Programme d'appui au renforcement de la gestion des finances publiques et des statistiques
(PAR-GS)

Formation du personnel non statisticien des ministères sectoriels

Novembre 2015

SUPPORT DE FORMATION EN STATISTIQUE DESCRIPTIVE

Préparé par Boureima OUEDRAOGO
Ingénieur Statisticien Economiste

Sommaire

Introduction

Ce cours est élaboré dans le cadre des formations continues du personnel des structures du système statistique national offertes par le Par-Gs. Elle s'adresse principalement à des professionnels non statisticiens exerçant dans les ministères et impliqués dans la production ou l'exploitation des données statistiques dans le cadre de leurs activités professionnelles.

A l'issue de la formation, les apprenant doivent maîtriser les principales notions de la statistique descriptive, être capables de produire et interprétés des tableaux et des indicateurs de synthèse statistique d'une ou plusieurs série de données. En particulier, le cours aborde les questions suivantes :

- Concepts de base ;
- Typologie des variables ;
- Étude d'une variable qualitative ;
- Étude d'une variable quantitative discrète ;
- Étude d'une variable quantitative continue ;
- Caractéristiques de tendance centrale et utilité ;
- Caractéristiques de dispersion et utilité ;
- Étude d'une série statistique bivariée quantitative ;
- Étude d'une série statistique bivariée qualitative ;
- Notions générales sur les indices et les séries temporelles.

Pour chacune des notions, des exemples concrets seront abordés avec des travaux pratiques sur Excel. Par conséquent, les apprenants doivent être munis chacun d'un ordinateur et avoir au moins des capacités de base à l'utilisation des logiciels bureautiques Excel et Word.

Chapitre 1 : Concepts de base

1. Définition

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

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

2. Objet et utilité de la statistique

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

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

Pour un pays, par exemple, la statistique permet de mesurer des agrégats afin de connaître la situation actuelle d'un phénomène (conjoncture économique), son évolution dans le temps, de prévoir son état futur (prévision des recettes de l'Etat), de comparer des entités, de décider de l'action à mener.

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

- les méthodes descriptives : elles comprennent les statistiques descriptives et l'analyse des données (analyses factorielles et classification). Elles servent à simplifier un ensemble de données (généralement vaste) sans trop perdre d'information par le biais de graphes, de tableaux et de nombres qui résument les données ;
- La statistique mathématique dont l'objet et de formuler les lois à partir d'échantillons et de sous-ensembles d'une population statistique.

3. Définition des concepts usuels de la statistique

3.1. Population et individu

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

Remarque :

- On emploiera les termes population et individu aussi bien lorsqu'il s'agit d'un ensemble d'êtres humains (les salariés d'une entreprise) ou d'objets inanimés ou bien d'un ensemble plus ou moins abstrait comme l'ensemble des accidents de la route au cours d'une période donnée.
- La population étudiée doit être définie de façon précise pour que tous les intervenants qui concourent à l'observation, au traitement, à l'analyse ou à l'utilisation de l'information statistique en aient la même compréhension.

Exemples :

- La population du Burkina Faso au 1er janvier 2015

  Préciser si les burkinabé de l'étranger et les étrangers vivant au BF en font partie.

- Les salariés de l'entreprise X au 31 décembre 2006
- Les étudiants inscrits pour la 1ère fois à l'Université de Ouagadougou en 2014.

3.2. Echantillon/Population mère

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

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

3.3. Variable statistique ou caractère

C'est le critère ou la propriété suivant lequel on étudie la population statistique.

Exemple :

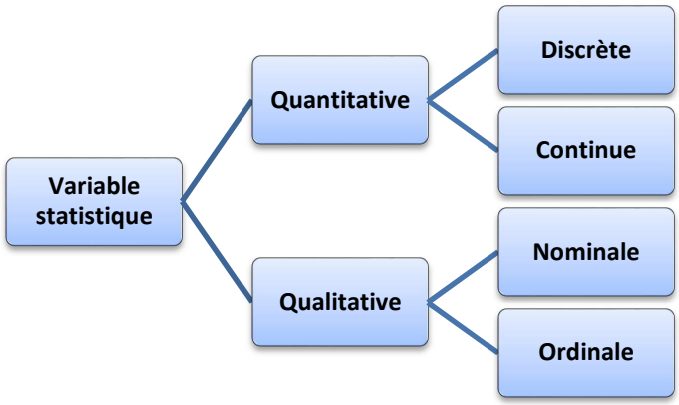
- L'âge des étudiants d'une université
- L'ancienneté des travailleurs d'une société
- La couleur des motocyclettes dans la ville de Ouagadougou
- Le degré d'appréciation d'une mesure gouvernementale par les populations.

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

Exemple : Couleur des yeux : noir, bleu, marron ou vert

La variable statistique peut être qualitative ou quantitative.

3.4. Types de variables statistiques



Variable quantitative : mesurable ou repérable

Exemples : âge, poids, ancienneté, température, taille, nombre d'enfants en charge.

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

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

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

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

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

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

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

Exemple :

Couleurs des yeux : *Bleu/Noir/Vert/Marron*

Situation matrimoniale :

- *Marié/Non marié*
- *Marié/Célibataire/Divorcé/Veuf*

Appréciation d'un cours par les étudiants : *Mauvais/Bon/Très Bon*

Remarque

Les modalités peuvent être représentées par des chiffres qui représentent des codes (codage) et non une mesure.

Exemple

Situation matrimoniale :

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

Variable qualitative nominale : les modalités ne présentent aucun ordre, aucune hiérarchie entre elles.

Exemple : situation matrimoniale, couleur des yeux

Variable qualitative ordinale : les modalités respectent un certain ordre

Exemple :

Appréciation d'un cours : $Mauvais < Bon < Très\ bon$

Catégorie socio professionnelle dans une entreprise :
$Personnel\ de\ soutien; cadre\ moyen; cadre\ supérieur$

4. Elaboration de statistiques

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

L'observation des faits peut se faire de façon instantanée (enquêtes par sondages et recensements) ou de façon continue (enregistrement des naissances à l'état civil, comptabilité d'une entreprise).

4.1. Recensement

C'est une méthode exhaustive, c'est-à-dire que toute la population fait l'objet d'observation suivant le ou les caractères étudiés.

Exemple : recensement de la population du Burkina Faso en décembre 2006 suivant des caractères démographiques (âge, sexe, etc.), économiques (activités économiques), sociaux (niveau d'éducation, alphabétisation, etc.), géographiques (lieu de résidence).

4.2. Enquête par sondage :

Elle porte sur un échantillon.

Exemples :

- Enquête sur les conditions de vie des ménages
- Enquête démographique et de santé
- Sondages d'opinion (CGD)
- Etudes de marché (par sondage)

4.3. Les grandes étapes d'une enquête statistique

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

1. **La conception** : Elle consiste à définir les objectifs de l'étude, définir l'ensemble de l'étude ainsi que les critères à étudier, à concevoir les outils nécessaires à la collecte des informations (questionnaires, guide d'entretien, manuels des agents, etc.). Elle doit également définir les résultats attendus, notamment les indicateurs essentiels à calculer.
2. **La phase de collecte** : Elle comprend la formation des acteurs, la sensibilisation des personnes cibles, l'observation et l'enregistrement de l'information à l'aide de questionnaires. La collecte peut se faire par interview directe, par courrier (poste, e-mail), par téléphone, etc.
3. **La phase de traitement** : Elle consiste à la validation des questionnaires, la codification des réponses, le dépouillement (manuel ou automatique) et le traitement éventuel des données manquantes, des erreurs de saisie, etc.
4. **La phase d'analyse et de diffusion** : Calcul des indicateurs, critique et interprétation des résultats, présentation des résultats obtenus.

5. Critiques de la statistique

A tort ou à raison, plusieurs griefs sont souvent faits à la statistique :

- « La statistique porte sur des faits passés et apporte trop tard ses enseignements »

  Pas toujours vrai puisqu'il existe des méthodes d'observation continue et des méthodes de prévision.

- *« Les statistiques sont fausses »*

  Bien sûr si les bases ont été faussées ou si les méthodes utilisées ne sont pas scientifiquement correctes. C'est pour cela il est nécessaire de comprendre les statistiques pour les interpréter.

- *« Les statistiques aboutissent à des conclusions relatives au comportement d'ensemble et non à celui de l'individu. »*

  C'est précisément l'objet de la statistique

- « *Une des formes les plus raffinées du mensonge.* »

  Nécessité de connaître clairement de quoi il s'agit, les concepts et les méthodes utilisées afin de mieux porter son jugement.

Chapitre 2 : Présentation des données

A l'issue de la collecte des données (lors d'une enquête par exemple), les informations recueillies ne sont pas immédiatement exploitables. Il est alors nécessaire de les organiser, les ordonner et les présenter de façon lisible et facilement compréhensible. Pour cela la statistique descriptive offre des techniques pour la représentation des données sous forme de tableaux ou de graphes.

1. Série statistique à une dimension

1.1. Définition

Une série statistique est la liste des valeurs de la variable statistique observées sur les individus d'un échantillon d'une population donnée. Lorsque plusieurs variables sont simultanément observées sur le même échantillon, la série obtenue sera à 2, 3, ou n dimensions.

1.2. Exemples :

- Série statistique du nombre d'enfants à charge de 20 employés d'une entreprise : 1 ; 0 ; 1 ; 2 ; 2 ; 5 ; 4 ; 4 ; 3 ; 1 ; 0 ; 1 ; 0 ; 0 ; 0 ; 6 ; 10 ; 7 ; 1 ; 7
- Langue maternelle des élèves d'une classe de 15 élèves : Mooré ; Mooré ; Dioula ; Mooré ; Français ; Dafing ; Gourmatché, Foulfouldé ; Foulfouldé ; Mooré ; Dioula ; Dioula ; Mooré ; Mooré ; Mooré.
- Salaire mensuel (en milliers de FCFA) des travailleurs d'une entreprise de 10 personnes : 112,0 ; 100,0 ; 215,2 ; 156,0 ; 100,2 ; 115,0 ; 50,1 ; 62,5 ; 150,0 ; 127,7.
- situation matrimoniale de 40 détenus d'une prison (Marié = 1, Célibataire = 2, Divorcé = 3, veuf = 4) : 1 ; 1 ; 3 ; 1 ; 2 ; 1 ; 2 ; 2 ; 4 ; 3 ; 1 ; 2 ; 2 ; 2 ; 1 ; 2 ; 2 ; 2 ; 2 ; 1 ; 3 ; 1 ; 1 ; 1 ; 4 ; 3 ; 1 ; 1 ; 2 ; 1 ; 2 ; 2 ; 3 ; 1 ; 1 ; 2 ; 4 ; 3 ; 2 ; 2.

2. Tableau de dénombrement

2.1. Définition :

La façon la plus simple de présenter de façon synthétique une série statistique est un tableau présentant en face de chaque modalité le nombre d'individus de l'échantillon qui portent cette modalité. Un tel tableau est appelé tableau de dénombrement.

Effectif : On appelle effectif ou encore fréquence absolue d'une modalité M, le nombre d'individus de l'échantillon qui possèdent cette modalité.

La constitution d'un tableau de dénombrement est immédiate dans le cas des variables qualitatives et des variables quantitatives discrètes. Par contre, dans le cas des variables continues, il existe une infinité (ou un très grand nombre) de modalités. Il est donc nécessaire dans ce cas de transformer les données en les regroupant dans des classes de valeurs (intervalles).

2.2. Exemples

Cas de variable quantitative discrète

Tableau 1 : Nombre d'enfants à charge des employés d'une entreprise

Nombre d'enfants	Effectif
0	5
1	5
2	2
3	1
4	2
6	1
6	1
7	2
10	1
Total	**20**

Cas d'une variable qualitative

Tableau 2 : Situation matrimoniale des détenus

Situation matrimoniale	Code	Effectif
Marié	1	14
Célibataire	2	17
Divorcé	3	6
Veuf	4	3
Total		**40**

Cas d'une variable continue

Tableau 3 : Salaire mensuel des employés de l'entreprise X

Salaire mensuel (en millier de franc cfa)	Effectif
[50 ; 100 [	2
[100 ; 150 [	5
[150 ; 200 [	2
[200 ; 250 [	1
Total	**10**

Remarques :

- Ce tableau indique par exemple que deux employés ont un salaire au moins égal à 50 mille mais inférieur à 100 mille.
- La largeur des classes (ou encore amplitude) est constante et égale à 50.
- La borne inférieure de la distribution (50) et la borne supérieure (250) ont été choisies de sorte que toutes les valeurs observées soient dans l'intervalle [50 ; 250 [
- Les classes sont disjointes (une valeur ne peut être à la fois dans deux classes différentes) et continues (il n'y a pas d'espace entre deux classes successives).

2.3. Choix des classes dans le cas continu

Le choix du nombre de classes et de leur amplitude dépend du domaine de variation de la variable étudiée et de la statistique à établir ; un trop faible nombre de classes peut conduire à des regroupements dans une même classe des mesures observées de la variable qui présenteraient entre elles des écarts sensibles, et en conséquence peut nuire à la précision des résultats obtenus. Inversement un trop grand nombre de classes conduirait sans doute à des résultats précis, mais entraînerait un grand nombre de calculs.

Il est conseillé d'avoir des classes d'amplitudes égales. Cependant, on pourrait être amené à regrouper plusieurs classes lorsqu'elles présentent des effectifs trop faibles ou nuls. Il existe quelques règles empiriques pour le choix optimal du nombre de classes :

- Règle de Sturge :

$$Nombre\ de\ classes\ =\ 1 + (3{,}3 \times\ log\ N)$$

- Règle de Yule :

$$Nombre\ de\ classes\ =\ 2{,}5\sqrt[4]{N}$$

N est la taille de l'échantillon.

L'amplitude de chaque classe (dans le cas où elle est constante) est alors calculée de la manière suivante :

$$a = \frac{X_{max} - X_{min}}{Nombre\ de\ classes}$$

Où X_{max} et X_{min} sont respectivement la valeur maximale et la valeur minimale de la série.

3. Tableaux des fréquences

On considère une série statistique sur un échantillon de taille N. les modalités (ou les classes) de la variable étudiée sont notées X_i et leurs effectifs sont notés N_i. On suppose qu'il existe K modalités (ou classes).

On a

$$\sum_{i=1}^{K} N_i = N$$

C'est-à-dire que la somme des effectifs des modalités (ou des classes) est égale à la taille de l'échantillon.

Le tableau de dénombrement d'une telle série à la forme suivante :

Modalités	**Effectifs**
X_1	N_1
X_2	N_2
.	.
.	.
.	.
X_i	N_i
.	.
.	.
.	.
X_K	N_K
Total	$\boldsymbol{N}$

3.1. Définitions

Fréquence :

La fréquence ou fréquence relative d'une modalité X_i est la proportion d'individus de la population qui présentent cette modalité. On la note f_i.

$$f_i = \frac{N_i}{N}$$

Remarque

- On a la relation suivante :

$$\sum_{i=1}^{N} f_i = \sum_{i=1}^{N} \frac{N_i}{N} = \frac{\sum_{i=1}^{N} N_i}{N} = \frac{N}{N} = 1$$

- La fréquence peut être exprimée en pourcentage

$$f_i(\%) = 100 \times f_i$$

- L'emploi des fréquences s'avère utile pour comparer deux distributions de fréquences établies à partir d'échantillons de tailles différentes.

Fréquence cumulée : la fréquence cumulée ou fréquence cumulée croissante à la modalité X_i est le nombre F_i tel que $F_i = \sum_{i=1}^{K} f_i$

Remarque :

On calcule aussi la fréquence cumulée décroissante par $F_i^{'} = \sum_{p=i}^{K} f_p$

Les fréquences cumulées (croissantes ou décroissantes) permettent de répondre aux questions du type :

- Quelle est la proportion d'individus qui possèdent une valeur inférieure à X_i pour la variable X ?
- Quelle est la proportion d'individus qui possèdent une valeur supérieure à X_i pour la variable X ?

3.2. Tableau des fréquences d'une distribution

Avec les notations ci-dessus, la forme générale d'un tableau de fréquences est la suivante :

Modalités du caractère (X_i)	**Effectifs** (Ni)	**Fréquences** $f_i = \frac{N_i}{N}$	**Fréquences cumulées croissantes**	**Fréquences cumulées décroissantes**
X_1	N_1	f_1	$F_1 = f_1$	$F_1^{'} = f_1 + f_2 + ... + f_K = 1$
X_2	N_2	f_2	$F_2 = f_1 + f_2$	$F_2^{'} = f_2 + f_3 + ... + f_K$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
X_i	N_i	f_i	$F_i = f_1 + f_2 + f_i$	$F_i^{'} = f_i + f_{i+1} + ... + f_K$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
X_K	N_K	f_K	$F_K = f_1 + f_K = 1$	$F_K = f_K$
Total	$\sum_{i=1}^{K} N_i = N$	$\sum_{i=1}^{K} f_i = 1$		

3.3. Exemples :

Tableau 4 : Distribution de la langue maternelle des élèves (voir paragraphe 1)

Langue maternelle	**Effectifs**	**Fréquence**	**Fréquence (%)**
Mooré	7	0,47	46,7
Dioula	3	0,20	20,0
Français	1	0,07	6,7
Dafing	1	0,07	6,7
Gourmantché	1	0,07	6,7
Foulfouldé	2	0,13	13,3
Total	**15**	**1,00**	**100,0**

Tableau 5 : Répartition du nombre d'enfants des salariés d'une entreprise

Nombre d'enfants	**Effectifs**	**Fréquence**	**Fréquence cumulée croissante**	**Fréquence cumulée décroissante**
0	5	0,25	0,25	1,00
1	5	0,25	0,50	0,75
2	2	0,10	0,60	0,50
3	1	0,05	0,65	0,40
4	2	0,10	0,75	0,35
5	1	0,05	0,80	0,25
6	1	0,05	0,85	0,20
7	2	0,01	0,95	0,15
10	1	0,05	1,00	0,05
Total	**20**	**1,00**		

Tableau 6 : Distribution du salaire mensuel des employés de l'entreprise X

Salaire mensuel	**Effectifs**	**Fréquences (%)**	**Fréquences cumulées (%)**
[50 ; 100[	2	20,0	20,0
[100 ; 150[	5	50,0	70,0
[150 ; 200[	2	20,0	90,0
[5200 ; 250[	1	10,0	100,0
Total	**10**	**100,0**	

3.4. Quelques règles de présentation d'un tableau statistique

La présentation d'un tableau statistique doit comporter les éléments suivants :

- le titre du tableau : renseigne sur le contenu du tableau. Il doit être précis et se place au-dessus du tableau ;
- les titres des lignes et des colonnes : doivent être aussi courts que possible pour ne pas encombrer le tableau ;
- les unités de mesure des variables ;
- la source : placée en bas du tableau, elle indique le ou les services qui publient les statistiques contenues dans le tableau.

Quelques règles usuelles de présentation des données à l'intérieur d'un tableau qui facilitent la lecture :

- Utiliser une police de caractères lisible pour les chiffres (exemple Arial) ;
- Aligner les chiffres à droite sans coller à la bordure du tableau ;
- Centrer verticalement les chiffres ;
- Utiliser les séparateurs de milliers pour les chiffres pour les valeurs dépassant 1 000 ;
- Harmoniser le nombre de chiffres après la virgule à l'intérieur de chaque colonne ;
- Limiter le nombre de chiffres après la virgule en fonction du degré de précision requis (en général un ou deux chiffres après la virgule) ;
- Utiliser de préférence un chiffre après la virgule pour les valeurs en pourcentage.

Les tableaux doivent être en pleine page s'ils ont suffisamment de colonnes ou sur la moitié de la page s'ils n'ont que quelques colonnes.

Les colonnes, hors celle de l'intitulé divent avoir une largeur identique.

4. Représentation graphique

La représentation graphique permet de renseigner immédiatement sur l'allure générale de la distribution. Elle facilite l'interprétation des données.

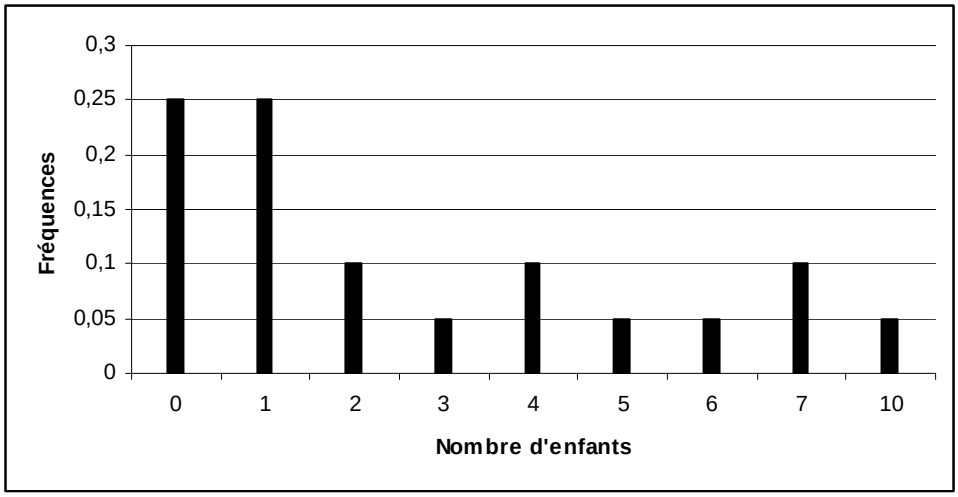
Le type de graphique à utiliser pour représenter une série statistique dépend de la nature discrète ou continue de la variable.

4.1. Le diagramme en bâtons et le diagramme circulaire.

Ils servent à représenter les variables qualitatives et les variables quantitatives discrètes.

Dans le cas du diagramme en bâtons, les modalités de la variable sont représentées par des bâtonnets ou des rectangles (tuyaux d'orgue) dont les hauteurs sont proportionnelles aux effectifs des modalités.

Exemple : Graphique en bâtonnets de la distribution du nombre d'enfants à charge des employés d'une entreprise.

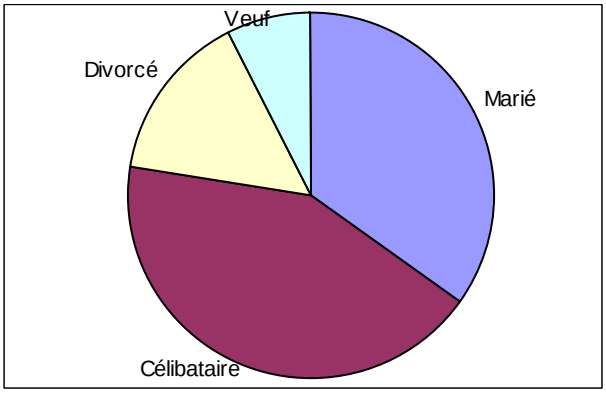


Dans le cas du diagramme circulaire ou par secteurs, chaque modalité est représentée par une portion de disque proportionnelle à l'effectif de la modalité (secteur). Par conséquent chaque secteur a un angle au centre proportionnel à l'effectif de la modalité qu'il représente.

Exemple : Etat matrimonial des détenus d'une prison

Etat matrimonial	Code	Fréquences	Angle (déférés)
Marié	1	0,35	126,0
Célibataire	2	0,43	153,0
Divorcé	3	0,15	54,0
Veuf	4	0,08	27,0
Total		1,00	360,0

Graphique : Représentation par le diagramme circulaire de l'état matrimonial des détenus d'une prison



Remarque

L'angle Ai de chaque modalité se calcule de la façon suivante

$A_i = 360 \times f_i$

Et $\sum A_i = \sum 360 \times f_i = 360 \times \sum f_i = 360 \times 1 = 360$

Remarque :

Le diagramme en secteurs circulaires permet mieux que le diagramme en bâtons de visualiser la part relative de chaque modalité dans l'ensemble de la population.

Pour des comparaisons dans l'espace et dans le temps, la représentation par secteurs permet de rendre sensible à la fois les différences en valeurs absolues et en valeurs relatives.

Exercice :

Comparer les structures de l'emploi par grands secteurs d'activité en France et aux Etats-Unis.

Tableau 7 : Structure de l'emploi civil par grands secteurs d'activités en France et aux Etats-Unis (1985)

Secteurs d'activités	**Etats-Unis**		**France**	
	N_i	f_i	N_i	f_i
Agriculture	3 338	3,1	1 583	7,6
Industrie	30 048	28,0	6 681	32,0
Transport, commerce, service	73 764	68,6	12 626	60,4
TOTAL	**107 150**	**100,0**	**20 890**	**100,0**

Source : OCDE

Tableau de calculs

Secteurs	**Etats-Unis**		**France**	
	f_i	A_i	f_i	A_i
Agriculture	3,1	11,21	7,6	27,36
Industrie	28,0	100,95	32,0	115,20
Transport	68,8	247,83	60,4	217,44
TOTAL	**100,0**	360,00	**100,0**	360,00

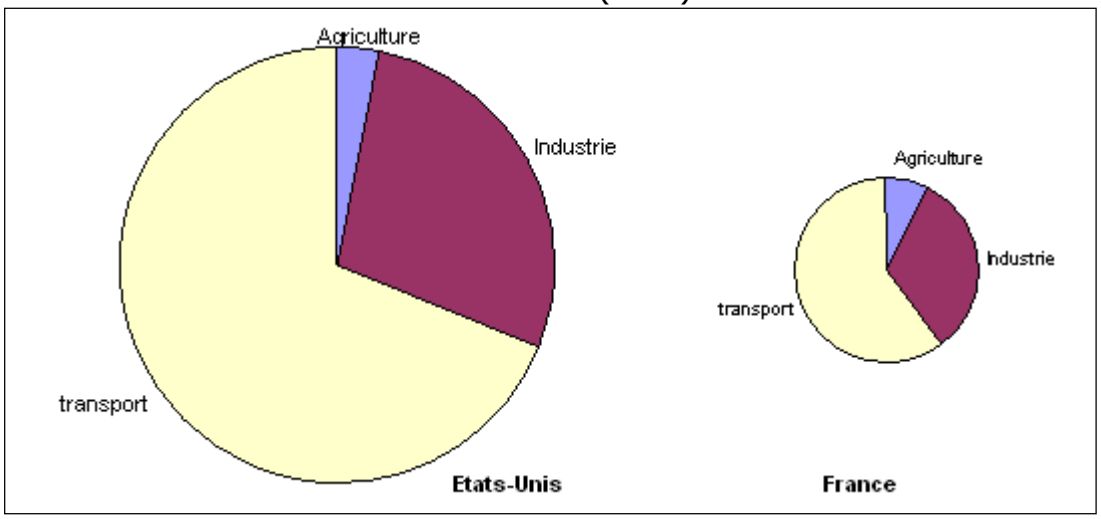
Pour comparer la structure de l'emploi dans les deux pays la situation de chaque pays sera représentée par un diagramme circulaire. Le principe de proportionnalité des superficies des secteurs représentatifs des modalités implique que les superficies des cercles soit également proportionnelles aux valeurs respectives de l'emploi dans les deux pays.

Ainsi, on a :

$$\frac{\pi R_{US}^2}{107150} = \frac{\pi R_{Fr}^2}{20890} \Rightarrow R_{US} = R_{Fr}\sqrt{\frac{107150}{20890}} = 2{,}26R_{Fr}$$

Où R_{US} et R_{Fr} désignent respectivement les rayons des cercles représentant les structures de l'emploi aux Etats-Unis et en France.

Graphique 1 : Structures comparatives de l'emploi civil par grands secteurs d'activités en France et aux Etats-Unis (1985)



Le graphique ci-dessus fait ressortir à la fois les structures internes de l'emploi en France et aux Etats-Unis et permet de comparer les deux structures. On constate : dans les deux cas une forte prédominance du secteur des services suivi de l'industrie et de l'agriculture. Cependant, le secteur de l'agriculture regroupe une plus forte proportion de population en France qu'aux Etats-Unis.

4.2. Histogramme et polygone de fréquences

Ils sont utilisés dans le cas des variables continus.

a. Histogramme

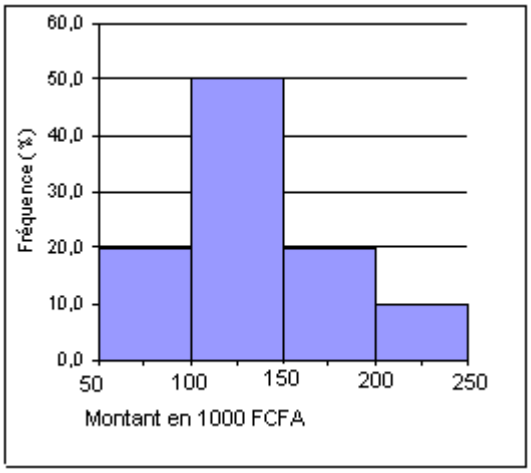
C'est la représentation graphique de la distribution des effectifs ou des fréquences d'une variable statistique continue. A chaque classe de valeurs de la variable portée en abscisse, on fait correspondre un rectangle basé sur cette classe.

Exemple 1 :

Tableau 8 : Salaire mensuel des travailleurs de l'entreprise X en janvier 2008

Salaire mensuel (en milliers de FCFA)	Effectif	Fréquences (%)
[50; 100[	2	20,0
[100; 150[	5	50,0
[150; 200[	2	20,0
[200; 250[	1	10,0
Total	10	100,0

Graphique 2 : Histogramme de la distribution du salaire mensuel des travailleurs de l'entreprise X en janvier 2008



Remarque : Les rectangles de l'histogramme ont des surfaces proportionnelles aux effectifs des classes qu'elles représentent. Dans l'exemple précédent, les classes sont de même amplitude égale à 50. De ce fait, les hauteurs des rectangles sont proportionnelles aux effectifs des classes.

Dans le cas où les classes ne sont pas de même amplitude, les hauteurs des rectangles de l'histogramme ne sont pas proportionnelles aux fréquences des classes mais aux fréquences corrigées f_i^c calculées de la façon suivante : $f_i^c = \frac{f_i}{a_i}$ où a_i est l'amplitude de la classe.

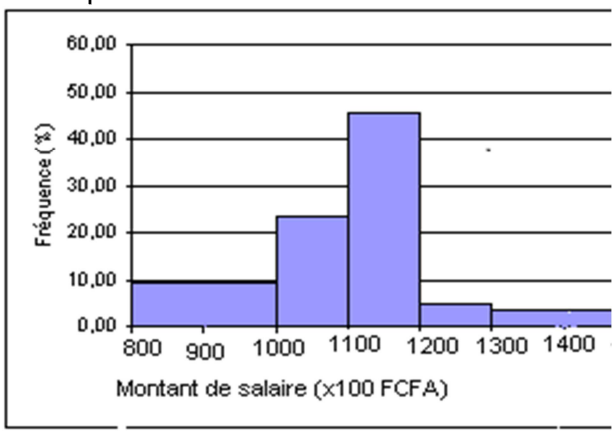
Exemple 2 : Niveau de salaire dans une entreprise

Tableau 9 : Salaires mensuels des employés de l'entreprise Y au 31 décembre 2007

Classe de salaire	n_i	f_i	f_i^c
[800, 1000 [	26	18,57	9,29
[1000, 1100	33	23,57	23,57
[1100, 1200	64	45,71	45,71
[1200, 1300	7	5,00	5,00
[1300, 1500	10	7,14	3,57
TOTAL	140	100,00	

Il y a des amplitudes de **100** et de **200.** Deux effectifs (ou fréquences) de deux classes ne sont comparables directement que si les classes concernées sont de même amplitude. Il faudra donc en tenir compte dans la représentation graphique.

Graphe 3 : Histogramme de la distribution des salaires mensuels des employés de l'entreprise Y au 31 décembre 2007



b. Courbe de fréquence

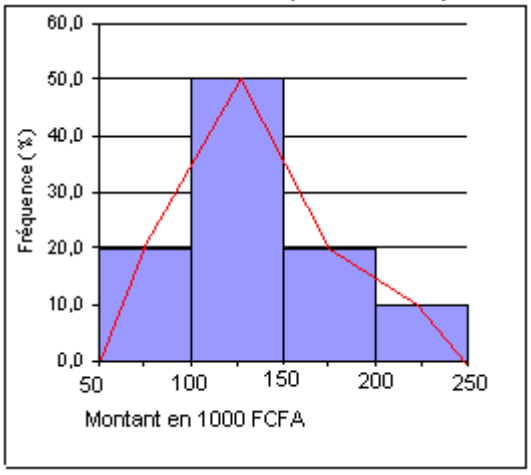
Elle provient de l'idée suivante :

- Si la population étudiée est très nombreuse, l'histogramme ne donne qu'une représentation imparfaite de celle-ci du fait du regroupement des observations en un nombre relativement petit de classes.
- Si l'on divise une première fois, l'amplitude de chaque classe par deux (02), on obtiendrait une représentation plus satisfaisante de la distribution. On peut recommencer l'opération une deuxième, troisième fois, etc. , c'est-à-dire à la limite, l'amplitude des classes tend vers 0 et l'histogramme tend vers une courbe continue appelée *courbe de fréquences ou polygone des fréquences.*

De façon pratique on construit le polygone des fréquences en joignant les milieux des segments des rectangles de l'histogramme.

Exemple

Graphique 4 : Polygone de fréquences de la distribution du salaire mensuel des travailleurs de l'entreprise X en janvier 2008



NB : La surface délimitée par la courbe des fréquences (en rouge sur le graphique) est égale à celle de l'histogramme de la série (surface en bleu).

4.3. Courbe des fréquences cumulées (ou courbe cumulative)

C'est la représentation graphique de la fonction de répartition de la variable statistique. Elle est utilisée dans le cas des variables qualitatives (discrètes et continues).

Soit X une variable statistique. La fonction de répartition de X est l'application

$$F : IR \to [0;1]$$
$$x \mapsto F(x) = P(X < x)$$

Tel que $P(X < x)$ est la proportion d'individus dont la valeur observée de X est inférieure à x.

La courbe cumulative se construit à partir des fréquences cumulées croissantes.

a. Courbe cumulative d'une variable discrète.

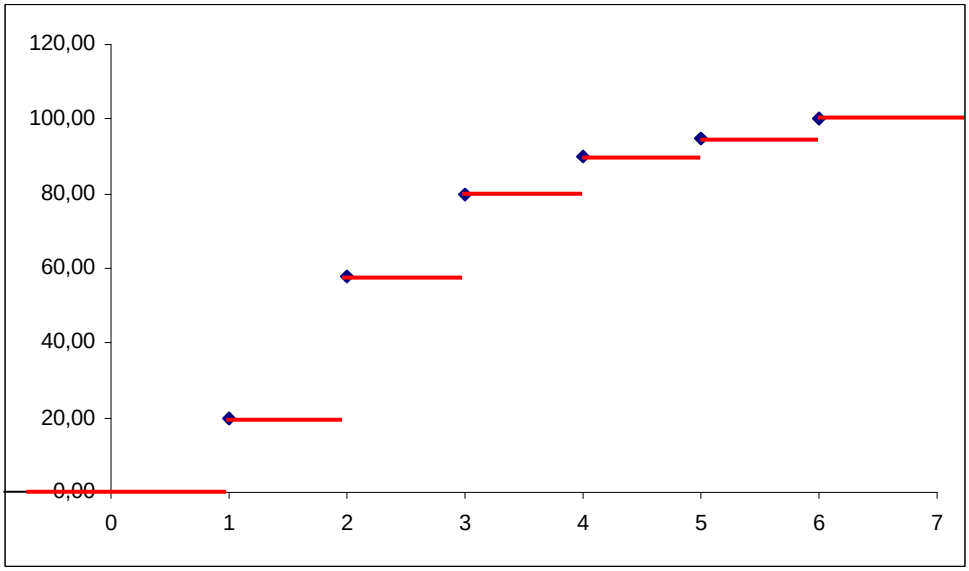
Dans le cas d'une variable discrète, la courbe cumulative se présente comme une courbe en escalier puisque la fonction de répartition F est dans ce cas une fonction constante par intervalles.

Exemple :

Tableau 10 : Répartition des familles des travailleurs d'un groupe industriel selon le nombre d'enfants.

Nombre d'enfants	Nombre de familles	Fréquences	Fréquences cumulées
0	1 390	19,81	19,81
1	2 654	37,82	57,63
2	1 571	22,39	80,02
3	713	10,16	90,18
4	334	4,76	94,94
5 et +	355	5,06	100,00
Total	**7 017**	**100,00**	

Graphe 5 : Diagramme cumulatif du nombre d'enfants pas famille



b. Courbe cumulative d'une variable continue

Dans le cas de la variable continue, la courbe des fréquences cumulées est une courbe continue joignant les points de coordonnées $(e_i ; F_i)$ dans un repère orthogonal où e_i est la borne supérieure de la classe i et F_i est la fréquence cumulée à la classe i.

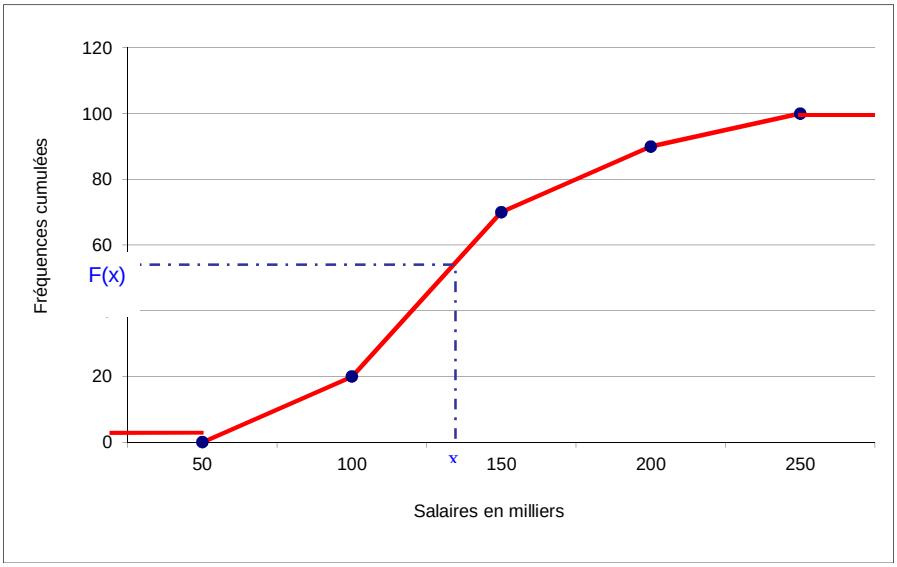
Remarques :

- La courbe cumulative est représentée pour des valeurs de la variable allant de $-\infty$ à $+\infty$.
- Pour les valeurs $x \leq e_{min}$ (borne inférieure de la 1ère classe) on a F(x) = 0, donc la courbe présente une partie constante d'ordonnée = 0
- Pour les valeurs $x \geq e_{max}$ (borne supérieure de la dernière classe) on a F(x) = 1 (ou 100 %). Donc la courbe présente une partie constante d'ordonnée = 1.

Exemple : Distribution des salaires mensuels des employés de l'entreprise X en janvier 2008.

Salaire mensuel (en milliers de FCFA)	Fréquences (%)	Fréquences cumulées croissantes
[50; 100[	20,0	20,0
[100; 150[	50,0	70,0
[150; 200[	20,0	90,0
[200; 250[	10,0	100,0
Total	100,0	

Graphique 6 : Courbe cumulative de la distribution des salaires mensuels des employés



Remarque :

- Le tracé de la courbe cumulative de la variable continue fait l'hypothèse d'une répartition uniforme des individus à l'intérieur des classes.
- La courbe cumulative permet de déterminer graphiquement, pour tout nombre réel x, la proportion d'individus dont la valeur pour la variable X est inférieure à x, (voir graphe ci-dessus).

4.3. Autres types de représentation graphique

- Les cartogrammes
- La pyramide des âges

Chapitre 3 : Caractéristiques de tendance centrale

L'objectif est de résumer à travers quelques indicateurs numériques ou paramètres caractéristiques la distribution d'une variable statistique. On les appelle des indicateurs de synthèse d'une distribution statistique. On utilise des indicateurs de position (ou de tendance centrale), des indicateurs de dispersion et des indicateurs de forme (voir chapitres suivants).

L'analyse numérique et l'analyse graphique d'une distribution sont complémentaires et non exclusives.

Les caractéristiques de tendance centrale sont des valeurs numériques, calculées à partir d'une série (ou d'une distribution) statistique et qui permettent de déterminer la valeur typique ou l'ordre de grandeur de la distribution. Les principales caractéristiques de tendance centrale sont : le mode, la médiane et la moyenne.

1. Le mode

1.1. Définition

Le mode est la valeur la plus fréquente dans une série d'observations. On le note M_o

Dans le cas d'une variable quantitative continue on appelle « **classe modale** » la classe qui présente l'effectif le plus élevé.

Remarque :

Le mode d'une série n'est pas nécessairement unique. Il peut ne pas exister

Exemple 1 : la série $\{1;7;2;4;5;3\}$ n'a pas de mode

Exemple 2 : la série $\{2;1;2;2;3;1;5;4;4;5;4\}$ a deux modes à savoir 2 et 4.

1.2. Cas des variables à modalités isolées (qualitatives et quantitatives discrètes)

Le mode est facile à déterminer dans ce cas à partir d'un tableau des fréquences ou d'un graphique de distribution. C'est la modalité qui présente l'effectif le plus élevé (ou la fréquence la plus élevée).

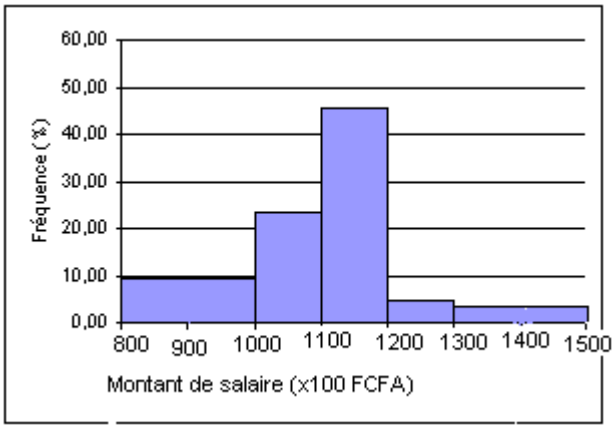
Exemple 1 : Langue maternelle (Exemple 1.2 du chapitre 2). Dans ce cas, le mode est Mooré.

Exemple 2 : Nombre d'enfants des travailleurs d'une entreprise (Exemple 1.2 du chapitre 2) : Il y a deux modes : 0 et 1.

1.3. Cas des données groupées (variables continues)

Lorsque les données sont groupées en classes, on détermine d'abord la classe modale.

Exemple : Salaires mensuels des employés de l'entreprise Y au 31 décembre 2007 (Voir Chap2 ; Paragraphe 4.2 – Exemple2)



Ici, la classe modale est la classe 1100-1200. Sa fréquence est égale à 45,7%.

1.4. Avantages et inconvénients du mode

- La détermination du mode est aisée (graphiquement)
- Son intérêt est évident puisqu'il désigne la valeur de la variable qui est la plus observée sur l'échantillon.
- Le mode n'a de signification véritable que si l'effectif correspondant est nettement supérieur aux effectifs des autres modalités.
- Le mode n'est intéressant que lorsqu'il est unique.

2. La médiane

2.1. Définition

C'est la valeur qui sépare une série d'observations ordonnées en ordre croissant ou décroissant, en deux parties comportant le même nombre d'observations. On la désigne par la notation $\mathbf{M_e}$.

2.2. Méthode de calcul – cas général

- Présenter les données sous forme de série. Lorsque les données sont présentées sous forme de tableau de distribution, les convertir en série.
- Ordonner la série par ordre croissant ou décroissant.
- Déterminer si la série comprend un nombre pair ou impair d'unités statistiques.

Soit N le nombre d'observations :

Cas où N est impair : Dans ce cas la médiane est la valeur qui occupe le rang $\frac{N+1}{2}$ dans la série ordonnée.

Exemple : Série S = 2 ; 4 ; 4 ; 6 ; 7 ; 8 ; 10 ; 10 ; 12

Ici, la médiane est égale à 7.

Cas où N est pair : Dans ce cas la médiane est la moyenne des valeurs de rangs $\frac{N}{2}$ et $\frac{N}{2}+1$

Exemple : S = 0 ; 1 ; 1 ; 2 ; 2 ; 3 ; 3 ; 3 ; 4 ; 5

$$M_e = \frac{2+3}{2} = 2{,}5$$

2.3. Méthode de calcul – cas des données groupées.

Si les données sont groupées par classes (cas des variables continues) il faut :

- localiser la classe médiane, c'est-à-dire celle qui contient la médiane.
- calculer par extrapolation linéaire la valeur de la médiane ;
- ou déterminer la médiane par projection à partir du diagramme des fréquences cumulées.

NB : La classe médiane est celle dont la fréquence cumulée est ≥ 50 % et dont la classe précédente à une fréquence cumulée < 50 %.

Si on note M_e la médiane, e_1 la borne inférieure de la classe médiane, F la fonction de répartition de la variable, et f_{Me} la fréquence de la classe médiane, on a alors $F(e_1)$ est la fréquence cumulée à la classe précédant la classe médiane, $F(e_2)$ la fréquence cumulée à la classe médiane et :

$$M_e = e_1 + \frac{0{,}5 - F(e_1)}{F(e_2) - F(e_1)} \times (e_2 - e_1)$$

Remarque :

- Si les fréquences sont exprimées en % on a :

$$M_e = e_1 + \frac{50 - F(e_1)}{F(e_2) - F(e_1)} \times (e_2 - e_1)$$

- On peut remplacer les fréquences par les effectifs cumulés. Dans ce cas

$$M_e = e_1 + \frac{\frac{N}{2} - N(e_1)}{N_{Me}} \times (e_2 - e_1)$$

Avec N_{Me} = effectif de la classe médiane et $N(e_1)$ = effectif cumulé à la classe précédant la classe médiane.

2.4. Avantages et inconvénients de la médiane

- Son calcul est facile.
- Donne une idée satisfaisante de la tendance centrale de la distribution.
- N'est pas influencée par les valeurs extrêmes de la distribution (valeurs aberrantes).
- La médiane M_e possède la propriété suivante : $\sum_i |x_i - M_e| \leq \sum_i |x_i - x_o|$ Pour toute valeur x_o de la série différente de la médiane.
- Elle ne tient pas compte des valeurs prises par la variable mais seulement de leurs ordres de grandeur.
- Elle concerne uniquement les variables quantitatives.

3. Généralisation de la notion de médiane – Les quantiles.

La médiane est la valeur qui divise la population en deux sous-populations de tailles égales. De la même façon on peut définir des valeurs qui divisent la population en quatre, dix, cent, ... sous-populations de tailles égales. On définit ainsi :

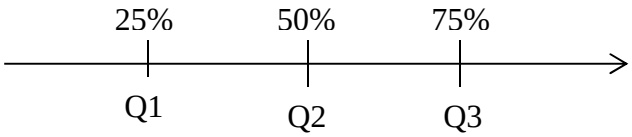
3.1. Les quartiles :

Ce sont les valeurs du caractère qui partagent la série en quatre sous-ensembles de tailles égales. Ils sont au nombre de 3 : Q1, Q2 et Q3

Q_1 : 25 % de valeurs inférieures et 75 % de valeurs supérieures.

Q_2 : 50 % de valeurs inférieures et 50 % de valeurs supérieures, Q2 est la médiane.

Q_3 : 75% des valeurs inférieures et 25% des valeurs supérieures.



3.2. Les quintiles :

Ils divisent la série en cinq sous-ensembles de tailles égales, soit 20 %. Ils sont au nombre de quatre.

3.3. Les déciles :

Ils divisent la série en dix sous-ensembles de tailles égales, soit 10 %.

3.4. Les centiles :

Ils divisent la série en cent sous-ensembles de 1 % de la population.

3.5. Détermination des quantiles.

Les quantiles sont déterminés de la même manière que la médiane par méthode graphique à partir de la courbe des fréquences cumulées ou par extrapolation linéaire (voir cas de la médiane).

Les quartiles sont les valeurs dont les fréquences cumulées sont respectivement :

$^1/_4$; $^2/_4$ et $^3/_4$; C'est-à-dire que : $F(Q_1) = \frac{1}{4} = 25\%$; $F(Q_2) = \frac{2}{4} = 50\%$; $F(Q_3) = \frac{3}{4} = 75\%$

De même :

$F(D_i) = \frac{i}{10}$, pour i = 1, 2,..., 9

$F(C_i) = \frac{i}{100}$, pour i=1,2,..., 99

4. La moyenne arithmétique

4.1. Définition :

La moyenne arithmétique d'un ensemble de données est la somme des valeurs obtenues divisée par le nombre d'observations. Elle est notée $\overline{X}$ pour une variable notée X.

Il existe deux façons courantes de calculer la moyenne arithmétique.

4.2. Moyenne arithmétique simple :

Sa formule est : $\overline{X} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i$ (1)

où les x_i sont les valeurs observées et N est le nombre d'observations ou la taille de la population.

Cette formule est utilisée dans le cas où les données sont présentées sous forme de série.

Exemple : série du nombre d'enfants des employés (voir chapitre 2)

Le nombre moyen d'enfants par employé est

$$\overline{X} = \frac{1+0+1+2+2+5+4+4+3+1+0+1+0+0+0+6+10+7+1+7}{20} = 2{,}75$$

4.3. La moyenne arithmétique pondérée

Sa formule est : $\overline{X} = \frac{\sum_{i=1}^{K} N_i x_i}{\sum_{i=1}^{K} N_i} = \frac{1}{N}\sum_{i=1}^{K} N_i x_i$ (2)

où les x_i sont les modalités (différentes valeurs) de la variable et N_i les effectifs de ces modalités et K le nombre de modalités de la variable.

Remarque :

- Cette formule est intéressante dans le cas où les données sont présentées sous forme d'un tableau de distribution des effectifs (ou des fréquences).
- La formule peut aussi s'écrire de la façon suivante :

$$\overline{X} = \sum_{i=1}^{K} \frac{N_i}{N} x_i = \sum_{i=1}^{K} f_i x$$

où les f_i sont les fréquences des modalités.

- La formule (2) diffère de la formule (1) par le fait que le calcul se fait dans le cas (2) sur les K valeurs distinctes de la variable et non sur les N individus. Les valeurs sont alors pondérées par les effectifs.

Exemple : Série du nombre d'enfants à charge avec tableau des fréquences.

i	**Nombre d'enfants (x_i)**	**Effectifs (N_i)**	**Fréquences (N_i / N)**	$N_i \times x_i$	$f_i \times x_i$
1	0	5	0,25	0	0
2	1	5	0,25	5	0,25
3	2	2	0,10	4	0,20
4	3	1	0,05	3	0,15
5	4	2	0,10	8	0,40
6	5	1	0,05	5	0,25
7	6	1	0,05	6	0,30
8	7	2	0,10	14	0,70
9	10	1	0,05	10	0,50
Total		**20**	**1,00**	**55**	**2,75**

On a bien donc : $\overline{X} = \frac{1}{20}\sum_{=1}^{9} N_i \times x_i = \sum_{i=1}^{9} f_i \times x_i = 2{,}75$

4.4. Calcul de la moyenne dans le cas des données groupées (variables continues)

Dans le cas où les données sont groupées par classes, on fait l'hypothèse que chaque observation à l'intérieur d'une classe a une valeur égale au centre de la classe. Ce qui constitue bien sûr une approximation.

Soit a_i et b_i respectivement les bornes inférieures et supérieures de la classe $N°i$, le centre c_i de la classe est $c_i = \frac{a_i + b_i}{2}$

Dans ce cas on a : $\overline{X} = \frac{\sum_{i=1}^{K} N_i \times c_i}{\sum_{i=1}^{K} N_i} = \frac{1}{N}\sum_{i=1}^{K} N_i \times c_i$

Où K est le nombre de classes et N_i les effectifs des classes.

Remarque :
La moyenne calculée sur les données groupées est généralement différente de la moyenne calculée sur la série initiale non groupée.

Exemple : Soit la série $\{4\ ;0\ ;1\ ;1\ ;2\ ;2\ ;2\ ;3\ ;3\ ;4\ ;2\ ;3\ ;4\ ;5\ ;2\ ;1\ ;3\ ;3\ ;4\ ;5\}$

Le tableau de distribution de la variable étudiée est comme suit :

Valeurs (xi)	0	1	2	3	4	5	**Total**
Effectifs (Ni)	1	3	5	5	4	2	**20**

Sa moyenne est $\overline{X} = \frac{1\times 0 + 1\times 3 + 2\times 5 + 3\times 5 + 4\times 4 + 5\times 2}{20} = 2{,}7$

Si on regroupe les données en classes d'amplitudes égales à 2, on obtient le tableau de distribution suivant :

Valeurs (xi)	[0 ; 2[	[2 ; 4[	[4 ; 6[	**Total**
Centre (Ci)	1	3	5	
Effectifs (Ni)	4	10	6	**20**

La moyenne de cette nouvelle distribution est :

$$\overline{X} = \frac{\sum N_i \times C_i}{N} = \frac{4\times 1 + 10\times 3 + 6\times 5}{20} = 3{,}2$$

Remarque :

Les données groupées ne doivent être utilisées pour les calculs que lorsque les données initiales ne sont pas disponibles.

4.5. Avantages et inconvénients de la moyenne arithmétique

- Du fait qu'elle utilise pour son calcul toutes les valeurs prises par la variable, la moyenne arithmétique est la meilleure des caractéristiques de position.
- La moyenne possède la propriété suivante :

$$\sum_{i=1}^{N}(x_i - \overline{X})^2 \leq \sum_{i=1}^{N}(x_i - x_o)^2 ,$$

pour toute valeur x_o de la série différente de la moyenne.

- La moyenne n'a de sens que pour des variables quantitatives.

- La moyenne arithmétique présente l'inconvénient d'être sensible aux valeurs extrêmes (valeurs aberrantes).

Exemple : soit la série S = {1 ; 1 ; 2 ; 1 ; 1000}

La moyenne de cette série est $\overline{X} = \frac{1+1+2+1+1000}{5} = 201$

Cette valeur est très éloignée de la majorité des observations qui se situent entre 1 et 2.

5. Généralisation de la notion de moyenne

On obtient d'autres types de moyenne en remplaçant dans la formule de la moyenne arithmétique, la variable X par f(X). La formule générale de la moyenne devient :

$$f(X) = \frac{1}{N}\sum_{i=1}^{K} N_i \times f(x_i)$$

5.1. Moyenne géométrique

Elle est utilisée dans le cas d'une variable positive (strictement >0). Sa formule est :

$G = \sqrt[N]{x_1 \times x_2 \times ... \times x_n}$: Moyenne géométrique simple

ou

$G = \sqrt[N]{x_1^{N_1} \times x_2^{N_2} \times ... \times x_K^{N_K}}$: Moyenne géométrique pondérée

Si l'on considère la fonction $f = \ln$, on a :

$$\ln(G) = \ln\left[(x_1 \times x_2 \times ... \times x_N)^{\frac{1}{N}} \right] = \frac{1}{N}\sum_{i=1}^{K} \ln(x_i)$$

Ainsi, la moyenne géométrique d'une série à valeurs strictement positives est le nombre dont le logarithme est égal à la moyenne arithmétique des logarithmes des valeurs de la série.

Remarque :

La moyenne géométrique est utilisée dans le cas des variables positives présentant une évolution géométrique telle que par exemple la population. Elle permet le calcul du taux de croissance moyen, du coefficient multiplicateur moyen. Par exemple, si une variable X croît au cours de N périodes à des taux $t_1;\ t_2;\ \ldots;\ t_n$, alors le taux de croissance moyen annuel est :

$$\bar{t} = -1 + \sqrt[N]{(1+t_1)(1+t_2)\ldots(1+t_N)}$$

Exemple :

Une banque propose à ses clients des taux d'intérêt sur épargne de la façon suivante : 3 % à la 1ère année ; 3,5 % les 2ème et 3ème années et 4 % au-delà de la 3ème année.

Quel est le taux d'intérêt moyen annuel d'un placement au bout de la 6ème année ?

Résolution :

Soit M_o le montant initial placé et M_i.le montant de la somme épargnée au bout de la $i^{ème}$ année et M le montant au bout des 6 ans. Soit $t_1;\ t_2;\ \ldots;\ t_6$, les taux d'intérêt annuels et t_m le taux moyen annuel.

On a :

au bout de la 1ère année : $M_1 = (1+t_1)M_o$
au bout de la 2ème année ; $M_2 = (1+t_2)M_1 = (1+t_2)(1+t_1)M_o$
au bout de la 6ème année : $M = (1+t_1)(1+t_2)\ldots(1+t_6)M_o$
Or $M = (1+t_m)^6 M_o$
Donc $1+t_m = \sqrt[6]{(1+t_1)(1+t_2)\ldots(1+t_6)}$

$$t_m = -1 + \sqrt[6]{1{,}03 \times 1{,}035^2 \times 1{,}04^3} = 3{,}67\%$$

5.2. Moyenne harmonique

$H = \frac{N}{\sum_{i=1}^{N} \frac{1}{x_i}}$: Moyenne arithmétique simple

Ou encore

$H = \frac{N}{\sum_{i=1}^{K} \frac{N_i}{x_i}}$: Moyenne arithmétique pondérée

Si on considère la fonction $f(x) = \frac{1}{x}$ alors la moyenne harmonique d'une série est l'inverse de la moyenne des inverses des valeurs de la série.

Remarque :

- La moyenne harmonique ne peut être calculée que lorsque la série a des valeurs non nulles.

- Elle est utilisée pour le calcul des durées moyennes, des distances moyennes, et de certains ratios.

Exemple 1:

Un cycliste parcourt une distance de 100 km avec les vitesses horaires suivantes : 40km/h les 1ers 25 km puis 30 km/h, 25km/h et 20km/h les 2ème, 3ème et 4ème 25 km suivants.

Quelle est la vitesse moyenne horaire du cycliste ?

Résolution :

Soit T la durée totale de la course, et V la vitesse moyenne horaire

T_i et V_i sont les durées et les vitesses respectives sur le tronçon N°i.

$T_i = \frac{25}{V_i}$

$T = T_1 + T_2 + T_3 + T_4$

$\frac{100}{V} = \frac{25}{V_1} + \frac{25}{V_2} + \frac{25}{V_3} + \frac{25}{V_4}$ d'où $V = \frac{100}{\frac{25}{V_1}+\frac{25}{V_2}+\frac{25}{V_3}+\frac{25}{V_4}}$

Exemple 2 : Les statistiques suivantes ont été observées sur 6 régions :

Population (milliers d'habitants)	250	450	800	150	1 200	600
Nombre d'habitants pour un médecin	1 000	1 500	2 000	1 250	2 500	900

Quel est pour l'ensemble des six villes le nombre moyen de médecin par habitant ?

5.3. Moyenne quadratique

$Q = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2}$: Moyenne quadratique simple

Ou encore

$Q = \sqrt{\frac{1}{N}\sum_{i=1}^{K} N_i x_i^2}$: Moyenne quadratique pondérée

Remarque :

On utilise la moyenne quadratique pour le calcul des écarts quadratiques moyens $Q = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - m)^2}$ où m est une mesure de tendance centrale. Si m est la moyenne, Q est l'écart-type de la série.

5.4. Comparaison des moyennes

On démontre que : $H \leq G \leq \bar{X} \leq Q$ pour une série à valeurs positives non nulles.

Chapitre 4 : Les caractéristiques de dispersion

Elles servent à mesurer la variabilité de la variable statistique et de juger de la pertinence (représentativité) de la caractéristique de tendance centrale.

1. L'étendue

1.1. Définition :

C'est la différence entre la plus grande et la plus petite valeur observée.

Exemple : Dans la série des salaires des travailleurs de l'entreprise (chap2), on a :

$$Etendue = 215\,000 - 50\,100 = 165\,100\ FCFA$$

1.2. Interprétation, avantages et inconvénients

La signification de l'étendue est claire et sa détermination facile. Cependant, elle présente des inconvénients sérieux. En effet, ne dépendant que des valeurs extrêmes qui sont souvent exceptionnelles voire aberrantes et non pas de tous les termes, elle est sujette à des fluctuations considérables d'un échantillon à un autre.

2. Intervalle interquartile

2.1. Définition

C'est la différence entre le 3ème et le 1er quartile. $I_Q = Q_3 - Q_1$

On définit de la même façon l'intervalle inter-décile (I_D=D_9-D_1) et l'intervalle inter-centile (I_C=C_{99}-C_1).

2.2. Interprétation, avantages et inconvénients

L'utilisation de ces intervalles permet d'éliminer l'influence des valeurs extrêmes qui sont des valeurs rares ou aberrantes.

La perte de l'information du fait de la diminution des observations qu'elle entraîne est compensée par l'homogénéité des données dans l'intervalle interquartile.

3. Ecart absolu moyen

3.1. Définition

C'est la moyenne des écarts absolus entre chaque observation et la moyenne.

$EM = \frac{1}{N}\sum_{i=1}^{N}|x_i - \bar{X}|$: Cas simple

Ou encore

$EM = \frac{1}{N}\sum_{i=1}^{N} N_i \times |x_i - \bar{X}|$: Cas pondéré

Remarque :

- On peut aussi calculer l'écart absolu moyen à partir de la médiane

$$EM = \frac{1}{N}\sum_{i=1}^{N}|x_i - M_e|$$

3.2. Interprétation, avantages et inconvénients

L'écart absolu moyen mesure la dispersion des valeurs observées d'une variable statistique autour d'une valeur centrale. Une valeur faible de l'écart absolu moyen traduit une faible dispersion des valeurs autour de la valeur centrale. Cependant la comparaison de cette caractéristique pour deux séries est difficile car sa valeur dépend de l'ordre de grandeur (échelle ou unité de mesure) des observations.

4. Variance et écart-type

4.1. Définition :

La variance est la moyenne des écarts (élevés au carré) des valeurs observées par rapport à la moyenne arithmétique de la série. On la note $V(X)$ pour une variable notée X.

$V(X) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{X})^2$: Cas simple

Ou encore

$V(X) = \frac{1}{N}\sum_{i=1}^{N} N_i \times (x_i - \bar{X})^2$: Cas avec pondération

L'écart-type est la racine carrée de la variance. On le note $\sigma(X)$ ou σ_X Sa formule est :

$$\sigma(X) = \sqrt{V(X} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{X})^2}$$

4.2. Interprétation, avantages et inconvénients

- L'écart-type et la variance mesurent la dispersion de la variable autour de la moyenne. Ainsi, des valeurs élevées (respectivement faibles) de ces caractéristiques traduisent une grande (respectivement faible) dispersion des valeurs autour de la moyenne.
- La variance est calculée à partir des valeurs de la série élevées au carré. Ainsi l'unité (de mesure) de la variance est le carré de celle de la variable. Par exemple, si la variable est mesurée en francs, en kg ou en mètre, la variance sera mesurée en francs au carré, en kg au carré ou en mètres au carré. Par contre l'écart-type a la même unité de mesure que la variable.

4.3. Méthode de calcul

- On calcule d'abord la moyenne arithmétique, puis les écarts entre chaque observation et la moyenne arithmétique.
- On élève les écarts au carré et on somme pour obtenir la variance.
- On extrait la racine carrée de la variance pour obtenir l'écart-type.

Dans le cas où les données sont groupées par classes, on calcule les écarts des centres de classes par rapport à la moyenne.

4.4. Autre méthode de calcul :

On utilise la formule suivante dite théorème de Kœnigs.

$$V(X) = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \bar{X}^2$$

En d'autres termes, la variance est égale à la différence entre la moyenne arithmétique des carrés et le carré de la moyenne arithmétique.

Dans ce cas, on calcule:

- la moyenne arithmétique des observations ;
- la moyenne arithmétique des carrés des observations et ;
- la variance en utilisant la formule de Kœnigs.

4.5. Exercice

Soit le tableau suivant représentant la distribution de la mesure du poids en kg de 100 personnes :

Poids en kg	Effectifs
[58,5 ; 62 ,5[	5
[62,5 ; 65,5[	18
[65,5 ; 68,5[	42
[68,5 ; 74,5[	27
[74,5 ; 80,5[	8
Total	**100**

1. Déterminer la moyenne et la médiane de cette distribution.
2. Calculer l'écart absolu moyen respectivement par rapport à la moyenne et à la médiane.
3. Déterminer l'écart interquartile.
4. Calculer la variance et l'écart-type de la distribution.

5. Les coefficients de variation

5.1. Définition :

Le coefficient de variation de l'écart-type est le rapport entre l'écart-type et la moyenne de la distribution. On le note CV_σ

$$CV_\sigma = \frac{\sigma(X)}{\bar{X}} = \frac{Ecart\ type}{Moyenne}$$

De façon analogue, on définit le coefficient de variation de l'intervalle interquartile par :

$$CV_Q = \frac{I_Q}{M_e} = \frac{Q_3 - Q_1}{Q_2}$$

5.2. Interprétation, avantages et inconvénients

- Contrairement aux autres indicateurs de dispersion, le coefficient de variation est sans unité de mesure. On l'exprime souvent en pourcentage.
- Du fait qu'elle est sans unité, le coefficient de variation présente l'avantage de ne pas être sensible à l'ordre de grandeur (ou à l'unité de mesure) de la variable mais seulement à la dispersion des valeurs autour de la moyenne. Ainsi on peut l'utiliser pour comparer la dispersion de deux séries dont les ordres de grandeur (ou les unités de mesure) sont différents.
- Un coefficient de variation élevé (respectivement faible) traduit une grande (respectivement faible) dispersion de la variable autour de la moyenne.
- L'appréciation du niveau (faible ou élevé) du coefficient de variation est laissée aux soins de l'utilisateur. Cependant une valeur du CV supérieure à 10 % doit susciter des questions quant à la représentativité de la moyenne comme caractéristique de tendance centrale.

Chapitre 5 : Les séries statistiques à deux dimensions

1. Introduction

Pour l'étude de certains phénomènes complexes, il s'avère insuffisant de prendre en compte un seul caractère. Il faut en considérer simultanément deux ou même davantage. Naturellement, l'analyse des tableaux correspondants et leur représentation graphique deviennent plus difficiles. La statistique descriptive à deux dimensions a essentiellement pour but de mettre en évidence les relations qui existent entre deux séries d'observations considérées simultanément. Ces données peuvent être de nature qualitative ou quantitative.

Il sera envisagé dans ce chapitre :

- l'élaboration de tableaux statistiques permettant de condenser les données sous forme de distributions de fréquences à deux dimensions ou distributions conjointes ;
- la représentation graphique des observations ;
- la mesure de la liaison entre deux variables.

2. Présentation générale des tableaux statistiques à double entrée (tableaux croisés)

Les observations relatives à deux variables sur N individus se présentent le plus simplement sous la forme d'une série statistique double, à savoir une suite de N couples de valeurs observées (x_i, y_i).

$$x_1, x_2, \dots, x_n$$
$$y_1, y_2, \dots, y_n$$

Exemple :

Individu (i)	1	2	3	4	5	6	7	8
Sexe (x_i)	1	1	2	2	1	2	1	2
Etat matrimonial (y_i)	2	3	1	1	1	2	1	4

où

- Sexe : 1=Homme ; 2 = Femme
- Etat matrimonial : 1 = Marié ; 2= Célibataire ; 3= Divorcé ; 4 = Veuf

Comme dans le cas des séries à une dimension, il peut être utile, lorsque N est grand de condenser les données en une distribution d'effectifs (ou de fréquences). Celle-ci se présente sous la forme d'un tableau où les modalités x_i de la variable X sont croisées avec les modalités y_i de la variable Y et dont chaque cellule présente l'effectif N_{ij} des individus correspondants à la fois au couple (x_i, y_j).

Exemple :

Sit. Matrimoniale / Sexe	Marié	Célibataire	Divorcé	Veuf	Total
Homme	2	1	1		4
Femme	2	1	0	1	4
Total	4	2	1	1	8

Ce tableau indique par exemple que la population étudiée comprend :

- 4 hommes dont 2 sont mariés, 1 est célibataire et 1 est divorcé
- 4 femmes dont 2 sont mariées, 1 est célibataire et 1 est veuve.

2.1. Définition : distribution conjointe

De façon générale, si on étudie simultanément deux caractères X et Y sur une population de taille N et si X et Y ont respectivement les modalités $x_1, x_2, \dots, x_k$ et $y_1, y_2, \dots, y_l$ alors le tableau de la distribution conjointe (ou tableau croisé) des deux variables se présente de la façon suivante :

X \ Y	y	y_2	…	y_j	…	y_l	Total
x_1	N_{11}	N_{12}		N_{1j}		N_{1l}	$N_{1.}$
x_2	N_{21}	N_{22}		N_{2j}		N_{2l}	$N_{2.}$
.						.	.
.			…		…	.	.
.						.	.
x_i	N_{i1}	N_{i2}		N_{ij}		N_{il}	$N_{i.}$
.	.	.		.		.	.
.	.	.		.		.	.
.	.	.		.		.	.
x_k	N_{k1}	N_{k2}		N_{kj}		N_{kl}	$N_{k.}$
Total	$N_{.1}$	$N_{.2}$	…	$N_{.j}$	…	$N_{.l}$	N

N_{ij} représente l'effectif des individus de la population qui possèdent à la fois la valeur x_i de la variable X et la valeur y_j de la variable Y.

2.2. Notations

$$N_{i.} = N_{i1} + N_{i2} + \cdots + N_{il} = \sum_{j=1}^{l} N_{ij}$$

C'est le total des effectifs de la ligne i ; c'est-à-dire l'effectif total des individus qui possèdent la valeur x_i de la variable X (indépendamment de la valeur de la variable Y).

$$N_{.j} = N_{1j} + N_{2j} + \cdots + N_{kj} = \sum_{i=1}^{k} N_{ij}$$

C'est le total des effectifs de la colonne j ; c'est-à-dire l'effectif total des individus qui possèdent la valeur y_j de la variable Y (indépendamment de la valeur de la variable X).

Remarque : On a :

$$N = \sum_{i=1}^{k}\sum_{j=1}^{l} N_{ij} = \sum_{i=1}^{k} N_{i.} = \sum_{j=1}^{l} N_{.j}$$

2.3. Fréquences (ou pourcentages)

La fréquence du couple (x_i, y_j) est

$$f_{ij} = \frac{N_{ij}}{N}$$

C'est la fréquence du couple (x_i, y_j) observée sur l'ensemble de la population. Elle peut être exprimée en pourcentage.

En adoptant la notation ci-dessus on a :

Fréquence conjointe

$$f_{ij} = \frac{N_{ij}}{N}$$

Total des fréquences conjointes de la ligne i

$$f_{i.} = \frac{N_{i.}}{N} = \sum_{j=1}^{l} f_{ij}$$

Total des fréquences conjointes de la colonne j

$$f_{.j} = \frac{N_{.j}}{N} = \sum_{i=1}^{k} f_{ij}$$

3. Distributions marginales et distributions conditionnelles

3.1. Distributions marginales

Les sommes des effectifs ou des fréquences en lignes définissent la distribution marginale (d'effectifs ou de fréquences) de la variable X. C'est la distribution définie par la colonne « Total » du tableau de distribution conjointe. C'est une distribution à une dimension puisque la variable Y n'intervient pas.

On définit de même la distribution marginale de la variable Y par les sommes des effectifs ou des fréquences par colonne. C'est la distribution définie par la ligne « Total » du tableau de distribution conjointe.

On a donc les distributions marginales suivantes :

Distribution marginale de X

Valeurs de X (x_i)	Effectifs marginaux	Fréquences marginales
x_1	$N_{1.}$	$f_{1.}$
x_2	$N_{2.}$	$f_{2.}$
.	.	.
.	.	.
.	.	.
x_i	$N_{i.}$	$f_{i.}$
.	.	.
.	.	.
.	.	.
x_k	$N_{k.}$	$f_{k.}$
Total	N	**1**

Distribution marginale de Y

Valeurs de Y (y_i)	Effectifs marginaux	Fréquences marginales
y_1	$N_{.1}$	$f_{.1}$
y_2	$N_{.2}$	$f_{.2}$
.	.	.
.	.	.
.	.	.
y_j	$N_{.j}$	$f_{.j}$
.	.	.
.	.	.
.	.	.
y_l	$N_{.l}$	$f_{.l}$
Total	N	**1**

On peut ainsi calculer des caractéristiques de tendance centrale (moyenne, médiane, etc.) ou de dispersion (variance, écart-type, etc.) pour chacune des variables à partir des distributions marginales.

3.2. Distributions conditionnelles

Lorsqu'on ne considère qu'une colonne (colonne j) du tableau de distribution conjointe de X et Y, on obtient une distribution à une dimension appelée distribution conditionnelle ou liée à X sous la condition $Y = y_j$ ou encore distribution de X sachant $Y = y_j$.

On définit la fréquence conditionnelle de x_i sachant y_j par :

$$f_{i/j} = f(x_i/y_j) = \frac{N_{ij}}{N_{.j}} = \frac{f_{ij}}{f_{.j}}$$

Tableau : Distribution conditionnelle de X sachant y_j

Valeurs de X (xi)	Effectifs condition-nels	Fréquences condition-nelles
x_1	N_{1j}	$f_{1/j}$
x_2	N_{2j}	$f_{2/j}$
.	.	.
.	.	.
.	.	.
x_i	N_{ij}	$f_{i/j}$
.	.	.
.	.	.
.	.	.
x_k	N_{kj}	$f_{k/j}$
Total	$N_{.j}$	1

De même on définit la distribution conditionnelle de Y liée à x_i (ou distribution conditionnelle de Y sachant x_i) en considérant la ligne i du tableau de distribution conjointe :

La fréquence conditionnelle de y_j sachant x_i est :

$$f_{j/i} = f(y_j/x_j) = \frac{N_{ij}}{N_{i.}} = \frac{f_{ij}}{f_{i.}}$$

Tableau : Distribution conditionnelle de Y sachant x_j

Valeurs de Y (y_i)	**Effectifs conditionnelles**	**Fréquences conditionnelles**
y_1	N_{i1}	$f_{1/i}$
y_2	N_{i2}	$f_{2/i}$
.	.	.
.	.	.
.	.	.
y_i	N_{ij}	$f_{j/i}$
.	.	.
.	.	.
.	.	.
y_l	N_{il}	$f_{l/i}$
Total	$N_{i.}$	1

3.3. Propriétés des fréquences marginales et conditionnelles

$$f_{ij} = f_{.j} \times f(x_i/y_j) = f_{j.} f_{i/j}$$

$$f_{ij} = f_{i.} \times f(y_j/x_i) = f_{i.} f_{j/i}$$

Cette propriété découle immédiatement des formules qui définissent les fréquences conditionnelles de X et de Y.

3.4. Exemple

Le tableau ci-dessous représente un échantillon de 1000 personnes étudiées suivant les caractères « Sexe » et « Situation matrimoniale ».

Sit. matrimoniale / Sexe	Marié	Célibataire	Divorcé	Veuf	Total
Homme	250	200	100	50	600
Femme	150	150	75	25	400
Total	400	350	175	75	1000

Calculons les fréquences conjointes des deux variables.

Tableau 11 : Répartition (en pourcentage) de la population étudiée selon le sexe et la situation matrimoniale

Sit. matrimoniale / Sexe	Marié	Célibataire	Divorcé	Veuf	Total
Homme	25,0	20,0	10,0	5,0	60,0
Femme	15,0	15,0	7,5	2,5	40,0
Total	40,0	35,0	17,5	7,5	100,0

Ce tableau de fréquences permet de connaître la structure de la population suivant les deux caractères étudiés. On peut lire par exemple que :

- 25% de la population est constituée d'hommes mariés ;
- les femmes représentent 40% de l'effectif total ;
- les femmes veuves constituent 2,5% de l'effectif total ;
- …

On peut aussi calculer les tableaux des fréquences (ou pourcentages) en lignes

Tableau 12 : Répartition (en pourcentage lignes) de la population étudiée selon le sexe et la situation matrimoniale

Sit. matrimoniale / Sexe	Marié	Célibataire	Divorcé	Veuf	Total
Homme	41,7	33,3	16,7	8,3	100,0
Femme	37,5	37,5	18,8	6,3	100,0
Total	40,0	35,0	17,5	7,5	100,0

Le tableau ci-dessus présente les fréquences en lignes ou encore les fréquences conditionnelles de la situation matrimoniale en fonction du sexe. On peut lire par exemple :

- 41,7% des hommes sont mariés contre 37,5% chez les femmes ;
- les célibataires sont proportionnellement plus nombreux chez les femmes que chez les hommes ;
- …

Tableau 13 : Répartition (en pourcentage lignes) de la population étudiée selon le sexe et la situation matrimoniale

Sit. matrimoniale / Sexe	Marié	Célibataire	Divorcé	Veuf	Total
Homme	62,5	57,1	57,1	66,7	60,0
Femme	37,5	42,9	42,9	33,3	40,0
Total	100,0	100,0	100,0	100,0	100,0

Ici, on peut lire les structures, selon le sexe, des sous-populations définies par les situations matrimoniales. C'est le tableau des fréquences conditionnelles de la variable « Sexe » ou tableau des profils colonnes ou encore tableau des pourcentages en colonnes. On peut lire par exemple que :

- les hommes constituent 62,5% des personnes mariées ;
- les femmes constituent 40% de la population totale ;
- …

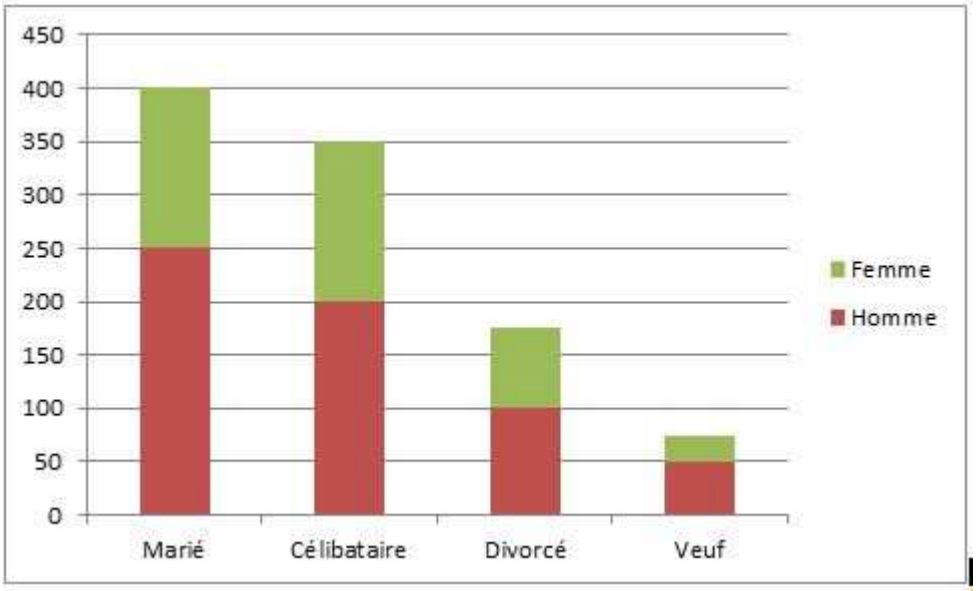
4. Représentation graphique

4.1. Exemple1 : Cas de variables discrètes

Tableau 14 : Répartition d'un échantillon de personnes suivant le sexe et le statut matrimonial

Sit. Matrimoniale / Sexe	Marié	Célibataire	Divorcé	Veuf	Total
Homme	250	200	100	50	600
Femme	150	150	75	25	400
Total	400	350	175	75	1000

Graphique 7 : Représentation de la répartition des effectifs de l'échantillon suivant le sexe et le statut matrimonial

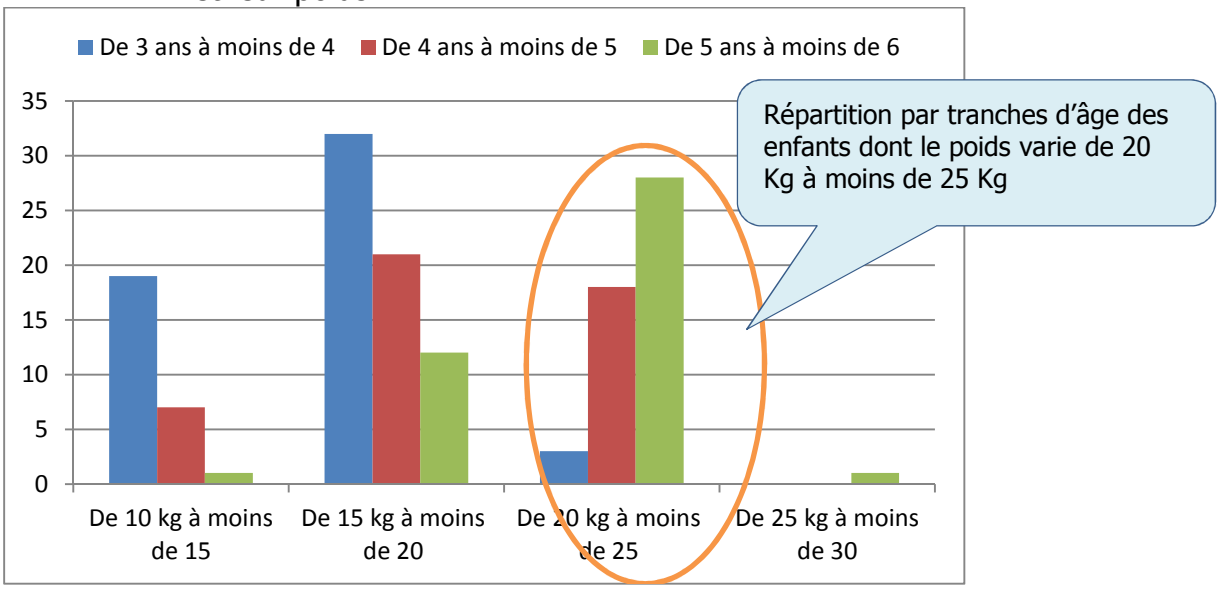


Dans cet exemple, le diagramme est un diagramme en barres (ou tuyeaux d'orgue) compilées d'orgue :

- la distribution marginale par catégorie socioprofessionnelle est représentée par les hauteurs des tuyaux d'orgue ;
- les effectifs N_{ij} (ou les fréquences f_{ij}) sont représentés par les hauteurs des rectangles intérieurs représentant chaque modalité de la variable « sexe ».

4.2. Exemple 2 : Cas où les 2 caractères sont des variables quantitatives

Graphique 8 : Répartition des enfants d'une école maternelle d'après leur âge et leur poids



Le diagramme dans cet exemple est un diagramme en barres groupées. Chaque groupe de barres représente la répartition par tranches d'âge (3 ans à moins de 4 ans ; 4 ans à moins de 5 ans ; 5 ans à moins de 6 ans).

4.3. Autres représentations graphiques

- Nuage de points
- Nuage des points pondérés

5. Mesure de la liaison entre deux variables

L'un des intérêts de l'étude simultanée de deux caractères est l'analyse des variations communes afin de détecter l'existence ou non d'une dépendance.

Les cas suivants peuvent se présenter :

- les variations des deux caractères n'ont aucun lien entre elles. On dira que les deux variables sont indépendantes : Exemple : la taille des élèves et leur moyenne en classe.
- les deux variables sont rigoureusement liées. On parle de liaison fonctionnelle. Exemple: le revenu et la dépense de consommation des ménages.
- les deux variables évoluent globalement dans le même sens (ou en sens contraire) sans être liées rigoureusement. On parle de corrélation positive (ou négative).

Il existe des indicateurs permettant de mesurer le niveau de la relation entre deux variables. Parmi ces indicateurs, on a :

- le khi deux (distance du khi deux)
- la covariance
- le coefficient de corrélation.

5.1. Notion d'indépendance de deux variables.

On dit que deux variable statistiques X et Y sont indépendantes si la réalisation de n'importe quel résultat de X n'influence d'aucune façon celle d'un résultat quelconque pour Y.

En considérant le tableau de contingence de X et Y et les distributions conditionnelles de X et Y, l'indépendance de X et Y se traduit par le résultat suivant : Pour i et j quelconques, la fréquence conditionnelle f_{ij} est égale à la fréquence marginale f_{ij}.

$$f_{i/j} = f_{i.} \Rightarrow f_{ij} = f_{i.}f_{.j}$$

Cette formule découle des propriétés du paragraphe 3.3.

Ainsi, l'indépendance entre les variables X et Y se traduit par le fait que les fréquences conjoints f_{ij} sont les produits des fréquences marginales.

5.2. Notions de covariance et indépendance de deux variables

La covariance de X et Y est le nombre $Cov(X,Y)$ défini par :

$$Cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{X})(y_i - \bar{Y})$$

Après transformation, cette formule s'écrit :

$$Cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N}x_i y_i - \bar{X}\bar{Y}$$

Si les variables X et Y sont indépendantes, alors on aura :

$$\frac{1}{N}\sum_{i=1}^{N}x_i y_i = \left(\frac{1}{N}\sum_{i=1}^{N}x_i\right)\left(\frac{1}{N}\sum_{i=1}^{N}y_i\right) = \bar{X}\bar{Y}$$

et donc $cov(X,Y) = 0$

L'indépendance de X et Y entraîne la nullité de $cov(X,Y)$. Autrement dit, si $cov(X,Y) \neq 0$ alors il existe une certaine dépendance entre X et Y.

Remarque : La nullité de la covariance n'implique pas forcement l'indépendance entre X et Y.

Exemple :

X	-2	-1	0	1	2
Y	2	1	0	1	2

Dans cet exemple on a $Cov(X,Y) = 0$ alors qu'il existe une relation fonctionnelle entre X et Y du type $Y = |X|$

Remarque :

- La covariance permet de déceler seulement les liaisons du type $y = ax + b$ (fonction affine). Par ailleurs, sa valeur est influencée par les unités de mesure des variables X et Y, on lui préfère le coefficient de corrélation linéaire défini par:

$$r_{XY} = \frac{Cov(X,Y)}{V(X)V(Y)}$$

Le coefficient de corrélation linéaire est un nombre compris entre -1 et 1.

- si $r \approx -1$ alors il existe une relation du type $y = ax + b$ avec $a < 0$. entre X et Y.
- si $r \approx 1$ alors il existe une relation du type $y = ax + b$ avec $a > 0$
- Si $0 < r < 1$ alors le nuage de points (x_i, y_j) s'allonge suivant une droite croissante.
- Si $-1 < r < 0$ le nuage de points (x_i, y_j) s'allonge suivant une droite décroissante.
- si $r \approx 0$ le nuage de points (x_i, y_j) n'a pas une forme allongée.

5.3. Distance du khi-deux et indépendance entre 2 variables

En considérant les notations précédentes (voir tableau de contingence et fréquence) on définit la distance du khi-deux par :

$$D_{\chi^2} = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{\left(f_{ij} - f_{i.} f_{.j}\right)^2}{f_{i.} f_{.j}}$$

ou encore:

$$D_{\chi^2} = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{\left(N_{ij} - N_{i.} N_{.j}\right)^2}{N_{i.} N_{.j}}$$

Remarques :

- La distance du Khi-deux vaut 0 si les variables X et Y sont indépendantes. En effet, si X et Y sont indépendantes, $f_{ij} = f_{i.} f_{.j}$

  et donc pour tous i et j

  $$f_{ij} - f_{i.} f_{.j} = 0$$

  La distance du Khi-deux est maximale s'il existe une dépendance systématique entre X et Y.
- Les quantités $f_{ij}^{*} = f_{i.} f_{.j}$ et $N_{ij}^{*} = N_{i.} N_{.j}$ sont appelées respectivement les fréquences et les effectifs théoriques de la distribution conjointe de X et Y. Ce sont les fréquences (ou les effectifs) qu'aurait la distribution conjointe si X et Y étaient indépendantes.
- Ainsi la distance du Khi-deux mesure l'écart entre la distribution empirique (fréquences observées) et la distribution théorique (fréquences théoriques dans le cas d'indépendance de X et Y).
- Pour le calcul du Khi-deux il est conseillé de regrouper les modalités, à voir pour lesquels il existe des effectifs $N_{ij} < 5$.

Chapitre 6 : Indices statistiques

Pour l'étude des phénomènes économiques, on a souvent besoin de décrire les évolutions de grandeurs simples (prix, produit, etc.). Ces valeurs sont à comparer dans le temps et dans l'espace. Les indices permettent ces comparaisons entre grandeurs simples (indices élémentaires) mais, aussi entre grandeurs complexes qui résultent de l'agrégation de composantes de natures diverses et dont le nombre peut être élevé (indice synthétique).

Exemples d'indices :

- l'indice des prix à la consommation (ou indice harmonisé des prix à la consommation des pays de l'UEMOA) ;
- l'indice de la production industrielle ;
- les indices boursiers : Indices BRVM10, BRVM Composite, CAC 40, … ;
- l'indice de confiance des consommateurs ;
- l'indice de développement humain durable (IDH).

1. Les indices élémentaires

1.1. Définition

Soit X_t une variable fonction du temps, l'indice d'évolution de X entre une date 0 et une date t est défini par le rapport $I_{t/0} = X_t/X_0$

Remarque :

Cette valeur est sans dimension.

En général, on exprime l'indice en (%). Dans ce cas on a $I_{t/0}(\%) = 100 \times (X_t/X_0)$

La période de référence est aussi souvent appelée « période base 100 ». La période pour laquelle on effectue le calcul est appelée « période courante ».

Exemple 1 : Le prix du sucre est passé de 350F en 2000 à 500F en 2005. L'indice élémentaire du prix du kg de sucre pour l'année 2005 base 100 en 2000 est : $I_{2005/2000} = 142,85$ soit une augmentation de 42,85 %.

Exemple 2: Le chiffre d'affaires d'une entreprise évolue de la façon suivante au cours du temps :

Année (t)	2011	2012	2013	2014
Chiffre d'affaires (millions de FCFA)	225	300	425	500

Si on considère l'année 2011 comme année de base, on peut mesurer l'évolution du chiffre d'affaires de l'entreprise par rapport à l'année de base. On obtient alors l'indice base 100 en 2011 du chiffre d'affaires.

Année (t)	2011	2012	2013	2014
Indice base 100 en 2011 ($I_{t/2011}$)	100,0	133,3	188,9	222,2

L'indice base 100 en 2011 indique le niveau du chiffre d'affaires de l'année t par rapport à l'année 2011. Ainsi par exemple le chiffre d'affaires de 2012 est 133,3% de celui de 2011, ceci traduit une croissance de 33,3% par rapport à l'année 2011.

1.2. Propriétés des indices élémentaires

Circularité

Si t1 et t2 sont deux dates et X une variable fonction du temps, on a :

$$I_{t2/0} = I_{t2/t1} \times I_{t1/0}$$

Par conséquent, on a :

$$I_{t2/t1} = I_{t2/0}/I_{t1/0}$$

Ce qui permet la comparaison entre deux dates quelconques.

Réversibilité

$$I_{t2/t1} = \frac{1}{I_{t1/t2}}$$

Cette propriété est surtout utile lorsque les comparaisons ne se font pas au cours du temps mais dans l'espace.

Enchaînement

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

On dit que l'on chaîne les évolutions.

Exemple : Variation du prix du sucre

Année	1980	1981	1982
Prix	200	250	400

Calculons : $I_{81/80} = \frac{250}{200} = 125$ $\quad I_{82/81} = 160$ $\quad I_{82/80} = 200$

$I_{82/81} * I_{81/80} = 125 \times 160 = 200$

2. Les indices synthétiques

Les indices considérés jusqu'à présent permettaient de suivre l'évolution des grandeurs simples et parfaitement définies. La plupart du temps, en économie, ce n'est pas l'évolution de grandeurs élémentaires qu'il est intéressant de suivre mais, celle de grandeurs complexes résultant de l'agrégation de plusieurs grandeurs. Les indices synthétiques se proposent de résumer en un seul nombre l'évolution conjuguée de toutes ces valeurs composites.

Par exemple, lorsqu'on veut mesurer l'augmentation du coût de la vie on considère un ensemble de biens de consommation dont on mesure l'évolution des quantités consommées et des prix d'achat. La combinaison de ces prix et de ces quantités permettent de calculer un indice synthétique.

De nombreuses formules d'indices synthétiques ont été proposées, mais seules les plus couramment utilisées sont présentées.

2.1. Indices des moyennes simples

Soit un panier de consommation composé de n biens économiques indicés par i. P_o^i est le prix du bien i à t = 0, P_t^i son prix à la date t. Pour évaluer l'évolution des prix des biens à la date t par rapport à la date 0, l'indice des moyennes simples a été proposé.

$$I_{t/o} = \frac{\frac{1}{n}\sum_{i=1}^{n} P_t^i}{\frac{1}{n}\sum_{i=1}^{n} P_o^i} = \frac{\sum_{i=1}^{n} P_t^i}{\sum_{i=1}^{n} P_o^i}$$

Cet indice est donc l'indice élémentaire des prix moyens.

2.2. Moyenne des indices élémentaires

Les indices élémentaires des prix des biens sont : $\frac{P_t^i}{P_o^i}$. Toujours pour évaluer l'évolution subie entre les dates 0 et t, l'indice moyen des indices élémentaires est défini par :

$$I_{t/o}\left(m_i\right) = \frac{1}{n}\sum \frac{P_t^i}{P_o^i}$$

L'inconvénient de ces deux indices est qu'ils accordent la même importance à tous les biens. Or, dans un panier de consommation, il est évident qu'il y a des biens dont l'achat est beaucoup plus fréquent que d'autres. Ces biens ne sauraient donc être pondérés de la même façon dans le calcul d'un indice synthétique. Aussi, existe-t-il des indices dits pondérés, notamment ceux de Laspeyres, de Paasche et de Fischer.

2.3. Indice de Laspeyres, de Paasche et de Fischer

Soit X la grandeur étudiée. Cette grandeur est complexe et plusieurs constituants indicés par $i = 1,\ldots,n$ interviennent dans sa composition.

Soit X_t: valeur du constituant i à la date t.

Soit X_0: valeur du constituant i à la date 0.

Soit W_t^i: importance du constituant i dans la grandeur X à la date t.

Soit W_0^i: importance du constituant i dans la grandeur X à la date 0.

On a : $\sum_{i=1}^{n} W_o^{\ i} = \sum_{i=1}^{n} W_t^{\ i} = 1$

L'indice de Laspeyres de la grandeur X à la date t par rapport à la date de référence 0 est :

$$L_{t/o}(X) = \sum_{i=1}^{n} W_o^i \frac{X_t^i}{X_o^i} = \sum_{i=1}^{n} W_o^i I_{t/o}^i$$

L'indice de **Laspeyres** est égal à la moyenne arithmétique des indices élémentaires pondérés par les coefficients de l'année de base.

L'indice de **Paasche** est défini par :

$$\frac{1}{P_{t/o}(G)} = \sum_{i=1}^{n} \frac{W_t^i}{X_t^i / X_o^i} = \sum_{i=1}^{n} W_t^i \cdot \frac{1}{I_{t/o}^i(X)}$$

L'indice de **Paasche** est la moyenne harmonique des indices élémentaires pondérés par les coefficients de l'année courante.

L'indice de **Fischer** est la moyenne géométrique simple des indices de **Laspeyres** et de **Paasche**. Il est défini par :

$$F_{t/o}(G) = \sqrt{L_{t/o}(X) \times P_{t/o}(X)}$$

2.4. Propriété des indices de Laaspeyres, de Paasche et de Fischer

- **circularité :** Aucun des trois (03) ne vérifie cette propriété
- **réversibilité** : seul l'indice de Fischer est réversible

On a : $L_{o/t}(X) = \sum_i W_t^i \frac{X_o^i}{X_t^i} = \frac{1}{P_{t/o}(X)}$

D'où

$$F_{o/t}(X) = \sqrt{L_{o/t}(X) \times P_{o/t}(X)} = \frac{1}{F_{t/o}(X)}$$

$$\frac{1}{P_{o/t}} = \sum W_o^i \frac{X_t^i}{X_o^i} = L_{t/o}(X)$$

Chapitre 7 : Séries chronologiques

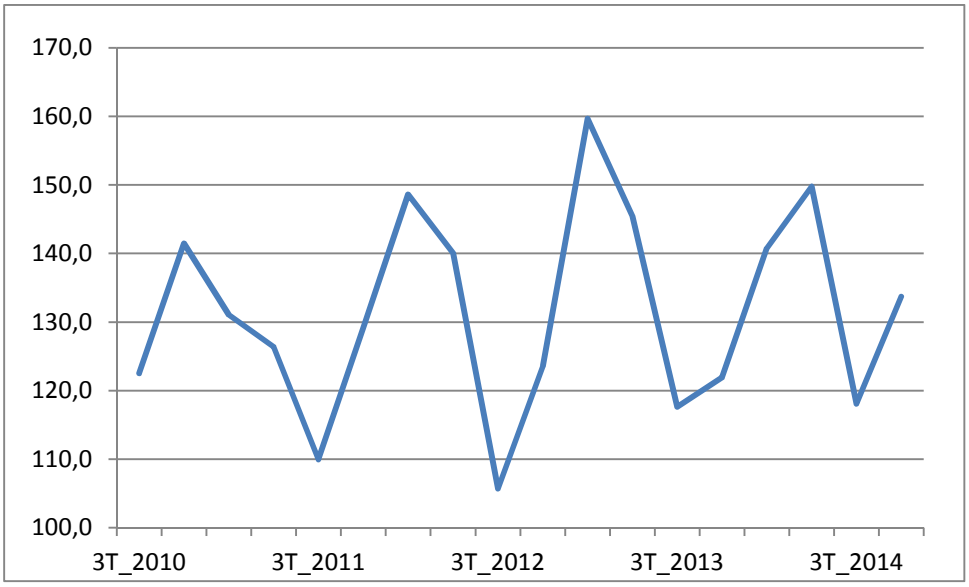
1. Définition

On appelle série chronologique (ou chronique, ou temporelle) une suite d'observations chiffrées d'un même phénomène, ordonnées dans le temps.

C'est une série statistique à deux variables dont une variable est obligatoirement le temps.

Tableau 15 : Evolution de l'indice harmonisé de la production industrielle du Burkina

	2010	2011	2012	2013	2014
Trimestre 1		131,0	148,6	159,7	140,7
Trimestre 2		126,4	140,0	145,4	149,8
Trimestre 3	122,5	109,9	105,7	117,6	118,1
Trimestre 4	141,5	129,1	123,5	121,9	133.7



Remarques :

- On distingue sur le graphique une tendance à l'augmentation de l'indice (c'est la tendance générale)
- On distingue des variations saisonnières (elles représentent les ressemblances entre les différentes périodes)

2. Composantes d'une série chronologique

L'analyse d'une série chronologique permet de distinguer dans l'évolution d'une série :

- une tendance générale (trend) à la hausse ou à la baisse voire constante.
- des variations saisonnières ou mouvements saisonniers qui se répètent chaque année à des moments bien déterminés
- des variations accidentelles ou résiduelles imprévisibles, exceptionnelles (grève, catastrophe naturelle, etc.)

2.1. Tendance notée C_t

La tendance correspond à l'évolution à long terme, l'évolution fondamentale de la série.

Dans l'exemple, l'IHPI augmente de 2010 à 2014.

La tendance est à la hausse (ou haussière), à l'inverse elle serait à la baisse (ou baissière)

2.2. Variations saisonnières S_t

Dans l'exemple, les indices les plus élevés sont au premier trimestre et les plus faibles au troisième trimestre.

Ces variations sont dues au rythme des saisons (production agricole, etc.)

2.3. Variations accidentelles ε_t

Les variations accidentelles sont des fluctuations irrégulières et imprévisibles.

Elles sont supposées en général de faible amplitude.

3. Modélisation d'une série chronologique

Un modèle de série chronologique est une équation précisant la façon dont les composantes s'articulent les unes par rapport aux autres pour constituer la série chronologique.

Il existe deux modèles classiques :

- un modèle additif
- un modèle multiplicatif

3.1. Modèle additif

Dans un modèle additif, on suppose que les trois composantes (tendance, variations saisonnières et variations accidentelles) sont indépendantes les unes des autres.

On considère que la série s'écrit comme la somme de ces trois composantes :

$$Y_t = C_t + S_t + \varepsilon_t$$

Graphiquement, l'amplitude des variations est constante autour de la tendance, la droite qui rejoint les maxima est parallèle (même coefficient directeur) à la droite qui rejoint les minima.

3.2. Modèle multiplicatif

Dans le modèle multiplicatif, on considère que les variations saisonnières dépendent de la tendance (accentuent celle-ci).

La série s'écrit sous la forme :

$$Y_t = C_t \times S_t + \varepsilon_t$$

Graphiquement, l'amplitude des variations saisonnières varie. La droite qui rejoint les maxima n'est pas parallèle (coefficient directeur différent) à la droite qui rejoint les minima.

3.3. Méthode de modélisation

Les principales étapes simplifiées sont les suivantes :

- Faire un graphique
- Identifier le modèle de composition (modèle d'analyse) : additif ou multiplicatif
- Identifier la tendance de la série
- Effectuer le calcul des coefficients de variations saisonnières
- Evaluer la tendance à un moment futur.

Références bibliographiques

Bernard Grais, 2003, Statistique descriptive, Dunod

Boureima Ouedraogo, Module de statistique descriptive, IAPM, 2008

Walder Masiéri 1996, Statistique et calcul des probabilités, éditions Sirey, 7ème édition

Bernard PY, 1987, Statistique descriptive, Economica

Jean Bégin, Résumé de cours de statistique (net)

Fabrice Mazerol, Statistique descriptive, 2008(net)