont de même nature et forment un ensemble homogène.

La population statistique doit être définie de manière précise car cela conditionne fortement l'étude statistique et ses résultats.

Le nombre d'individus composant une population statistique s'appelle ***taille*** de la population ou ***effectif total***, désigné par "**N**".

Remarque : en pratique, la taille des populations à étudier est tellement importante qu'on a souvent recours à des sous-ensembles de populations qu'on appelle ***échantillon***.

2. Caractère

C'est l'aspect, la propriété, l'attribut ou le trait particulier que l'on désir étudier. Il est observable sur tous les individus de la population et est susceptible de varier, ce qui lui confère également le nom de ***variable statistique*** (Cf, § 2.3.1.2 ci-dessous). Par exemple ; l'âge des étudiants de première année, la couleur des voitures dans un parking, le nombre de pièces par logement dans une ville, etc.

Sur une population statistique donnée, on peut étudier ou observer un ou plusieurs caractères.

3. Modalités

Notées (x_i), ce sont toutes les situations ou valeurs possibles du caractère. Tous les caractères présentent au moins deux modalités (deux ou plus), sinon l'étude statistique n'aurait pas de sens. Par contre, sur chaque individu on ne retrouve qu'une et une seule modalité (un individu ne peut pas avoir en même temps 18 ans et 20 ans à la fois) (Py, 1996).

Il y a autant de modalités que d'individus dans la population. Cependant, le nombre de modalités différentes ("k") est toujours inférieur ou égale au nombre d'individus : $k \leq N$.

Remarque : xi signifie la modalité numéro « i » (« i » est appelé " *indice* ", il désigne le numéro de la ligne), les modalités étant ordonnées par ordre croissant.

Ainsi :

- x_1 signifie la modalité numéro 1 ou la modalité la plus petite, elle est portée à la première ligne du tableau statistique.
- x_2 désigne la modalité numéro 2, celle portée à la deuxième ligne du tableau.
- x_k désigne la modalité numéro « k » ou la plus grande modalité, elle est portée à la dernière ligne du tableau.

- x_i désigne une quelconque modalité, (la numéro i), parmi toutes les modalités possibles du caractère portées sur le tableau statistique. Elle est portée à la ligne $k^{ème}$ ligne.

4. Les différents types de caractères

Un caractère peut être de type qualitatif (qualité) ou de type quantitatif (quantité).

4.1. Le caractère qualitatif

Un caractère est dit qualitatif lorsque ses modalités ne sont pas mesurables ou pas quantifiables. Ses modalités sont alors simplement constatées ou qualifiées, repérées par des mots ou des numéros (étiquettes, codes, numéros, …).

De même, les opérations arithmétiques ne sont pas possibles sur les modalités d'un caractère qualitatif, et débouchent sur des résultats irrationnels et vides de sens (Py, 1996).

Lorsque les modalités du caractère qualitatif ne reflètent pas un ordre de grandeur ou de hiérarchie, on dit qu'il s'agit d'un caractère qualitatif *nominal* (nom). Par exemple ; la couleur, l'origine géographique, nature des missions, etc.

Par contre, lorsque les modalités du caractère qualitatif reflètent un ordre de grandeur et peuvent être hiérarchisées, on dit qu'il s'agit d'un caractère qualitatif *ordinal* (ordre), c'est le cas des codes numériques, des adjectifs, des catégories,…. Par exemple, les dates de naissance, les numéros d'assurance, les numéros des salles, les catégories socioprofessionnelles, le stade d'avancement d'une maladie, la mention au bac, etc. (Py, 1996).

En pratique, les modalités d'un caractère qualitatif forment les différentes rubriques d'une nomenclature établie de telle sorte que chaque individu figure dans une et une seule rubrique (une et une seule modalité).

4.2. Le caractère quantitatif

On dit qu'un caractère est quantitatif lorsque ses modalités sont quantifiables ou mesurables, c'est-à-dire, reflètent une mesure ou une quantification. Ses modalités sont, par conséquent, toujours traduites par des données numériques (chiffres) sur lesquelles les opérations arithmétiques sont possibles et débouchent sur des résultats rationnels (Py, 1996).

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

Les caractères quantitatifs ou les variables statistiques sont de deux natures :

4.2.1. La variable statistique discrète

Lorsque les modalités d'une variable statistique reflètent un *dénombrement* ou un *comptage*, c'est-à-dire désignent *le nombre de* quelque chose, on dit qu'il s'agit d'une variable statistique *discrète* ou *discontinue*. Ses modalités sont exprimées alors par des nombres entiers ou isolés appartenant à l'ensemble des *nombres naturels* (Ñ), reflétant des réalités indivisibles. Exemple ; le nombre d'enfants par ménage, le nombre d'étudiants par salle, le nombre de pièces par logement, le nombre de SMS reçus au cours d'une période donnée, etc. Dans ce cas, les modalités 0,3 enfants ; 2,5 SMS ; 3,6 pièces, comme exemples, ne sont pas admises. Ce sont des valeurs qui n'appartiennent pas à l'ensemble des nombres naturels. (Py, 1996).

Remarque : En pratique, il peut arriver que le nombre de modalités soit trop important. Les données sont alors présentées ou résumées sous forme de classes et sont alors traitées comme des variables statistiques continues. Mais les résultats obtenus à partir des traitements gardent la même nature que la variable étudiée. Ex ; si la moyenne calculée est de 1,5 étudiants, on lira 1 à 2 étudiants ou entre 1 et 2 étudiants.

4.2.2. La variable statistique continue

Une variable statistique est dite continue lorsqu'elle prend ses valeurs dans l'ensemble des nombres réels (R), autrement dit, dans un intervalle infinis de valeurs, ou lorsque ses modalités sont présentées sous formes d'intervalles ou de classes (Boudia, 2008).

Ainsi, à l'exception du dénombrement, toutes les opérations qui consistent en la mesure, à savoir ; la pesée, le métrage, le chronométrage, le calcul, … ; représentent des caractères quantitatifs continus.

Même si dans ce cas les modalités sont exprimées sous forme de nombres entiers, on peut toutefois pousser les mesures à un nombre infini de décimales. Ainsi, un poids de 10Kg représente en réalité un poids entre 9,999…9 Kg et 10,000….01 Kg. De même, une taille de 170 cm représente en réalité une taille entre 169,9999…..9 et 170,000….01 cm, etc. ce sont donc en réalité, des valeurs qui n'appartiennent pas à l'ensemble des nombres naturels (N), elles sont donc par nature continues.

Section 3 : Notions d'effectif, fréquence et de distribution des fréquences

Cette section a pour objet de définir, respectivement, les notions d'effectif absolu ou fréquence absolue, fréquence relative ou effectif relatif et de distribution de fréquences.

1. Effectif absolu

On appelle effectif absolu, ou fréquence absolue, noté "n_i", le nombre de fois qu'une modalité est observée. Ou bien encore, le nombre d'individus de la population présentant la même modalité (x_i). On écrit : ($x_i \longrightarrow n_i$).

Ainsi :

- n_1 désigne le nombre d'individus présentant la modalité x_1.
- n_2 désigne le nombre d'individus présentant la modalité x_2.

La somme des effectifs absolus nous donne l'effectif total de la population étudiée (Py, 1996).

$$\boldsymbol{\Sigma n_i = n_1 + n_2 + n_3 + \ldots\ldots + n_K = N}$$

2. Fréquence relative

Appelée aussi effectif relatif et notée « f_i », elle représente la proportion ou le pourcentage de chaque effectif absolu « n_i » par rapport à l'effectif total « N ». C'est donc la proportion ou le pourcentage d'individus présentant la même modalité « x_i » (Boudia, 2008). On écrit alors :

$$\boldsymbol{f_i = n_i / N}$$

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

$$\boldsymbol{\Sigma f_i = f_1 + f_2 + \ldots\ldots + f_K = 1 \text{ ou } 100\%}$$

Avec : $\boldsymbol{0 \leq f_i \leq 1}$

3. Effectif cumulé et fréquence cumulée

Il désigne l'effectif ou la fréquence d'une modalité quantitative, augmenté des effectifs ou fréquences des modalités précédentes. Les modalités quantitatives étant toujours ordonnées par ordre croissant dans le tableau statistique.

Les effectifs ou fréquences cumulés se déterminent à partir du tableau statistique par *sommation* ou *cumul par ligne* des effectifs respectifs des modalités ordonnées.

On peut avoir des effectifs ou fréquences cumulés *croissants*, notés « Ni » ou « Fi », c'est-à-dire le cumul croît du premier effectif ou fréquence (porté à la première ligne du tableau) vers le dernier effectif ou fréquence jusqu'à avoir la somme totale des effectifs (portée à la dernière ligne du tableau). Autrement dit, le cumul croît du haut vers le bas du tableau (de n_1 jusqu'à N). On écrit alors :

$$N_i = N_{i-1} + n_i \quad \text{ou} \quad F_i = F_{i-1} + f_i$$

N_{i-1} et F_{i-1} étant respectivement l'effectif cumulé et la fréquence cumulée de la modalité ou de la ligne avant la modalité ou la ligne « i ».

On peut également avoir des effectifs ou fréquences cumulés *décroissants*, notés « $Ni\downarrow$ » ou « $Fi\downarrow$», c'est-à-dire le cumul décroit de l'effectif total (la somme des effectifs), porté à la première ligne du tableau, jusqu'au dernier effectif, porté à la dernière ligne du tableau (de N jusqu'à n_k). On écrit alors :

$$N\downarrow_i = N_{i-1}\downarrow - n_{i-1} \quad \text{ou} \quad F_i\downarrow = F_{i-1}\downarrow - f_{i-1}$$

$N_{i-1}\downarrow$ et $F_{i-1}\downarrow$ étant respectivement l'effectif cumulé et la fréquence cumulée décroissants de la modalité ou de la ligne avant la modalité ou la ligne « i ».

On note en général que :

- $N_1 = n_1$ ou $F_1 = f_1$
- $N_k = N$ ou $F_k = 1$
- $N_1\downarrow = N$ ou $F_1\downarrow = 1$
- $N_k\downarrow = n_k$ ou $F_k\downarrow = f_k$

Les effectifs ou fréquences cumulés présentent trois grandes utilités en statistique :

- ils reflètent, dans un ordre croissant des données, le classement, le numéro, la position ou le rang de chaque modalité, de la première (la plus petite) jusqu'à la dernière ou la $n^{ième}$ (la plus grande) ;
- les effectifs ou fréquences cumulés croissants permettent de répondre à la question : *quel est l'effectif ou la proportion d'individus qui ont* ***moins de*** *ou* ***au plus*** *…. ?*
- les effectifs ou fréquences cumulés décroissants permettent de répondre à la question : *quel est l'effectif ou la proportion d'individus qui ont* ***plus de*** *ou* ***au moins*** *…. ?* (Py, 1996).

4. Distribution des fréquences

Le paragraphe suivant permet, à partir d'un exemple concret de définir la notion de *distribution des fréquences*, et d'éclairer l'étudiant sur les différentes notions statistiques définies plus haut.

Une enquête auprès d'un groupe de 35 étudiants de première année LMD, concernant leur âge, a permis de collecter les données suivantes :

20	18	21	19	19	18	18
22	20	19	17	18	18	19
17	18	19	22	21	21	20
23	22	23	21	20	19	18
18	17	25	18	20	20	23

On détermine d'abord, la population statistique, l'unité, le caractère, les modalités et la nature du caractère :

Population : c'est l'ensemble des 35 étudiants du groupe enquêtés.
Unité : un étudiant.
Caractère : âge.
Modalités : 17 - 18 - 19 - 20 - 21 - 22 - 23 - 25. (il y a donc 8 modalités différentes: k=8).
Nature du caractère : variables statistique continue.

Il s'agit là des premiers éléments de l'enquête. Ce sont des données brutes, désordonnées et indéchiffrables. Pour rendre ces données lisibles et instructives, il faudrait transformer ces données en statistiques, en leur faisant subir un traitement par les méthodes de la statistique descriptive. Celle-ci, comme on l'a expliqué plus haut, consiste à ordonner, classer, résumer et présenter ces données de la manière la plus synthétique et la plus lisible possible. Nous allons donc procéder étape par étape.

- *Ordonner les données*

  Il s'agit de données numériques, donc la manière la plus logique de les ordonner c'est de suivre un ordre croissant des données. On obtient alors ce qu'on appelle une *série statistique ordonnée*, comme suit :

17- 17- 17- 18- 18- 18- 18- 18- 18- 18- 18- 18-
19- 19- 19- 19- 19- 19- 20- 20- 20- 20- 20- 20-
21- 21- 21- 21- 22- 22- 22- 23- 23- 23- 25.

Remarque

La modalité 17 est la première et plus petite modalité de la série statistique ordonnée, elle est généralement notée « X_{min} ». La modalité 25 est la dernière, ou $35^{ème}$ modalité, et aussi la plus grande valeur de la série ordonnée. Elle est généralement notée « X_{max} ».
La différence entre la plus grande et la plus petite valeurs de la série nous donne l'***étendue*** de la série, notée « e ». dans ce cas elle est égale à 25 – 17 = 8 ans. C'est l'écart d'âge ente le plus âgé des étudiants du groupe et le plus jeune.

La série présentée de cette manière est ordonnée mais non encore résumée. On peut donc la simplifier davantage et la rendre encore plus lisible. On va dans ce cas l'*individualiser*, c'est-à-dire supprimer les répétitions en faisant apparaître chaque modalité une seule fois, en lui associant le nombre d'étudiants correspondant (effectif).

- Classer et présenter les données

  En classant les données, on obtient ce qu'on appelle une *distribution statistique*, appelée aussi *série individualisée ou pondérée*, ou *distribution des fréquences (xi ⟶ ni)*.
  Cependant, une distribution statistique est présentée dans un tableau statistique, où ne sont représentés que le caractère étudié, ses modalités et les effectifs correspondant, comme suit :

Age (x_i)	**Nombre d'étudiants (n_i)**	**Fréquences relatives (f_i)**	**Effectifs cumulés (N_i)**	**Fréquences cumulées (F_i)**	**Effectifs cumulés décroissants (N_i ↓)**	**Fréquences cumulées décroissantes (F_i ↓)**
17	3	0,086	3	0,086	35	1
18	9	0,257	12	0,343	32	0,914
19	6	0,172	18	0,515	23	0,657
20	6	0,172	24	0,687	17	0,485
21	4	0,114	28	0,801	11	0,313
22	3	0,086	31	0,887	7	0,199
23	3	0,086	34	0,973	4	0,113
25	1	0,027	35	1	1	0,027
Total	35	1	-	-	-	-

Remarques :

- *Dans la colonne des fréquences relatives (f_i), on arrondis les chiffres à deux ou trois nombres après la virgule, de telle sorte à avoir au total 0,999 ou 1.*
- *Les colonnes des effectifs et fréquences cumulés ne contiennent pas de totaux car, s'agissant de cumuls, les totaux dans la dernière ligne du tableau n'ont pas de sens.*

On remarquera, à partir du tableau, que, comme nous l'avons souligné plus haut :

- $\Sigma fi = 1$
- $F_1 = f_1$
- $N_1 = n_1$
- $\Sigma n_i = N$
- $F_k\downarrow = f_k$ ou $N_k\downarrow = n_k$
- $F_1\downarrow = 1$ ou $N_1\downarrow = N$
- ΣN_i ou ΣF_i et $\Sigma N_i\downarrow$ ou $\Sigma F_i\downarrow$ n'ont pas de sens, ils ne signifient rien !

Notons, enfin, que la distribution des fréquences cumulées s'appelle *fonction de répartition* ou *fonction cumulative,* où à chaque modalité x_i est associé un effectif ou une fréquence cumulé :

$\{ x_i \longrightarrow N_i \}$ ou $\{ x_i \longrightarrow F_i \}$.

Conclusion au chapitre 1

Au terme de ce premier chapitre, l'étudiant est censé avoir pris connaissance du jargon ou vocabulaire usuel en statistique. La définition précise et détaillée des différentes notions fondamentales lui permettront, dans la suite du cours, de ne pas tomber dans la confusion et de bien cerner les contours des problèmes et applications statistiques qui lui seront posées.

Ce premier chapitre fait également l'objet d'une application en T.D (série n°1), où l'étudiant sera éclairé davantage sur la terminologie statistique.

Après ce préalable terminologique il est opportun de s'intéresser aux techniques de présentation des données statistiques résumées. C'est l'objet du chapitre suivant.

Chapitre 2 : Présentation des données dans des tableaux

Introduction au chapitre 2

L'objet du présent chapitre est d'apprendre à l'étudiant comment présenter les données statistiques collectées et ordonnées auparavant. Les tableaux les premières et l'une des principales formes de présentation des données privilégiées en Statistique. Cependant, ces derniers relèvent d'un certains nombre de règles et de normes que l'étudiant doit impérativement connaitre dans la perspective de son cursus universitaire.

La présentation des données sous forme de tableau statistique vise à présenter de manière simplifiée et claire les données pour le lecteur; tandis que les graphiques permettent un aperçu encore plus simple et surtout plus rapide du phénomène étudié (Py, 1996).

Dans le présent chapitre, nous examinons, successivement la structure usuelle d'un tableau statistique (section1). Puis en second lieu, nous présentons les différentes structures de tableaux statistiques en fonction du type de caractère (section 2). Pour finir, nous abordons la question des classes en statistique et les règles de leur construction (section 3).

Section 1 : Structure d'un tableau statistique

Une fois ordonnées et classifiées, les données sont présentées, de manière résumée, dans un tableau statistique. Par définition, un tableau statistique a pour objet, dans un souci de synthèse et de clarté, de montrer le *caractère étudié*, ses *modalités* (généralement présentées dans la première colonne du tableau) et les *effectifs et fréquences* correspondants : c'est-à-dire faire apparaitre les couples { xi ; ni } ou { xi ; fi }.(Hamdani, 2006).

1.1- Des données brutes au tableau statistique : le *tri à plat*

Dans certains cas, les données brutes collectées sont présentées dans un tableau élémentaire dans lequel les individus statistique sont identifiés individuellement chacun par sa modalité. C'est ce qu'on appelle le *tableau des données ponctuelles* ou *tableau des données élémentaires ou brutes,* qui n'est pas encore un tableau statistique à proprement dit. Si on reprend l'exemple du chapitre I, sur l'âge des 35 étudiants du groupe enquêtés, nous pouvons dire que les 35 chiffres (modalités) donnés représentent ce tableau de données élémentaires. En procédant, comme nous l'avons fait, à l'ordonnancement et à la classification (individualisation) de ces données, nous avons aboutit au tableau statistique proprement dit. On appelle cette opération du passage des données brutes au tableau statistique le *tri à plat* (Anderson *& al.,* 2006).

1.2- Forme usuelle d'un tableau statistique

Quelque soit le type de caractère et la nature de la variable, le tableau statistique se présente toujours de la même manière et suivant le même principe. La structure d'un tableau statistique de base se présente généralement comme suit :

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

(Unité de mesure)

Caractère (xi)	**Effectifs (ni)**	**Fréquences (fi)**	**Effectifs cumulés (Ni)**
x_1	n_1	f_1	N_1
x_2	n_2	f_2	N_2
.	.	.	.
.	.	.	.
x_i	n_i	f_i	n_i
.	.	.	.
.	.	.	.
x_k	n_k	f_k	n_k
Total	**N**	**1**	**-**

Source :

Tout tableau statistique établi dans le cadre d'une recherche scientifique doit obligatoirement être présenté selon la structure définie ci-dessus et doit contenir les indications suivantes (Hamdani, 2006) :

- *Intitulé ou titre du tableau*

  Il a pour rôle d'indiquer ce qui est représenté globalement par le tableau.

- *Le numéro du tableau*

  Dans tout travail de recherche, la présentation des tableaux doit être numérotée, ce qui facilite les renvois et facilite la lecture du manuscrit.

- *Intitulés des colonnes*

  Ils renseignent sur le contenu de chaque colonne, de telle sorte à clarifier encore davantage l'intitulé du tableau lui-même.

- *L'unité de mesure*

  Elle offre un renseignement supplémentaire en indiquant au lecteur dans quelle unité sont mesurées les modalités étudiées (kilogramme, milligramme, gramme,...).

- *La date*

  Elle renseigne le lecteur sur la période ou la date à laquelle sont collectées les données présentées dans le tableau. L'obligation de préciser la date des données, répond au souci que certaines données sont très variables dans le temps, la date permet d'apporter une certaine consistance et une certaine pertinence aux données présentées

par le chercheur. En méthodologie, cela reflète l'honnêteté intellectuelle du chercheur (ex ; on ne peut pas parler de la situation de la population algérienne en 2013, en s'appuyant sur des données de 1965 !).

- *La source*

  Elle indique d'où proviennent les données présentées. Cette indication obligatoire permet, selon le cas, de confirmer ou d'infirmer la pertinence de ces données. Elle reflète aussi l'honnêteté intellectuelle du chercheur. Par exemple, certaines sources sont réputées pour la non fiabilité de leurs données, des données constituées par le chercheur lui-même ne sont pas forcément acceptables par la communauté scientifique,…

Remarque :

L 'étudiant doit, dès maintenant, assimiler ces indications qu'il est appelé à appliquer à l'avenir, notamment lors de la rédaction de son mémoire de fin d'études.

Section 2 : Présentation des tableaux statistiques selon le type de caractère

Voici, ci-dessous, des exemples de tableaux statistiques en fonction du type de caractère et de la nature de la variable.

2.1 Tableau d'un caractère qualitatif

S'agissant d'un caractère qualitatif, il faut ordonner les modalités à l'intérieur du tableau soit (Py, 1996) :

- par ordre alphabétique s'il s'agit d'un caractère qualitatif nominal,

- par ordre d'importance ou hiérarchique s'il s'agit d'un caractère qualitatif ordinal, ce qui facilite encore davantage leur ordonnancement.

Tableau n° 1 : « Répartition des employés du groupe industriel «X» au 31-12-2012 »

Catégorie socioprofessionnelle	Nombre d'employés (Effectifs ou ni)	Fréquences (fi)
Cadre supérieur	5	0,035
Cadre moyen	15	0,105
Personnel administratif	30	0,210
Ouvrier professionnel	85	0,594
Agent d'entretien	8	0,056
Total	**143**	**1**

Source : Groupe industriel «X», Direction générale, Mars 2013.

On peut également codifier les modalités, suivant la nomenclature officielle en cours si elle existe comme suit (Py, 1996) :

Tableau n° 1 : « Répartition des employés de l'entreprise «X» au 31-12-2012 »

(millier d'employés)

Code (N°)	**Catégorie socioprofessionnelle**	**Nombre d'employés (Effectifs ou ni)**	**Fréquences (fi)**
01	Agent d'entretien	8	0,056
02	Ouvrier professionnel	85	0,594
03	Personnel Administratif	30	0,210
04	Cadre moyen	15	0,105
05	Cadre supérieur	5	0,035
-	**Total**	**143**	**1**

Source :

Le plus souvent aussi le tableau d'un caractère qualitatif ne comporte pas de colonne d'effectifs cumulés.

Comment faire des lectures à partir d'un tableau statistique d'un caractère qualitatif ?

Voici quelques exemples à partir du tableau n°1 :

- 3,5 *% des employés de l'entreprise X sont des cadres supérieurs (f_1 = 0,035 = 3,5%).*

- *La majorité (ou 59,4 %) des employés de l'entreprise X sont des ouvriers professionnels.*

- *14 % ((15 + 5)/143) des employés de l'entreprise X sont des cadres dont 3,5 % sont des cadres supérieurs.*

2.2- Tableau d'un caractère quantitatif

Selon qu'il s'agisse de variable statistique discrète ou de variable continue, le tableau diffère légèrement.

2.2.1- Tableau d'une variable statistique discrète

Le tableau d'une variable statistique discrète, où les données sont présentées individuellement ou de manière isolée, se présente comme indiqué dans le tableau n° 2, ci-dessous. A la différence du tableau précédent, dans ce cas on ajoute la colonne des effectifs ou fréquences cumulés car dans le cas d'un caractère quantitatif, cette colonne trouve tout son sens. Cependant, les modalités doivent impérativement être ordonnées par ordre croissant de la première ligne du tableau jusqu'à la dernière ligne.

Tableau n°2 : « Répartition des ménages de la cité X selon le nombre d'enfants » au 31-12- 2012»

Nombre d'enfants (xi)	Nombre de ménages (ni)	Fréquences (fi)	Effectifs cumulés (Ni)	Fréquences cumulées (Fi)	Effectifs cumulés décroissants (Ni ↓)	Fréquences cumulées décroissantes (Fi ↓)
0	80	0,040	80	0,040	2000	1
1	150	0,075	230	0,115	1920	0,960
2	430	0,215	660	0,330	1770	0,885
3	670	0,335	1330	0,665	1340	0,670
4	550	0,275	1880	0,940	670	0,335
5	100	0,050	1980	0,990	120	0,060
6 et plus	20	0,010	2000	1	20	0,010
Total	**2000**	**1**	**-**	**-**	**-**	**-**

Source : Enquête auprès des ménages de la cité X, Juin 2013.

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

- 660 ménages, soit 33 %, ont moins de 3 enfants, ou au plus 2 enfants.

- 1980 ménages, soit 99 %, ont moins de 6 enfants, ou au plus 5.

- 1340 ménages, 67%, ont plus de 2 enfants, ou au moins 3 enfants.

- 120 ménages, soit 6%, ont plus de 4 enfants, ou au moins 5 enfants.

2.2.2. Tableau d'une variable statistique continue

Le tableau d'une variable statistique continue, où les données sont présentées de manière groupée, sous forme de *classes* ou d'*intervalles*, se présente comme indiqué dans le tableau n° 3, ci-dessous. A la différence du tableau de la variable statistique discrète, dans ce cas on ajoute la colonne des « *centres de classes* » (xi). Cependant, les modalités doivent impérativement être ordonnées par ordre croissant de la borne inférieure de la première classe jusqu'à la borne supérieure de la dernière classe.

Tableau n° 3 : « Répartition des salaires des employés dans l'entreprise «X» au 31-12-2012 »

(Unité : 10^3 DZ)

Classes	Centres de classes (xi)	ni	fi	Ni	Fi	Ni ↓	Fi ↓
[20 ; 40[	30	42	0,420	42	0,420	100	1
[40 ; 60[	50	30	0,300	72	0,720	58	0,580
[60 ; 80[	70	18	0,180	90	0,900	28	0,280
[80 ; 100[	90	6	0,060	96	0,960	10	0,100
[100 ; 200[	150	4	0,040	100	1	4	0,040
Total	**-**	**100**	**1**	**-**	**-**	**-**	**-**

Source : Enquête auprès la Direction générale de l'entreprise «X», Avril 2013.

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

- 42 employés, soit 42%, touchent moins de 40.000DZ de salaire.
- 96 employés, 96%, touchent moins de 100.000DZ de salaire.
- 58 employés, 58%, touchent plus de 40.000DZ (ou 40.000DZ et plus).
- Tous les employés, soit 100%, touchent plus de 20.000DZ (ou 20.000DZ ou plus).

Remarque

Les tableaux présentés ici représentent la *structure de base* d'un tableau statistique. Cependant, en pratique, le nombre de colonnes à ajouter, au-delà de ce qui est présenté ici, dépend du phénomène étudié et des paramètres statistiques que le chercheur prévoit de déterminer.

Section 3 : Les classes et les règles de leur construction

En statistique la classe désigne un intervalle borné de modalités susceptibles d'être prises par « n_i » parmi les « N » individus de la population étudiée. Cet intervalle est délimité par des extrémités qu'on appelle « *bornes* » ou « *limites* » de classes selon le cas.

Par convention, en statistique, on considère des intervalles semi-ouverts ou ouverts à droite, afin d'éviter le chevauchement des classes.

La distance entre les deux extrémités de l'intervalle, mesurée par la différence « *borne supérieure moins la borne inférieure* », s'appelle l'***amplitude*** de classe et notée « $\mathbf{a_i}$ ».

En mathématiques, par définition, un intervalle contient un nombre infini de valeurs possibles. Aussi, en statistique, pour parer à cette contrainte, chaque classe est résumée ou représentée par son ***centre***, noté « $\boldsymbol{x_i}$ » et qu'on appelle aussi « ***centre de classe*** ».

Le centre de classe est la moyenne des deux extrémités de la classe :

$$\mathbf{Xi} = \frac{\textbf{Borne inférieure + Borne supérieure}}{\textbf{2}}$$

Selon que l'intervalle soit ouvert ou fermé, l'amplitude de classe se calcule différemment :

- Intervalle semi-ouvert : $[b_i ; b_{i+1}[$, alors $\mathbf{a_i = (b_{i+1} - b_i).}$
- Intervalle fermé : $[b_i ; b_{i+1}]$, alors ai = $\mathbf{(b_{i+1} - b_i) + 1}$ (Cette deuxième règle n'est cependant valable que lorsque les extrémités des classes sont des nombres entiers) . Ou bien, a_i est égale aux nombre de valeurs entières qu'il ya entre b_i et b_{i+1} ou de b_i jusqu'à b_{i+1}.

Ou bien encore ; à la différence entre la borne inférieure de la classe suivante (b_{i+2}) moins la borne inférieure de la classe, soit : $\mathbf{a_i = (b_{i+2} - b_i).}$

Le choix du nombre de classes, noté (**Z**), relève, en générale, des intensions et des compétences du chercheur. Cependant, il est recommandé en statistique de choisir un nombre raisonnable de classes, entre 5 et 20, et une amplitude de classes de préférence constante.

Le nombre de classes doit ainsi être assez grand pour refléter le phénomène étudié, et assez réduit pour permettre des lectures et des opérations aisées. Le mieux serait alors de prendre un nombre de classes entre 6 et 12.

Toutefois, il faut veiller à ce que la plus petite valeur (modalité) observée de la série appartienne à la première classe, et que la plus grande valeur (modalité) observée de la série appartienne à la dernière classe, ce qui éviterait de construire des classes inutiles ou fictives.

Il existe, en théorie statistique, plusieurs formules pour déterminer le nombre de classes (Boudia, 2008) :

- Formule de STURGE : Z = 1+3,33(logN).
- Formule de YULE : Z = 2,5 $\sqrt[4]{N}$.
- Formule la plus utilisée : Z = $\sqrt{N}$.

Remarque

Lorsque l'amplitude des classes est constante, on peut établir une relation entre le nombre de classes (**Z**), l'étendue de la série (**e**) et l'amplitude constante des classes ($\mathbf{a_i}$) comme suit :

$$\mathbf{a_i = e/Z \quad \Rightarrow \quad Z = e / a_i}$$

Le nombre d'individus composant une classe s'appelle effectif, noté n_i , déjà défini plus haut. Cependant, le nombre d'individus par *unité d'amplitude* (pour chaque $a_i = 1$) désigne la ***densité*** de la classe, notée « $\mathbf{d_i}$ », que l'on peut définir de deux manières :

- En termes absolus :

$$\mathbf{d_i = n_i / a_i}$$

- En termes relatifs :

$$\mathbf{d_i = f_i / a_i}$$

1.3.3.1- Notions de bornes, limites et limites réelles

Comme nous l'avons souligné plus haut, les constructions de classes utilisées en statistique sont généralement faites par convention ou de manière arbitraire, voire pas tout à fait réelle (c'est-à-dire simplifiée ou personnelle). Cependant, il arrive dans certaines études d'être contraint de construire des classes bornées par des valeurs réelles. Dans ce cas l'attention de l'étudiant est attirée par quelques renseignements supplémentaires et certaines règles de prudence qu'il doit observer.

Il ya en générale, en statistique descriptive, trois manières de présenter ou de borner les classes :

- Les bornes conventionnelles,
- Les limites,
- Les limites réelles.

a/- ***Les bornes conventionnelles***

Ce sont les bornes des intervalles *semi-ouverts* (ouverts à droite) qu'on utilise les plus fréquemment en statistique descriptive, en particulier en sciences humaines, notamment en économie. Ces bornes sont appelées conventionnelles car elles résultent du choix délibéré du chercheur qui construit ses classes, dans un but de synthèse et de simplification, et en fonction du problème étudié. Les classes étant personnellement choisies, le chercheur se simplifie la tache en prenant au passage des nombres entiers comme bornes et une amplitude de préférence constante.

b/- Les limites

Ce sont les bornes des intervalles *fermés* qu'on rencontre également en statistique. Cette manière de présenter est également arbitraire ou personnelle. Cependant, le fait de fermer les intervalles entraine systématiquement des pertes d'informations réelles qui, dans certains cas, peut être préjudiciable aux résultats de l'étude statistique. En effet, toutes les valeurs possibles situées entre la limite supérieure d'une classe et la limite inférieure de la classe suivante sont systématiquement perdues. Les limites supposent toujours un *degré d'erreur* systématique.

Il faut noter que les limites de classes sont des nombres de même nature que les bornes conventionnelles. C'est-à-dire, si les bornes conventionnelles sont des nombres entiers, les limites le seront également. Si les bornes conventionnelles sont des nombres décimaux avec un chiffre après la virgule, les limites seront également des décimaux avec un chiffre après la virgule, et ainsi de suite.

c/- Les limites réelles

Ce sont les bornes des intervalles *semi-ouverts* (ouverts à droite) qu'on est amené à utiliser parfois en statistique. Comme leur appellation l'indique, elles sont plus réelle que les limites et les bornes conventionnelles. Autrement dit, elles sont toujours plus précises que ces

dernières. Ainsi, si les limites et les bornes conventionnelles sont des nombres entiers, les limites réelles seront des décimaux à un chiffre après la virgule. Si les limites et les bornes conventionnelles sont des nombres décimaux à un chiffre après la virgule, les limites réelles seront des décimaux à deux chiffres après la virgule, ainsi de suite. C'est-à-dire que les limites réelles contiennent toujours un chiffre de plus après la virgule que les limites et les bornes conventionnelles. Ce chiffre de plus après la virgule implique plus de précision dans les limites réelles, d'où leur appellation limites réelles.

Il y a, cependant, une relation entre les trois types de bornes, c'est-à-dire qu'on peut passer d'une forme à une autre comme illustrer par l'exemple suivant :

Bornes conventionnelles	**Limites**	**limites réelles**
[10 ; 20[	[10 ; 19]	[9,5 ; 19,5[
[20 ; 30[	[20 ; 29]	[19,5 ; 29,5[
[30 ; 40[	[30 ; 39]	[29,5 ; 39,5[
[40 ; 50[	[40 ; 49]	[39,5 ; 49,5[

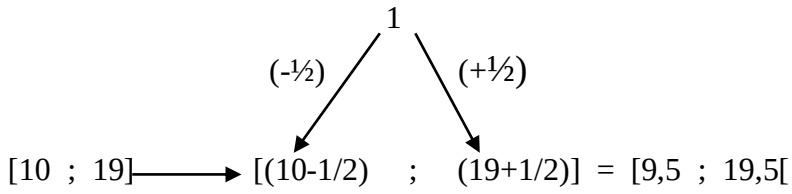
Dans cet exemple, on est passé des bornes conventionnelles vers les limites, puis, vers les limites réelles.

Nous avons des bornes conventionnelles formées par des nombres entiers, alors automatiquement, les limites seront aussi des entiers. Pour construire les limites, il faut fermer les intervalles tout en gardant des bornes avec des nombres entiers. Dans ce cas, pour fermer l'intervalle de la première classe, la limite supérieure de classe à choisir doit être un nombre entier immédiatement inférieure à 20 (20 ne faisant pas partie de la première classe puisque l'intervalle est ouvert à droite). Ensuite, la première classe étant fermée, la limite inférieure de la deuxième classe ne doit pas être la même que la limite supérieure de la première classe (la classe précédente). Donc la limite inférieure de la deuxième classe restera telle qu'elle 20. On applique le même raisonnement pour toutes les classes.

Ensuite, une fois les limites constituées, on détermine les limites réelles. Celles-ci doivent obligatoirement être plus précises que les limites et les bornes conventionnelles. Puisque nous avons des entiers, nous aurons donc des nombres décimaux à un chiffre après la virgule. On détermine ces derniers de deux manières comme suit :

- Soit on fait la moyenne de la limite supérieur d'une classe et de la limite inférieure de la classe suivante. (19 + 20)/2 = 19,5 ; (29 + 30)/2 = 29,5 ; …..

- Soit en récupérant le degré d'erreur causé par les limites. Dans cet exemple, en passant d'une classe à une autre on perd 1 (ex ; on passant de 19 à 20 on perd 1 ; en passant de 29 à 30 on perd 1 ; ….). Nous allons récupérer ce 1, en le répartissant équitablement (1/2 ; 1/2) de part et d'autre de l'intervalle, sachant que le sens positif va dans le sens croissant de la borne inférieure vers la borne supérieure. On aura :



Remarque :

En passant des bornes conventionnelles, aux limites et limites réelles, l'amplitude de classes ne change jamais et demeure constante.

[10 ; 20[⟶ [10 ; 19] ⟶ [9,5 ; 19,5[

Bornes conventionnelles — *Limites* — *limites réelles*

(*ai = 20 -10 = 10*) (*ai = (19 - 10) +1 = 10*) (*ai = 19,5 - 9,5 = 10*)

Conclusion au chapitre 2

A l'issue de ce deuxième chapitre, l'étudiant aura pris connaissance et est ainsi initié à la première des principales manières de présentation synthétique des données statistiques, à savoir les tableaux.

L'usage fréquent de ces derniers en statistique, tant durant le cursus universitaire que professionnel, impose à l'étudiant la nécessaire maitrise des techniques, principes et normes de leur construction. C'est dans cet esprit que le présent chapitre a été conçu.

Ce chapitre permet ainsi de rappeler l'étudiant la nécessité de savoir lire et interpréter les tableaux.

Au terme du présent chapitre, l'étudiant est désormais apte à réaliser seul les premières étapes de l'analyse statistique, de la collecte des données jusqu'à leur présentation sous forme de tableaux. Il doit, cependant, découvrir et s'initier à l'autre manière de présenter les données, à savoir les graphiques, c'est l'objet du chapitre suivant.

Chapitre 3 : Représentation graphique des données

Introduction au chapitre 3

La représentation graphique est l'image ou la représentation schématique du tableau statistique dont les couples $\{x_i ; n_i\}$. Elle permet, par un simple coup d'œil et de manière moins contraignante à la lecture que les tableaux, d'avoir un aperçu de l'allure générale de la distribution statique ou du phénomène étudié. On dit souvent chez les statisticiens qu' « *un beau graphique vaut mieux qu'un long discours* » (Py, 1996).
La représentation graphique, en statistique, repose sur le principe de la *proportionnalité des effectifs aux surfaces.* Tout graphique, en statistique, est caractérisé par :

- les coordonnées qu'il utilise (cartésiennes ou polaires),
- l'échelle retenue (arithmétique, logarithmique, ordinale,…),
- la nature du caractère étudié (Boudia, 2008).

Dans ce cours, nous utilisons essentiellement :

- Deux systèmes de coordonnées : cartésien, où le point M(x ; y) est représenté par le point M(xi ; ni), et polaire, où le point M(xi ; ni) est remplacé par l'angle α par rapport à l'axe horizontal et le rayon ou le vecteur $\overrightarrow{OM}$ par rapport au centre du cercle.

- Deux types d'échelles : arithmétique pour les caractères quantitatifs et ordinal pour les caractères qualitatifs.
- Deux types de caractères : déjà étudiés plus haut : qualitatif et quantitatif.

Remarque

Comme pour les tableaux statistiques, les graphiques doivent comporter, en plus du titre, de la source et de la date ; l'échelle de mesure indiquant ce que représente chaque centimètre sur le schéma et une légende qui définit toutes les nuances ou les couleurs utilisées dans les graphiques.

Section 1 : Représentation graphique d'un caractère qualitatif

Ils existe en pratique plusieurs types de graphiques pour représenter le caractère qualitatif : le diagramme circulaire, les diagrammes à barres ou tuyaux d'orgue, le cartogramme, le diagramme figuratif, etc. Cependant, pour des raisons de praticabilité, nous n'étudions ici que les plus courants et les plus simple à confectionner sur papier. Aussi, on étudiera le diagramme circulaire, les diagrammes à barres ou tuyaux d'orgue.

1.1. Le diagramme circulaire

Appelé aussi diagramme à secteurs ou camembert, ce graphique repose sur le système de coordonnées polaires. Le principe de sa représentation consiste (conformément au principe de proportionnalité des surfaces aux effectifs) à considérer que l'angle total formé par le cercle, soit 360°, correspond à l'effectif total (N) de la distribution à la somme des fréquences relatives, soit 1. On traduit alors la fréquence correspondant à chaque modalité par la mesure d'angle équivalente (α_i), ce qui nous permet d'avoir des parties de cercle ou des secteurs (ou portions).

L'angle (α_i) se calcule comme suit (Py, 1996) :

$$\alpha_i = (n_i / N) \cdot 360 = (f_i \cdot 360)°$$

Chaque secteur ou portion est ensuite singularisée par une couleur ou une nuance propre. Une légende, indiquant la signification de chaque nuance, facilitera la lecture du graphique.

Exemple 1

La répartition des employés d'une entreprise, au 31-12-2012, selon la catégorie socioprofessionnelle se présentait comme suit :

Catégorie socioprofessionnelle	**Nombre d'employés**
Ouvriers	50
Employés	25
Cadre moyens	20
Cadres supérieurs	5
Total	100

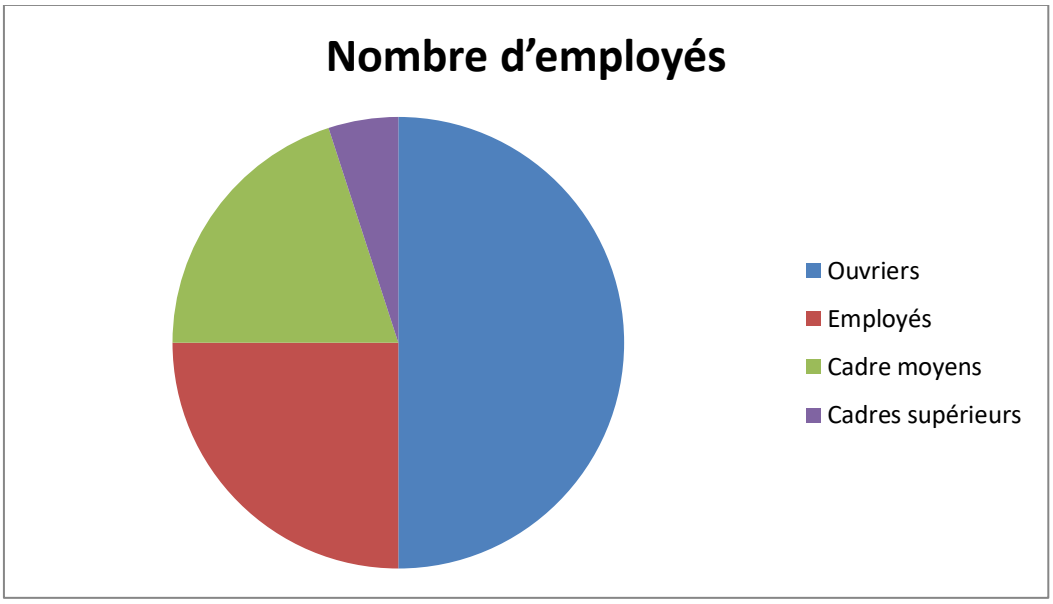
- Représenter graphiquement cette distribution statistique par un diagramme circulaire.

Réponse

Pour représenter graphiquement cette distribution, il faut d'abord compléter le tableau en calculant les fréquences et les angles (α_i), comme suit :

catégorie socioprofessionnelle	***Nombre d'employés (ni)***	***Fréquences (fi)***	***Angles (α_i°)***
Ouvrier	*50*	*0,50*	*180*
Employé	*25*	*0,25*	*90*
Cadres moyen	*20*	*0,20*	*72*
Cadres supérieur	*5*	*0,05*	*18*
Total	*100*	*1*	*360*

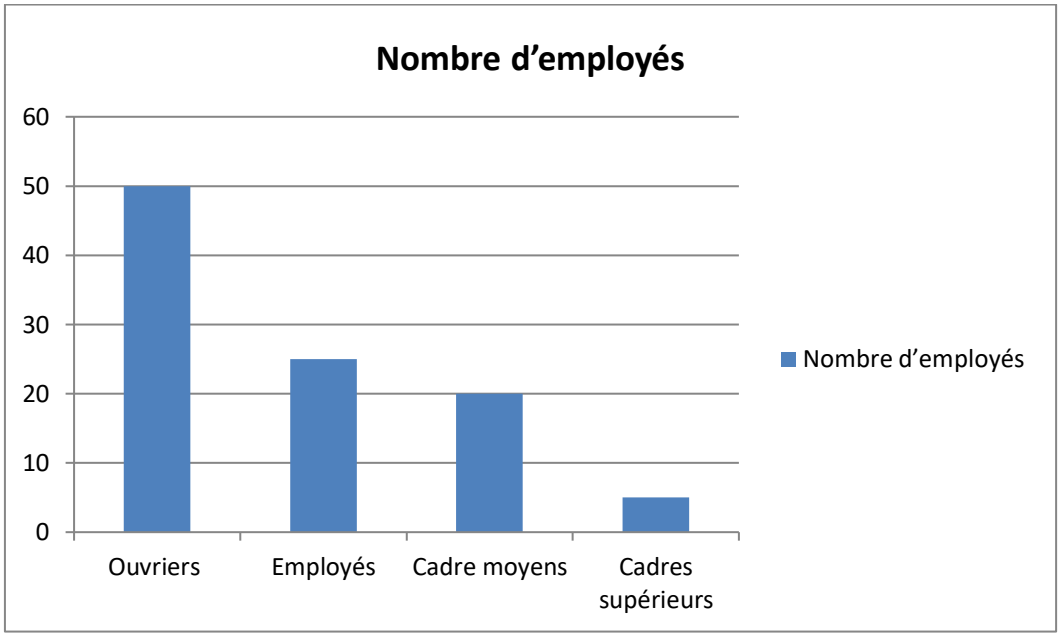
Représentation graphique



1.2. Le diagramme à barres ou tuyaux d'orgue

Ce type de graphique repose sur un système de coordonnées cartésiennes avec une échelle ordinale. Sur un plan orthonormé, on représente chaque modalité du caractère par une barre ou bande dont la hauteur, proportionnelle à l'effectif ou à la fréquence, est représentée par l'axe vertical yy', et la largeur indiquant la modalité, de dimension arbitraire puisqu'il s'agit d'un caractère qualitatif qui ne reflète pas une mesure, et de préférence constante pour l'esthétique, sur l'axe horizontal xx'.

Reprenant l'exemple 1 précédent, on aura graphiquement :



On remarquera que les bandes de ce diagramme peuvent être verticales

({xx' ⟶ modalités ; yy' ⟶ ni } ou horizontales, il suffit pour cela d'inverser le plan : {xx' ⟶ ni ; yy' ⟶ modalités}.

Section 2 : La représentation graphique d'une variable discrète

La représentation diffère selon qu'il s'agisse d'une variables statistique discrète (VSD) ou de variable statistique continue (VSC).

Dans ce cas, la fonction de distribution $\{x_i ; n_i\}$ ou {xi ; fi} est représentée par un *diagramme en bâtons* et la fonction de répartition ou cumulative {xi ; Ni} ou {xi ; Fi} est représentée par la *courbe cumulative en escaliers* (Py, 1996).

2.1. Le diagramme en bâtons

Ce type de graphique repose sur le système de coordonnées cartésiennes. Sur un plan orthonormé, on représente chaque modalité discrète par un *bâton* ou une ligne verticale, dont la hauteur est proportionnelle à son effectif ou sa fréquence. On porte ainsi, en abscisses les modalités « x_i » et, en ordonnées, les effectifs « ni » ou les fréquences « fi ».

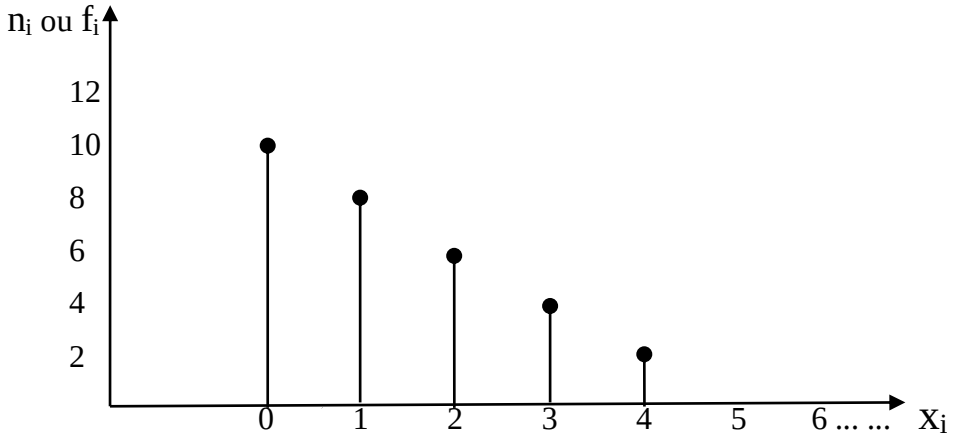
Exemple 2

Le tableau suivant donne la répartition du nombre de terminaux de connexion internet par bureau dans un établissement public.

Nombre de terminaux (x_i)	0	1	2	3	4	Total
Nombre de bureaux (n_i)	10	8	6	4	2	30

- Représenter graphiquement la distribution des fréquences.

Il s'agit dune variable statistique <u>*discrète*</u> puisque c'est un dénombrement (*le nombre de terminaux par bureau*). Ce type de variable est représenté par un <u>*diagramme en bâtons*</u>, comme suit (Py, 1996) :



« Diagramme en bâtons »

2.2. La courbe cumulative en escaliers

Ce graphique est utilisé pour représenter la fonction de répartition {xi ; Ni} ou {xi ; Fi}. Dans le cas d'un cumul croissant (Ni ou Fi), il représente la proportion ou l'effectif d'individus pour lesquels la valeur de la variable est strictement inférieure (<) à x_i (*Moins de*). Ces derniers apparaissent alors sous forme de paliers horizontaux, ouverts à gauche et fermés à droite, donnant à la courbe son allure en escaliers (Py, 1996).

La fonction cumulative est nulle pour toute valeur xi inférieure à la plus petite modalité observée, ce

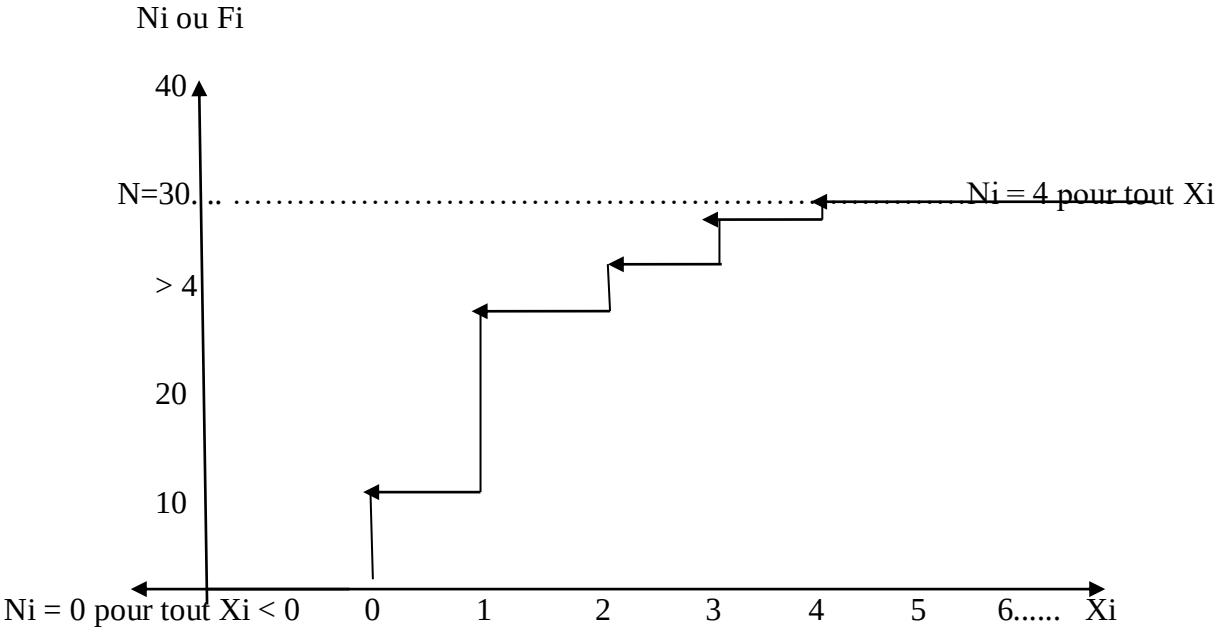
qui fait que le premier palier se confond toujours avec l'axe horizontal ou l'axe des abscisses, et reste ouvert vers $(-\infty)$.

La fonction cumulative est égale à 1 pour toute valeur supérieure à la plus grande modalité x_i observée, ce qui fait que le dernier palier a pour coordonnées Ni = N ou Fi = 1 et stagne à 1 ou N jusqu'à $+\infty$. (Py, 1996).

Exemple 3

Reprendre l'exemple 2a précédent, et représenter graphiquement sa fonction de répartition.

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



« Courbe cumulative en escaliers »

Section 3 : La représentation graphique d'une variable continue

Dans ce cas les données sont regroupées sous forme de classes. La fonction de distribution {xi ; ni } ou {xi ; fi} est représentée par un *histogramme* (à partir duquel on peut déduire le *polygone*), et la fonction cumulative {xi ; Ni} ou {xi ; Fi} est représentée par la *courbe cumulative en « S »*. (Hamdani, 2006).

3.1. L'histogramme

Ce type de graphique repose sur un système de coordonnées cartésiennes. Dans ce cas, chaque classe est représentée par un rectangle vertical dont la largeur représente l'amplitude de la classe, et la longueur l'effectif ou la fréquence de la classe.
C'est la surface de l'histogramme (des rectangles) qui intéresse le chercheur. Elle doit être proportionnelle aux effectifs. Cette proportionnalité se vérifie suivant deux situations :

A/- Amplitude de classe constante

Dans ce cas tous les rectangles ont la même largeur ($a_i = C^{te}$). Donc les surfaces sont proportionnelles aux seules longueurs (n_i ou f_i). Alors, l'histogramme peut être tracé directement avec les n_i ou les f_i. (Hamdani, 2006).

Exemple 3

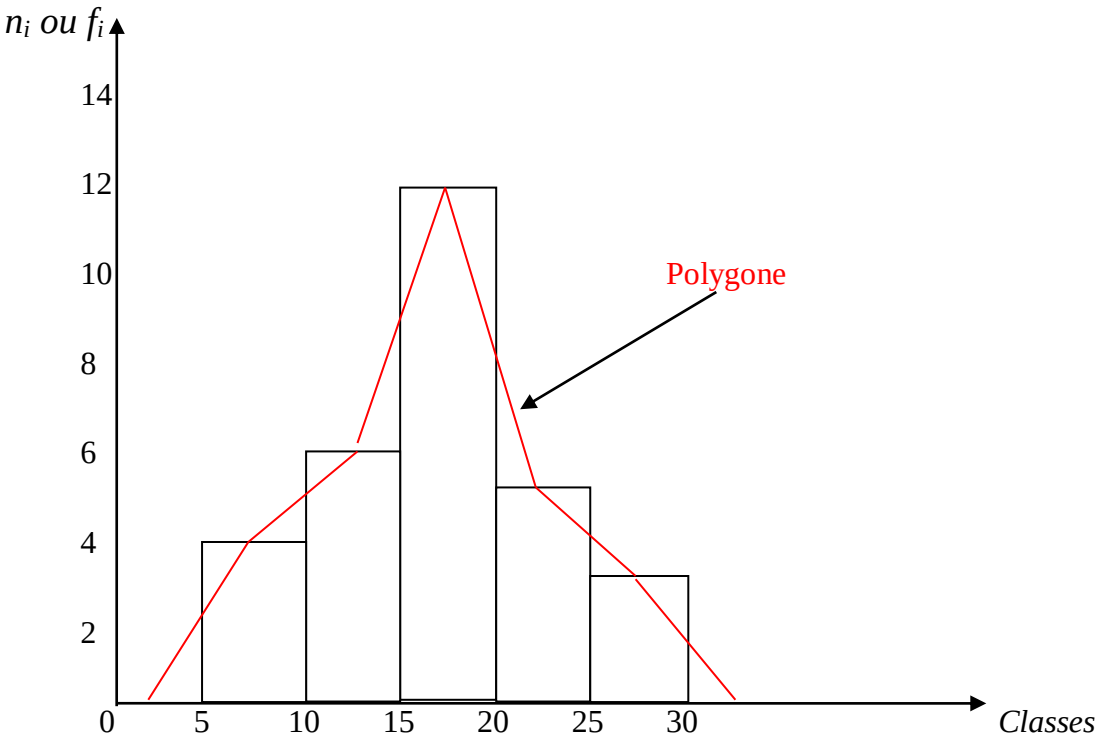
Soit la distribution suivante des salaires horaires (en euro) dans une entreprise donnée :

Salaire horaire	[5 ; 10[	[10 ; 15[	[15 ; 20[	[20 ; 25[	[25 ; 30[	Total
Nombre d'employés	4	6	12	5	3	30

- Représenter graphiquement cette distribution.

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

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



« Histogramme »

B/- Amplitude de classe non constante

Dans ce cas la surface des rectangles n'est pas proportionnelle aux seuls effectifs ou fréquences (longueurs), mais aussi aux amplitudes de classes (largeurs). Il faudrait dans ce cas corriger cet handicap. L'opération qui permet de le corriger s'appelle la correction des effectifs. En fait, il s'agit de rendre les surfaces proportionnelles aux *densités*. Ces densités se conçoivent de deux manières (Py, 1996) :

- soit on ramène chaque effectif à l'amplitude de la classe correspondant, c'est-à-dire aux densités que nous avons définis plus haut ($d_i = ni/a_i$ ou $d_i = f_i/a_i$),
- soit on ramène tous les effectifs (ou fréquences) de classes à une *même* et *commune* amplitude, appelée *amplitude de base*, notée « a_0 ». Cette opération s'appelle la "*correction des effectifs*" (étant donnée sa complexité, elle sera développée en cours magistral et en T.D). On obtient alors une autre forme de densité appelée effectifs ou fréquences corrigés, notés « n_{ic} » ou « f_{ic} ».

Exemple 4

Soit la distribution suivante :

Salaire horaire	[5 ; 10[	[10 ; 15[	[15 ; 20[	[20 ; 30[	Total
Nombre d'employés	4	6	12	8	30

- Représenter graphiquement cette distribution.

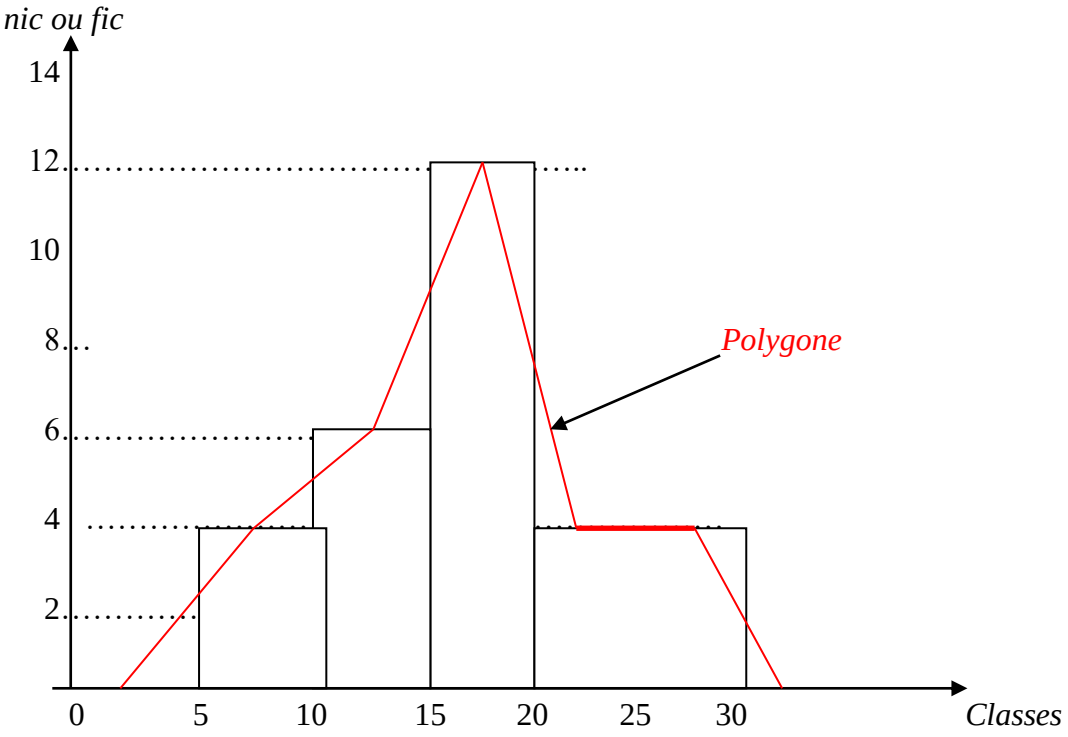
Il s'agit d'une distribution sous forme de classes (forme continue), donc la représentation graphique correspondante est l'histogramme. Cependant, avant de tracer celui-ci, il faut d'abord vérifier les amplitudes de classes si elle sont constantes ou pas, c'est-à-dire vérifier le principe de la proportionnalité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

On remarque dans cet exemple que l'amplitude de classe n'est constante : ex ; les première, deuxième et troisième classes ont une même amplitude ($a_i = 5$), alors que la quatrième classe a une amplitude ai = 10. Autrement dit, $a_i \neq C^{te}$. Il faudrait donc, avant de tracer l'histogramme, corriger les effectifs ou calculer les densités.

Les densités (ni/ai ou fi/ai) sont faciles à calculer et sont données dans le tableau. Par contre, pour les effectifs corrigés, on doit les calculer. On précisera que l'amplitude de base « a_0 », qui est la plus petite amplitude de classe, est $a_0 = 5$. On aura donc le tableau suivant :

Classes	x_i	n_i	a_i	N_i	$N_i\downarrow$	d_i	a_i/a_0	n_{ic}
[5 ; 10[	7,5	4	5	4	30	08	1	4
[10 ; 15[	12,5	6	5	10	26	12	1	6
[15 ; 20[	17,5	12	5	22	20	24	1	12
[20 ; 30[	25	8	10	30	8	08	2	4
Total	-	**30**	-	-	-	-	-	-

Notre histogramme sera donc comme suit :



Remarque

Etant donné le principe de proportionnalité des effectifs aux surfaces des rectangles, on en déduit que la surface totale de l'histogramme (la somme des surfaces des rectangles)

est égale à la somme des ni, soit N, ou à la somme des fi, soit 1.

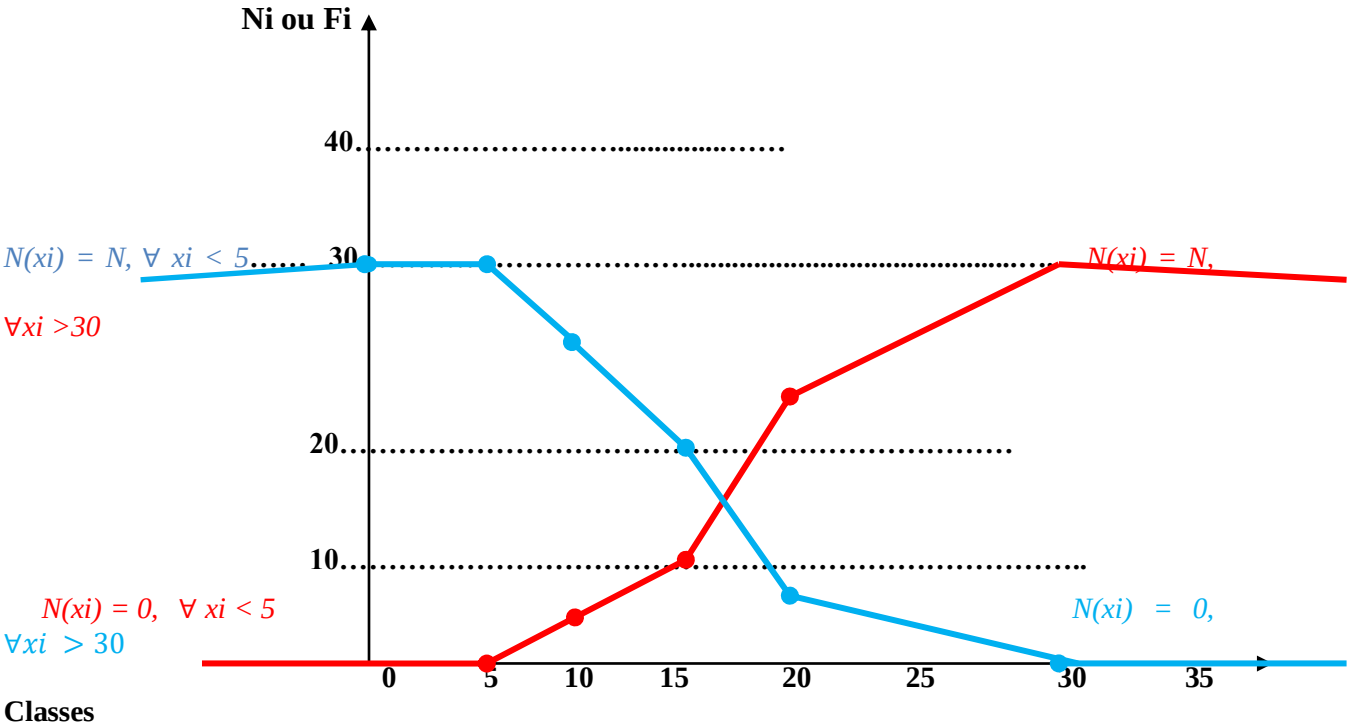
3.2. Le polygone

Parfois l'histogramme, aussi parfait soit-il, ne permet pas des comparaisons entre différentes distributions. Aussi, dans ce cas, on a recours au *« polygone »*, construit à partir de l'histogramme lui-même (*Cf. exemples 3 et 4, supra*). Le polygone est un ensemble de segments qui relient les centres des sommets des rectangles, à distance égale à « $\mathbf{a_i/2}$ » si l'amplitude de classe est constante, et à « $\mathbf{a_0/2}$ » si l'amplitude de classe n'est pas constante ; tout en ajoutant deux classes fictives l'une avant la première classe, l'autre après la dernière classe.
Le polygone nous donne une courbe continue dont la surface délimitée avec l'axe des abscisses représente la même surface que celle de l'histogramme, celle-ci étant égale à N ou 1, et ce, conformément au principe de *compensation des aires* (ce dernier sera explicité en cous magistral).

3.3- Les courbes cumulatives en « *S* »

La fonction de répartition d'une variable statistique continue est représentée par deux courbes cumulatives en « s ». L'une croissante, pour les fréquences ou effectifs cumulés croissants, l'autre décroissante, pour les fréquences ou effectifs cumulés décroissants. Les deux courbes sont représentées sur un même graphique (même plan). La courbe croissante relie les bornes supérieures des classes (Moins de), la courbe décroissante relie les bornes inférieures des classes (Plus de) (Py, 1996).

On peut reprendre l'exemple 4 précédent, dont nous avons déjà calculé les effectifs cumulés, et tracer sa courbe cumulative.



Remarques

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

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

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

Conclusion au chapitre 3

A l'issue de ce troisième chapitre, l'étudiant aura pris connaissance et est ainsi initié aux deux principales manières de présentation synthétiques de données statistiques, à savoir les tableaux et les graphiques.

L'usage fréquent de ces derniers en statistique, tant durant le cursus universitaire que professionnel, impose à l'étudiant la nécessaire maitrise des techniques, principes et normes de leur construction. C'est dans cet esprit que s'inscrit le présent chapitre a été conçu.

Au-delà de la conception des graphiques, ce chapitre visait également à rappeler à l'étudiant la nécessité de savoir lire et interpréter ces représentations.

Ainsi, au terme du présent chapitre, l'étudiant est désormais apte à réaliser seul les quatre premières étapes de l'analyse statistique. Il doit, cependant, découvrir et s'initier à la cinquième et dernière étape, à savoir ; le calcul des paramètres pertinents et leur interprétation. L'objet du chapitre suivant ce sont les paramètres de tendance centrale.

Chapitre 4 : Les paramètres de tendance centrale

Introduction au chapitre 4

Le chapitre précédent nous a permis de comprendre comment faire la synthèse des données statistiques et comment les présenter de manière lisible et compréhensible par un large éventail de lecteurs. Cependant, la synthèse et la présentation des données statistiques ne s'arrête pas uniquement à ce qui est développé dans le dit chapitre. Au-delà, le statisticien ou le chercheur dispose encore d'une multitude de techniques de synthèse qui permettent de renseigner sur le phénomène étudié, à travers la caractérisation de la série. Cette caractérisation, consiste à représenter la série statistique par des chiffres illustratifs ou représentatifs, appelés *paramètres*, que le chercheur calcule lui-même, selon la nature du phénomène qu'il veut mettre en évidence.

Il existe en statistique descriptive quatre grands groupes de paramètres : *les paramètres de tendance centrale, de dispersion, de forme et les paramètres de concentration.*

L'objet de ce troisième chapitre est l'étude des *paramètres de tendance centrale*. Nous y étudions successivement : le *mode* (section 1), la *médiane* (section2), la *moyenne arithmétique* (section 3) et les *autres types de moyennes* (section 4).

Section 1 : Le Mode

1. Définition

Le Mode d'une série statistique est la modalité dominante ou la plus fréquente, c'est-à-dire la modalité qui correspond au plus grand effectif ou plus grande fréquence. Il est noté « **Mo** ».

2. Détermination pratique

Le Mode se détermine de manières *algébrique* et *graphique*, et la détermination diffère selon qu'il s'agisse de variable statistique discrète (VSD) ou variable statistique continue (VSC) (Py, 1996).

2.1- Variable statistique discrète (VSD)

- *De manière algébrique* :

Il s'agit de repérer dans la colonne (ni) ou (fi) du tableau statistique l'effectif le plus élevé (ou le chiffre le plus élevé), la modalité (xi) (dans la colonne xi) correspondant à celui-ci est le mode.

***Exemple 1* :**

Soit la distribution suivante :

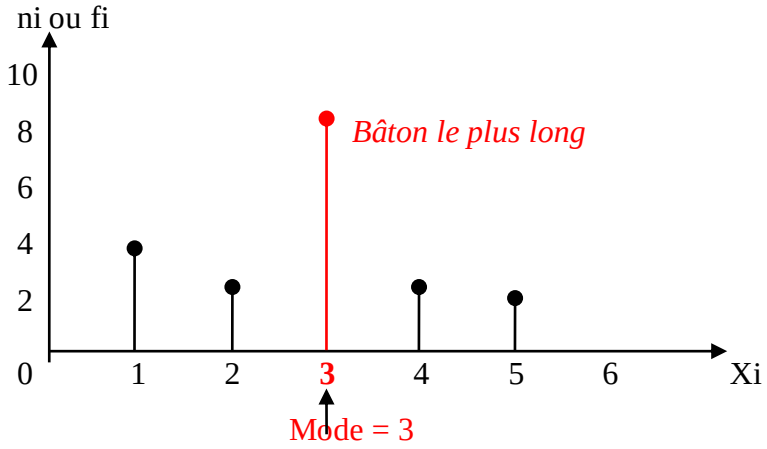
x_i	1	2	3	4	5	Total
n_i	4	3	8	3	2	20

- *Calculer le mode de cette distribution.*

Il suffit de regarder la colonne (ni), le plus grand effectif c'est ni = 8, la modalité correspondante dans la colonne (xi) c'est 3. Donc le Mode = 3.

- *De manière graphique* :

Il s'agit d'une variable statistique discrète, donc le graphique correspondant est le *diagramme en bâtons* . Le principe est de repérer le bâton le plus long (dont la hauteur indique l'effectif), la valeur (xi) correspondante sur l'axe horizontal (l'axe xx') est le mode. On aura alors le graphique suivant :



2.2- Variable statistique continue (VSC)

Dans ce cas les données sont regroupées sous forme de classes. La valeur (xi) correspondant au mode appartient forcément à une classe. Cette classe s'appelle *classe modale*. Elle est repérée comme la classe « ***la plus dense*** », c'est à dire celle correspondant à la plus grande densité (d_i).

Deux situations peuvent se présenter (Boudia, 2008) :

2.2.1- Amplitude des classes constante

Dans ce cas la classe la plus dense est celle aussi correspondant au plus grand effectif, car dans ce cas di = ni/ai revient à diviser tous les effectifs par la même amplitude, donc la classe qui a le plus grand effectif sera aussi la plus dense.

- *De manière algébrique* : la détermination du mode suit les étapes suivantes :
  - repérer dans la colonne (n_i) ou (f_i) le plus grand effectif ou fréquence,
  - repérer la classe modale correspondant à ce plus grand effectif ou fréquence,
  - appliquer la formule du mode réservée au cas continue comme suit (Py, 1996) :

$$\mathbf{Mo = X_o + a \left[\frac{(n_{mo} - n_{mo-1})}{(n_{mo} - n_{mo-1}) + (n_{mo} - n_{mo+1})} \right]} \quad \text{.....(1)}$$

Avec :

- X_o = borne inférieure de la classe modale.
- a = amplitude la classe modale ou des classes puisque a_i est constante.
- n_{mo} = effectif ou fréquence (f_{Mo}) de la classe modale.
- $n_{mo\ -1}$ = effectif ou fréquence (f_{Mo+1}) de la classe avant ou précédant la classe modale.
- $n_{mo+}1$ = effectif ou fréquence (f_{Mo+1})de la classe après ou suivant la classe modale.

Exemple 2a *: Calculer le mode de la distribution*

Classes	10 - 20	20 - 30	30 - 40	40 - 50
Effectifs	*10*	*10*	*15*	*5*

Réponse :

Il faut d'abord, au préalable, calculer les amplitudes et voir si elles sont constantes ou pas. On ajoute alors une colonne (ai) au tableau, on obtient :

Classes	ni	fi	ai
10 - 20	10	0,250	10
20 - 30	10	0,250	10
30 - 40	15	0,375	10
40 - 50	5	0,125	10
Total	40	1	-

On constate donc que l'amplitude (ai) est constante (ai =10). Par conséquent, la classe modale est celle qui correspond au plus grand effectif. Si on regarde dans la colonne des (ni), la plus grande valeur c'est 15, (ni = 15). La classe correspondant (c'est-à-dire sur la même ligne) c'est [30 - 40[, c'est la classe modale, donc $\Longrightarrow$ Xo = 30.

On applique la formule du mode développée précédemment ((1) ou (2)):

$$Mo = Xo + a\left[\frac{(n_{mo} - n_{mo-1})}{(n_{mo} - n_{mo-1}) + (n_{mo} - n_{mo+1})}\right] \quad \text{……..(1)}$$

$$Mo = 30 + 10\left[\frac{(15 - 10)}{(15 - 10) + (15 - 5)}\right] = 33{,}33 \longrightarrow \underline{Mo = 33{,}33}$$

Ou bien on peut aussi le calculer en utilisant les fréquences.

2.2.2- Amplitude des classes non constante (ai ≠ C^te)

Dans ce cas la classe la plus dense n'est pas forcément celle qui correspond au plus grand effectif. En effet, di = ni/ai, (ai≠ C^te), signifie que l'on divise les effectifs ou fréquences par des valeurs de ai différentes, ce qui implique que l'on peut tomber sur des situations où la classe correspondant au plus grand effectif ne soit pas aussi la classe la plus dense (Py, 1996).
Aussi, avant de calculer le Mode il faut calculer les densités (ou calculer les effectifs corrigés n_{ic}) que utilisera dans la formule du mode précédente à la place des ni ou fi. La plus grande densité nous donnera la classe modale, et on suivra les mêmes étapes que précédemment.

Exemple 2b *(Amplitude de classe non constante)*

Calculer le mode de la distribution

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

Réponse :

Pour calculer le mode, il faut d'abord vérifier les amplitudes (ai). On construit comme précédemment une colonne (ai) de laquelle dépendra le nombre de colonnes à ajouter au tableau, selon que ai soit C^{te} ou pas.

Classes	**Effectifs**	$\mathbf{a_i}$	$\mathbf{d_i}$	$\mathbf{a_i/a_o}$	$\mathbf{n_{ic}}$
10 - 20	10	10	1	1	10
20 - 30	20	10	2	1	20
30 - 50	30	20	1,5	2	15
50 - 80	25	30	0,83	3	8,33
Total	85	-	-	-	-

Si on regarde la colonne (a_i), on constate que l'amplitude de classe n'est pas constante. La classe modale est donc la classe la plus dense, c'est-à-dire celle qui a la plus grande densité ou le plus grand effectif corrigé, et qui n'est pas forcément celle qui a le plus grand effectif. Par conséquent, avant de calculer le mode il faut corriger les effectifs, c'est-à-dire ; soit calculer les densités (d_i), soit calculer les effectifs corrigés (n_{ic}) (Cf, chapitre 2).

L'amplitude de base (la plus petite amplitude), d'après le tableau est $a_o = 10$. On en déduit les (n_{ic}). Ou bien, on calcule directement les densités (d_i).

Dans la colonne (di), la plus grande valeur c'est 2. Sur la même ligne, la plus grande valeur correspondante dans la colonne (n_{ic}) doit <u>automatiquement</u> être la plus grande valeur de la colonne (n_{ic}). Si ce n'est pas le cas, cela voudrait dire qu'il y a erreur de calcul quelque part que l'étudiant doit vérifier et corriger.

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

Donc Mo sera :

$$Mo = Xo + a\left[\frac{(n_{icmo} - n_{icmo-1})}{(n_{icmo} - n_{icmo-1}) + (n_{icmo} - n_{icmo+1})}\right] \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(3)$$

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

Ou bien ;

$$Mo = Xo + a\left[\frac{(d_{mo} - d_{mo-1})}{(d_{mo} - d_{mo-1}) + (d_{mo} - d_{mo+1})}\right] \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..(5)$$

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

De manière graphique

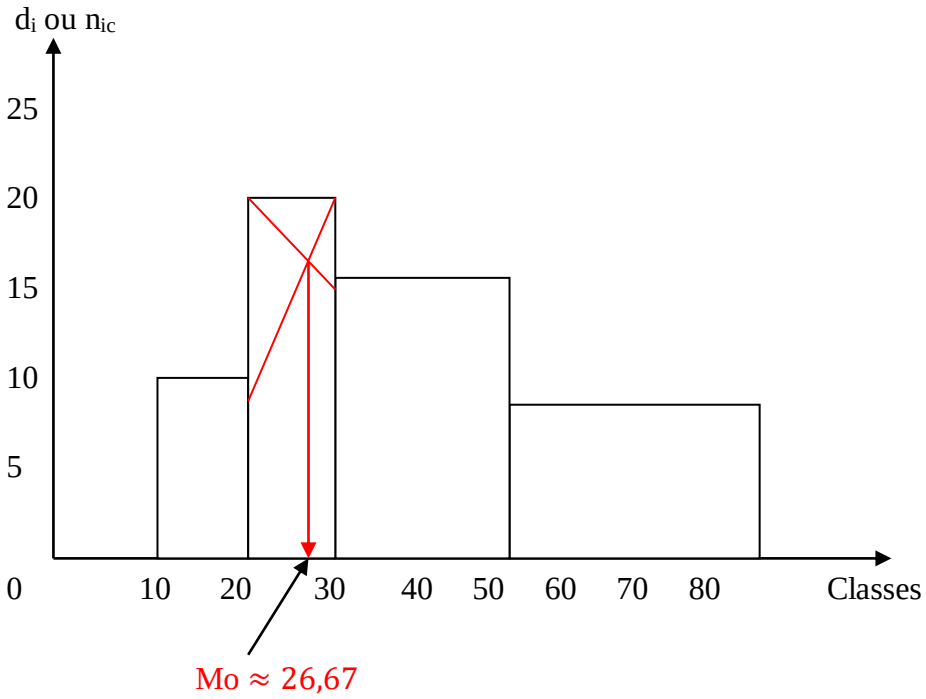
Il s'agit d'une VSC, donc la fonction de distribution est représentée graphiquement par l'histogramme. Compte tenu de l'amplitude de classe (constante ou pas), le mode se détermine de la même manière à partir du rectangle de la classe modale (la plus dense) dans l'histogramme.

Il suffit pour cela de relier par une ligne (ou un segment) la borne supérieure du rectangle de la classe modale à la borne supérieure du rectangle de la classe précédente. Ensuite, relier par une autre ligne ou segment la borne inférieure du rectangle de la classe modale à la borne inférieure de la classe suivante. Le point de croisement des deux lignes ou deux segments est ensuite projeté sur l'axe des abscisses (l'axes des classes ou horizontal). La valeur (xi) correspondante sur cet axe est le mode. Voyons cela avec l'exemple 2c ci-dessous.

Exemple 2c

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

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



Remarque

- Le mode satisfait aux 1ère, 3ème et 4ème conditions de YULE mais pas aux autres conditions.

- Son principal avantage est d'être déterminé très rapidement et sa signification est simple. On l'utilise surtout dans le cas des variables statistiques discrètes et pour faire apparaitre facilement une première estimation de valeur centrale de la distribution.

- Une série statistique peut présenter un, deux ou plusieurs modes à la fois, comme elle peut ne pas en présenter du tout (ni = C^{te}), on l'appelle alors une distribution ou série « *amodale* ».

Section 2 : La Médiane

2.1- Définition

Etymologiquement[1], médiane signifie « *milieu* ». C'est le milieu de la distribution ordonnée. Notée « **Me** », elle est la modalité (xi) qui est située dans la série ordonnée, de telle sorte que le nombre de modalités situées avant la médiane (ou inférieures à la médiane) est le même que celui des modalités situées après la médiane (ou supérieure à la médiane). Autrement dit, la médiane est la modalité qui partage l'effectif total des modalités en deux groupes de même taille(50% chacun).

C'est donc la valeur (x_i) pour laquelle 50% des individus présentent des modalités inférieures à Me, et 50% des individus restant présentent des modalités supérieures à Me. Les modalités étant, bien entendu, obligatoirement ordonnées par ordre croissant. La médiane est donc aussi un *paramètre de position* (Hurlin & Mignon, 2018).

2.2- Détermination pratique

La médiane se détermine de deux manières ; algébrique et graphique, et la détermination diffère selon qu'il s'agisse de VSD ou VSC (Mazerolle, 2006).

2.2.1- Variable statistique discrète

2.2.1.1- De manière algébrique

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

A/- Si N est un nombre impair

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

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

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

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

Exemple 1a

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

Il y a sept (7) modalités ou individus dans la série, soit N= 7.

7 étant un nombre impair, il s'écrit donc N = 2k+1 = 7 ⟶ k = N – 1 / 2 = 3

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

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

B/- Si N est un nombre pair

Dans ce cas il existe non pas une modalité médiane mais un *intervalle médian* délimité par

[1] Du point de vue de l'origine ou des racines du mot.

[$k^{ème}$ et $(k+1)^{ème}$] modalités.

Exemple 1b

Soit la série ordonnée suivante : 3 - 6 - 12 - 18 - 20 - 23 - 28 - 30.

Déterminer la médiane de cette série.

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

N peut s'écrire mathématiquement : N = 2k = 8.

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

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

[18 ; 20[, soit Me ≈ (18 + 20) = 19.

Remarque

Dans le cas où N est trop élevé, on a recours alors au tableau statistique, où la colonne (Ni) nous permet de repérer les positions des modalités qu'on cherche, en suivant la même logique.

Exemple 1c

Déterminer la médiane de la distribution suivante :

xi	ni	Ni
0	10	10
1	32	42
2	36	78
3	15	93
4	5	98
5	2	100
Total	**100**	**-**

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

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

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

K étant un effectif cumulé (Ni), on peut lire dans le tableau que $50^{ème}$ et $51^{ème}$ se situent entre les effectifs cumulés N_2 = 42 N_3 = 78, c'est-à-dire entre la $42^{ème}$ et la $78^{ème}$ modalités. Or, d'après le tableau on peut lire que de la $43^{ème}$ jusqu'à la $78^{ème}$ ce sont toutes des modalités égale à 2. Autrement dit, les deux modalités que nous cherchons (la $50^{ème}$ et $51^{ème}$) font partie de ces modalités puisqu'elles sont situées entre la $43^{ème}$ et $78^{ème}$ modalité. Donc notre intervalle médian est [2 ; 2], ce qui donne Me = 2 !

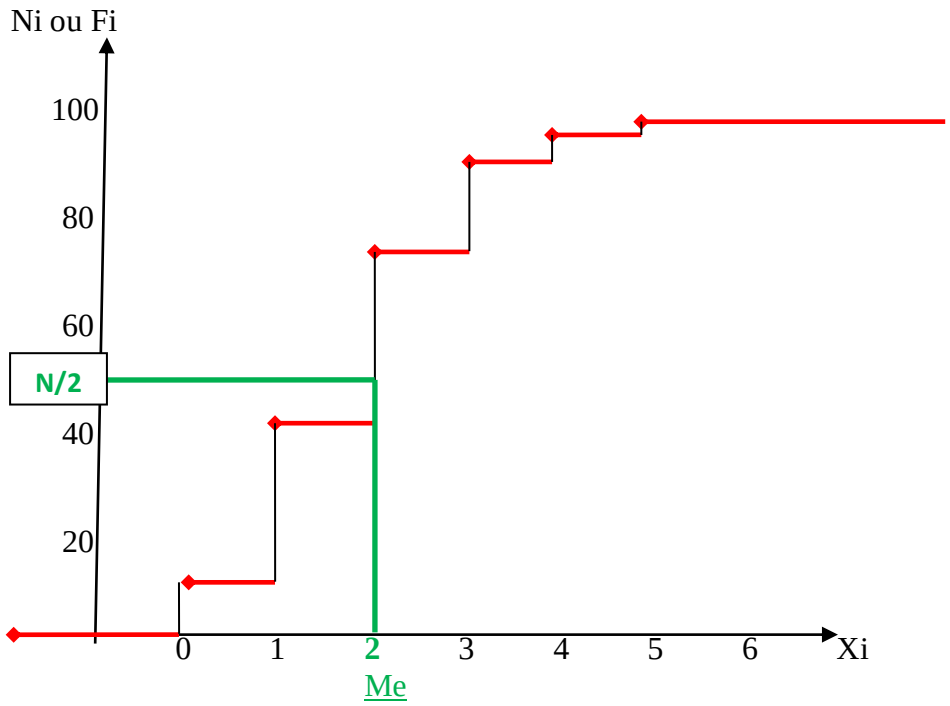
Ou bien de manière plus simple encore, il suffit de calculer N/2, celui-ci nous donne l'effectif cumulé ou la position de la modalité médiane. Si cette position se situe entre deux effectifs cumulés ou entre deux ligne de la colonne Ni, on prend directement la ligne du bas, c'est-à-dire la valeur immédiatement supérieure ou égale (≥) à (N/2).

C'est ce même principe qu'on utilise pour la détermination graphique.

2.2.1.2- Détermination graphique

Dans le cas d'une VSD la fonctions cumulative, {xi ; Ni} ou {xi ; Fi} est représentée graphiquement par la *courbe cumulative en escalier*. La médiane se déterminant par les effectifs cumulés, elle est donc logiquement déterminée graphiquement par la courbe représentant ces derniers.

On peut illustrer la méthode de détermination à partir de la courbe cumulative de l'exemple 1c précédent.



Sur l'axe vertical (axe des Ni ou Fi), repérer la valeur correspondant à N/2. On projette, ensuite, ce point à l'horizontal parallèlement à l'axe des Xi, sur la courbe en escaliers. La première contremarche sur laquelle on tombe est directement projeté par une ligne verticale sur l'axe des Xi, la valeur sur laquelle on tombe est la médiane (Mazerolle, 2006).

NB/- Si on tombe sur un palier au lieu d'une contremarche, cela voudrait dire que nous avons à faire à un intervalle médian délimité par deux valeurs différentes, ce qui n'est le cas de l'exemple1c.

2.2.2- Variable statistique continue

Dans ce cas, les données sont présentées sous forme de classes, la modalité médiane appartient forcément à une classe, appelée intervalle ou *classe médiane*, à partir de laquelle il faudrait la déterminer. (Mazerolle, 2006).

Comme pour la VSD, la détermination de la médiane pour une VSC se fait également de deux manières possibles : algébrique et graphique.

2.2.2.1- De manière algébrique

La méthode se déroule suivant trois étapes comme suit (Hurlin & Mignon, 2018) :

- Repérer dans la colonne Ni ou Fi, la valeur ou la position N/2 ou Fi = 0,5 (soit 50%), notée « Th_2 ».
- Déterminer la classe correspondant à cette position, c'est-à-dire la classe médiane.
- Appliquer la formule de la médiane suivante :

$$\mathbf{Me = X_0 + a}\left[\frac{\mathbf{Th_2 - N_{me}}}{\mathbf{n_{me}}}\right] \quad \mathbf{ou} \quad \mathbf{Me = X_0 + a}\left[\frac{\mathbf{0{,}5 - F_{me}}}{\mathbf{f_{me}}}\right]$$

Avec :

X_0= borne inférieure de la classe médiane.

a = amplitude de la classe médiane.

Th_2 = N/2 ou 0,5.

N_{me} ou F_{me} = effectif ou fréquence cumulé correspondant à la classe médiane.

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

Exemple 2

Déterminer la médiane de la distribution suivante :

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

Réponse :

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

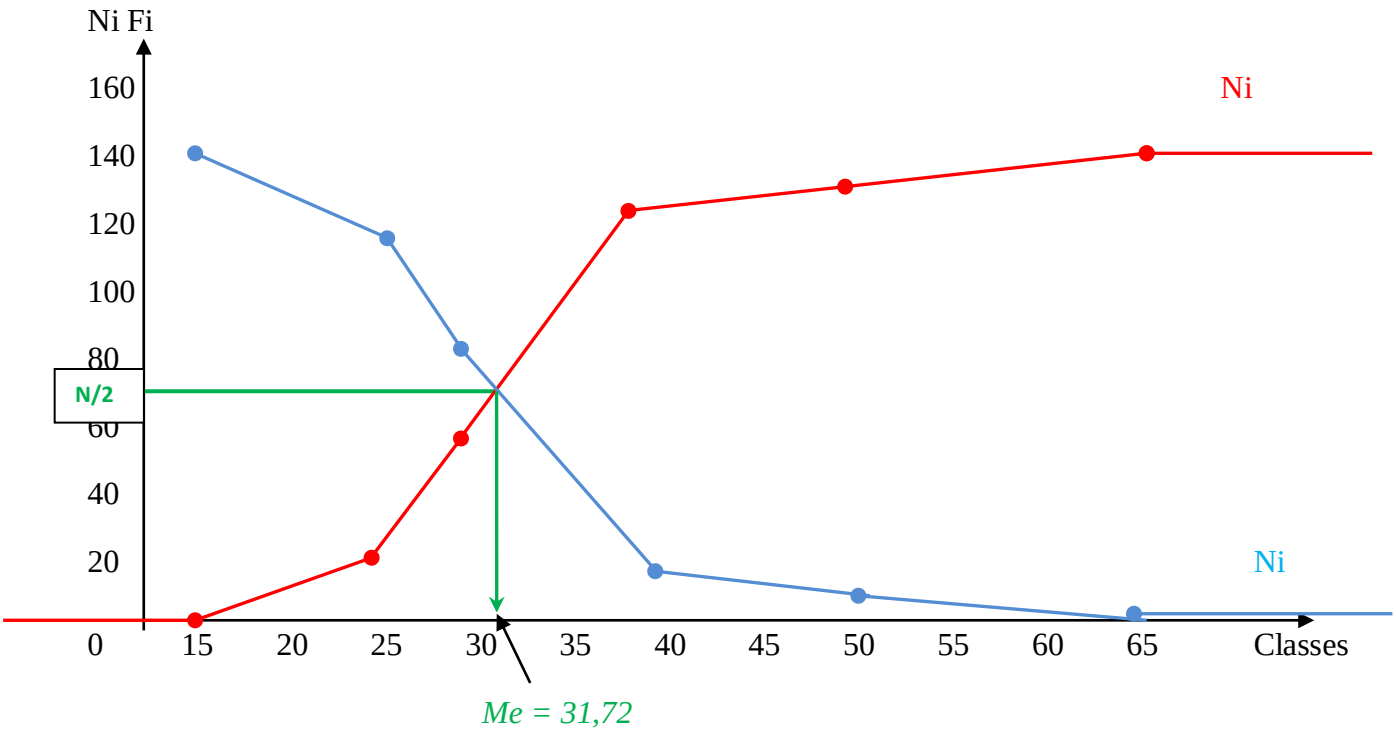
Classes	Xi	ni	Ni	Ni ↓
15 - 25	20	26	26	140
25 - 30	27,5	33	59	114
30 - 40	35	64	123	81
40 - 50	45	7	130	17
50 - 65	57,5	10	140	10
Total		140	-	-

Th_2 = N/2 = 70, cette valeur se trouve entre les valeurs 59 et 123, c'est-à-dire entre deux lignes, dans la colonne Ni, ce qui implique qu'on prendra directement la ligne du bas, soit celle correspondant à Ni = 123, la classe correspondante, soit [30 - 40[, est la classe médiane. ⟶ Me ∈ [30 - 40[.

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

2.2.2.2- De manière graphique

A partir des deux courbes cumulatives en « S », on détermine la médiane en projetant le point de rencontre des deux courbes sur l'axe horizontale. La valeur correspondante est la médiane (Py, 1996). Le point où se rencontrent les deux courbes est le point correspondant sur l'axe vertical à Ni = N/2 ou Fi = 0,5 (50%). Il a donc pour coordonnées (xi ; Ni) = (Me ; N/2).



Remarques

- La médiane satisfait assez bien les conditions de YULE à l'exception de la 6ème ; elle ne se prête pas aux calculs algébriques.

- Elle n'est pas affectée par les valeurs extrêmes, puisqu'elle ne dépend des valeurs que par leur position (leurs effectifs cumulés). (Py, 1996).

2.3- Généralisation de la médiane : les Quantiles

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

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

2.3.1- Les Quartiles

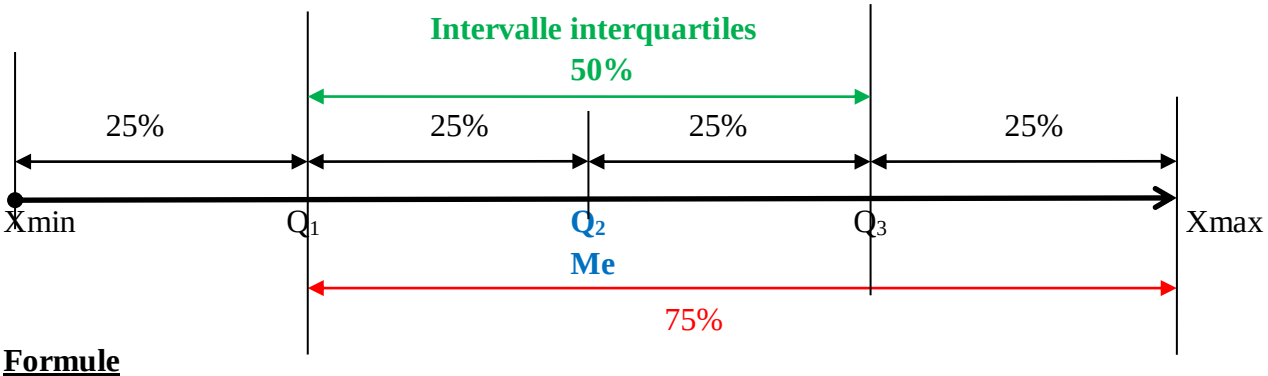
Notés « Qi », ce sont, selon Hurlin & Mignon (2018). des paramètres de position et aussi des modalités xi qui, par leurs positions (notées « **Thi** ») , partagent la série statistique en quatre groupes de mêmes effectifs, soit 25% chacun. Ils sont donc au nombre de trois : Q_1 ; Q_2 ; Q_3.

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

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

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

L'intervalle [Q_1 ; Q_3[s'appelle ***l'intervalle interquartile***. Il contient **50%** des modalités centrales (Hurlin & Mignon, 2018). On peut schématiser ces quartiles comme suit :



Formule

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

2.3.2- Les Déciles

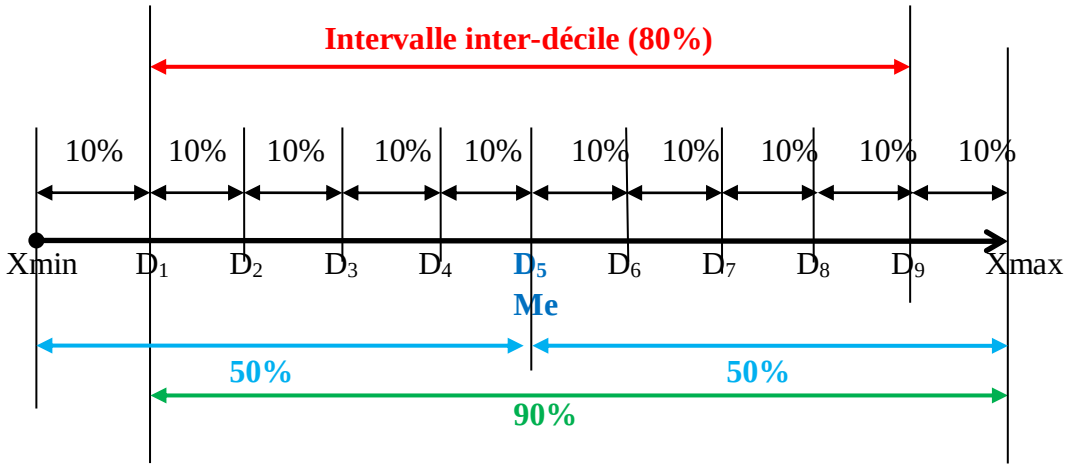
Notés « Di », ce sont des paramètres de position et aussi des modalités xi qui partagent la série statistique en dix groupes de même effectif, soit 10% chacun. Ils sont donc au nombre de neuf (9) : D_1 ; D_2…….. ; D_9.

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

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

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

L'intervalle [**D_1 ; D_9[** s'appelle ***intervalle inter-décile***. Il contient **80%** des modalités centrales (Hurlin & Mignon, 2018). On peut schématiser ces déciles comme suit :



Formule

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

2.3.3- Les Centiles

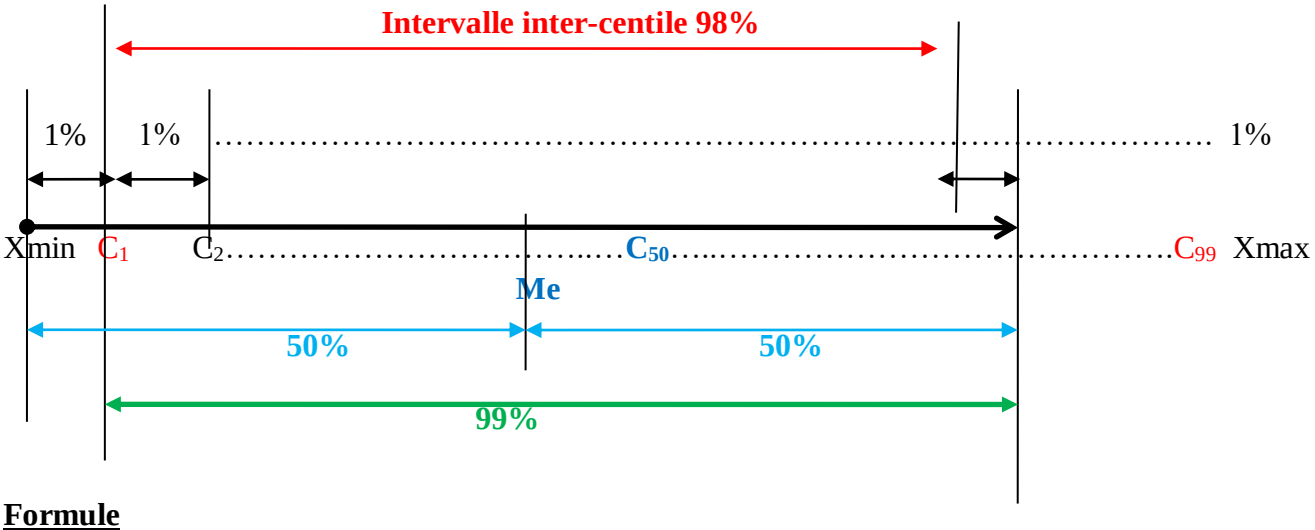
Notés « **Ci** », ce sont des paramètres de position et aussi des modalités xi qui partagent la série statistique en cent groupes de même effectif, soit 1% chacun. Ils sont donc au nombre de 99 : C_1 ; C_2……C_{50} ;… ; C_{99}.

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

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

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

L'intervalle **[C_1 ; C_{99}[** s'appelle ***intervalle inter-centile***. Il contient **99%** des modalités centrales (Hurlin & Mignon, 2018). On peut schématiser ces quartiles comme suit :



Formule

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

Remarque

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

Exemple

Calculer Q_1 ; D_2 et C_{80} de la distribution de l'exemple 2 précédent.

Réponse : ***Exemple***

Calculer Q_1 ; D_2 et C_{80} de la distribution de l'exemple 2 précédent.

1/- Q_1 :

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

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

2/- D_2 :

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

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

3/ C_{80} :

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

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

Section 3 : La Moyenne arithmétique

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

3.1- Définition

Notée « $\overline{\mathbf{X}}$ **»,** la moyenne arithmétique correspond au rapport de la somme des modalités par leur effectif total. (Hurlin & Mignon, 2018).

- On dit qu'une moyenne arithmétique est « *simple* » ou non pondérée lorsque chaque modalité x_i ne se répète qu'une seule fois ($n_i = 1 = Cte$) (Mazerolle, 2006). On écrit alors :

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

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

- On dit qu'une moyenne arithmétique est « *pondérée* » lorsqu'à chaque modalité x_i correspond un effectif (n_i). (Mazerolle, 2006). On écrit alors :

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

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

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

3.2- Méthode de calcul

3.2.1- Variable statistique discrète

3.2.1.1- Cas d'une série simple

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

N= 6 ; toutes les modalités se répètent une seule fois, c'est donc une série simple. Dans ce cas la moyenne arithmétique sera : $\overline{\mathbf{X}} = \left(\sum_{i=1}^{N} \boldsymbol{Xi}\right) / \mathbf{N}$ (Mazerolle, 2006).

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

3.2.1.2- Cas d'une série pondérée

Soit la distribution suivante :

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

Dans ce cas à chaque modalité (x_i) est associé un effectif (n_i). Il s'agit donc d'une série pondérée ou distribution statistique. Dans ce cas la moyenne arithmétique sera (Mazerolle, 2006) :

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

Pour pouvoir appliquer cette formule, nous devons compléter le tableau, en ajoutant une nouvelle colonne « n_ix_i » ou « f_ix_i ». Le tableau sera :

x_i	n_i	f_i	n_ix_i	f_ix_i
0	20	0,200	0	0
1	35	0,350	35	0,35
2	10	0,100	20	0,2
3	30	0,300	90	0,9
4	5	0,050	20	0,20
Total	100	1	165	1,65

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

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

3.2.2- Variable statistique continue

Dans ce cas on retient pour les calculs les centres de classes (xi), et on ajoute les colonne (n_ix_i) ou (f_ix_i). (Hamdani, 2006). Soit la distribution suivante :

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

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

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

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

3.3- Propriétés de la moyenne arithmétique

3.3.1 Propriétés algébriques : d'après B. Py (1996), elles résument aux six suivantes :

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

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

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

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

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

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

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

Démonstration :

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

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

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

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

5- La moyenne arithmétique d'une population scindée en deux ou plusieurs sous-populations est égale la moyenne arithmétique des moyennes des sous-populations, pondérées par leurs effectifs respectifs. Ainsi si une population P d'effectif total N est subdivisée en P_1, P_2, P_3,….Pn sous-populations de moyennes respectives X_1 ; X_2 ; X_3….., Xn et d'effectifs respectifs N_1 ; N_2 ; N_3 ;…… ; Nn ; alors la moyenne arithmétique de cette population est :

$$\mathbf{\overline{X}_P = 1/N\ \Sigma[X_1.N_1 + X_2.N_2 + X_3.N_3 + ;\ldots\ldots\ldots ; X_k.N_k]}$$

Ou bien $\qquad \mathbf{\overline{X}_P = \Sigma[X_1f_1 + X_2f_2 + X_3f_3 + ;\ldots\ldots\ldots + X_kf_k]}$

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

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

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

3.3.2- Propriétés générales : d'après B. Py (1996), on peut citer les suivantes :

- La moyenne arithmétique satisfait à toutes les conditions de YULE, ce qui fait qu'elle est le paramètre le plus utilisé en statistique.

- La moyenne arithmétique dépend de toutes les modalités par leurs valeurs et leurs effectifs.

- On apprécie par rapport à la moyenne arithmétique, la faiblesse ou l'importance d'un phénomène.

- Elle a une signification concrète, c'est la valeur de la variable qui égalise toutes les autres.

- C'est un paramètre qui n'entraine pas de perte d'informations du fait qu'il touche toutes les modalités.

- le seul inconvénient de la moyenne arithmétique c'est qu'elle est très sensible aux influences des valeurs (modalités) extrêmes qui la rendent peu significative. De même, dans le calcul des durées ou vitesses moyennes, des taux ou des pourcentages et des valeurs élevées au carré, elle est très mal appropriée. On lui préfère dans ce cas d'autres types de moyennes que nous développons dans la quatrième section. (Py, 1996).

3.4- Méthode rapide de calcul de $\overline{X}$: le *changement de variable*

Appelée aussi « *méthode de l'origine arbitraire* » (Boudia, 2008), cette méthode permet de calculer $\overline{X}$ en réduisant l'importance des modalités, notamment lorsque celles-ci sont trop volumineuses.

Cette méthode consiste, selon Py (2007) à utiliser une nouvelle variable, notée « $\mathbf{X_i'}$ », qu'on obtient en faisant subir à la variable (x_i) :

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

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

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

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

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

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

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

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

On en déduit que :

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

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

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

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

Exemple

Soit la distribution suivante :

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

- Calculer la moyenne arithmétique par la méthode classique et par le changement de variable.

Réponse :

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

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

NB/- On a ajouté la colonne (fi) à titre supplémentaire pour rappeler à l'étudiant que toutes les applications que l'on fait avec les ni (i.e en termes absolus) on peut aussi les faires avec les fi (i.e en termes relatifs).

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

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

$\overline{X} = a.\overline{X'} + x_0$; (avec a et x0 respectivement l'amplitude et le centre de la classe modale).

$a_i \neq Cte \longrightarrow$ la classe modale est celle qui correspond à la plus grande densité ($d_3 = 4$)

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

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

On retrouve donc la même moyenne calculée précédemment.

Section 4 :- Généralisation de la moyenne

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

moyens, de la moyenne harmonique, pour le calcul des rapports ou vitesses moyennes, et de la moyenne quadratique pour la moyenne des valeurs élevées au carrée.

4.1- La moyenne géométrique

4.1.1- Définition

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

4.1.2 - Calcul de la moyenne géométrique

4.1.2.1- Moyenne géométrique simple

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

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

Le calcul peut également se faire par les logarithmes :

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

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

Exemple 10

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

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

4.1.2.2- La moyenne géométrique pondérée

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

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

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

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

Exemple 11

Soit la distribution suivante :

Xi	1	2	3	4
ni	4	6	8	2

- Calculer la moyenne géométrique.

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

4.1.3- Propriétés de la moyenne géométrique : d'après Py (2007), on peut énoncer :

- La moyenne géométrique du produit de deux variables statistiques est égale au produit de leurs moyennes géométriques respectives :

Si Z = X.Y ⟶ $\mathbf{G_z = G_x.G_y}$

- La moyenne géométrique du rapport de deux variables statistiques est égale au rapport de leurs moyennes géométriques respectives :

Si Z = X/Y ⟶ $\mathbf{G_z = G_x/G_y}$

- La moyenne géométrique du produit d'une constante et d'une variable statistique est égale au produit de la constante et la moyenne géométrique de la variable :

Si Y = a.X ⟶ $\mathbf{G_y = a.\ G_x}$

- La moyenne géométrique élevée à la puissance *p* est égale à la moyenne géométrique de la variable qui est élevée à la puissance *p* :

$$\mathbf{(Gx)^p} \left[= \sqrt[N]{\prod_{i=1}^{k}(xi)^{ni}}\right]^p = \sqrt[N]{\prod_{i=1}^{k}(xi)^{ni.p}} = \underline{\mathbf{Gx_i^p = (G_x)^p}}$$

- La moyenne géométrique est toujours inférieure à $\bar{X}$.

- La moyenne géométrique d'une constante « a » est égale à la constante elle-même :

$$\mathbf{Ga = a \qquad (a = C^{te})}$$

4.2- La moyenne harmonique

4.2.1- Définition

Notée « **H** », elle est la valeur de la variable pour laquelle son inverse est la moyenne arithmétique de l'inverse des modalités de la variable ($1/x_i$) (Hamdani, 2006). On l'exprime le plus souvent par son inverse ($1/\bar{H}$) pour faire apparaitre sa logique de construction semblable à celle de $\bar{X}$.

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

4.2.2- Méthode de calcul

Selon la formule donnée par Hamdani (2006), elle se calcule comme suit :

4.2.2.2- Moyenne harmonique simple

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

4.2.2.3- Moyenne harmonique pondérée

$$1/H = \left(\sum_{i=1}^{k} ni/xi\right)/N = 1/N\ \Sigma(N_1/X_1 + N_2/X_2 + \ldots + N_K/X_K) \longrightarrow \mathbf{H = N/\sum_{i=1}^{k}(ni/xi)}$$

Exemple 12

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

Ce n'est pas (20 + 14)/2 = 17€ !

Le cours moyen est = (valeur totale des actions)/(nombre total d'actions).

Soient :

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

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

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

CM = ΣVi/ΣQi……..(1)

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

Dans chaque opération : CMi = Vi/Qi ⟶ Qi = Vi/CMi ; on remplace dans (1) :

CM = ΣVi/Σ(Vi/CMi) , on retrouve donc la formule de la moyenne harmonique : H = N/Σ(ni/xi)

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

CM = (20.000 + 10.000)/[(20.000/20) + (10.000/14)] = **17,65€.**

4.3- La moyenne quadratique

4.3.1- Définition

Notée « **Q** », elle est la racine carrée de la moyenne arithmétique des carrés des variables (x_i^2) de la série. On l'emploi surtout pour le calcul des moyennes des écarts à une variable centrale ; $(Xi - \overline{X})^2$ ou $(Xi - Me)^2$; où l'élévation au carré permet de ne pas avoir à manipuler des écarts négatifs (première propriété de $\overline{X}$). On l'exprime souvent au carré (Q^2) pour éviter de manipuler la racine carrée (Boudia, 2008).

4.3.2- Méthode de calcul

La moyenne harmonique se calcule, selon le type de série, avec l'une ou l'autre des deux formules suivantes :

$$\mathbf{Q^2 = 1/N \sum_{i=1}^{N} (Xi)^2} \longrightarrow \textbf{Moyenne simple}$$

$$\mathbf{Q^2 = 1/N \sum_{i=1}^{k} ni(xi)^2} \longrightarrow \textbf{Moyenne pondérée}$$

Exemple 13

Calculer la moyenne quadratique de la distribution suivante :

Xi	1	2	3	4	5	6
ni	20	30	15	10	5	2

Pour calculer « Q », nous devons ajouter deux autres colonnes au tableau : une colonne x_i^2 et une autre colonne $ni(xi)^2$:

xi	ni	$(xi)^2$	$ni(xi)^2$
1	20	1	20
2	30	4	120
3	15	9	135
4	10	16	160
5	5	25	125
6	2	36	72
Total	82	-	632

- Il s'agit d'une moyenne quadratique pondérée :

$Q^2 = 1/N \sum_{i=1}^{k} ni(xi)^2 = 1/82\ (632) = 7,71$

⟶ $Q = \sqrt{7,71} =$ **2,78**

Remarque

On vérifie toujours la relation suivante :

H< G< X< Q

Conclusion au chapitre 4

Au terme de ce chapitre l'étudiant aura saisit les principaux paramètres de tendance centrale qui permettent de résumer une série statistique à travers un seul paramètre, exprimé par un chiffre. Nous les avons passé en revue en détail, en insistant sur les avantages et les inconvénients de chacun, en indiquant à l'étudiant les précautions à prendre et les pièges à éviter lors de la manipulation de ces paramètres.

Résumer ou caractériser une distribution statistique par un paramètre de tendance centrale est à la fois facile et compliqué. Cependant, parfois c'est dans la facilité que réside la complexité du calcul de ces paramètres. La facilité dans le calcul de ces paramètres peut distraire l'étudiant des pièges et erreurs inhérents à ce calcul, et peut contrarier le les résultats, voire même les décisions à prendre, qui en découlerai dans le cas d'un gestionnaire d'entreprise.

Cependant, il faut savoir que les séries statistiques ne se résument pas seulement par des paramètres de tendance centrale, mais également par d'autres paramètres, notamment les paramètres de dispersion ou les écarts, qui constituent l'objet du chapitre suivant.

Chapitre 5 : Les paramètres de dispersion

Introduction au chapitre 5

Après les paramètres de tendance centrale étudiés dans le chapitre précédent, il s'agit dans le présent chapitre d'étudier d'autres paramètres qui consistent à évaluer ou à calculer l'éloignement des valeurs par rapport à leur valeur centrale, le plus souvent leur moyenne arithmétique. Ce sont les paramètres de « *dispersion* » qu'on appelle aussi les « *écarts* ».

Si l'on considère l'exemple suivant de B. Py (2007), de deux séries des notes des étudiants dans deux groupes A et B :

Groupe A : 2- 2- 2- 2- 10- 18- 18- 18- 18.

Groupe B : 9- 9- 9- 9- 10- 11- 11- 11- 11.

Par un calcul simple on peut déduire que les deux séries ont la même moyenne arithmétique (et la même médiane). Pourtant, elles reflètent deux réalités différentes. En effet, alors que dans le groupe B les étudiants présentent un niveau général moyen, voire bon, pour tous les étudiants, dans le groupe A nous avons près de la moitié des étudiants qui sont très loin en dessous de la moyenne et une autre moitié qui présente un niveau très loin au-dessus de la moyenne. Autrement dit, le même paramètre $\bar{X}$ ne reflète pas la même réalité dans les deux séries.

Par conséquent, il faut se méfier des raisonnements ayant pour seul support les paramètres de tendance centrale. Ces derniers sont intéressants mais sont insuffisants car, le plus souvent, il est nécessaire d'avoir des renseignements sur la répartition (l'éloignement) des valeurs entre elles et autour de leur valeur centrale (la moyenne le plus souvent), c'est-à-dire sur leur « *dispersion* ».

Aussi, en statistique, on dépasse le raisonnement par les seuls paramètres de tendance centrale en analysant la dispersion et/ou la concentration des valeurs de la série.

La dispersion, objet du présent chapitre, analyse la fluctuation ou l'éloignement des valeurs par rapport à une valeur centrale (généralement la moyenne arithmétique) ou dans un intervalle. Cette dispersion est appréhendée par la notion d' « *écarts* » (Py, 1996). Dans le présent chapitre nous étudions successivement :

- la dispersion dans un intervalle, mesurée par les écarts simples (section 1),
- la dispersion autour d'une valeur centrale, mesurée par les écarts moyens (section 2),
- la comparaison de la dispersion de deux ou plusieurs séries statistiques, à savoir le coefficient de variation (Section 3).

Section 1 : La dispersion dans un intervalle ou les écarts simples

Ce sont essentiellement l'intervalle de variation, appelé aussi « *étendue* » (e), et les intervalles inter-quantiles.

1.1. Intervalle de variation (ou étendue)

Notée (e), elle est la différence entre la plus grande et la plus petite valeurs de la série statistique ordonnée par ordre croissant. On comprend, par conséquent, qu'elle est sujette à des fluctuations considérables d'un échantillon à un autre et très sensible aux influences des valeurs extrêmes aberrantes. Aussi, on ne l'utilise que pour avoir une idée sommaire et rapide de la dispersion de la série. Pour éviter l'influence des valeurs extrêmes aberrantes, on choisit de les écarter de la série, on a alors recours aux intervalles inter-quantiles ; on perd en informations mais on gagne en homogénéité. (Py, 1996).

1.2. Les intervalles inter-quantiles

Les plus utilisés sont l'intervalle interquartile, l'intervalle inter-décile et l'intervalle inter-

centile. Toutefois, ce dernier intervalle exclut difficilement les valeurs aberrantes.

La définition et la détermination des intervalles interquartiles étant déjà développées dans le chapitre précédent, il faut cependant savoir que ces paramètres sont aussi imparfaits. Ils sont simples et rapides à calculer et à interpréter mais ont l'inconvénient de ne tenir compte que de la position des modalités et pas de leurs valeurs. Or, il est indispensable d'avoir recours à des paramètres permettant de tenir compte de toutes les modalités de la série. C'est ce que nous permettent les écarts autour d'une valeur centrale ou *écarts moyens*.

Section 2 : La dispersion autour d'une valeur centrale : les écarts moyens

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

- considérer la valeur absolue des écarts : $| Xi - \overline{X} |$, afin d'éviter les écarts négatifs,

- considérer les carrés des écarts $(Xi - \overline{X})^2$, permettant également d'éviter les écarts négatifs. (Mazerolle, 2006).

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

2.1. L'écart absolu moyen

Noté **Ex**, il désigne la moyenne arithmétique des valeurs absolues des écarts des modalités par rapport à leur moyenne arithmétique ($|Xi - X|$). On écrit alors, selon Hurlin & Mignon (2018):

$\mathbf{Ex = 1/N\ \Sigma|Xi - \overline{X}|}$ pour l'écart absolu moyen simple.

$\mathbf{Ex = 1/N\ \Sigma ni|Xi - \overline{X}|}$ **ou** $\mathbf{E_x = \Sigma fi|Xi - \overline{X}|}$ pour l'écart absolu moyen pondéré.

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

$\mathbf{E_{Me} = 1/N\ \Sigma|Xi - Me|}$ pour l'écart absolu médian simple.

$\mathbf{E_{Me} = 1/N\ \Sigma ni|Xi - Me|}$ **(ou** $\mathbf{E_{Me} = \Sigma fi|Xi - Me|}$**)** pour l'écart absolu médian pondéré.

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

2.2. La variance et l'écart quadratique moyen

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

2.2.1. La Variance

Notée Vx, la variance est la moyenne arithmétique des carrés des écarts à la moyenne arithmétique (Hurlin & Mignon, 2018). On écrit alors dans une première formule appelée *formule de définition* :

$$\mathbf{Vx = 1/N\Sigma(Xi - \bar{X})^2}$$ ……Variance simple.

$$\mathbf{Vx = 1/N\Sigma ni(Xi - \bar{X})^2 = \Sigma fi(Xi - \bar{X})^2}$$……Variance pondérée.

2.2.2. L'Ecart quadratique moyen

Noté δx et appelé aussi **écart-type** du fait qu'il constitue le paramètre de dispersion le plus pertinent à l'heure actuelle (Hurlin & Mignon, 2018), il est la racine carré de la variance. On écrit:

$$\boldsymbol{\delta x = \sqrt{Vx} = \sqrt{(1/N[\Sigma(Xi - \bar{X})^2]}}$$ ……Ecart-type simple.

$$\boldsymbol{\delta x = \sqrt{Vx} = \sqrt{1/N[\Sigma ni(Xi - X)^2]} = \sqrt{\Sigma fi(Xi - X)^2}}$$……Ecart-type pondéré.

Or, comme on l'a vu dans le chapitre précédent, la racine carré des carrés des modalités est une moyenne quadratique (Q), d'où l'appellation d'écart quadratique moyen. Cet écart est le plus petit écart que l'on puisse avoir : $\Sigma(Xi - X) \longrightarrow$ Min), deuxième propriété de la moyenne arithmétique (Py, 1996). C'est l'écart le plus utilisé en statistique.

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

Exemple 1

Soit la distribution des salaires (en 10^3DA) dans une entreprise suivante :

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

- Calculer l'écart absolu moyen, la variance et l'écart-type de la distribution.

Réponse

1. Ecart absolu moyen : $Ex = 1/N\Sigma|Xi - \bar{X}|$

$\bar{X} = 1/N(\Sigma nixi) = 2848/140 =$ ***20,343*** $\longrightarrow$ *Compléter le tableau.*

$Ex = 1/140\ |484,52| = 3,46$ $\longrightarrow$ ***$Ex = 3,46 .10^3$ DA.***

2. La Variance : $Vx = 1/N\Sigma ni(Xi - \bar{X}) = 1/140(25595442)$ $\longrightarrow$ ***$Vx=18,282 .10^3$ DA.***

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

2.2.3. Autres méthodes de calcul de la variance et de l'écart-type

En plus de la formule de définition que l'on vient d'expliciter, il existe deux autres manières de déterminer la variance et l'écart-type.

2.2.3.1. La Formule développée

C'est une autre formule plus simplifiée et moins lourde permettant un calcul plus rapide

avec des risques d'erreurs de calcul minimisés. Cette formule résulte du développement mathématique de la formule de définition (Hurlin & Mignon, 2018).

$$Vx = 1/N\Sigma ni(Xi - X)2 = 1/N\Sigma ni(Xi^2 - 2Xi\overline{X} + \overline{X}^2) = 1/N(\Sigma ni(Xi^2)) - 1/N(2\Sigma niXi\overline{X}) + 1/N(\Sigma ni\overline{X})$$

$$= 1/N(\Sigma niXi^2) - 2\overline{X}\overline{X} + 1\overline{X}.$$

$$= 1/N(\Sigma niXi^2) - 2\overline{X}^2 + \overline{X} = 1/N(\Sigma niXi^2) - \overline{X}^2$$

On aura donc : $\mathbf{Vx = 1/N(\Sigma ni(Xi^2)) - \overline{X}^2 = \Sigma fi(xi^2) - \overline{X}^2}$

Exemple 2

Recalculer la variance et l'écart-type de la distribution de l'exemple 1 précédent en utilisant la formule développée.

Réponse

Pour cela il suffit d'ajouter les deux colonnes suivantes au tableau statistique :

Xi^2	$ni(Xi^2)$
196	5090
324	10692
484	30976
676	4732
900	9000
-	**60496**

$Vx = 1/N[\Sigma ni(Xi^2)] - \overline{X}^2 \longrightarrow \overline{X^2} = (20,343)^2 = 413,115.$

$Vx = 1/140(60496) - 413,115 \longrightarrow \mathbf{Vx = 18,28 .10^3}$

$\longrightarrow \boldsymbol{\delta x = 4,275 .10^3.}$

2.2.3.2. La méthode de changement de variable

Comme pour la moyenne arithmétique, la variance et l'écart-type peuvent se calculer par la méthode de changement de variable. En suivant le même principe, il s'agit d'utiliser une nouvelle variable (Xi'), site variable centrée (Xi – Xo) réduite [(Xi – Xo)/a]. Suivant B. Py (2007), la méthode est la suivante :

$Xi \longrightarrow Xi' = (Xi - Xo)/a$ Avec Xo et a comme respectivement le centre et l'amplitude de la classe modale.

On aura alors :

$\mathbf{Vx = Vx'.(a^2)}$. Avec $\mathbf{Vx' = 1/N\ [\Sigma ni(Xi'^2)] - (\overline{X'})^2}$

Ou bien :

$\mathbf{Vx' = 1/N[\Sigma ni(Xi' - \overline{X'})^2]}$.

Exemple 3

Recalculer la variance de la distribution de l'exemple1 précédent en utilisant la méthode de changement de variable.

Réponse

Pour cela il faut construire un autre tableau statistique avec la nouvelle variable Xi'.

Salaires	Xi	Xi'	niXi'	$(Xi' - \overline{X'})^2$	$ni(Xi' - \overline{X'})^2$	$(Xi')^2$	$ni(Xi'^2)$
12 - 16	14	-2	-52	2,4964	64,9064	4	104
16 - 20	18	-1	-33	0,3364	11,1012	1	33
20 - 24	22	0	0	0,1764	11,2896	0	0
24 - 28	26	1	7	2,0164	14,1148	1	7
28 - 32	30	2	20	5,8564	58,564	4	40
Total	-	-	-58	-	159,976		184

$Vx = Vx' . (a)^2$. *La classe modale : $Mo \in [20 - 24[\longrightarrow Xo = 22$ et $a = 4$.*

~~X'~~ $= 1/N(\Sigma niXi') = -58/140 = -0,42 \longrightarrow (\overline{Xi'})^2 = 0,1681$

a/- Formule de définition

$Vx' = 1/N\Sigma ni(Xi' - \overline{X'})^2 = 1/140\ (159,976) =$ ***1,143*** $\longrightarrow$ *$Vx = 1,143 . (4)^2 =$* ***$18,28 .10^3$DA*** $\longrightarrow$ ***$\delta x = 4,27 .10^3$DA.***

b/- Formule développée

$Vx' = 1/N\ [\Sigma ni(Xi'^2)] - (\overline{X'})^2 = 1/140[\ (184) - (0,1681)] = 1,14$

$Vx = Vx' . (a)^2 = 1,14. (4)^2 =$ ***$18,28 .10^3$DA*** $\longrightarrow$ ***$\delta x = 4,27. 10^3$DA***

Section 3: La comparaison des dispersions des séries statistiques

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

Pour pouvoir faire ces comparaisons sans difficulté, on calcul un paramètre appelé « *Coefficient de variation* » (Hurlin & Mignon, 2018).

Le *Coefficient de variation*, noté C.V, est un nombre sans dimension (ou sans unité de mesure) exprimé de ce fait en pourcentage. Il mesure le rapport de la moyenne arithmétique à l'écart-type :

$$\mathbf{C.V = (\delta x / \overline{X}). 100}$$

Plus ce pourcentage (C.V) est élevé, plus la dispersion est forte, et inversement (Hurlin & Mignon, 2018).

Exemple 4

Reprendre l'exemple 1 précédent et calculer le Coefficient de variation, puis, Comparer la dispersion des salaires de notre entreprise avec celle d'une autre entreprise dont le salaire moyen est de 30. 10^3 DA et l'écart-type de 15,68 DA.

Réponse

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

$CV_A = \delta x_A / \overline{X_A} = (4,27/20,34) . 100 = 21\%$.

Pour pouvoir comparer, il faut calculer le Coefficient de variation de l'autre entreprise et qu'on va noter $CV_{B.}$

$CV_B = \delta x_B / \overline{X_B} = (15{,}68/30) .100 = 52{,}27\%$.

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

Conclusion au chapitre 5

Au terme de ce quatrième chapitre, l'étudiant aura pris connaissance du contenu du deuxième grand groupe de paramètres que sont les paramètres de dispersion. Deux paramètres sont ainsi été mis en évidence : l'écart absolu moyen et l'écart-type. Ces paramètres indiquent l'éloignement moyen des modalités par rapport à leurs valeurs centrales médiane ou moyenne arithmétique.

Cependant, ces paramètres nous indiquent pas la manière dont s'éloignent les modalités par rapport à leur valeur centrale. C'est-à-dire de manière symétrique ou asymétrique. Ceci est indiqué par d'autres paramètres qu'on appelle les *paramètres de forme*. Ceux-ci sont l'objet du chapitre suivant.

Chapitre 6 : Les caractéristiques de forme

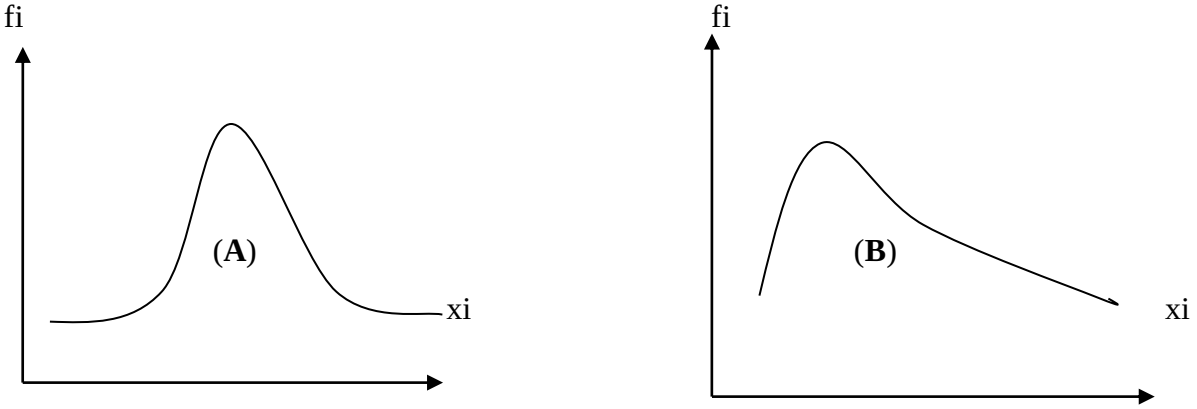
Introduction au chapitre 6

Pour résumer et caractériser les séries statistiques, nous avons jusque-là analysé :

- les paramètres de tendance centrale, qui donnent une idée de la grandeur de la série,

- les caractéristiques de dispersion qui mesurent l'intensité de l'éloignement ou de la fluctuation des modalités autour d'une valeur centrale. (Py, 1996).

Nous avons ainsi à notre disposition une multitude de paramètres qui nous renseignent sur l'allure générale de la série et qu'on retrouve en traçant la courbe des fréquences.



On voit bien que les deux séries ne se ressemblent pas. La série A est symétrique et très peu aplatie. La série B est asymétrique et assez aplatie. Il est par conséquent logique de chercher à caractériser ou à mesurer ces différences.

Les paramètres de forme permettent de préciser l'allure de la courbe des fréquences sans avoir besoin de la tracer. On repère généralement deux mesures de la forme d'une distribution statistique (Py, 1996) :

- celle de l'asymétrie qui a pour objet de nous renseigner sur la façon régulière ou non dont les observations se répartissent de part et d'autre d'une valeur centrale. C'est ce que nous étudions dans la section 1,

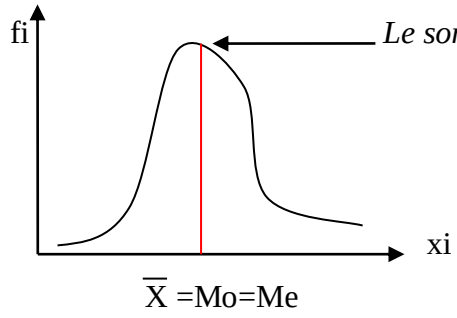
- celle de l'aplatissement qui a pour objet de faire apparaitre si une faible variation de la valeur de la variable entraine ou non une forte variation des fréquences relatives. C'est ce que nous étudions dans la section 2.

Section 1 : Mesure de la symétrie

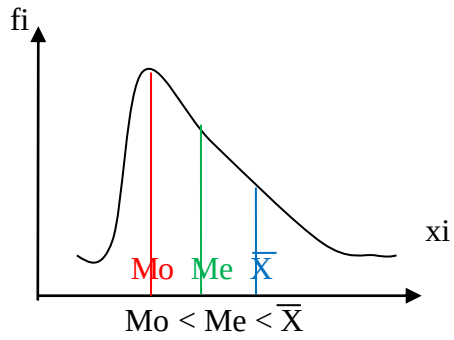
1.1. Définition

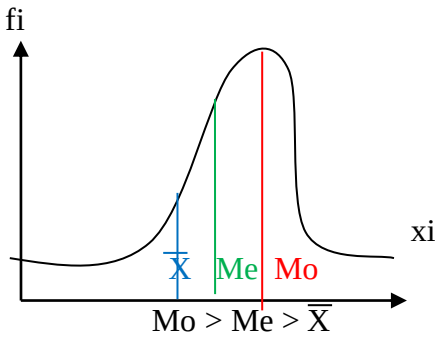
Une série statistique est dite symétrique si les modalités repérées par leurs fréquences sont également dispersées de part et d'autre de leur valeur centrale ($\bar{X}$, Me, Mo).

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



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





1.2. Calcul des paramètres d'asymétrie

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

1.2.1. Le Coefficient de YULE

YULE propose une mesure de la symétrie en comparant l'étalement à gauche et l'étalement à droite, tous deux repérés par les trois quartiles de la série. La formule est :

$$S = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)}$$

Si :

- S = 0 ⟶ La distribution est symétrique.
- S > 0 ⟶ $(Q_3 - Q_2) > (Q_2 - Q_1)$ ⟶ Etalement à droite, donc distribution oblique à gauche.
- S < 0 ⟶ $(Q_3 - Q_2) < (Q_2 - Q_1)$ ⟶ Etalement à gauche, donc distribution oblique à droite (Hurlin & Mignon, 2018).

1.2.2- Le coefficient de Pearson

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

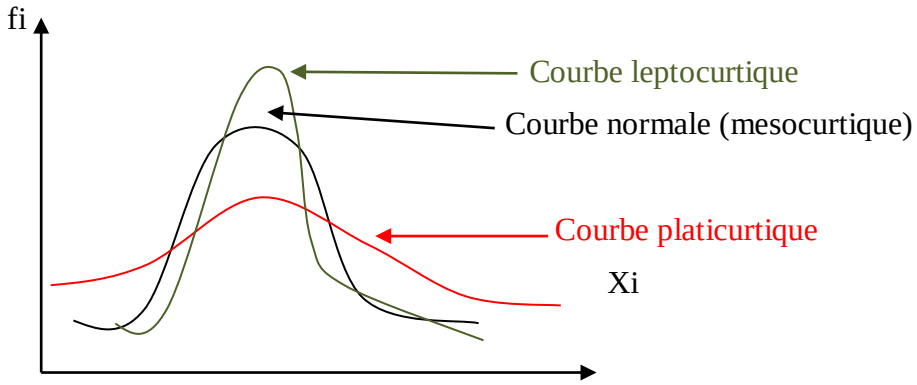
$$S = \frac{\overline{X} - Mo}{\delta_x}$$

- S = 0 ⟶ La distribution est symétrique, $\overline{X}$ = Mo (cette situation est très rare en pratique).
- S < 0 ⟶ $\overline{X}$ < Mo ⟶ La courbe est oblique à droite (le sommet de la courbe est situé à droite après $\overline{X}$).
- S > 0 ⟶ $\overline{X}$ > Mo ⟶ La courbe est oblique à gauche (le sommet de la courbe est situé à gauche avant $\overline{X}$). (Hurlin & Mignon, 2018).

Section 2 : Mesure de l'aplatissement

2.1 Définition

L'aplatissement indique si une faible variation des fréquences de la variable entraine ou non une forte variation des fréquences relatives et inversement (ex ; une variation de1cm en x_i correspond à une variation de 5cm en f_i, et inversement). L'aplatissement étant mesuré par rapport à la courbe *normale* ou *mesocurtique* (courbe en forme de cloche, appelée aussi courbe de la *loi normale*).



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

Remarque : En latin :

- le préfixe « *lepto* » signifie « grêle » ou aigue ou anse,
- le préfixe « *platus* » signifie large ou étalée,
- le préfixe « *kurtosis* » signifie bosse ou saillie (Py, 1996).

2.2. Calcul des paramètres d'aplatissement

Comme pour la mesure de l'asymétrie, en statistique il existe plusieurs paramètres, appelés aussi coefficients, qui permettent de mesurer l'aplatissement d'une distribution

statistique. La logique est de comparer si la distribution est plus ou moins aplatie par rapport à une courbe normale ou *mesocurtique* (Labrousse, 1995). Nous en proposons deux des plus utilisés en statistique, ils reposent sur la notion de « *Moment* », plus précisément le *moment centré d'ordre 'r'*, noté ***Ur***.

Remarque

Le moment général par rapport à une origine 'a' s'écrit :

$$\mathbf{Mr = 1/N[\Sigma ni(Xi - a)^r]}$$

Il existe, selon C. Labrousse (1995), deux types de moments d'ordre 'r' :

a/- Le moment simple

C'est le moment général pour lequel l'origine a = 0. On écrit alors :

$$\mathbf{M_r = 1/N[\Sigma ni(xi)^r]}$$

- Si r = 0 ⟶ Mo = 1.
- Si r = 1 ⟶ $M_1 = \overline{X}$.
- Si r = 2 ⟶ $M_2 = \overline{X^2}$.

b/- Le moment centré d'ordre 'r'

C'est le moment général pour lequel l'origine a est égale à $\overline{X}$. On écrit alors :

$$\boldsymbol{U_r = 1/N\Sigma ni(Xi - \overline{X})^r}$$

- Si r = 0 ⟶ $U_0 = 1$.
- Si r = 1 ⟶ $U_1 = 0$.
- Si r = 2 ⟶ $U_2 = Vx$ (*la variance*) ou $(\delta_x)^2$.

2.2.1. Le coefficient d'aplatissement de Pearson

Noté P, le coefficient de Pearson s'écrit comme suit :

$$\mathbf{P = U_4/(U_2)^2 = U_4/(\delta x)^4}$$

P = 3 ⟶ La distribution est normale ou mésocurtique.

P > 3 ⟶ La courbe est leptocurtique.

P < 3 ⟶ La courbe est platicurtique.

Remarque

P est toujours positif et supérieur à 1. Plus il tend vers 1, plus la courbe est aplatie. (Labrousse, 1995).

2.2.2- Le coefficient d'aplatissement de Fisher

Noté F, le coefficient d'aplatissement de Fisher est égale au coefficient de Pearson moins 3. Il se calcule comme suit (Hurlin & Mignon, 2018) :

$$\mathbf{F = [U_4/(\delta x)^4] - [3] = [P - 3]} \qquad \textbf{Avec } [(\delta_x)^2]^2 = (\delta_x)^4 = (U_2)^2$$

- Si F = 0 ⟶ La courbe est normale ou mesocurtique.
- Si F > 0 ⟶ La courbe est leptocurtique.
- Si F < 0 ⟶ La courbe est platicurtique.
-

Conclusion au chapitre 6

Par son contenu, le présent chapitre permet à l'étudiant de savoir, après avoir déterminé ses paramètres de tendance centrale et l'éloignement des modalités de la série par rapport à ces derniers (la dispersion), comment se répartissent les modalités de part et d'autres des valeurs centrales et si elles présentent une distribution symétrique ou asymétrique (oblique). De même que l'étudiant est désormais capable de déterminer de quel côté l'obliquité de la distribution est prononcée.

Il reste cependant à l'étudiant de découvrir le quatrième groupe de paramètres, à savoir ; les paramètres de concentration. C'est l'objet du chapitre suivant.

Chapitre 7 : Les paramètres de concentration

Introduction au chapitre 7

Une autre manière d'analyser la dispersion pourrait être aussi l'analyse de sa conséquence. En effet, s'il y a dispersion, il y a forcément concentration des modalités. Si celles-ci se dispersent de part et d'autre de leur valeur centrale, c'est pour se concentrer de part et d'autre des deux cotés de la série (Py, 1996). Dispersion et concentration sont donc deux notions interdépendantes.

En économie, la notion de concentration est très importante : on parle de concentration des chiffres d'affaires, des salaire, des populations,….

En Statistique, l'analyse de la concentration met en confrontation, ou la comparaison, les effectifs (n_i) ou les fréquences (f_i) des individus de la population statistique étudiée et leurs « masses » (n_ix_i ou f_ix_i) respectives (Anderson & al., 2007) . Ainsi, par exemple, la répartition des salaires est analysée par la confrontation des effectifs des employés aux masses salariales dans l'entreprise.

Autrement dit, la concentration mesure la proportionnalité des fréquences (proportions ($f_{i\ =\ ni/N}$)) des individus aux fréquences (proportions) de leurs « masses » ($n_ix_i/\Sigma n_ix_i$). Une disproportion entre les deux fréquences indique l'existence de concentration (Grais, 2000).

Il existe généralement, en Statistique, deux méthodes d'analyse de la concentration :

- la méthode algébrique ou mathématique,
- la méthode graphique ou géométrique (Py, 2007).

Ce sont ces deux méthodes que nous développerons dans le présent chapitre. En premier lieu on développera la méthode algébrique qui consiste à comparer deux paramètres de position centrale : l'un relatif aux effectifs ou fréquences, n_i ou f_i, appelé « *Médiane* » (déjà développée au chapitre III), l'autre relatif aux « masses », n_ix_i, appelé « *Médiale* ». Cette comparaison permettra de donner une première idée sur l'importance de la concentration des modalités. (Section 1).

En deuxième lieu, on développera la méthode graphique qui consiste à tracer une courbe, dite « *courbe de Gini* » ou « *courbe de concentration* », ou encore, « *courbe de Lorenz* », à partir de laquelle on déduit un paramètre ou indicateur appelé « *indice de Gini* ». Cet indice, qui a d'abord pour objet de confirmer le constat de la méthode algébrique, nous renseigne également sur l'intensité de la concentration des modalités. (Section 2).

Section 1. L'analyse algébrique de la concentration

L'analyse algébrique de la concentration revient à calculer et à comparer des paramètres, en suivant les quatre étapes suivantes (Py, 1996) :

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

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

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

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

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

1.1- Notion de Médiale

Notée ML, la Médiale est la modalité de série qui partage la somme des masses ou la masse totale, $(\Sigma n_i x_i)$, en deux parties identiques : $(\Sigma n_i x_i)/2$ chacune (Hurlin & Mignon, 2018). Elle se détermine suivant le même principe que la médiane.

1.1.1- Calcul de la Médiale

Mis à part le fait que l'on se réfère à la colonne des $(n_i x_i)$ cumulés, notée $(n_i x_i)\uparrow$, on calcul la Médiale en suivant les même étapes que la médiane (Py, 1996) :

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

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

Ou bien :

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

Avec:

X_0 = Borne inférieure de la classe médiale.

a = amplitude de la classe médiale.

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

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

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

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

1.2- L'écart Médiale-Médiane

Noté ΔM, il mesure la différence |ML - Me|. Cet écart est généralement exprimé en valeur absolu pour rappeler qu'il est positif, puisque la Médiale est toujours supérieure à la Médiane (Bressoud & Kahané, 2009).

1.3- Comparaison de ΔM à l'étendue

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

- Si (ΔM/e) est inférieur à 50%, la concentration est dite ***<u>faible</u>***.

- (ΔM/e) est supérieur à 50%, la concentration est dite *forte*.
- (ΔM/e) est égale à 1, cela signifie que ΔM = e, dans ce cas la concentration est dite *nulle* . C'est une situation de parfaite répartition des modalités ou d'« équi-répartition ». Il y a une proportionnalité entre les fréquences d'individus et les fréquences des masses. Autrement dit, à chaque proportion d'individus correspondrait la même proportion de masse : $f_i = 10\% \longrightarrow f_i' = 10\%$; $f_i = 30\% \longrightarrow f_i' = 30\%$ ... En pratique une telle situation ne se rencontre qu'exceptionnellement (Hamdani, 2006).

Remarque

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

Exemple 1

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

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

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

Réponse :

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

Classe	x_i	n_i	N_i	f_i	F_i	$n_i x_i$	$f_i' = nixi/\Sigma nixi$	$F_i' = (nixi/\Sigma nixi)\uparrow$
0 - 4	*2*	*6*	*6*	*0,06*	*0,06*	*12*	*0,0092*	*0,0092*
4 - 8	*6*	*25*	*31*	*0,25*	*0,31*	*150*	*0,1154*	*0,1246*
8 - 12	*10*	*24*	***55***	*0,24*	***0,55***	*240*	*0,1846*	*0,3092*
12 - 16	*14*	*17*	*72*	*0,17*	*0,72*	*238*	*0,1831*	*0,4923*
16 - 22	*19*	*14*	*86*	*0,14*	*0,86*	*266*	*0,2046*	***0,6969***
22 - 30	*26*	*11*	*97*	*0,11*	*0,97*	*286*	*0,2200*	*0,9169*
30 - 42	*36*	*3*	*100*	*0,03*	*1*	*108*	*0,0831*	*1*
Total	-	*100*	-	*1*	-	*1300*	*1*	-

1- Calcul de la médiane

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

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

2- Calcul de la Médiale

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

⟶ $ML \in [16 - 22[$

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

3- Comparer ΔM à l'étendue de la distribution

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

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

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

Remarque

Ce résultat devra être confirmé par la méthode graphique. C'est l'objet de la section suivante.

Section 2 : L'analyse graphique de la concentration

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

Elle consiste à représenter graphiquement la concentration en traçant une courbe, dite *courbe de Lorenz* ou *courbe de Gini*, ou encore, *courbe de concentration.* A partir de cette courbe est déduit, par des opérations géométriques, un paramètre permettant de mesurer l'intensité de la concentration, appelé « *indice de Gini » (Boursin, 1991).*

2.1- La courbe de concentration

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

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

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

En reliant les points de coordonnées (0 ;0) et (1 ;1), on obtient une diagonale qui divise le carré en deux triangles de même surface, soit ½ ou 0,5 chacun (Boudia, 2008).

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

La diagonale représente la courbe où la concentration est nulle, c'est-à-dire, à chaque proportion d'individus correspond la même proportion de masse : $f_i = f_i'$. On l'appelle la droite d' « *équi-répartition* » ou droite de répartition équitable (Hubler, 2007).

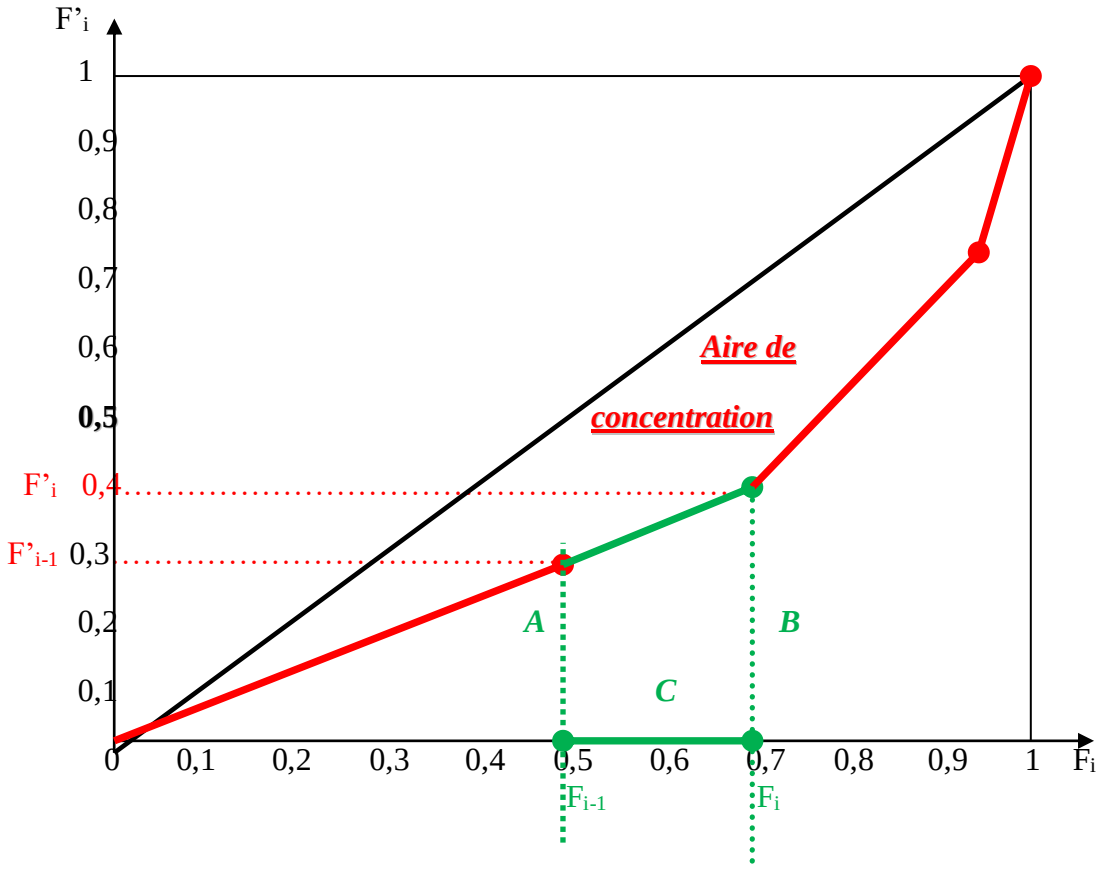
Par conséquent, plus la courbe de concentration se rapproche de la diagonale, plus la concentration est faible et plus l'*aire de concentration* se réduit, et vis versa. (Hubler, 2007).

Remarque

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

Comment tracer la courbe de concentration ?

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



2.2- L'indice de Gini

Noté I_G, il mesure l'intensité de la concentration, en mesurant le rapport de l'aire de concentration à la surface du triangle du bas. On aura alors (Hamdani, 2006) :

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

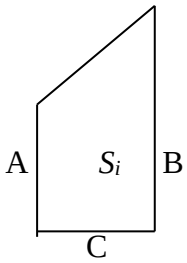
A partir de la courbe de ci-dessus, on déduit que l'aire de concentration est égale à :

S = 1/2 – (la somme des surfaces des trapèzes situés sous la courbe).

En effet, la surface sous la courbe est formée par des trapèzes dont nous illustrons l'un d'eux qui est formé par les cotés A, B et C (en couleur verte sur la courbe).

Illustration :

La surface d'un trapèze étant : $Si = \frac{(A+B)C}{2}$



A partir de la courbe, on déduit que :

$A = F'_{i-1}$

$B = F'_i$

$C = F_i - F_{i-1} = f'_i$

Il es résulte que :

$$Si = \frac{(F'_{i-1}+F'_i)\, f_i}{2}$$

I_G = 2. Aire de concentration = $2\left(1/2 - \left[\frac{\Sigma\,(F'_{i-1}+F'_i)\, f_i}{2}\right]\right)$

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

Remarque : On appelle cette méthode de détermination de I_G, la « *méthode des trapèzes* » (Anderson & al., 2007).

Interprétation de l'indice de Gini

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

- Si $I_G = 0$, ⟶ la concentration est nulle, l'aire de concentration est nulle. Autrement dit, I_G = *2. Aire de concentration* = 2.0 = 0. La courbe de concentration se confond avec la diagonale (la courbe de concentration est dans ce cas la diagonale elle même).

- Si = 1, ⟶ la concentration est maximale, l'aire de concentration est égale à toute la surface du triangle sous la diagonale qui est elle même égale à 1/2. Autrement dit, I_G = *2. Aire de concentration* = 2. 1/2 = 1. Dans ce cas la courbe de concentration s'éloigne au maximum de la diagonale.

- Ainsi ;
  - Si I_G tend vers 1, ($I_G > 0{,}5$), la concentration est dite *forte*.
  - Si I_G tend vers 0, ($I_G < 0{,}5$), la concentration est dite *faible* (Monino, 2017).

Exemple 2

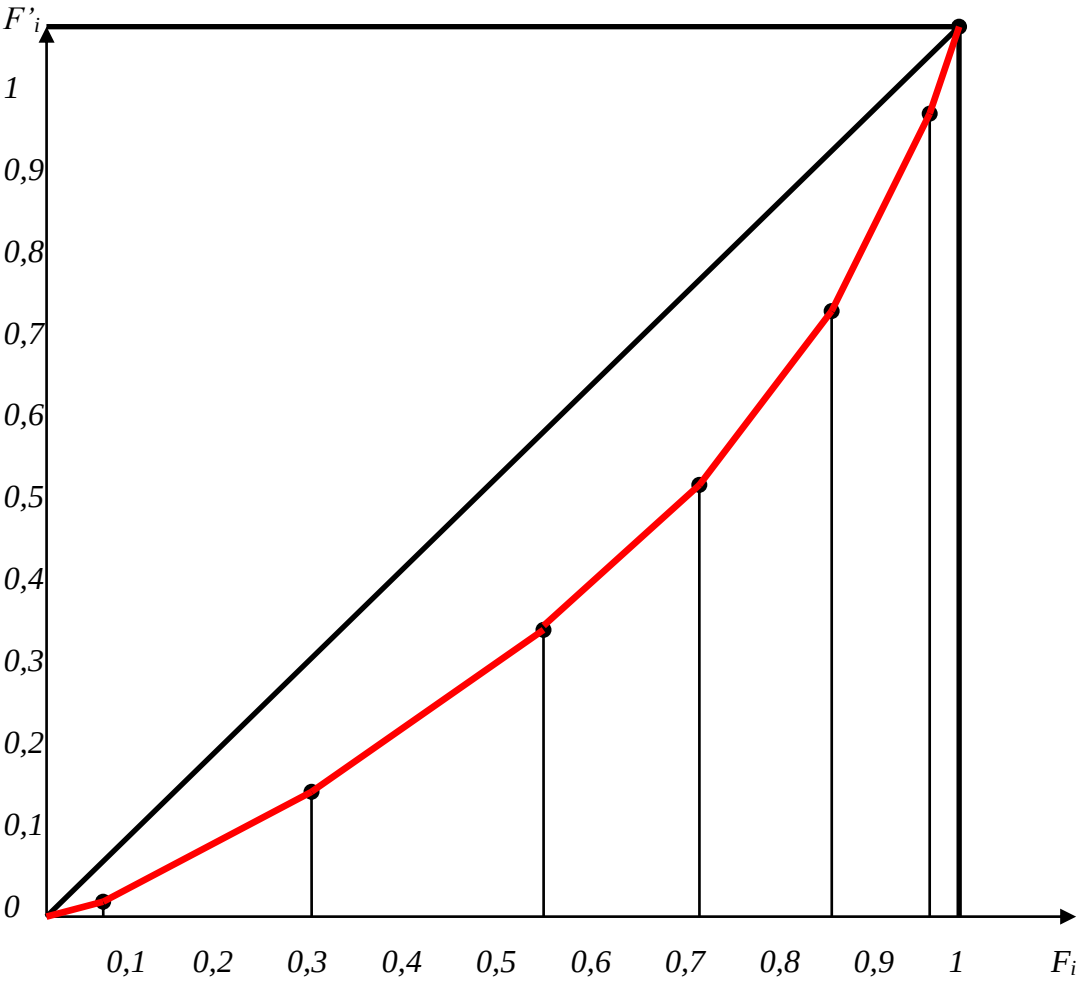
Reprendre l'exemple précédent et analyser la concentration par la méthode dite graphique.

Réponse : *Il s'agit, en fait, en plus de ce que l'on a déterminé dans l'exemple 1, de tracer la courbe de concentration et de calculer l'indice de Gini. Cependant, dans l'exemple précédent,*

nous avons constaté une concentration faible. Nous devons impérativement confirmer cette faiblesse par la méthode graphique. Les deux méthodes ne doivent pas se contrarier.

1- Tracer la courbe de concentration

Avec le contenu du tableau précèdent on peut d'abord tracer la courbe de concentration, comme suit :



On constate donc que la courbe de concentration est proche de la diagonale, ce qui nous permet de dire que la concentration est plutôt faible. On confirme donc le constat établi par la méthode algébrique.

2- Calcul de l'Indice de Gini

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

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

Remarque

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

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

$$\boxed{\mathbf{I_G = 0{,}325}}$$

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

Conclusion au chapitre 7

Au terme du présent chapitre, l'étudiant aura pris connaissance de l'essentiel des paramètres d'analyse et de synthèse, ou que l'on peut appeler simplement les caractéristiques des séries statistiques, notamment les plus couramment utilisés en statistique descriptive, à savoir ; les paramètres ou les caractéristiques de tendance centrale, de dispersion, de forme et de concentration.

La concentration est, comme on vient de le voir, ce paramètre qui donne une idée de la proportionnalité des modalités aux effectifs, plus précisément, il renseigne sur la proportionnalité des proportions d'individus aux proportions des masses (f_i par rapport à f'_i).

Ainsi, en plus de la caractérisation des distributions statistiques par des valeurs centrales (moyenne, médiane, mode), on peut également caractériser leur dispersion, comme on peut aussi caractériser leur concentration.

Au delà de ces trois types de paramètres, il serait temps de passer, dans la suite du

présent cours, à d'autres thèmes, plus large, de la statistique descriptive, notamment les distributions statistiques à deux caractères, …, qui seront donc développés dans le chapitre 9.

Cependant, avant d'entamer ces dernières, il serait utile pour l'étudiant, de prendre connaissance de la manipulation mathématique des pourcentages, à travers l'étude des *indices* ou les *nombres indices.* C'est l'objet du chapitre suivant.

Chapitre 8 : Les Indices de la vie économique

Introduction au chapitre 8

Dans la vie économique et sociale les grandeurs représentatives de phénomènes varient dans le temps et dans l'espace. De même que très souvent ces grandeurs sont difficiles à apprécier et à comparer.

Si on dit, par exemple, que la production d'un bien X passe de 106984 unités à 128380 unités, et que celle du bien Y passe de 385 unités à 462 unités, la comparaison immédiate n'est pas facile. Or, quand on fait le calcul on s'aperçoit que les productions des deux biens varient dans la même proportion : 20% chacune.

Aussi, en statistique, pour fournir des informations plus simples et plus faciles à lire, on fait recours à des paramètres ou ratios, appelés « *indices* », qui reflètent directement l'importance des variations. On comprend alors qu'un *indice* est un rapport et un indicateur sans dimension (sans unité de mesure). Ce n'est qu'une indication chiffrée caractérisant une évolution.

Le calcul des indices n'a de sens que si l'on veut comparer les valeurs d'une grandeur dans deux ou plusieurs états (dates ou lieux) différents (Py, 1996)

Le raisonnement le plus simple pour faire ces comparaisons serait de diviser une valeur par une autre et d'en tirer le résultat sous forme d'un nombre sans dimension ou en pourcentage.

La notion d'indice présente deux avantages : des facilités de lecture et des possibilités d'élargissement.

Les indices que l'on calcule sur une seule grandeur, ou grandeur simple, sont appelés « *indices élémentaires* » : ex ; indice du prix du café, indice du prix du lait, indice du prix de l'acier, indice du prix du blé, …

On peut également calculer des indices sur des grandeurs complexes, composées de plusieurs grandeurs simples. Ces indices sont appelés « *indices synthétiques* ». Ils sont le résultat de la synthèse ou de l'agrégation de plusieurs grandeurs simples (ex ; indice général des prix, indice des prix de la consommation, indice de la production industrielle, etc.).

Dans le présent chapitre, nous examinerons successivement ces deux types d'indices. Nous commencerons par les « *indices élémentaires* » (Section1), pour enchainer, ensuite, avec les « *indices synthétiques* » (Section 2).

Section 1 : Les indices élémentaires

On s'intéresse dans cette première section à la manière d'écrire un indice élémentaire et, surtout, à leurs propriétés qui sont très utiles en pratique. Avant cela, il convient de définir d'abord la notion d'*indice élémentaire*.

1.1. Définition

On appelle indice élémentaire, le nombre sans dimension résultant du rapport de deux valeurs prises par une même grandeur simple, soit à deux ou plusieurs dates différentes, soit sur deux ou plusieurs espaces (lieux) différents. (Py, 1996)

Si V_0, la valeur de la variable ou grandeur au temps, ou à la date, « 0 », et V_t sa valeur au temps ou à la date, « t », l'indice élémentaire « I » s'écrit :

$$\mathbf{I_{t/0} = V_t/V_0 . 100 \quad ou \quad i_{t/0} = V_t/V_0}$$

Et on l'énonce comme suit :

« *L'indice de la variable au temps t, base 100 (ou date de référence) en temps 0.* ».

Notons :

- La date « t = 0 » s'appelle date de *référence ou date de base* pour les indices temporels ou chronologiques, et devient *situation de base ou de référence* dans le cas des indices spatiaux ou de lieux.

- La date « t » s'appelle date ou période *courante* dans le cas d'indices temporels, et devient situation *courante* dans le cas des indices spatiaux ou de lieux.

Exemple

Le prix d'un bien X passe de 150 DA à 180 DA, de 2010 à 2013. L'indice du prix de ce bien X, en 2013, base 100 en 2010 est :

$I_{13/10}$ = (Prix en 2013 / Prix en 2010) . 100 = (180/150). 100 = 120 (ou 1,2), soit une augmentation de 20%.

Remarque

$I_{13/10} \neq I_{10/13}$

1.2. Propriétés des indices élémentaires

Les indices élémentaires remplissent plusieurs propriétés très utiles en pratique et que l'étudiant doit maîtriser pour mieux les manipuler (Py, 1996).

1.2.1. L'identité

Lorsque la date de référence est la même que la date de base, l'indice élémentaire est alors égale à 1 : $\mathbf{I_{0/0} = 100 ;\ I_{t1/t1} = 100}$; …(ou $\boxed{\mathbf{i_{0/0} = 1 ;\ i_{t1/t1} = 1}}$)

1.2.2. La circularité ou transférabilité

Si une grandeur économique prend les valeurs V_0, V_1, V_2,… ; aux temps 0, 1, 2, … ; l'indice élémentaire satisfait la relation suivante (Py, 1996) :

$$\mathbf{I_{2/0} = (I_{2/1} \,.\, I_{1/0}) \,.\, 1/100 = (V_2/V_1).\ 1/100 ;\ (ou\ \boxed{i_{2/0} = i_{2/1}.i_{1/0}})}$$

On peut aussi déduire: $\mathbf{I_{2/1} = [(I_{2/0})/(I_{1/0})].\ 100 \rightarrow (ou\ \boxed{i_{2/1} = i_{2/0} \,/\, i_{1/0}})}$

Autrement dit, pour comparer deux grandeurs simples à deux dates t = 1 et t = 2, il suffit de faire le rapport de leurs indices, fois 100.

Ainsi, on peut opérer des changements de bases sur les indices élémentaires, en substituant à la date 0 la date t = 1 qui est une date intermédiaire entre les dates 0 et 2.

On peut généraliser la propriété de circularité et démontrer que les indices sont « *enchaînables* », ils s'enchainent :

$$\mathbf{I_{t/0} = [(I_{1/0})/100.\ I_{2/1}/100 \,.\, I_{3/2} \,.\, \ldots\ldots\ldots I_{t/t-1}].\ 100}$$

Ou

$$I_{t/0} = [i_{1/0} . i_{2/1} . i_{3/2} . \ldots i_{t/t-1}]$$

Par exemple, l'indice du prix d'un bien en 2014, base 100 en 2009, est :

$$I_{14/09} = [I_{14/13}/100 . I_{13/12}/100 . I_{12/11}/100 . I_{11/10}/100 . I_{10/09}/100]. 100$$

Ou

$$i_{14/09} = [i_{14/13} . i_{13/12} . i_{12/11} . i_{11/10} . i_{10/09}]$$

Autrement dit, la propriété de circularité permet d'obtenir l'indice élémentaire de la date « *t* » par rapport à la base, en effectuant *le produit des indices élémentaires intermédiaires successifs*.

1.2.3. La réversibilité

Cette propriété s'énonce comme suit : quand on inverse le rôle de la base et de la période courante**,** l'indice élémentaire s'inverse à $(100)^2$ près. Ou bien, on obtient l'inverse de l'indice initial. On écrit cela comme suit (Py, 1996) :

$$I_{t/0} . I_{0/t} = (100)^2 \rightarrow I_{t/0} = (100)^2 / I_{0/t}$$

ou

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

1.2.4. Indices élémentaires de grandeurs liées par un produit

Si une grandeur simple (A) est le produit de deux autres grandeurs simples (B et C), l'indice élémentaire de A est égale au produit des indices élémentaires de B et C. On écrit :

$$A = B . C \longrightarrow I_{t/0\,(A)} = I_{t/0\,(B)} . I_{t/0\,(C)} . 1/100$$

ou

$$i_{t/0\,(A)} = i_{t/0\,(B)} . i_{t/0\,(C)}$$

C'est le cas par exemple de l'indice de valeur. La valeur étant égale au prix fois la quantité, on aura *l'indice élémentaire de valeur* égale :

$$I_{t/0(PQ)} = [I_{t/0(P)} . I_{t/0\,(Q)}]. 1/100$$

Ou

$$i_{t/0}\,(PQ) = i_{t/0}(P) . i_{t/0}(Q)$$

L'indice de valeur peut également s'écrire, si nous connaissons les valeurs de P_i et Q_i , sous la forme du rapport des produits des deux valeurs à deux dates différentes :

$$=[P_t^i Q_t^i / P_0^i Q_0^i]. 100^2. 1/100$$

$$= [P_t^i Q_t^i / P_0^i Q_0^i]. 100 \quad \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

Ou

$$I_{t/o} = P_t Q_t / P_0 Q_0 = P_t / P_0 . Q_t / Q_0$$

$$= \boxed{i_{t/0}\,(PQ) = i_{t/0}(P)\,.\,i_{t/0}(Q)}$$

1.2.5. Indices élémentaires de grandeurs liées par un rapport

Quand une grandeur simple (A) est le rapport de deux autres gradeurs simples (B et C), l'indice élémentaire de A est égale au rapport des indices élémentaires de B et C. On écrit :

$$A = B/C \longrightarrow I_{t/0}\,(A) = I_{t/0}\,(B) \,/\, I_{t/0}\,(C)\,.\,100$$

ou

$$\boxed{i_{t/0}\,(A) = i_{t/0}\,(B) \,/\, i_{t/0}\,(C)}$$

Remarque

Il est conseillé à l'étudiant de retenir et de privilégier dans ses exercices les formules encadrées car elles sont plus simples et donc plus pratiques.

Section 2 : Les indices synthétiques

Il s'agit dans la présente section de présenter successivement l'écriture d'un indice synthétique et ses différents types. Cependant, au préalable, il convient de définir d'abord ce que l'on entend par indice synthétique.

2.1- Définition

Un indice synthétique mesure la variation d'une grandeur complexe constituée par un ensemble de plusieurs grandeurs simples. Autrement dit, un indice synthétique résume une série d'indices élémentaires (Bressoud & Kahané, 2009).

Il existe trois types d'indices synthétiques : l'indice des prix, l'indice des quantités et l'indice des valeurs (PQ). L'indice de valeur dépend de celui des deux autres.

Si l'indice de valeur varie, on ne peut savoir si c'est dû à une variation des prix ou à celle des quantités. Une manière de lever ce doute consiste à poser l'hypothèse qu'une des deux variables (P ou Q) est fixe pendant que l'autre varie. Autrement dit, pour calculer l'indice de valeur d'un bien ***i***, entre deux dates différentes, il suffit d'éliminer respectivement l'influence de l'une et de l'autre des variables. Ainsi, pour faire ressortir les variations de prix d'un bien entre deux dates, il suffit d'éliminer l'influence des quantités, c'est-à-dire calculer ce qu'aurait été la valeur globale à l'année ***t*** si les quantités étaient restées fixes et seuls les prix avaient variés, et inversement.
On aura donc, conformément à l'équation (1) ci-dessus (Py, 1996) :

- <u>*Quantités constantes*</u> :

$$I_{t/0(PQ)} = [\ \cancel{P_t^i}\cancel{Q_0^i}/\cancel{P_0^i}\cancel{Q_0^i}\ ].\ 100 = [\ P_t^i/P_0^i\ ].1/100$$

Ou

$$i_{t/0(PQ)} = [\ \cancel{P_t^i}Q_0^i/\cancel{P_0^i}Q_0^i\ ] = [\ P_t^i/P_0^i\ ].$$

On revient donc à l'indice élémentaire des prix.

- <u>*Prix constants*</u> :

$$I_{t/0(PQ)} = [\ \cancel{P_0^i}Q_t^i/\cancel{P_0^i}Q_0^i\ ].\ 100 = [\ Q_t^i/Q_0^i\ ].1/100$$

Ou

$$i_{t/0(PQ)} = [\cancel{P_0^i} Q_t^i / \cancel{P_0^i} Q_0^i] = [Q_t^i / Q_0^i]$$

On revient donc à l'indice élémentaire des quantités.

Passer de l'indice élémentaire à l'indice synthétique, revient à considérer non pas un seul bien mais un ensemble des biens, l'ensemble des biens composant la grandeur complexe. En suivant cette méthode et en fonction de la date de référence (ou de base) choisie, on peut avoir trois type d'indices synthétiques construits par trois auteurs différents. Le choix de la date de référence constitue la spécificité de chaque indice (Hamdani; 2006).

2.2- Les formules simplifiées des indices synthétiques

2.2.1. L'indice de Laspeyres

Laspeyres propose, en prenant comme date de référence (ou de base) une date antérieure à la date d'observation (ou actuelle), c'est-à-dire la date « 0 », deux indices selon que l'on fixe les prix ou les quantités à cette date « 0 » (Bressoud & Kahané, 2009).

2.2.1.1- L'indice de Laspeyres des prix

Noté $L_{t/0}^{p}$, il s'obtient en figeant les quantités à la date « 0 » , on écrit :

$$L_{t/0}^{p} = [\sum_{i=1}^{k} P_t^i Q_0^i] / [\sum_{i=1}^{k} P_0^i Q_0^i] \ . 100$$

2.2.1.2- L'indice de Laspeyres des quantités

Noté $L_{t/0}^{q}$, il s'obtient en figeant les prix à la date « 0 », on écrit :

$$L_{t/0}^{q} = [\sum_{i=1}^{k} P_0^i Q_t^i] / [\sum_{i=1}^{k} P_0^i Q_0^i] \ . 100$$

2.2.2- L'indice de Paasche

Paasche propose, en prenant comme date de référence (ou de base) la date actuelle ou la date d'observation, c'est-à-dire la date « t », deux indices selon que l'on fixe les prix ou les quantités à cette date « t ». (Bressoud & Kahané, 2009).

2.2.2.1- L'indice de Paasche des prix

Noté $P_{t/0}^{p}$, il s'obtient en figeant les quantités à la date « t », on écrit :

$$P_{t/0}^{p} = [\sum_{i=1}^{k} P_t^i Q_t^i] / [\sum_{i=1}^{k} P_0^i Q_t^i] \ . 100$$

2.2.2.2- L'indice de Paasche des quantités

Noté $P_{t/0}^{p}$, il s'obtient en figeant les prix à la date « t », on écrit :

$$P_{t/0}^{q} = [\sum_{i=1}^{k} P_t^i Q_t^i] / [\sum_{i=1}^{k} P_t^i Q_0^i] \ . 100$$

2.2.3- L'indice de Fisher

C'est un indice intermédiaire entre les deux précédents. Noté F^p, il est la *moyenne géométrique* des indices de Laspeyres et de Paasche (Bressoud & Kahané, 2009) ;

L'indice de Fisher des prix :

$$F^p = \sqrt{(L^p \ . \ P^p)}$$

L'indice de Fisher des quantités :

$$\mathbf{F^q = \sqrt{L^q . P^q}}$$

On note en général que :

- L'indice de Fisher est compris entre ceux de Laspeyres et de Paasche :

$$\mathbf{P \leq F \leq L}$$

- L'indice de valeur ou des dépenses, noté (I(v), I(PQ) ou I(D), calculé par Laspeyres est égale à celui calculé par Paasche, et est égale au produit de l'indice de Fisher des prix par l'indice de Fisher des quantités. Ce qui nous permet d'écrire :

$$\mathbf{I^{(PQ)} = L^{pq} = P^{pq} = F^p . F^q}$$

Cependant, il faut souligner que les formules que l'on vient de développer sont dites « *formules simplifiées* » et ne sont pas uniques. D'autres formules existent et on les applique fréquemment en pratique, appelées « *formules pondérées ou définition* ». C'est ce qu'on montre dans ce qui suit.

2.3- Les formules pondérées des indices synthétiques

Il faut savoir que l'on peut écrire, par un jeu de manipulations mathématiques, les indices de Laspeyres et de Paasche sous forme de moyennes arithmétique et/ou harmonique (Py, 1996).

On peut alors écrire ce qui suit.

2.3.1- Les formules de définition de l'indice de Laspeyres

2.3.1.1- L'indice de Laspeyres des prix

- L'indice de Laspeyres des prix est la ***moyenne arithmétique*** de la série des indices élémentaires de prix ($i^p_{t/0} = P_t/P_0$), pondérée par les **valeurs globales de la date de base « 0 » ; ($P_0Q_0/\Sigma P_0Q_0$),** notées **« a_0 »**. (Py, 1996) :

$$\boxed{L^p_{t/0} = [\sum_{i=1}^{k} P^i_0 Q^i_0 \ (P_t / P_0)] / [\sum_{i=1}^{k} P^i_0 Q^i_0]}$$

On reconnait la forme usuelle de la moyenne arithmétique : $X = \overline{\Sigma}\, n_i x_i / \Sigma n_i$. En effet ;

. $a_0 = P_0Q_0/\Sigma P_0Q_0$ correspond à l'écriture $n_i/\Sigma n_i = f_i$.

Les paramètres ($P_0Q_0 / \Sigma P_0Q_0$) s'appellent ***coefficients de pondération*** à la date « 0 », notés « a_0 ». Dans le cas des indices de prix on les appelle aussi ***coefficients budgétaires***. Ce qui permet d'écrire également (Py, 1996) :

$$\boxed{L^p_{t/0} = \Sigma[(i^p_{t/0}) \ . \ (a_0)]}$$

On reconnait encore la formule de la moyenne arithmétique : $\overline{X} = \Sigma\, f_i x_i$. En effet ;

. $a_0 = P_0Q_0 / \Sigma P_0Q_0$ correspond à l'écriture $n_i / \Sigma n_i = f_i$.

❖ L'indice de Laspeyres des prix est aussi la *moyenne* ***harmonique*** de la série des indices élémentaires de prix (P_t/P_0), pondérée par les valeurs globales de la **date de base « 0 » pour les quantités** et de la **date courante « t » pour les prix** (***P_tQ_0 / ΣP_tQ_0***). (Py, 1996).

$$L_{t/0}^{p} = [\sum_{i=1}^{k} P_t^i Q_0^i] / \sum_{i=1}^{k} [P_t^i Q_0^i)(P_0/P_t)]$$

On reconnait la formule de la moyenne harmonique : $H = \Sigma n_i/\Sigma n_i/x_i$

2.3.1.2- L'indice de Laspeyres des quantités

❖ L'indice de Laspeyres des quantités est la *moyenne* ***arithmétique*** de la série des indices élémentaires de quantités ($i_{t/0}^{q} = Q_t/Q_0$), pondérée par les **valeurs globales de la date de base « 0 »** ; (***$P_0Q_0 / \Sigma P_0Q_0$***). (Py, 1996).

$$L_{t/0}^{q} = [\sum_{i=1}^{k} P_0^i Q_0^i \ (Q_t / Q_0)] / [\sum_{i=1}^{k} P_0^i Q_0^i]$$

On reconnait la forme usuelle de la moyenne arithmétique : $\overline{X} = \Sigma\, n_i x_i/\Sigma n_i$.

Les paramètres ($P_0Q_0/\Sigma P_0Q_0$) s'appellent ***coefficients de pondération***, notés « a_0 », sont les mêmes que ceux définis plus haut. Ce qui permet d'écrire également (Py, 1996) :

$$L_{t/0}^{q} = \Sigma[(i_{t/0}^{q}) \,.\, (a_0)]$$

On reconnait encore la formule de la moyenne arithmétique : $\overline{X} = \Sigma\, f_i x_i$.

❖ L'indice de Laspeyres des quantités est également la *moyenne* ***harmonique*** de la série des indices élémentaires de quantités (Q_t/Q_0), pondérée par les valeurs globales de la **date de base « 0 » pour les prix** et de la **date courante « t » pour les quantités** (***$P_0Q_t/\Sigma P_0Q_t$***). (Py, 1996) :

$$L_{t/0}^{q} = [\sum_{i=1}^{k} P_0^i Q_t^i] / [\sum_{i=1}^{k} P_0^i Q_t^i (Q_0/Q_t)]$$

On reconnait la formule de la moyenne harmonique : $H = \Sigma n_i/\Sigma n_i/x_i$.

2.3.2- Les formules développées de l'indice de Paasche

2.3.2.1- L'indice de Paasche des prix

❖ L'indice des prix de Paasche est la ***moyenne arithmétique*** de la série des indices élémentaires des prix (P_t/P_0), **pondérée par les valeurs globales de la date courante « t » pour les quantités, et de la date de base « 0 » pour les prix** ; (***$P_0Q_t / \Sigma P_0Q_t$***). (Py, 1996).

$$P^{p}_{t/0} = [\sum_{i=1}^{k} P^{i}_{0} Q^{i}_{t} \ (P_t / P_0)] / [\sum_{i=1}^{k} P^{i}_{0} Q^{i}_{t}]$$

- L'indice des prix de Paasche est aussi la ***moyenne harmonique*** de la série des indices élémentaires des prix (P_t/P_0), pondérée par les **valeurs globales de la date courante « t » pour les prix et les quantités; ($P_tQ_t / \Sigma P_tQ_t$).** (Py, 1996).

$$P^{p}_{t/0} = [\sum_{i=1}^{k} P^{i}_{t} Q^{i}_{t}] / [\sum_{i=1}^{k} P^{i}_{t} Q^{i}_{t} (P_0 / P_t)]$$

Ou encore en termes de fréquences :

$$P^{p}_{t/0} = 1 / \Sigma[\frac{P_t Q_t}{\Sigma P_t Q_t} / P_t/P_0] = H = \Sigma fi/\Sigma(fi/xi)$$

Cette formule se résume comme suit :

$$P^{p}_{t/0} = 1/ \Sigma \ (a_t / \ i^{p}_{t/0}) \ = \ 1/ \Sigma \ a_t \ (P_0/P_t)$$

Avec $a_t = \frac{P_t Q_t}{\Sigma P_t Q_t}$ et $\Sigma \ a_t = 1$.

2.3.2.2- L'indice de Paasche des quantités

Sous forme de moyenne arithmétique on aura : (Py, 1996).

$$P^{q}_{t/0} = [\sum_{i=1}^{k} P^{i}_{t} Q^{i}_{0} \ (Q_t / Q_0)] / [\sum_{i=1}^{k} P^{i}_{t} Q^{i}_{0}]$$

Sous forme de moyenne harmonique on aura :

$$P^{q}_{t/0} = [\sum_{i=1}^{k} P^{i}_{t} Q^{i}_{t}] / [\sum_{i=1}^{k} P^{i}_{t} Q^{i}_{t} (Q_0 / Q_t)]$$

Remarque 1

- L'indice de Fihcer étant toujours la moyenne géométrique des deux indices de Laspeyres et Paasche.

- Les formules pondérées ou de définition sont très utiles lorsque à la place des valeurs nominales ou absolues on dispose des indices élémentaires et des proportions des valeurs.

Remarque 2

On résume souvent les formules de définition (pondérées) en écrivant :

- L'indice de Laspeyres des prix est une moyenne arithmétique des indices élémentaires de prix ($P_t/P_0 = i_{t/0}$), pondérée par les coefficients de pondération ou budgétaires (valeurs globales : $P_0Q_0/\Sigma P_0Q_0 = a_0$) de la date de base « 0 ». On écrit (Py, 1996) :

$$L^{p}_{t/0} = \Sigma\,(a_0 \,.\, i_{t/0}) = \Sigma\,[a_0 \,.\, (P_t/P_0)]$$

- L'indice de Paasche des prix est une moyenne harmonique des indices élémentaires de prix, pondérée par les coefficients de pondération ou budgétaires de la période courante « t » : $P_tQ_t / \Sigma P_tQ_t = a_t$. On écrit (Py, 1996).:

$$P^{p}_{t/0} = 1/\,\Sigma\,(a_t / i^{p}_{t/0}) = 1/\,\Sigma\, a_t\,(P_0/P_t)$$

Remarque 3

Les coefficients de pondération (a_0 et a_t) étant des proportions, il en résulte que leurs sommes sont toujours égales à 1.

2.4. Complément

2.4.1. Indice synthétique de valeur (ou des dépenses)

Comme nous l'avons souligné plus haut (Cf, §2.2.3), l'indice de valeur $I_{t/0}(V)$ ou des dépenses $I_{t/0}(D)$ ou $I_{t/0}(PQ)$ est le produit d'un indice de prix et d'un indice de quantité. A partir de là on peut écrire ce qui suit (Py, 1996). :

$$I_{t/0}(V) = L^{p}_{t/0} \cdot P^{q}_{t/0} = P^{p}_{t/0} \cdot L^{q}_{t/0} = L^{pq} = P^{pq} = F^{p}.F^{q} = [\textstyle\sum_{i=1}^{k} P^i_t Q^i_t] / [\sum_{i=1}^{k} P^i_0 Q^i_0]$$

Cela signifie qu'à partir de la dernière équation, c'est-à-dire celle de l'indice de valeur (qui est, rappelons-le, le rapport entre la valeur (ou dépense) totale de la période courante « t » et la valeur (ou dépense) totale de la période de base ou de référence « 0 »), on peut déduire l'indice de Laspeyres ou de Paasche (Cf, question n° 3, exercice n° 6, série n°7).

2.4.2. Indice du pouvoir d'achat

Le pouvoir d'achat est ce que l'on peut acheter en fonction des variations des prix, de la valeur de la monnaie ou du salaire. Selon les situations (énoncés des exercices pratiques), l'indice du pouvoir d'achat se détermine différemment. On peut écrire dans un premier temps (Boudia, 2008) :

$$i_{(PA)t/0} = i_{(Q)\,t/0} / i_{(P)t/0}$$

- D'abord, s'il s'agit du pouvoir d'achat de la monnaie, c'est-à-dire lorsque l'on ne dispose que des variations de prix. On suppose alors que les quantités sont fixes ($i_{(Q)\,t/0} = 1$). Dans ce cas, l'indice du pouvoir d'achat est l'inverse de l'indice des prix. Il concerne la capacité d'achat d'une monnaie (Boudia, 2008) :
:

$$i_{(PA)t/0} = 1/\, i_{(P)t/0}$$

il est clair que le pouvoir d'achat varie inversement avec les variations de prix. Lorsque les prix augmentent (*toute chose étant égale par ailleurs*), le pouvoir d'achat diminue, et inversement. Donc, si l'indice des prix augmentent, l'indice du pouvoir d'achat diminue, et inversement.

Dans ce cas aussi, il peut être exprimé par le rapport (l'évolution) des quantités entre la période de base « 0 » et la période de référence « t » :

$$i_{(PA)t/0} = Q_t/Q_0 = i_{(Q)t/0}$$

Une augmentation de l'indice des quantités implique une augmentation du pouvoir d'achat, et inversement. C'est-à-dire, le raisonnement inverse que par rapport à l'indice des prix.

Conclusion au chapitre 8

Au terme du présent chapitre, l'étudiant aura pris connaissance des fondamentaux et des éléments de base de définition et de calcul des indices économiques. Cependant, dans un souci de gestion du temps impartis, ne sont présentés que les conceptions théoriques des indices, les conceptions pratiques, utiles dans la compréhension des phénomènes économiques, notamment au niveau macro, ne sont pas abordés. Il en est de même des indices utilisés par les pays, notamment en Algérie, pour le calcul de certains agrégats conjoncturels qui sont également indispensables à connaitre. Aussi, nous invitons l'étudiant à approfondir et à enrichir ses connaissances en la matière par des lectures supplémentaires, notamment par la consultation de revues et sites internet spécialisés (ex ; l'ONS[1] en Algérie ; l'INSEE en France,…).

Ce chapitre constitue la transition entre deux parties du cours : celle développée jusque-là à travers les six premiers chapitres, et qui ont porté sur les distributions statistiques à un seul caractère ou une seule variable, et les chapitres suivants qui porteront sur les distributions statistiques à deux caractères ou deux variables. Le chapitre 9 suivant est une introduction à ces dernières.

[1] Office National des Statistiques (www.ons.dz) et ses publications périodiques, notamment les annuaires statistiques et les bulletins.

Chapitre 9 : Distributions à deux caractères, corrélation et régression

Introduction au chapitre 9

On l'aurait compris au début du cours que l'étude d'une population statistique peut porter sur un ou plusieurs caractères à la fois. Nous avons étudié jusque-là les distributions statistiques à un seul caractère ou une seule dimension, qu'on appelle aussi, les distributions *univariées* (une seule variable). L'objet du présent chapitre est d'initier l'étudiant aux distributions statistiques à deux caractères, ou *bivariées* (à deux variables). L'étude des distributions statistiques à plusieurs caractères, ou *multivariées*, étant réservée aux semestres 3 et 4.

Etudier une population d'étudiants suivant leurs poids et leurs tailles, le revenu et la consommation des ménages, les quantités d'engrais et la production agricole,… ; constituent autant d'exemples de distributions statistiques à deux caractères.

Pour savoir étudier ce type de distribution, l'étudiant doit d'abords apprendre à les présenter, notamment dans un tableau qu'on appelle le *tableau de contingence,* comme il doit aussi se familiariser avec le langage propre à ce type de distribution. Construire, compléter et lire correctement ces tableaux, avec le langage approprié sont les premiers éléments traités dans le présent chapitre (Section 1).

La section 2 initie l'étudiant au calcul des caractéristiques des distributions statistiques à deux caractères, notamment les *moyennes arithmétiques* et les *variances*, qui sont indispensables à l'analyse de la relation entre les deux caractères.

Il faut savoir, en effet, que si l'on s'encombre à construire des séries à deux caractères, alors que l'on peut étudier chaque caractère séparément dans des séries à un seul caractère, c'est justement pour pouvoir déterminer la relation ou le lien qui unit les deux caractères ou les deux variables. L'étude de cette relation qu'on appelle la *régression* est énoncée dans la section 3.

Section 1 : Présentation et notions fondamentales des distributions à deux caractères

Les distributions statistiques à deux caractères, appelées aussi distributions à deux dimensions, ou encore, distributions bi-variées, disposent, comme les distributions à un seul caractère, de leurs méthodes de présentation (tableau et graphique), de leurs propres caractéristiques, ainsi que de leur propre vocabulaire (notions fondamentales). Nous examinons dans ce qui suit, respectivement, la présentation et les notions fondamentales propres aux distributions à deux dimensions.

1.1- Le tableau de contingence

C'est le tableau statistique qui permet de présenter simultanément les deux caractères (variables) de la distribution statistique conjointe ou bi-variée. Il est structuré comme indiqué ci-dessous.

- « i » désigne la ligne et « j » désigne la colonne. On représente un caractère dans les lignes (généralement X) et l'autre caractère en colonne (généralement Y). Ainsi, l'effectif « n_{ij} » représente le nombre d'individus présentant à la fois la modalité « x_i » et la modalité « y_j ». Cet effectif s'appelle « *effectif partiel* ».

Remarques :

1. ***n_{ji} n'existe pas !***
2. ***Lorsque les données ne sont pas pondérées (pas d'effectifs associés), le tableau de contingence***

« Structure d'un tableau statistique d'une distribution statistique à deux caractères pondérés par leurs effectifs (ou tableau de contingence) »

Y / X	Y_1	Y_2	.....	y_j	.....	y_p	$n_{i.}$
X_1	n_{11}	n_{12}	.....	n_{1j}	.....	n_{1p}	$\mathbf{n_{1.}}$
X_2	n_{21}	n_{22}	.....	n_{2j}	.....	n_{2p}	$\mathbf{n_{2.}}$
.	.	.		.		.	.
.	.	.		.		.	.
.	.	.		.		.	.
.	.	.		.		.	.
.	.	.		.		.	.
X_i	n_{i1}	n_{i2}	.....	n_{ij}	.....	n_{ip}	$\mathbf{n_{i.}}$
.	.	.		.		.	.
.	.	.		.		.	.
.	.	.		.		.	.
.	.	.		.		.	.
.	.	.		.		.	.
X_k	n_{k1}	n_{k2}	.....	n_{kj}	.....	n_{kp}	$\mathbf{n_{k.}}$
$\mathbf{n_{.j}}$	$\mathbf{n_{.1}}$	$\mathbf{n_{.2}}$	.....	$\mathbf{n_{.j}}$	.....	$\mathbf{n_{.p}}$	$\mathbf{n_{..}}$

« Structure d'un tableau statistique d'une distribution statistique à deux caractères pondérés par leurs effectifs (ou tableau de contingence) »

- « i » désigne la ligne et « j » désigne la colonne. On représente un caractère dans les lignes (généralement X) et l'autre caractère en colonne (généralement Y). Ainsi, l'effectif « n_{ij} » représente le nombre d'individus présentant à la fois la modalité « x_i » et la modalité « y_j ». Cet effectif s'appelle « *effectif partiel* ». (Py, 1996).

Remarques :

1. ***n_{ji} n'existe pas !***

2. Lorsque les données ne sont pas pondérées (pas d'effectifs associés), le tableau de contingence n'aura que des 1 sur la diagonale et tous les autres effectifs partiels seront égaux à 0. Dans ce cas on peut se passer du tableau de contingence pour calculer les paramètres de la série.

3. Dans certains cas, il arrive qu'on présente le tableau avec les fréquences conditionnelles selon x (voir plus bas sous-titre 1.2.5), toutes les sommes en lignes sont égales à 100% ou 1. Le tableau s'appelle alors « tableau des profils en lignes ». On peut également, suivant la même logique, présenter le tableau avec les fréquences conditionnelles selon y et avoir le « tableau des profils en colonnes » où toutes les sommes en colonnes sont égales à 100% ou 1.

- La dernière ligne et la dernière colonne du tableau s'appellent les « *Marges* ». Les effectifs

portés sur ces marges s'appellent « *effectifs marginaux* ». Le point « **.** » désigne l'élément (la ligne i ou la colonne j) qui varie.

- Si on associe la première colonne du tableau à la dernière colonne, on obtient la distribution à un seul caractère suivant le caractère X. On l'appelle la ***distribution marginale de X*** (en vert sur le tableau). Elle est constituée par les couples (x_i ; $n_{i.}$).

- Si on associe la première ligne du tableau de contingence (y_j) à la dernière ligne ($n_{.j}$) on obtient tout simplement la distribution à un seul caractère suivant le caractère « Y » qu'on appelle ***distribution marginale de Y*** (en rouge sur le tableau). Elle est formée par les couples (y_j ; $n_{.j}$).

- La somme de la dernière colonne est égale à la somme de la dernière ligne et est égale à l'effectif total « $n..$ ». On peut écrire (Py, 1996) :

$$\sum_{i=1}^{k} n.j = \sum_{j=1}^{p} ni. = n.. = N$$

- La répartition des $n..$ individus de la population suivant les couples (x_i ; y_j) s'appelle "*distribution conjointe*".

1.2- Notions fondamentales des distributions bivariées

1.2.1- Notion d'effectif marginal

Ils sont notés « $n_{i.}$ » pour les lignes, et « $n_{.j}$ » pour les colonnes. On les obtient en faisant la somme par ligne pour les « $n_{i.}$ », et en faisant la somme par colonne pour les « $n_{.j}$ ». On écrit alors :

$$n_{i.} = \sum_{j=1}^{p} nij \quad \text{et} \quad n_{.j} = \sum_{i=1}^{k} nij$$

Le point « **.** » désigne l'élément (la ligne i ou la colonne j) qui varie.

1.2.2- Distribution conditionnelle

Elles sont déterminées ; soit selon X, soit selon Y.

1.2.2.1- Distribution conditionnelle de X selon Y

Elle signifie la distribution du caractère X selon, ou sous condition, que $y = y_j$, c'est-à-dire à la condition que y soit fixé à l'une de ses modalités. On obtient alors la distribution suivant les couples (x_i ; n_{ij}). (Py, 1996).

Par exemple, la première colonne du tableau ($y = y_1$) représente la distribution conditionnelle de x selon $y = y_1$ (ou sous condition $y = y_1$). On obtient alors la distribution suivant les couples (x_i ; n_{i1}), comme suit :

« Distribution conditionnelle de X selon Y = y_1

X	Y_1
x_1	n_{11}
x_2	n_{21}
.	.
.	.
x_i	n_{i1}
.	.
.	.
x_k	n_{k1}
$n_{.j}$	$n_{.1}$

1.2.2.2- Distribution conditionnelle de Y selon X

Elle signifie la distribution du caractère Y selon, ou sous condition, que x = xi, c'est-à-dire à la condition que X soit fixé à l'une de ses modalités. On obtient alors la distribution suivant les couples (y_j ; n_{ij}). (Py, 1996).

Par exemple, la première ligne du tableau ($x = x_1$) représente la distribution conditionnelle de Y selon $x = x_2$ (ou sous condition que $x = x_1$). On obtient alors la distribution conditionnelle de Y suivant les couples (y_j ; n_{1j}).

« Distribution conditionnelle de Y selon X = x_1»

Y	y_1	y_2	.....	y_j	.....	y_p	$n_{i.}$
X_1	n_{11}	n_{12}	.....	n_{1j}	.....	n_{1p}	$n_{1.}$

Au final, on aura donc autant de distributions conditionnelles de y selon x qu'il y a de modalités pour x, et autant de distributions conditionnelles de x selon y, qu'il y a de modalités pour y.

1.2.3- Notion de fréquence marginale

On appelle fréquences marginales, notées « $f_{i\ .}$ » ou « $f_{.j}$ », les rapports des effectifs marginaux (« $n_{i.}$ » et/ou « $n_{.j}$ ») à l'effectif total (« n.. »). On écrit :

$$\boxed{f_{i.} = n_{i.}/n..} \quad \text{et} \quad \boxed{f_{.j} = n_{.j}/n..}$$

Avec $f_{i.}$ et $f_{.j} \in [0 ; 1]$ et $\Sigma f_{i.} = \Sigma f_{.j} = 1$.

$f_{i.}$ représente la proportion des individus présentant la modalité x_i, par rapport à l'effectif total, quelque soit les modalités de y. (Hamdani, 2006).

$f_{.j}$ représente la proportion des individus présentant la modalité y_j, par rapport à l'effectif total, quelque soit les modalités de x.

1.2.4- Notion de fréquence partielle sur effectif total

On appelle fréquence partielle sur effectif total, notée « f_{ij} », le rapport de l'effectif partiel « n_{ij} » sur l'effectif total « **n..** ». On écrit :

$$\boxed{f_{ij} = n_{ij} /n..} \qquad \text{avec } \Sigma f_{ij} = 1$$

Ainsi, « f_{ij} » représente la proportion des individus présentant simultanément la modalité x_i et la modalité y_j, par rapport à l'effectif total de la population étudiée.

1.2.5- Notion de fréquence conditionnelle

On appelle fréquence conditionnelle « $f_{i/j}$ » ou « $f_{j/i}$ », le rapport de l'effectif partiel à l'effectif marginal correspondant. On écrit alors (Hamdani, 2006) :

$$\boxed{f_{i/j} = n_{ij}/n_{.j}} \qquad \text{et} \qquad \boxed{f_{j/i} = n_{ij}/n_{i.}}$$

« $f_{i/j}$ » se lit : *f_i si j,* soit ; *la fréquence conditionnelle de x = xi si y = y_j = constante*. Ou bien *fréquence conditionnelle de x = x_i si y est fixé (ou constant) à la modalité ou colonne « j », ou sous condition que y = y_j,* ou ; *par rapport à la colonne j ».* La lecture se fait dans ce cas à la verticale. Autrement dit, on croise toutes les lignes par la colonne « j ». (Py, 1996).

$f_{j/i}$ se lit « *f_j si i* », soit ; « *fréquence conditionnelle de y = y_j si x = x_i = constante* ». ou bien, « *fréquence conditionnelle de y = y_j sous condition que x = x_i* ou *par rapport à ligne i* ».

Remarque 1

- Quand il est demandé de calculer par exemple $f_{3/4}$, on ne peut pas savoir s'il est de type « $f_{i/j}$ » ou « $f_{j/i}$ » . Aussi, dans ce cas, on ajoute toujours une mention pour préciser quelle est la

colonne et quelle est la ligne. On dira alors, « $f_{3/4}$ si i fixé ou constant », ce qui signifie que l'on raisonne par rapport à la ligne. Autrement dit, on est devant la lecture de type *f_j si i est constant* ou *f_j sous condition que i = 4*. Donc, notre fréquence est de type « $f_{j/i}$ ». On écrit (Py, 1996).:

$$f_{j/i} = n_{ij}/n_{i.} \quad \rightarrow \quad f_{3/4} = n_{43}/n_{4.}$$

Si, au contraire, on précise que j est fixé, cela signifie que $f_{3/4}$ se lit : fréquence conditionnelle de i sous condition que j = 4 ou fixé à 4. Dans ce cas, elle est donc, de type $f_{i/j}$. On raisonne donc par rapport à la quatrième colonne. On écrit (Py, 1996).:

$$f_{i/j} = n_{ij}/n_{.j} \quad \rightarrow \quad f_{3/4} = n_{34}/n_{.4}$$

Remarque 2

- Le produit des fréquences marginales par les fréquences conditionnelles est égale aux fréquences partielles sur effectif total :

$$\mathbf{f_{ij} = f_{i/j} \, . \, f_{.j}} \quad \text{et} \quad \mathbf{f_{ij} = f_{j/i} \, . \, f_{i.}}$$

- Si les fréquences conditionnelles sont identiques aux fréquences marginales, cela signifie que les deux variables X et Y sont <u>*indépendantes*</u> ou <u>*ne sont pas liées*</u> :

$$\mathbf{f_{i.} = f_{i/j}} \quad \text{et} \quad \mathbf{f_{.j} = f_{j/i}}$$

Ou

$$\mathbf{n_{i.}/n_{..} = n_{ij}/n_{.j}} \quad \text{et} \quad \mathbf{n_{.j}/n_{..} = n_{ij}/n_{i.}} \qquad \text{(Voir §1.2.6 plus bas)}$$

- Lorsqu'il y a indépendance entre X et Y : $f_{i/j} = f_{i.}$ et $f_{j/i} = f_{.j}$ c'est-à-dire tous les <u>profils en colonne sont identiques</u> ($f_{i/j}$ = constante) et sont égaux aux fréquences marginales de Y ($f_{.j}$) , avec $f_{.j} = 1$ (somme par colonne du haut vers le bas). Ce qui revient à dire que :

$f_{1/1} = f_{1/2} = f_{1/3} \ldots\ldots\ldots\ldots = f_{i.}$

$f_{2/1} = f_{2/2} = f_{2/3} \ldots\ldots\ldots f_{2.}$; etc. Avec $f_{.j} = 1$ et $f_{i.} = 1$. (Py, 1996).

Ce qui veut dire que la fréquence conditionnelle de X_1 pour Y_1 est la même que celle de X_1 pour Y_2 et la même aussi que celle de X_1 pour Y_3 ;… et celle de X_1 pour Y_p, etc. Au final, les fréquences conditionnelles de X_i quelque soient les modalités de Y_j , ce qui fait que tous les profils en colonne sont identiques et aussi égaux aux fréquences marginales de X ($\mathbf{f_{i.}}$) :

$$f_{i/j} = f_{i.}$$

Autrement dit, les modalités de X_i ne dépendent pas de celles de Y_i et inversement : X et Y sont <u>indépendantes</u>.

On peut aussi, par symétrie, retrouver la même indépendance entre X et Y si tous les <u>profils en ligne sont identiques</u> et égaux aux fréquences marginales de Y ($\mathbf{f_{.j}}$) ; $f_{j/i} = f_{.j}$; la somme des $f_{i.} = 1$. Autrement dit, les fréquences conditionnelles de Y_i ; ($f_{j/i}$) ; quelque soient les modalités de X_i , sont identiques et aussi égalent aux fréquences marginales de Y ; ($f_{.j}$). (Py, 1996).

Pour résumer, nous pouvons dire que le *tableau des profils en ligne* donne les $\mathbf{f_{i.} = 1}$, c'es à dire les sommes à l'<u>horizontale</u>, et le *tableau des profils en colonne* donne les $\mathbf{f_{.j} = 1}$, c'est à dire les sommes à la <u>verticale</u>.

- Dans le cas de la <u>*dépendance totale*</u> ou <u>*liaison fonctionnelle et réciproque*</u>, il n'y a qu'un seul effectif partiel « n_{ij} » ($\neq 0$) par ligne et par colonne. C'est-à-dire sur chaque ligne et chaque colonne on ne trouve qu'un et un seul effectif partiel «n_{ij}» $\neq 0$, tous les autres sont nuls ou égaux à 0. Dans ce cas aussi, les moyennes conditionnelles sont égales aux valeurs des variables :

$$\mathbf{\overline{X_j} = x_i} \quad \mathbf{et} \quad \mathbf{\overline{Y_i} = y_j} \quad \textit{(voir plus bas paragraphe 2.2.2)}$$

Section 2 : Caractéristiques des distributions à deux caractères

Comme pour les distributions à un seul caractère, on peut calculer sur les distributions à deux caractères plusieurs paramètres. Cependant, comme on l'a vu dans la section précédente, il y a plusieurs niveaux de calcul. Aussi, pour les besoins de notre cours, nous ne retiendrons que les moyennes arithmétiques et les variances qui nous serons utiles pour l'analyse de la relation entre les deux variables X et Y.

2.1- Caractéristiques des distributions marginales

Il existe deux distributions marginales ; l'une suivant X, l'autre suivant Y. Aussi, nous déterminerons deux moyennes et deux variances.

2.1.1- Les moyennes marginales

- Suivant la distribution marginale de X, c'est-à-dire quelque soit Y, on obtient la moyenne arithmétique de la distribution à un seul caractère, suivant le caractère X, notée X. Elle est représentée par les couples (x_i ; n_i**.**). On écrit alors :

$$\overline{\mathbf{X}} = 1/n.. \sum_{i=1}^{k} xi \,.\, ni. \;=\; 1/n.. \sum_{i=1}^{k} \sum_{j=1}^{p} nijxi \quad (\text{ Puisque } n.. = \Sigma_{j=1}^{p} n_{ij})$$

- Suivant la distribution marginale de Y, c'est-à-dire quelque soit X, on obtient la moyenne arithmétique de la distribution à un seul caractère, suivant le caractère Y, notée Y. Elle est définie par les couples (y_j ; $n_{.j}$). On écrit alors (Py, 1996) :

$$\overline{\mathbf{Y}} = 1/n..\sum_{j=1}^{p} yjn.j \;=\; 1/n..\Sigma_{j=1}^{p} \sum_{i=1}^{k} nijyj$$

2.1.2- Les variances marginales

- Suivant la variable X, on aura :

$$\mathbf{V(X)} = 1/n.. \Sigma_{i=1}^{k} n_i.(x_i - \overline{X})^2 \quad = \quad 1/n.. \Sigma_{i=1}^{k} ni.(x_i^2) - (\overline{X})^2 \quad (\textit{formule développée})$$

- Suivant la variable Y, on aura :

$$\mathbf{V(Y)} = 1/n..\Sigma_{j=1}^{p} n_{.j}(y_j - \overline{Y})^2 \;=\; 1/n..\Sigma_{j=1}^{p} n_{.j}(y_j)^2 - (\overline{Y})^2 \quad (\textit{formule développée})$$

2.2- Caractéristiques des distributions conditionnelles

Comme signalé plus haut, il y a autant de distributions conditionnelles qu'il y a de modalités pour chacun des deux caractères. On les résume en deux écritures : distribution conditionnelle de X selon Y (y = constante) et distribution conditionnelle de Y selon X (x = constante). (Hamdani, 2006).

2.2.1- Distribution conditionnelle de X selon Y

Dans ce cas on déterminera autant de moyennes et de variances conditionnelles qu'il y a de modalités de Y ou de colonnes (c'est-à-dire p), soit une moyenne et une variance conditionnelles par colonne. On écrit alors :

$$\overline{\mathbf{X_j}} = 1/n_{.j} \sum_{i=1}^{k} nijxi \qquad (\boldsymbol{j} \textit{ indique la colonne ou la modalité de Y retenue})$$

Les variances seront de type :

$$\mathbf{V(X)_j} = 1/n_{.j}\Sigma_{i=1}^{k} n_{ij}(x_i - \overline{x_j})^2 \;=\; 1/n.j\Sigma_{i=1}^{k} n_{ij}(x_i^2) - (\overline{X_j})^2$$

Ainsi, si j varie de 1 à *p*, on aura *p* moyennes conditionnelles et *p* variances conditionnelles possibles de X selon Y.

2.2.2- Distributions conditionnelles de Y selon X

Dans ce cas on déterminera autant de moyennes et de variances qu'il y a de modalités de X ou de lignes, c'est-à-dire *k*. On écrit alors (Hamdani, 2006) :

$$\overline{Y}_i = 1/ni.\sum_{j=1}^{p} nij(yj) \qquad (i \text{ indique la ligne ou la modalité } x_i \text{ retenue})$$

Les variances seront de type :

$$V(Y)_i = 1/n_i.\Sigma_{j=1}^{p} nij(y_j - \overline{Y}_i)^2 \quad = \quad 1/n_i.\Sigma_{j=1}^{p} n_{ij}(y_j)^2 - (\overline{Y}_i)^2 \quad (\textit{formule développée})$$

Remarque 1

- Si les moyennes conditionnelles sont égales aux moyennes marginales pour les deux variables, cela signifie que les deux variables sont *indépendantes* ou *non liées* :

$$\overline{X}_j = X \quad \text{et} \quad Y_i = \overline{Y}$$

- Si les moyennes conditionnelles sont égales aux valeurs des variables, cela signifie qu'il y a *dépendance totale* ou *fonctionnelle* entre les deux variables :

$$\overline{Y}_i = Y_j \quad \text{et} \quad \overline{X}_j = X_i$$

Remarque 2

- Pour les distributions conjointes, au lieu de la variance, on calcule la « ***covariance*** ». Notée « **cov(x,y)** », elle est la moyenne arithmétique du produit des écarts à la moyenne des deux variables X et Y. Elle indique la tendance à la « co-variation » ou la variation simultanée, voire réciproque, des deux variables. On l'utilise surtout pour mesurer le degré de dépendance linéaire des deux variable. C'est donc un paramètre inévitable dans l'étude de la relation entre deux variable. On écrit alors :

- Dans le cas de séries simples ou non pondérées (n_{ij} = constantes = 1) :

$$Cov(x.y) = 1/N \sum_{i=1}^{N} (x_i - \overline{X})(y_j - \overline{Y}) = 1/N \sum_{i=1}^{N} (x_i y_j) - (\overline{X}.\overline{Y})$$

- Dans le cas de séries pondérées ($n_{ij} \neq$ constantes) :

$$Cov(x,y) = 1/n..\Sigma_{i=1}^{k}\Sigma_{j=1}^{p} n_{ij}(x_i - \overline{X})(y_j - \overline{Y})$$

Ou bien

$$Cov(x,y) = [1/n..\Sigma_{i=1}^{k} \Sigma_{j=1}^{p} n_{ij}(x_i y_i)] - [(\overline{x}\,\overline{y})] = \overline{XY} - \overline{X}\,\overline{Y}$$

Ou bien encore

$$Cov(x,y) = \sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij}(x_i - \overline{X})(y_j - \overline{Y}) = \sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij}(x_i y_j)(\overline{X}.\overline{Y})$$
$$= \overline{XY} - \overline{X}\,\overline{Y}$$

Ainsi :

- Si cov(x,y) est positive, cela indique que X et Y varient simultanément et dans le même sens

- Si cov(x,y) est négative, cela signifie que X et Y varient simultanément en sens inverse,

- Si X = Y, cela signifie que cov(x,y) = V(x) = V(y), c'est-à-dire elle est égale à la variance de chacune des variables (Py, 2006).

Exemple 1

Les résultats d'une enquête, en vue d'étudier la relation entre l'âge et les salaires (10^2€) des employés d'une entreprise, sont résumés dans le tableau de contingence suivant :

Salaire(Y) / Age (X)	6 - 10	10 - 14	14 - 18	18 - 22	22 - 26	26 - 30	30 - 34	34 - 38	$\mathbf{n_{i\bullet}}$
22 - 28	5	4	1	0	0	0	0	0	**10**
28 - 34	4	5	8	6	3	2	1	0	**29**
34 - 40	0	2	8	10	9	6	4	2	**41**
40 - 46	1	2	6	8	10	8	6	2	**43**
46 - 52	0	0	2	4	8	6	8	4	**32**
52 - 58	0	0	1	7	10	10	8	4	**40**
$\mathbf{n_{\bullet j}}$	**10**	**13**	**26**	**35**	**40**	**32**	**27**	**12**	**195**

1- Que représente la colonne « $n_{i\bullet}$ » ?

2- Quel est l'âge moyen des employés dans l'entreprise ?

3- Donner les valeurs de $n_{\bullet 5}$; $n_{4\bullet}$; n_{45} ; n_{54} et $n_{\bullet\bullet}$.

4- Calculer et donner les significations de $f_{5\bullet}$; $f_{\bullet 7}$; f_{24} ; $f_{3/4}$ (avec *i* fixé) et $f_{3/4}$ (avec *j* fixé).

5- Dégager la distribution conditionnelle de X selon Y ∈ [14 - 18[et calculer la moyenne correspondante.

6- Dégager la distribution conditionnelle de Y selon X = X_4 . Expliciter et calculer $Y_i = Y_4$.

Réponses

1- La colonne « $n_{i\bullet}$ » représente les effectifs marginaux de la distribution marginale (à un seul caractère) de l'âge ou de X . C'est-à-dire les effectifs correspondants à chaque modalité de X, quelque soit les modalités de Y (ou bien sans tenir compte de Y).

2- L'âge moyen des salariés de l'entreprise est la moyenne marginale de la distribution marginale (à un seul caractère) de X ou de l'âge. Cette distribution est comme suit :

Age	x_i	$n_{i\bullet}$	$n_{i\bullet}x_i$
22 - 28	*25*	*10*	*250*
28 - 34	*31*	*29*	*899*
34 - 40	*37*	*41*	*1517*
40 - 46	*43*	*43*	*1849*
46 - 52	*49*	*32*	*1568*
52 - 58	*55*	*40*	*2200*
Total	-	*195*	*8283*

La moyenne marginale se calcule comme dans le cas des distributions à un caractère.

$\overline{X} = 1/n..\sum_{i=1}^{k} ni.xi = 1/195\ (8283) =$ ***<u>42,48 ans</u>***

3- « $n_{\bullet 5}$ » représente le totale de la dernière ligne, colonne 5. Il est égale, d'après le tableau, à ***40*** *.*

« $n_{4\bullet}$ » représente le total de la quatrième ligne. Il est égale, d'après le tableau, à ***43*** *.*

« n_{45} » représente l'effectif correspondant au croisement de la ligne 4 et de la colonne 5 ; Il est égale, d'après le tableau, à ***10*** *.*

« n_{54} » représente l'effectif correspondant au croisement de la ligne 5 et de la colonne ; Il est égale, d'après le tableau, à ***4*** *.*

« $n_{\bullet\bullet}$ » représente l'effectif total de la population étudié, et représente, à la fois, le total de la dernière ligne et le total de la dernière colonne. Il est égale d'après le tableau à ***195*** *.*

4- $f_{5\bullet} = n_{5\bullet}/n_{\bullet\bullet} = 32/195 =$ ***0,164 = 16,4%*** *. Elle signifie la proportion des salariés de l'entreprise âgés de 46 ans à 52 ans, quelque soit leur salaire.*

$f_{\bullet 7} = n_{\bullet 7}/ n_{\bullet\bullet} = 27/195 = 0,138 =$ ***13,8%*** *. C'est la proportion des salariés de l'entreprise qui touchent entre 3000 à 3400 euros par mois, quelque soit leur âge.*

$f_{24} = n_{24}/n_{\bullet\bullet} = 6/195 = 0,031 =$ ***3,1%*** *. C'est la fréquence partielle sur effectif total. Elle représente la proportion des salariés âgés entre 18 et 22 ans et touchant entre 2800 3400 euros par rapport à l'ensemble des salariés de l'entreprise.*

$f_{3/4}$ (avec i fixé) : i étant constant, on devrait lire donc fréquence conditionnelle de j si i est constant (ou sous condition que i est constant), elle est donc de la forme $f_{j/i}$. Autrement dit, j = 3 et i = 4.

Donc : $f_{3/4}$ (avec i fixé) = $n_{ij}/n_{.j} = n_{43}/n_{\bullet 3} = 6/43 = 0,139 =$ ***13,9%*** *. C'est la proportion des salariés touchant entre 1400 et 1800 euros parmi tous les salariés âgés de 40 à 46 ans.*

$f_{3/4}$ (avec j fixé) : j étant constant, dans ce cas on devrait lire fréquence conditionnelle de i si j est constant (ou sous condition que j est constant), elle est donc de la forme $f_{i/j}$. Autrement dit, i = 3 et j = 4 .

Donc : $f_{i/j} = n_{i/j}/n_{.j} = n_{34}/n_{\bullet 4} = 10/35 = 0,286 =$ ***2,86%*** *. c'est la proportion des salariés âgés de 34 à 40 ans, parmi tous les salariés touchant entre 1800 et 2200 euros.*

5- La distribution conditionnelle de X selon Y ∈[14 - 18[(ou Y= Y_3) se présente comme suit :

Age	x_i	n_{i3}	$n_{i3}\, x_i$
22 - 28	*25*	*1*	*25*
28 - 34	*31*	*8*	*248*
34 - 40	*37*	*8*	*296*
40 - 46	*43*	*6*	*258*
46 - 52	*49*	*2*	*98*
52 - 58	*55*	*1*	*55*
$n_{\bullet 3}$	-	***26***	***980***

La moyenne conditionnelle correspondante :

$\overline{X_j} \equiv \overline{X_3} = 1/n_{.j}\, \Sigma n_{ij}\, x_i = (1/26)\,(980) =$ ***37,69 ans*** *.*

Elle se lit comme l'âge moyen des employés touchant entre 1400 € et 1800 €.

6- La distribution conditionnelle de Y selon X = x_4 ∈ [40 - 46[se présente comme suit :

Salaire	*6 - 10*	*10 - 14*	*14 - 18*	*18 - 22*	*22 - 26*	*26 - 30*	*30 - 34*	*34 - 38*	***n_4.***
Yj	*8*	*12*	*16*	*20*	*24*	*28*	*32*	*36*	*-*
n_{4j}	*1*	*2*	*6*	*8*	*10*	*8*	*6*	*2*	***43***
n_{4j} y_j	*8*	*24*	*96*	*160*	*240*	*224*	*192*	*72*	***1016***

$Y_4 = 1/n_4 . \sum_{j=1}^{p} n4jyj = (1/43)(1016) =$ ***23,63 .10² €*** = ***2363 €.***

Elle se lit comme suit : « le salaire moyen des salariés âgés de 40 à 46 ans est de 2363 € ». C'est aussi la moyenne conditionnelle de Y sous condition ou si X ∈ [40 - 46[ans.

Section 3 : Analyse de la relation entre deux variables

Dans tous les domaines socio-économiques, la recherche de la relation entre deux (ou plusieurs) variables est fondamentale. En statistique, on appelle cette analyse l'***analyse de la régression***. Lorsque celle-ci porte sur deux variables seulement, elle est dite *régression simple*, lorsqu'elle porte sur plus de deux variables, elle est dite *régression multiple*. L'objet de la présente section est la régression simple.

3.1- La régression linéaire simple

Etudier la relation entre deux variables revient à étudier leur plus ou moins dépendance. Autrement dit, la régression simple consiste à expliquer les variations d'une variable, dite variable *dépendante, endogène* ou *expliquée*, généralement notée (y_i), par une autre variable dite *indépendante, autonome, exogène* ou *explicative*, généralement notée (x_i) : c'est-à-dire expliquer les variations de y_i par les variations de x_i (Anderson & al., 2006).

La forme mathématique la plus simple permettant de relier deux variables est la forme linéaire de type « $y = ax+b$ », c'est-à-dire une droite, d'où l'appellation linéaire.

On peut rencontrer trois types de relations linéaires simples :

- la relation linéaire positive, où les deux variables varient en même temps et dans le même sens (*Figure 1)*,

- la relation linéaire négative, où les deux variables varient simultanément mais en sens inverse : l'une augmente pendant que l'autre diminue et vis-versa (*Figure 2)*,

-l'absence de relation, où l'une des variables varie (la variable indépendante) pendant que l'autre demeure constante (*Figure 3)*. (Anderson & al., 2006).

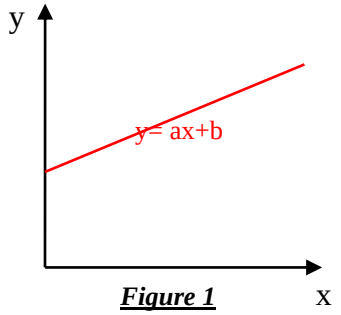


Figure 1

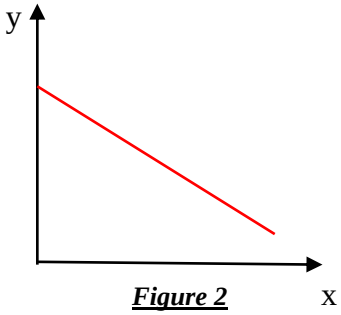


Figure 2

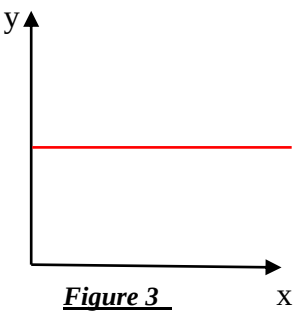


Figure 3

L'***ajustement*** est l'un des moyens d'*estimer* la relation entre deux variables, c'est aussi le moyen le plus utilisé dans l'analyse de la régression. Il existe plusieurs méthodes d'ajustements.

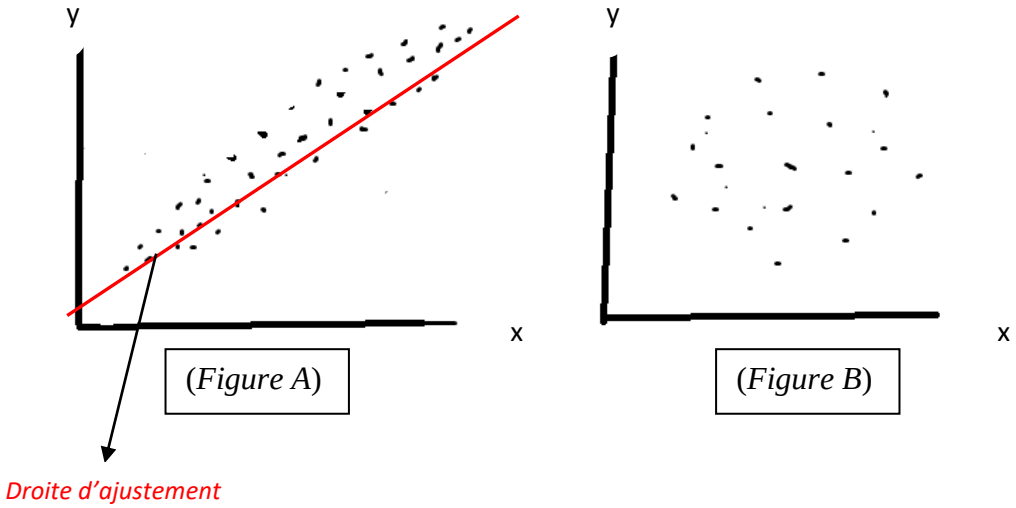
3.2- Les méthodes d'ajustement

Il existe en statistique plusieurs manières d'ajuster les données relatives aux deux caractères étudiés. Nous en exposons dans ce qui suit les plus usitées.

3.2.1- L'ajustement graphique

Une autre manière d'exposer les distributions statistiques à deux caractères, après le tableau de contingence, est le graphique : on l'appelle dans ce cas le *nuage de points*, où les points représentent les couples (x_i ; y_i).

La répartition des points du nuage, ou bien l'allure de ce dernier, est la première étape qui permet de nous indiquer s'il y a (*figure A*) ou pas (*figure B*) relation entre les deux variables. Autrement dit, voir si les deux variables varient globalement dans le même sens et suivent la même tendance dans leurs variations. Lorsque une tendance à la liaison entre les deux variables se dégage, on s'intéressera alors de savoir si la forme de celle-ci peut être linéaire ou pas. C'est-à-dire regarder si le nuage de points *suggère* une droite ou une tendance à la linéarité *(figure A).* Lorsque c'est le cas, on cherchera alors une *droite d'ajustement* (appelée aussi *droite de régression).* (Py, 2007). C'est justement ce cas précis de la relation linéaire qui nous intéresse dans la présente section.



3.2.2- L'ajustement mécanique

Appelé aussi ajustement empirique, il consiste en la *décomposition* de la série des données (xi ; yi) en sous-séries ou groupes. Deux techniques principales sont généralement utilisées pour cela (Hamdani, 2006).

3.2.2.1- Ajustement par les moyennes échelonnées

Cette technique consiste à remplacer des groupes de modalités, d'ordre t (ou de taille t), de la distribution bivariée par leurs médianes, pour la variable X, et par leurs moyennes arithmétiques pour la variable Y. Afin de faciliter la détermination des médianes, on retient un ordre t impair, généralement 3, et on prendra soin d'ordonner les modalités x_i par ordre croissant.

Exemple 2

Soit la distribution bivariée suivante :

Xi	80	85	90	92	98	105
Yi	106	110	124	120	101	102

$\overline{Y}_1$ $\overline{Y}_2$

- Déterminer les moyennes échelonnées d'ordre 3 (ou décomposer, ou ajuster, la série suivante par la méthode des moyennes échelonnées d'ordre 3).

Réponse

$\overline{Y}_1 = (y_1 + y_2 + y_3)/3 = 113,3 \longrightarrow X_2 = 85$ *(Médiane)*

$\overline{Y}_2 = (y_4 + y_5 + y_6)/3 = 107,7 \longrightarrow X_5 = 98$ *(Médiane)*

On obtient ainsi les couples de points (xi ; yi) = {(85 ; 113,3) ; (107,7 ; 98)} par lesquels on peut ajuster le nuage de points.

3.2.2.1- Ajustement par les moyennes mobiles

Cette technique est surtout utilisée dans le cas de séries chronologiques où l'une des variables est le temps. Elle consiste à remplacer la série brute (informations collectées) ou les modalités par une série de moyennes et de médianes. Ce principe rappelle beaucoup celui des moyennes échelonnées. Il s'agit de décomposer la série en sous-séries ou groupes, d'ordre (ou de taille) t. On prendra les médianes pour la variable X et les moyennes pour la variable Y. Afin de faciliter la détermination des médianes, on retient un ordre t impair, généralement 3, et on prendra soin d'ordonner les modalités x_i par ordre croissant (Py, 1996).

La formule des moyennes mobiles est simple :

Si Y = y ; y ; y ;…… ; yt , les moyennes mobiles, d'ordre 3, seront :

$$\overline{Y}_t = (y_{t-1} + y_t + y_{t+1})/3$$

$\overline{Y}_2 = (y_1 + y_2 + y_3)/3$

$\overline{Y}_3 = (y_2 + y_3 + y_4)/3$

Cette méthode fait perdre de l'information aux extrémités de la série, et ce, d'autant plus que le nombre de modalités observées est grand. Cette méthode est particulièrement intéressante pour corriger les variations saisonnières dans les séries chronologiques (Cf, Chapitre 9).

Exemple 3

Reprendre les données de l'exemple 2 et déterminer les moyennes mobiles d'ordre 3 (ou décomposer ou ajuster la série en utilisant la méthode des moyennes mobiles d'ordre 3).

Réponse

$\overline{Y}_2 = (106 + 110 + 124)/3 = 113,3 \longrightarrow x_2 = 85$ *(Médiane)*

$\overline{Y}_3 = (110 + 124 + 120)/3 = 118 \longrightarrow x_3 = 90$ *(Médiane)*

$\overline{Y}_4 = (124 + 120 + 101)/3 = 115 \longrightarrow x_4 = 92$ *(Médiane)*

$\overline{Y}_5 = (120 + 101 + 102)/3 = 107,7 \longrightarrow x_5 = 98$ *(Médiane)*

On aura donc à représenter sur le graphique les quatre points suivants :

$(x_i ; y_i) = \{ (113,3 ; 85) ; (118 . 90) ; (115 ; 92) ; (107,7 ; 98) \}$.

3.2.3- L'ajustement analytique

C'est la méthode la plus utilisée en statistique, en particulier pour des séries non chronologiques. Elle consiste à traiter le nuage de points obtenu par la méthode graphique en l'ajustant par une droite. En effet, lorsque le nuage de points *suggère* une droite ou une tendance à la linéarité *(figure A),* on cherchera alors une *droite d'ajustement* (appelée aussi *droite de régression),* parmi toutes les droites possibles, qui minimise les écarts entre les points situés sur la droite (notés y_c), qui sont des points *estimés,* et les points non situés sur la droite (notés y_i), qui sont des points *observés*. (Anderson & al., 2006). Pour ce faire, on fait recours au principe des *moindres carrés* (Cf, Chapitre 3-Section 3).

Le principe des moindres carrés stipule que la somme des carrés des écarts est minimum. Si on note nos écarts $\boldsymbol{e_i}$, on obtient : $\boldsymbol{e_i} = (y_i - y_c)$, avec ; $\Sigma \boldsymbol{e_i}^2 \longrightarrow$ Min .

L'ajustement analytique consiste donc à construire l'équation mathématique de la droite d'ajustement qui permet de minimiser ces écarts. C'est une équation simple de type y = ax+b permettant de relier les deux variables (xi et yi). Cette équation, construite à partir de données réelles (informations recueillies) d'une période donnée, permettra par la suite d'établir des prévisions pour la période à venir (Anderson & al., 2006).

Il suffit alors de déterminer la valeur des paramètres $\boldsymbol{a}$ et $\boldsymbol{b}$ de l'équation telles que :

$\boldsymbol{e_i} = (y_i - y_c) \longrightarrow$ On pose $\Sigma\, e_i^2 = G$, on aura :

$$G = [y_i - (ax + b)]^2 = \text{Minimum.}$$

Du point de vue mathématique, pour que la fonction G admette un minimum, il faut que sa dérivée première (dérivées partielles par rapport à a et à b) s'annule et que sa dérivée seconde soit positive :

$\frac{\partial' G}{\partial\ a} = 0$ et $\frac{\partial' G}{\partial\ b} = 0$ Avec $\frac{\partial'' G}{\partial a} > 0$ et $\frac{\partial'' G}{\partial b} > 0$

En déterminant ces dérivées partielles et après développement mathématique, on aura deux formules de $\boldsymbol{a}$ (Py, 1996) :

- ***Formule développée***

$$\mathbf{a} = \frac{\overline{XY} - \overline{X}.\overline{Y}}{\overline{X^2} - (\overline{X})^2} = \frac{\mathbf{Cov\ (x\ ;\ y)}}{\mathbf{V(x)}} = \frac{(xi - \overline{X})(yi - \overline{Y})}{\overline{X^2} - (\overline{X})^2}$$

- ***Formule de définition***

$$\mathbf{a} = \frac{\Sigma(xi - \overline{X})(yi - \overline{Y})}{\sum_{i=1}^{k} (xi - \overline{X})^2}$$

Et

. $\mathbf{b = \overline{Y} - a\overline{X}} \quad \Rightarrow \mathbf{\overline{Y} = a\overline{X} + b}$ *(Remarque 1).*

Remarque 1

- La droite d'ajustement passe par le point moyen de coordonnées ($\overline{X}$; $\overline{Y}$), d'où la valeur de ***b***.

- ***a*** est la pente de la droite d'ajustement. S'il est positif, cela signifie que les deux variables varient (croissent) dans le même sens. S'il est négatif, cela signifie que les deux variables varient en sens inverse (l'une décroit pendant que l'autre croît).

- ***b*** est la valeur de Y quand X = 0.

- Une fois les valeurs de a et b déterminées, on peut à partir des valeurs de xi prévoir ou estimer les valeurs de y_c. (Anderson & al., 2006)

Remarque 2

Ce que nous venons de déterminer c'est Y à partir de X avec la droite d'ajustement de Y en X. On peut aussi déterminer X à partir de Y, de même qu'on peut trouver la droite d'ajustement de X en Y. Cela revient à échanger les variables de sorte que X devienne la variable dépendant et Y la variable indépendante. On aura (Py, 1996) :

. $x = a'y + b'$

Avec :

. $\mathbf{a' = \dfrac{Cov\,(x\,;\,y)}{V(y)}}$

Ou bien

. $\boldsymbol{a' = \dfrac{\overline{XY} - \overline{X}.\overline{Y}}{\overline{Y^2} - (\overline{Y})^2}}$

. $\overline{X} = a'\overline{Y} + b'$ $\qquad\qquad$ $\mathbf{b' = \overline{X} - a'\overline{Y}}$

Exemple 4

Les responsables d'une chaîne de restauration rapide, implantée dans plusieurs régions d'un pays, constatent que les restaurants les plus fréquentés sont situés près des campus universitaires. Ils estiment que les ventes (chiffres d'affaires) trimestrielles de ces restaurants (notés yi) sont liées à la taille (nombre d'étudiants) des campus universitaires (notée xi). Ils veulent, par conséquent, établir la relation entre les deux variables.

En mettant à votre disposition les informations recueillies lors de l'enquête auprès d'un échantillon d'une dizaine de ces restaurants (Anderson & al., 2006) :

Restaurant n°	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Taille du campus	*2*	*6*	*8*	*8*	*12*	*16*	*20*	*20*	*22*	*26*
Chiffre d'affaire 10^3\$	*58*	*105*	*88*	*118*	*117*	*137*	*157*	*169*	*149*	*202*

On vous demande de :

1- représenter le nuage de points,

2- déterminer le type de relation suggérée par celui-ci,

3- tracer une droite d'ajustement,

4- déterminer l'équation de la droite d'ajustement de y en x en utilisant le principe des moindres carrés.

5- déterminer l'équation de la droite d'ajustement de x en y.

6- estimer le chiffre d'affaire d'un restaurant situé près d'un campus de 10.000 étudiants.

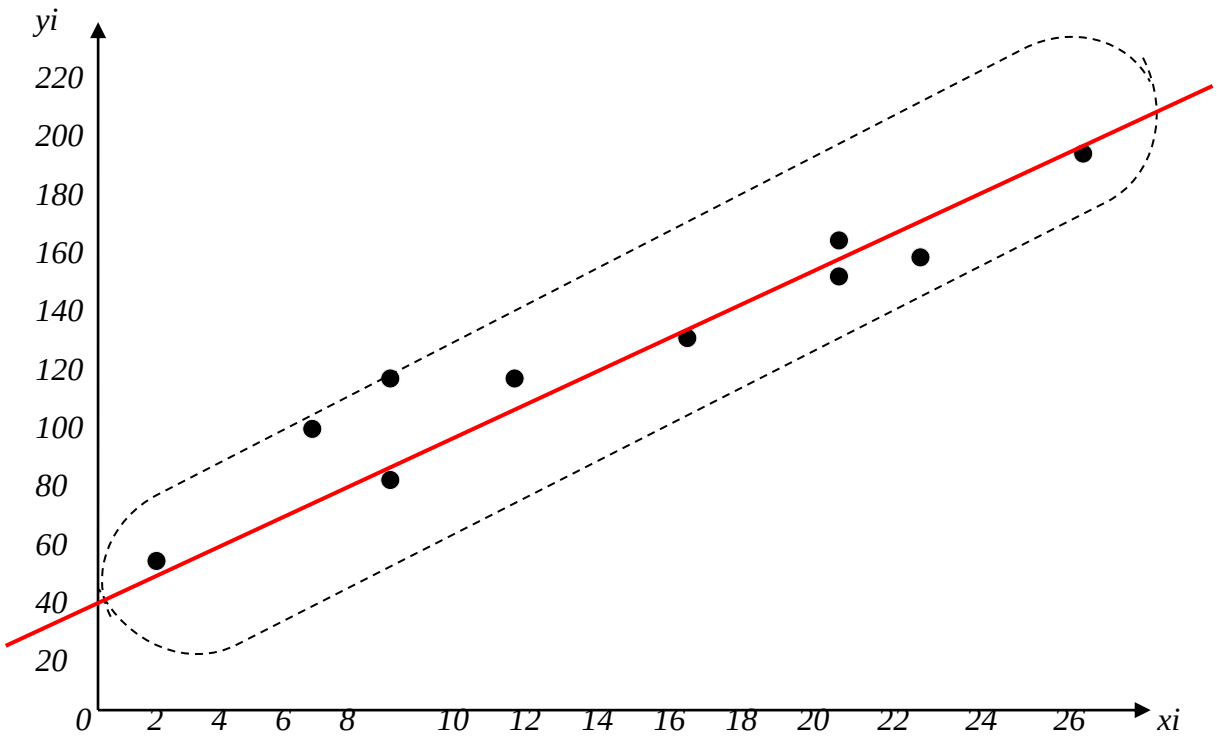
Réponse

Remarque : la première ligne indique le numéro ou l'indentification de chaque restaurant (individuellement), il ne s'agit ni des effectifs, ni des modalités, c'est une codification délibérée (volontaire) qui n'intervient pas dans les calculs. On peut très bien ne pas l'inclure dans le tableau, mais elle donne toutefois plus de lisibilité aux données.

1- Le nuage de points :

Il est présenté sur le graphique ci-dessous.

2- La forme du nuage de points suggère une tendance linéaire à la croissance, ainsi qu'une liaison relative positive entre les ventes des restaurants et la taille des campus.



3- Pour tracer une droite, deux points suffisent. Aussi, on peut tracer plusieurs droites d'ajustement possibles. Nous représentons l'une d'elles en rouge sur le graphique. Mais, attention, cette droite n'est pas la plus parfaite des droites possibles, nous l'avons tracé au hasard en nous basant sur deux points du plan.

4- Pour déterminer l'équation de la droite d'ajustement de y en x en utilisant le principe des moindres carrés, il faut trouver les paramètres ***a*** *et* ***b*** *de la droite d'ajustement Y = aX + b.*
Nous savons que :

$$a = \frac{\overline{XY} - \overline{X}.\overline{Y}}{\overline{X^2} - (\overline{X})^2}$$

et $b = \overline{Y} - a\overline{X}$

Il suffit alors de remplir les colonnes du tableau statistique avec les éléments de la formule dont nous avons besoin. Le tableau est le suivant :

Restaurant n°	*Xi* (10^3)	*Yi* (10^3)	*XY*	Xi^2	Yi^2
1	*2*	*58*	*116*	*4*	*3364*
2	*6*	*105*	*630*	*36*	*11025*
3	*8*	*88*	*704*	*64*	*7744*
4	*8*	*118*	*944*	*64*	*13924*
5	*12*	*117*	*1404*	*144*	*13689*
6	*16*	*137*	*2192*	*256*	*18769*
7	*20*	*157*	*3140*	*400*	*24649*
8	*20*	*169*	*3380*	*400*	*28561*
9	*22*	*149*	*3278*	*484*	*22201*
10	*26*	*202*	*5252*	*676*	*40804*
Total	***140***	***1300***	***21040***	***2528***	***184730***

Nous pouvons directement calculer :

$\overline{X} = 140 / 10 = 14$

$\overline{Y} = 1330 / 10 = 130$

$\overline{XY} = 21040 / 10 = 2104$

$\overline{X^2} = 2528 / 10 = 252{,}8$

$(\overline{X})^2 = (14)^2 = 196$

$\overline{Y^2} = 18473$

$(\overline{Y})^2 = 16900$

En remplaçant dans la formule, on aura :

$$a = \frac{2104 - (14)(130)}{252{,}8 - 196} = 5 \longrightarrow \underline{\boldsymbol{a = 5}}$$

$b = \overline{Y} - a\overline{X} = 130 - 5(14) = 60 \longrightarrow \underline{\boldsymbol{b = 60}}$

Donc la droite d'ajustement s'écrit sous forme de l'équation linéaire suivante :

$$Y = 5X + 60.$$

On l'appelle aussi l'équation « estimée » de Y en X, parce qu'elle permet de faire des estimations (prévisions) de Y à partir de valeurs données de X.
Cette équation permet de tracer la seule droite qui minimise les écarts entre les points (xi ;yi) du plan (les valeurs collectées ou réelles) et les points situés sur la droite tracée (les points estimés).

5- Pour déterminer l'équation de la droite d'ajustement de x en y, il faut inverser le rôle des variables : x devient la variable dépendante et Y la variable indépendante (en termes clairs X devient Y et Y devient X).
L'équation est de type : X = a'Y + b'

Avec :

$$a' = \frac{\overline{XY} - \overline{X}.\overline{Y}}{\overline{Y^2} - (\overline{Y})^2}$$

$$\overline{X} = a'\overline{Y} + b' \longrightarrow b' = \overline{X} - a'\overline{Y}$$

$$a' = \frac{284}{18473 - 16900} = 0,18 \longrightarrow \underline{a' = 0,18}$$

$$b' = 14 - (0,18)130 = -9,4 \longrightarrow \underline{b' = -9,4}$$

donc la droite d'ajustement de X en Y est de la forme :

$$X = 0,18Y - 9,4$$

On remarque que la pente de cette droite (a' = 0,18) est très faible par rapport à la droite d'ajustement de Y en X (a = 5). Cela signifie si les ventes des restaurants (Y) sont expliquées en partie par la taille des campus, la taille des campus est difficilement explicable par les ventes des restaurants Autrement dit, si X permet d'expliquer en partie Y, Y ne permet pas vraiment d'expliquer X, même si dans les deux cas la tendance est à la croissance dans le même sens (positive).
De même, lorsque Y est nulle, X est négatif (- 9,4), ce qui n'a pas de sens sachant que X indique le nombre d'étudiants dans le campus.

6- *Estimer le chiffre d'affaire d'un restaurant situé près d'un campus de 10.000 étudiants. Autrement dit, là il s'agit de prévoir ou d'estimer la valeur de Y pour une valeur donnée de X (x = 10. 10^3). Il faut donc utiliser l'équation d'ajustement de Y en X.*

Y = 5X + 60 $\longrightarrow$ *si X =10 .10^3* $\longrightarrow$ *Y = 5(10) + 60 = 110. 10^3 $.* (Anderson & al., 2006)

3.3- L'analyse de la Corrélation

En général, on rencontre trois situations concernant la relation entre deux variables (Anderson & al., 2006) :

- une *relation totale,* appelée aussi *relation fonctionnelle*, où les variations d'une variable sont expliquées exclusivement et totalement par les variations de l'autre. Dans ce cas les deux variables sont totalement dépendantes. Ce type de relation est très rare en pratique, notamment dans le domaine socio-économique ;

- *absence de relation*, signifiant que les variations d'une variable n'ont aucun effet sur les variations de l'autre. Les deux variables sont dans ce cas indépendantes ;

- une *relation relative,* appelée aussi liaison ou *dépendance partielle*, où les variations d'une variable sont en grande partie, mais non en totalité, expliquées par celles de l'autre. C'est le cas le plus fréquent dans le domaine socio-économique. Une part des variations est laissée au hasard ou aux influences de l'environnement. La dépendance partielle suppose donc une partie certaine importante expliquée par les variations de l'une des variables (y = f(x)), et une partie incertaine ou aléatoire, peu importante et réduite non expliquée par les variations de l'une des variables mais par le hasard, on l'exprime souvent par le symbole *Epsilon* « $\boldsymbol{\varepsilon}$ » pour signifier son caractère négligeable dans la formulation mathématique ($y = f(x) + \varepsilon$).

Dans la relation linéaire simple, on parle d'analyse de la corrélation linéaire simple. Celle-ci mesure le *degré* ou l'*intensité* de la liaison entre deux variables. Elle consiste en la mesure d'un paramètre, appelé *coefficient de corrélation* linéaire de Pearson. (Anderson & al., 2006).

3.3.1- Le coefficient de corrélation

Noté $\boldsymbol{r}$, il est le rapport de la covariance entre X et Y au produit des écarts-types de X et Y.

C'est un nombre sans dimension. On écrit (Bressoud & Kanahé, 2009) :

$$\boldsymbol{r = Cov(x,y) \,/\, \delta_x.\delta_y}$$

On peut également écrire que (Hamdani, 2006) :

$$\boldsymbol{r = \sqrt{a.a'}}$$

ou bien encore (Hamdani, 2006) :

$$\boldsymbol{r = a.(\delta x/\delta y)}$$

$\boldsymbol{r}$ varie entre -1 et 1: $\boldsymbol{r} \in [-1 ;\ 1]$.

Si :

- r =1, les deux variables varient dans le même sens et la liaison est *totale* ou *fonctionnelle*. Cela signifie que la droite de régression (données estimées) s'ajuste parfaitement aux données réelles.
- r = -1, les deux variables varient en sens inverse.
- r = 0, pas de liaison entre les deux variables. La droite de régression est alors une droite de pente a = 0, soit parallèle à l'axe des abscisses, soit parallèle à l'axe des ordonnées : les variations d'une variable n'influence pas celles de l'autre.
- r proche de 0, la relation entre les deux variables est faible.

- r < 0, il existe une corrélation relative et inverse ou négative entre les deux variables, toute variation d'une variable entraine une variation en sens inverse de l'autre variable.
- r > 0, il existe une corrélation *relative* ou *partielle* et positive entre les deux variables, toute variation d'une variable entraine une variation relative, dans le même sens, de l'autre variable. Dans ces deux derniers cas, une grande partie de la variation de Y est expliquée par les variations de X. L'autre partie, infime, est expliquée par le hasard.
- r proche de ±1, la relation est forte. (Anderson & al., 2006).

Remarque 1

Il ne faut cependant pas oublier qu'on ne peut juger de la qualité de l'ajustement, comme on vient de le démontrer, à partir du seul examen du coefficient de corrélation. L'examen graphique et la signification des variables sont des compléments nécessaires au constat établi avec ***r***.

Remarque 2

Au-delà de la mesure de l'intensité de la relation entre les deux variables, on peut aussi

mesurer la partie de Y expliquée par les variations de X. C'est à dire sur la totalité des variations de Y, la part qui est induite par les variations de X, (Py, 2007), soit :

% = variation expliquée / variation totale

Il existe un paramètre qui permet de mesurer cela, on l'appelle le ***coefficient de détermination***, noté « r^2 », et qui est exprimé en pourcentage**.** Il n'est rien d'autre que le carré du coefficient de corrélation. On écrit (Py, 2007) :

$$r^2 = (r)^2 . 100$$

Ainsi, par exemple, si r = 0,6 $\longrightarrow$ $r^2 = 0,36 . 100 = 36\%$. Autrement dit, 36% de la variation totale de Y est due à sa relation avec X ou due à la variation de X. Ou bien encore, que 36% des variations de Y dépendent des variations de X.

On peut également écrire : $r^2 = a.a'$

Exemple 5

Reprendre les données de l'exemple 4 et calculer les coefficients de corrélation et de détermination.

Réponse

$r = Cov(x , y)/\delta_x.\delta_y = \overline{XY} - \bar{X}.\bar{Y})/ \delta_x.\delta_y$

$\delta_x = 7,54$

$\delta_y = 39,66$ $\qquad r = 284 / (7,54).(39,66) = 0,95$

$cov(x,y) = 284$ $\qquad \underline{r = 0,95}$

Ou bien : $r = \sqrt{a.a'} = \sqrt{0,18 . 5} = 0,95.$

r est très proche de 1, donc la corrélation entre les ventes des restaurants et la taille des campus est très forte. Autrement dit, une bonne partie des chiffres d'affaires des restaurants est due à l'effectif des étudiants dans les campus.
*Cependant, on peut mesurer l'importance de l'influence de la taille des campus sur les ventes des restaurants. Autrement dit, la part des chiffres d'affaires induite par la taille des campus à proximité des restaurants. C'est ce que nous permet de savoir le coefficient de détermination (**r^2**) :*

$$r^2 = (r)2 \,.\, 100 = (0,95)^2 = 0,9025 \,.100 = \underline{\mathbf{90,25\%}}$$

Autrement dit, une part de 90,25% des chiffres d'affaires des restaurants serait due à la taille des campus situés à proximité. (Anderson & al., 2006). *On peut dire aussi que les chiffres d'affaires des restaurants dépendent, pour 90,25% , de la taille des campus universitaires situés à proximité.*

Remarque

En réalité il n'est pas tout à fait juste de croire que les chiffres d'affaires des restaurants est à 90,25 % due à la « seule présence » des étudiants dans les campus. D'autres facteurs peuvent entrer en ligne de compte (le niveau de ressources financières des étudiants, leur propension à manger dans les restos, la politique marketing des restos, etc. L'étudiant ne doit donc pas se presser de tirer des conclusions simples et hâtives. Un futur économiste doit toujours avoir cela à l'esprit.

Conclusion au chapitre 9

Au terme du présent chapitre, l'étudiant aura pris connaissance des éléments de base de l'analyse de la relation entre les deux caractères de la distribution bivariée. De même, qu'il se trouvera initié aux techniques de prévision à partir de données présentes réelles.

Par ailleurs, l'étudiant doit comprendre que cette analyse des distributions à deux caractères de manière générale, et de la régression simple de manière particulière, n'est qu'une introduction à l'analyse, encore plus complexe, des distributions à plusieurs caractères ou multivariées (*régression multiple*) qu'il découvrira aux semestres 3 et 4.

Cependant, un type particulier de distribution, très étudié en sciences économiques et de gestion, est nécessaire à aborder en première année, avant de passer à son application directe dans les études de cas en troisième et quatrième années consécutives, ce sont les *séries chronologiques* ou *temporelles*. Celles-ci sont des distributions bivariées où l'une des variables est le temps (chrono), d'où leur appellation. Celles-ci n'étant plus enseignées dans le nouveau programme, nous renvoyons les étudiants désireux de les découvrir de consulter notre autre polycopié élaboré suivant le programme du système classique.

Conclusion générale

A travers la lecteur et l'examen des chapitres du présent cours l'étudiant aura ainsi pris connaissance, de façon assez élargit, des soubassements, notions, formules usuelles, règles et des techniques de la statistique descriptive. Ces connaissances sont indispensables et nécessaires, tant pour la poursuite de son cursus LMD que dans sa future vie professionnelle.

Le cours débute par une introduction au domaine de la statistique descriptive, en exposant les définitions et les notions de base, puis vers la présentation des données avant de passer au calcul des paramètres de mesure des variations des données. Suite à cela on élargit le champs d'analyse en étudiant les nombres indices et, en dernier lieu, les distributions à deux caractère et l'analyse de la relation entre elles. Nous sommes ainsi passé du simple au complexe.

Les chapitres exposés de cette façon, c'est à dire enchaînée et suivant un raisonnement cohérent et logique, répondant largement aux principes de la statistique descriptive, ont pour but de faciliter à l'étudiant l'assimilation des connaissances proposées et de l'imprégner du raisonnement de la statistique qui est, on ne peut plus, rationnel et scientifique. Cela permet d'accompagner l'étudiant dans sa transition de l'enseignement du secondaire à l'enseignement supérieur.

Au terme de ces chapitres, l'étudiant peut passer à l'étape plus complexe de l'analyse statistique, à savoir ; les probabilités et les variables aléatoires. C'est l'objet de l'enseignement dispensé en Statistique 2 au second semestre.

BIBLIOGRAPHIE

(Avec les côtes pour les ouvrages disponibles au niveau de la bibliothèque de la faculté)

1. Anderson, R., Dennis J. Sweeney, D.J., & Williams, T.A. (2006). *Statistiques pour l'économie et la gestion* . 2e Edition. De Boeck, Paris. 779 p.

A / 3664 2°ED

2. Anderson, D.R., Camm, J., Cochran, J.J., Sweeney, D.J, & Williams, T.A. (2015). *Statistiques pour l'économie et la gestion.*5e Edition. De Boeck Supérieur, Paris. 944 p.

3. Bailly, P. (1993). L'Economie et les chiffres : *exercices corrigés de statistique descriptive.* OPU, Alger. 141 p. **B / 1418**

4. Benmessaoud, M., & Oukacha, B. (2008). *Statistiques descriptives et calculs des probabilités : cours et exercices corrigés.* Pages bleues. Alger. 323 p. **A / 4449**

5. Boudia, M C. (2008). *Statistique descriptive*. Casbah, Alger. 315 p.

A / 4427

6. Boukella-Bouzaouane, M. (2001). *Statistique descriptive : rappels de cours avec exercices corrigés*. Casbah. Alger. 171 p. **B / 2089**6.

7. Boursin, J-L. (1991). *Comprendre la statistique descriptive*. AC. Paris. 163 p.

B / 0865

8. Boursin, J-L. (2000). *L'essentiel de la statistique pour l'économie et la gestion*. Gualino. Paris. 127 p.

C / 0972

9. Bressoud, É., & Kahané, J-C. (2009). *Statistique descriptive*. Collection Synthex. Pearson Education France. 258p.

10. Chauvat, G. (1992). *Statistiques descriptives*. Paris : AC. 205 p.

B / 2390

11. Dhuin, C : P*roblèmes corrigés de statistiques : posés aux examens du Deug de sciences économiques (1ère et 2ère année)*. Paris : Ellipses, 192 p. **A / 3369**

12. Dussaix, A-M. (1995). *Statistique pour la gestion*. Alger : Chihab. 340 p.

B / 1495

13. Duthil, G. *Initiation à la statistique descriptive*. Paris : Ellipses. 191p.

A / 0893

14. Grais, B. (1998). *Statistique descriptive avec rappels de cours*. Paris : Dunod. 234 p.

A / 0869

15. Golfarb, B., & Paradoux, C. (2011). *Introduction à la méthode statistique,* 5ème édition. Dunod.

16. Grais, B. (2000). *Statistiques descriptive : Techniques statistique 1*. 3e Ed. Paris : Dunod. 280 p.

A / 0871

17. Hamdani, H. (1988). Statistique descriptive et expression graphique. Alger : OPU. 381 p.

A / 3143

18. Hamdani, H. (2006). *Statistique descriptive avec initiation aux méthodes d'analyse de l'information économique : exercices corrigés.* 5éme éd.. Alger : OPU. 259 p.

A / 3143 5°ED

19. Hubler, J. (2007). Statistique descriptive : appliquée à la gestion et à l'économie. 2ème éd. Paris : Bréal. 219 p. **B / 2346**

20. Hurlin, C., & Mignon, V. (2015). *Statistique et probabilités en économie-gestion*. Dunod. Paris. 370p.

21. Hurlin, C., & Mignon, V. (2018). *Statistique et probabilités en économie-gestion*. Dunod. Paris. 370p.

22. Hurlin, C., & Mignon, V. (2022). *Statistique et probabilités en économie-gestion*. Dunod. 2ème éd°. Collection : Openbook. Paris. 416p.

23. Yves Tillé. (2010). *Résumé du Cours de Statistique Descriptive*. Université de Neuchâtel. https://www.unine.ch/files/live/sites/statistics/files/shared/documents/cours_statistique_descriptive.pdf

24. Janvier, M. (1999). *Statistique descriptive avec ou sans tableur : cours et exercices corrigés*. Paris : Dunod. 276 p. **A / 4026**

25. Labenne, C. (1995). *Introduction à la statistique descriptive et probabilités*. Paris : Economica. 197 p. **A / 0320**

26. Labrousse, C. (1987). Statistique : exercices corrigés avec rappels de cours / Christian Labrousse. - 4°éd.. - Paris : Dunod, 1987. - 292 p. **A / 1904 II**

27. Labrousse, C. (1991). *Statistique : exercices corrigés avec rappels de cours*. 5°éd. Paris : Dunod. 292 p.

28. Lasary, J. (2001). *La statistique descriptive à portée de tous*. Alger : [s.n.]. 188 p. **B / 1570**

29. Lecoutre, J-P. (1990). *Statistique descriptive : exercices corrigés avec rappels de cours*. Paris : Masson, 1990. - 222 p. ; 24 cm **A / 2478**

30. Leboucher, L., & Voisin, M-J. (2011). *Introduction à la statistique descriptive (cours et exercices avec tableur*). Cépaduès-Éditions. 208p. http://livre21.com/LIVREF/F12/F012029.pdf

31. Doane, G.P., & Seward, L. (2016). *Applied Statistics in Business and Economics*, Fifth Edition. MacGraw-Hill. 864 p.

32. Mazerolle, F. (2006). *Statistique descriptive : séries statistiques à une et deux variables; séries chronologiques ; indices.* Paris : Gualino. 172 p. **A / 3823**

33. Monino, J.L., Kosianski, J.M., & Le Cornu, F. (2000). *Statistique descriptive rappels de cours, questions de réflexion, exercices d'entrainement, Annales corrigés*. Paris : Dunod. 248 p. **A / 0870**

34. Monino, J.L. (2017). *TD de statistique descriptive* - 5e éd. Dunod. Paris. 353p.

35. Py, B. (1994). *Exercices corrigés de statistique descriptive : problèmes exercices et Q.C.M.* 2e éd. Paris : Economica. 177 p. **A / 0814**

36. Py, B. (1996). *Statistique descriptive : nouvelle méthode pour bien comprendre et réussir*. 4 ° éd. Paris : Economica. 353 p. **A / 0813 4°ED**

37. Py, B. (2007). *Statistique descriptive : nouvelle méthode pour bien comprendre et réussir* / 5éme éd. Paris : Economica. 353 p. **A / 0813 5°ED**